Aston University
Birmingham

OFSET_*mine* : An Integrated Framework for Cardiovascular Diseases Risk Prediction based on Retinal Vascular Function

Karma Mohamed Gaber FATHALLA

Doctor of Philosophy

Aston University
School of Engineering and Applied Science
ALICE

June 2018

Declaration of Authorship

I, Karma Mohamed Gaber FATHALLA, declare that this thesis titled, "OFSET_*mine* : An Integrated Framework for Cardiovascular Diseases Risk Prediction based on Retinal Vascular Function" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Signed: Karma M.Fathalla
Date: 29 June 2018

Abstract

OFSET_*mine* : An Integrated Framework for Cardiovascular Diseases Risk Prediction based on Retinal Vascular Function

by Karma Mohamed Gaber FATHALLA

As cardiovascular disease (CVD) represents a spectrum of disorders that often manifest for the first time through an acute life-threatening event, early identification of seemingly healthy subjects with various degrees of risk is a priority.

More recently, traditional scores used for early identification of CVD risk are slowly being replaced by more sensitive biomarkers that assess individual, rather than population risks for CVD. Among these, retinal vascular function, as assessed by the retinal vessel analysis method (RVA), has been proven as an accurate reflection of subclinical CVD in groups of participants without overt disease but with certain inherited or acquired risk factors. Furthermore, in order to correctly detect individual risk at an early stage, specialized machine learning methods and feature selection techniques that can cope with the characteristics of the data need to be devised.

The main contribution of this thesis is an integrated framework, OFSET_mine, that combines novel machine learning methods to produce a bespoke solution for Cardiovascular Risk Prediction based on RVA data that is also applicable to other medical datasets with similar characteristics. The three identified essential characteristics are 1) imbalanced dataset, 2) high dimensionality and 3) overlapping feature ranges with the possibility of acquiring new samples. The thesis proposes FiltADASYN as an oversampling method that deals with imbalance, DD_Rank as a feature selection method that handles high dimensionality, and GCO_*mine* as a method for individual-based classification, all three integrated within the OFSET_*mine* framework.

The new oversampling method FiltADASYN extends Adaptive Synthetic Oversampling (ADASYN) with an additional step to filter the generated samples and improve the reliability of the resultant sample set. The feature selection method DD_Rank is based on Restricted Boltzmann Machine (RBM) and ranks features according to their stability and discrimination power. *GCO_mine* is a lazy learning method based on Graph Cut Optimization (GCO), which considers both the local arrangements and the global structure of the data.

OFSET_*mine* compares favourably to well established composite techniques. It exhibits high classification performance when applied to a wide range of benchmark medical datasets with variable sample size, dimensionality and imbalance ratios. When applying OFSET *_mine* on our RVA data, an accuracy of 99.52% is achieved. In addition, using OFSET, the hybrid solution of FiltADASYN and DD_Rank, with Random Forest on our RVA data produces risk group classifications with accuracy 99.68%. This not only reflects the success of the framework but also establishes RVA as a valuable cardiovascular risk predictor.

Keywords: Cardiovascular disease, Retinal Vessel Analysis, feature selection, oversampling, lazy classification.

Acknowledgements

First of all, I would like to express my deepest gratitude to Dr .Aniko Ekart who relentlessly provided the needed support and diligently added value to this work through her constructive critique. Dr.Aniko has set an example to me in dedication, time management and prioritisation. Also, I would like to deeply thank Dr. Doina Gherghel for her much appreciated guidance and innovative ideas which made this work possible.

I am very thankful to Dr Andras Joo for reviewing earlier versions of Chapter 9 and providing very helpful recommendations for improvement. In addition to the research office team at Aston University, especially Sandra Mosley, for being timely responsive and cooperative. Additionally, I would like to express my appreciation to the esteemed examiners (Dr. Dympna O'Sullivan and Dr. Christopher Buckingham) for their comprehensive comments that helped me produce a better structured and narrative thesis.

Thanks are due to the presidency of the Arab Academy for Science and Technology (AAST) for approving my studies financial and academic support. I am also thankful to my professors for understanding my research needs and to my colleagues and staff of the AAST for providing a friendly working environment.

Above all, I show my profound gratefulness to my mother and father for their unconditional encouragement and endless sacrifices. Special thanks go to my sister and her family for being a source of such comfort in my life. Last but not least, I thank my husband for backing me up throughout our journey together.

Contents

List of Figures

List of Tables

List of Abbreviations

CVD	**C**ardio **V**ascular **D**isease
CVR	**C**ardio **V**ascular **R**isk
CAD	**C**oronary **A**rtery **D**isease
RVA	**R**etinal **V**essel **A**nalysis
DVA	**D**ynamic **V**essel **A**nalysis
FRS	**F**ramingham **R**isk **S**core
GA	**G**enetic **A**lgorithms
kNN	**k** **N**earest **N**eighbours
ML	**M**achine **L**earning
MLP	**M**ulti **L**ayer **Perceptron**
NB	**N**aive **B**ayes
RF	**R**andom **F**orest

To my beloved AYSEL and SELIM
and
my grand father, dearest ABD EL AZIZ

Chapter 1

Introduction

Cardiovascular diseases (CVDs) can be defined as a group of disorders of the heart and blood vessels [207]. CVDs often manifest for the first time as acute life threatening events (heart attacks and strokes). CVDs related deaths accounted for 31% of the total global deaths in 2016 [207], where 85% of these were due to heart attacks and strokes. In Europe (2017), CVDs related deaths represented 45% of the total deaths (37% of which are in the EU) [203]. As for the UK in 2018, one death every three minutes can be attributed to CVD [35].

Besides, CVDs are one of the leading diseases in creating burden on populations. The burden is quantified using Disability-Adjusted Life Year (DALY) metric, where DALY is defined as the sum of years of potential life lost due to premature mortality and the years of productive life lost due to disability. CVDs account for the loss of more than 64 million DALYs in Europe (23% of all DALYs lost) and 26 million DALYs in the EU (19%) [203].

According to the World Health Organisation (WHO) [206, 207], 80% of premature heart disease and stroke is preventable through early detection and effective management. The development of most cardiovascular diseases can be stopped by altering some modifiable risk factors related to life style. These factors include tobacco use, unhealthy diet, physical inactivity and excessive use of alcohol. Hence, early prediction of cardiovascular risk is needed due to its anticipated individual and population benefits. An example of these benefits can be seen where population wide interventions in several countries have led to considerable reduction in CVD related deaths [206].

1.1 Cardiovascular Disease Risk Prediction

Several CVD risk prediction models [43, 49, 87] were developed to predict risk. Although these models are useful in estimating long term risk, the learned models are not readily transferable from one population to another. They tend to overestimate or underestimate the risk when applied to a population different than the original population used in constructing the model [12] due to differences in influential factors such as ethnicity and socioeconomic factors.

Cardiovascular risk prediction receives substantial attention [9, 108, 152, 159] due to the threats imposed by cardiovascular diseases on human life [142]. Despite the dedicated efforts, a need still exists for establishing new individual based risk markers to more accurately predict cardiovascular risk in an early stage. New risk markers that directly relate to and convey vascular health are required. In addition, the acquisition of the new risk markers would preferably be non-invasive to encourage individuals to undergo the examinations needed for early CVD risk prediction.

According to Hlatky et al [88], the phases of establishing a novel marker for cardiovascular risk prediction can be summarised as: a) Test whether it can separate risk groups; b) Validate that it predicts the development of hard outcomes; c) Assess its incremental predictive value over the established risk markers; d) Assess its clinical utility; e) Evaluate whether its use in individuals' management improves clinical outcomes.

In this study, we investigate the prospect of a relatively new technique namely Retinal Vessel Analysis (RVA) in separating cardiovascular risk groups. Hence, this study can serve in the first phase of establishing RVA as a new risk marker, through a cross sectional investigation on apparently healthy subjects applying predictive data mining. In cross section investigations, observational data are analysed from a representative sample of a population at a specific point.

1.1.1 Retinal Vessel Analysis for Cardiovascular Risk Prediction

RVA assesses the function of retinal vessels based on changes in diameter measurements when the retina is subjected to flickering light. The RVA [169] is a relatively new technique compared to static retinal image analysis. It gives a clearer picture of the dynamics of the vessels wall (dilation and constriction) when challenged by varying external conditions (flickering light). The RVA provides automated, objective, and continuous recording of the diameter of a retinal vessel. Accordingly, it enables the observation and subsequent evaluation of local (segmental) fluctuations of the vessel diameter and small periodic changes (e.g., pulse waves).

The justification behind investigating the RVA for cardiovascular risk prediction stems from the following:

1. RVA directly relates to the definition of CVDs (as vessels-related diseases), since it assesses the behaviour of the vessels and subsequently reveals vascular health status. Therefore, RVA can be regarded as a low cost non-invasive method, which could be superior to some established risk markers such as Body Mass Index and smoking status that indicate general health status.

2. RVA captures the dynamic functional behaviour of retinal vessels , which provides a deeper insight into the vessels' health status compared to static retinal vessels analysis.

3. Several techniques, for example retinal fundus examination, for visualising the vessels are already included in primary health care venues. The techniques are provided as part of the routine care given to patients suffering from other diseases

such as diabetics [68], who top the list of high risk groups to cardiovascular diseases. Hence, RVA can be readily incorporated as an additional screening tool for various health conditions.

1.2 Predictive Data Mining for Cardiovascular Risk Prediction

Prediction of CVD risk is a critical task [176]. The successful early risk prediction is directly related to the reduction of complications and increase in probability of survival. Automated prediction and diagnosis systems can aggregate knowledge and expertise of physicians in various sub-specialties creating a comprehensive system. In addition, automated systems can be used in some cases to compensate for limitation in human resources [156]. Successful diagnosis and risk prediction through machine learning approaches is highly anticipated.

Machine Learning (ML) and data mining methods are able to discover unknown (hidden) multi-factorial associations [11] and to produce reliable individual judgments. This property makes ML methods particularly suitable for exploratory cross sectional investigations, where the capability of the new marker to differentiate between risk groups still needs to be ascertained. Cross sectional studies present a fundamental initial stage in the process of establishing a new risk marker [88]. However, the data collected through these studies is often imperfect. Thus, the application of additional ML methods could be needed to enable the construction of reliable prediction models.

Risk prediction models estimate the risk level of developing an outcome using a set of predictors (characteristics) of the subjects [149]. The models comprise a set of techniques to process the data and remedy the data imperfections. These techniques include methods for prediction, missing values handling, dimension reduction,etc.

ML and predictive data mining methods have been applied for cardiovascular risk prediction, however, the acceptance of clinical experts to adopt these methods is influenced by various factors. Obviously, the accuracy of the produced predictions would significantly impact their willingness to use the methods. Health care specialists target high accuracy models that can correctly identify a subject's risk level. Another important factor that influence their acceptance to adopt these methods is whether they can understand how the predictions have been generated by the prediction model. The ability to provide a human interpretable explanation for the prediction (classification) decision increases the experts confidence in the resulting judgments [119].

Various factors can contribute to the accuracy of the model. A primary factor is the characteristics (profile) of the data used for model construction. The RVA data available in our study manifest characteristics that hinder the success of standard ML solutions. Thus, specialised ML solutions are required to successfully handle

RVA data properties. Different ML techniques can be applied and adapted to handle the various characteristics of the problem under study, leading to better prediction.

In the next section, we will provide a summary of the available RVA data characteristics that mandate ML handling together with the main limitations of existing techniques.

1.3 Thesis Motivation

The RVA data profile is the main drive for developing our ML framework called Oversampling Feature SElecTion to *mine* data (OFSET_*mine*) for processing the RVA data and generating the risk predictions.

1.3.1 RVA Data Characteristics Description

The study includes 236 participants, who were divided based on an existing risk score into three groups of low, medium and high risk of sizes 212, 14 and 10 respectively. For each participant in our study, a set of 104 features were generated from the collected RVA data. In addition to the RVA data, a set of measures for systolic blood pressure, total cholesterol and hdl-cholesterol were measured. Age, gender, ethnicity, whether smoker or not and family history of cardiovascular disease were also recorded. The collected data and the labeling process will be described in detail in Chapter 3 section 3.2.

The application of data mining techniques on the available RVA data is of special nature [98] due to the following factors:

1. The availability of relatively large number of features generated based on RVA subjects response. The generation of a large number of features is common practice when new markers for disease risk stratification are being investigated. The features are generated to explore their association with the disease, which may lead to overloading with possibly irrelevant features. In addition, the generated features are interdependent and involved in numerous interactions. Hence, an approach is needed to reduce the dimension of the data while keeping the most informative features and providing an explanation for the determined features relative importance.

2. The limited sample size, mainly due to the fact that only a small number of participants meet the study's inclusion criteria. Another reason for the limited sample size, is that the cost effectiveness together with the motivation of the subjects and subsequently their willingness to participate are not high. This is due that the benefit of the exploratory study of RVA as a prospective risk marker is still to be affirmed.

3. The presence of imbalanced classes. The skewed distribution of the examined disease risk among participants further exacerbate the limited sample size. The imbalance degrades the performance of standard classifiers that assume similar class

sizes. Both characteristics of limited sample size and class imbalance need intervention to compensate for them, while at the same time maintaining the representativenss of the data.

4. The presence of severe class overlap that hinders the effective separation of risk groups [71, 182], which is a prevalent problem in medical data.

5. The expectation to acquire new data during the course of the study as the data will remain to be collected, which means the model needs to be easily updated. This encourages the use of an easily adaptive model capable of generating individual-based judgments.

Straightforward application of established classification techniques on such data can be ineffective, thus combined approaches are needed.

1.3.2 Current Methods and Associated Limitations

Various ML methods can be applied to handle the available RVA data and improve predictions accuracy.

For dimensionality reduction, feature selection can be applied for choosing the most relevant features to increase the accuracy and simplify the model. However, the existing feature selection algorithms anticipate the relevance of the predictors (measures/features) either through theoretical measures or based on the performance of a specific classifier. Thus, the features may not have high predictive performance or have low generalisation with other classifiers and the target concept to be learned. In addition, a characteristic that undermines the suitability of some existing methods for clinical use is that they do not provide an explanation for the relative importance of the features, which lead to their selection.

For handling small sample size and imbalance, oversampling can offer a solution through generating synthetic samples of the small (minority) classes. A concern that rises when oversampling medical data is whether the synthesised samples are true representatives of the generating classes. This can achieved by verifying the validity of the synthesised samples after oversampling. However, current approaches do not address post oversampling validation satisfactorily.

For handling class overlap and augmenting dataset, lazy (distance-based) learning is considered the most appropriate approach [209]. Nevertheless, the purely local approaches adopted in the available lazy methods make them vulnerable to noise and outliers and inadequately handle border samples.

The RVA characteristics together with the limitations of the existing methods motivate the development of a framework of ML methods adapted to the characteristics of the collected RVA data to effectively handle them.

1.4 Thesis Objectives and Contribution

This study has two main objectives. The first objective is to investigate the prospect of using RVA for determining the cardiovascular risk group for different subjects.

The second objective is to select and apply ML techniques effectively to handle the characteristics of the available data and increase the reliability of risk level prediction.

Therefore, *the main contribution of this thesis is an integrated framework, OFSET_mine, that through combination of ML methods remedies previous methods' shortcomings to produce a bespoke solution for Cardiovascular Risk Prediction based on RVA data, that is also applicable to other medical datasets with similar characteristics.* The four identified essential characteristics are (1) imbalanced dataset (2) high dimensionality of data (3) interleaving feature ranges and (4) possibly expanding dataset to increase the sample size.

The framework comprises FiltADASYN for oversampling, to deal with imbalance, DD_Rank for feature selection, to deal with high dimensionality, and *GCO_mine* an easily adaptive learning method for individual-based classification, all integrated within the OFSET_*mine* framework. FiltADASYN improves an existing oversampling method through adding a filtering step to verify the representativeness of the generated samples. DD_Rank assigns a selection score to features combining both plausible aspects of predictive performance and the features co-occurrences with the classes. *GCO_mine* is a distance-based learner that merges the concepts of locality and global decision making to improve the performance of existing lazy methods.

The framework is a multi stage approach which includes independent methods, that can be applied or removed from the framework according to the characteristics of the collected data. In the future, when substantially more data are collected, the oversampling stage can be omitted and the proposed *GCO_mine* can adapt to the newly acquired samples

1.5 Thesis Outline

The thesis is organised as follows:

In Chapter 2, traditional cardiovascular risk prediction models using machine learning techniques are discussed. Also, the association of cardiovascular risk with altered retinal vascular function is presented to provide the motivation behind investigating RVA for risk prediction.

Chapter 3 includes the problem statement together with the specific characteristics of RVA data that drive the choice of ML solution.

Chapter 4 reviews several recent approaches in the field of predictive data mining. First, we briefly review the common predictive data mining approaches to tackle the problems of missing data, data imbalance, small sample size and high dimensionality, all identified characteristics of RVA data. The design requirements and the suitability criteria for assessing candidate methods are introduced next. A set of selected approaches are then described in detail and evaluated against the suitability criteria.

Chapter 5 first verifies the practical suitability of ML approaches for cardiovascular risk prediction using our data. Then, an outline of the proposed framework OFSET_*mine* and a summary of each of the proposed methods is provided. We describe the additional medical datasets used to validate the general applicability of the proposed techniques.

Chapters 6, 7 and 8 explain in detail the proposed techniques within the OFSET _*mine* framework along with an individual evaluation of the proposed methods performance. All these chapters comprise a description of the experimental study including the experiments objectives, used data, evaluation metrics.

Chapter 6 presents the proposed oversampling method FiltADASYN in detail together with its evaluation against one of the state of the art methods namely ADASYN.

Chapter 7 describes Restricted Boltzmann Machines (RBMs) as they present the foundation for the developed feature selection method DD_Rank. The method is then detailed and compared against two of the well established feature selection algorithms ReliefF and Correlation Coefficient (CorrCoeff).

Chapter 8 portrays the proposed lazy classification method *GCO_mine* where its performance is verified against a set of eager and lazy classification methods.

Chapter 9 presents the overall performance of the hybrid approach combining Oversampling and Feature SElecTion in OFSET and OFSET_*mine* for cardiovascular risk prediction using RVA-based measures and on other selected benchmark datasets.

Chapter 10 summarises the main results and findings of the study, provides an evaluation of the clinical utility of the proposed methods and presents the suggested future work directions.

Ethical approval for the study was received from Aston's University Ethics Committee. Written informed consent was received from all participants prior to study enrolment and all study procedures were designed and conducted in accordance with the tenets of the Declaration of Helsinki.

Chapter 2

Cardiovascular Risk Prediction State of the Art

Cardiovascular diseases represent one of the main causes of human morbidity with rising trends in large areas of the world [142]. For many patients, the first sign of cardiovascular disease is a severe event that is often fatal. The early identification of seemingly healthy subjects at risk through non-invasive methodologies is a priority.

Extensive studies have revealed an association between several factors and increased risk of cardiovascular diseases. The identified risk factors [66] can be categorised based on whether a factor is modifiable or not. Modifiable risk factors include smoking, alcohol consumption, stress, high cholesterol diet, psychological factors and obesity. The alteration of modifiable factors may help decrease the related risk, thus such categorisation is crucial. Age, gender, menopausal state for women and personality type are examples of non modifiable risk factors. New computing and sensor technologies enabled the collection of huge amounts of cardiovascular related data in a non-invasive manner. Recently, extracting knowledge and insight from this huge volume of data and building efficient decision support systems through data mining became necessary.

Existing studies that perform cardiovascular risk prediction are reviewed and the need for further studies is justified. Also, studies that associate changes in retinal calibres to cardiovascular risk are described to provide justification for the investigation of RVA for CVD risk prediction.

2.1 Traditional Models

Large efforts have been dedicated to the study of several risk factors and the associated cardiovascular risk. The Framingham heart study is an ongoing longitudinal cohort study since 1948. The study's aim is to build risk models for predicting 10 year risk of developing cardiovascular diseases. It is considered one of the most well established studies [139]. The Framingham Heart Study started with 5,209 adults from the city of Framingham, Massachusetts. Since then, several versions for risk score calculation were constructed. Originally, the Framingham Risk Score

(FRS) was developed to estimate the 10-year risk of developing coronary heart disease only. In a later version of FRS score calculation developed by D'Agostino et al. [49], cerebrovascular events, peripheral artery disease and heart failure were subsequently added as disease outcomes together with coronary heart disease. The more recent risk model by D'Agostino et al. presents a single sex-specific multivariable risk function that predicts risk of developing CVD and all of its constituents, unlike other function that predict single specific events. The derived model by D'Agostino et al. applied Cox proportional-hazards regression on 8491 Framingham study participants between 30 and 74 years of age (mean age, 49 years; 4522 women), who attended a routine examination and were free of CVD. The developed "general CVD" risk function incorporated age, total and high-density lipoprotein cholesterol, systolic blood pressure, treatment for hypertension, smoking, and diabetes status. Cox proportional-hazards regression [47] was used to build the models and evaluate the risk of developing a first CVD event.

Over 12 years of follow-up, 1174 participants (456 women) developed a first CVD event. The general CVD algorithm demonstrated good discrimination (C statistic, 0.763 [men] and 0.793 [women]) and calibration. Framingham Risk (FRS) calculators exhibit limitations [12, 34, 46, 89] in terms of transportability to different populations , where re-calibration of the algorithms would be needed. Moreover, they tend to overestimate or underestimate the risk. In addition, the acquisition of some of the FRS measures is invasive, such as blood analysis for total and high-density lipoprotein cholesterol determination, which may lead to the reluctance of subjects to undergo the examination [210]. Also, the limited resources of some developing countries may hinder performing blood tests for screening [24]. Hence, the exploration of non-invasive alternative risk markers is justifiable.

Hippisley-Cox et al. [87] conducted a prospective cohort study for 15 years using data of 2.3 million patients aged 35-74, collected from general practice records. The aim was to construct a risk model (QRisk) that estimates cardiovascular risk in patients from different ethnic groups in England and Wales. Cox hazard proportional regression was applied to build the model. The same measures as for Framingham Risk were used. In addition, other measures such as body mass index and ethnicity were included in the QRisk score calculation with the aim of increasing the accuracy of the model relative to Framingham Risk model. Hippisley-Cox et al. concluded the extra measures into the QRisk algorithm provides better calibration and discrimination compared to FRS in a nationally representative population. Despite this relative advantage, the QRisk shares with FRS the same limitations of transportability to different populations and invasive examination. Also, in the early generations of the QRisk study both the derivation and validation populations were the same, which implies that further studies are still required to increase the reliability of the risk score calculator.

Other example studies are reviewed next, which study the association of relatively novel risk factors with cardiovascular risk either directly or indirectly. Some

of these studies consider features such as arterial stiffness and inter-ventricular septum thickness as indirect assessment of cardiovascular risk. The studies are chosen because they are either cross sectional with similar sample size to our study or most importantly apply machine learning methods for prediction.

C4.5 decision tree was utilised to develop a system to assess the ability of several risk factors for predicting Coronary Heart Disease (CHD) in the study of Karaolis et al. [108]. The examined risk factors include non modifiable factors: age, gender, family history, and a group of modifiable factors such as smoking, diabetes, hypertension treatment, blood pressure and lipids. Karaolis et al. [108] tackled the issue of splitting criteria effect on decision trees performance. Five splitting criteria were examined in building the decision trees namely: information gain [128], Gini index [62], likelihood ratio chi-squared statistics [73], gain ratio, and distance measure. The generated models were investigated accordingly including, event and non-event subjects. Data collection, cleaning and missing values filling were performed.

The experiments included the records of 528 cases in which the before-event and after-event risk factors were registered. The available cases were categorised as medium risk and high risk groups using FRS and thresholding leading to group sizes of 35 and 493 respectively. Wilcoxon rank sum test [202] [1] was applied to differentiate between the splitting criteria in terms of statistical significance. After tree construction, a set of rules are extracted. In the experimental analysis, all the applied splitting criteria were found to have comparable results. Finally, the presented results were compared to the results of the European Action on Secondary and Primary Prevention by Intervention to Reduce Events (EUROASPIRE) [120] study. Similar findings regarding the most important risk factors namely: sex, age, smoking, blood pressure, and cholesterol were reached by both studies. Also, similar findings on the percentages of the participants having high blood pressure, high cholestrol low density lipoprotein and smoking after an event. The investigated five splitting criteria were not independent but relying on each other. For example information gain, likelihood ratio chi-squared and gain ratio include information gain in their computation while distance measure is calculated based on Gini index. This may explain the comparable performance with an average classification accuracy of 72%. Therefore, investigating further splitting criteria maybe a path worth following. Also, the methods that were used for filling missing values, selecting features and rules were not described, although they may have greatly influenced the performance.

Alty et al. [9] studied the problem of arterial stiffness prediction, as an indicator of CVD risk, from digital volume pulse waveform (DVP). DVP is used as an approximation measure for pulse wave velocity (PWV) since the acquisition of PWV is more complex and invasive. The study compared SVM-based classification

[1] Wilcoxon rank test: a non-parametric statistical hypothesis test used to compare two related samples to assess whether their population mean ranks differ. It can be used to determine whether two dependent samples were selected from populations having the same distribution.

[188], SVM-based Regression [188] and Artificial Neural Network (ANN) [78] performances (to be described in Chapter 4. The study included 461 subjects with PWV and DVP measured for each subject. The subjects were categorised into low (PWV < 9 m/s) and high stiffness (PWV > 11m/s) subjects, eliminating subjects with intermediate stiffness. Two types of feature extraction methods were applied namely physiologically motivated feature extraction and signal subspace-based feature extraction. While physiologically motivated features are parameters associated with the physiological properties of the aorta and arterial characteristics in general such as peak to peak time, crest time and stiffness index, signal subspace-based features are extracted applying information theoretic approaches. The applied approaches for signal subspace feature extraction included kernel principal component analysis, wavelet packet decomposition, and signal subspace analysis. The extracted features were fed into the prediction algorithms. SVM–based techniques outperformed ANN approaches reaching 87.6% correct PWV prediction rate employing a combination of features from both applied methods. The study showed that the prediction rate largely depends on the feature extraction methods, either physiologically motivated or signal subspace-based method, applied. DVP is also used to predict PWV status and accordingly indicate CVD risk. The direct relation between DVP and CVD risk prediction still needs to be established through further analysis.

Ramirez-Villegas et al.[100] investigated the application of ANN and SVM on Heart Rate Variability (HRV) series in the prediction of cardiovascular risk in general, not related to any specific cardiac disease. The study analysed 90 HRV records only: 45 of healthy subjects and 45 of known cardiovascular risk patients. The HRV series were analysed by statistical, spectral, multi resolution and non-linear methods generating 52 features in the feature extraction stage. After feature extraction, two sample Kolmogorov-Smirnov (KS)–test [2] [141] was employed for feature selection . The KS-test selected five, 10 and 15 features to be used in the experiments. The selected features were fed into prediction algorithms, Multi-Layer Perceptron (MLP) network, Radial Basis Function (RBF) network (to be described in Chapter 4) and SVM [188] were utilised in the experiments.

First, the performance of the predictors was assessed given different feature subset sizes to determine the best feature subset size. MLP achieved the highest accuracy of 96.6% when applied on the top five selected features by KS-Test. The second experiment involved testing the performance using all the generated features with the classifiers to evaluate the significance of feature selection, here SVM outperformed ANNs. The authors argue that this is due to the overfitting of ANN with high dimensional data. Despite that SVM outperformed ANN with the complete feature set; MLP with top five features selected by KS test retained the best overall performance. The top five features ultimately selected were not reported which makes the reproduction and the interpretation of the work impossible.

[2] KS-test is a non-parametric test that uses cumulative distribution function to test whether two samples share the same distribution.

Pfaff et al. [152] investigated the application of Fuzzy C-Means (FCM) clustering [23] to the problem of inter-ventricular septum thickness (IVS), considered as a quantitative cardiovascular risk factor, prediction in haemodialysis patients. Since clustering only groups data based on a measure of similarity, a rule extraction algorithm is needed to assist in the prediction process and create rules with a specific target variable. The experiments studied long term follow up records of 63 patients for four years with known and unknown IVS thickness. Year 1 and Year 4 records were used in the study and 42 preselected variables for each year per patient including patient data, anamnesis and diagnosis data, laboratory and medication data. Two target clusters were generated for large and small IVS thicknesses. Uniconditional rules were generated and rated by Normalised Empirical Rating (NER) [72] and Confident Hit Rate (CHR) [99]. Three rule bases were formulated using the top 10 rated rules from NER, CHR and NER/CHR rating. In the prediction stage, the NER rulebase combined with the FCM clusterbase had the best results. The results of the patients with unknown IVS thickness were medically validated. The paper suggested a solution for the problem of unknown target value prediction that can be applied in comparable cases, but limited results interpretation for this investigation was provided and no clarification for the medical validation procedure was given. It also utilised a well known clustering algorithm and rule rating measures, these can be used with similar data sets. Preprocessing and preselection of the variables were mentioned briefly without specifying the used techniques although these would be significant steps. The small size of the test set should have been compensated in the evaluation stage.

Other examples of cardiovascular risk prediction applying machine learning techniques on larger cross sectional datasets can be found in [104, 115]. Juarez-Orozco et al. [104] use simple available predictors from 1,241 participants with no previous myocardial infarction or revascularisation to predict PET-measured hampered myocardial perfusion reserve (MPR), which conveys a significant risk for adverse cardiovascular outcomes. Demographic, clinical and complementary diagnostic data were collected for each participant. MLP Network and ensemble boosting of the ANN were employed with the partitioned data. These lead to accuracies of 74% (AUC =0.77) and 85%, respectively.

Kim et al. [115] used Deep Belief Networks [95] (DBN) and statistical feature selection to predict cardiovascular risk level (low or high) from the sixth Korea National Health and Nutrition Examination Survey (KHANES-VI) data set. The dataset was split into 70% for training and 30% for testing. The study included 4,244 participants, for whom eight primary care features were recorded. Statistical feature selection (Mann-Whitney U-test [3] and Chi-square [4]) was applied and six features were used for learning. The performance of statistical DBNs was compared to Naïve

[3]Mann-Whitney U-test: Examines two groups samples means to check whether they are coming from the same population. It can be used when the normal distribution assumption is not met.

[4]Chi-square: a statistical measure that quantifies the independence between variables (e.g.predictor and class). The higher the value of Chi-square the more useful the feature is considered.

Bayes, logistics regression, back propagation network, support vector machine, random forest and DBN with the full feature set. Statistical DBNs attained the highest accuracy (83.9%), sensitivity (87.6%) and ROC curve (0.79 ± 0.016) performance, while SVM scored 100% specificity indicating that SVM could separate low risk subjects particularly well. The labeling procedure of the participants was rather opaque and the adoption of two well separated (low and high risk) classes is likely to have produced optimistic results.

Weng et al. [201] conducted an extensive prospective cohort study using routine clinical data of 378,256 patients from UK family practices. The study's objective was to investigate the prospect of machine learning methods against an established algorithm by the American College of Cardiology (ACC) guidelines in predicting first cardiovascular events in 10 years. Four machine-learning algorithms were used: random forest, logistic regression, neural networks (described in detail in Chapter 4) and gradient boosting [5]. The study comprised eight features used in ACC risk model calculations and 22 variables that had the potential of being associated to cardiovascular risk from other studies. Predictive accuracy was assessed by area under the receiver operating curve (AUC), and sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) to predict cardiovascular risk. All applied machine learning methods showed performance improvement over the ACC algorithm. Neural Networks achieved the highest accuracy improvement of 3.6%. The study stipulated on the ability of machine learning algorithms to improve cardiovascular risk prediction through capturing complex interactions between risk factors and expanding current risk models by exploring a wider range of risk indicators.

The presented studies attempt to address the problem of cardiovascular risk prediction using various approaches. Overall, they provide variable quality solutions that tackle the problem from different perspectives. The studies prove the suitability of different machine learning approaches in CVD risk prediction. Despite the value of the presented studies, they manifest some shortcomings. An example of these shortcomings is that some of the studies provide further validation for already known risk factors [104, 108, 201]. This further validation of known risk factors is useful, but it does not take advantage of one of the principle capabilities of data mining which is detecting hidden (unknown) relationships within data. Another shortcoming of some of the reviewed models is that they predict outcomes (related to CVD risk) as indirect indicators of CVD risk (no direct prediction) [9, 104, 152]. Also, some studies predict the risk for clearly separated groups where one of the groups is already known to be of high risk (e.g: previously diagnosed vs control subjects , suffered from cardiac events vs cardiac events free subjects, etc) [9, 108, 115, 152, 159]. The utilisation of clearly separated groups with previously known risk and lack of the examination of variability in apparently healthy groups does not

[5]Gradient boosting is an ensemble learning algorithm which combines the predictions of a group of weak learners to optimise a differentiable loss function.

serve early CVD risk detection. Early detection can help reduce the threat of death dramatically and ML methods can help advance the state of the art in this direction. In addition, the study of Ramirez-Villegas et al. [159] and Pfaff et al. [152] had a small sample size of 90 and 63 records respectively which limits the validity of the produced models.

The success of ML for CVD risk prediction, together with the limitations of existing studies established foundations for further work in the field of early cardiovascular risk prediction using machine learning. In this study, we aim to examine (a) the use of a relatively new technique namely Retinal Vessel Analysis (RVA) as a potential risk marker (b) to separate seemingly healthy participants into risk groups through an individual based model.

2.2 Retinal Vascular Function

Investigating the use of RVA for CVD risk prediction is justified by the fact that the retina offers an easily accessible site to the evaluation of ocular microcirculation. The coexistence of ocular microvascular and systemic macrovascular abnormalities in early stage of many morbid conditions has been recently studied [97, 111, 117, 146, 148]. Subtle changes in the retinal vasculature is assumed to carry information useful for predicting cardiovascular risk [64, 183, 192] independent of traditional risk factors.

The diameter of a vessel is an important quantitative parameter for describing it. Originally, the vessel diameter was measured semi-automatically in single static images. The association of differences in retinal vascular calibres to cardiovascular risk in still fundus images, which picture the center and peripheral retina and its vessels, has been explored in the past decade [168, 181, 195].

In the following subsection, we will review studies that associate variations in retinal vessels calibres from static images to CVD risk, using population-based cohort study records and statistical analysis. Afterwards, we will present investigations that studied the relation between dynamic RVA and cardiovascular risk.

2.2.1 Static Retinal Vessels Calibres for Cardiovascular Risk Prediction

Wang et al. [195], studied the capability of retinal vascular calibre to independently predict risk of coronary heart disease (CHD) -related death. Retinal arteriolar and venular calibres of 3654 Australians aged above 49 years were measured from baseline retinal photographs. The arteriole to venule ratio (AVR) was calculated. A follow up of nine years was conducted and CHD-related death was confirmed from the Australian National Death Index. An association between wider venules, narrower arterioles and smaller AVR and coronary heart disease (CHD) -related death was found, especially in women and middle aged groups. It was found that larger retinal venular calibre independently predicted a 1.5–2-fold higher risk of CHD death

in both men and women aged 49–75. Also, smaller arteriolar calibre and AVR predicted a 1.5–2-fold higher risk of CHD death, but only in women aged 49–75. These results suggest the possible usefulness of retinal vessels changes in predicting cardiovascular risk.

Wong et al. [181] examined the relation between retinal vessels calibres and incident coronary heart disease (CHD) and stroke in elderly persons. The venular and arteriolar calibres were measured from the retinal photographs of 1992 men and women aged 69 to 97 years from four US communities. After five years follow up, an analysis was carried out controlling for several known risk factors. The study findings ascertained the correlation between larger retinal venular caliber and the increased risk of CHD, where participants with larger retinal venular caliber had a higher incidence of CHD (11.7% vs 8.1%). Smaller retinal arteriolar caliber was associated with incident CHD (rate ratio 2.0; comparing largest with smallest arteriolar caliber quartiles).

Seidelmann et al. [168] conducted an extensive 16 year cohort study to determine whether retinal vessel calibres are associated with long-term cardiovascular outcomes. The study also investigated whether retinal vessel calibres can provide incremental value over the 2013 American College of Cardiology/American Heart Association Pooled Cohort Equations in predicting atherosclerotic cardiovascular disease events. The study included 10 470 men and women without prior atherosclerotic cardiovascular disease events or heart failure in the ARIC Study (Atherosclerosis Risk in Communities). All the participants underwent retinal photography to extract the vessels measurements. The study's results included that: a) rates of all outcomes were higher in those with wider retinal venules and narrower retinal arterioles. b) Retinal vessel caliber reclassified 21% of low-risk women (11% of all women) as intermediate risk. The women identified at higher risk by the use of retinal vessels would not be recognised using existing practice guidelines. In conclusion, Seidelmann et al. confirmed the relation between narrower retinal arterioles and wider retinal venules and higher long-term risk of mortality and ischemic stroke in both sexes and coronary heart disease in women. They also stipulated on the fact that these measures can provide a cost effective and reproducible marker that added incremental value to current practice guidelines.

The described studies present clear diversity in terms of age groups, population communities and the associated cardio-related outcomes. The constant affirmed association between deviations in retinal vessels calibres and elevated cardio-related risk and the diversity of the presented studies encouraged Poplin et al. [154] to employ new approaches to establish the use of retinal vessels calibres as novel added value markers. Poplin et al. employed deep learning and demonstrated high effectiveness in predicting cardiovascular-related risk markers (such as smoking, body mass index and systolic blood pressure) from retinal static images [154]. In addition, they could predict the onset of so-called Major Adverse Cardiovascular Events (MACE) within five years, with Area Under Curve (AUC) values similar to the well

recognised risk calculator SCORE[44].

2.2.2 Dynamic Retinal Vessels Analysis for Cardiovascular Risk Prediction

The reviewed studies repeatedly ascertain the relation of retinal vessels calibres differences to CVD risk. Some studies suggest possible usefulness of incorporating retinal-based measures for cardiovascular risk prediction in current practice [168, 181]. However, the changes that can be observed in static images only indicate the presence of a pre-clinical disease that already exists, therefore, the prediction is actually anticipated. A step forward from the use of static images is the analysis of dynamic, real time vascular functional changes that can occur in individuals at risk even when their static images show no vascular alterations.

Retinal Vessel Analysis (RVA) [169] can perform dynamic analysis and captures the vessels' functional behaviour based on stimulation with flickering light, according to an established protocol. The vessels behaviour when subjected to flickering light include a dilation response followed by constriction (recovery) after flicker cessation, returning to the baseline vascular fluctuations.

Various studies assessed retinal vascular function using RVA of participants with CVD, with risk factors for CVD, or with cardiovascular-related pathology such as diabetes.

Lim et al.[132] studied the dynamic response of retinal vessel caliber in association to Diabetic Retinopathy (DR)(cardiovascular-related pathology). The study included 15 subjects with Type I Diabetes Mellitus (DM), 216 subjects with Type II DM and 45 control subjects. The results showed that reduced arteriolar and venular dilation is associated with DR progression. This relation was persistent in both univariate and multivariate analysis adjusted for age, gender and smoking status.

The alteration in retinal vessels responses of patients with Coronary Artery Disease (CAD) was investigated by Heitmar et al.[83]. Subjects with CAD (24 subjects) were recruited and 30 control participants were age and gender matched. The principle finding of the study was that the retinal vessels reaction time was significantly different ($p - value = 0.016$ where significance threshold θ set to 0.05) between patients and controls.

Heitmar et al.[82] examined the variability in retinal vessel reactivity of subjects with DM and/or CVD. The examination involved 36 participants with DM only, 43 participants with CVD only and 37 participants with both CVD and DM. The study revealed subtle reactivity differences between groups suffering from CVD with and without DM. Examples of these differences are 1) inconsistent constriction response in CVD + DM participants between different flickers, unlike the other groups, 2) different dilation time between flicker cycles in CVD + DM participants, 3) different reaction pattern and lack of arterial constriction in DM group. These findings may provide a suitable marker to monitor progression.

Seshadri et al. [172] compared retinal vascular function in asymptomatic individuals with and without a positive Family History (FH), which is a known risk factor for CVD. The comparison was carried on all apparently healthy individuals, of which 38 subjects have a positive FH of CVD and 37 with negative FH. Individuals with FH of CVD showed significantly reduced arterial dilation response and decreased constriction slope ($p - value$ = 0.001 and $p - value$ <0.001 respectively). Similar findings were recorded for venular response. Seshadri et al. concluded that even though macrovascular function does not convey impairment in individuals with positive FH of CVD and low FRS, functional changes can be detected at the retinal microvascular level.

Fathalla et al. [60] were the first to apply machine learning to demonstrate the capability of the RVA data in discriminating between cardiovascular risk groups in apparently healthy participants. Fathalla et al. applied machine learning techniques on 236 participants divided into low, medium and high risk with sizes 212, 14, 10 respectively. However, their reported results showed room for accuracy improvement, which motivate further research on developing more effective approaches that use RVA for risk group prediction.

Based on the aforementioned outcomes of the reviewed studies and the ability of RVA to capture dynamic vessel functionality, the RVA is seen to have the potential to enable the detection of early signs of vascular ill-health, before any changes in static images occur. Therefore, further investigation for the prospect of RVA as early and direct cardiovascular risk indicator is justifiable.

Chapter 3

Rationale

As cardiovascular disease (CVD) represents a spectrum of disorders that often manifest for the first time through a sudden life-threatening event, early identification of seemingly healthy subjects with various degrees of risk is a priority.

3.1 Problem Statement

More recently, traditional scores used for early identification of CVD risk are slowly being replaced by more sensitive biomarkers that assess individual, rather than population risks for CVD. Among these, retinal vascular function, as assessed by the retinal vessel analysis method (RVA), has been proven as an accurate reflection of subclinical CVD in groups of participants without overt disease but with certain inherited or acquired risk factors. Therefore, RVA has the potential to be used for early cardiovascular risk prediction and ML prediction methods can be applied for this purpose. However, the available RVA data for our study has various characteristics that hinder the success of standard prediction methods. As a result, specialised machine learning methods that can cope with the characteristics of the data need to be devised in order to correctly detect individual risk at an early stage.

In order to produce accurate cardiovascular risk prediction using the collected RVA data, we need to 1) identify the collected RVA data characteristics that mandate handling; and 2) develop and apply the necessary methods that increase the dependability of the obtained predictions.

In the next section, the collected RVA data is fully described to define the characteristics of the data and provide a solid clarification for the requirements of the needed machine learning methods.

3.2 The RVA Data

3.2.1 Data Collection and Feature Generation

Asymptomatic volunteers were recruited and investigated at the Aston University Vascular Research Laboratory, Birmingham, UK by the assigned personnel [146] and was not part of this PhD contribution. Our study includes 236 participants, eliminating subjects who had a positive diagnosis of severe cardio- or cerebro-vascular

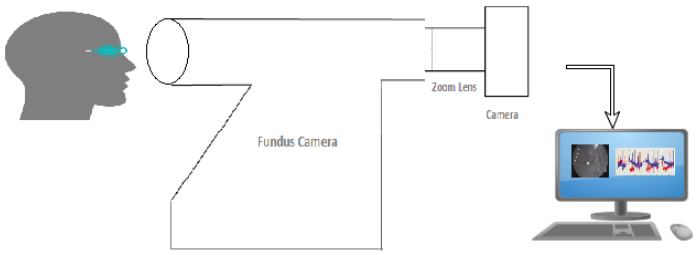


FIGURE 3.1: The RVA Acquisition Setup

disease. All measurements were taken between 8.00 am and 11.00 am following a 12-hr overnight fast (no alcohol or caffeine). Retinal vessel reactivity was measured using the dynamic retinal vessel analyser (DVA; IMEDOS GmbH, Jena, Germany) as described in [146]. The setup for RVA is illustrated in Figure 3.1. The subjects are at a sitting position with the head leaned toward the fundus camera. Regions of interest from both arterial and venular retinal vessels are marked (as shown in Figure 3.2) and their diameters are recorded. On the screen of the retinal vessel analyser, a medical expert marks the segments of the retinal vessels to be examined with a rectangular area. If any small eye movements occur, the retinal vessel analyser performs on-line correction and readjusts the recording to the new location of the eye as long as it is within the marked area. During the RVA, a flicker light impulse is generated to test vessels walls responses in terms of diameters fluctuations. The attached computer automatically records the readings and calculates an average response using the respective analysis software (DVA; IMEDOS GmbH, Jena, Germany). The recruitment of subjects along with the corresponding RVA and the subsequent curve smoothing and feature generation were performed in the Aston University Vascular Research Laboratory, Birmingham, UK by the designated personnel.

Retinal vessel diameters were recorded over a 350-second time period. This period consisted of 50 seconds of baseline measurements under still illumination (25 Hz), followed by three cycles $F1$, $F2$ and $F3$ of 20 seconds flicker stimulation (opto-electronically generated at 12.5 Hz) each interrupted by 80 seconds of still illumination (recovery). Retinal vessel diameters were recorded at a frequency of 25 readings/sec. According to the protocol recommended by Nagel et al. [147], the period from -30 to -5 seconds prior to each flicker cycle was taken as baseline. All measurements were performed in a quiet, temperature controlled room (22^oC) following full pupil dilation (1% tropicamide; Chauvin Pharmaceuticals Ltd, Surrey, UK) and were taken from the inferior temporal vessel branches approximately one and a half disc diameters from the optic nerve head. For more details on the instrument, followed procedure and protocol, refer to Seshadri (Chapter 4) [170].

The response of each subject to the three flicker cycles was recorded and segmented as $F1$, $F2$ and $F3$ for both arterial and venular vessels where $_A$ and $_V$ denote arterial and venular flickers respectively. An averaged flickers response for each individual subject was computed; averaged flickers were used to ensure reliability as some recordings of each individual flicker cycle may be missing.

Polynomial regression was applied on the individual segmented and averaged flicker responses for curve fitting to obtain an approximate response for each participant [146]. Polynomial regression lead to smoothing of the obtained responses (curves) and hence better visualisation of the response. Figure 3.3 shows the three flicker cycles response as (A) original vessels diameter fluctuations (response) during RVA and as (B) smoothed response by polynomial regression. The polynomial degree used for polynomial curve fitting was chosen by Mroczkowska et al. [146] to be 20 as this provided the best fit to the available data points (the recorded diameter values in response to flickering light). The polynomial degree is an adjustable parameter depending on the fitted data points. The smoothing effect of polynomial regression may slightly alter the values of the generated features. However, since the same polynomial fitting is performed on all the responses from all subjects, the relativity of their responses (to each other and to the classes) is maintained.

Figure 3.4 shows the segmented smoothed response for each vessel (arterial and venular) and for each flicker cycle as well as the averaged response, producing eight response segments. A set of features are calculated by Seshadri et al. [171] and Mroczkowska et al. [146] for the eight responses demonstrated in Figure 3.4. The features are derived to quantify variation in baseline vessels states, response times and response amplitudes. They are chosen to create vasodilation and vasoconstriction response profiles for each volunteer that would map to various risk groups. All the generated measures are numeric continuous valued features. The same set of features calculated by Seshadri et al. [171] are used in this study, where 13 features are computed per response segment (8 segments), which are shown in Figure 3.4. In total, 104 RVA features are created per subject. The generated features per segment are shown below:

- Baseline: Mean diameter before flicker cycles start.
- Baseline Diameter Fluctuation BDF = difference between maximum baseline vessel diameter and minimum baseline vessel diameter.

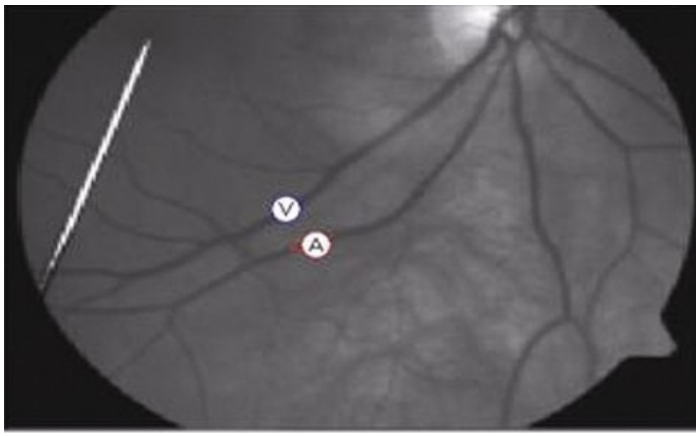
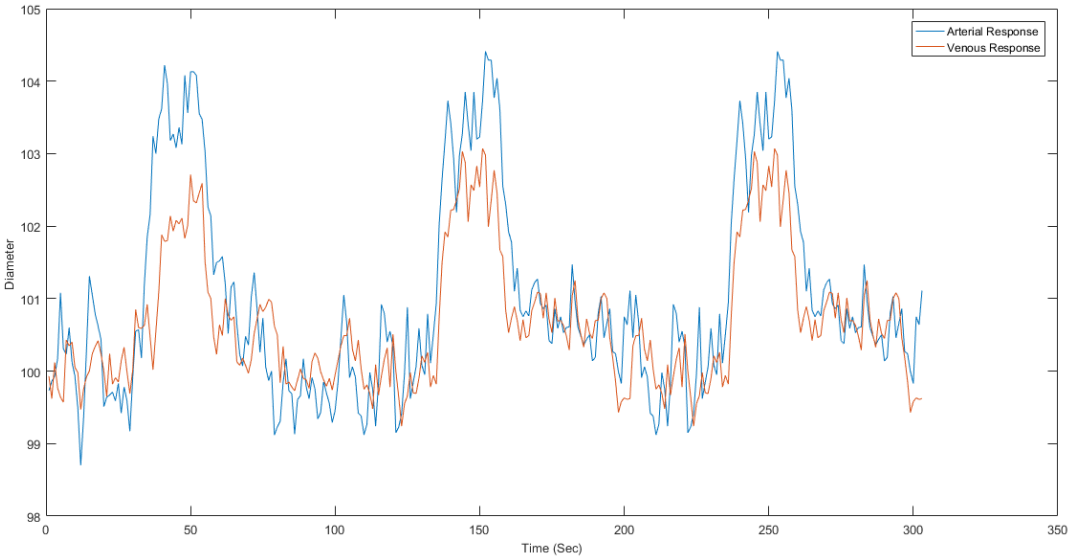
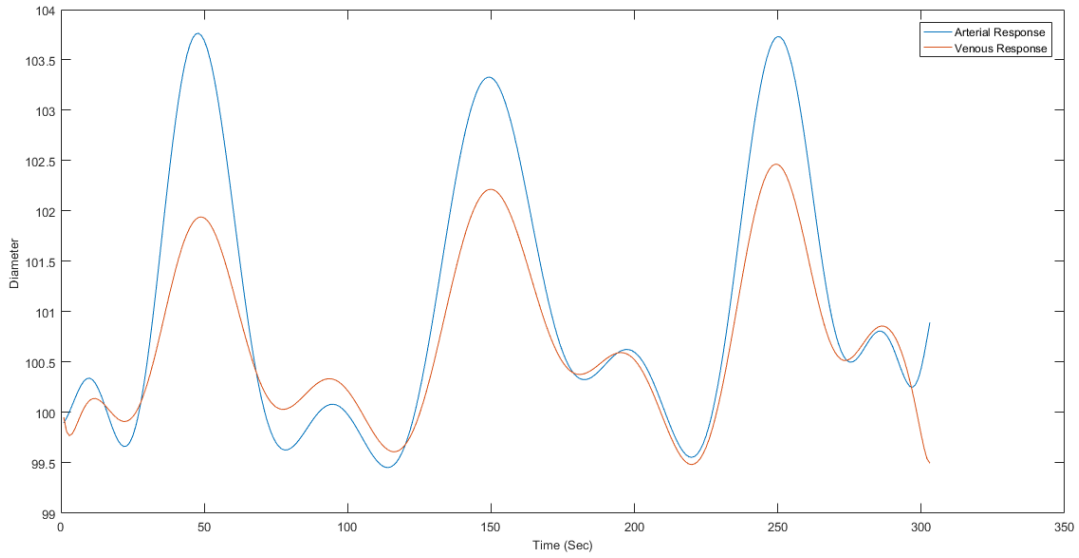


FIGURE 3.2: Vessels marking for RVA



(A) Sample Original Response



(B) Sample Smoothed Response

FIGURE 3.3: The smoothing effect of polynomial regression when applied to a sample (A) Original Recorded Response leading to (B) Smoothed Response

- Maximum Diameter MD: Maximum vessel diameter value after flicker start.
- Reaction Time to Maximum Dilation tMD: Time to reach maximum diameter value after flicker start (sec).
- Percentage Dilation $PerDil = \frac{MD-Baseline}{Baseline} \times 100\%$.
- Maximum Constriction MC: Minimum diameter after maximum dilation.
- Time to Maximum Constriction tMC: Time to reach maximum vessel constriction diameter (sec).
- Percentage Constriction $PerCon = \frac{MC-Baseline}{Baseline} \times 100\%$.

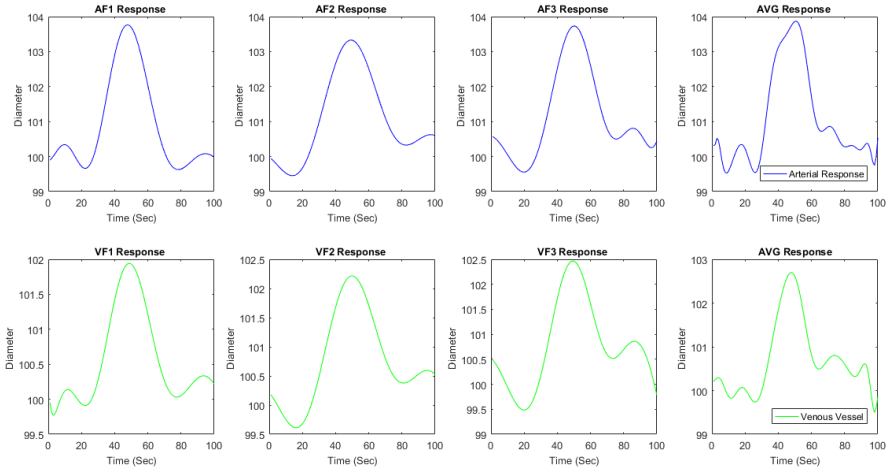


FIGURE 3.4: Segmented flickers responses from which the respective features will be generated.

- Dilation Amplitude $DA = MD - MC$.
- Baseline Corrected Flicker Response

  $BCFR = DA - BDF$.
- Time between Maximum Dilation and Maximum Constriction $tMDMC = tMC - tMD$.
- Dilation Slope $upslope = \frac{MD - Baseline}{tMD}$.
- Constriction Slope $downslope = \frac{MC - MD}{tMDMC}$.

3.2.2 Characteristics of the Collected RVA Data

From the described calculations, 104 features are generated where some of these may be less relevant than others and this can degrade the performance and/or complicate the model. The generated features are often interdependent and exhibit multi-way interactions. The value ranges of the features and their standard deviations are given in Table 3.1. The standard deviation is calculated to estimate the degree of dispersion of each feature. In addition, ANalysis Of Variance (ANOVA) [1] is performed using each feature across risk groups with the null hypothesis that the means for all risk groups are equal. The hypothesis would be rejected if the $p - value$ is less than the set threshold θ of 0.05. The hypothesis was rejected with three features only namely: $Downslope_VF1$, TMC_VF1 and MC_VF1. The low and comparable standard deviation of the features together with affirmed ANOVA equal mean hypothesis show that feature selection could not be performed on simple statistical basis such as standard

[1] One-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of two or more groups given a response variable (feature). The response variable needs to be numeric and the groups are categorical.

deviation [215] or ANOVA [20]. Therefore, a more sophisticated method is needed to reduce the dimensionality of the data. In addition, it would better be capable of detecting the features interactions and reduce the number of features accordingly.

TABLE 3.1: RVA Measures Ranges and Standard Deviation (std-dev)

Measurement	*Range*	$std - dev$
Baseline	(78.66 - 105.47)	0.53
Baseline Diameter Fluctuation	(0.33 - 12.43)	1.59
Maximum Dilation	(93.24 - 124.65)	0.26
Reaction Time to Maximum Dilation	(5 - 43)	2.03
Percentage Dilation	(7.25 - 34.33)	0.26
Maximum Constriction	(75.70 - 113.89)	0.47
Time to Maximum Constriction	(12 - 62)	0.72
Percentage Constriction	(-22.27 - 9.81)	0.38
Dilation Amplitude	(2.62 - 25.32)	0.25
Baseline Corrected Flicker Response	(4.49 - 21.07)	1.35
Time between Maximum Dilation and Maximum Constriction	(3 - 52)	0.83
Dilation Slope	(0.17 - 9.34)	0.17
Constriction Slope	(-5.02 - (-0.02))	0.13

Additional general investigations were performed in Aston University Vascular Research Laboratory, Birmingham, UK by the designated personnel. These investigations included measuring systolic blood pressure, total cholesterol and hdl-cholesterol. Also, several other parameters were recorded such as age, gender, ethnicity, whether smoker or not and family history of cardiovascular disease. The recorded measures from the general investigations are the measures used for calculating FRS and QRisk scores. These measures are collected to allow assessing the prediction quality of RVA based features in comparison to FRS measures and QRisk measures. FRS is chosen for comparison as it is a long established risk score calculator based on an extensive cohort multi-ethnic study, while QRisk presents a risk score derived from UK population similar to the subjects who volunteered in our study.

3.2.3 Labeled Classes Properties and Representatives Visualisation

Despite the affirmed association between changes in retinal vessels calibres and cardiovascular risk, there are still no reference ranges of the RVA-based measures for normal and pathological cases. The lack of reference ranges hinders the manual risk labeling of the participants. Hence, a validated estimate for labeling each participant record (containing RVA measures and general investigations measurements) is needed to allow building an appropriate prediction model for the data. A scheme based on the FRS [49] is adopted for labeling the participants. The FRS provides a validated means of estimating CVD risk in asymptomatic patients. It presents a 10-year risk score for each subject given physical examination findings and laboratory evaluations. The FRS scores are calculated using the FRS calculator provided by D'Agostino et al. [49], where Cox hazards proportional regression was used to

generate the model and the related predictors regression coefficients. The regression coefficients (β) used in the D'Agostino et al. Framingham model [49] with each risk measure are shown in Table 3.2.

TABLE 3.2: Regression Coefficients specified by the Framingham Study and used in D'Agostino et al. calculator

Measurement	β
Women	
Log of age	2.32888
Log of total cholesterol	1.20904
Log of HDL cholesterol	0.70833
Log of SBP if not treated	2.76157
Log of SBP if treated	2.82263
Smoking	0.52873
Diabetes	0.69154
Men	
Log of age	3.06117
Log of total cholesterol	1.12370
Log of HDL cholesterol	0.93263
Log of SBP if not treated	1.93303
Log of SBP if treated	1.99881
Smoking	0.65451
Diabetes	0.57367

After calculating the FRS, thresholds are applied to create risk groups. The applied thresholds to the calculated risk score are provided and used for primary care [28]. In case of missing measurements, the risk score can not be calculated and the labels are not given (subject omitted from the study) unless a known risk factor exists (such as smoker, FH of CVD, Diabetes Prone). Three groups are defined and the available subjects are labeled accordingly:

- **Low Risk (LR):** Subjects with FRS $<$10% (212 participants).
- **Medium Risk (MR):** Subjects with 10%$\leq$ FRS $<$20% (14 participants).
- **High Risk (HR):** Subjects with FRS $\geq$ 20% and subjects with unknown FRS but have one or more risk factors (smoker, FH of CVD, Diabetes Prone) (10 participants).

The labeled risk groups clearly exhibit high class imbalance, where the sizes of low, medium and high risk groups are 212, 14 and 10 respectively. Such characteristic imposes difficulties on learning [6], which mandates the use of an appropriate method to alleviate the effects of imbalance. Such difficulties arise since most of the classifiers are accuracy driven and assume equal or similar distribution of input classes [6]. Accuracy driven classifiers can simply maximise their performance as classifying all samples as the majority class.

Another essential aspect is the consistency and separability of the classes samples within the feature space. Three recognised criteria are used for this purpose namely: Silhouette index (S), Davies Bouldin (DB) index and Calinski-Harabasz

(CH) criterion [135]. Silhouette index measures the consistency within classes. S can take values ranging from -1 to 1 where 1 indicates perfect consistency within classes. While DB and CH assess the scatter within the classes relative to the separation in between them. DB uses the ratio between the intra- and inter- class distances, a lower value of DB shows better compactness and separation between classes. CH compares inter- to intra- class variances, where a higher value reflects better clustering of data points.

The values of the criteria are reported in Table 3.3. The indexes values (negative S and particularly low CH) reveal lack of consistency and representativeness of the classes samples. Figure 3.5 shows the distribution of each class samples using Venular Minimum Constriction for Flicker 1 (MC_VF1) and Arteriolar Maximum Dilation for Flicker 1 (MD_AF1) to illustrate the degree of overlap within the feature space. The small sample size aggravated by the skewed class distribution and coupled with features ranges overlap lead to low separability of the classes. In addition, the low cohesion per class may indicate the presence of disjoint features sub-spaces within each class [101] magnified by the sparse representation of samples, where each minority sample represents a disjunct subspace. Hence, the methods to be applied need to handle the presence of possibly overlapping (inter-class) and disjoint (intra-class) spaces and enhance the representativeness of the samples relative to the classes. Also, the applied methods need to account for the small sample size.

TABLE 3.3: Original Classes Partitions Quality Evaluation

	Original Classes
Silhouette	-0.12
DaviesBouldin	6.80
Calinski − Harabasz	0.76

Illustration of representative samples' responses is shown in Figure 3.6 to enable the visual inspection of the differences between the retinal vessel responses of different risk groups. A representative arterial and venular flicker response for each class is illustrated. Instead of calculating a simple average of flickers in each group, a centroid is chosen to avoid the effect of outliers which may influence the average and also to have a real participant as representative. The centroid (both arterial and venular) flicker response of each risk group is designated as the flicker of the participant with the minimum distance to all participants' flickers within the same risk group. The Fréchet distance [2] [84] is employed to calculate the distance between curves. The averaged responses of the representatives for each risk group are shown in Figure 3.6. The shown response is calculated as the average of of the three flickers responses ($F1$, $F2$ and $F3$) (previously illustrated in Figure 3.4). The arterial response of the low risk and medium risk groups exhibit the 'two humped' dilation response previously reported by Lanzl et al. [124]. The behaviour of the vessels of the high

[2]Fréchet distance: defined as the minimum cord-length sufficient to join two points traveling forward along two different paths, where their rate of travel is not necessarily uniform.

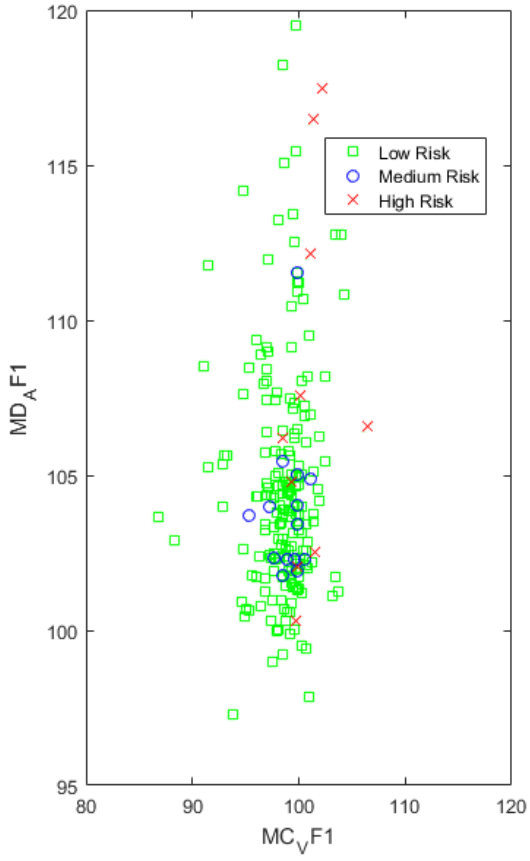
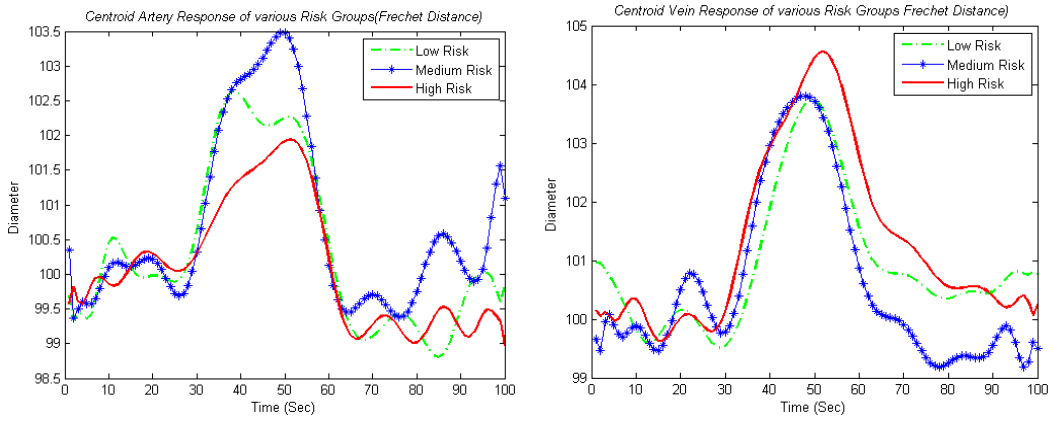


FIGURE 3.5: Samples two-dimensional distribution given two randomly chosen features namely MC_VF1 and MD_AF1



(A) Arterial Response (B) Venular Response

FIGURE 3.6: Vessels Averaged Response of flickers F1, F2 and F3 responses for Centroid Representatives per Risk Group

risk representative shows loss of elasticity expected with reduced vascular health. The artery has a restricted dilation reaction while the vein overdilates. The vein response shows delayed recoil of the vessels. The shown responses are in line with previous findings [131], where narrowed arteriolar and widened venular calibres are associated with cardiovascular problems and helps assure the existence of response differences between risk groups.

3.3 Identified Essential Characteristics of the RVA Data

After examining the generated RVA-based measures (features) and the resultant labeled groups, various characteristics of the data can be ascertained, which hinder effective risk group prediction through direct application of standard prediction (classification) methods. These characteristics are summarised as:

1) High Dimension of interdependent features with the possibility of being irrelevant, which lowers the model's performance and reduces its understandability. Therefore, dimensionality reduction is required to eliminate non-informative features.

2) Small sample size, especially of the critical risk groups of medium and high risk, which can jeopardise these participants for whom an accurate risk prediction is crucial.

3) Imbalanced classes with sparsity of the minority classes samples, which can lead to the treatment of critical minority samples as noise by the prediction algorithm. Thus, a method is needed to increase the representativeness of the samples and account for the small sample size and skewness.

4) Overlapping class boundaries, which is when combined with class imbalance increases the difficulty of prediction.

Another characteristic of the available data is that new data will be continuously collected to allow further validation, hence the developed model needs to easily adapt to the newly arriving data.

3.4 Handling of the Identified RVA Characteristics

Overall, the presented RVA characteristics require the application of a variety of machine learning techniques. The aim of applying these methods is to enhance the dependability of the available data for effective risk prediction. Chapter 4 section 4.1 provides a description of the common approaches to handle the identified RVA characteristics with a justification for the most suitable approaches for our study.

The criticality of cardiovascular risk prediction and the involvement of medical experts in the process entail further requirements on the ML methods to be applied. Requirements such as transparency and high performance are crucial in medical practice. These main requirements drive the formulation a set of criteria to assess existing ML methods. A framework is introduced in section 4.2 for establishing the

suitability of the various methods within each chosen approach. The methods are evaluated based on the framework (defined in section 4.2) and the most appropriate ones are selected to be applied. The evaluation of the available methods, as will be shown in sections 4.3, 4.4 and 4.5, motivates the development of other purpose-built techniques to handle the RVA data.

Chapter 4

Predictive Data Mining

At the beginning of this chapter, a description of the stages of the predictive data mining is given. Approaches that tackle the characteristics of our data are overviewed, then the suitability of the approaches chosen to be applied, namely oversampling, feature selection and instance based learning, is declared (in section 4.1). A set of criteria will be specified in section 4.2 to evaluate the suitability of the reviewed methods. The state of the art within the approaches of prediction, oversampling and feature selection is detailed with an evaluation on their suitability to the RVA data.

The prediction process consists of an interactive and iterative (feedback generated) procedure. The incorporation of expert knowledge into prediction represents the interactive aspect of the procedure. The iterative approach of the prediction procedure stems from the need for high performance and accurate results. Internal and external evaluation [144] of the prediction model may be required to update the model according to its performance. The general outline of a prediction process largely conforms to the Knowledge Discovery in Database (KDD) process steps [61]. The procedure can be summarised in the following steps.

- Data preprocessing and cleaning: In this step, the aim is to promote the dependability of the data in terms of usability and reliability. Outlier removal, noise handling, data sampling and missing values filling are examples of the commonly applied methods. Outlier removal and noise handling can help attain high consistency (reliability) while data sampling and missing values filling enhances the usability of the data [144].

- Data Reduction: The selection step can be viewed from two perspectives [26]. The first perspective is reducing number of features through feature construction or feature selection. The other perspective is choosing the cases that aid the prediction process and this is often called example (subject) selection.

- Prediction: This stage is the core stage in the predictive data mining process, where actual outcomes are inferred from the given features through the learned models. The learned models can be constructed through classification or regression models. Classification is used when the predicted variable is binary or categorical and regression is used for the prediction of continuous variables.

- Assessment and Evaluation: Performance is measured using a set of traditional and novel measures [178].
- Interpretation: The possible interpretation of the mined patterns through expert examination would allow establishing which patterns can be considered as valid added knowledge. Expert knowledge can account for individuals' and populations' variability in health phenomena, while statistical measures may fail in explaining such situations. Also, the approval of a risk model by an expert is crucial for launching its wide practical usage. Relevant valid healthcare specialist interpretation is needed to assure the value of mining health data.

Prediction and evaluation are compulsory steps in the process while data preprocessing and cleaning and data reduction steps are optional steps depending on the nature of the input data. Also, the evaluation and interpretation steps are often merged.

The prediction performance can be evaluated through a set of measures that can be calculated from the confusion (contingency) matrix shown below:

TABLE 4.1: Confusion Matrix 2x2 displaying outcome of prediction

		True Class	
		Positive	Negative
Predicted Class	Postive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Ngative (TN)

Examples of the most commonly used measures are:

- *Overall Accuracy* (OA) : which is defined as the number of correctly classified samples relative to the total number of samples.

$$OA = \frac{TP + TN}{TP + FP + FN + TN} \tag{4.1}$$

- *Sensitivity* (Sn) also known as *Recall* (R) is calculated as:

$$Sn = \frac{TP}{TP + FN} \tag{4.2}$$

- $Specificity\ (Sp)$ also known as *True Negative Rate* is calculated as:

$$Sp = \frac{TN}{FP + TN} \tag{4.3}$$

- *Precision* (P) also known as *Positive Predictive Value* is calculated as:

$$P = \frac{TP}{TP + FP} \tag{4.4}$$

- $F - measure$ sometimes called F_{score} which is the harmonic mean of precision and sensitivity and calculated as:

$$F - measure = 2 \times \frac{P \times Sn}{P + Sn} \tag{4.5}$$

- $G - measure$ sometimes called $G - mean$ which is the geometric mean of precision and recall and calculated as:

$$G - measure = \sqrt{P \times Sn} \tag{4.6}$$

- *Area Under ROC Curve* (AUC): The ROC curve plots the sensitivity against (1 - specificity) using different decision thresholds. AUC measures the goodness of predictions and how well the prediction model separates groups.

- *Area Under Precision* $-$ *Recall Curve* $(AUPRC)$: The PRC shows the relation between the P and R measures for model performance . A high $AUPRC$ shows both high precision and high recall.

4.1 Common Approaches for Data Preprocessing, Reduction and Prediction

When dealing with RVA, similarly to medical datasets in general, one faces several characteristics that may demand special processing. Examples of these characteristics are the presence of missing values, imbalance, high dimensionality and continuous data collection during the course of the study due to recruiting new volunteers to increase sample size.

For handling the issue of missing values, the common approaches are record deletion, imputation and maximum likelihood estimation [185, 8]. Since the percentage of missing data in our study is less than 5% (precisely = 2.97%), therefore it is considered insignificant [56] and does not need special attention. The most significant characteristics are 1) high dimensionality with possibly irrelevant features; 2) imbalanced classes and small sample size; 3) overlapping class boundaries; and 4) expanding dataset. Hence, we will focus on the approaches for handling these four characteristics.

4.1.1 Handling of High Dimension and Irrelevant Features

There are two major ways to address the first characteristic and reduce the dimensionality of a dataset: feature construction and feature selection. Feature construction [113] works by mapping the original feature space into a lower dimension space where each transformed feature is a linear or non-linear combination of a group of original features. The constructed features are aimed to have higher discrimination significance. Notable feature construction methods include Principle Component

Analysis (PCA) [1], Independent Component Analysis [193], Singular Value Decomposition [107] and Linear Discriminant Analysis (LDA) [48]. Constructed features have a serious limitation, especially for medical data: they lose any physiological meaning, thus models based on constructed features are difficult or often impossible to interpret by medical professionals. In addition, feature construction does not identify the key features or show the relative importance of the features, which is required for enriching medical knowledge.

Instead of constructing features, feature selection [26] simply provides a reduced set of original key features, enabling easier model interpretation. When applying feature selection, one can expect benefits such as model with enhanced generalisation ability, reduced risk of over-fitting, enhanced process performance, increased algorithm accuracy and reduced computational cost [26]. The interaction between the feature selection method and the learning algorithm is an important criterion for categorising feature selection algorithms. According to this criterion, there are three categories, namely, embedded, filter and wrapper feature selection algorithms. Embedded feature selection is incorporated within the learning algorithm where an explicit function is used for feature evaluation. In case of filter algorithms, feature set reduction is done as a pre-processing step before learning. Wrapper algorithms are regarded as wrapping around the learning process where the learning algorithm is used to evaluate the selection method. Chandrashekar et al. [37], present several successful feature selection approaches and give examples of these approaches, which we will discuss later in section 4.5 of this chapter.

4.1.2 Handling of Imbalanced Classes and Small Sample Size

As for handling imbalanced classes, a dataset exhibiting an unequal distribution of classes can be adjusted by sampling methods (under– and oversampling) and learning methods. Given a skewed two-class dataset, undersampling removes samples belonging to the majority class [212]. Although undersampling is fast, it may lead to loss of information and is not suitable for highly skewed datasets with extremely small minority classes. Oversampling is based on either 1. the random replication of minority class instances or 2. synthesising of new instances derived from minority class instances. As oversampling inflates the dataset, it may lead to increased computational cost. However, this is not a concern in this study as increasing the sample size is actually required. Thus, oversampling is considered to provide a solution for small sample size as well. Another concern that may arise with oversampling is the validity of the synthesised samples as true representatives of the generating class.

Common learning methods to handle class imbalance are cost-sensitive learning (CSL) [200] and one-class learning. CSL is based on the assumption that correctly classifying the minority class instances is more important than correctly classifying the majority class instances. CSL can be performed by data space weighting or cost-sensitive classification [79]. Data space weighting assigns different misclassification

costs to the dataset; cost-sensitive classifiers work by combining an ensemble of classification methods with cost-minimising techniques. Cost-sensitive learning uses all the available data and hence no information is lost. At the same time, tuning the misclassification penalty might be difficult as this information is rarely available upfront. In One-class learning, the classifier (e.g. one-class SVM) is trained with a single target class. Then, the class of a testing (query) sample is established by determining its similarity to the target class through a predefined threshold. This method is shown to achieve better predictive power than standard multiclass class learners, but it is not practical with multiple minority classes.

We regard oversampling as the most appropriate approach for our study as it provides a solution for both imbalance and small sample size and its approach is simpler compared to the learning methods approach.

4.1.3 Handling of Overlapping Classes and Continuous Data Collection

For classification and prediction, eager and lazy instance-based methods can be used. In eager learning, classification models are constructed from training samples then applied to the test sample set. The learned abstraction models are not easily adaptive to newly arriving data, due to the separate model construction stage. Hence, when new data is collected a new model is built. Coinciding feature ranges can also degrade the performance of global abstraction models.

Alternatively, for lazy learning, also known as instance-based learning, the classification decision for a test sample is based on its neighbourhood. At testing time, local arrangements are constructed using similar training instances in the vicinity of a test sample, to induce a classification decision for the test tuple. Therefore, lazy methods are particularly suited for studies where data collection is expected to extend for long periods. Moreover, lazy non-parametric methods do not rely on distribution assumptions. Hence, lazy methods can provide a suitable solution to some slightly skewed datasets where distribution estimation is not possible. Moreover, it was shown by Xiong et al.[209] that lazy distance-based classification (such as K-nearest neighbour algorithm) better handle class overlap than rule-based methods (C4.5 Decision Tree). Thus, we consider that instance-based lazy learning would offer an adequate solution in the case of overlapping feature value ranges and new data are continuously collected. This is because instance-based learning has the ability to better adapt its classification decision to previously unseen data [199] and base its decision on the individual sample neighbourhood.

In this study, we focus on the approaches of lazy learning, oversampling and feature selection to handle the challenges set by our RVA data. Prediction methods potentially suitable for our data are reviewed, next lazy methods are discussed to illustrate their potential applicability and limitations when applied to our problem. Oversampling provides a feasible solution to the imbalance problem as our dataset is relatively small, hence the performance of our solution will not be significantly impaired by the added computation cost. Feature selection approach is selected due

to its intrinsic advantage of providing a reduced set of human interpretable features which is highly recommended within the medical domain. In the next section, the framework for assessing the suitability of the various methods is described. Following the evaluation of the methods, hybrid approaches that combine oversampling and feature selection are discussed.

4.2 Framework for Assessing Methods Suitability

When developing a prediction model, general criteria such as high accuracy and generalisation of the model apply. For cardiovascular risk prediction using RVA-based measures other additional criteria and requirements apply. These added requirements either stem from the characteristics of the data and/or the recommendations of medical experts. Details of the requirements for each method of the cardiovascular risk prediction solution are presented next.

4.2.1 Prediction Method

The prediction model to be applied has to show:

a) *High performance* : In case of prediction, users seek solutions that offer accuracy as high as possible. Often, several approaches are applied and tested and the one with the highest accuracy is chosen. This is especially important for cardiovascular risk prediction, as an erroneous prediction may lead to detrimental effects on participants well being.

b) *Transparency* : The degree of transparency of the prediction model highly influences whether medical experts accepts to use it. Transparency of a model can be defined as the ability of the user to understand how the patterns were generated and explain how the conclusion (prediction judgment) was reached. This property increases the user's confidence in the model and therefore the likelihood of using it.

c) *Generalisation and ability to accommodate expanding datasets* : Since the collection of medical data is sometimes expensive, as a result only a limited sample size is initially available for exploratory studies. At the same time, the data is continuously collected to expand the dataset and increase the samples reliability. Thus, it is desirable to have a model that is able to reliably predict risk on the initial data set, as well as readily accommodate expanding data sets during the course of the study.

d) *Effectiveness at handling overlap and inconsistencies*: Classes overlap and inconsistencies is a problem known to impose difficulty on the prediction model. Since our RVA-based measures were shown to overlap across the classes, a prediction model that effectively handles this property is required.

4.2.2 Oversampling Method

The oversampling method needs to retain all the information that can be derived from the available samples and verify the validity of the generated samples by fulfilling the below requirements:

a) *Consideration of all minority samples for synthesis* : Such requirement aims at maintaining the information provided by all the original real instances in the oversampled dataset to avoid loss of details. Also, the inclusion of all minority samples for synthesis avoid secluding borderline samples (if only inner instances are oversampled) and/or reversing the features values distribution (if only borderline instances are oversampled).

b) *Suitability to Medical data* : Several assumptions of some oversampling methods are invalid, when handling medical data. Feature independence is assumed in some methods, while in many medical datasets (including our RVA) features are interdependent.

c) *Independence of a specific classifier* : Separating the oversampling method from the prediction (classification) process allows better generalisation of the generated samples, hence lead to improved representativeness of the resultant data set.

d) *Post-Oversampling Validation* : A major concern when oversampling medical data is whether the synthesised sample truly represents its target class and would be a valid sample (mimic a real sample). Hence, a post validation step for the generated samples is needed to address this concern.

4.2.3 Feature Selection Method

The feature selection method would preferably provide a clinically informative and highly performing reduced set, hence this could be achieved through:

a) *Measurement of predictive performance together with theoretical heuristic relevance* : The merge between statistical relevance and actual predictive performance aims at combining the benefits of filter and wrappers methods to attain a highly informative feature subset [92].

b) *Accounting for feature interaction and the combined effect of features* : The available RVA data exhibit multi-way inter-dependencies and interactions between the generated features. Thus, a feature selection that could capture these dependencies and their combined effect is required.

c) *Independence of a Specific Classifier* : Similar to requirement c) for oversampling, feature selection being independent of a classifiers is believed to select features that represent the target concept in absolute terms (better generalisation) rather than optimise the performance of a given classifier [150].

d) *Production of a Ranked List of features* : Feature selection algorithms that output a ranked feature list aid medical experts in interpreting the relative importance of the features. Also, methods that produce feature subsets often depend on a user

defined limit for the size of the subset, which may hinder the discovery of more optimal search spaces [143].

In conclusion, for prediction the model has to present high performance, transparency, generalisation ability and class overlap handling. For oversampling, the method needs to include all minority samples, accounts for feature interdependence, generates samples independent of a specific classifier and performs post oversampling validation. Moreover, the feature selection approach is required to combine the advantages of filter and wrapper methods, account for feature interaction, be independent of the classifier and produce a ranked feature list.

4.3 Prediction

Different approaches of regression and classification for prediction will be reviewed and their suitability to the studied problem will be discussed.

4.3.1 Regression Analysis

Regression analysis [163] constructs a mathematical model relating a set of features called independent variables to a dependent variable (outcome). The dependent variable can be categorical or continuous. In linear regression, a straight line relation between a single independent variable and the dependent variable is assumed. Later, this concept was extended to model the relation between multiple variables and the outcome, which is known as multiple linear regression. Another extension was to assume non-linear relation between the independent variables and the dependent variable, leading to polynomial regression analysis. In polynomial regression, a curved relation is assumed between variables. All these analysis methods accept numeric input variable and output a continuous outcome. Regression analysis allow the identification of relevant risk factors and the calculation of risk scores. It can characterise the relationships among multiple factors. On the other hand, for the regression model to be robust and informative it needs independent relevant variables to be input to the model with an adequate sample size [166], which is not always the case in most applications. In addition, although the mechanism of constructing of regression models is explainable, the reasoning behind each sample decision is not perfectly clear to a medical expert. However, regression presents high performance and handles inconsistencies in data relatively well. Therefore, the properties of regression based on the defined suitability criteria can be summarised as exhibiting high performance, handling inconsistencies and providing a partially explainable model.

4.3.2 Naive Bayes

Naive Bayes (NB) is a probabilistic supervised classification approach [102] that can be viewed as a special type of Bayesian network. The term 'naive' describes two

simplifying assumptions that underlie the model: conditional independence of predictive features and absence of hidden or latent variables that would influence the prediction process. Continuous variables are assumed to follow a normal distribution within each class. Maximum likelihood estimation is used to estimate the mean and the standard deviation of the distribution. To classify a test instance, the simple Bayes rule is used to calculate the probability of each class and it is assigned the label of the class with the highest probability. Despite the simplicity of the approach, NB provides comparable results to more sophisticated and complex algorithms depending on the problem and the data. Hence, it is applied in several applications when a simple, yet effective model is sought as it also provides a relatively understandable model. Thus, according to our criteria NB presents (possibly) high performance that is data dependent and an understandable procedure for model construction.

4.3.3 Decision Trees and Random Forest

Decision Trees [32, 136, 138, 157] depend on a 'divide and conquer' approach for prediction model construction, they offer structural descriptions of what is learned. They provide a visual representation of the attributes that are considered relevant to the model and the decision making process. Decision trees comprise decision and leaf nodes. At a decision node, an attribute is selected to be tested. Several criteria can be used to select an attribute. For example, information gain [204] can be used for this purpose. For a given decision node, it is calculated for each candidate attribute and the attribute with highest gain is selected for a given decision node. Gain can be interpreted as the informational value (partitions' purity) of creating a branch on a specific attribute. Other measures can be used including information entropy and gini index [62]. At a decision node, the aim is to split the input data into subgroups of relatively pure class labels. A branch is created for each possible value splitting up the dataset into subsets, one for every value of the attribute. The process is then repeated recursively for each branch, using only those instances that correspond to the branch. The construction of the tree stops when a group of instances have the same classification label. After tree construction, an instance attributes' values determine the path it follows between the decision nodes until it reaches the leaf node and acquires its class label.

Decision Trees provide knowledge representation that is easily interpretable by humans [153] and as a result provide good support to expert decisions. Also, they present high performance in terms of the number of correctly classified instances. On the other hand, decision trees will encounter a problem handling missing values [204] (if not filled in the data preprocessing step). Any instance with a missing attribute value will fail to be classified i.e, reach a leaf node if that attribute happens to be tested on the route followed by the instance in the decision tree. This is a significant limitation in the application of decision trees in practical problems. Several modifications are suggested in the literature such as treating a missing value as an attribute value and constructing a route for it. Also, imputing missing data can offer

a solution to this limitation. Another problem is over fitting [204] which may lead to serious degradation of decision trees performance when tested on a different dataset than the one used for model construction. Pruning can be used to reduce over fitting by eliminating certain sub trees and replacing them with a leaf node carrying the majority class label.

Random Forest [31] can be considered as an ensemble classifier where a large number of random trees are constructed and each tree votes for the assigned class for a test instance. In a random tree, random selection of features is performed to determine the decision split. The simplest procedure adopted for growing a random tree is to select at random, at each node, a fixed number of input features to split on, grow the tree using Classification and Regression Trees (CART) methodology [33] to and do not allow pruning.

Given the identified suitability criteria, the advantages of decision trees and subsequently Random Forest can be summarised as: (a) Perform implicit feature selection during model building, (b) Provide models that can be interpreted by human experts and (c) Produce high accuracy results. On the other hand, their disadvantages are: Constructing complex models in case of large number of features, instability in the constructed model when new data becomes available and vulnerability to data inconsistencies.

4.3.4 Artificial Neural Networks

Artificial Neural networks (ANN) are said to imitate the brain function through simple computation nodes (neurons). The neurons are simple processors interconnected through links of adjustable weights. The weights of the links are updated in the learning phase of model construction. ANN have different structures depending on the node layers and the interconnection between them. The node layers can be categorised as input layer, hidden layer and output layer.

MultiLayer Perceptron networks (MLP) and Radial Basis Function Networks (RBF) [78] are examples of the most extensively used networks. MLP are supervised feed forward neural networks that require a specified output for training. MLP consists of at least three layers of nodes (input, hidden and output layers) and may include one or two hidden layers. Each node, in the hidden and output layers, is a neuron that uses a nonlinear activation function that is used to transform the input data to the specified output. MLP adopts a supervised learning technique called backpropagation for training. Its multiple layers and non-linear activation leads to the capability to distinguish data that is not linearly separable. RBF neural network has an input layer, a hidden layer and an output layer. The neurons in the hidden layer contain Gaussian (most commonly used) transfer functions whose outputs are inversely proportional to the distance from the center of the neuron. Radial Basis Function networks may operate in the supervised or unsupervised mode. The supervised approach produces better results and both are faster and require less training samples than other ANN [204].

Restricted Boltzmann machines (RBMs) are stochastic artificial neural networks that can learn probability distributions of their set inputs or the joint distribution between a set of variables and a target outcome variable. RBM layers can be stacked to construct deep learning models called Deep Belief Networks [95]. RBMs were initially developed as generative models then discriminative variants were introduced. RBMs consist of layers of visible and hidden nodes. Connections between nodes of the same layer (in-between hidden nodes or in-between visible nodes) are not allowed, giving it the restricted sense compared to Boltzmann Machines. This property allows the use of more efficient training algorithms such as gradient-based contrastive divergence algorithm [36]. RBMs are known to capture the interactions between the input feature and to implicitly determine the significant features within the learning process. Therefore, RBMs have been applied successfully for classification [126] and dimensionality reduction [85] through feature construction in the hidden layers. Despite their success in dimensionality reduction, the reduction process lacks transparency to the user. This is because the interaction and the significance of the features are determined by the weights of the nodes connections during the learning process in the hidden layers, hence they are not readily output to the user. Also, the standard architecture used does not provide direct association between the reconstructed features and the output class. As a result, the significant features used in learning are not explicitly ranked as part of RBMs standard output.

When evaluating ANNs [10] against the suitability criteria, we find that ANNs are considered one of the highest accuracy algorithms. Also, they handle noisy data and new instances adequately. On the other hand, the accuracy of ANN depends on several parameters that need to be set such as the number of hidden layers and the number of neurons [214]. The selection of these parameters is a complex task since it is application specific and the ANNs performance is sensitive to these parameters [18]. Also, the model produced by ANNs lack transparency, which makes it particularly inappropriate in health care applications.

4.3.5 Instance-Based (IB) Learning

Within the lazy approach, the k-Nearest Neighbour algorithm (kNN) is an example of a well established lazy classifier. In kNN [5], the k nearest neighbours of a test instance are determined. The nearest neighbours are located within the training samples set using Euclidean distance and majority vote from the k-neighbour is used as the classification decision of the test instance. Efforts have been dedicated to improve the performance of kNN and study the different aspects of the algorithm. Examples of such aspects are: introducing decision rules other than majority vote, using more efficient search strategies for neighbours search, determining the best value of k and investigating the effect of the utilised distance function on performance.

Shang et al. [173] combined fuzzy set theory with classical kNN. The effect of the neighbouring samples is weighted by their distance to the test instance. A fuzzy

membership is computed to every class based on neighbours membership weighted by their distance. The test sample receives the class label of the highest membership. Another attempt for decision rule improvement can be found in the study of Kaveh et al. [110]. The distance of the neighbours is weighted by the size and dispersion of their class, where neighbours which belong to larger and more dispersed classes are allocated a higher weight. This is applied to determine the impact of the neighbours more accurately.

Kaveh et al.[110] address the issue of search space reduction using linear discriminant analysis. An approach for efficient space search was introduced [194] using particle swarm intelligence to determine k-nearest neighbours and eliminate outliers quickly.

For determining the best value of k, the most common strategy is cross validation (brute force) [191]. However, Wang et al. [191], determine k locally using statistical confidence, while Hassanat et al. [77] employ ensemble classification to reduce the influence of a single k selection. Hassanat et al. apply weak kNN classifiers of different k followed by weighted sum rule to combine the classifications of the weak classifiers.

The effect of the utilised distance function was studied by Hu et al. [94] who showed that the performance is dependent on feature data types of the dataset. Moreover, Bao et al. [16] combined several distance functions such as heterogeneous Euclidean-Overlap metric and discretised value difference metric to determine different k-nearest neighbours groups, then applied majority vote. The K^* algorithm [42] used entropic transformation function to determine the samples similarity. K^* was shown to handle categorical data better than kNN due to its similarity function.

Another line of study was to merge the concept locality (nearest neighbours) decision with Naive Bayes (NB) classifier [65], where a Naive Bayes model is constructed locally (LWNB) based on k-nearest neighbours test samples. Another similar variant is presented by Xie at al[208], where multiple NB models are locally constructed with different K. Then, the most accurate model is selected to classify a test instance.

According to our criteria, we evaluate lazy learning as a candidate solution. In lazy learning, simple understandable models are constructed that perform well in many applications. Moreover, they can accommodate new data when they are collected. However, existing lazy learning methods, purely local approaches, are vulnerable to noise and they remain the most prevalent. In these purely local approaches, however the global resultant structure is overlooked. Hence, an improvement on the existing models can be sought.

4.3.6 Reviewed Prediction Methods Evaluation

Table 4.2 assesses the presented prediction methods according to the criteria introduced in subsection 4.2.1. From the outlined comparison, it can be seen that Instance-based (IB) learning methods partially fulfills the highest proportion of the

requirements. However, their performance is considered sensitive to the data on which they are applied. Hence, an enhancement to the existing instance based models is needed to address this issue and improve the performance of the approach.

To this end, we combined the concepts of lazy mining (classification) algorithm and Graph Cut Optimisation into the *GCO_mine* algorithm. Graph Cut Optimisation [30] is a combinatorial optimisation technique that relies on max flow/ min cut principle [29] to minimise the formulated problem energy function. The proposed algorithm aggregates local connectivities into a globally connected graph on which a global classification decision is taken. The *GCO_mine* approach strikes a favourable balance between merely local instance-based lazy methods and the eager techniques which build global latent models of the training data in a separate phase.

TABLE 4.2: Reviewed Prediction Models Properties with respect to the defined Design Requirements and Suitability Criteria

Method	a) High Performance	b)Transparency	c)Expanding Dataset	d) Handle Inconsistencies
Regression [163]	✓	*		✓
NB [102]	*	✓		
DT and RF [157, 31]	✓	✓		
ANN [78]	✓			*
IB [5]	*	✓	✓	*

*Can be achieved dependent on the data and the application / To some extent
✓Present property

4.4 Oversampling

Oversampling enlarges the minority class through generation of synthetic samples. Although this would change the original prior probabilities of the classes, this is necessary in some cases (where the datasets exhibit highly skewed distributions), since the classifiers treat the minority class as noise and completely fail to classify them. The effect of prior probabilities alteration is likely to be ameliorated by the synthesis of good representative samples to enable the separation of the different categories and enhance the generalisation of the model. However, classifiers that are dependent on prior probability of classes might be less suitable to be used in conjunction with oversampling compared to other methods such as non parametric instance-based methods [165]. Various oversampling methods are reviewed and evaluated according to the criteria specified in subsection 4.2.2, at which we identified that the methods need to: use all the minority samples for synthesis, maintain the features relations and dependence for the new sample, be unbound to a specific classifier and perform post synthesis validation.

The simplest oversampling approach is Random oversampling (ROS), which replicates existing minority instances to achieve an acceptable class balance ratio. Although this technique is easy to apply, it suffers from a high risk of overfitting, therefore different heuristics are usually used in conjunction with oversampling .

One of the most considered heuristic approaches is the Synthetic Minority Oversampling TEchnique (SMOTE) [38], which generates equal numbers of synthetic samples for each minority data example considering all minority samples. The synthetic samples are generated on the straight line interpolation of two minority instances, which preserves the feature relations. Obviously, SMOTE fulfills three of the defined requirements as it is independent of the classification, uses all samples and regards for feature dependence.

Han et al. [76] limited the SMOTE oversampling process to borderline samples as these are more likely to be misclassified and called their approach Borderline-SMOTE. Borderline samples are determined as instances with more majority class neighbours and fewer minority class neighbours within a sample's neighbourhood, while samples having only majority class neighbours are considered noise and not oversampled. When compared to SMOTE and random oversampling on an artificial dataset and three UCI ML Repository datasets, Borderline-SMOTE showed favourable results in terms of true positive (TP) and F-measure. Borderline-SMOTE, achieves the same criteria achieved by SMOTE except for limiting the samples to borderline samples.

Another approach for handling samples that can be easily misclassified is the ADAptive SYNthetic sampling (ADASYN) [80] algorithm. It generates more samples for the examples that are more difficult to learn, so the resulting dataset forces the learning algorithm to focus on those samples. Nevertheless, all minority class samples are used to generate new synthetic tuples, which better preserves the samples feature values distribution compared to Borderline-SMOTE. The ADASYN oversampling process is illustrated in Algorithm 1. Instances $x_i \in X$ with class labels C are input to the algorithm, where n_{mn} and n_{mj} are the number of minority class and majority class instances, respectively. The maximum allowed imbalance threshold d_{th} and the desired balance level after generating the synthetic data β are preset. ADASYN relies on density distribution $\hat{r}_i$ (line 7) to determine the number of synthetic samples that need to be generated for each minority data example. $\hat{r}_i$ depends on $\triangle_i$, which is the number of majority class samples in the set of k-nearest neighbours of x_i. Thus, the minority class samples that are surrounded by more majority class samples, and are hence more difficult to learn, are used to generate a proportionally greater number of synthetic samples g_i (line 8).

An example of the generation process is shown in Figure 4.1, where we assume the presence of a majority class (indicated by black dots) and two clusters of a minority class (given as red triangles) at the borders of the majority class. Sample x_{zi} would be considered as a neighbour of x_i if, for example, the number of neighbours k is set to four. The minority sample x_i and the randomly chosen minority neighbouring sample x_{zi} are used to generate a synthetic sample s_j. The generated samples are at a random distance λ between the original minority samples. A limitation of the ADASYN approach is that it may generate samples within the feature space cloud of a class other than the intended (generating) minority class. Figure 4.1

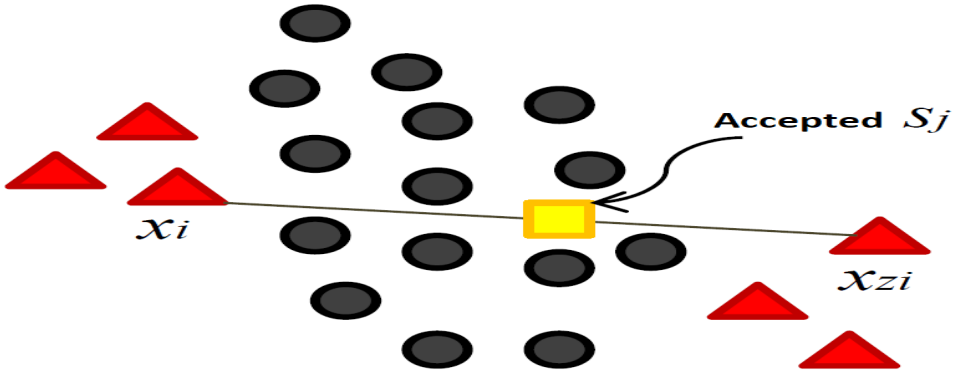


FIGURE 4.1: An example of a synthetic sample (yellow square) that would be created and accepted by ADASYN

illustrates this: the generated sample s_j would be **accepted** by ADASYN although it has been created within the cloud of the majority class. ADASYN was compared against SMOTE in terms of OA , $Precision$, $Recall$, $F-measure$ and $G-mean$ using five datasets. ADASYN outperformed SMOTE in the total number of wins and achieved the best $G-mean$ with all datasets, while realising the same requirements as SMOTE.

Algorithm 1 ADASYN Oversampling Algorithm

procedure ADASYN($X, C, n_{mn}, n_{mj}, d_{th}, \beta$)
  $d \leftarrow n_{mn}/n_{mj}$
  if $d < d_{th}$ **then**
    $G \leftarrow (n_{mj} - n_{mn}) \times \beta$
    for all $x_i \in C_{mn}$ **do**
      Find k-nearest neighbours for x_i
      $\hat{r}_i \leftarrow \triangle_i / k$
      $g_i \leftarrow \hat{r}_i \times G$
      repeat
        Randomly choose one minority sample x_{zi}
        from k-nearest neighbours of x_i
        Generate Sample $s_j \leftarrow x_i + (x_{zi} - x_i) \times \lambda$
      until $j > g_i$
    end for
  end if
end procedure

Majority Weighted Minority Oversampling (MWMOTE) [17] is another method that focuses on difficult to learn instances, defined in this work as borderline samples, intended to overcome ADASYN's limitation. In MWMOTE, minority samples set is first cleaned by excluding samples that are entirely surrounded by majority class neighbours to avoid using noisy data for synthesis. Nevertheless, this procedure may remove informative samples from the minority set that may be representing disjunct subspaces and not suitable for extremely small sample sizes. Then, the cleaned minority set is partitioned using agglomerative clustering. For each border

sample, a selection weight is assigned considering its closeness to other majority samples and the dispersion of the cluster. The results of 20 real datasets and 4 artificial datasets with continuous features only are reported [17]. The results showed promising performance except for Recall. MWMOTE relies heavily on tuning parameters and user defined thresholds (6 variables). Similarly to Borderline-SMOTE, MWMOTE totally neglects samples that are not on the borders, which is likely to further disrupt the original features distribution and limits the information carried by the synthesised samples.

An approach that attempts to keep the original minority feature values distribution and expand class boundaries is called Random Walk Oversampling (RWO) [219]. It generates a new attribute value for each instance and for each attribute, based on the original features mean, standard deviation, number of samples and a sampling value. Given m attributes and n instances, a new attribute value $a_i^{'}$ is generated for each attribute a_i per instance j, The new attribute is calculated using the equation:

$$a_i^{'}(j) = a_i(j) - r_j \times \frac{\sigma_i}{\sqrt{n}}, i \in \{1, 2, 3, ...m\}, j \in \{1, 2, 3, ...n\} \tag{4.7}$$

where r_j is a sampling value of distribution $N(0, 1)$ and σ_i is the standard deviation of the $i - th$ attribute. The new sample is generated by concatenating the synthetic attribute values into a single feature vector. Thus, each new sample is generated by randomly walking from one real sample. Since each real instance is used to generate another synthetic sample and there is no criteria for varying the number of generated samples per instance, the oversampling rate of RWO must be an integer multiple of 100% [219]. RWO was applied on 21 datasets and it showed good results in terms of $OA, F - measure, G - mean$ and *Sensitivity* depending on the classifier used and the oversampling rate adopted. RWO presents classifier independent performance and considers all samples equally for oversampling. However, the generation process assumes independence between the features which is often an invalid assumption for medical datasets. Also, it can be argued that having the generated samples at a random walk from a single sample *on attribute basis* and not being bound between two actual samples would reduce their prospect to mimic real life samples.

A common point between all of the described oversampling techniques is that they are originally formulated and developed for two-class problems. Efforts have been directed to oversampling techniques dedicated to multi-class problems. Dynamic SMOTE Radial Basis Function (DSRBF)[63], Dynamic Sampling (DyS) [133] and Mahalanobis Distance-based Oversampling (MDO) [2] are examples of such techniques.

DSRBF incorporates two stage dynamic SMOTE oversampling into Radial Basis

Function (RBF) learning. DSRBF [63] optimises the performance of RBF neural networks using a memetic algorithm (MA)[1]. First, the training data are oversampled prior to learning, then the MA is run and the samples of the class with the lowest sensitivity in every generation of the evolution are oversampled. Lin et al. [133] present a similar concept with Dys. The minority samples are randomly oversampled with a different rate at each training epoch and assigned a higher probability to train a Multi-Layer Perceptron (MLP) classifier. In this way, the MLP is directed to focus on a different minority class during learning. Although the described techniques directly handle multi-class imbalance, their applicability is bounded to specific learning algorithms.

The more recent approach of MDO, presented by Abdi et al. [2], does not depend on a particular learning algorithm. In MDO, synthetic samples are generated along the probability contours, creating samples with the same Mahalanobis distance to the mean of the class (to be oversampled) as the existing samples. This procedure aims at preserving the covariance structure of the oversampled class and reducing class overlap. To limit class overlap, MDO generates synthetic samples from "safe" instances, which lie in dense regions. However, this may result in jeopardising borderline instances as they become more isolated.

4.4.1 Reviewed Oversampling Methods Evaluation

On the whole, the discussed oversampling methods each provides a plausible solution to synthetic samples generation. Table 4.3 summarises the reviewed methods in terms of the assessment criteria provided in subsection 4.2.2. All the reviewed methods do not validate the representativeness of the synthesised data, which is a critical step for our RVA data to limit the overgeneralisation that can result from oversampling. Both SMOTE and ADASYN fulfill the same set of criteria, but ADASYN focuses on difficult to learn samples and was shown to attain better results [80]. ADASYN generates synthetic samples from both safe and borderline samples with higher proportion to be generated using borderline samples. This allows focusing on rare borderline samples without totally neglecting more dense safe samples, unlike Borderline-SMOTE [76] and MWMOTE [17], which entirely use borderline samples or MDO [2] which uses only safe samples further isolating border samples. Another important aspect that it does not assume independence between features, which is an invalid assumption for our data, in contrast to RWO [219] which has this assumption. Also, the application of ADASYN is not bounded to a specific learning algorithm such as DSRBF [63] or DyS [133], which allows its use with better interpretable classifiers. Hence, we regard ADASYN as the most suitable candidate approach for oversampling. Nevertheless, it shares the limitation of lacking post-oversampling validation with all the reviewed methods, which motivates the development of a method that appropriately addresses post-validation. The proposed

[1]Memetic algorithms: present a class of optimisation algorithms that combine evolutionary algorithms with local search.

method FiltADASYN adds a step for filtering out synethetic samples that are likely to be invalid.

TABLE 4.3: Reviewed Oversampling Methods Properties with respect to the defined Design Requirements and Suitability Criteria

Method	a)Uses All Minority	b)Considers Features Dependence	c)Unbound to a classifier	d) Performs Post-Validation
SMOTE [38]	✓	✓	✓	
BorderSMOTE [76]		✓	✓	
ADASYN [80]	✓	✓	✓	
MWMOTE [17]		✓	✓	
RWO [219]	✓		✓	
DSRBF [63]	✓	✓		
DyS [133]	✓	✓		
MDO [2]		✓	✓	

✓Present property

4.5 Feature Selection

Feature selection can be defined as choosing a subset of features that is most relevant and best describes the target concept of to be learned from the data. Several definitions of relevance are provided by Blum et al.[26]. The definition applied may vary depending on the motivation and approach utilised. Feature selection algorithms are adopted aiming at gaining various benefits such as enhancing generalisation ability of the model and reducing the risk of over-fitting through low dimensional representation. A low dimensional representation can enhance the process performance as it might increase the algorithms accuracy. Also, determining key features helps understand the model better. Feature selection can be viewed as a search procedure with a starting point, scheme of the search, evaluation measure and a halting condition.

The known feature selection algorithms can be categorised based on several criteria, including, (1) the type of training data whether it is labeled, unlabeled or partially labeled, (2) the interaction between the feature selection and the learning algorithm, leading to three algorithm categories namely embedded, filter and wrapper feature selection algorithms and (3) the type of output which can be a reduced feature set or a ranked list of the input features.

In the following subsections, various feature selection approaches are reviewed, including their performance. The used classifiers and the feature selection methods compared against are specified to provide full illustration and allow cross comparisons [2]. The reviewed feature selection methods are assessed based on the suitability criteria given in subsection 4.2.3.

[2]In some cases, the impact of the feature selection approach could not be distinguished from that of the applied classifier due to the limited results reported in the original articles.

4.5.1 Generic Feature Selection Methods

A large number of generic feature selection methods are available, we shall discuss those most relevant to our study. First, traditional and well established algorithms are presented then the recently introduced methods with promising results are discussed.

Filter ranking methods are common approaches for feature selection. Examples of these methods include ReliefF [96], Correlation Coefficient [204] and PCA-entropy [160] feature selection.

ReliefF [96] is a filter ranking method based on Relief [116] for feature selection. The basic idea of the algorithm is to assign a score to each feature depending on its capability to distinguish and separate the given samples. ReliefF feature scoring basically assigns a high score to features that for a given sample have different values from its nearest neighbour from a different class (nearest miss) and similar value to a neighbour of the same class (nearest hit). The ReliefF extension enables the application of Relief on multi-class problems. Moreover, it reliably handles missing data. ReliefF elaborates the idea of nearest miss to k-nearest neighbours from the k different classes. In case of incomplete data, the conditional probabilities, that two given samples have different values for the feature being evaluated, are approximated with relative frequencies from the training set. The scoring function S for each attribute A can be expressed as in equation 4.8, where n_i is the number of instances approximating the probabilities, h is the nearest hit and $m(C)$ is the nearest miss from a class C other than the class of instance I:

$$S(A) = S(A) - diff(A, I, h)/n_i + \sum_{C \neq Class(I)} [P(C) \times diff(A, I, m(C))]/n_i, \quad (4.8)$$

where $diff(A, I_1, I_2)$ calculates the difference in values of attribute A between two instances I_1 and I_2. For discrete values the difference is either 0 or 1, while for continuous values it is the actual normalised difference in the range [0,1]. In case of incomplete data, where I_1 has a missing value, $diff(A, I_1, I_2)$ is calculated as:

$$diff(A, I_1, I_2) = 1 - P(value(A, I_2)|class(I_1)). \quad (4.9)$$

If both instances have missing data, $diff(A, I_1, I_2)$ is approximated as:

$$diff(A, I_1, I_2) = 1 - \sum_{v}^{\#values(A)} (P(v|class(I_1)) \times P(v|class(I_2))). \quad (4.10)$$

As discussed, ReliefF is a filter selection method that produces a ranked list of features with relevance score. Hence, it achieves two of the designated criteria.

Another filter method for feature selection is Correlation Coefficient (CorrCoeff) feature selection [204], where a simple ranking approach is adopted. The ranking method depends on the feature correlation with target class. The feature with the

largest correlation score with the given classes is ranked highest. This approach is quite simple yet effective, especially with balanced data. Pearson's correlation ρ is used such that the correlation between attribute A and class C is calculated as shown in equation4.11:

$$\rho(A, C) = \frac{cov(A, C)}{\sigma_A \sigma_C}, \tag{4.11}$$

where $cov(A, C)$ is the covariance and σ is the standard deviation. For nominal attributes, each value is treated individually by considering each value as an indicator. Then, a weighted average is computed to provide an overall correlation for the nominal attribute. For missing values, the mean value is used for continuous features and the most frequent value for discrete attributes. Similar to ReliefF, CorrCoeff attains the same criteria of being unbound to a classifier and producing features ranking.

Rao et al. [160] presented an unsupervised feature ranking algorithm called PCA-entropy. Representation entropy [53] is used to assess the degree of uncertainty in the data set, hence determine the importance of a given feature accordingly. The idea behind the algorithm is that the importance of a particular feature is directly proportional to the increase in representation entropy (uncertainty) of the data set calculated without that feature. The representation entropy RE_D is calculated for the full data set and RE_{FR} after feature removal for each feature. A score is calculated as $S = RE_{FR}$ - RE_D. The features are ranked in descending order of S. The performance of the ranking algorithm is compared to ReliefF, SUD [51]and SVD-Entropy based ranking [189] using J48 decision tree classifier applied on four benchmark data sets (Iris, Glass, Pima and Bupa) from UCI ML repository. PCA-entropy provided comparable accuracy to SVD-Entropy and SUD on the five datasets. However, ReliefF manifested higher accuracy on all datasets with varying number of features. The method produces a ranked list of features, independent of a specific classifier satisfying two of our recommended suitability criteria. Nevertheless, PCA-entropy presented lower performance than Relieff, which fulfills the same criteria, rendering PCA-entropy as an inappropriate candidate for our study.

Other popular subset filter methods that rely on information theory include: Information Gain (IG) feature selection [128], the classical approach of minimum Redundancy and Maximum Relevance (mRMR)[150], Fast Correlation Based Filter (FCBF) [216] and Dynamic Relevance Joint Mutual Information Maximisation ($DRJMIM$) [93] .

In IG feature selection [128], the mutual information between the features and the classes is measured, then the features are sorted in descending order by mutual information MI. A subset is obtained from the sorted list of features using a user defined threshold specifying the number of required features. The issue that rises with IG is that it only considers the relevance of the features to the class and disregards the redundancy between features. With mRMR [150], this issue is addressed as it introduces a weighted term to minimise feature redundancy and balances the impact of the redundancy term. The mRMR selection criterion $J(.)$ for a candidate

feature X_k is formulated as:

$$J(X_k) = MI(X_k; Y) - \frac{1}{|S|} \sum_{X_j \in S} MI(X_j; X_k), \tag{4.12}$$

where S is the set of selected features subset, X_j denotes a selected feature, X_k represents a candidate feature, Y represents the class and $|S|$ is the size of the subset. The mRMR approach, unlike the previous methods of ReliefF, CorrCoeff and IG, addresses the issue of feature interaction as a criteria for the developed feature selection methods. However, it fails to produce a ranked list of features desirable for medical interpretation.

An algorithm similar in concept to mRMR for feature selection FCBF is proposed by Liu et al. [216], where predominant features are selected and the rest are removed. Predominant features are defined as features with highest symmetric uncertainty [3] to a target class and least redundancy (symmetric uncertainty) to the rest of features, whereas in mRMR mutual information is used for this purpose. The selection process is completed in two passes. The set of features are ordered descending based on degree of relevance to the target concept. Then, starting from the highest relevant feature (Fp) all its redundant peers are removed from the list. This process is repeated until all the features are examined. The performance of the algorithm is evaluated in terms of running time and accuracy on 10 datasets from the UCI ML Repository and the UCI KDD archive. The datasets are chosen to exhibit a wide range of classes number, dimensionality and sample size. Liu et al. computed accuracy for all the experimented datasets, FCBF showed data dependent performance comparable to ReliefF, CONSSF and CORRSF.

Hu et al. [93] propose another filter approach namely Dynamic Relevance Joint Mutual Information Maximisation $DRJMIM$ for feature selection based on information theory, they address two common drawbacks in existing algorithms based on information theory. The drawbacks are: no distinction between candidate feature relevancy and selected feature relevancy and lack of discrimination between feature redundancy and interdependency. In order to overcome these drawbacks, the $DRJMIM$ approach merges the concepts of dynamic relevance [179] and joint mutual information maximisation (JMIM) [22]. For JMIM, the selection criterion $J(.)$ is expressed as:

$$J(X_k) = argmax_{X_k \in F-S}(\min_{X_j \in S}(CI(X_k, X_j; Y))) \tag{4.13}$$

, where S is the set of selected features subset, F is the set of the remaining candidate features, X_j denotes a selected feature, X_k represents a candidate feature, Y represents the class and $|S|$ is the size of the subset, CI is the conditional mutual

[3] Symmetric uncertainty [205] is defined as the normalised mutual information by the product of the entropy of both variables.

information and calculated as:

$$CI(X_k, X_j; Y) = MI(X_k; Y) + MI(X_j; X_k|Y) \tag{4.14}$$

Although similar to mRMR, $DRJMIM$ employs the 'maximum of the minimum' non-linear criterion instead of the linear summation method of mRMR. Also, it uses conditional mutual information, while mRMR uses solely mutual information. The method is applied on 12 real-world benchmark datasets coming from different fields and an artificial dataset. The datasets showed diversity in the number of features, number of instances covering binary and multi class learning. The results show higher accuracy on average and lower numbers of features selected when compared to mRMR. However, the results were data set dependent. Similar to mRMR approach, both Liu et al. and Hu et al. methods accomplish two of the criteria previously defined in subsection 4.2.3, of being independent of the classifier and detecting features dependency.

All of the presented methods are filter methods that rely on theoretic measures for evaluating features importance. Although, these measures have shown good performance, they do not account for the practical predictive performance of the features.

C.Yun et al. [217], an example of a wrapper approach, which selects feature subsets based on the classifier's performance. In addition, C.Yun et al. [217] merge mRMR theoretical feature relevance with the predictive performance of the features, which is of interest in our study. Four feature selection algorithms are proposed namely: Genetic Algorithm [4] Feature Subset Selection (GAFSS), Particle Swarm Optimisation [5] Feature Subset Selection (PSOFSS), GAFSS+mRMR and PSOFSS+mRMR.

The algorithms either rely on genetic algorithms (GA) and particle swarm optimisation (PSO) or combine GA and PSO with mRMR for feature selection. In GAFSS and PSOFSS, the search for the best features subset is performed based on standard approaches of GA and PSO aiming at maximising the classification accuracy. The search is halted when the number of selected features reaches a predetermined threshold, or when none of the alternatives improve the performance.

The other two proposed variants (GAFSS+mRMR and PSOFSS+mRMR) combine mRMR with accuracy for feature evaluation. The features are divided into three categories depending on the mRMR value, this categorisation is used to guide the optimisation process. In GAFSS+mRMR, the mutation process is adjusted to control the inclusion of features of high significance (High mRMR) and the exclusion of the features of low significance (low mRMR) to improve performance of GAFSS. In PSOFSS+mRMR, the random process is modified. The particles category is checked

[4]Genetic Algorithms (GAs) is an adaptive heuristic search technique suitable for optimisation problems. GAs are inspired by natural evolution and genetics of biological organisms.

[5]Particle swarm optimisation (PSO) is a population based stochastic optimisation technique, inspired by social behavior of bird flocking or fish schooling.

and the velocity of the particle is adjusted accordingly. All the experiments were conducted using 20 benchmark datasets from UCI ML Repository chosen to be of variable number of samples, features and classes. Three classification algorithms namely Nearest neighbour, C4.5 decision tree and SVM were applied. The performance of GAFSS and PSOFSS was investigated against five existing methods (mRMR [151], MI [57], I-RELIEF [180], INTERACT [221] and PAM [184]) in terms of the number of features and classification accuracy. PSOFSS accuracy surpassed all its counterparts using the three classifiers in at least 17 datasets. The differences in the number of selected features was not significant. Another set of experiments compared GAFSS and PSOFSS against GAFSS +mRMR and PSOFSS +mRMR. PSOFSS +mRMR outperformed GAFSS +mRMR using NB in 15 cases with differences ranging from 0.63% to 6.45%, while their accuracy using SVM and C4.5 was similar. Moreover, GAFSS +mRMR and PSOFSS +mRMR provided non-significant accuracy improvement over PSOFSS and GAFSS. The GAFSS +mRMR and PSOFSS +mRMR variants presented by C.Yun et al. satisfy two of the suitability criteria defined in subsection 4.2.3 namely combining predictive and heuristic measures for feature selection and accounting for features interaction, while PSOFSS and GAFSS fail to satisfy any of our criteria. However, GAFSS + mRMR and PSOFSS + mRMR selected subsets vary based on the classifier they optimise and do not provide a relevance score, which leads to opacity of the selection.

4.5.2 Feature Selection Methods for Imbalanced Data

Since high dimensionality aggravate the imbalance problem [25], the interest in developing feature selection methods specific to imbalanced data has risen. Various methods were developed and applied specifically on imbalanced data. The developed methods followed various approaches such as: using measures better adapted (invariant) to class skewness [7, 213], combining theoretic and predictive relevance [27, 143] or adopting data segmentation [27, 213].

A Density-Based Feature Selection (DBFS) method is proposed by Albeigi et al. [7]. For each feature, DBFS estimates a separate probability density function per class. Feature are ranked according to a calculated discrimination-ability measure. The measure relies on the heuristic that features with minimum overlapping class probability density functions and minimum class changes are best candidates for selection. Datasets from various domains, including biological (9 datasets), text analysis (13 datasets), and UCI ML Repository [130] (2 datasets) are used for performance evaluation. DBFS is compared to IG [128], chi-square [73], signal to noise ratio [73],and FAST [39] methods applying 1-NN, linear SVM and NB classifiers. AUC and $F-measure$ metrics were utilised for evaluation. DBFS was shown to offer a significant improvement over baseline and surpass the presented standard methods. Considering our set of suitability criteria, DBFS satisfies two criteria, namely it produces a ranked list that is unbound to a specific classifier. Nevertheless, its

performance was only validated on two class problems. Such experimental decision leaves its suitability to multi-class problems questionable with the potential difficulty of handling the overlap estimation of multiple density functions. Also, in many practical scenarios (including our study) the size of the minority class is very small, thus the estimation of the probability distributions would not be practically possible. Hence, the application of this approach in our study is deemed inappropriate

A different technique (SYMON) that merges SYmetric uncertainty (theoretic relevance) and harMONy search [6] (predictive performance) is discussed by Moayedikia et al. [143]. In SYMON, the first step is to rank the features using symmetric uncertainty. Afterwards, the vectors with most significant features undergo a harmony search. The improvised vectors containing real valued features are evaluated according to a given fitness function (such as kappa statistic [175] [7]) and the vectors are tuned and replaced accordingly. The search terminates after a preset number of iterations, where the size of the output feature subset is controlled by a user defined threshold. At the end of harmony search, the best vector with highest fitness is selected. The performance of the algorithm is demonstrated on eight large-scale DNA microarray data sets and one dataset containing images called Olivetti Faces. SVM is used as the underlying classifier with AUC and G-mean as the evaluation metrics. SYMON manifest variable, but highly competitive performance compared to other algorithms of Guyon et al. and Yin et al. [74, 213], also SMOTE combined with Principle Component Analysis (PCA)[1] using the SVM classifier. This method was shown by Moayedikia et al. [143] to be suitable for handling high dimensional data with features of comparable importance and accounts for feature interdependence through symmetric uncertainty. Nevertheless, the vector tuning operations are highly dependent on the required subset size, which hinders the discovery of more optimal parts of the solution space. Also, it shares with other wrapper approaches the limitation of lack of generalisation of the selected features as its fitness is bound to a specific classifier.

Differently from the previous methods, which use the whole dataset at once, Bolón-Canedo et al. [27] proposed a distributed approach. The feature set is partitioned into subsets of size k. The subsets are constructed either randomly creating the Distributed Filter (DF) approach, or based on a ranking approach generating the Distributed ranking Filter (DRF) approach, where features of similar rank are grouped in the same subset. A selection from the features subset is obtained using a filter method (such as Correlation Feature Selection, IG, ReliefF. etc), followed by a

[6] Harmony search [67] is a heuristic optimisation technique inspired by musicians' work when composing a new tune, where musicians adjust stored music pitches in their memory to find the perfect harmony (optimal solution). Harmony search has two distinguishing operators harmony memory considering rate (HMCR) and pitch adjusting rate (PAR) that are used to generate and further mutate a solution. The candidate vectors are adjusted depending on their fitness evaluation.

[7] Kappa Statistic: measures the agreement for qualitative (categorical) values between the actual labels and the assigned classifications of a classifier. It takes into account the possibility of the agreement occurring by chance.

merging procedure which updates a feature subset according to the improvements in classification accuracy. The method was demonstrated on eight gene expression datasets with varying number of attributes, samples and imbalance. Extensive experiments are conducted comparing the non-distributed filter methods and a wrapper method to their distributed counterparts applying SVM, k-NN and Naive Bayes Classifier. DRF and DF satisfied a single criteria of the ones defined in subsection 4.2.3, which is considering both theoretic and practical relevance of a feature. Over and above, the decomposition approach is hard to apply on small classes as in our study, which makes this approach in appropriate.

Yin et al. [213] proposes two approaches for feature selection, one based on class decomposition and the other approach proposes the Hellinger distance [8] as a measure for feature selection. The first approach divides the majority class into subclusters using K-means and assigns pseudo-labels to them. The number of subclusters is manually preset to obtain fairly balanced subsets (pseudo-classes), however this does not guarantee achieving the targeted balance. After decomposing the dataset, correlation, mutual information and Fisher score criteria are used to quantify the relevance of each feature with respect to the output pseudo-class label. In this approach, correlation and mutual information calculate the relation of the features to the artificially constructed pseudo classes only. Fisher score determines the discrimination ability of a feature, where the numerator indicates the discrimination between classes and the denominator indicates the scatter within each class. The larger the value of Fisher score, the higher the importance of the feature. The algorithm returns the top n features, where n is a user-defined threshold.

The second approach proposed by Yin et al. is based on Hellinger distance to assess the importance of the features. The Hellinger distance between two sample groups is calculated using a given feature. If the calculated distance is small then this feature is considered of low importance. It is proposed as a feature selection measure because it is insensitive to classes skewness. Five datasets from different application domains were used applying SVM, C4.5 and Bayesian learning. The measures used were $F - measure$ and AUC. The experiments were carried out in two folds, the first fold determined the effect of the class decomposition approach and the second fold compared the proposed Hellinger distance to the other measures (correlation, mutual information and Fisher).

The results of the approaches were inconsistent across the datasets, implying that the performance of the proposed approaches is data dependent. The approaches have two shortcomings when considered with our data. First, the clustering scheme using a pre-defined threshold may lead to inconsistent pseudo classes. Also, the idea of dividing the majority class into multiple pseudo classes of similar size to the minority class is not applicable with extremely small minority classes as it will lead to over-segmentation. As a result, the dependability of the selected features would

[8] Hellinger distance [41] is the probabilistic analog to euclidean distance between two probability distributions.

be degraded. Secondly, the Hellinger distance measure can not readily be applied to datasets with small sized classes as it is hard to estimate a probability distribution from small samples. On the whole, both methods proposed by Yin at al. achieve only the requirement of classifier independence.

4.5.3 Reviewed Feature Selection Methods Evaluation

Table 4.4 summarises the properties of the reviewed feature selection methods in relation to our criteria of subsection 4.2.3. There are no clear winners if we simply count the number of criteria satisfied, 10 methods satisfy two criteria each. However, an important criterion for the feature selection method is to be independent of the classifier. This criterion leads to the selection of informative relevant features that well describe the target concept, which can enhance the existing medical knowledge, rather than features that optimise the performance of a specific classifier. Based on the importance of this criterion, we consider the candidate feature selection methods to be: ReliefF, CorrCoeff, mRMR, FCBF, DRJMIM, PCA-entropy and DBFS. However, mRMR, FCBF and DRJMIM all basically rely on MI for redundancy and relevance computation, MI is declared to be sensitive to noise and outliers [13, 106]. This limits the possible candidate methods to ReliefF, CorrCoeff, PCA-entropy and DBFS. Despite that the four methods satisfy the same criteria, PCA-entropy was shown by Rao et al. to provide lower performance than ReliefF. In addition, DBFS was also previously evaluated (in DBFS review) as inappropriate for our RVA data due to the small number of samples in our minority classes.

Hence, we consider ReliefF and CorrCoeff as candidate solutions because they satisfy the two requirements of classifier independence and production of a ranked list. Both methods are well established and widely used, with readily available implementations, thus would allow comparison with other methods in the future. However, as these methods fulfill our set suitability criteria only in part, we propose a different feature ranking approach that satisfies all the criteria. The proposed method is based on Restricted Boltzmann Machines (RBMs) previously described in subsection 4.3.4. RBMs have not been previously used for the purpose of feature selection, hence they can not be compared here against the reviewed methods for feature selection such as ReliefF and CorrCoeff.

We propose RBMs for feature selection because they are high performing prediction methods capable of detecting features synergy (interaction and interdependence) and modeling the co-occurrences of the features and the output class. Thus, they can be used to select predictive features with high correlation to the target concept. Hence, RBMs-based feature ranking can account for both predictive and heuristic relevance of the features irrespective of the classifier used for actual prediction. Additionally, they have the intrinsic advantages of handling missing data [123] and accommodating large data sets [86], when data become available. The proposed new method will be described in detail in Chapter 7, together with a full evaluation of its performance compared to the two feature ranking methods from

the literature, ReliefF and CorrCoeff, that are most suitable according to our defined criteria.

TABLE 4.4: Reviewed Feature Selection Methods Properties with respect to the defined Design Requirements and Suitability Criteria

Method	(a) Predictive + Heuristic	(b) Feature Interdependence	(c) Independent of classifier	(d) Ranked List
ReliefF [96]			✓	✓
CorrCoeff [204]			✓	✓
IG [128]			✓	
mRMR [150]		✓	✓	
FCBF [216]		✓	✓	
DRJMIM [93]		✓	✓	
PCA-entropy [160]			✓	✓
GAFSS [217]				
PSOFSS				
GAFSS+mRMR [217]	✓	✓		
PSOFSS+mRMR	✓	✓		
DBFS [7]			✓	✓
SYMON [143]	✓	✓		
DF and DRF [27]	✓			
Decomposed Approach [213]			✓	
Hellinger Distance [213]			✓	

✓Present property

4.6 Hybrid Approaches

Traditionally, imbalance and high dimensionality have been addressed separately [25]. A recently emerging approach is to adopt hybrid solutions that simultaneously use oversampling and feature selection to tackle the compound effect of imbalance and high dimensionality [70, 129, 161].

A combination of wrapper feature selection techniques and sampling methods together with k-nearest neighbour (kNN) classifier was investigated with the aim to assess different biomarkers combinations for the classification of Alzheimer's disease [161]. Rodrigues et al. [161] found that kNN with oversampling and forward feature selection resulted in higher precision than Support Vector Machines (SVM) and weighted kNN classifier.

Another hybrid approach is described by Gourdeau et al [70]. The authors used mutual information feature selection and SMOTE oversampling for the analysis of clinical data from extremely preterm infants, in order to determine if they are ready to be removed from endotracheal mechanical ventilation. The results show improved *AUC* and lowered False Positive Rate.

A further example of how feature selection and sampling methods can be paired successfully is the work of Lian et al [129]. Working on cancer treatment outcome prediction, the authors proposed a feature selection method based on Dempster-Shafer theory (also known as Evidence-based theory) that incorporates the prior knowledge of must-include features. For feature selection, a loss function is constructed and optimised by genetic algorithms. The loss function searches for an

informative feature subset according to three requirements: high classification accuracy, small overlaps between different classes in the reduced feature space and sparsity to reduce risk of overfitting. Mean squared error is used to account for the classification accuracy. While the overlap is measured relying on a mass function of $1 - \exp^{-\gamma_q d_{ij}^2}$, where $\gamma = [\gamma_1 ... \gamma_c]$ are set as the inverse of the mean distance between instances of the same class, c is the number of classes and d_{ij} is the weighted Euclidean distance between sample i and its neighbour j. Also, based on prior knowledge, some features are predetermined to be included as fixed elements in the selected subset. The study found that feature selection applied on rebalanced data outperforms feature selection alone when considering OA, AUC and robustness (measured as the relative weighted consistency of the selected subsets).

These hybrid approaches are more successful compared to approaches only handling imbalance or high dimensionality alone. In our study, the application of hybrid approaches is expected to have advantageous effects, as our RVA data manifest both class imbalance and relative high dimensionality.

Based on the requirements driven by data characteristics together with the pointed out inadequacies of the available methods and the favourable effect of hybrid approaches, we establish that a purpose-built machine learning framework is needed. The framework should be suitable for the characteristics of the RVA data and address the requirements stated in section 4.2 to produce reliable risk predictions.

In the next Chapter, a preliminary set of experiments applying standard machine learning approaches to our collected RVA data and their results are summarised to clarify the practical need for a purpose-built solution. The foundations for the envisaged OFSET_*mine* framework are illustrated. A description of the other benchmark datasets used to illustrate the generalisation of the proposed framework is given.

Chapter 5

Proposed Solution Framework for Cardiovascular Risk Prediction

Traditional markers used for identification of CVD risk, such as the FRS [49] and QRisk [87] markers exhibit limitations in terms of individual stratification, although they are good for measuring risk within a population [46, 34, 12, 89].

In this study, we address this problem by exploring RVA [147] as a CVD risk predictor. In contrast to FRS and QRisk measures, the acquisition of RVA measures is non-invasive, therefore RVA can be readily incorporated in primary care venues.

We develop novel computation methods for RVA data to successfully predict cardiovascular risk level on individual basis and fulfill the suitability criteria specified in section 4.2. The devised methods are integrated in a framework called Oversampling Feature SElecTion to *mine* data (OFSET_*mine*).

The collected RVA data are used to construct a vascular profile for each subject. Based on several risk factors and FRS scoring, the risk group (low, medium and high) is determined. Existing data mining techniques are applied on the available data and a brief evaluation of their effectiveness is presented.

To illustrate the potential of the proposed methods as candidate solutions to a range of problems, additional medical benchmark data sets from the UCI ML repository [130] are used. The details of the datasets are briefly outlined within this Chapter.

5.1 Standard ML Solution for CVD Risk Prediction

5.1.1 Data Description

In the current study, the collected data include 236 subjects records with 104 RVA-based measurements (features) for each subject (as described in Chapter 3). Other measures were collected namely: systolic blood pressure, total cholesterol and hdl-cholesterol. Also, several other parameters were recorded such as age, gender, ethnicity, whether smoker or not and family history of cardiovascular disease. The subjects are categorised into three risk groups low, medium and high of sizes 212, 14 and 10 respectively according to their FRS (as described in Chapter 3 subsection 3.2.3).

The recorded measures are the ones used for calculating FRS and QRisk score. These measures are collected to allow assessing the prediction quality of RVA based features in comparison to FRS measures and QRisk measures. FRS is chosen for comparison as it is a long established risk score calculator based on an extensive cohort multi-ethnic study, while QRisk presents a risk score derived from UK population similar to the subjects who volunteered in our study. The FRS comparison set contains the factors used in FRS score calculation : age, gender, total cholesterol (Chol), HDL cholesterol, systolic blood pressure (SBP) and SBP treatment (yes or no), DM (yes or no) and Current smoking (yes or no) status. The QRisk factors comparison set includes age, gender, FH of CVD, Chol/HDL,Smoker, DM, SBP, BMI and Ethnicity.

As discussed earlier, the study's data exhibit small sample size with imbalanced class sizes and relative high dimensionality, which justified the need for intervention using ML techniques. Oversampling and feature selection were found as appropriate methods to compensate for small sample size and reduce dimensionality as previously discussed in Chapter 4. Another characteristic is that for a given RVA-based feature, no clear demarcation can be drawn across the risk groups, which means the presence of risk groups overlap. Instance-based learning can handle such characteristic as it builds the classification decision on the similarity between individual subjects using all features, rather than using a global abstraction model.

Existing oversampling, feature selection and prediction methods are applied, here, to provide additional practical foundation for the need of developing novel methods.

5.1.2 ML Methods

With the aim of establishing the suitability of ML techniques and the need for new devised methods for cardiovascular risk prediction using RVA, a set of preliminary experiments are conducted. The experiments utilise some existing well recognised machine learning methods to address the challenges of high dimensionality and imbalance met with RVA data. The obtained results are shown in this section.

Quadratic regression, similar in concept to Cox Hazard regression, is first applied on FRS measures to elucidate its performance using the available data in a preliminary experiment. Quadratic regression is applied *only* on the FRS measures. This is done to emphasise the limitations imposed by the available data and investigate whether a similar regression technique to Cox Hazard would be capable to derive a representative model from the available dataset.

Afterwards, a set of three established classification algorithms, namely NB, MLP and RF, are used to generate the risk class prediction models using the assigned labels (labeling procedure described in Chapter 3). A smaller set of experiments were conducted using a range of classifiers by Fathalla et al [60], where RF, MLP and NB presented the best performance. Hence, they were selected for the current experiments. Also, RF and MLP are utilised due to their known effectiveness and robustness, while NB offers low computation complexity. The chosen classifiers are

used throughout all the experiments. A range of established classifiers with different underlying principles were used, where NB represents probabilistic learning, RF presents ensemble learning with decision trees and MLP is a frequently used model of ANNs. Classifiers of various principles are chosen to eliminate the effect of incompatibility of a single classifier on our findings. This is to avoid bias in the subsequent observations which would be induced by the selection of a wrong classifier. Hence, the achieved results would mainly reveal the characteristics of the data and the limitations they impose on the classification performance.

The detailed experiments are conducted applying classification algorithms rather than regression techniques for several reasons: 1) The objective of this study is to show whether RVA measures can separate participants into risk groups, which is a classification task; 2) The classification into risk groups more closely resembles the approach of a medical expert when determining risk category rather than generating a risk score; 3) The FRS score was originally calculated using Cox Hazard Regression and thus it would be biased towards the FRS measure to use a similar technique of regression (to the one that generated the score) for determining risk. The aim is to compare the performance of the measures using a different classification (risk stratifying) technique than the one used for FRS generation, creating unbiased and uniform conditions.

ADASYN is applied for the purpose of data balancing and sample size increase. In addition, ReliefF and CorrCoeff are used for feature selection. The default Weka [75] implementation of RF, MLP, NB, ReliefF and CorrCoeff is utilised in the experiments, while ADASYN [80] Matlab implementation and the default matlab settings for quadratic regression is used.

The performance measures used are : OA, AUC, Sn for high risk (Sn_H) and F_{score}. OA is commonly reported and hence allow cross comparison with other solutions, while AUC and F_{score} are more suitable for imbalance data evaluation. Sn_H is reported in this study as a misclassification in high risk group could be detrimental (i.e. not capturing the level of risk may lead to missing out on treatment and subsequently deteriorating health). Also, Sensitivity for Medium Risk Group is reported for RVA data only (Sn_M). Other measures that are frequently reported are area under precision-recall curve and $G-mean$ measure. Since area under precision-recall curve was used as a part of the proposed feature selection criteria in DD_Rank to be reported in Chapter 7, it was not reported to avoid bias and maintain the consistency of results across the reported methods. Also, $G-mean$ had negligible differences compared to F_{score} hence its values were not stated to avoid redundancy. Precision P was not reported to avoid redundancy as it had similar results to Sn. Also, precision in included in F_{score} calculation so it is indirectly reported.

5.1.3 Experiments and Results

FRS Measures and Regression for Risk Prediction

The default MATLABR2016b implementation of quadratic regression is used. Regression is applied, here *only* on the FRS measures, to explore whether the performance of FRS measures with quadratic regression can lead to a 100% accuracy or the characteristics of the data would hinder such attainment. Thus, enable focusing on the impact of the inherent challenges of the data set itself.

After generating a score using quadratic regression, subjects are mapped to the predicted classes using the same procedure described earlier where subjects with quadratic regression score < 10% are predicted as low risk, medium risk with score > = 10% and < 20% and high risk with score > = 20. Then, the resultant predicted class using quadratic regression is compared against the actual risk class.

The obtained confusion matrix is illustrated in Table 5.1. FRS measures and quadratic regression fail to attain a 100% accuracy and do not meet the requirements of high performance and transparency . This can be due to differences between Cox regression and quadratic regression models. These differences [164] conclude that Cox regression better accommodate right censored data, easily accommodate time varying data that may change over the course of time and does not require a specific probability model for estimating survival time. More importantly, due to the highly imbalanced dataset with the very small minority class it was impossible to reproduce a 100% correct predictive model from the available data set.

TABLE 5.1: Sample Confusion Matrix from applying Quadratic Regression on Framingham Risk Measures

		Classified		
		low	medium	high
Labeled	low	**206**	6	0
	medium	3	**9**	2
	high	1	6	**3**

RVA, FRS and QRisk Measures for Risk Prediction

After illustrating the impact of the data characteristics on the performance of FRS measures with regression, we need to examine the relative performance of RVA, FRS and QRisk measures using the selected range of classifiers. The classification methods are applied directly to the labeled data without preprocessing and the results are illustrated in Table 5.2. It shows the OA and Sn_H across the three sets. The results demonstrate the existence of imbalanced classes problem with high OA and low Sn_H. The low Sn_H (zero in six out of nine experiments) indicate that the classifiers treated the minority samples as noise and failed to construct a representative model. This is attributed to the high skewness in class distribution and extremely

small sample size, which justifies the adoption of oversampling as class imbalance solution.

TABLE 5.2: Classification Overall Accuracy and High Risk Group Sensitivity on Original(non-oversampled all feature) Dataset

	FRS		QRisk		RVA_data	
	OA	Sn_H	OA	Sn_H	OA	Sn_H
RF	92.01	0.0	92.43	0.0	90.25	0.0
MLP	95.79	0.0	92.01	0.0	82.20	0.0
NB	92.01	0.1	84.06	0.1	49.57	0.1

A sample confusion matrix of FRS with Naive Bayes is given in Table 5.3 to show the distribution of the predicted classes and better illustrate the results. It is to be noted that FRS measures with regression (Table 5.1) showed better results in terms of medium and high risk stratification compared to FRS with NB. Nevertheless, both results are unsatisfactory and there is no significant performance gap, which clarifies the challenges imposed by the data characteristics regardless of the applied learning model.

TABLE 5.3: Sample Confusion Matrix from applying Naive Bayes on Framingham Risk Measures

		Classified		
		low	medium	high
Labeled	low	**209**	2	1
	medium	3	**7**	4
	high	3	6	**1**

ADASYN oversampling [80] is applied to generate synthetic instances from the available sets for data balancing and increasing sample size. A preliminary setting of $d_{th} = 0.9$ and $\beta = 1$ is used to obtain a fully balanced data set. A uniform class distribution is sought to explore the potential of RVA measures in absolute terms, while eradicating the adverse effect of class skewness. The POST-ADASYN datasets containing real and synthesised samples include 212 low risk, 211 medium risk and 208 high risk samples. The generated sets are not of equal size due to rounding. After oversampling, two feature selection approaches are adopted: Correlation and ReliefF. Feature set size is chosen to be in the range 7 to 10 to have similar size to FRS and QRisk sets, to limit the differences in the comparison in this preliminary experiment. The subset that scores the highest accuracy is selected and its number of features #F is recorded.

The results of the various classifiers on these datasets are shown in Table 5.4. Four recognised classification performance measures are recorded: OA, AUC, Sn_H and F_{score}. QRisk Measures exhibit the highest OA, F_{score} with MLP, while RVA measures and RF manifest the highest Sn_H of 0.99 and the second best for OA.

The results demonstrate considerable performance improvement and that the classification performance of RVA-based features is comparable to the well-known

TABLE 5.4: Various Classifiers Performance using ADASYN oversampled FRS, QRisk and RVA measures

	FRS Measures				QRisk Measures				RVA Measures				
	OA	AUC	Sn_H	F_{score}	OA	AUC	Sn_H	F_{score}	OA	AUC	Sn_H	F_{score}	$\#F$
RF	91.82	0.98	0.89	0.92	93.7	0.99	0.94	0.94	**95.42**	0.99	0.99	0.95	10
MLP	94.49	0.97	0.95	0.95	**97.33**	0.99	0.98	0.97	92.58	0.96	0.99	0.93	10
NB	65.09	0.89	0.22	0.65	**77.83**	0.90	0.77	0.71	77.60	0.90	0.89	0.78	9

measures of FRS and QRisk scores. The results also show the potential superiority of RVA-based measures at detecting high risk (Higher Sn_H) when compared to FRS measures in particular.

Despite that the FRS measures were the ones used in score calculation using Cox hazard regression, they don't attain a 100% accuracy in classification and show lower performance compared to QRisk and RVA measures in some cases. The repeated finding with different classifiers suggest that the reason lies within the data. Possible reasons behind this can be:

a) The difference in representation of some FRS measures when used for risk score calculation (log form) and for classification (scalar form).

b) The difference in the underlying principles, techniques and handling of the measures for the model construction between classification and regression. Each technique attempts to relate the input variables to the target outcome using a different technique.

c) The mapping from a continuous risk score (FRS score) to discrete class labels (low, medium and high) leads to confusion at border samples especially because of the presence of four categorical variables within FRS measures.

On the other hand, all RVA measures are continuous values that directly convey the vessels status. Overall, the obtained results not only elucidate the positive impact of oversampling and feature selection on the overall performance, but also illustrate the potential for further improvement if machine learning techniques that best exploit the available RVA data are used.

5.2 OFSET_*mine*: The Proposed Framework

The presented characteristics of the collected RVA data in Chapter 3 and the specified methods requirements at section 4.2 in Chapter 4 together with the preliminary results of RVA-based measures in the previous section, all highlight the potential usefulness of a tailored solution. The existing standard solutions manifest some limitations according to our criteria. We do not only seek a higher predictive accuracy than the ones obtained using the existing methods, but we aim for fulfilling the set criteria specified in section 4.2. Therefore, we propose OFSET_*mine* framework, which integrates three novel approaches for oversampling, feature selection

and lazy learning. The target of the proposed framework is to predict cardiovascular risk group with the highest attainable accuracy and attain the requirements recommended for our study.

Figure 5.1 illustrates the framework stages and its incorporation within the cardiovascular risk prediction solution. The stages are described in brief, as follows, to highlight the main contributions of the presented study. First, we describe the proposed oversampling method, followed by the feature selection method. Then, a description of the developed classification technique is given. For each method, the motivation for applying the proposed method is shown, an outline of the method and its associated performance are portrayed.

5.2.1 FiltADASYN oversampling

The relative risk groups of the available data is highly imbalanced (highest imbalance ratio 21.2), which demands a solution for imbalance learning. Since the size of the dataset is small, the computation of synthetically increasing the number of samples will not be high. In addition, learning-based imbalanced class solutions such as cost sensitive learning are unsuitable to apply as the misclassification costs are hard to be assumed in such context. Hence, oversampling is selected as the solution for imbalance learning.

FiltADASYN oversampling is proposed to generate well-representative synthetic samples from all of the available minority samples, that would enhance the consistency and the separability of the original data set. The proposed method aims at preserving the scatter of the original data and ensuring the reliability and usability of the created samples. In FiltADASYN, synthetic samples are generated using ADASYN method which interpolates feature vectors to maintain feature dependence. Then, a post processing step depending on nearest neighbours is applied to perform post synthesis validation and filter the generated samples that are susceptible to be noise and/or outliers.

The performance of FiltADASYN is evaluated individually and as a component of hybrid solutions. Its applicability compared to ADASYN is shown to be dependent on the dataset under study with constant improvement over baseline (non-oversampled) results. The details of FiltADASYN method and the associated evaluation are described in Chapter 6.

5.2.2 DD_Rank and Linear Search Feature Selection

The derived features from the collected RVA data are expected to include redundant or irrelevant features which would hinder the risk group labeling process. Also, reducing the feature set size promotes computational efficiency and provides a better understandable model for medical professionals. Therefore, feature selection is applied to handle the characteristic of high dimensionality.

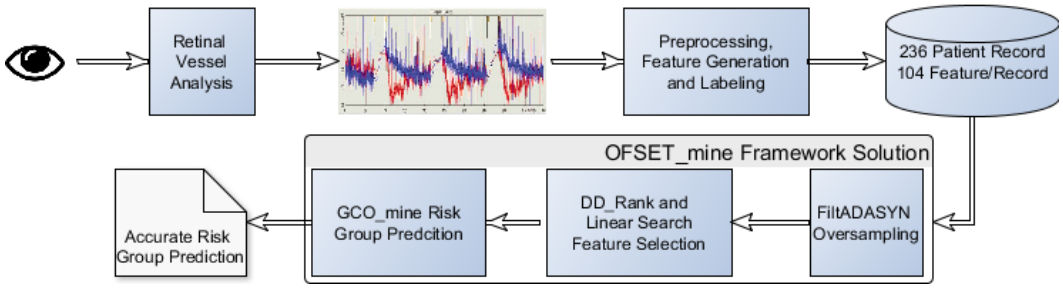


FIGURE 5.1: Proposed OFSET_*mine* Framework Solution for Cardiovascular Risk Prediction

An explicit feature ranking approach is proposed based on an existing deep disjunct architecture of Restricted Boltzmann Machines. RBMs are known to detect latent relations between the input features and create higher order representations. Thus, the proposed approach would be capable of discovering the features interrelations. The ranking approach seeks to rank stable and predictive features high. The values of Area under precision-recall curve and reconstruction errors are used per feature to combine measures on predictive performance and theoretic heuristics for ranking features. A ranked feature list is output that determines the relative importance of the features according to DD_Rank. After ranking the features, a linear search is followed to select the feature subset that marks the highest accuracy.

The impact of the proposed feature selection approach is assessed independently and collaboratively with oversampling. The performance of the proposed feature ranking technique is highly competitive with two well established ranking algorithms, namely ReliefF and Corrcoeff. The results reveal that DD_Rank alone can help combat the problems of high dimensionality and imbalance in some applications. Also, when combined with oversampling a statistically significant improvement is attained. Chapter 7 provides a full explanation of the proposed feature selection approach along with the related experiments.

5.2.3 *GCO_mine* Prediction

The pathological and normal ranges of physiological features values are known to overlap, hence a single decision value that can separate groups does not exist. As a result, eager classification methods that tend to create global approximation models based the features values may not be best suited to our problem. This motivates the development of learning algorithms that perform classification decisions on instance (individual) basis as it was shown by Xiong et al. [209] to be well suited for the overlap problem. Also, although the available dataset is currently small, data is expected to remain to be collected for further verification of the success of RVA data in cardiovascular risk prediction. Therefore, the developed learning method should accommodate future expansion of the dataset. Such factors direct our adopted learning approach to instance-based lazy learning.

A partially lazy learning technique is developed based on the established technique of Graph Cut Optimisation (GCO) [30, 118]. The classification decision is based on the global structure of the resultant output classes, while considering the local neighbourhood of each sample in classification cost calculation. GCO_*mine* offers a balance between purely local traditional lazy approaches and over generalised eager approaches, aiming for high performance predictions. The classification decision taken by GCO_*mine* is transparent to the user as it can be explained by its vicinity to specific samples and class representatives.

The developed learning method is applied directly on the selected datasets and within the OFSET_*mine* framework. The developed technique saves model construction time and better accommodate new instances, when they become available, compared to established eager approaches. It also improves the performance relative to purely local lazy approaches. The algorithm is depicted in depth in Chapter 8 with its performance assessment.

In Chapter 9, DD_Rank and FiltADASYN are brought together as Oversampling Feature SElecTion (OFSET) solution, then combined with *GCO_mine* as OFSET_*mine* integrated framework. Both solutions OFSET and OFSET_*mine* are discussed . The proposed techniques are validated, as stand alone methods and as part of the OFSET_ *mine* framework. The results show the methods to be competitive to established approaches.

5.3 The Applicability of OFSET_*mine* to Medical Problems

Despite that OFSET_*mine* is designed for cardiovascular risk prediction using RVA, similar characteristics to that of RVA-based measures can be found in other medical problems. Thus, the appropriateness of OFSET_*mine* on a chosen set of additional benchmark medical datasets is to be shown.

The general applicability of the methods is demonstrated using 13 medical benchmark datasets from the UCI ML Repository. The datasets depict various medical conditions, which are critical to accurately predict. The provision of accurate predictions for these problems would support effective health care.

The datasets are selected here on the basis of similar characteristics to our RVA-based dataset in terms of size and/or imbalance ratio and/or number of target classes. The sizes of the datasets range from 106 to 768 records, while the imbalance ratio (ratio between the size of the largest and smallest class) from 1.68 to 71.5. Imbalanced datasets are selected to determine the effectiveness of the proposed algorithms in case of skewed as well as balanced datasets.

The details of the utilised datasets are described in Table 5.5 in terms of number of features (#F), number of classes (#C), Imbalance ratio (I_r), number of samples in each class (# Samples/Class) and a brief description of the studies objectives. The related methods are applied when similar characteristics to our RVA data exist.

TABLE 5.5: Medical Dataset Characteristics: number of features (#F), number of classes (#C), Imbalance ratio (I_r), number of samples per class and Classification Objective

Dataset	# F	#C	I_r	#Samples / Class	Objective Brief Description
RVA_data	104	3	21.2	(212, 14, 10)	Predict cardiovascular risk (low, medium and high risk) using RVA-based measures
Pima Diabetes	8	2	1.87	(268, 500)	Predict based on diagnostic measurements whether a patient has diabetes or not
Ecoli	8	8	71.5	(2, 2, 6, 20, 35, 52, 77, 143)	Determine the localisation site of the E.coli
Colic(SurgLesion)	22	2	1.70	(136, 232)	Classify whether the problem (lesion) was surgical or not
Colic (outcome)	22	3	4.33	(52, 89, 225)	Determine what eventually happened to the horse : lived,died or was euthanised
Lymph	18	4	41	(2, 4, 61, 82)	Determine histological types of lymph cells: normal find,metastases,malign lymph and fibrosis
Dermatology	34	6	5.65	(20, 49, 52, 61, 72, 113)	Differential diagnosis of erythemato-squamous diseases
Parkinson	22	2	3.06	(48, 147)	Detect health status of a subject either Parkinson's or healthy
HeartCleveland 2C	13	2	2.98	(55, 164)	Diagnosis of heart (angiographic) disease status: less or more than 50% diameter narrowing
HeartCleveland AllC	13	5	12.6	(13, 35, 37, 55, 164)	Determine the degree of presence of heart disease: Value ranging from 0 (no presence) to 4
WisconsinBreast Diagnostic	30	2	1.68	(212, 357)	Classify whether breast tumor is benign or malignant
Breast Tissue	9	6	1.57	(14, 15, 16, 18, 21, 22)	Classify the breast tissue as carcinoma,fibro-adenoma, mastopathy, glandular, connective or adipose
VertebColumn 2C	6	2	2.1	(100, 210)	Classify orthopaedic patients into normal or abnormal
VertebColumn 3C	6	3	2.5	(60, 100, 150)	Classify orthopaedic patients into 3 classes: normal, disk hernia or spondilolysthesis .

Chapter 6

Filtered ADASYN Oversampling Method

A common drawback in existing oversampling techniques (such as: Random Oversampling, SMOTE and ADASYN) is overgeneralisation, where the synthesised minority class instances overlap erroneously with majority samples. When applying oversampling to medical data, handling this legitimate concern of over generalization becomes more critical as the synthesised samples turn into false representatives of the minority class (risk group).

In this chapter, we propose Filtered ADASYN (FiltADASYN) a technique that allows for the conservative expansion of the minority class and restricts the degree of over generalization. At the same time, it focuses on difficult to learn samples at the borders. A filtering phase is added to ADASYN oversampling to improve the dependability of the data and reduce inconsistency within the resultant classes. The filtering stage ensures that the nearest neighbours of the produced samples are of the same class and eliminates any samples that violate this condition.

6.1 The Proposed Approach

FiltADASYN, similar to ADASYN relies on density distribution $\hat{r}_i$ to determine the number of synthetic samples that need to be generated for each minority data example. The detailed procedure is outlined in Algorithm 2. Instances $x_i \in X$ with class labels of a single minority class C_{mn} and the majority class C_{mj} where both $\in C$ are input to the algorithm and n_{mn} and n_{mj} are the number of minority class and majority class instances, respectively. The maximum allowed imbalance threshold d_{th} and the desired balance level after generating the synthetic data β are preset and input to the FiltADASYN procedure. The density distribution $\hat{r}_i$ (line 10) depends on $\triangle_i$, the number of majority class samples in the set of k-nearest neighbours of x_i. Thus, the minority class samples that are surrounded by more majority class samples, and are hence more difficult to learn, are used to generate a proportionally greater number of synthetic samples g_i (line 11). This would enable the classification model to correctly identify higher risk samples despite their apparent similarity to low risk

Algorithm 2 Filtered ADASYN Oversampling Algorithm

1: **procedure** FILTADASYN($X, C, n_{mn}, n_{mj}, d_{th}, \beta, k, l$)
2: $d \leftarrow \frac{n_{mn}}{n_{mj}}$
3: **if** $d >= d_{th}$ **then**
4: **return**
5: **end if**
6: $G_{Acc} \leftarrow C_{mn}$
7: $G \leftarrow \beta(n_{mj} - n_{mn})$
8: **for all** $x_i \in C_{mn}$ **do**
9: Find k-nearest neighbours for x_i
10: $\hat{r}_i \leftarrow \frac{\triangle_i}{k}$
11: $g_i \leftarrow \hat{r}_i G$
12: $j \leftarrow 0$
13: **repeat**
14: Randomly choose one minority sample x_{zi}
15: from k-nearest neighbours of x_i
16: Generate Sample $s_j \leftarrow x_i + \lambda(x_{zi} - x_i)$
17: Check l-nearest neighbours nn of s_j *in* $G_{Acc} \cup C_{mj}$
18: **if** $\exists\ nn\ \in C_{mj}$ **then**
19: Reject s_j
20: **else**
21: $G_{Acc} \leftarrow G_{Acc} \cup s_j$
22: $j \leftarrow j + 1$
23: **end if**
24: **until** $j > g_i$
25: **end for**
26: **end procedure**

instances. On the other hand, these samples are likely to lead to generation of samples that have majority samples in their neighbourhood which would increase the difficulty. Thus, the use of small l avoids over rejection and at the same time filters the generated samples and pushes them within the minority class boundary. This is achieved through allowing the generation of synthetic boundary samples, while confining them within the generating minority cloud. The synthetic sample s_j is at a random distance λ between x_i and another randomly chosen minority sample x_{zi} (as shown in lines 14-16). The l nearest neighbours nn of the generated sample are determined within G_{Acc} and C_{mj}. G_{Acc} is defined as the set of accepted minority samples (original and synthesised). It is initialised as the minority class samples C_{mn}. G_{Acc} is chosen instead of C_{mn} to allow for natural expansion and avoid over clustering of the resultant oversampled minority class. If a majority sample is found within the l nearest neighbours of s_j, s_j is rejected and the sample generation procedure is repeated with another random minority sample x_{zi} and random factor λ. Otherwise, s_j is accepted and added to G_{Acc} (line 21). The value of l for the nearest neighbours is preferably kept small to avoid over rejection. The low value for

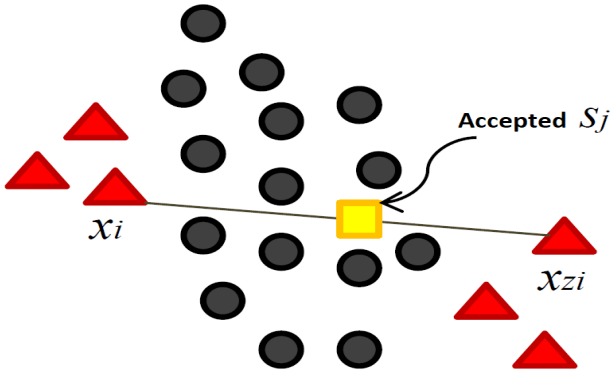
l and the employment of G_{Acc} instead of C_{mn} allow for boundary extension as we keep the rejection rate low and consider new synthetic samples as safe to expand on. Another sample x_i is drawn for the oversampling procedure until the number of accepted generated samples j exceeds the designated number g_i (line 24), where g_i is calculated based on the targeted balance and the density distribution. The algorithm terminates when all the minority samples are used to generate other synthetic instances. Failure to generate a valid synthetic sample from existing minority samples is unlikely as there are two random parameters, which are λ and x_{zi} that can be changed to fulfill the condition. The added filtering step (outlined at lines 18 to 23) aims at preserving the original structure of the generated minority class through ensuring that the generated samples are within the boundaries of the minority class and limiting the overlap between synthesised samples and the majority class. This is achieved as it only accepts samples that do not have majority class samples within their neighbourhood.

Figure 6.1 illustrates an example of a generated sample s_j that would be (A) accepted by ADASYN as a legitimate synthetic sample and (B) rejected by FiltADASYN in the post validation filtering step. In Figure 6.1, we assume the presence of a majority class (indicated by black dots) and two clusters of a minority class (given as red triangles) at the borders of the majority class. Sample x_{zi} would be considered as a neighbour of x_i if, for example, number of neighbours k is set to four. As a result, s_j will be generated within the majority class. FiltADASYN treats such shortcoming in ASDASYN as it rejects sample s_j and selects another neighbour for oversampling. In addition, in case of the presence of minority class disjunct subclusters surrounding the majority class, FiltADASYN implicitly directs the synthesised samples to be within the clusters of the minority class.

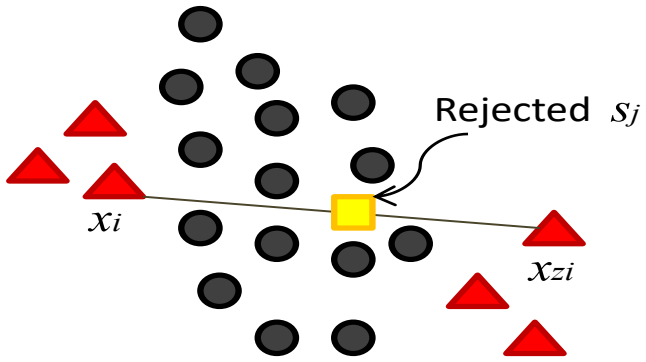
The proposed FiltADASYN procedure is originally designed for two class imbalanced problem but can be readily extended for our three class problem. Given that we have one majority class (Low Risk) and two minority classes (Medium and High Risk), we apply the oversampling procedure twice following a simple policy of 'one versus all', where FiltADASYN is applied once for each pair of (minority, majority) samples, i.e. (High,{Low $\cup$ Medium}) and (Medium,{Low $\cup$ High}).

6.2 Results and Evaluation

In this section, the usefulness of the proposed FiltADASYN is validated through a set of experiments. The general applicability of the proposed method is verified using standard benchmark datasets, and its performance is shown to be competitive with ADASYN oversampling. The experiments were done using Weka [75] and MATLAB R2016b.



(A) ADASYN



(B) FiltADASYN

FIGURE 6.1: Illustration of an accepted synthesised sample by (A) ADASYN vs a rejected synthesised sample by (B) FiltADASYN in two dimensional space

6.2.1 Experimental Study

In the following we briefly present the experiments objectives, the data, the tools, the implemented experimental design and the evaluation metrics used.

Experiments Objectives: The objective of experiments described in this section is to elucidate the impact of FiltADASYN oversampling on RVA-based measures and compare it to ADASYN. Also, FiltADASYN competitive performance is to be illustrated when applied to benchmark datasets.

Experimental data: The performance of the proposed method needs to be validated on the specific data it was designed to handle and on available benchmark datasets; to show its range of applicability. Therefore, FiltADASYN is applied on RVA-based measures dataset to jointly evaluate their capability in cardiovascular risk prediction. Also, the method is applied on a set of benchmark medical datasets from UCI ML Repository [130] previously outlined in Table 5.5 section 5.3.

Methods implementation: FiltADASYN relies on the MATLAB implementation of AD-ASYN [79].

Experimental tools: In order to verify the performance of the proposed framework,

three well established classifiers are used: Random Forest (RF), Multilayer Perceptron (MLP) and Naive Bayes (NB). The default Weka implementation of RF, MLP and NB is utilised in the experiments. The chosen classifiers are used throughout all the experiments. Other classification algorithms can be readily applied as there are no restrictions imposed by our proposed approaches. The settings of the classifiers are described in Table 6.1. Same settings are used throughout all the conducted experiments.

TABLE 6.1: Classifiers Settings in Weka

Classifier	Default Settings
Given any classifier, uniform misclassification costs are assigned for all risk groups.*	
RF	Random Number Seed = 1 Number of trees = 100 Number of threads = 1 Calculate Out of Bag Error Unlimited Depth (No Pruning) Number of Randomly Chosen Attributes = $\log_2(Numberof Features) + 1$
MLP	Momentum = 0.2 Learning Rate = 0.3 Normalise Attributes Number of training Epochs = 500 Number of Hidden Layers = $\frac{Numberof Features + Numberof classes}{2}$
NB	Apply Normal Distribution Assumption No Discretisation for Numeric Attributes

*This is to avoid any possible harmful effects of having false positives from Low risk [190, 158].These harmful effects include but are not limited to: Health Implication from over treating healthy subjects and Failure in prioritising subjects for re-assessment/follow up

Experimental procedure: For RVA-based measures, several balance levels β are investigated. First, the effect of maintaining the base rate with oversampling is shown. Then, the performance on partially balanced datasets with variable balance levels reaching fully balanced oversampling dataset is presented.

For the described benchmark medical datasets(Table 5.5), the conducted experiments investigate the independent effect of oversampling. To clarify the effect of the proposed techniques on overall performance, experiments are conducted on the original datasets (non-oversampled full features set) to provide reference baseline performance. FiltADASYN is compared to ADASYN oversampling on the benchmark medical datasets with their parameters set as $d_{th} = 0.9$, $\beta = 1$ (for fully balanced data) and k set to 5. The value of l for FiltADASYN is chosen to be 3.

The values of k and l are chosen based on experimental trials on the RVA data. First, the best value of k for ADASYN is chosen and set, then l is allowed to vary for the benefit of FiltADASYN performance. The same value of k was used for both ADASYN and FiltADASYN for comparability and to limit the interacting effect of several parameters and highlight the effect of the added parameter l.

The reported results are the average of five 10-fold cross validation runs on the available datasets. The statistical significance of the achieved results is evaluated using Friedman test [90]. It is a non-parametric multiple comparison test that can

be used with continuous data to determine if any statistical difference exist between the results.

Evaluation metrics: The utilised performance indicators are OA, AUC, F_{score} and Sn_H when it is clearly marked as High Risk. These metrics are used for the evaluation of all the methods performance.

6.2.2 Cardiovascular Risk Prediction Results based on RVA Measurements

Initially, the results of classification methods on the original non-oversampled full feature set of RVA-based measures are reported to serve as reference baseline performance and to help identify the requirements for designing a specialised solution for this problem.

Table 6.2 shows these results, which are typical for imbalanced datasets: high OA (with RF and MLP) is observed with low Sn of minority classes. Thus, a method specifically designed to account for class imbalance is needed. The effect of the proposed method (FiltADASYN) is examined separately to determine its individual influence on the performance and its corresponding significance.

TABLE 6.2: Evaluation of Models Performance on non-oversampled full feature set (Baseline) RVA data

	All Features Non-Oversampled set				
	OA	AUC	Sn_H	Sn_M	F_{score}
RF	90.25	0.40	0.0	0.0	0.86
MLP	82.20	0.46	0.0	0.0	0.82
NB	49.57	0.52	0.1	0.57	0.62

Oversampling: ADASYN and FiltADASYN

Oversampling is commonly applied to compensate for classes skewness and perform data balancing. Nevertheless, it is to be argued that enlarging the sample size while maintaining the original classes distribution of the condition under study is preferable. Therefore, we investigate both effects of (a) increasing the sample size while keeping the base rate (original class priors) and (b) performing class balancing (partial and full) with oversampling.

(a) Base Rate Oversampling

We start with applying ADASYN and FiltADASYN in three fold manner to maintain the original imbalance ratio (base rate) of the dataset. First, the minority classes (medium and high risk) are oversampled with $\beta = 0.25$ resulting in a total of (66, 60) instances respectively. Then, the majority class (low risk) is randomly partitioned to smaller subsets and oversampled against the merged syntheic minority classes keeping the original base rate with a total of 1321 low risk samples.

The corresponding results of ADASYN and FiltADASYN are shown in Table 6.3. The results show considerable improvement in Sn_M and Sn_H but the overall performance is below the targeted performance level, especially the true positive rate Sn with RF and MLP , as well as the overall accuracy OA with NB. Another note is that FiltADASYN presents slightly better results with RF and MLP so the further investigations are performed applying FiltADASYN.

(b) Class Balancing Oversampling

(b.1) Partial Balance Oversampling

Since the available data is highly skewed with imbalance ratio (low to high risk groups) of 21.2, this per se would disrupt the performance of the classifiers [137] even if the sample size of the minority class is increased. Hence, we apply FiltADASYN to partially balance the dataset creating **only** medium and high synthetic samples.

The same β value of 0.25 as the previous experiment of base rate oversampling is chosen to have the same number of samples and to vary a single factor for comparison, which is reducing the imbalance ratio. The results are shown in Table 6.4. Although the achieved OA is lower than the base rate scenario with RF and MLP, the Sn_M and Sn_H are higher indicating that partial balancing lead to better stratification of critical risk groups, which encourages further balancing using FiltADASYN.

FiltADASYN is applied on RVA data with varying β values, to elucidate the effect of increasing sample size together with partial re-balancing of the data. The resulting (synthetic + original) risk groups sizes for {Medium,High} risk groups using β values of (0.25, 0.5 and 0.75) is {66, 60}, {118, 114} and {160, 160} respectively.

The achieved sensitivity values for the resulting Medium and High risk sets are shown in Figure 6.2. The results show that increasing the targeted balance level, leads to higher sensitivity of both Medium and High risk groups approaching one. Clearly, it is desirable to attain a sensitivity of one especially when the set contains real and synthetic samples to ensure the real samples are correctly classified. Also, the original distribution of the classes of the collected dataset does not necessarily represent the actual natural distribution of risk groups within a population.

According to a WHO report [155], the natural imbalance ratio (low to high risk) vary dramatically depending on the population, gender and ages group. etc. Hence, when new data is collected or when applied on other samples populations, the underlying distribution might be different. This directed the following experiments towards a fully balanced dataset to completely eliminate the adverse effects of imbalance applying ADASYN and FiltADASYN to allow further comparison.

(b.2) Full Balance Oversampling

FiltADASYN and ADASYN are then applied on the original dataset to generate synthetic instances so that the risk categories contain similar numbers of samples. The resultant oversampled datasets then contain both real and synthesised samples,

TABLE 6.3: Oversampling Performance on RVA data (All Features set) keeping the classes base rate

	Post ADASYN Measures					Post FiltADASYN Measures				
	OA	AUC	Sn_H	Sn_M	F_{score}	OA	AUC	Sn_H	Sn_M	F_{score}
RF	96.23	0.98	0.68	0.42	0.96	**96.40**	0.99	0.76	0.53	0.96
MLP	98.62	0.98	0.81	0.79	0.97	**98.76**	0.98	0.81	0.80	0.98
NB	**58.29**	0.87	0.83	0.80	0.68	53.49	0.82	0.83	0.78	0.64

TABLE 6.4: Evaluation of Models Performance on Partially Balanced with FiltADASYN oversampling ($\beta = 0.25$) full feature RVA dataset

	All Features Base Rate Set				
	OA	AUC	Sn_H	Sn_M	F_{score}
RF	92.33	0.97	0.82	0.74	0.92
MLP	88.79	0.96	0.84	0.84	0.89
NB	67.55	0.89	0.85	0.82	0.70

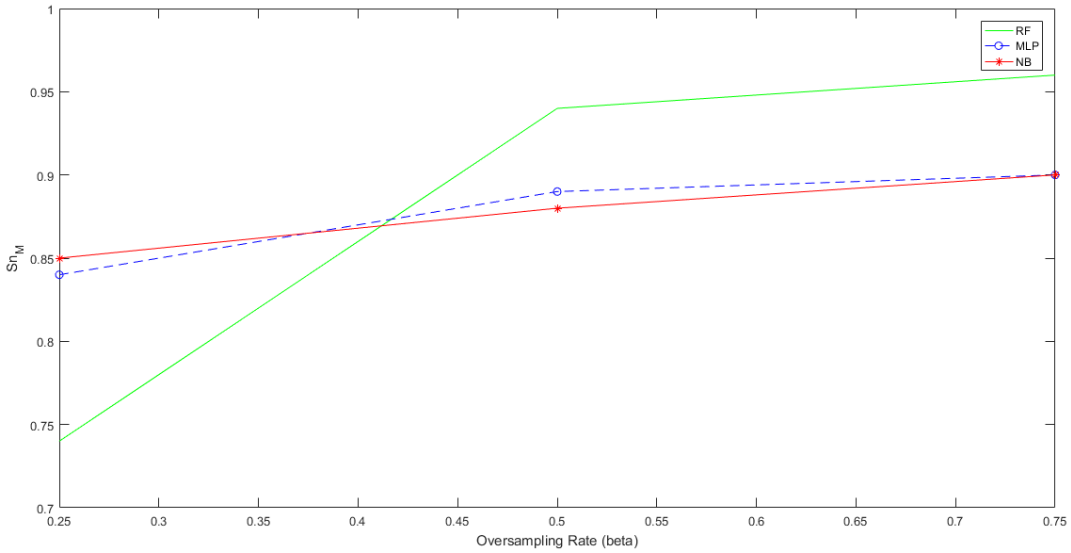
as follows: 212 low risk, 211 medium risk and 208 high risk samples. The generated sets are not of equal size due to rounding.

RF, MLP and NB are applied on the oversampled fully balanced datasets and the classification results are reported in Table 6.5. The results also show higher accuracies achieved by PostFiltADASYN data with RF and MLP. Therefore, any following results are reported using PostFiltADASYN data. The results depict the immense performance improvement attained through oversampling. The remarkable increase in Sn_H and Sn_M can be attributed to the oversampling methods capability of generating representative samples, successfully presenting the corresponding classes. However, these results are possibly optimistic due to the imposed similarity between the samples in the training and testing phases. Although, Sn_H reaches one meaning that all the original high risk samples are correctly classified, with Sn_M a chance still exists that the original medium risk samples are misclassified as their true positive rate doesn't reach one. Such finding motivate further improvement attempts to provide correct classifications for all real samples. Therefore, a further solution is needed to promote the performance and dependability of the results; feature selection will be applied for this purpose and for the other targeted benefits for feature selection.

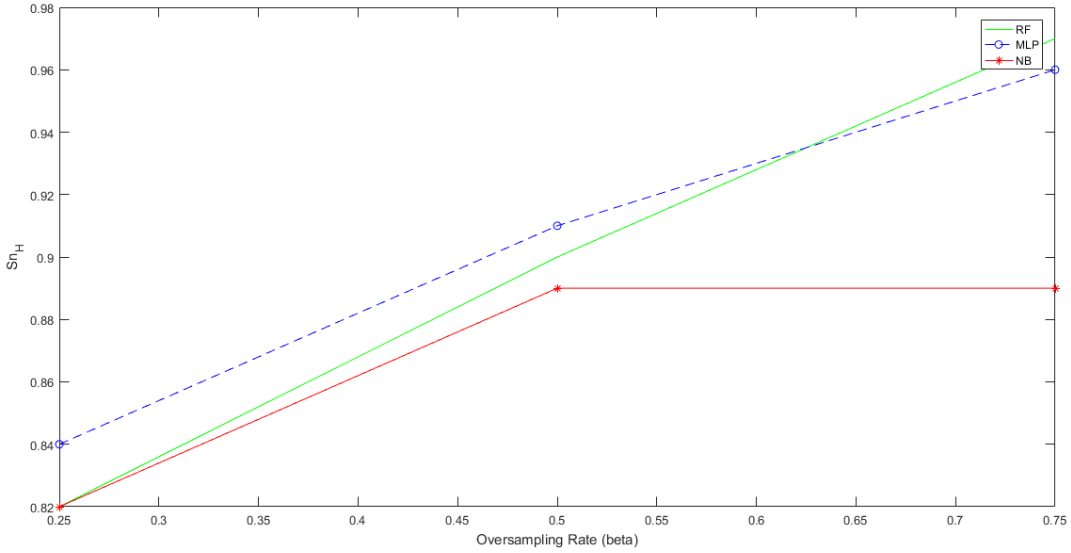
In order to quantify the quality of the resultant classes after oversampling, three

TABLE 6.5: Oversampling Performance on RVA data (All Features set) creating fully balanced datasets

	Post ADASYN Measures					Post FiltADASYN Measures				
	OA	AUC	Sn_H	Sn_M	F_{score}	OA	AUC	Sn_H	Sn_M	F_{score}
RF	98.42	0.99	1	0.98	0.98	**99.52**	0.996	1	0.99	0.995
MLP	93.84	0.97	1	0.90	0.94	**94.19**	0.97	1	0.92	0.95
NB	**85.80**	0.96	1	0.93	0.87	79.59	0.94	0.92	0.92	0.80



(A) Medium Risk Group Sensitivity



(B) High Risk Group Sensitivity

FIGURE 6.2: Sensitivity for (A) Medium and (B) High Risk Groups in relation to Different Balancing Levels (β).

partition quality measures are used. Silhouette index (S), Davies Bouldin (DB) index and Calinski-Harabasz (CH) criterion [135] are utilised to assess the compactness and separability of the produced synthetic classes by ADASYN and FiltADASYN. The obtained mean values are summarised in Table 6.6. Based on the shown results, we claim that both ADASYN and FiltADASYN generate similar class partitions although FiltADASYN outperform ADASYN given DB and CH with about 2% and 3% respectively. Moreover, the oversampling enhances the labels consistency and builds better structured classes. This can be observed from the improvement in S, DB and CH indexes values when compared to the original dataset (shown in subsection) 3.2 : Table 3.3).

Figure 6.3 shows the samples spatial distribution of two features namely MD_AF1

TABLE 6.6: Classes Partitions Quality Evaluation after Oversampling

	ADASYN	FiltADASYN
Silhouette	0.15	0.14
DaviesBouldin	5.30	5.20
Calinski − Harabasz	21.81	22.54

and MC_VF1 for original, POST ADASYN and POST FiltADASYN datasets. Despite the apparent similarity, it illustrates some differences in samples scatter pattern between the resultant synthetic sets (FiltADASYN Figure 6.3 (c) vs ADASYN Figure 6.3 (b)) when compared against the original set. Given the Original, POST ADASYN and POST FiltADASYN chosen features scatter plots, it is evident that FiltADASYN better handles instances that may be outliers as shown at Figure 6.3. At Figure 6.3(c) no synthetic samples are created at the vicinity of the marked (circled) Medium Risk instance producing a more compact synthetic medium risk group which better mimics the original dataset. While ADASYN Figure 6.3 (b) portrays several medium risk synthetic instances created near the marked instance. Also, it is observable that some of the generated medium risk samples in Figure 6.3 (b) and some of the synthetic high risk samples in Figure 6.3(c) protrude away from the samples bulk in a similar pattern. By visual inspection, we can say that the high risk samples in Figure6.3(c) conform better with the original distribution, as no samples were protruding from the medium risk bulk in Figure 6.3 (a).

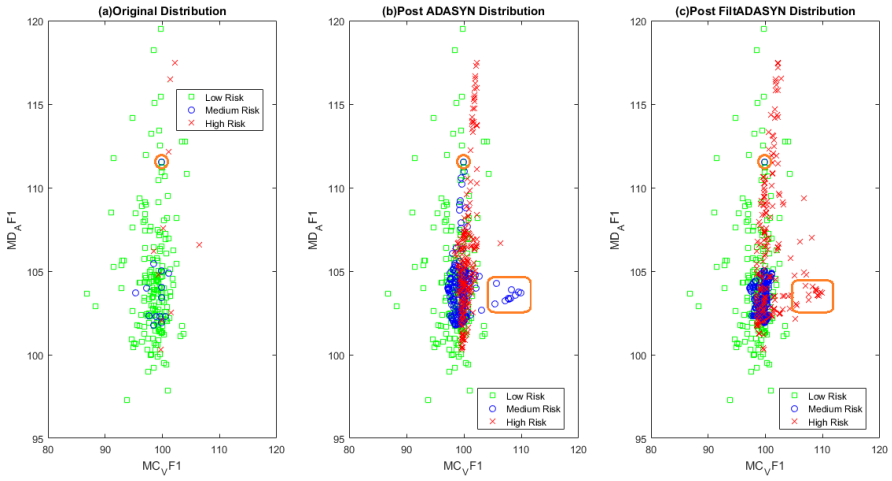


FIGURE 6.3: The distribution of sample features MC_VF1 (x-axis) which is the minimum constriction for Venular respsonse Flicker 1 vs MD_AF1 (y-axis) which is the maximum dilation for Arterial response Flicker 1 in (a) Original Dataset (b) Post ADASYN Oversampling dataset (c) Post FiltADASYN Oversampling dataset

6.2.3 FiltADASYN General Applicability on Benchmark Medical Datasets

The performance of the proposed oversampling method is demonstrated on 7 out of 13 medical benchmark datasets from the UCI ML Repository. The sets are chosen such that they have a maximum of four classes where two or three classes would be oversampled, to allow for the direct application of the oversampling algorithms and avoid significant overfitting. Highly skewed datasets with an imbalance ratio of over 1.8 are chosen to ensure a need for intervention.

Baseline results of the classifiers RF, MLP, and NB on all the 13 benchmark datasets are presented in Table 6.7. Baseline results are the results for the imbalanced dataset with full list of features without preprocessing. All baseline results are reported for future reference, when processing solutions other than oversampling are applied.

Oversampling: ADASYN and FiltADASYN

Oversampling methods ADASYN and FiltADASYN are applied on the selected datasets using fixed settings of $\beta = 1$ creating fully balanced datasets. This is done to draw a rough picture for their general performance.

The results achieved on the seven datasets are outlined in Table 6.8. We note that the two methods manifest comparable performance on these benchmark datasets with minute differences. Both methods provide an OA improvement with most datasets. In some cases, an improvement of more than 10% over baseline (Original non-oversampled data) accuracy is achieved as in the case of Colic (outcome) dataset with RF. On the other hand, both methods worsens NB baseline accuracy with three sets: Pima Diabetes, HeartCleveland 2C and VertebColumn 3C. This may indicate the incompatibility of these two oversampling techniques with NB.

Based on the results in Tables 6.5 and 6.8, it is noticeable that the success rate of FiltADASYN is higher with RVA measures (2/3) against the other medical datasets (7/21). This is because the parameter setting was first tuned based on the RVA measures and then directly applied on the remaining data sets.

6.2.4 Statistical Significance Analysis of Performance

In order to establish whether the performance of proposed method FiltADASYN, is significantly different from that of existing ADASYN method, we apply the Friedman test for continuous data [90]. The significance of the difference is checked only for the fully balanced sets, as this is the only case applied on both RVA and benchmark medical datasets, hence would enable significance analysis. The results are presented in Table 6.9. For each of the measures OA, AUC, Sn_H and F_{score}, the actual values of the measures are input to the Friedman test. The significance values ($p - values$), together with the rankings of the baseline (Bl) and the compared methods are output by the test. A higher attained ranking score manifest better performance. In each case, first, an overview of the overall performance is presented, by collating the results across all three classifiers. Then, the performance of each

TABLE 6.7: Baseline Performance on Benchmark Datasets. Datasets 1-7 are candidates for oversampling, with no more than four classes and imbalance ratio larger than 1.8. Datasets 8-13 will not be oversampled, as they either have more than four classes or smaller imbalance ratio.

ID#	Dataset		Performance Measures			
			OA	AUC	Sn_H	F_{score}
Candidates for Oversampling						
1	Pima	RF	75.26	0.81	0.60	0.75
	Diabetes	MLP	75.13	0.79	0.61	0.75
		NB	76.30	0.81	0.61	0.76
2	Colic	RF	69.94	0.83		0.67
	(outcome)	MLP	69.39	0.77		0.69
		NB	68.30	0.83		0.69
3	Lymph	RF	83.11	0.93		0.83
		MLP	89.96	0.93		0.90
		NB	85.13	0.89		0.85
4	Parkinson	RF	91.28	0.96	0.75	0.91
		MLP	91.28	0.96	0.83	0.91
		NB	69.23	0.86	0.61	0.75
5	Heart	RF	77.62	0.78	0.38	0.77
	Cleveland	MLP	76.71	0.80	0.56	0.77
	2C	NB	78.89	0.80	0.51	0.79
6	Verteb	RF	84.19	0.93	0.76	0.84
	Column	MLP	84.51	0.93	0.70	0.85
	2C	NB	77.74	0.88	0.87	0.80
7	Verteb	RF	83.54	0.96		0.84
	Column	MLP	85.48	0.96		0.86
	3C	NB	83.23	0.95		0.83
No Oversampling						
8	Ecoli	RF	86.09	0.96		0.86
		MLP	85.71	0.95		0.86
		NB	85.41	0.96		0.86
9	Colic	RF	85.32	0.89	0.77	0.85
	(Surg_Lesion)	MLP	81.25	0.87	0.74	0.82
		NB	77.1	0.82	0.74	0.78
10	Dermatology	RF	96.44	1.00		0.97
		MLP	97.54	1.00		0.98
		NB	97.54	1.00		0.98
11	Heart	RF	58.41	0.81		0.55
	Cleveland	MLP	54.78	0.75		0.54
	All Classes	NB	56.43	0.81		0.56
12	Wisconsin	RF	95.78	0.99	0.98	0.96
	Breast	MLP	96.66	0.99	0.95	0.97
	Diagnostic	NB	92.97	0.98	0.94	0.93
13	Breast	RF	71.69	0.93		0.72
	Tissue	MLP	64.15	0.88		0.65
		NB	70.75	0.93		0.71

TABLE 6.8: Oversampling Performance using Random Forest (RF), Multilayer Layer Perceptron (MLP) and Naive Bayes (NB) on Selected Benchmark Data Sets

ID#	Dataset		Post ADASYN Data				Post FiltADASYN Data			
			OA	AUC	Sn_H	F_{score}	OA	AUC	Sn_H	F_{score}
1	Pima Diabetes	RF	**80.89**	0.89	0.84	0.81	80.69	0.89	0.85	0.81
		MLP	**75.89**	0.81	0.79	0.76	75.18	0.81	0.78	0.75
		NB	*70.58*	0.79	0.63	0.71	*70.58*	0.79	0.65	0.71
2	Colic (Outcome)	RF	81.34	0.95		0.82	**82.78**	0.95		0.84
		MLP	**78.62**	0.89		0.79	75.17	0.88		0.75
		NB	71.73	0.86		0.72	**72.59**	0.87		0.73
3	Lymph	RF	*92.72*	0.98		0.93	*92.72*	0.98		0.93
		MLP	**92.71**	0.97		0.93	91.26	0.97		0.91
		NB	87.38	0.94		0.88	**87.86**	0.95		0.88
4	Parkinson	RF	95.30	0.99	0.92	0.95	**95.63**	0.99	0.94	0.96
		MLP	96.30	0.98	0.93	0.97	**96.64**	0.98	0.95	0.97
		NB	**77.18**	0.85	0.60	0.78	76.84	0.84	0.59	0.79
5	HeartCleveland2C	RF	**88.81**	0.96	0.96	0.90	87.88	0.95	0.91	0.88
		MLP	**86.64**	0.90	0.94	0.87	84.16	0.88	0.90	0.85
		NB	**75.46**	0.80	0.73	0.76	70.81	0.80	0.63	0.71
6	VertebColumn2C	RF	87.76	0.96	0.93	0.88	**89.15**	0.96	0.94	0.89
		MLP	**86.14**	0.91	0.94	0.87	85.45	0.92	0.92	0.86
		NB	*79.2*	0.87	0.87	0.80	79.20	0.87	0.87	0.80
7	VertebColumn3C	RF	88.61	0.98		0.88	**89.50**	0.98		0.89
		MLP	**86.83**	0.95		0.87	85.49	0.94		0.85
		NB	**80.58**	0.93		0.80	79.24	0.93		0.79

TABLE 6.9: Friedman Test Significance Results when applied on actual Baseline and Oversampling Performance Measures collectively (All) and per classifier

		$p-value$	Algorithms Rankings		
			Bl	ADASYN	FiltADASYN
OA	All	2.1×10^{-5}	1.19	2.45	2.36
	RF	3.7×10^{-3}	1	2.35	2.64
	MLP	3.8×10^{-5}	1	2.71	2.29
	NB	> 0.05	–	–	–
AUC	All	1.19×10^{-3}	1.43	2.31	2.26
	RF	1.7×10^{-3}	1	2.5	2.5
	MLP	4.2×10^{-2}	1.29	2.43	2.29
	NB	> 0.05	–	–	–
Sn_H	All	3.1×10^{-4}	1.2	2.46	2.33
	RF	1.6×10^{-2}	1	2.3	2.7
	MLP	1.6×10^{-2}	1	2.3	2.7
	NB	> 0.05	–	–	–
F_{score}	All	1.2×10^{-4}	1.26	2.40	2.33
	RF	2.5×10^{-3}	1	2.29	2.71
	MLP	5.8×10^{-3}	1.07	2.71	2.21
	NB	> 0.05	–	–	–

method with each classifier is detailed separately. The significance threshold θ for the $p - value$ is set to 0.05.

Table 6.9 shows the comparison for oversampling methods ADASYN and FiltADASYN on both RVA and benchmark data. A statistical difference exists for all considered performance measures using the collated classifiers results. The performance of FiltADASYN is particularly good with RF for all measures and it also has higher Sn_H ranking with MLP. It can be noted that the performance of NB shows negligible improvement with oversampling as its $p - value > \theta$. Although both FiltADASYN and ADASYN presents comparable results, FiltADASYN is used in the following chapters experiments as it presented better results on our RVA data.

6.3 Summary

In this chapter, an improved version of ADASYN oversampling called FiltADASYN is proposed to handle the characteristic of classes imbalance. It satisfies all of the identified criteria in subsection 4.2.2.

The impact of FiltADASYN is compared against ADASYN oversampling and the baseline classifiers performance on our RVA data. The general applicability of the suggested technique is additionally verified, where FiltADASYN is applied on selected benchmark medical datasets, using the same parameters settings used for RVA data.

FiltADASYN exhibits similar performance to ADASYN, while it provides significant improvement over the baseline. FiltADASYN offers a minimum of 9.27% accuracy improvement (100% improvement in Sn_H) over Baseline with RVA data. In addition, FiltADASYN with RF and MLP succeed in scoring the highest OA across all classifiers in five out of seven datasets.

Chapter 7

DD_Rank Feature Selection Method

In this Chapter, we propose a feature selection algorithm that ranks features according to their predictive performance and the ability to correctly reconstruct the features. Correct feature reconstruction relative to the input class labels indicates the presence of informative co-occurrence of the features and the class labels. We call this informative co-occurrence as the stability of the feature. A Restricted Boltzmann Machine (RBM) Architecture is used for classification and subsequent feature ranking. Restricted Boltzmann Machines (RBMs) are extensively applied in the field of computer vision and image segmentation to attain high accuracy levels [59, 81, 220]. RBMs are known to create higher order representation of the input features and to capture the interactions between the given features in the hidden layers. The created higher order representation and the captured interactions aid creating high performance models. On the other hand, the process lacks transparency as the significance of the features and their interactions are not explicitly output.

We use an existing Deep Disjunct Boltzmann machine to Rank features creating DD_Rank. In DD_Rank approach, a feature scoring and ranking approach is proposed that uses the properties of RBMs to decide on the predictive features and produce a ranked feature list that would aid the understandability of the model. To our knowledge, limited research was directed to their use for explicit feature selection in data mining domain. The proposed approach is presented in detail in section 7.1 and the related evaluation is discussed in section 7.2.

7.1 The Proposed Approach

The basis of RBMs [86, 21] can be presented as follows: an RBM is a two-layer stochastic neural network consisting of m visible units denoted with v, and n hidden units denoted with h. The two layers are fully connected. In contrast to Boltzmann Machines, connections are not available between units of the same type. The visible nodes often correspond to input measurements while hidden nodes are conditionally dependent on the measurements. RBMs have been originally developed to model binary inputs where values of $(v, h) \in \{0, 1\}^{m+n}$. The joint probability of (v, h)

can be expressed as:

$$p(v,h) = \frac{1}{Z}\exp(-E(v,h)) \tag{7.1}$$

where Z is the normalising constant

$$Z = \sum_{v}\sum_{h}\exp(-E(v,h)) \tag{7.2}$$

and E is the joint energy function that is minimised through the learned weights and biases:

$$E(v,h) = -\sum_{i=1}^{m}\sum_{j=1}^{n} w_{ij}v_i h_j - \sum_{i=1}^{m} b_i v_i - \sum_{j=1}^{n} c_j h_j \tag{7.3}$$

w_{ij} is a real valued weight between nodes v_i and h_j. b_i and c_j are real valued bias terms related to the i-th visible and j-th hidden variables, respectively. As shown in Equations (7.4) and (7.5), the probability of the hidden units is dependent on the visible units and vice versa.

$$p(v \mid h) = \prod_{i=1}^{m} p(v_i \mid h) \tag{7.4}$$

$$p(h \mid v) = \prod_{j=1}^{n} p(h_j \mid v) \tag{7.5}$$

To estimate the conditional probability at the hidden and visible nodes given binary input, neural network propagation rules are applied using Equations (7.6) and (7.7):

$$p(v_i = 1 \mid h) = f\left(b_i + \sum_{j=1}^{n} h_j w_{ij}\right) \tag{7.6}$$

$$p(h_j = 1 \mid v) = f\left(c_j + \sum_{i=1}^{m} v_i w_{ij}\right) \tag{7.7}$$

The most commonly used algorithm for learning model's parameters is contrastive divergence [211]. The contrastive divergence algorithm approximates the gradient of the log likelihood and performs gradient descent to minimise the energy function in the parameter space. The gradient is estimated using the difference between the actual values and the reconstructed ones on the visible nodes. The reconstructions are basically the sum of products of the values resulting on the hidden nodes and the learned machine parameters (weights w_{ij} and the biases b_i and c_j).

RBMs have been adapted to classification problems [125], as they could learn the association between the features and the classes. The reconstructions on the visible layer are used to signal how likely it is the feature vector belongs to a certain class.

RBMs were initially built to handle binary input data at visible units. Changes in the energy function formulation were applied in the literature to extend the use of RBMs to real-valued Gaussian input at visible nodes. Despite this, training Gaussian visible units remains a challenging task with Gaussian distributions at input visible

units [197]. The difficulty in training Gaussian RBMs and tuning their parameters can be attributed to several factors [40]. One of these factors can be the addition of a quadratic term to the energy function with a variance parameter to estimate [196]. The energy function formulation becomes as shown in equation 7.8.

$$E\left(v,h\right)=\sum_{i=1}^{m}\frac{\left(v_i-b_i\right)^2}{2\sigma_i^2}-\sum_{i=1}^{m}\sum_{j=1}^{n}w_{ij}\frac{v_i}{\sigma_i}h_j-\sum_{j=1}^{n}c_jh_j \tag{7.8}$$

where w_{ij} is a real valued weight between nodes v_i and h_j. b_i and c_j are real valued bias terms related to the i-th visible and j-th hidden variables, respectively. σ_i is the variance parameter. Other factors are (1) despite that the conditional independence (restriction) assumption for nodes simplifies several computations such as with probability and gradient calculation, it hinders the use of covariance information useful in reducing the input space complexity and better modeling the distribution and (2) the lack of upper bound on components size in reconstruction, both leading to unstable dynamics in the activity and weight update of the machine. Therefore, binary visible units are used, in this study, instead of the continuous extensions due to the known learning stability of binary-binary nets and the reported failures with Gaussian units [86]. Another reason that encouraged the use of binary-binary nets is that results in favor of data discretisation prior to classification have been recently reported [105, 140].

Based on the aforementioned reasons, an Equal Width (EW) binning procedure is adopted for discretisation of the input features into n_b bins. The same n_b is chosen for all features such that all the features have similar contribution (weights) within the machine to avoid any imposed bias in estimating the predictive value of each feature. A binarisation step follows where each feature is encoded as n_b bits and the bin value is indicated as 1 at the corresponding bit location and 0 at the remaining bits. Missing values are not omitted but considered as bin n_b+1. The bit values are assigned accordingly to maintain all the information the data may provide. Figure 7.1 illustrates the binarisation of a given value v_j within a feature vector f_i. The class labels are also binarised resulting in n_c bits corresponding to the number of classes per record.

Although single layer RBMs can approximate any binary distribution, their use is often impractical as it may require high number of hidden units and large amounts of training data [59]. Stacked RBMs (Deep Boltzmann Machine DBM, Figure 7.2(b)) offer a solution to this issue [59] since additional layers of latent variables are introduced that can capture high order dependencies between the hidden variables of previous layers, providing a rich model of complex structures using relatively few hidden units.

The RBM architecture used to perform explicit feature selection is shown conceptually in Figure 7.2(c)). The used architecture contains n_f separate RBMs in $layer_1$, where n_f corresponds to the number of features. Each RBM block represents a binarised feature Bf_i and binarised class label BC_i that are input at visible layer v of

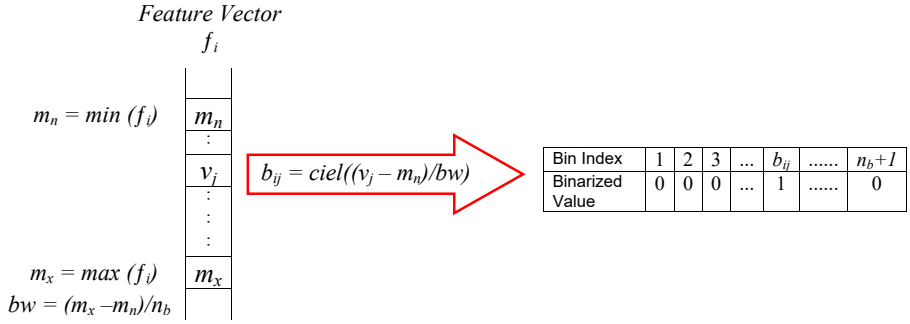


FIGURE 7.1: Binarisation of a sample value v_j of feature f_i, through calculating minimum feature value m_n, maximum feature value m_x and bw to determine the asserted bit index b_{ij}.

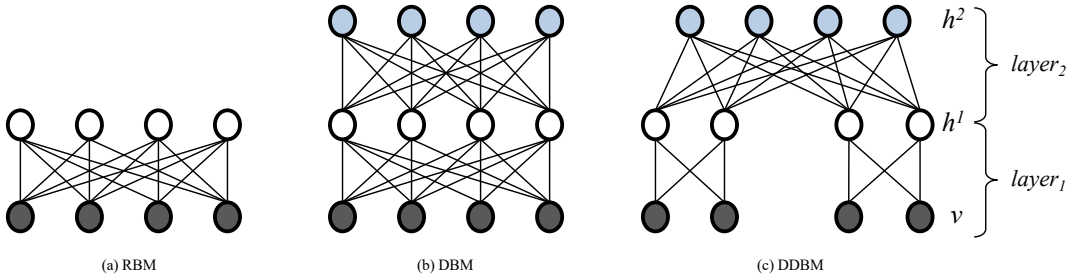


FIGURE 7.2: Restricted Boltzmann Machines Architectures

$layer_1$. At layers v and h^1, each RBM block comprises $(n_b + n_c + 1)$ visible and hidden nodes that are fully connected per block, where n_b is number of bins for feature binarisation and n_c is the number of classes. The hidden nodes h^1 of $layer_1$ are aggregated by a second continuous stacked RBM $layer_2$ of size $n_f \times (n_b + n_c + 1)$, creating the targeted Deep Boltzmann Machine structure. Every node at h^1 is connected to all nodes of h^2. The resultant architecture is called Deep Disjunct Boltzmann Machine (DDBM) architecture. The links between v and h^1 capture the dependencies between the features and the classes, while the links between h^1 and h^2 capture the synergy between the input features. The Disjunct RBM structure in $layer_1$ is selected such that the obtained classifications can be localised and attributed to a specific feature.

The proposed DD_Rank feature ranking procedure is outlined in Algorithm 3. Between lines 2 and 8 of the algorithm, the features and their respective classes are binarised and aggregated by the stacked RBM structure. Using the described architecture, input B_i which is the concatenation of Bf_i and BC_i to RBM_i of $layer_1$ is reconstructed using contrastive divergence, resulting in reconstructed value RB_i corresponding to input B_i. RB_i comprises the reconstructed binarised feature RBf_i and reconstructed binarised classes RBC_i (line 9). For each feature f_i, a selection score sc_i is calculated (line 16) given the reconstruction results. The score depends on two measures, the normalised reconstruction error and weighted value of area under precision-recall curve ($AUPRC$). The reconstruction error re_{ij} is calculated as in line 12, where rb_{ij} is the index of the set bit of the reconstructed RBf_i for record j, b_{ij} is the index of the asserted bit in the original Bf_i for record j and n_r is the number of

Algorithm 3 DD_Rank Algorithm: Explicit Feature Ranking using DDBM

1: **procedure** DD_RANK(F, C)
2: Binarise C into BC with n_c bits
3: **for all** $f_i \in F$ **do**
4: Binarise f_i into Bf_i with $n_b + 1$ bits
5: Concatenate Bf_i and BC into B_i
6: Feed B_i into RBM_i of $layer_1$
7: **end for**
8: Aggregate the n_f $RBMs$ of $layer_1$ using RBM_{n_f+1} of $layer_2$
9: Reconstruct inputs of $layer_1$ onto RB_i with RBf_i and RBC_i
10: Given $1 \leq i \leq n_f$
11: Compute reconstruction error re_i as:
12: $re_i \leftarrow \sum_j |rb_{ij} - b_{ij}|$, where $1 \leq j \leq n_r$
13: Normalise re_i as $nre_i \leftarrow \frac{re_i}{n_r n_b}$
14: Compute the area under precision-recall curve $AUPRC_i$ per f_i
15: Compute feature selection score sc_i as:
16: $sc_i \leftarrow AUPRC_i + (1 - nre_i)$
17: Rank f_i based on sc_i descending
18: **end procedure**

records.

In line 13, re_i is normalised by dividing re_i by n_r and maximum possible error n_b. Also, the area under precision-recall curve $AUPRC_i$ is calculated per feature based on RBC_i (line 14). The reconstruction error estimates how the learned weights and biases given this feature approximate the input and how they model the co-occurence (joint distribution) of the inputs and outputs. The precision-recall curve is used as a measure of performance suitable for balanced and imbalanced datasets. The objective of combining these performance measures is to determine the most stable (low re) and discriminative (high $AUPRC$) [198] features.

Following the ranking procedure, an exhaustive linear (incremental) search is carried out to establish the selected feature subset. In the incremental search, the selected feature subset is initialised as an empty set and the features are added linearly (one by one) by the same order of the ranking starting from the highest rank. The cut-off point for the subset size is chosen to be the point with the highest overall accuracy.

We argue that the DD_Rank feature selection procedure combines the advantages of wrapper and filter feature selection methods. This is achieved through ranking the features according to their classification performance based on DDBM reconstruction output, at a separate preprocessing step, where the produced ranking is independent of the actual classifier used for risk group prediction. In addition, upper hidden layer (h^2) capture the interaction between the features by the modeled weights, biases and subsequent reconstructions.

7.2 Results and Evaluation

In this section, we evaluate the performance of DD_Rank feature selection when applied on RVA data and standard benchmark datasets. The experiments were done using Weka [75] and MATLAB R2016b.

7.2.1 Experimental Study

In the following we briefly present the experiments objectives, the data, the tools, the implemented experimental design and the evaluation metrics used.

Experiments Objectives: The experiments objective is to clarify the impact of the proposed feature selection method (DD_Rank) on cardiovascular risk prediction using RVA data and on other benchmark datasets.

Experimental data: DD_Rank is applied on RVA-based measures dataset to evaluate its capability in cardiovascular risk prediction. Also, the method is applied on a set of benchmark medical datasets from UCI ML Repository. The details of the selected datasets were described in Table 5.5 in section 5.3.

Methods implementation: DD_Rank is implemented using MATLAB R2016b. The DeeBNet V3.0 Toolbox [112] for stacked RBMs implementation is employed for DD_Rank development.

Experimental tools: In order to verify the performance of the proposed method, three well established classifiers are used: RF, MLP and NB. Other classification algorithms can be readily applied as there are no restrictions imposed by our proposed approach. Same settings as depicted in Table 6.1 are applied.

Experimental procedure: The experiments are organised to evaluate the individual competence of the proposed method on RVA-based dataset and the described benchmark datasets 5.5.
DD_Rank is compared against well recognised methods to elucidate its effectiveness at handling the addressed problems. DD_Rank is evaluated against ReliefF and correlation coefficient (CorrCoef) feature ranking methods using Weka default implementation. The reported results are the average of five 10-fold cross validation runs on the available datasets. The statistical significance of the achieved results is evaluated using Friedman test [90].

Evaluation metrics: The utilised performance indicators are overall Accuracy (OA), Area under the ROC curve (AUC), F_{score}, and Sensitivity for High Risk Group when it is clearly marked as High Risk (Sn_H). Also, Sensitivity for Medium Risk Group is reported for RVA data only (Sn_M).

7.2.2 Cardiovascular Risk Prediction Results based on RVA Measurements

The effect of the proposed method DD_Rank is examined separately to determine the individual influence of feature selection on the performance and its corresponding significance. The results of classification methods on the original non-oversampled

full feature set of RVA-based measures are previously reported in Table 6.2 subsection 6.2.2 to serve as reference baseline performance and to help identify the requirements for designing the specialised solution for this problem.

Feature Selection: DD_Rank, ReliefF and CorrCoeff

DD_Rank using arbitrarily set $n_b = 10$, ReliefF and CorrCoeff are applied on the original dataset and the results are shown in Table 7.1. These results were selected based on highest Sn_H to point out whether feature selection can improve Sn_H in specific. Negligible Sn_H increase is achieved while the results show considerable OA improvement with MLP and NB using ReliefF and DD_Rank features (minimum improvement 6%) with average reduction of 56.7% in feature set size. This denotes the difficulty of classifying high risk samples from the original data as the OA increase is accounted for by slight increase in Sn of medium risk (especially with MLP using ReliefF and DD_Rank selected features) and increase in Sn of low risk. Another observation is that ReliefF, DD_Rank and CorrCoeff produce comparable results to baseline performance given single features namely RT_AF3, $tMDMC_AF2$ and MC_VF1 respectively which gives an indication of the importance of these measures.

Although the achieved results applying FiltADASYN and DD_Rank separately show improvement, the results do not meet the requirements of risk prediction. Hence, the impact of applying hybrid solution (OFSET) that addresses both characteristics of imbalance and high dimensionality is investigated. The composite effect of applying DD_Rank feature selection on POST FiltADASYN dataset to rank RVA measures will be presented in Chapter 9.

TABLE 7.1: Feature Selection Performance using Random Forest (RF), MultiLayer Perceptron (MLP) and Naive Bayes (NB) on Non-oversampled RVA data

	ReliefF Features						CorrCoeff Features					
	OA	AUC	Sn_H	Sn_M	F_{score}	$\#F$	OA	AUC	Sn_H	Sn_M	F_{score}	$\#F$
RF	**90.25**	0.40	0.0	0.0	0.86	7	86.44	0.60	0.3	0.07	0.86	1
MLP	**90.68**	0.62	0.0	0.42	0.86	1	82.62	0.51	0.1	0.0	0.83	55
NB	58.89	0.55	0.1	0.5	0.69	97	**60.17**	0.55	0.4	0.57	0.71	26

	DD_Rank Features					
	OA	AUC	Sn_H	Sn_M	F_{score}	$\#F$
RF	**90.25**	0.40	0.0	0.0	0.86	1
MLP	88.13	0.59	0.1	0.35	0.86	13
NB	56.81	0.54	0.3	0.42	0.68	67

7.2.3 DD_Rank General Applicability on Benchmark Medical Datasets

The performance of DD_Rank is demonstrated on 13 medical benchmark datasets from the UCI ML Repository. Baseline results of the classifiers RF, MLP, and NB

are reported in Table 6.7 subsection 6.2.3. Additionally, the OA of known recent solutions for these benchmark datasets are shown in Table 7.2. The methods in Table 7.2 can be grouped into methods that employ novel classification techniques [114, 167, 186] and approaches that develop new feature selection algorithms [3, 58, 121, 134, 145].

TABLE 7.2: Recent Methods Accuracy on Benchmark Datasets

Dataset	Method Source	OA(%)
Pima Diabetes	Seera et al.[167]	78.39
Ecoli	Kumar et al.[121]	77.00
Colic(SurgLesion)	Abe et al. [3]	85.33
Lymph	Emary et al.[58]	87.5
Dermatology	Liu et al.[134]	95.91
Parkinson	Khan et al. [114]	92.93
Heart Cleveland AllC	Kumar et al.[121]	55.00
Heart Cleveland 2C	Moradi et al. [145]	80.06
WBreast Diagnostic	Moradi et al. [145]	95.84
Breast Tissue	Kumar et al.[121]	56.00
Verteb Column 2C	Unal et al.[186] (raw data)	81.93
Verteb Column 3C	Unal et al.[186] (raw data)	83.23

*No results were reported in the literature for Colic (Outcome) Dataset

Feature Selection: DD_Rank, ReliefF and CorrCoeff

To illustrate and compare the impact of different feature selection methods on these benchmark datasets, ReliefF, DD_Rank and CorrCoeff feature selection algorithms are applied on the original datasets.

The results are reported in Table 7.3 where we mark as Baseline Performance (green shaded cells) the lack of performance improvement by the various feature selection methods over baseline performance. DD_Rank achieves the highest accuracy in 15 out of 39 of the experiments (39%) and comes second in 10 out of 39 of the cases (25%). ReliefF leads to the best overall accuracy in 14 out of 39 of the cases (36%). CorrCoeff is best in 28% of the cases (11 out of 39), thus ranking third. For the cases of Breast Tissue dataset with MLP learning and Lymph dataset with NB learning, all feature selection methods perform equally in terms of OA. DD_Rank improves OA over the baseline in all but three cases. At the same time, ReliefF fails to improve over baseline in 5 and CorrCoeff in 6 out of the 39 cases, respectively. In terms of AUC, Sn_H and F_{score}, ReliefF and DD_Rank achieve similar performance, and CorrCoeff slightly worse. Statistical significance analysis of the differences is detailed in subsection 7.2.4.

To compare the selection methods in terms of the number of features selected, we define the selected feature set proportion φ as the number of features selected over the number of features in the full feature set. Figure 7.3 shows a 3D scatter plot of the φ values for all methods and datasets. Each dimension represents the selected feature proportion (ratio) for a single classifier and the different colors represent the 13

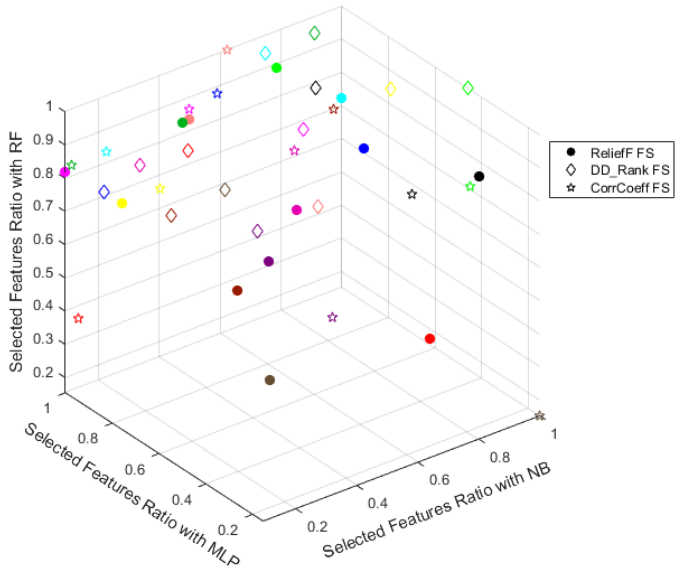


FIGURE 7.3: The Selected Features Proportions (Ratios) φ by each Feature Selection Method per Dataset where each dataset is indicated by a different colour

datasets. It can be observed that DD_Rank tends to select higher proportions of the feature set, especially with RF and MLP, as indicated by the clustering of DD_Rank points near the upper part of the z-plane where φ for RF and MLP approaches 1 (indicating no selection). In contrast, ReliefF and CorrCoeff points are more scattered, with lower φ in some cases, indicating selectiveness.

Next, we compare DD_Rank against the performance of existing known methods . We compare to DD_Rank performance on non-oversampled datasets only (see Table 7.3), to ensure a fair comparison given that none of the known methods apply oversampling. DD_Rank in combination with different classifiers achieves higher OA in 9 out of 12 benchmark datasets with the improvement ranging from 1.08% to 13.8%. This proves the capability of DD_Rank to better identify the key features of these datasets. For the three datasets Pima Diabetes, Lymph and Parkinson, DD_Rank produces lower OA, with differences of 1.83%, 1.02% and 0.11%, respectively. For Pima Diabetes [167] and Parkinson [114], the reported solutions are novel classification techniques. In Seera et al. work[167], a computation intensive multi-stage classifier is constructed from fuzzy min-max Neural Network, classification and regression tree and Random Forest. The multi-stage classifier is expected to give better results than a single classifier, especially for Pima Diabetes, which has a small feature set and hence is not likely to benefit significantly from feature set reduction. We consider the difference of 0.11% for Parkinson negligible, in particular, given that the solution offered in Khan et al. study [114] is specifically developed and adapted to the Parkinson dataset. In the case of Lymph, Emary et al. [58] propose three variants of a bio-inspired algorithm that uses Binary Ant Lion Optimisation (BALO) for feature selection and test it with three different initialisation methods, creating nine possible combinations. Only one combination is better than

DD_Rank (with fixed settings), while DD_Rank produces higher OA than seven of the remaining combinations, with a difference between 0.04% - 8.88% and a tie with the eighth combination. This indicates the potential of improving DD_Rank's OA through dataset specific customised settings.

7.2.4 Statistical Significance Analysis of Performance

In order to establish whether the performance of the proposed approach DD_Rank is significantly different from that of existing methods, we apply the Friedman test for continuous data [90].

For each of the measures OA, AUC, Sn_H and F_{score}, the actual values were used as input to the Friedman test. The significance values ($p-values$), together with the rankings of the baseline performance (classifiers performance on non-processed original dataset) (Bl) and the compared methods are provided are output by Friedman test. In each case, first, an overview of the overall performance is presented, by collating the results across all three classifiers. Then, the performance of each method with each classifier is detailed separately. The significance threshold θ for the $p-value$ is set to 0.05.

Table 7.4 compares the significance results of the feature selection methods ReliefF, DD_Rank (DD_R) and CorrCoef ($Corr$). In the collective case of all classifiers together, a significant difference is detected with all measures. Despite the fact that DD_Rank improved the OA in a larger number of cases, DD_Rank scores second after ReliefF using the Friedman test. This is attributed to its considerable incompatibility with NB, which degraded its overall ranking. An explanation for this incompatibility can be found as NB relies on the naive assumption of feature independence which does not conform with the DD_Rank assumptions.

Based on the significance results and the associated rankings, it can be concluded that DD_Rank is best suited to the MLP classifier, with which it achieved the best position for all measures.
To sum up, DD_Rank presents considerable improvement to baseline performance and provide a competitive performance when compared to state of the art methods.

7.3 Summary

In this Chapter, a feature selection technique dependent on Disjunct Deep Boltzmann Machine namely DD_Rank is proposed. The presented method demonstrates overall performance on par with well-established methods in feature selection (ReliefF and CorrCoeff). At the same time, DD_Rank provides better classifications with MLP as shown by the statistical significance results.

The general applicability of DD_Rank on benchmark medical datasets is additionally verified, where it outperforms previously reported solutions with 9 out of 12

TABLE 7.3: Feature Selection Performance on Original Non-Oversampled Benchmark Datasets

ID#	Dataset		ReliefF Features					DD_Rank Features					CorrCoeff Features				
			OA	*AUC*	Sn_H	F_{score}	#F	*OA*	*AUC*	Sn_H	F_{score}	#F	*OA*	*AUC*	Sn_H	F_{score}	#F
1	Pima Diabetes	RF	*75.65*	0.82	**0.61**	0.75	7	**76.56**	0.82	0.60	**0.77**	7	75.39	0.82	0.59	0.75	**5**
		MLP	*76.43*	0.81	0.50	0.76	2	76.30	0.82	**0.61**	0.76	7	**76.69**	**0.83**	0.60	0.77	**5**
		NB	**76.82**	0.80	**0.65**	**0.77**	7	Baseline Performance					*76.43*	**0.81**	0.62	0.76	3
2	Colic (outcome)	RF	*73.49*	**0.85**		**0.74**	18	73.22	**0.85**		0.71	**17**	**73.77**	0.84		0.73	19
		MLP	Baseline Performance					**70.49**	**0.79**		**0.71**	**19**	Baseline Performance				
		NB	**69.39**	**0.82**		0.70	18	*69.13*	0.81		0.70	17	68.85	0.81		0.70	**10**
3	Lymph	RF	*85.13*	0.94		0.85	**13**	**86.48**	0.94		**0.86**	17	83.78	**0.95**		0.84	15
		MLP	Baseline Performance					Baseline Performance					Baseline Performance				
		NB	*85.13*	0.89		0.84	17	*85.13*	0.89		0.85	17	*85.13*	0.89		0.85	16
4	Parkinson	RF	**93.33**	**0.96**	**0.79**	**0.93**	**5**	91.79	0.95	0.77	0.92	17	*92.07*	**0.96**	**0.79**	0.92	8
		MLP	**93.84**	0.96	0.88	**0.94**	**12**	92.82	0.96	**0.90**	0.93	21	*93.33*	0.96	**0.90**	0.94	21
		NB	*84.10*	0.88	0.56	0.84	2	81.03	0.65	0.63	0.72	2	**85.12**	**0.89**	**0.65**	**0.85**	2
5	Heart Cleveland 2C	RF	**80.82**	0.77	**0.52**	**0.81**	6	*79.90*	**0.78**	0.44	0.79	9	79.45	0.70	0.31	0.79	**2**
		MLP	79.9	0.76	0.44	0.79	4	**81.28**	**0.84**	**0.60**	**0.81**	11	*80.36*	0.74	0.34	0.80	**2**
		NB	**82.19**	**0.79**	**0.56**	**0.82**	6	80.82	0.75	0.51	0.81	4	*81.27*	0.76	0.46	0.81	**3**
6	Verteb Column 2C	RF	*84.51*	0.91	**0.77**	**0.85**	**4**	**84.83**	**0.93**	0.76	**0.85**	5	Baseline Performance				
		MLP	*86.45*	0.93	0.73	0.87	5	**86.77**	**0.94**	0.74	0.87	5	*86.45*	0.93	**0.75**	0.87	5
		NB	**80.32**	**0.91**	0.89	0.82	3	*79.67*	0.87	**0.90**	0.82	3	*79.67*	0.87	**0.90**	0.82	3
7	Verteb Column 3C	RF	*84.19*	**0.96**		0.84	5	**84.83**	0.95		**0.85**	**4**	Baseline Performance				
		MLP	Baseline Performance					**86.45**	**0.97**		**0.87**	4					
		NB						**84.19**	**0.96**		**0.84**	**4**					
8	Ecoli	RF	**86.31**	0.96		0.85	6	*86.09*	0.96		**0.86**	7	**86.31**	0.96		**0.86**	6
		MLP	*86.60*	0.96		0.86	7	**86.62**	0.96		**0.87**	6	*86.60*	0.96		**0.87**	7
		NB	*85.41*	0.96		0.86	6	**86.01**	0.96		0.86	5	*85.41*	0.96		0.86	6
9	Colic (SurgLesion)	RF	*86.14*	**0.89**	0.79	0.86	20	**86.41**	**0.89**	0.79	0.86	15	*86.14*	0.88	0.79	0.86	14
		MLP	*83.15*	**0.86**	0.74	0.83	**4**	*83.15*	0.85	0.74	0.83	14	**83.41**	**0.85**	0.74	0.83	5
		NB	*82.88*	**0.86**	0.75	0.83	5	82.33	0.85	0.76	0.82	**4**	**83.41**	0.85	**0.78**	**0.84**	**4**
10	Dermatology	RF	**98.09**	1.00		0.98	**29**	*97.54*	1.00		0.98	33	**98.09**	1.00		0.98	**29**
		MLP	*97.81*	1.00		0.98	33	**98.09**	1.00		0.98	33	**98.09**	1.00		0.98	33
		NB	**98.09**	1.00		0.98	32	*97.54*	1.00		0.98	33	*97.54*	1.00		0.98	**27**
11	Heart Cleveland All Classes	RF	Baseline Performance					Baseline Performance					**59.08**	**0.80**		**0.55**	**12**
		MLP	**59.07**	**0.79**		**0.57**	11	*58.74*	0.77		0.56	6	58.08	0.76		0.53	**2**
		NB	**60.06**	**0.81**		**0.58**	10	*57.42*	0.79		0.55	6	**60.06**	0.80		0.56	**3**
12	Wisconsin Breast Diagnostic	RF	**96.84**	0.99	**0.97**	0.97	**16**	96.66	0.99	0.94	0.97	27	*96.83*	0.99	0.94	0.97	20
		MLP	*96.84*	0.99	**0.97**	0.96	26	**97.53**	**1.0**	0.95	0.98	**24**	96.83	0.99	0.95	0.98	28
		NB	*95.78*	0.99	**0.96**	0.96	4	**96.30**	0.99	0.94	0.96	4	94.37	0.98	0.91	0.96	**3**
13	Breast Tissue	RF	**74.35**	0.93		**0.75**	**6**	*73.58*	0.93		0.74	8	*73.58*	0.93		0.74	7
		MLP	*69.81*	0.90		**0.71**	**4**	*69.81*	0.90		0.70	5	*69.81*	0.90		0.70	8
		NB	*70.75*	0.93		0.71	6	*70.75*	0.93		**0.72**	**4**	**71.70**	0.93		**0.72**	8

TABLE 7.4: Friedman Test Significance Results when applied on actual Baseline and Feature Selection Performance Measures collectively (All) and per classifier using significance threshold θ for p_value

		$p-value$	Algorithms Rankings			
			Bl	*ReliefF*	*DD_R*	*Corr*
OA	All	1.9×10^{-13}	1.3	**3.04**	2.93	2.71
	RF	8.8×10^{-5}	1.3	**3.20**	2.93	2.57
	MLP	1.1×10^{-5}	1.23	2.73	**3.3**	2.73
	NB	2×10^{-3}	1.4	**3.2**	2.57	2.83
AUC	All	1.1×10^{-2}	2.13	**2.74**	2.70	2.4
	RF	> 0.05	–	–	–	–
	MLP	3.84×10^{-4}	1.7	2.67	**3.27**	2.37
	NB	> 0.05	–	–	–	
Sn_H	All	1.0×10^{-2}	1.83	2.6	**2.85**	2.7
	RF	> 0.05	–	–	–	–
	MLP	7.4×10^{-3}	1.81	2.18	**3.13**	2.88
	NB	> 0.05	–	–	–	
F_{score}	All	2.38×10^{-11}	1.51	2.94	**2.97**	2.56
	RF	2.1×10^{-5}	1.5	**3.2**	3.13	2.17
	MLP	1.7×10^{-4}	1.47	2.67	**3.2**	2.67
	NB	9.9×10^{-4}	1.57	**2.97**	2.57	2.9

Best ranking is bolded

benchmark datasets with the improvement ranging from 1.08% to 13.8%. DD_Rank outperforms ReliefF and CorrCoeff in terms of accuracy in 15 out of 39 of the performed experiments and combines statistical relevance and actual predictive power for feature ranking.

The proposed method DD_Rank proves suitable to small and medium sized problems, while maintaining the capability to accommodate large amounts of data since this is an established advantage of deep Boltzmann machines [177]. The OFSET hybrid approach of FiltADASYN oversampling and DD_Rank feature selection, proposed to handle the imbalanced high dimensional RVA-based measures, will be discussed in Chapter 9.

Chapter 8

GCO_*mine* Classification Method

Graph Cut Optimisation *GCO* [30], which is a combinatorial optimisation technique, relies on max flow/ min cut principle to minimise a formulated problem.

GCO has been used as a solution to clustering problems [109, 55, 54]. Dhillon et al. [54] employed kernel *k*-means to optimise weighted graph cuts and overcome the equal-sized cluster restriction of Karypis and Kumar [109]. Despite the success of *GCO* in clustering problems, its use for classification problems in the data mining domain remains limited.

Extensive application of *GCO* can be found in image segmentation and object classification: *GCO* has been widely used with hyperspectral images [14, 50], retinal images, where *GCO* produces bi-labels (artery/vein) from segmented images [52, 103, 162] and flower segmentation from colour images [218].

Even though *GCO* has been included within a broad set of learning models, not all the presented models use *GCO* as a stand alone classifier. The described approaches either utilise pre-classifiers (e.g. SVM and K-means) [14, 50] or apply *GCO* on presegmented images [52, 103, 162] or apply significant image-specific preprocessing before *GCO* [218].

We consider that further development of *GCO*-based generic classification methods is a promising avenue as (1) *GCO* has lead to high accuracy classification for a range of problems and (2) it is a low cost method with guaranteed optimality bounds as shown by Boykov et al [29, 30]. Also, utilising its formulation to modify the typical purely local instance-based learning approach would achieve better generalisation of the learning decision.

The main objective is to propose a classification method that can correctly identify subjects from different cardiovascular risk groups based solely on RVA data, while providing an understandable model to the user. We propose a partially lazy mining (classification) algorithm based on Graph Cut Optimisation *GCO_mine* that aggregates local connectivities into a globally connected graph, on which a global classification decision is taken. In addition, *GCO_mine* introduces the concept of a sample's direct membership (distance) to the given classes, which does not exist in traditional lazy approaches such as kNN. Figure 8.1 depict the impact of adding class membership to the formulation, where two class labels exist (Class 1 and Class

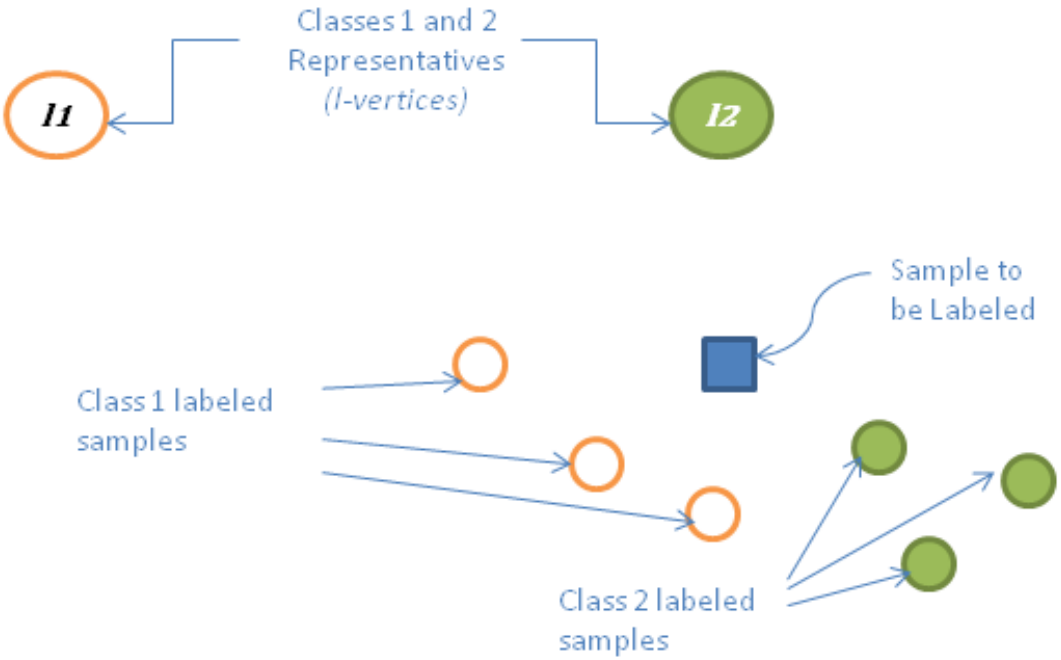


FIGURE 8.1: A two dimensional graph segment illustrating adding class representatives to Lazy Approaches

2) and a test sample (denoted by a square) is to be labeled and its ground truth label is Class 2. In the neighbourhood of the test sample, the majority instances come from class 1. Thus, if we apply simple kNN classification (for k = 3 for example), the test sample will be labeled as Class 1. The addition of class representatives (l1 and l2) will direct the classification towards class 2 as the distance between the test sample and the representatives overrides the majority vote. The *GCO_mine* approach strikes a favourable balance between merely local instance-based lazy methods and eager techniques, which build global latent models of the training data in a separate phase.

The novel classification algorithm *GCO_mine* can produce reliable risk prediction and can handle the specifics of the data. The collected RVA data, similar to medical data, have fuzzy overlapping boundaries between risk groups where the limits are not crisply defined and various interactions exist between the features. Despite that currently the data set size is small and extremely imbalanced and that oversampling was used to rectify this problem, data is expected to remain to be collected. Therefore, a method that can adapt to newly arriving data when they become available is needed. Thus, instead of a global abstraction classification model using eager methods, an instance-based approach is proposed here. Our *GCO_mine* method (similar to other instance-based lazy methods) is particularly suitable, because it respects the individual variation within one risk group and manages the overlapping pre-morbid and normal ranges of features, with the aptitude to accommodate expanding datasets. *GCO_mine* is devised to better handle border samples than existing lazy methods as discussed earlier on Figure 8.1

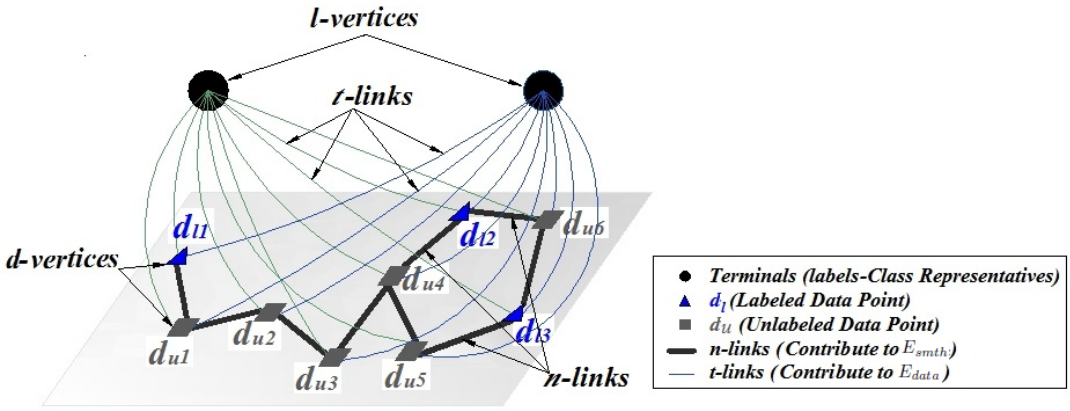


FIGURE 8.2: Graph Formulation of Classification Problem

8.1 The Proposed Approach

In this section, we introduce *GCO_mine*, a partially lazy classification method that employs multi-label graph cut optimisation (*GCO*). *GCO_mine* formulates the classification model as a graph cut minimisation problem. *GCO_mine* reaches a solution by incorporating a smoothness function, which employs similarity information from *both* training and test instances. The graph formulation introduces connectivity among the local neighbourhoods, thus allowing for the study of the global structure of the classes. We consider *GCO_mine* as a partially lazy classifier as it includes caching intermediate results for parameter settings [4].

8.1.1 Graph Cut Formulation of *GCO_mine*

Using graph cut optimisation [30, 118], we formulate our classification problem as an undirected graph. In an undirected graph, there exists a set of vertices v and edges e that connect these vertices. Each edge e_i is assigned a non-negative cost (weight) c_i denoting the penalty of cutting the edge e_i between two vertices. There is a special type of vertices called the terminals. Each terminal of the graph represents a class label creating l-vertices for l class problems. The other vertices correspond to the data points (records). Given a dataset, its data points are represented by d-vertices. In our formulation, there exist d_l-vertices which belong to the set of labeled data points D_l and d_u-vertices for unlabeled data points D_u. Each d-vertex is connected to all the terminals (l-vertices) through t-links of different costs. Also, the neighbouring data points have weighted links called n-links. The neighbouring data points are defined as the nearest m data points to a sample in terms of Euclidean distance.

The described graph cut formulation structure is outlined in Figure 8.2. The energy of a cut per iteration depends on the costs of the severed t-links and n-links in a single α-expansion move. In α-expansion, random labeling f is initially performed. Then, for each label α a new labeling $\hat{f}$ is generated within one α-expansion, where the data points labeled as α increases while considering all other labels as $\bar{\alpha}$. The α-expansion is guided by the principle of minimum flow between source and sink. The

new labeling is approved and adopted when the energy of $\hat{f}$ is lower than energy of f. The relabeling process through α-expansion is repeated until no improvement in energy is achieved. (see [30] for details). Only the expansion move is utilised for optimisation due to its guaranteed proven optimality properties within a factor of the global minimum. The factor depends on the pairwise interaction between neighbouring data points [30]. The objective is to find the cut with minimum energy E_{Total}, that partitions the data points d-vertices such that each d-vertex is associated with a single l-vertex corresponding to its label. When the designated cut is reached, a suboptimal labeling $Class_{labels}$ is realised.

8.1.2 Proposed Energy Function Definition for *GCO_mine*

The energy of a cut E_{Total} is defined as the sum of the costs of the edges it severs when the cut is formulated as will be shown in this section. E_{Total} is based on the data energy E_{data} and smoothness energy E_{smth}:

$$E_{Total} = E_{data} + \delta E_{smth} \tag{8.1}$$

where E_{data} measures the conformity of the data points with each label represented by a l-vertex (class representative), δ determines the contribution weight of E_{smth}, and E_{smth} quantifies the interaction penalty of the neighbouring data points (represented by the d_l and d_u vertices) with each other. For each candidate cut, E_{data} and E_{smth} are computed corresponding to the costs of the t-links and n-links, respectively. Both energy components are calculated using the standardised Euclidean distance ξ between data points (equations 8.3 and 8.7). A standardised distance metric is chosen to balance the contribution of each feature to the cost, as all features are converted to the same scale.

The data energy E_{data} (equation 8.2) comprises two cost functions: C_u and C_l, where C_u computes the cost of assigning an unlabeled data point d_i to a class l_i, where $d_i \in D_u$ (equation 8.3) and C_l sets the cost of classifying a labeled training data point d_i to class l_i, where $d_i \in \mathcal{T}$ and $\mathcal{T} \subset D_l$ (equation 8.4). In order to avoid the use of noise or outliers in the classification and energy calculations (guided sampling), only a good representative subset $\mathcal{T}$ of the labeled samples D_l can be chosen through guided sampling and used for establishing the neighbourhoods and cost calculation. Another possible aim for using a subset of labeled samples can be to reduce computational cost, which can be achieved through random sampling. The remaining (not sampled) data points are left out by the method. The sampling rate is allowed to reach 1 using all the set of labeled samples depending on performance. E_{data}, C_u and C_l are defined as:

$$E_{data} = \sum_{d_i \in D_u} C_u(l_i|\, d_i) + \sum_{d_i \in \mathcal{T}} C_l(l_i|\, d_i) \tag{8.2}$$

$$C_u(l_i|\, d_i) = \xi(d_i, \eta) \tag{8.3}$$

$$C_l(l_i|\, d_i) = \begin{cases} 0, & \text{if } l_i = \mathfrak{C} \\ \infty, & \text{otherwise.} \end{cases} \tag{8.4}$$

C_u is measured as the distance between d_i and the representative η of class l_i. η is selected from the training subset $\mathcal{T}$. For a labeled data point d_i, the cost C_l is set to zero when l_i is the 'ground truth' target class $\mathfrak{C}$ and to ∞ (practically the largest integer) in all other cases, to direct the chosen cut and guide the classification process. This extremely large distance acts as a high penalty imposed for misclassification.

E_{smth} is calculated as the sum of the normalised costs of assigning two neighbours d_i and d_j to different classes (cutting their n-link), $\omega(d_i, d_j)$ (equation 8.5). This cost is calculated as the difference between the local maximum distance and the pairwise distance normalised to the local maximum distance (equation 8.7). The local maximum distance $\max_{p\in\aleph}(\xi(d_i, d_p))$ is the largest pairwise distance among the calculated distances between d_i and a number m of its nearest neighbours $\aleph$, where $\aleph \subset \{D_u \cup \mathcal{T}\}$. Traditional lazy learning methods include only labeled samples for establishing neighbourhoods. Unlike these methods, in *GCO_mine* both labeled and unlabeled data can contribute to the classification decision of a point d_i, to achieve better structured classes.

$$E_{smth} = \sum_{d_i} \sum_{d_j \in \aleph} V(l_i, l_j|\, d_i, d_j).\, \omega(d_i, d_j) \tag{8.5}$$

$$V(l_i, l_j|\, d_i, d_j) = \begin{cases} 0, & \text{if } l_i = l_j \\ 1, & \text{otherwise} \end{cases} \tag{8.6}$$

$$\omega(d_i, d_j) = \frac{\max_{p\in\aleph}(\xi(d_i, d_p)) - \xi(d_i, d_j)}{\max_{p\in\aleph}(\xi(d_i, d_p))} \tag{8.7}$$

After proposing these energy definitions, it can be seen that two aspects would affect the classification process: the choice of the class representative η and the selection of the training samples subset $\mathcal{T}$. Firstly, the choice of a class representative significantly influences the C_u contributing to E_{data}, therefore it is important to investigate the impact of the representative choice. Two class representatives are studied, namely:

- Centroid (*Cent*): η is the data point with minimum overall (sum) distance to all $\mathfrak{C}$ members
- Closest-point (*Close*): η is a data point from $\mathfrak{C}$ with minimum distance to point d_i.

Secondly, the selection (sampling) method for the labeled subset ($\mathcal{T}$) and the number of selected instances need to be considered. The commonly used sampling strategies are random and guided. In random selection ($_r$), the training samples are drawn arbitrarily from the training set. This sampling process is fast, simple and easy to apply, but may lead to the selection of clustered samples or outliers. A guided selection ($_g$) method is proposed, here, to address this issue and select a well representative subset. The aim of the developed guided method is to sample labeled data points of each class that are uniformly distributed in the feature space, while avoiding outliers. For this purpose, the centroid of each class is determined and the angle space around the centroid is partitioned into n_s slices, where n_s is the number of samples to be drawn. The angle between the class training samples and the centroid is calculated to locate each training instance within an angle partition. Then for each angle partition, the training instance closest to the median is selected to ensure that it is an appropriate representative (i.e. not an outlier). The number of selected instances n_s by both random and guided sampling depends on a predefined sampling rate S_r and the number of training samples per class n_{tc} such that $n_s = S_r \times n_{tc}$.

The main steps of the *GCO_mine* method are outlined in Algorithm 4. The differences in subset selection of the training samples and the class representative choices lead to variations in $Sample\_Training(D_l, S_r)$ (line 2) and $Compute\_EData(D_u, \mathcal{T})$ (line 3), respectively. Thus, there are four variants of *GCO_mine*: $GCO\_mine_{r,Cent}$, $GCO\_mine_{g,Cent}$, $GCO\_mine_{r,close}$ and $GCO\_mine_{g,close}$.

Algorithm 4 *GCO_mine*: Non-Parametric Classification via GCO

1: **procedure** $GCO\_mine(D_l, D_u, \delta, m, S_r)$
2: $\quad \mathcal{T} = Sample\_Training(D_l, S_r)$
3: $\quad E_{data} = Compute\_EData(D_u, \mathcal{T})$
4: $\quad E_{smth} = Compute\_ESmooth(D_u, \mathcal{T}, m)$
5: $\quad Class_{labels} = GCO(D_u, E_{data}, E_{smth}, \delta)$
6: **end procedure**

8.1.3 Parameter Setting for *GCO_mine* variants

The performance of *GCO_mine* is controlled by hyper parameters: the contribution weight of E_{smth} (δ), the number of nearest neighbours to establish the neighbourhoods m and sampling rate S_r. The optimisation of these parameters can be done by search methods such as grid or heuristic search. Despite its simplicity, blind grid search becomes inefficient as the number of parameters increases and is not practical for searching continuous spaces. In contrast, genetic algorithms (GA) scale well with the increase in parameter numbers. Since the parameter space of *GCO_mine* variants is limited, both blind search and genetic algorithms can be applied in conjunction with the proposed variants to reach a near optimal setting. We employ both search techniques and compare them. The performance with each setting is evaluated using cross validation. In blind search, the parameter space is uniformly

TABLE 8.1: Medical Datasets Partitions Quality Characteristics: Silhouette (S), Davies Bouldin (DB) and Calinski Harabasz (CH) together with each dataset number of classes (#C)

Dataset	#C	S	DB	CH
Oversampled RVA-based Measures	3	0.15	5.30	21.81
Pima Diabetes	2	0.16	4.42	24.29
Ecoli	8	0.35	1.57	81.17
Parkinson	2	0.25	2.85	13.05
Wisc.Breast Diagnostic	2	0.61	0.72	633.63
Breast Tissue	6	-0.36	3.69	6.73
VertebColumn 2C	2	0.09	1.56	55.65
VertebColumn 3C	3	0.08	2.09	97.71
Colic#1 (Surg_lesion)	2	-0.10	5.1	4.78
Colic#2 (Outcome)	3	0.04	2.94	26.13
Lymph	4	0.15	1.86	11.50
Dermatology	6	-0.13	5.82	2.07
HeartCleveland 2C	2	-0.05	6.30	2.77
HeartCleveland AllC	5	-0.11	10.30	3.30

partitioned and the accuracy of the variants of *GCO_mine* with each setting is considered. With GAs, each chromosome encodes a possible set of values for δ, m and S_r and the fitness of the solution is defined as the overall accuracy.

8.2 Results and Evaluation

We first introduce our experimental study and then discuss the results on both our RVA data and UCI ML Repository datasets, using blind search and genetic algorithms for parameter setting.

8.2.1 Experimental Study

The experiments employ the Weka [75] and MATLAB environments. *Experiments Objectives:* An extensive set of experiments are conducted to validate the effectiveness and suitability of *GCO_mine* on our RVA dataset. Additionally, the general applicability of the proposed method is demonstrated using 13 benchmark datasets.

Experimental data: GCO_mine is applied on both the original and the oversampled RVA-based measures. In addition, it is applied on the original 13 benchmark medical datasets that are selected from UCI ML Repository outlined in Table 5.5 section 5.3.

Methods Implementation and Settings: Variants of *GCO_mine* are implemented using MATLAB R2016b. They rely on the MATLAB implementation of GCO Toolbox v3.0 [30, 29, 118]. For setting the parameters of *GCO_mine* variants, the allowed ranges are: [0.1, 20] for δ, $\{1, 2, \ldots, 10\}$ for m and [0.1, 0.9] for S_r. These ranges are selected to offer a balance between performance and computation complexity.

Experimental Tools: We compare the results of the proposed algorithm with those of four well established classifiers: RF, MLP, NB (same settings as depicted in Table 6.1 are applied) and K-nearest neighbours (kNN). Standardised Euclidean metric is

used for distance calculation in kNN and k us allowed the same range of values of m in *GCO_mine*. kNN is used for the current experiments only, as it is a popular lazy classifier, to which the proposed method is similar, because it bases the classification decision on the neighbourhood of the instance. Therefore, we report its results to show the influence of the proposed modification on the traditional lazy formulation, rather than the difference between lazy and eager method that would be illustrated by RF, MLP and NB. The Weka implementation of RF, MLP, NB and kNN is utilised in the experiments.

Experimental Procedure: The proposed algorithm variants are first applied on the original non-oversampled RVA data to confirm the need for imbalanced data solution using different classifiers and to decide on the winning variant to continue the experiments with. Then, the performance of the winning variant when applied on the oversampled RVA dataset keeping its original imbalance ($\beta = 0.25$) is reported to be able to compare its performance against the other classifiers in all of the studied cases. Also, it is applied on the fully balanced oversampled RVA data including 212 low, 211 medium and 208 high risk samples of 104 features. In addition, the variant with the best accuracy is used to compare the performance of FRS, QRisk and RVA measures. Moreover, the best performing variant is applied on the benchmark datasets and its performance is compared to the selected well-established classifiers.

In order to study how the proposed method handles the differences in the structure of feature spaces, the cohesion and separation properties of the classes in each dataset are depicted as shown in Table 8.1. Three recognised evaluation measures are reported: Silhouette index (S), Davies Bouldin index (DB) and Calinski-Harabasz criterion (CH) [135].
The reported results are the average of five 10-fold cross validation runs on the available datasets.

Evaluation metrics: The utilised performance indicators are OA, AUC, F_{score}, and Sn_H when the class is clearly marked as High risk. Also, Sn_M is reported for RVA data only.

For RVA data, the total computation time t_c needed for parameter space search and training subset selection for the winning *GCO_mine* variant is recorded. Also, the execution times of a single run t_{sr} for all classifiers when applied on the fully balanced RVA data are compared.

8.2.2 Cardiovascular Risk Prediction based on RVA measurements

GCO_mine Variants Comparison on RVA Data

The *GCO_mine* variants are applied on the original RVA-based measures in order to both examine whether it could provide a sound cardiovascular risk prediction, relative to established methods, from the available data and to uncover the need for a composite solution that handles the characteristics of RVA data. The OA and Sn_H results are shown in Table 8.2. The results further illustrate the imbalance problem

TABLE 8.2: *GCO_mine* Variants Performance (Overall accuracy and High Risk group Sensitivity) on Original RVA data

Variant	Blind Search (*b*)		Genetic Algorithm (*i*)	
	OA	Sn_H	OA	Sn_H
$GCO\_mine_{r,Cent}$	82.04	0	83.21	0
$GCO\_mine_{g,Cent}$	83.46	0	85.89	0
$GCO\_mine_{r,close}$	84.72	0	88.69	0.1
$GCO\_mine_{g,close}$	88.89	0	90.11	0.1

TABLE 8.3: Classifiers Performance using the selected performance mtrics on Oversampled RVA measures while keeping the initial imbalance ratio

	OA	AUC	Sn_H	Sn_M	F_{score}
RF	96.40	0.99	0.76	0.53	0.96
MLP	98.76	0.98	0.81	0.80	0.98
NB	53.49	0.82	0.83	0.78	0.64
kNN	95.64	0.92	0.86	0.87	0.96
hGCO_mine	99.58	0.99	0.98	0.95	

TABLE 8.4: Classifiers Performance using the selected performance measures and time of a single run on fully balanced Oversampled RVA data $\beta = 1$

	OA	AUC	Sn_H	Sn_M	F_{score}	t_{sr}
RF	**99.52**	0.99	*1*	**0.99**	0.99	4.98
MLP	93.84	0.97	*1*	0.92	0.95	235.76
NB	79.59	0.94	0.92	0.92	0.80	1.26
kNN	84.45	0.89	0.84	0.85	0.83	**0.98**
hGCO_mine	**99.52**	**0.996**	*1*	**0.99**	**0.992**	2.13

as previously shown in subsection 6.2.2: Table 6.2, where RF and MLP manifested the same pattern of high OA and negligible Sn_H . Also, the *GCO_mine* variants present superior performance compared to NB and MLP when applied to original (non-oversampled full feature set) RVA data (Table 6.2).

Moreover, when comparing the performance of the *GCO_mine* variants shown in Table 8.2, it can be observed that *close* variants outperform their counterpart *cent* variants with an improvement of OA ranging from 2.68% to 5.48%. This can be explained by the fact that *cent* fails to represent scattered and partially coinciding classes, characteristic to our RVA dataset due to the presence of subjects at the borderline between two risk groups. Also, guided sampling contributes to better classification accuracy, at the expense of computation time. Thus, the selection of the sampling method becomes a design decision between high accuracy ($_g$) and low computation time ($_r$). For CVD risk prediction, high accuracy is essential to avoid consequences of misclassification, which can be detrimental for missed high and medium risk patients or costly for allocation of low risk patients onto unnecessary treatment plans. Therefore, the highest accuracy ($GCO\_mine_{g,close}$ with GA parameter setting) denoted as (*hGCO_mine*) is chosen for further investigation and comparison to RF, MLP, NB and kNN.

The winning *GCO_mine* variant comparison to RF, MLP, NB and kNN

The RVA-based measures are oversampled such that the original base rate (Imbalance Ration IR) is maintained. The winning variant of *GCO_mine* with heuristic genetic algorithms denoted as *hGCO_mine* together with the set of classifiers RF, MLP, NB and kNN are applied on the base rate post-oversampling resultant data set. The result of these experiments are shown in Table 8.3. The results portray the superiority of *hGCO_mine* over its counterparts considering all the evaluation metrics except when compared to the AUC of RF. Despite the high results presented by *hGCO_mine*, it fails to attain Sn_M and Sn_H of 1 leading to possible misclassifications of real subjects critically at risk. In addition, the performance of all the other classifiers shows the same limitation. Hence, we will apply *hGCO_mine* on the fully balanced dataset aiming for higher sensitivities.

Oversampling is applied on the minority classes only creating three fully balanced classes. The results of *hGCO_mine* and the set of classifiers RF, MLP, NB and kNN on the fully balanced data set are shown in Table 8.4. The results portray the superiority of *hGCO_mine* over its counterparts (except RF) considering the OA, AUC and F_{score} evaluation metrics, while a Sn_H of 1 is achieved by all algorithms except kNN and NB. On the other hand, the least execution time t_{sr} is offered by kNN, while $GCO\_mine_{g,close}$ presents an average OA improvement of 14% over the accuracies of the faster kNN and NB alternatives. Compared to RF, *GhGCO_mine* improves AUC, F_{score} and reduces t_{sr} to 43%. It is to be noted that the total time for parameter setting and representative subset selection is 2056.42 sec which is a considerable overhead.

FRS, QRisk and RVA Measures Comparison

The proposed *hGCO_mine* and the used classifiers are applied on to the fully balanced FiltADASYN oversampled FRS measures [49] and QRisk measures [87] for the same samples. The results are shown in Table 8.5 then compared to that of RVA measures, to clarify the prospect of RVA-based features in predicting cardiovascular risk. The FRS measures set contains the traditional factors of age, gender, total cholesterol (Chol), HDL cholesterol, systolic blood pressure (SBP) and SBP treatment (yes or no), Diabetes Mellitus (DM) (yes or no) and Current smoking (yes or no) status. The QRisk factors comparison set includes age, gender, FH of CVD, Chol/HDL, Smoker, DM, SBP, BMI and Ethnicity. The results re-affirm the superiority of RVA measures in risk stratification (as illustrated in Table 8.4 in contrast to 8.5 and as previously shown in Table 5.4). Also, *hGCO_mine* maintains its lead over kNN and NB.

TABLE 8.5: Various Classifiers Performance using FiltADASYN oversampling on FRS and QRisk measures producing fully balanced datasets

	FRS Measures					QRisk Measures				
	OA	AuC	Sn_H	Sn_M	F_{score}	OA	AuC	Sn_H	Sn_M	F_{score}
RF	90.89	.98	0.89	0.97	0.90	**94.50**	0.99	0.95	0.97	0.95
MLP	93.56	0.98	0.93	0.99	0.94	**97.95**	0.99	0.97	0.99	0.98
NB	64.69	0.88	0.24	0.88	0.61	**78.96**	0.90	0.69	0.995	0.79
kNN	72.68	0.80	0.18	1.0	0.67	**76.76**	0.82	0.30	1.0	0.73
hGCO_mine	85.87	0.96	0.86	0.9	0.80	**93.17**	0.98	0.91	0.95	0.90

The relative accuracies of the FRS, QRisk and RVA-based measures with the applied classifiers are shown in Figure 8.3. The illustrated accuracies convey the advantage of RVA-based measures over the known risk measures (FRS and QRisk) as it surpasses FRS and QRisk measures with *hGCO_mine*, kNN, RF and NB, while it is slightly worse than QRisk with MLP. On the whole, the shown results demonstrate the high potential of RVA-based measures in cardiovascular risk prediction.

8.2.3 *GCO_mine* General Applicability on Benchmark Medical Data

$GCO\_mine_{g,close}$, RF, MLP, NB and kNN are applied on the benchmark datasets previously outlined in Table 8.1. Blind search and heuristic search using genetic algorithms denoted by *b* and *h* respectively are employed for parameter setting with $GCO\_mine_{g,close}$. Table 8.6 and Table 8.7 depict the performance evaluation results. Table 8.6 illustrates the results of the datasets with features having real continuous values, while Table 8.7 present results on the datasets that include categorical variables. As shown, $GCO\_mine_{g,close}$ is particularly effective on datasets of continuous real features and it presents a competitive performance on categorical datasets. This can be attributed to the utilisation of standardised Euclidean distance metric in data cost calculation, since this metric is designed for real valued samples. An

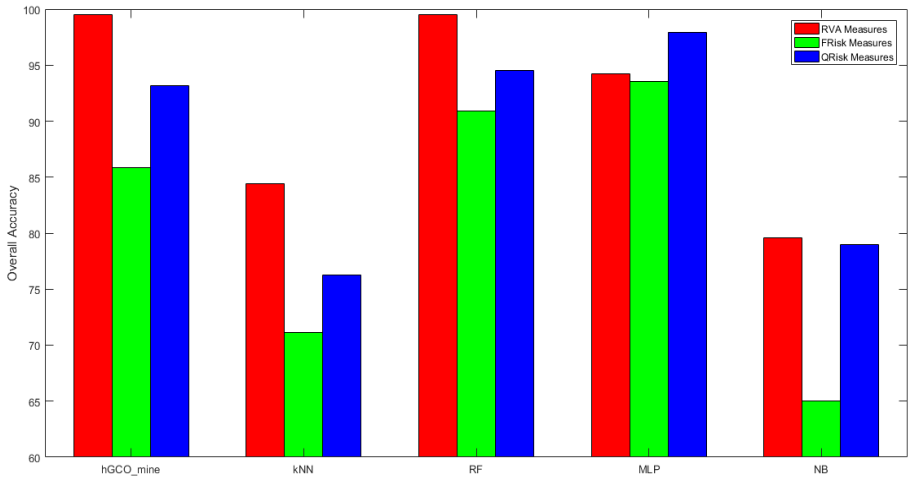


FIGURE 8.3: Classifiers Accuracy using Various Oversampled Risk Measures

improvement could be achieved for categorical data through the future adoption of a specially designed distance function.

Overall, $GCO\_mine_{g,close}$ has higher OA on 8 out of 13 datasets with differences ranging from 0.13 % to 2.56 % to the second highest accuracy value. In some cases, $GCO\_mine_{g,close}$ shows a remarkable accuracy increase such as when applied on the Parkinson dataset, a 25 % increase is recorded when compared to NB. In comparison to kNN (a nonparametric classifier of similar principle), $GCO\_mine_{g,close}$ is superior in 10 out of 13 datasets; the improvement reaches 16.12 %. Even though GCO_mine and kNN rely on a similar concept for data cost determination, the inclusion of unlabeled samples in smoothing energy E_{smth} together with adding direct class membership through E_{data} and the adoption of graph cut optimisation lead to better classification on the overall compared to kNN. However, the performance of $GCO\_mine_{g,close}$ is lower for datasets of low partitions quality: Colic# 1, Dermatology and HeartClevel2C. Their low partition quality is indicated by negative silhouette index values, relatively high values of Davies Bouldin index and low Calinski-Harabasz criterion values.

For continuous variables (Table 8.6) GA performs better than blind search on all datasets with the exception of Verteb Column 2C, while in the case of categorical variables (Table 8.7) the difference between the results of the two methods is not consistent, i.e., equal in two cases, in one case blind search is better, in two cases GA is better. In order to understand the reasons for these, in Figure 8.4 we plot the error (z-axis) against the two parameters $\delta \in [0.1, 20]$ (y-axis) and $S_r \in [0.1, 0.9]$ (x-axis), for the number of neighbours m chosen by the two methods, blind search (left) and GA (right), respectively. The two datasets chosen for comparison are Breast Tissue, as a representative with difference in accuracy and Colic#1, as representative for equal performance. The resulting phenotype fitness landscape in Figure 8.4 (a) is more rugged and more complex, with several local optima, whereas the landscape in Figure 8.4 (b) is smoother and simpler, with less local optima. Thus, it can be seen

TABLE 8.6: Classifiers Models Performance using the selected performance measures on Continuous Datasets

		OA	AUC	Sn_H	F_{score}
Pima Diabetes	RF	75.26	0.81	0.60	0.75
	MLP	75.13	0.79	0.61	0.75
	NB	76.30	0.81	0.61	0.76
	kNN	72.52	0.79	0.58	0.72
	bGCO_mine	76.32	0.80	0.59	0.76
	hGCO_mine	**77.24**	0.83	0.65	0.77
Ecoli	RF	86.09	0.96		0.86
	MLP	85.71	0.95		0.86
	NB	85.41	0.96		0.86
	kNN	86.90	0.95		0.86
	bGCO_mine	88.48	0.97		0.87
	hGCO_mine	**89.39**	0.98		**0.89**
Parkinson	RF	91.28	0.96	0.75	0.91
	MLP	91.28	0.96	0.83	0.91
	NB	69.23	0.86	0.61	0.75
	kNN	93.84	0.98	0.81	0.94
	bGCO_mine	93.16	0.98	0.77	0.93
	hGCO_mine	**94.21**	0.98	0.81	0.94
Wisconsin Breast Diagnostic	RF	95.78	0.99	0.98	0.96
	MLP	96.66	0.99	0.95	0.97
	NB	92.97	0.98	0.94	0.93
	kNN	95.95	0.95	0.97	0.96
	bGCO_mine	96.25	0.99	0.96	0.97
	hGCO_mine	**96.79**	0.99	0.98	0.97
Breast Tissue	RF	71.69	0.93		0.72
	MLP	64.15	0.88		0.65
	NB	70.75	0.93		0.71
	kNN	71.69	0.83		0.72
	bGCO_mine	69.00	0.83		0.70
	hGCO_mine	**72.00**	0.93		0.73
Verteb Column 2C	RF	84.19	0.93	0.76	0.84
	MLP	84.51	0.93	0.70	0.85
	NB	77.74	0.88	0.87	0.80
	kNN	80.00	0.86	0.68	0.80
	bGCO_mine	**87.00**	0.93	0.78	0.86
	hGCO_mine	86.67	0.94	0.78	0.86
Verteb Column 3C	RF	83.54	0.96		0.84
	MLP	85.48	0.96		*0.86*
	NB	83.23	0.95		0.83
	kNN	77.42	0.91		0.78
	bGCO_mine	86.33	0.98		0.86
	hGCO_mine	**87.00**	0.98		0.86

*Highest overall accuracy per dataset is in bold

that the GA performs better than blind search on the more complex landscape and both methods perform similarly on the simpler landscape.

Moreover, we compare the performance of *GCO_mine* to existing recent solutions (reported in Table 7.2.3 subsection 7.2.3) on the described benchmark datasets to further establish its relative performance. *GCO_mine* presents lower performance with four datasets, namely Pima Diabetes, Colic (SurgLesion), Lymph and Dermatology, out of 13, where the difference in OA range from 1.15% to 11.6%. The features of three of these datasets (Colic (SurgLesion), Lymph and Dermatology) are

TABLE 8.7: Classifiers Models Performance using the selected performance measures on Categorical Datasets

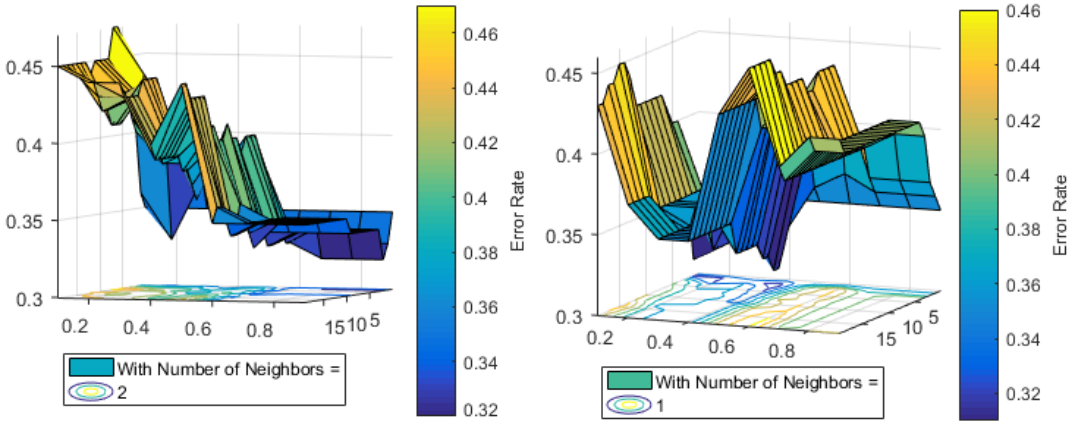
		OA	AUC	Sn_H	F_{score}
Colic #1 (Surg_Lesion)	RF	**85.32**	0.89	0.77	0.85
	MLP	81.25	0.87	0.74	0.82
	NB	77.1	0.82	0.74	0.78
	kNN	84.23	0.88	0.73	0.84
	bGCO_mine	73.61	0.81	0.72	0.74
	hGCO_mine	73.61	0.81	0.72	0.74
Colic #2 (outcome)	RF	69.94	0.83		0.67
	MLP	69.39	0.77		0.69
	NB	68.30	0.83		0.69
	kNN	69.12	0.82		0.66
	bGCO_mine	**72.50**	0.79		0.66
	hGCO_mine	71.39	0.78		0.65
Lymph	RF	83.11	0.93		0.83
	MLP	**89.96**	0.93		0.90
	NB	85.13	0.89		0.85
	kNN	69.59	0.84		0.69
	bGCO_mine	85.71	0.88		0.84
	hGCO_mine	85.71	0.88		0.84
Dermatology	RF	96.44	*1.00*		0.97
	MLP	**97.54**	1.00		*0.98*
	NB	**97.54**	1.00		*0.98*
	kNN	95.62	0.99		0.95
	bGCO_mine	91.39	0.94		0.90
	hGCO_mine	91.67	0.94		0.91
Heart Cleveland 2C	RF	77.62	0.78	0.38	0.77
	MLP	76.71	0.80	0.56	0.77
	NB	78.89	0.80	0.51	0.79
	kNN	**81.74**	0.77	0.44	0.80
	bGCO_mine	77.14	0.77	0.35	0.79
	hGCO_mine	80.48	0.79	0.44	0.80
Heart Cleveland All Classes	RF	**58.41**	0.81		0.55
	MLP	54.78	0.75		0.54
	NB	56.43	0.81		0.56
	kNN	54.45	0.78		0.50
	bGCO_mine	54.22	0.77		0.50
	hGCO_mine	55.00	0.79		0.51

*Highest overall accuracy per dataset is in bold

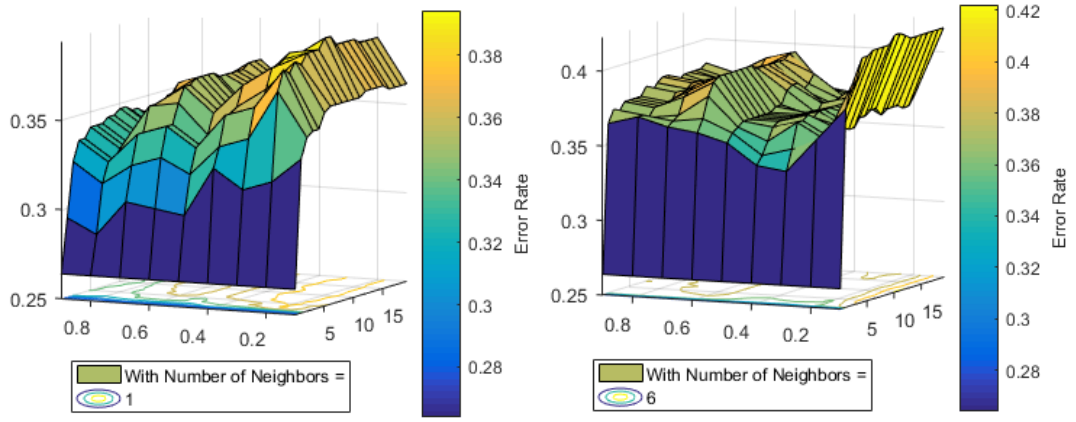
categorical, which explains the lagging performance. For Pima Diabetes, *GCO_mine* achieves lower accuracy by 1.15% in OA. As previously explained, Seera et al. [167] developed a multistage classifier which is expected to perform better than a single classifier as the case with *GCO_mine*. For Heart Cleveland All Cases, there is a tie with Kumar et al's method [121]. For the eight remaining datasets, *GCO_mine* presents higher classification accuracy with OA differences ranging from 0.42% to 16%.

8.2.4 Statistical Significance Analysis

Friedman test is applied to assess whether there is significant statistical difference between the classifiers performance. The continuous results of all the experiments are input to the Friedman test. The significance results are shown in Table 8.8. For



(A) Breast Tissue dataset



(B) Colic #1 dataset

FIGURE 8.4: Error surface plots against δ and S_r for the number of neighbours given by blind search (left) and GA (right)

each of the measures OA, AUC, TP_H and F_{score}, the actual values are input to the Friedman test. The significance values ($p-values$), together with the rankings of the compared methods are provided as output of the Friedman test. The significance threshold θ for the $p-value$ is set to 0.05.

For OA and AUC, the $p-value$ shows significant difference with the illustrated mean ranks, where *hGCO_mine* presents the highest rank in terms of OA and the second best for AUC. On the other hand, Sn_H and F_{score} manifest no significant difference given Friedman test results.

TABLE 8.8: Friedman Test Significance Results on Classification Algorithms Performance

	$p-value$	Algorithms Rankings					
		RF	MLP	NB	kNN	*bGCO_mine*	*hGCO_mine*
OA	1.4×10^{-2}	3.57	3.21	2.60	2.82	*3.86*	**4.93**
AUC	1.3×10^{-2}	**4.46**	3.25	3.61	2.21	3.21	*4.25*
Sn_H	> 0.05	–	–	–	–	–	–
F_{score}	> 0.05	–	–	–	–	–	–

8.3 Summary

In this Chapter, a partially lazy classification method has been proposed to accurately predict cardiovascular risk level from RVA-based measures and provide an understandable model. *GCO_mine* has been created to accommodate continuous data collection and produce accurate instance based classification, as needed in this context.

GCO_mine merges the concepts of decision locality and global optimisation and adds the effect of direct class membership to the traditional lazy formulation. Thus, it handles the presence of boundary cases and noise better than the traditional lazy alternative kNN. It also presents a transparent model that can be understood by medical experts. Indeed, compared to the kNN lazy classifier and the RF, MLP and NB well established eager classifiers, *GCO_mine* together with *heuristic* parameter setting (*hGCO_mine*) presents the highest accuracy on RVA data. Furthermore, *GCO_mine*'s general utility is demonstrated on 13 benchmark medical datasets from the UCI ML Repository. *GCO_mine* manifests superior performance relative to NB on 9 out of 13 datasets, while showing similar results to MLP and RF.

GCO_mine not only accurately predicts cardiovascular risk level based on RVA data, it also offers a competitive solution to a broad range of medical classification problems, with the additional intrinsic advantage of applicability to newly collected samples.

Chapter 9

Bringing it all together: The OFSET_*mine* Integrated Framework

In chapters 6, 7 and 8, the proposed methods were presented and evaluated as individual independent algorithms, with their associated impact on classification assessed. In this Chapter, the proposed integrated framework is described, first as an integration of FiltADASYN and DD_Rank into OFSET, then followed by the addition of *GCO_mine* into the complete OFSET_*mine*. We focus on OFSET, the hybrid solution of oversampling and feature selection, in particular, before the addition of *GCO_mine* as the characteristics of high dimensionality and imbalance often coexist in many real life problems. Hence, the compound effect of OFSET needs to be evaluated separately to establish its potential applicability as a stand alone solution.

The frameworks' performance on RVA data and selected medical benchmark datasets is demonstrated. OFSET and OFSET_*mine* are compared to other developed composite machine learning solutions which include oversampling and feature selection stages. The comparison is conducted as such, to avoid any bias in comparing two solutions with different processing stages. All the composite solutions rely on FiltADASYN as the oversampling method and the variability is in the feature selection methods.

9.1 On RVA data

For RVA-based measures, detailed experiments are conducted to reveal the effectiveness of the proposed approaches (OFSET and OFSET_*mine*) with RVA data compared to existing techniques. The performance of OFSET and OFSET_*mine* is evaluated in two cases. The first (a) when producing synthetic samples for all the classes to maintain the original imbalance ratio (base rate), while the second case (b) is when the oversampling is applied only on the minority classes producing fully balanced classes. Further verification of the methods performance is done by applying the resultant models on the original data.

TABLE 9.1: Comparison of OFSET Hybrid solution (FiltADASYN Oversampling and DD_Rank Feature SElecTion) Performance to other Hybrid Solutions on Base Rate Oversampled RVA data

	ReliefF Features + FiltADASYN						CorrCoeff Features + FiltADASYN					
	OA	AUC	Sn_H	Sn_M	F_{score}	$\#F$	OA	AUC	Sn_H	Sn_M	F_{score}	$\#F$
RF	96.54	0.99	0.51	0.73	0.96	84	**96.61**	0.98	0.59	0.67	0.97	72
MLP	**99.17**	0.97	0.89	0.95	0.99	41	98.96	0.98	0.88	0.97	0.99	80
NB	**91.98**	0.83	0.37	0.41	0.92	22	91.22	0.80	0.36	0.53	0.91	12

	DD_Rank Features + FiltADASYN					
	OA	AUC	Sn_H	Sn_M	F_{score}	$\#F$
RF	**96.61**	0.98	0.54	0.71	0.97	98
MLP	99.03	0.97	0.88	0.95	0.99	83
NB	91.63	0.64	0.12	0.26	0.89	5

*Highest overall accuracy per classifier is in bold

9.1.1 OFSET

The compound effect of OFSET on the RVA-based measures is first investigated using the oversampling case where the base rate is maintained. Table 9.1 compares various hybrid solutions performance to OFSET , when applied on the base rate post-oversampling RVA-based measures. The results show high classification accuracy, with potential of further improvement in terms of sensitivity Sn, especially with RF and NB. The highest recorded Sn_H and Sn_M with RF are 0.59 and 0.73 respectively. As for NB the highest value of Sn_H is 0.37 and Sn_M is 0.53. This observation further illustrates that increasing the sample size alone, may not resolve the problem of the minority classes due to high skewness. Therefore, we continue our detailed experiments on OFSET performance with the fully balanced option.

Detailed experiments on different aspects using the fully balanced dataset are conducted: the effect of bin number n_b used at the binarisation step (Algorithm 3 in section 7.1) on DD_Rank performance, the features ranking relative to the bin distribution. Also, the significance analysis in case of fully balanced dataset. However, when verifying the constructed models with the aid of OFSET on the original (real) data, we reapply the models from both the fully balanced and the base rate oversampled reduced feature set. This is done as it is critical to verify whether keeping the base rate would lead to better results when applying the resultant models on the original data.

In DD_Rank feature selection, the number of bins n_b is an important parameter. Therefore, before proceeding with our experiments, the effect of n_b on DD_Rank performance needs to be investigated. For this purpose, the overall accuracy with varying number of bins applying RF, MLP and NB on the POST FiltADASYN RVA measures is illustrated in Figure 9.1. DD_Rank shows a relatively stable performance with limited fluctuations with MLP and NB, while it depicts constant performance with RF. The highest accuracies are attained with MLP at n_b= 10. Hence, n_b is set to 10 with all classifiers as it provides a reasonable compromise between performance and complexity.

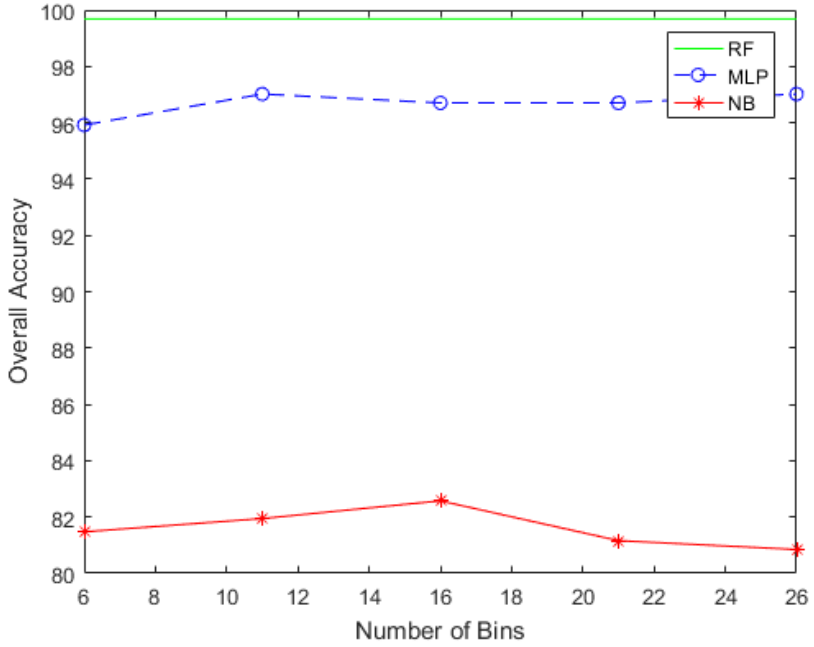
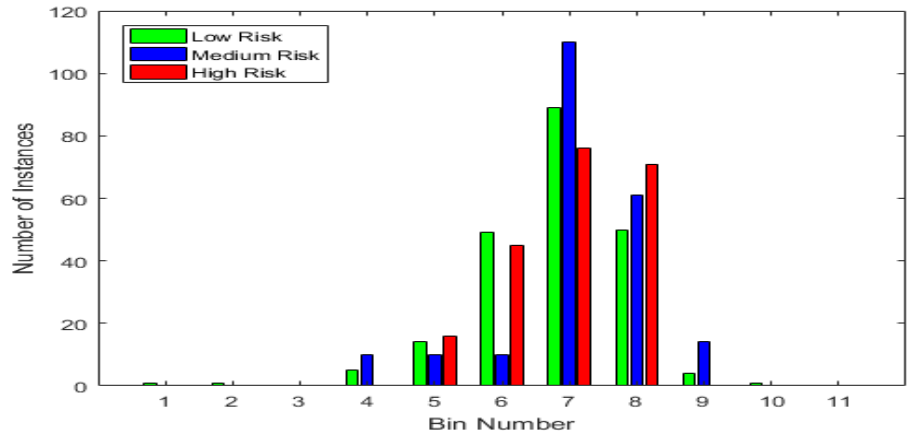


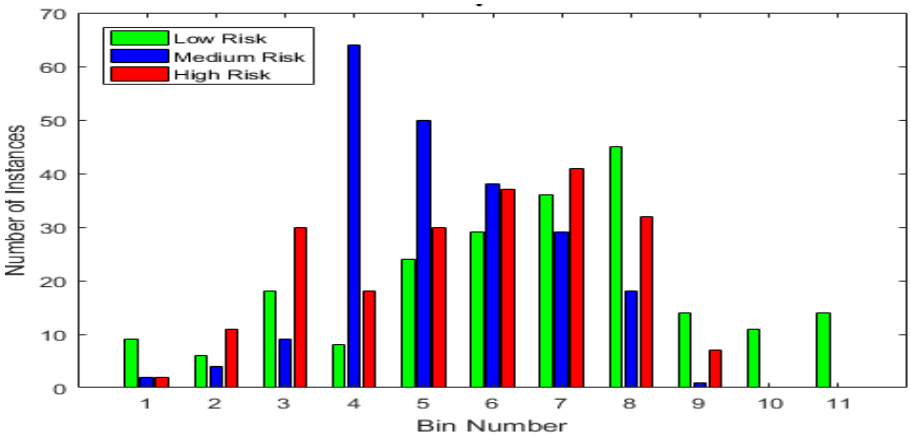
FIGURE 9.1: The Effect of the Number of Bins n_b on Performance (Overall Accuracy) for RVA data

Using Post FiltADASYN RVA data and DD_Rank with n_b set to 10, the RVA data measures are discretised and binned accordingly as was shown in Figure 7.1, resulting in a different bin distribution for each feature. Then, a ranked list of the measures is produced by DD_Rank. In an attempt to picture the effect of the ranking criteria on the produced list, the bins frequency distribution of three sample features: $MC_A Avg$ ranked first, $tMDMC_V F3$ ranked second and $DownSlope_V Avg$ ranked 104th is illustrated in Figure 9.2. Their relative distributions and rankings clarify an aspect of the produced order. The produced ranking indicates the interacting effect of reconstruction error (*re*) and precision-recall value (*prc*) on the calculated selection score (*sc*) (Algorithm 3 in section 7.1). Although, $DownSlope_V Avg$ Figure 9.2(c) is expected to have low *re*, it is ranked last due to its poor differentiating power between classes. On the other hand, $MC_A Avg$ and $tMDMC_V F3$ which are high ranked manifest relatively variable bin distribution per class (differentiating features) while all classes co-exist with comparable counts in most bins (relatively stable). In fact, the distributions shown in Figure 9.2 can not alone account for the produced rankings, which is attributed to the features synergy captured by the stacked RBMs.

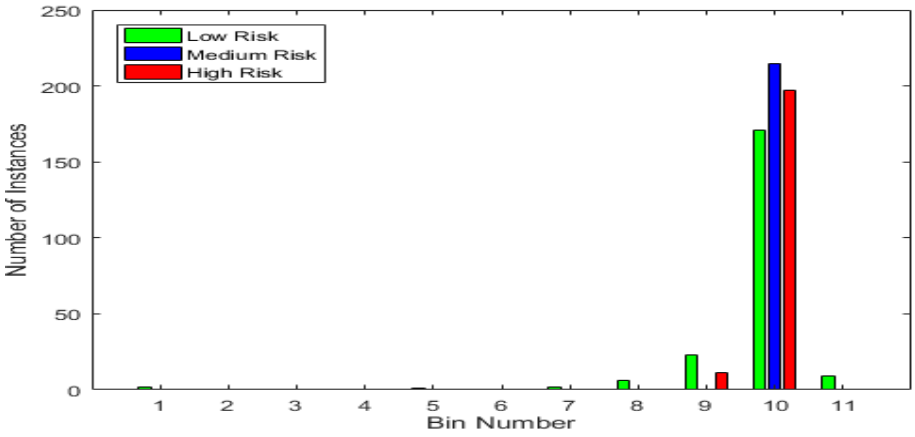
After ranking the given measures using ReliefF, CorrCoeff and DD_Rank methods, the ranked lists are input into a linear search procedure. The feature subset that marks the highest classification accuracy is selected. Table 9.2 demonstrates the performance of the feature subsets of ReliefF, CorrCoeff and DD_Rank methods. DD_Rank when applied on POST FiltADASYN RVA-based measures exhibits the best performance with RF and MLP, while comes second with NB. An overall accuracy of 99.68% with Sn_H of 1.0 using 75 features is reached. Also, the addition



(A) Ranked First: "$MC_A Avg$" Feature



(B) Ranked Second: "$tMDMC_V F3$" Feature



(C) Ranked 104th: "$Downslope_V Avg$" Feature

FIGURE 9.2: The Bin Distribution per Class of Sample Features that are fully balanced and ranked by DD_Rank

TABLE 9.2: Comparison of OFSET Hybrid solution (FiltADASYN Oversampling and DD_Rank Feature SElecTion) Performance to other Hybrid Solutions on fully balanced RVA data

	ReliefF Features						CorrCoeff Features					
	OA	AUC	Sn_H	Sn_M	F_{score}	$\#F$	OA	AUC	Sn_H	Sn_M	F_{score}	$\#F$
RF	99.52	0.996	1.0	0.99	0.996	58	99.52	0.996	1.0	0.99	0.995	93
MLP	95.76	0.98	1.0	0.94	0.96	12	96.86	0.98	1.0	0.95	0.97	39
NB	**85.40**	0.95	0.91	0.95	0.85	67	79.90	0.93	0.82	0.93	0.80	66

	DD_Rank Features					
	OA	AUC	Sn_H	Sn_M	F_{score}	$\#F$
RF	**99.68**	0.996	1.0	0.995	0.997	75
MLP	**97.02**	0.996	0.995	0.97	0.97	19
NB	81.94	0.94	0.92	0.95	0.82	64

*Highest overall accuracy per classifier is in bold

of DD_Rank after FiltADASYN oversampling helped improve Sn_M, small percentages increase of 0.007%, 0.05% and 0.03% is achieved with RF, MLP and NB respectively. Although, the percentages increase seem negligible, they show reasonable improvement when expressed in absolute terms as 1.48, 10.55 and 6.3 participants. It is to be argued that DD_Rank tend to produce larger sized feature subsets than its counterparts specifically ReliefF. To examine this point, the overall accuracy with the utilised classifiers is plotted against subset sizes in Figure 9.3. In order to attain the same accuracy of 99.52% with Random Forest as ReliefF and Corrcoeff, 60 features are needed with DD_Rank (two features more than ReliefF and 33 features less than CorrCoeff). As for MLP, a 1.26% accuracy improvement is attained at the expense of a moderate (7 features; 6.7% of total features count) increase in subset size. Also, to be noted that RF scores a Sn_H of 1.0 and OA of 98.59% at 29 features which is a moderate subset size. Hence, it can be debated that the choice of the method and subset size cutoff point depends on the application requirement; whether a discriminative (higher accuracy) or a simple (low feature number) model is needed. In cardiovascular risk estimation, higher prediction accuracy is required given that a tractable model is constructed.

To clarify the prospect of RVA measures in CVD risk prediction, the classifications based on RVA measures (applying OFSET) are compared to those based on the well-known QRisk and FRS Measures oversampled by FiltADASYN previously reported in Table 8.5 (Chapter 8). RVA measures score higher accuracies with RF+OFSET and NB+OFSET, namely 99.68% and 81.94%, while QRisk measures lead to a higher accuracy of 97.95% compared to RVA measures accuracy of 97.02% with MLP. The results indicate the potential of RVA measures in discriminating risk groups and the effectiveness of the hybrid approach in predicting cardiovascular risk with high accuracy when applied to RVA data.

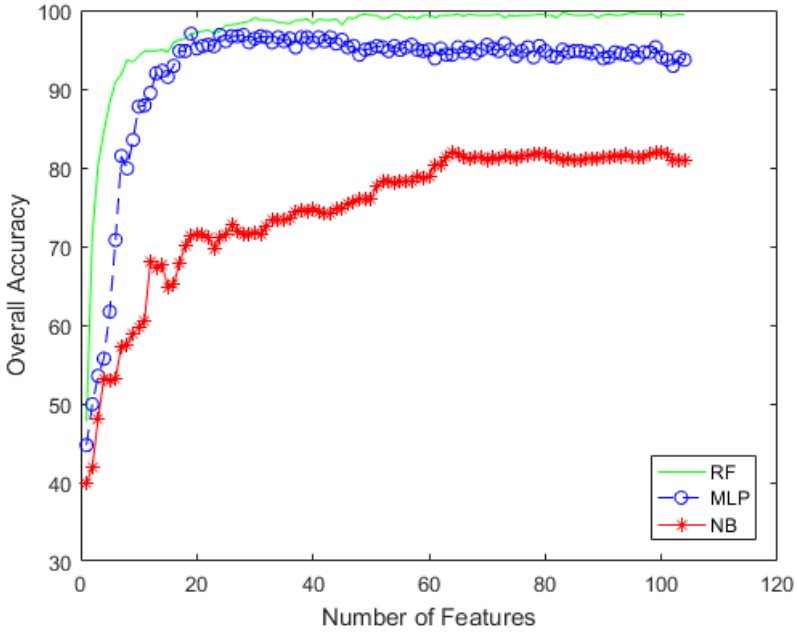


FIGURE 9.3: Classifiers Accuracies obtained with different number of features given by DD_Rank

9.1.2 OFSET_*mine*

The complete framework OFSET_*mine* is applied on (a) base rate oversampled and (b) fully balanced RVA data to predict cardiovascular risk category. The results of the two oversampling scenarios (a) and (b) are shown in Tables 9.3 and 9.4. The best results attained by other solutions (previously reported in Tables 9.2 and 9.1) are also summarised in the tables to allow for handy comparison. Also, the performance of kNN as an example of a well established lazy learner is illustrated, to show the considerable improvement achieved by *hGCO_mine* over a traditional instance-based method.

In Table 9.3 where the performance with the base rate scenario is illustrated, OFSET_*mine* achieves the best performance (highest OA and Sn_H) at 62 features. This was achieved at the expense of higher low risk mis-classifications compared to MLP. The results of OFSET_*mine* and MLP + ReliefF are relatively good and comparable. However, no solution managed to achieve the targeted sensitivity values of 1.

As for Table 9.4, *hGCO_mine* with OFSET (OFSET_*mine*) scores the second best accuracy after RF + OFSET with a minute difference of 0.16%, while it offers a 14.6% reduction in feature set size. This is in addition to its advantages as an instance-based lazy classifier. Moreover, OFSET_*mine* is shown to provide substantial performance improvement over NB and kNN based solutions. Both solutions RF + OFSET and OFSET_*mine* present the targeted sensitivity with Sn_H of 1 and Sn_M of 0.99. The results, also, show the particular suitability of OFSET (DD_Rank +FiltADASYN) to the RVA-based measures as it has the leading performance with all classifiers except with Naive Bayes.

In order to summarise the performance improvement attained by applying OFSET and OFSET_*mine* solutions to the fully balanced RVA data, the percentage improvement over the classifiers baseline performance is shown. The improvement in terms of feature set size reduction and accuracy increase is illustrated in Figure 9.4. The positive effect of OFSET is demonstrated by an accuracy increase of at least 5.37% (obtained with kNN) and a set size reduction of at least 27.8% (obtained with RF). OFSET_*mine* offers a 9.41% accuracy improvement reaching the second best accuracy while reducing the feature set by 38.5%.

9.1.3 Generated Models Verification on Original Real Samples *Only*

After applying the OFSET and OFSET_*mine* solutions to our (a) base rate oversampled and (b) fully balanced RVA data, the produced models are further verified on the original dataset with real samples only. The produced models are generated using the DD_Rank selected features on the oversampled (real + synthetic) RVA dataset. The saved RF, MLP and NB models are reapplied on the real non-synthesised dataset with the reduced feature set. While kNN and *GCO_mine* classifiers are applied to generate new classifications on the reduced real dataset while using the features previously selected by OFSET (DD_Rank and FiltADASYN).

Base Rate Oversampling

The OFSET based models constructed with RF, MLP and NB using 98, 83 and 5 features respectively (as previously shown in Table 9.1) are reapplied on the original

TABLE 9.3: Briefing of the *Best Performing* Hybrid Solutions (full results previously reported in Table 9.1) on the base rate oversampled RVA data compared to OFSET_*mine* Performance

	OA	AUC	Sn_H	Sn_M	F_{score}	$\#F$
RF + Corr	96.61	0.98	0.59	0.67	0.97	72
MLP + ReliefF	99.17	0.97	0.89	0.95	0.99	41
NB + ReliefF	91.98	0.83	0.37	0.41	0.92	22
kNN + OFSET	97.58	0.92	0.85	0.85	0.98	43
OFSET_*mine*	**99.58**	0.99	0.98	0.95	0.99	62

*Highest overall accuracy is in bold

TABLE 9.4: Briefing of the *Best Performing* Hybrid Solutions (full results previously reported in Table 9.2) on the fully balanced RVA data compared to OFSET_*mine* Performance

	OA	AUC	Sn_H	Sn_M	F_{score}	$\#F$
RF + OFSET	**99.68**	0.996	1.0	0.995	0.997	75
MLP + OFSET	97.02	0.996	0.995	0.97	0.97	19
NB + ReliefF	85.40	0.95	0.91	0.95	0.85	67
kNN + OFSET	90.11	0.93	0.89	0.92	0.90	20
OFSET_*mine*	99.52	0.996	1.0	0.99	0.996	64

*Highest overall accuracy is in bold

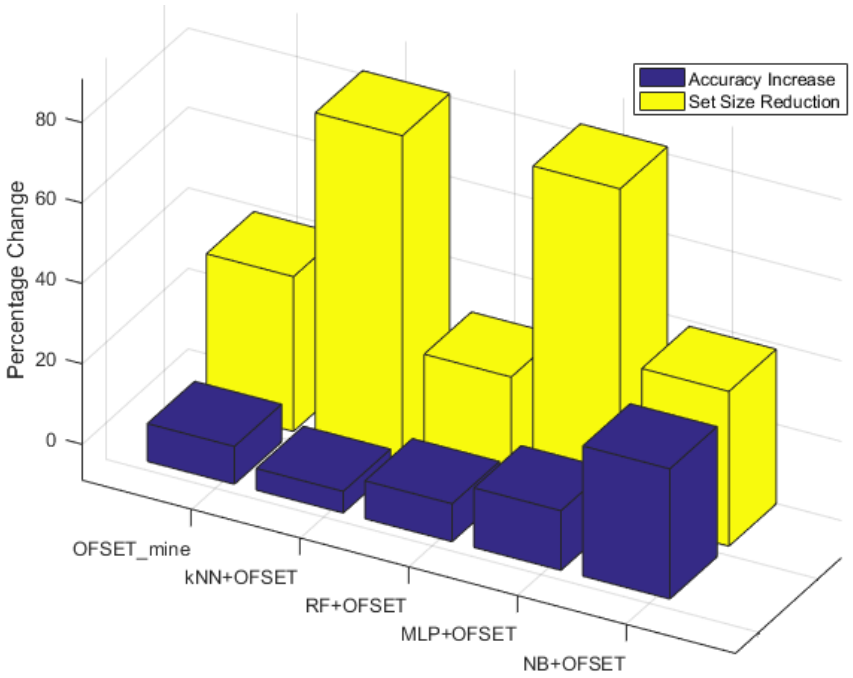


FIGURE 9.4: Percentage Improvement achieved by OFSET and OFSET_mine over Baseline Performance of the classifiers

(real) samples. Sample results from Weka are shown in Figures 9.5, 9.6, 9.7 and 9.8. On the overall, the sample results reveal a substantial improvement in performance when applying the base rate OFSET-based models on the original real data. With RF and MLP, the sensitivity values of Sn_H and Sn_M reaches (1 and 0.93) and (0.9 and 0.857) respectively. On the other hand, NB shows no significant improvement in sensitivity with $Sn_H = 0$ and $Sn_M = 0.1$. Also, a negligible improvement in sensitivity is attained using the **OFSET-based** selected features from the oversampled set with kNN as shown in Figure 9.8 over Baseline.

The performance of *GCO_mine* when applied on the original real set using the **OFSET-based** features selected from the base rate oversampled data set is better compared to kNN and NB. The results obtained given our evaluation metrics: OA, AUC, Sn_H, Sn_M and F_{score} are 89.26, 0.78, 0.6, 0.5 and 0.89. The results show lower improvement attained by *GCO_mine* and kNN compared to the other eager methods over the Baseline. The lack of model construction on the generalised oversampled datasets can explain this finding.

Full Balance Oversampling

Sample results obtained from Weka when reapplying the fully balanced models are shown in Figures 9.9, 9.10, 9.11 and 9.12. The sample results reveal considerable improvement in terms of Sn_H and Sn_M when applying the OFSET-based models on the original real dataset. A minimum improvement in Sn_H and Sn_M is achieved with NB, while with RF values of 1 and 0.93 values for Sn_H and Sn_M respectively are scored. Despite that NB is a probability-based classifier, the performance of its

```
=== Re-evaluation on test set ===

User supplied test set
Relation:     BaseRateRFandDD_RankedOrigFeatureList
Instances:     unknown (yet). Reading incrementally
Attributes:   99

=== Summary ===

Correctly Classified Instances         232               98.3051 %
Incorrectly Classified Instances         4                1.6949 %
Kappa statistic                          0.9061
Mean absolute error                      0.026
Root mean squared error                  0.1031
Coverage of cases (0.95 level)          99.1525 %
Total Number of Instances              236

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.995    0.125    0.986      0.995    0.991      0.905    0.997     1.000     l
                 0.857    0.005    0.923      0.857    0.889      0.883    0.946     0.881     m
                 0.900    0.000    1.000      0.900    0.947      0.947    0.999     0.977     h
Weighted Avg.    0.983    0.113    0.983      0.983    0.983      0.905    0.994     0.992

=== Confusion Matrix ===

   a   b   c   <-- classified as
 211   1   0 |   a = l
   2  12   0 |   b = m
   1   0   9 |   c = h
```

FIGURE 9.5: Verification of the OFSET-based (base rate) RF model on Original RVA data

```
=== Re-evaluation on test set ===

User supplied test set
Relation:     BaseRateMLPandDD_RankedOrigFeatureList
Instances:     unknown (yet). Reading incrementally
Attributes:   84

=== Summary ===

Correctly Classified Instances         233               98.7288 %
Incorrectly Classified Instances         3                1.2712 %
Kappa statistic                          0.9282
Mean absolute error                      0.0118
Root mean squared error                  0.0845
Coverage of cases (0.95 level)          99.1525 %
Total Number of Instances              236

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.125    0.986      1.000    0.993      0.929    0.983     0.998     l
                 0.857    0.000    1.000      0.857    0.923      0.922    0.970     0.933     m
                 0.900    0.000    1.000      0.900    0.947      0.947    0.912     0.905     h
Weighted Avg.    0.987    0.112    0.987      0.987    0.987      0.929    0.979     0.990

=== Confusion Matrix ===

   a   b   c   <-- classified as
 212   0   0 |   a = l
   2  12   0 |   b = m
   1   0   9 |   c = h
```

FIGURE 9.6: Verification of the OFSET-based (base rate) MLP model on Original RVA data

```
=== Re-evaluation on test set ===

User supplied test set
Relation:     BaseRateNBandDD_RankedOrigFeatureList
Instances:     unknown (yet). Reading incrementally
Attributes:   6

=== Summary ===

Correctly Classified Instances         208               88.1356 %
Incorrectly Classified Instances        28               11.8644 %
Kappa statistic                          0.039
Mean absolute error                      0.1395
Root mean squared error                  0.2752
Coverage of cases (0.95 level)          94.0678 %
Total Number of Instances              236

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.976    0.958    0.900      0.976    0.937      0.035    0.563     0.924     l
                 0.000    0.000    0.000      0.000    0.000      0.000    0.670     0.162     m
                 0.100    0.022    0.167      0.100    0.125      0.100    0.634     0.156     h
Weighted Avg.    0.881    0.862    0.816      0.881    0.847      0.035    0.572     0.846

=== Confusion Matrix ===

   a   b   c   <-- classified as
 207   0   5 |   a = l
  14   0   0 |   b = m
   9   0   1 |   c = h
```

FIGURE 9.7: Verification of the OFSET-based (base rate) NB model on Original RVA data

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         195               82.6271 %
Incorrectly Classified Instances        41               17.3729 %
Kappa statistic                          0.0556
Mean absolute error                      0.1259
Root mean squared error                  0.3422
Relative absolute error                 97.1395 %
Root relative squared error            136.6911 %
Coverage of cases (0.95 level)          82.6271 %
Mean rel. region size (0.95 level)      33.8983 %
Total Number of Instances              236

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.906    0.875    0.901      0.906    0.904      0.031    0.534     0.904     l
                 0.071    0.068    0.063      0.071    0.067      0.004    0.528     0.070     m
                 0.200    0.022    0.286      0.200    0.235      0.211    0.612     0.100     h
Weighted Avg.    0.826    0.791    0.826      0.826    0.826      0.037    0.537     0.820

=== Confusion Matrix ===

   a   b   c   <-- classified as
 192  15   5 |   a = l
  13   1   0 |   b = m
   8   0   2 |   c = h
```

FIGURE 9.8: Verification of the DD_Rank Selected Features from the (base rate) oversampled data using kNN Classifier reapplied on Original RVA data

full balance model is noticeably better than its base rate model. Such controversial finding may indicate that the better representativeness of the full balance data have compensated for altering the priors. This shall be worth further investigation in the future to determine the optimum balancing ratio for each classifier performance. Also, a slightly higher improvement is attained using the OFSET-based selected features from the fully balanced set with kNN as shown in Figure 9.12 over base rate scenario.

GCO_mine is applied on the original real set using the **OFSET-based** features derived from the full balance set. The results obtained given our evaluation metrics: OA, AUC, Sn_H, Sn_M and F_{score} are 95.33, 0.94, 0.7, 0.71 and 0.96. In contrast to the base rate selected features, with the fully balanced scenario higher sensitivity values are scored which shows the positive impact of eliminating imbalance leading to better performance than NB and kNN. However, the same pattern of lower improvement attained by *GCO_mine* and kNN compared to RF and MLP over the Baseline exists. The increase in Sn is achieved by changing the balancing ratio indicates the importance of this parameter and the need to investigate its effect in the future as it may lead to improving the relative performance of *GCO_mine*. Also, the hyper parameters of *GCO_mine* (δ, m and S_r previously described in Chapter 8) may need to be readjusted in the future on the original real dataset to enhance its performance. These future attempts are supported by the previously reported results where *GCO_mine* showed superior performance compared to RF and MLP.

Despite this deterioration in relative performance compared to RF and MLP, *GCO_mine* offers other advantages over RF and MLP such as the ability to accommodate new data to be collected in addition to the transparency of the model.

9.2 On Benchmark Medical datasets

In this section, we demonstrate the competitive performance of the proposed solutions (OFSET and *OFSET_mine*) when applied to benchmark datasets using the fully balanced scenario only. This decision was taken as the full balance scenario attained comparatively better results and due to the fact that only a rough picture on their applicability on the benchmark datasets is needed.

9.2.1 OFSET

OFSET is applied on the benchmark datasets 1-7 previously shown in Table 6.7 in subsection 6.2.3. Feature selection algorithms (ReliefF, DD_Rank and CorrCoeff) are applied on the resulting FiltADASYN fully balanced datasets.

The results are shown in Table 9.5. The fields marked as "FiltADASYN performance" indicate that the applied feature selection algorithms failed at reducing the feature set size and improving performance over FiltADASYN. It can be noted that the effectiveness of all the feature selection approaches degraded when increasing

```
=== Re-evaluation on test set ===

User supplied test set
Relation:     RFandDD_RankedOrigFeatureList
Instances:     unknown (yet). Reading incrementally
Attributes:   76

=== Summary ===

Correctly Classified Instances         234               99.1525 %
Incorrectly Classified Instances         2                0.8475 %
Kappa statistic                          0.9549
Mean absolute error                      0.0551
Root mean squared error                  0.0852
Coverage of cases (0.95 level)         100      %
Total Number of Instances              236

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.995    0.042    0.995      0.995    0.995      0.954    1.000     1.000     l
                 0.929    0.005    0.929      0.929    0.929      0.924    0.999     0.986     m
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     h
Weighted Avg.    0.992    0.038    0.992      0.992    0.992      0.954    1.000     0.999

=== Confusion Matrix ===

   a   b   c   <-- classified as
 211   1   0 |   a = l
   1  13   0 |   b = m
   0   0  10 |   c = h
```

FIGURE 9.9: Verification of the OFSET-based (fully balanced) RF model on Original RVA data

```
=== Re-evaluation on test set ===

User supplied test set
Relation:     MLPandDD_RankedOrigFeatureList
Instances:     unknown (yet). Reading incrementally
Attributes:   20

=== Summary ===

Correctly Classified Instances         228               96.6102 %
Incorrectly Classified Instances         8                3.3898 %
Kappa statistic                          0.8296
Mean absolute error                      0.0301
Root mean squared error                  0.1444
Coverage of cases (0.95 level)          97.4576 %
Total Number of Instances              236

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.986    0.000    1.000      0.986    0.993      0.936    0.995     0.999     l
                 0.643    0.000    1.000      0.643    0.783      0.793    0.847     0.715     m
                 1.000    0.035    0.556      1.000    0.714      0.732    0.988     0.707     h
Weighted Avg.    0.966    0.001    0.981      0.966    0.969      0.919    0.986     0.970

=== Confusion Matrix ===

   a   b   c   <-- classified as
 209   0   3 |   a = l
   0   9   5 |   b = m
   0   0  10 |   c = h
```

FIGURE 9.10: Verification of the OFSET-based (fully balanced) MLP model on Original RVA data

```
=== Re-evaluation on test set ===

User supplied test set
Relation:     NBandDD_RankedOrigFeatureList
Instances:     unknown (yet). Reading incrementally
Attributes:   65

=== Summary ===

Correctly Classified Instances         164               69.4915 %
Incorrectly Classified Instances        72               30.5085 %
Kappa statistic                          0.2105
Mean absolute error                      0.2047
Root mean squared error                  0.4383
Coverage of cases (0.95 level)          74.1525 %
Total Number of Instances              236

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.712    0.250    0.962      0.712    0.818      0.296    0.771     0.965     l
                 0.571    0.077    0.320      0.571    0.410      0.380    0.753     0.370     m
                 0.500    0.217    0.093      0.500    0.156      0.136    0.671     0.099     h
Weighted Avg.    0.695    0.238    0.887      0.695    0.766      0.294    0.766     0.893

=== Confusion Matrix ===

   a   b   c   <-- classified as
 151  17  44 |   a = l
   1   8   5 |   b = m
   5   0   5 |   c = h
```

FIGURE 9.11: Verification of the OFSET-based (fully balanced) NB model on Original RVA data

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         199               84.322  %
Incorrectly Classified Instances        37               15.678  %
Kappa statistic                          0.1801
Mean absolute error                      0.1106
Root mean squared error                  0.3223
Relative absolute error                 85.3248 %
Root relative squared error            128.7308 %
Coverage of cases (0.95 level)          84.322  %
Mean rel. region size (0.95 level)      33.4746 %
Total Number of Instances              236

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.915    0.708    0.919      0.915    0.917      0.203    0.606     0.919     l
                 0.071    0.068    0.063      0.071    0.067      0.004    0.505     0.061     m
                 0.400    0.022    0.444      0.400    0.421      0.397    0.662     0.186     h
Weighted Avg.    0.843    0.641    0.848      0.843    0.846      0.199    0.602     0.837

=== Confusion Matrix ===

   a   b   c   <-- classified as
 194  14   4 |   a = l
  12   1   1 |   b = m
   5   1   4 |   c = h
```

FIGURE 9.12: Verification of the performance of the DD_Rank Selected Features from the (fully balanced) oversampled data using kNN Classifier reapplied on Original RVA data

the instances to features ratio (caused by oversampling). DD_Rank position with FiltADASYN subsides to the hybrid solution of ReliefF and FiltADASYN where ReliefF-based models outperform their counterparts in 10 out of 21 and achieve the second rank in 6 out of 21. OFSET manifests superior performance in 8 out of 21 cases and comes second in 6 cases. CorrCoeff maintains the third rank with three wins only. Overall, OFSET's performance is competitive with the ReliefF hybrid solution and better than the CorrCoeff-based solution.

To compare the selection methods in terms of the number of features selected after oversampling, we illustrate the φ values (previously defined as the number of features selected over the number of features in the full feature set) for all methods and the seven oversampled datasets in a 3D scatter plot in Figure 9.13. A pattern different from Figure 7.3 (illustrated in subsection 7.2.3) is observed, where most of φ values for all methods are clustered in the upper inner corner of the space, which means that they tend to choose high proportions of the feature set. This indicates the prospective scalability of DD_Rank for larger sizes as it is maintaining the same pattern, whereas its counterparts are not. It is also noticeable that with MLP, all methods manifest high φ above 0.6 as shown by the limits of the x-axis.

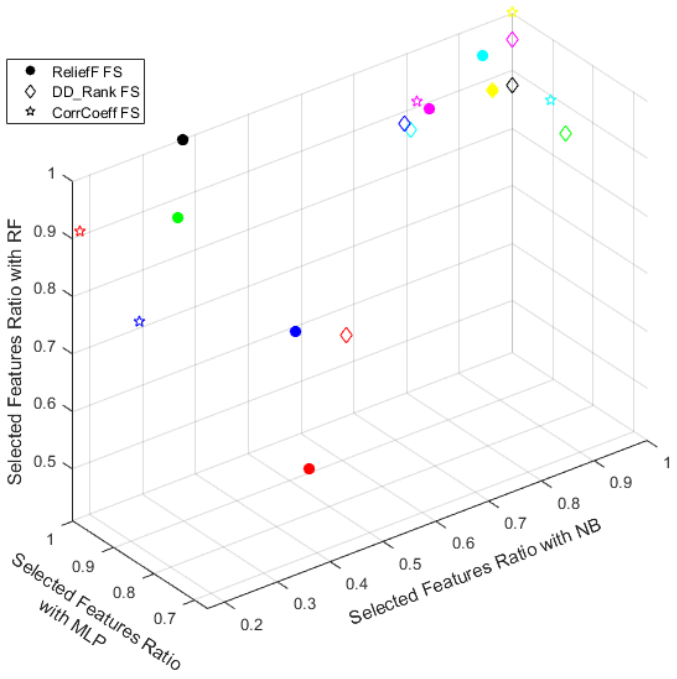


FIGURE 9.13: Selected Features Proportions φ per dataset with Hybrid Approaches (Different Feature Selection methods + FiltADASYN)

9.2.2 OFSET_*mine*

To evaluate the applicability of OFSET_*mine* as a general purpose solution, its performance on the same seven datasets (which were candidates for oversampling) is

TABLE 9.5: Hybrid Solutions Performance on Oversampled Benchmark Datasets comparing ReliefF + FiltADASYN, OFSET and CorrCoeff + FiltADASYN using RF, MLP and NB classifiers

ID#	Dataset		ReliefF + FiltADASYN					OFSET					CorrCoeff + FiltADASYN				
			OA	*AUC*	Sn_H	F_{score}	#F	*OA*	*AUC*	Sn_H	F_{score}	#F	*OA*	*AUC*	Sn_H	F_{score}	#F
1	Pima Diabetes	RF	FiltADASYN Performance					**81.76**	**0.89**	**0.86**	**0.82**	**7**	FiltADASYN Performance				
		MLP	FiltADASYN Performance					FiltADASYN Performance					FiltADASYN Performance				
		NB	**71.39**	**0.81**	**0.67**	**0.73**	**3**	FiltADASYN Performance					**71.39**	**0.81**	**0.67**	**0.73**	**3**
2	Colic (outcome)	RF	82.48	0.95		0.83	21	*82.78*	0.95		0.83	21	**82.92**	0.95		0.83	**20**
		MLP	**80.63**	**0.89**		**0.81**	**20**	FiltADASYN Performance					FiltADASYN Performance				
		NB	**73.31**	**0.87**		0.73	**17**	FiltADASYN Performance					73.02	**0.87**		**0.74**	18
3	Lymph	RF	**94.17**	0.98		**0.94**	17	**94.17**	0.98		**0.94**	17	*93.20*	0.98		0.93	17
		MLP	FiltADASYN Performance					**92.71**	**0.97**		**0.93**	16	*91.74*	**0.97**		0.92	**15**
		NB	*87.86*	0.95		0.88	17	**88.84**	0.95		**0.89**	**13**	*87.86*	0.95		0.88	17
4	Parkinson	RF	**96.64**	0.99	0.97	0.96	**9**	*95.97*	0.99	**0.98**	0.96	15	95.97	0.99	**0.98**	0.96	20
		MLP	**97.65**	0.98	0.99	**0.98**	20	*96.97*	0.98	0.99	0.97	**18**	FiltADASYN Performance				
		NB	*82.89*	0.88	0.93	0.83	12	80.87	0.85	**0.97**	0.81	12	**83.22**	**0.90**	0.93	**0.84**	**4**
5	Heart Cleveland 2C	RF	*88.19*	**0.96**	**0.93**	0.88	**12**	**88.50**	**0.96**	0.92	**0.89**	**12**	FiltADASYN Performance				
		MLP	**86.38**	**0.89**	**0.94**	**0.87**	12	*84.47*	0.87	0.87	0.85	**10**	FiltADASYN Performance				
		NB	*72.98*	0.77	0.68	0.73	**4**	**74.53**	0.80	**0.72**	0.75	12	72.05	**0.81**	0.64	**0.81**	10
6	Verteb Column 2C	RF	**89.61**	**0.96**	**0.94**	**0.89**	**5**	FiltADASYN Performance					*89.37*	**0.96**	0.93	**0.89**	**5**
		MLP	*86.37*	0.92	0.93	0.86	**4**	**87.29**	0.92	**0.94**	**0.87**	5	*85.68*	0.92	**0.94**	0.86	5
		NB	**84.29**	**0.88**	0.95	**0.85**	2	79.21	0.83	0.88	0.80	4	*82.67*	0.86	**0.96**	0.83	**1**
7	Verteb Column 3C	RF	FiltADASYN					FiltADASYN					FiltADASYN Performance				
		MLP	**87.05**	**0.95**		**0.87**	**5**	*86.39*	**0.95**		**0.87**	**5**					
		NB	*82.81*	**0.95**		**0.87**	5	**84.15**	**0.95**		0.84	**5**					

depicted in Table 9.6 together with a briefing on the best performing hybrid solutions. The results demonstrate the competitive performance of OFSET_*mine* where it surpasses NB solutions with all datasets and overcomes kNN solutions in 4 out of 7 datasets at a minimum accuracy difference of 0.39%. It also manifests similar performance to RF- and ML- based composite solutions in 5 out of 7 datasets.

On the other hand, there are cases such as Pima Diabetes and Heart Cleveland 2C where a considerable performance gap is present between OFSET_*mine* and its counterparts. Since *hGCO_mine* adopts the simple approach of kNN median imputation for handling missing values, this may explain such performance gap for oversampled Pima Diabetes dataset as it has high percentage of missing values (approximately 9.2%). Also, the use of standardised euclidean distance in energy computation of *hGCO_mine* can account for the performance difference with categorical Heart Cleveland 2C dataset.

9.3 Statistical Significance Analysis

The results attained with the **fully balanced** datasets only are used to determine whether OFSET and OFSET_*mine* provide a recognisable improvement over the other solutions, this is done as the majority of the experiments are done using the full balance scenario. Friedman significance test is applied for this purpose. The significance results for OFSET and OFSET_*mine* are presented in Tables 9.7 and 9.8 respectively.

9.3.1 OFSET

In Table 9.7, for each of the actual values of the evaluation metrics of OA, AUC, Sn_H and F_{score}, the significance values ($p - values$), together with the rankings of the baseline (Bl) and the compared methods are provided. In each case, first, an overview of the overall performance is presented, by collating the results across all three classifiers. Then, the performance of each method with each classifier is detailed separately. The significance threshold θ for the $p - value$ is set to 0.05. While negligible improvement is attained with NB in terms of OA and AUC, significant improvement is achieved in the case of RF and MLP. In particular, OFSET with RF scores the best mean rank using the OA, Sn_H and F_{score} evaluation measures. On the whole, the significance results show that OFSET provides a viable solution to the problems of imbalance and relatively high dimensionality.

In summary, the Friedman test results confirm the existence of significant difference in performance between the baseline and the tested Hybrid solutions. The proposed methods when applied with the classifiers RF, MLP and NB are competitive with existing ReliefF-based and Corr-based hybrid solutions as shown by the output rankings.

A notable finding is that when comparing the different integrated methods to each other (omitting Baseline), no statistical significance is found except for OA

TABLE 9.6: OFSET_*mine* Performance compared to the other *Best Performing* Hybrid Solutions on Benchmark Datasets

		OA	AUC	Sn_H	F_{score}	$\#F$
Pima Diabetes	RF + OFSET	81.76	0.89	0.86	0.82	7
	MLP	FiltADASYN Performance				
	NB + CorrCoeff	71.39	0.81	0.67	0.73	3
	kNN + ReliefF	**82.32**	0.83	0.78	0.83	7
	OFSET_*mine*	74.61	0.81	0.69	0.83	7
Colic (outcome)	RF + CorrCoeff	**82.92**	0.95		0.83	20
	MLP + ReliefF	80.63	0.89		0.81	20
	NB + ReliefF	73.31	0.87		0.73	17
	kNN + ReliefF	67.28	0.77		0.68	6
	OFSET_*mine*	80.58	0.88		0.73	20
Lymph	RF + OFSET	**94.17**	0.98		0.94	17
	MLP + OFSET	92.71	0.97		0.93	16
	NB + OFSET	88.84	0.95		0.89	13
	kNN + CorrCoeff	89.32	0.94		0.89	10
	OFSET_*mine*	91.5	0.96		0.97	17
Parkinson	RF + ReliefF	96.64	0.99	0.97	0.96	9
	MLP + ReliefF	97.65	0.98	0.99	0.98	20
	NB + CorrCoeff	83.22	0.90	0.93	0.84	4
	kNN + CorrCoeff	**98.32**	0.98	0.96	0.98	19
	OFSET_*mine*	95.86	0.96	0.95	0.98	18
Heart Cleveland 2C	RF + OFSET	**88.50**	0.96	0.92	0.89	12
	MLP + ReliefF	86.38	0.89	0.94	0.87	12
	NB + OFSET	74.53	0.80	0.72	0.75	12
	kNN + CorrCoeff	87.57	0.88	0.91	0.89	11
	OFSET_*mine*	80.31	0.84	0.86	0.81	12
Verteb Column 2C	RF + ReliefF	**89.61**	0.96	0.94	0.89	5
	MLP + OFSET	87.29	0.92	0.94	0.87	5
	NB + ReliefF	84.29	0.88	0.95	0.85	2
	kNN + ReliefF	87.75	0.87	0.89	0.89	5
	OFSET_*mine*	88.14	0.94	0.94	0.86	5
Verteb Column 3C	RF	FiltADASYN				
	MLP + ReliefF	87.05	0.95		0.87	5
	NB + OFSET	84.15	0.95		0.84	5
	kNN + ReliefF	87.72	0.91		0.88	5
	OFSET_*mine*	**92.27**	0.98		0.94	5

*Highest overall accuracy per dataset is in bold

across all classifiers. The $p-value$ obtained is 0.047 and the relative ranking for ReliefF-based, DD_Rank-based and Corr-based results are 2.23, 2.15 and 1.63 respectively. Such finding shows that the hybrid solutions provide comparable performance in general, where OFSET ranks second on the overall performance.

9.3.2 OFSET_*mine*

For OFSET_*mine* significance analysis, Table 9.8 shows the results. After the actual values of OA, AUC, Sn_H and F_{score} are input to the continuous Friedman significance test, the significance values ($p-values$), together with the rankings of the integrated approaches are provided. The significance threshold θ for the $p-value$

TABLE 9.7: Friedman test Significance Results on Baseline and Hybrid Solution Performance

	$p-value$	Algorithms Rankings			
		Bl	*FiltADASYN* +*ReliefF*	*OFSET*	*FiltADASYN* +*Corr*
OA					
All	3.3×10^{-8}	1.20	3.14	3.10	2.54
RF	7.0×10^{-4}	1.00	3.06	3.25	2.69
MLP	1.1×10^{-5}	1.00	3.25	3.31	2.44
NB	> 0.05	–	–	–	–
AUC					
All	6.0×10^{-4}	1.71	2.96	2.64	2.67
RF	8.9×10^{-5}	1.00	3.00	3.00	3.00
MLP	> 0.05	–	–	–	–
NB	> 0.05	–	–	–	–
Sn_H					
All	1.4×10^{-5}	1.00	2.92	3.17	2.92
RF	2.0×10^{-2}	1.00	2.75	3.5	2.75
MLP	3.0×10^{-2}	1.00	3.00	3.00	3.00
NB	$= 0.05$	–	–	–	–
F_{score}					
All	7.54×10^{-8}	1.28	3.02	3.04	2.67
RF	1.0×10^{-4}	1.00	2.94	3.44	2.63
MLP	5.3×10^{-3}	1.31	3.06	3.12	2.50
NB	> 0.05	–	–	–	–

TABLE 9.8: Friedman Test Significance Results for OFSET_*mine* and other Hybrid Solutions Performance

	$p-value$	Algorithms Rankings				
		RF Solution	MLP Solution	NB Solution	kNN Solution	OFSET_*mine*
OA	9.0×10^{-4}	**4.50**	*3.13*	1.13	*3.13*	*3.13*
AUC	4.0×10^{-4}	**4.81**	*3.38*	1.81	1.94	3.06
Sn_H	> 0.05	–	–	–	–	–
F_{score}	4.4×10^{-3}	**4.00**	3.00	1.25	3.13	*3.63*

is set to 0.05. A significant difference exists between the solutions in terms of OA, AUC and F_{score}. For these measures, RF-based solutions present the highest ranking. When considering OA, all the remaining solutions come at the following rank except for NB solutions. While OFSET_*mine* scores the third ranking in terms of AUC, it presents the second best position after RF-based solutions for F_{score}. It is to be noted that 8 out of 28 integrated solutions competing with OFSET_*mine* include OFSET as part of their compound solution, which signifies the impact of *GCO_mine* against RF, MLP, NB and kNN and the competitiveness of OFSET.

For cardiovascular risk prediction using RVA data, the presented findings reveal that OFSET_*mine* is a high performing solution for this problem. Using base rate oversampling, it achieves the highest performance on the real + synthesised samples. While with the fully balanced data (real+synthesised), OFSET_*mine* follows RF+OFSET solution with a minute OA difference of 0.16%, while it surpasses all its counterparts with a minimum OA difference of 2.5% to MLP. However, it is to be noted that when applied on the original data it subsides to RF and MLP OFSET-based solutions. In addition, OFSET_*mine* offers a good solution to a range of real world problems, as it provides similar performance to well established methods and a significant improvement over Baseline. In summary, OFSET and OFSET_*mine* provide solutions that are on par with established methods and surpass them in some cases.

Chapter 10

Summary, Conclusion and Future Work

In this study, an integrated machine learning framework OFSET_*mine* was proposed for cardiovascular risk prediction using Retinal Vessels Analysis (RVA) data. The framework handled the characteristics of the available RVA data to overcome the challenges they impose. In this chapter, a summary and conclusion of the developed methods and the obtained results are given, followed by a brief discussion on the clinical utility of the methods . Finally, the future work directions are suggested.

10.1 Summary of the Results

Following data collection and features generation, several existing machine learning methods were applied on the RVA-based measures to produce cardiovascular risk prediction. The results, previously reported in Chapter 5, of the preliminary set of experiments conducted on the the available data suggested the potential of RVA data for cardiovascular risk prediction. In addition, the results showed that there is still room for improvement to produce more accurate predictions and better fulfill the requirements of health care specialists. Consequently, a need has risen to develop a customised solution to better handle the characteristics of RVA data.

The main properties that imposed difficulties on learning were skewed class distribution, relative high dimensionality and overlapping features ranges. Also, it would be preferable to have a classification method that can easily adapt to new samples once collected.

The RVA data characteristics together with critical nature of accurate cardiovascular risk prediction imposes the set of requirements illustrated in Chapter 4 section 4.2. An integrated framework OFSET_*mine* is devised to satisfy these requirements and effectively handles the RVA data to produce accurate cardiovascular risk predictions. OFSET_*mine* comprises two stages of preprocessing and data reduction (FiltADASYN and DD_Rank) and the classification stage of *GCO_mine*.

10.1.1 FiltADASYN oversampling

We tackle the problem of class imbalance by introducing FiltADASYN oversampling. In FiltADASYN, A **Filt**ering step is added to **ADASYN** oversampling. The aim is to ensure the representativeness of the generated samples and improve the reliability of the resultant samples of the synthesised classes. The filtering step rejects a generated sample with instances from the majority class within its neighbourhood to increase the likelihood that the synthesised samples truly belong to the generating minority class. The added step was shown to improve the consistency of the post oversampling data and promote the performance of the applied classifiers on the RVA data.

The developed FiltADASYN method performs oversampling independent of the applied classifier, uses all of the available minority samples to conserve the information provided by the available samples, preserves feature dependence and validate the generated sample after synthesis. Thus, FiltADASYN oversampling complies with the designated requirements specified in subsection 4.2.2.

10.1.2 DD_Rank Feature Selection

In order to select the most informative features and eliminate the curse of dimensionality, DD_Rank feature selection method is proposed. An existing architecture of **D**eep **D**isjunct Belief Network is used to capture the synergy between the features and **Rank** the features to perform explicit feature selection.

The ranking is based on the features stability and predictive ability. The features stability is considered as the ability to reconstruct/reproduce the same feature distribution and is measured by the reconstruction error resulting from the learning (learned weights and biases) of the Deep Restricted Boltzmann machine. We aim for low reconstruction error as this indicates that the RBMs constructed model well approximates the co-occurrences between this feature and the outcome. A feature's predictive ability is estimated using the value of the area under precision-recall curve for this feature. Some of the experiments reveal the favourable effect of DD_Rank, as it managed alone to combat the compound effect of imbalance and high dimensionality with a number of datasets. This was shown by the considerable accuracy improvement it achieved with datasets that showed imbalance and high dimension.

DD_Rank is unbound to a given classifier and combines theoretic aspect of feature relevance, which is the ability to correctly reconstruct the features distribution relative to the classes measured by the reconstruction error, with the actual predictive value of the feature, which is measured by area under precision-recall curve. In addition, it accounts for feature interaction in the hidden layers producing a ranked list of features. Therefore, DD_Rank is shown to satisfy the specified suitability criteria in subsection 4.2.3.

10.1.3 *GCO_mine*

For the characteristic of expected data expansion and overlapping risk group ranges, we develop an instance-based lazy learner that depends on the principles of local neighbourhood mining and global graph cut optimisation.

GCO_mine introduces several amendments to existing lazy learners. It adds global connectivity between samples, defines direct membership between an unlabeled sample and a class representative and it construct neighbourhoods using both labeled and unlabeled samples. The proposed *GCO_mine* presents higher performance than the traditional lazy classifiers as depicted by its comparison to kNN classifier.

GCO_mine shows high performance comparable to the performance achieved by well established classifiers namely: RF, MLP and NB when applied to the RVA data and the additional real benchmark datasets. Moreover, *GCO_mine* takes individual-based decisions for classification, which is suitable for overlapping classes. Also, it can accommodate expanding datasets and give an understandable reasoning for the prediction decision. Hence, *GCO_mine* manages to realise the design requirements previously defined in subsection 4.2.1.

10.1.4 OFSET_*mine* Performance

On RVA Data

The proposed integrated framework OFSET_*mine* has provided an effective solution that both handles the characteristics of RVA data and produces an accurate cardiovascular risk prediction. When the developed framework methods are applied on the RVA-based measures, satisfactory results are obtained in absolute terms and relative to the Baseline (No intervention). The OFSET_*mine* methods, FiltADASYN in specific, are applied creating two resultant scenarios (a) base rate and (b) fully balanced oversampling scenarios.

With the base rate scenario, OFSET_*mine* succeeded in achieving the highest OA and Sn_H when applied on the real and synthesised samples while MLP + ReliefF + FiltADASYN comes next. However, when the derived models are applied on the original real data only , the MLP + OFSET model attains the highest OA, Sn_H and Sn_M, while OFSET_*mine* relative performance ranking deteriorates reaching the third position.

When using our proposed methods with the fully balanced RVA measures, the highest classification accuracy is achieved when employing RF + OFSET, while OFSET_*mine* scores second with a minute accuracy difference. In addition, when applying the **OFSET-based** models on the real RVA data, a considerable increase over Baseline (direct application) in {OA,Sn_H and Sn_M} is attained by RF and MLP models. Also, OFSET_*mine* (when applied on the original data) comes third showing better performance compared to its performance in the base rate case.

Overall, the lack of model construction in *GCO_mine* resulted in its relative performance deterioration when applied on the original data. However, *GCO_mine* still surpassed the performance of kNN and NB classifiers. In addition, the application of OFSET (DD_Rank + FiltADASYN) succeeded in improving the performance of the classifiers derived models when they are used with the original data. Also, it can be observed that eliminating the class imbalance (full balance scenario) better aided the learning algorithms in producing more accurate classifications as was shown by the higher Sn_H and Sn_M obtained on the real data with *GCO_mine*, RF and NB.

Another aspect that was shown is the superiority of RVA measures over FRS and QRisk measures in risk group prediction accuracy, with the exception of QRisk Measures with MLP.

In conclusion, we have succeeded in proving the capability of RVA measures in differentiating between cardiovascular risk groups by applying the newly proposed OFSET hybrid approach and the OFSET_*mine* framework. According to Hlatky et al. [88], differentiating between risk groups is the first step in establishing a risk marker. Hence, this study verifies the initial proof of principle needed for establishing a new risk marker through developing a tailored machine learning framework that satisfies the requirements defined in section 4.2, which enabled the stratification of risk groups.

On Additional Medical Benchmark Data

The applicability of the OFSET_*mine* methods as general purpose solutions was demonstrated. The methods were applied on a range of medical benchmark datasets from the UCI ML Repository exhibiting similar properties of imbalance and relative high dimensionality to the collected RVA measures. The performance of the methods is illustrated individually in Chapters 6, 7 and 8 and as a compound framework in Chapter 9.

The methods have shown on par performance (as individual components and as composite solutions) to the established methods used in the comparison. Therefore, the proposed methods can be considered as competitive solutions for a broad spectrum of applications.

10.2 Evaluation of the Clinical Utility of OFSET_*mine*

OFSET_*mine* was developed for medical and clinical use for predicting cardiovascular risk using RVA data. It was also tested on benchmark medical datasets. Here, their potential clinical utility is discussed, including the merits and the limitations of each element of OFSET_*mine*.

10.2.1 The proposed *GCO_mine* Classification

GCO_mine, the classification method purpose-built for the suitability criteria defined in section 4.2, has the following properties relevant to clinical use:

Merit 1: GCO_mine is considered a suitable classification method for clinical decision making as it mimics the way a health care specialist allocates a risk level to a new participant [69].
Evidence: GCO_mine considers the degree of resemblance of a new participant to other similar participants of known risk level, as well as assesses the *differences* between the new participant and a risk group representative (reference).

Merit 2: GCO_mine can be readily applied to other medical decision making problems, where a simple individual-based decision is needed.
Evidence: GCO_mine does not include any particular assumptions on the data and does not incorporate domain specific (cardio-related) knowledge into its decision making, which leads to a generic formulation.

Limitation 1: GCO_mine does not account for the varying importance of different biological measures when it allocates a risk level to a participant. In real life, some biological measures may have higher influence (importance) on the allocated risk than others.
Evidence: GCO_mine computes similarity between participants using a distance metric of uniform weights for all features (measures), treating all the measures with equal importance. When the relative importance of features is known, this can be incorporated by allocating weights accordingly.

10.2.2 The proposed FiltADASYN Oversampling

The applicability of the proposed FiltADASYN oversampling in the clinical field is evaluated as follows:

Merit 1: FiltADASYN verifies the representativeness of the synthetic samples to eliminate samples that were generated within the majority class region.
Evidence: FiltADASYN checks the validity of the sample after synthesis through a simple filtering step, rejects the overgeneralised samples and removes them from the resultant set.

Merit 2: FiltADASYN improves accuracy in exploratory studies, where recruiting new subjects would be expensive or take a long time and the benefit of the study outcome is not yet confirmed.

Evidence: FiltADASYN synthetically generates new realistic samples and hence increase the representativeness of the sample.

Limitation 1 FiltADASYN can not be considered as a substitute for further data collection, in order to be able to produce final clinical judgments.
Evidence: FiltADASYN, similarly to other oversampling methods, has the tendency to produce optimistic results in terms of sensitivity due to the imposed similarity between the original and the generated samples and therefore induces bias as it artificially alters the prior probabilities.

10.2.3 The DD_Rank Feature Selection

The suitability of DD_Rank feature selection for clinical use is assessed below:
Merit 1: DD_Rank produces a ranking of the features, which aids the understandability of the selected subset.
Evidence: DD_Rank outputs a ranked list, based on two plausible estimates of theoretic and practical relevance of a measure (feature), and with this it gives a comprehensible explanation for the assigned importance of the features.

Limitation 1: The underlying operation of DD_Rank is considered as a black box, which could lower the willingness of clinicians to adopt such method.
Evidence: DD_Rank uses RBMs for feature selection, where a stochastic process takes place within the RBMs layers.

10.3 Directions of Future Work

Since the objective of this study was two fold namely: 1) To investigate the prospect of Retinal Vascular Function assessed by Retinal Vessel Analysis (RVA) for determining cardiovascular risk category for apparently healthy subjects; 2) To effectively process the available data using machine learning techniques to produce a reliable risk level prediction. The future work of this study shall adopt two similar directions. The first is to improve the performance of OFSET_*mine* methods, to provide a reliable solution for prediction and to widen the range of applicability of the methods. The second direction is to continue with establishing RVA as an early cardiovascular risk marker through machine learning.

10.3.1 Enhancement of OFSET_*mine* methods Performance

In order to fulfill the first direction, the impact of several modifications to the proposed approaches should be investigated in the future:

In **FiltADASYN oversampling**, a more sophisticated procedure instead of the used 'one versus all' policy is to be considered to extend the original two class approach. A procedure that preserves the distinction between classes and avoid class

merging into a single class would be preferable. Also, the rejection process can be modified such that each generated sample is assigned a rejection score weighted by its distance to the other classes samples rather than a binary decision for inclusion or rejection of the sample. Besides, a more adaptive approach shall be considered, where the detailed structure of the classes (subclusters) can be used to guide the parameters setting using different parameters k (number of neighbours to randomly select from for synthesis) and l (used to determine neighbours after oversampling to decide whether a sample should be rejected) for each different subspace of a class. In addition, a thorough investigation on the effect of different balance ratios and their relation to the classifiers performance shall be carried out.

In **DD_Rank feature selection**, there is a set of variations that can be considered to analyse the effect of different aspects on its performance. The performance of deeper architectures of DDBM is worth examination, as it is expected to boost the latent representation of the measures on the hidden nodes to even higher-level, more abstract form that leads to better generalisation. But, the added time and complexity should be adequately compensated with substantial accuracy increase. Moreover, in the current formulation of the selection score, equal contribution for reconstruction error or precision-recall curve were used in calculating the selection score, thus the effect of formulations of variable weights can be investigated. Also, measures other than reconstruction error or precision-recall curve can be experimented.

In *GCO_mine* **classification**, there are several points that are worth further examination. One of the points is the adoption of other approaches for missing values handling. An approach that better preserves the structure of the data and avoid distorting its variance is required instead of the median imputation approach adopted [15]. This is required to improve its performance with datasets with high percentage of missing values. In addition, distances functions that are better adapted to categorical data can be utilised when needed to enhance its range of applicability. Heterogeneous Value Decomposition Metric [127] is an example of a proximity metric that handles mixed data and can be applied for this purpose. Another point that is worth consideration is the development of a *GCO_mine* variant specifically designed for cardiovascular risk prediction through varying the importance of RVA-based measures (features) used in differences and similarity calculation. The incorporation of domain knowledge regarding the relative importance of the measures (features) in the weights for distance calculation would be useful.

10.3.2 Establishment of RVA as a Cardiovascular Risk Marker

All the measures generated from RVA to create a vascular profile are measures that study the response parameters of each flicker independently. Other measures that analyse the inter-relations, differences and similarities between different response cycles might be more predictive. Examples of these measures are correlation and mutual information between flicker responses, differences between two signals maxima and minima and rate of change in peak values of the three consecutive flicker

signals ($F1, F2, F3$). The generation of these measures would be supported by the findings of Heitmar et al. [82] (Chapter 2), where differences in flicker responses was used to distinguish risk groups. In addition, experimentation with transform domain features can reveal other aspects of the responses. Fourier transform [45] can be utilised to illustrate a signal's frequency components, while the signal's localised transient time and frequency characteristics is obtained through Wavelet analysis [187, 193].

Moreover, thorough validation is further required to initiate the use of RVA as a screening risk marker in medical care venues. After illustrating the capability of RVA to distinguish cardiovascular risk groups, the potential of RVA to predict actual disease development (hard outcomes) needs to be validated. This validation is a necessary step for establishing RVA as a screening risk marker in medical venues. In order to proceed with this step, prospective data needs to be collected for the current subjects, to allow follow up and determine whether the classified high risk subjects truly develop a cardiovascular related disease. In addition, new participants need to be recruited to enlarge the sample size while preserving the natural distribution of the data to allow reliable analysis of the data.

Machine learning methods have the capability to consider a greater number and complexity of interacting variables than conventional statistical methods. Hence, further application of machine learning methods customised to the data to be used in the next steps of establishing RVA as a screening risk marker is anticipated to be of high importance.

Appendix A

Genetic Algorithms

Genetic Algorithms (GAs) [19] is an adaptive heuristic search technique suitable for optimization problems. GAs are inspired by natural evolution and genetics of biological organisms. They rely on the principles of natural selection and survival of the fittest. GAs mimic the natural behaviour of the organisms, where a population of individuals exist and the survival of the best fit individual is promoted. The individuals of the population represent candidate solutions to a defined problem and a fitness value is assigned to each candidate. The fitness score assesses the individual's capability of survival and his effectiveness in providing a solution to the problem in hand. The highly fit individuals in a given generation of the population are selected and let to reproduce and/or pass to the next generation. A new offspring is produced which combine features taken from highly fit parents. Hence, over many generations the good features of the individuals are allowed to evolve and promising areas of the search space are explored. while the least fit members of the population are left to die out and their associated search spaces ignored.

GAs may converge to an optimal solution if appropriately formulated and designed, but no guarantees for global optimality are provided. On the other hand, GAs are robust and provide a near optimal solution to a range of problems where specialised solutions are hard to be designed. There are several important aspects of GA design that are outlined below:

Individual Representation

For GA execution, each potential solution has to be encoded as a set of parameters. These parameters (known as genes) constitute a single applicable solution (often referred to as Chromosome). The encoding of the genes values (called alleles) can follow several schemes such as binary encoding, real value encoding, permutation encoding and tree encoding. Details of the encoding schemes can be found in [122]. For a given design problem, the set of parameters defining a solution is called the genotype, while the performance of an individual based on its genotype is called phenotype.

Fitness Function

For each problem, an appropriate fitness function needs to be formulated to assign a merit score to each chromosome and evaluate its ability in solving the problem. Some design tasks have multiple objectives and hence the fitness function may be required to combine several performance measures.

Selection and Reproduction

During the evolution process, the selection criteria of the highly fit individuals is critical as it affects the convergence speed and the final offered solution. The adopted selection methods include [174]:

- Proportionate Roulette Wheel Selection: The probability of selecting an individual for the next generation is equal to its fitness value over the sum of all chromosomes fitnesses.
- Ranking-based Selection: The individuals are sorted according to their fitness values. Then, the selection probability is assigned either linearly related or exponentially weighted depending on their ranks.
- Tournament Selection: It is one of the most popular techniques where a a number of individuals are randomly chosen from the entire population, then the one with the best fitness is passed to the next generation.

After the selection of the fit chromosomes, a set of operators are applied to allow for the evolution of the generations. The main operators are crossover and mutation. In crossover, two chromosomes are split at a random cut point (or multiple points) and their genes are swapped creating two new offspring. There is a crossover probability that controls the portion of the population that is subjected to crossover. Mutation is applied to each chromosome individually to alter single genes with a given probability providing a small amount of random search and ensuring the exploration of search space.

Convergence and Search Termination

As the populations evolve over successive generations, the average fitness of the population approaches the fitness of the best fit individual heading towards convergence. Convergence can be defined as the progression towards increasing fitness uniformity leading to the global optimum.
Termination of GA search procedure can be based on several criteria [91], which include maximum number of generations, minimum improvement in the best provided fitness value over a number of generations and a minimum value for the sum of deviations between individuals within a generation.

Bibliography

[1] Hervé Abdi and Lynne J Williams. "Principal component analysis". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 2.4 (2010), pp. 433–459.

[2] L. Abdi and S. Hashemi. "To Combat Multi-Class Imbalanced Problems by Means of Over-Sampling Techniques". In: *IEEE Transactions on Knowledge and Data Engineering* 28.1 (2016), pp. 238–251.

[3] Hidenao Abe and Takahira Yamaguchi. "Constructive Meta-level Feature Selection Method Based on Method Repositories". In: *Advances in Knowledge Discovery and Data Mining: 10th Pacific-Asia Conference, PAKDD 2006, Singapore, April 9-12, 2006. Proceedings*. Ed. by Wee-Keong Ng et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 70–80.

[4] David W Aha. "Lazy Learning". In: ed. by David W Aha. Norwell, MA, USA: Kluwer Academic Publishers, 1997. Chap. Lazy Learn, pp. 7–10.

[5] David W Aha, Dennis Kibler, and Marc K Albert. "Instance-based learning algorithms". In: *Machine Learning* 6.1 (1991), pp. 37–66.

[6] Aida Ali, Siti Mariyam Shamsuddin, and Anca Ralescu. "Classification with class imbalance problem: A review". In: *Int. J. Advance Soft Compu. Appl* 7 (Jan. 2015), pp. 176–204.

[7] Mina Alibeigi, Sattar Hashemi, and Ali Hamzeh. "DBFS: An effective Density Based Feature Selection scheme for small sample size and high dimensional imbalanced data sets". In: *Data and Knowledge Engineering* 81-82 (2012), pp. 67 –103.

[8] T Aljuaid and S Sasi. "Proper imputation techniques for missing values in data sets". In: *2016 International Conference on Data Science and Engineering (ICDSE)*. 2016, pp. 1–5.

[9] Stephen R Alty et al. "Predicting arterial stiffness from the digital volume pulse waveform". In: *IEEE Transactions on Biomedical Engineering* 54.12 (2007), pp. 2268–2275.

[10] James A Anderson and Joel Davis. *An introduction to neural networks*. Vol. 1. MIT Press, 1995.

[11] P.K. Anooj. "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules". In: *Journal of King Saud University - Computer and Information Sciences* 24.1 (2012), pp. 27 –40.

[12] E E A Arts et al. "Performance of four current risk algorithms in predicting cardiovascular events in patients with early rheumatoid arthritis". In: *Annals of the Rheumatic Diseases* 74.4 (2015), pp. 668–674.

[13] Benjamin Auffarth, Maite López, and Jesús Cerquides. "Comparison of Redundancy and Relevance Measures for Feature Selection in Tissue Classification of CT Images". In: *Advances in Data Mining. Applications and Theoretical Aspects*. Ed. by Petra Perner. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 248–262.

[14] J Bai, S Xiang, and C Pan. "A Graph-Based Classification Method for Hyperspectral Images". In: *IEEE Transactions on Geoscience and Remote Sensing* 51.2 (2013), pp. 803–817.

[15] M R Baneshi and A R Talei. "Does the missing data imputation method affect the composition and performance of prognostic models?" eng. In: *Iranian Red Crescent medical journal* 14.1 (2012), pp. 31–36.

[16] Yongguang Bao, Naohiro Ishii, and Xiaoyong Du. "Combining Multiple k-Nearest Neighbor Classifiers Using Different Distance Functions". In: *Intelligent Data Engineering and Automated Learning – IDEAL 2004: 5th International Conference, Exeter, UK. August 25-27, 2004. Proceedings*. Ed. by Zheng Rong Yang, Hujun Yin, and Richard M Everson. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 634–641.

[17] S. Barua et al. "MWMOTE–Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning". In: *IEEE Transactions on Knowledge and Data Engineering* 26.2 (2014), pp. 405–425.

[18] M. Bashiri and A. Farshbaf Geranmayeh. "Tuning the parameters of an artificial neural network using central composite design and genetic algorithm". In: *Scientia Iranica* 18.6 (2011), pp. 1600 –1608.

[19] D. Beasley, D. R. Bull, and R. R. Martin. "An overview of genetic algorithms: Part 1, fundamentals". In: *University Computing* 15.2 (1993), pp. 56–69.

[20] Mehdi Bejani, Davood Gharavian, and Nasrolah Charkari. "Audiovisual emotion recognition using ANOVA feature selection method and multi-classifier neural networks". In: *Neural Computing and Applications* 24 (Feb. 2014).

[21] Yoshua Bengio et al. "Greedy Layer-Wise Training of Deep Networks". In: *Advances in Neural Information Processing Systems 19*. Ed. by P B Schölkopf, J C Platt, and T Hoffman. MIT Press, 2007, pp. 153–160.

[22] M. Bennasar, Y. Hicks, and R. Setchi. "Feature selection using Joint Mutual Information Maximisation". In: *Expert Systems with Applications* 42.22 (2015), pp. 8520 –8532.

[23] James C Bezdek, Robert Ehrlich, and William Full. "FCM: The fuzzy c-means clustering algorithm". In: *Computers & Geosciences* 10.2 (1984), pp. 191–203.

[24] Asaf Bitton and Thomas A Gaziano. "The Framingham Heart Study's impact on global risk assessment." eng. In: *Progress in cardiovascular diseases* 53.1 (2010), pp. 68–78.

[25] Rok Blagus and Lara Lusa. "Class prediction for high-dimensional class-imbalanced data". In: *BMC Bioinformatics* 11.1 (2010), pp. 523 –539.

[26] Avrim L Blum and Pat Langley. "Selection of relevant features and examples in machine learning". In: *Artif. Intell.* 97.2 (1997), pp. 245–271.

[27] Verónica Bolón-Canedo, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos. "Distributed feature selection: An application to microarray data classification". In: *Applied soft computing* 30 (2015), pp. 136–150.

[28] N John Bosomworth. "Practical use of the Framingham risk score in primary prevention Canadian perspective". In: *Canadian Family Physician* 57.4 (2011), pp. 417–423.

[29] Y Boykov and V Kolmogorov. "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26.9 (2004), pp. 1124–1137.

[30] Y Boykov, O Veksler, and R Zabih. "Fast approximate energy minimization via graph cuts". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23.11 (2001), pp. 1222–1239.

[31] L. Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32.

[32] L. Breiman et al. "Classification and Regression Trees". In: *Machine Learning* 19 (1984), pp. 293–325.

[33] Leo Breiman. *Classification and regression trees*. Belmont, Calif: Wadsworth International Group, 1984, p. 358.

[34] Peter M Brindle et al. *The accuracy of the Framingham risk-score in different socioeconomic groups: a prospective study*. eng. 2005.

[35] British Heart Foundation. *Heart statistics, BHF Statistics Factsheet - UK*. https://www.bhf.org.uk/statistics, Last accessed on 2018-11-15. 2018.

[36] Miguel Á. Carreira-Perpiñán and Geoffrey E. Hinton. "On Contrastive Divergence Learning". In: *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, AISTATS 2005, Bridgetown, Barbados, January 6-8, 2005*. 2005.

[37] Girish Chandrashekar and Ferat Sahin. "A survey on feature selection methods". In: *Computers and Electrical Engineering* 40.1 (2014), pp. 16–28.

[38] N.V. Chawla et al. "SMOTE: Synthetic Minority Oversampling TEchnique". In: *Journal of Artificial Intelligence Research* 16 (2002), pp. 321–357.

[39] Xue-wen Chen and Michael Wasikowski. "FAST: A Roc-based Feature Selection Metric for Small Samples and Imbalanced Data Classification Problems". In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '08. New York, NY, USA: ACM, 2008, pp. 124–132.

[40] K. H. Cho, T. Raiko, and A. Ilin. "Gaussian-Bernoulli deep Boltzmann machine". In: *The 2013 International Joint Conference on Neural Networks (IJCNN)*. 2013, pp. 1–7.

[41] David A. Cieslak and Nitesh V. Chawla. "Learning Decision Trees for Unbalanced Data". In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by Walter Daelemans, Bart Goethals, and Katharina Morik. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 241–256.

[42] John G Cleary, Leonard E Trigg, and Others. "K*: An instance-based learner using an entropic distance measure". In: *Proceedings of the 12th International Conference on Machine learning*. Vol. 5. 1995, pp. 108–114.

[43] R M Conroy et al. "Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project". In: *European Heart Journal* 24.11 (2003), pp. 987–1003.

[44] R.M. Conroy and et al. "Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project". In: *European Heart Journal* 24.11 (2003), pp. 987–1003.

[45] J W Cooley and J W Tukey. "An algorithm for the machine calculation of complex Fourier series". In: *Mathematics of Computation* 19.90 (1965), pp. 297–301.

[46] Marie Therese Cooney, Alexandra L Dudina, and Ian M Graham. "Value and Limitations of Existing Scores for the Assessment of Cardiovascular Risk: A Review for Clinicians". In: *Journal of the American College of Cardiology* 54.14 (2009), pp. 1209–1227.

[47] D.R. Cox and D. Oakes. *Analysis of Survival Data*. Chapman and Hall, 1984, p. 212.

[48] John P Cunningham and Zoubin Ghahramani. "Linear Dimensionality Reduction : Survey , Insights , and Generalizations". In: *Journal of Machine Learning Research* 16 (2015), pp. 2859–2900.

[49] R. B. D'Agostino et al. "General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study". In: *Circulation* 117.6 (2008), pp. 743–753.

[50] B B Damodaran, R R Nidamanuri, and Y Tarabalka. "Dynamic Ensemble Selection Approach for Hyperspectral Image Classification With Joint Spectral and Spatial Information". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8.6 (2015), pp. 2405–2417.

[51] M Dash, H Liu, and J Yao. "Dimensionality reduction of unsupervised data". In: *Proceedings of Ninth IEEE International Conference on Tools with Artificial Intelligence*. 1997, pp. 532–539.

[52] B Dashtbozorg, A M Mendonça, and A Campilho. "An Automatic Graph-Based Approach for Artery/Vein Classification in Retinal Images". In: *IEEE Transactions on Image Processing* 23.3 (2014), pp. 1073–1083.

[53] P.A. Devijver and J Kittler. *Pattern Recognition: A statistical approach*. Prentice Hall Englewood Cliffs, NJ, 1982.

[54] I S Dhillon, Y Guan, and B Kulis. "Weighted Graph Cuts without Eigenvectors A Multilevel Approach". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.11 (2007), pp. 1944–1957.

[55] C H Q Ding et al. "A min-max cut algorithm for graph partitioning and data clustering". In: *Proceedings 2001 IEEE International Conference on Data Mining*. 2001, pp. 107–114.

[56] Yiran Dong and Chao-Ying Joanne Peng. *Principled missing data methods for researchers*. eng. 2013.

[57] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. 2nd Editio. New York: Wiley-Interscience, 2000.

[58] E Emary, Hossam M Zawbaa, and Aboul Ella. "Binary ant lion approaches for feature selection". In: *Neurocomputing* 213 (2016), pp. 54–65.

[59] S M Ali Eslami et al. "The Shape Boltzmann Machine: A Strong Model of Object Shape". In: *International Journal of Computer Vision* 107.2 (2014), pp. 155–176.

[60] K M Fathalla et al. "Cardiovascular risk prediction based on Retinal Vessel Analysis using machine learning". In: *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 2016, pp. 880–885.

[61] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "The KDD process for extracting useful knowledge from volumes of data". In: *Commun. ACM* 39.11 (1996), pp. 27–34.

[62] Usama M Fayyad and Keki B Irani. "The Attribute Selection Problem in Decision Tree Generation". In: *Proceedings of the Tenth National Conference on Artificial Intelligence*. AAAI'92. AAAI Press, 1992, pp. 104–110.

[63] Francisco Fernández-Navarro, César Hervás-Martínez, and Pedro Antonio Gutiérrez. "A dynamic over-sampling procedure based on sensitivity for multi-class problems". In: *Pattern Recognition* 44.8 (2011), pp. 1821 –1833.

[64] Josef Flammer et al. "The eye and the heart." In: *European heart journal* 34.17 (2013), pp. 1270–1278.

[65] Eibe Frank, Mark Hall, and Bernhard Pfahringer. "Locally Weighted Naive Bayes". In: *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*. UAI'03. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003, pp. 249–256.

[66] J M Gaziano. *Atlas of Cardiovascular Risk Factors*. Current Medicine Group, 2005.

[67] Zong Woo Geem. *Music-Inspired Harmony Search Algorithm: Theory and Applications*. 1st. Springer Publishing Company, Incorporated, 2009. ISBN: 364200184X, 9783642001840.

[68] "Global Guideline for Type 2 Diabetes". In: *Diabetes Research and Clinical Practice* 104.1 (2014), pp. 1–52.

[69] Cleotilde Gonzalez, Javier F Lerch, and Christian Lebiere. "Instance-based learning in dynamic decision making". In: *Cognitive Science* 27.4 (2003), pp. 591 –635.

[70] P Gourdeau et al. "Feature selection and oversampling in analysis of clinical data for extubation readiness in extreme preterm infants". In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2015, pp. 4427–4430.

[71] Shivani Gupta and Atul Gupta. "Handling class overlapping to detect noisy instances in classification". In: *The Knowledge Engineering Review* 33 (2018), e8.

[72] R Guthke and B Ludwig. "Generation of rules for expert systems by statistical methods of fermentation data analysis". In: *Acta biotechnologica* 14.1 (1994), pp. 13–26.

[73] Isabelle Guyon and André Elisseeff. "An Introduction to Variable and Feature Selection". In: *Journal of Machine Learning Research* 3 (2003), pp. 1157–1182.

[74] Isabelle Guyon et al. "Gene Selection for Cancer Classification using Support Vector Machines". In: *Machine Learning* 46.1 (2002), pp. 389–422.

[75] Mark Hall et al. "The WEKA Data Mining Software: An Update". In: *SIGKDD Explorations* 11.1 (2009), pp. 10–18.

[76] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning". In: *Advances in Intelligent Computing*. Ed. by De-Shuang Huang, Xiao-Ping Zhang, and Guang-Bin Huang. Springer Berlin Heidelberg, 2005, pp. 878–887.

[77] Ahmad Basheer Hassanat, Mohammad Ali Abbadi, and Ahmad Ali Alhasanat. "Solving the Problem of the K Parameter in the KNN Classifier Using an Ensemble Learning Approach". In: *International Journal of Computer Science and Information Security* 12.8 (2014), pp. 33–39.

[78] Simon S Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall Englewood Cliffs, NJ, 2007.

[79] Haibo He and Edwardo a. Garcia. "Learning from imbalanced data". In: *IEEE Transactions on Knowledge and Data Engineering* 21.9 (2009), pp. 1263–1284.

[80] Haibo He et al. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning". In: *Proceedings of the International Joint Conference on Neural Networks* 3 (2008), pp. 1322–1328.

[81] Nicolas Heess, Nicolas Le Roux, and John Winn. "Weakly Supervised Learning of Foreground-Background Segmentation Using Masked RBMs". In: *Artificial Neural Networks and Machine Learning – ICANN 2011: 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14-17, 2011, Proceedings, Part II*. Ed. by Timo Honkela et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 9–16.

[82] R. Heitmar et al. "Retinal vessel diameters and reactivity in diabetes mellitus and/or cardiovascular disease". In: *Cardiovascular Diabetology* 16.1 (2017), pp. 56 –66.

[83] Rebekka Heitmar et al. "Altered blood vessel responses in the eye and finger in coronary artery disease." In: *Investigative Ophthalmology & Visual Science* 52 (9 2011), pp. 6199 –6205.

[84] Alt Helmut and Michael Godau. "Computing the Fréchet distance between two polygonal curves". In: *International Journal of Computational Geometry and Applications* 5.1-2 (1995), pp. 75–91.

[85] G. E. Hinton and R. R. Salakhutdinov. "Reducing the Dimensionality of Data with Neural Networks". In: *Science* 313.5786 (2006), pp. 504–507.

[86] Geoffrey E Hinton. "A Practical Guide to Training Restricted Boltzmann Machines". In: *Neural Networks: Tricks of the Trade: Second Edition*. Ed. by Grégoire Montavon, Geneviève B Orr, and Klaus-Robert Müller. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 599–619.

[87] J. Hippisley-Cox et al. "Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2". In: *BMJ* 336.7659 (2008), pp. 1475–1482.

[88] Mark A Hlatky et al. "Criteria for Evaluation of Novel Markers of Cardiovascular Risk: A Scientific Statement From the American Heart Association". In: *Circulation* 119.17 (2009), pp. 2408–2416.

[89] F D R Hobbs et al. "Barriers to cardiovascular disease risk scoring and primary prevention in Europe". In: *QJM: An International Journal of Medicine* 103.10 (2010), pp. 727–739.

[90] M. Hollander and D. A. Wolfe. *Nonparametric Statistical Methods*. Hoboken, NJ: John Wiley & Sons, Inc., 1999.

[91] C. R. Houck, J. Joines, and M. G. Kay. "A genetic algorithm for function optimization: a Matlab implementation". In: *Ncsu-ie tr* 95.09 (1995), pp. 1–10.

[92] Hui-Huang Hsu, Cheng-Wei Hsieh, and Ming-Da Lu. "Hybrid feature selection by combining filters and wrappers". In: *Expert Systems with Applications* 38.7 (2011), pp. 8144 –8150.

[93] L. Hu et al. "Feature selection considering two types of feature relevancy and feature interdependency". In: *Expert Systems with Applications* 93 (2018), pp. 423 –434.

[94] Li-Yu Hu et al. *The distance function effect on k-nearest neighbor classification for medical datasets*. eng. 2016.

[95] Yuming Hua, Junhai Guo, and Hua Zhao. "Deep Belief Networks and deep learning". In: *Proceedings of 2015 International Conference on Intelligent Computing and Internet of Things*. 2015, pp. 1–4.

[96] Kononenko Igor. *Estimating attributes: analysis and extensions of RELIEF*. Catania, Italy, 1994.

[97] M K Ikram et al. "Retinal vessel diameters and risk of stroke: The Rotterdam Study". In: *Neurology* 66.9 (2006), pp. 1339–1343.

[98] Yoo Illhoi et al. "Data Mining in Healthcare and Biomedicine: A Survey of the Literature". In: *J. Med. Syst.* 36.4 (2011), pp. 2431–2448.

[99] Holger Jessen and Timo Slawinski. "Test-and Rating Strategies for Data Based Rule Generation". In: *Computational Intelligence* (1998).

[100] J.F.Ramirez-Villegas et al. "Heart Rate Variability Dynamics for the Prognosis of Cardiovascular Risk". In: *PLOS One* 6 (2011), pp. 1–15.

[101] Taeho Jo and Nathalie Japkowicz. "Class Imbalances Versus Small Disjuncts". In: *SIGKDD Explor. Newsl.* 6.1 (June 2004), pp. 40–49. ISSN: 1931-0145.

[102] G. H. John and P. Langley. "Estimating Continuous Distributions in Bayesian Classifiers". In: *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. UAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 338–345.

[103] Vinayak Joshi et al. "Automated Method for Identification and Artery-Venous Classification of Vessel Trees in Retinal Vessel Networks". In: *PloS one* 9 (2014), e88061.

[104] L.E. Juarez-Orozco et al. "Improving the value of clinical variables in the assessment of cardiovascular risk using Artificial Neural Networks". In: *European Heart Journal* 38.1 (2017), pp. 227–228.

[105] Segun Jung, Yingtao Bi, and Ramana V Davuluri. *Evaluation of data discretization methods to derive platform independent isoform expression signatures for multiclass tumor subtyping*. eng. 2015.

[106] Jan Kalina and Anna Schlenker. "A Robust Supervised Variable Selection for Noisy High-Dimensional Data." eng. In: *BioMed research international* 2015 (2015), pp. 320–330.

[107] K V Ravi Kanth et al. "Dimensionality Reduction for Similarity Searching in Dynamic Databases". In: *Computer Vision and Image Understanding* 75 (1999), pp. 59–72.

[108] M A Karaolis et al. "Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees". In: *IEEE Transactions on Information Technology in Biomedicine* 14.3 (2010), pp. 559–566.

[109] George Karypis and Vipin Kumar. "Multilevelk-way Partitioning Scheme for Irregular Graphs". In: *Journal of Parallel and Distributed Computing* 48.1 (1998), pp. 96–129.

[110] Fatemeh Kaveh-Yazdy, Mohammad-Reza Zare-Mirakabad, and Feng Xia. "A Novel Neighbor Selection Approach for KNN: A Physiological Status Prediction Case Study". In: *Proceedings of the 1st International Workshop on Context Discovery and Data Mining*. ContextDD '12. New York, NY, USA: ACM, 2012, 2:1–2:7.

[111] Ryo Kawasaki et al. "Retinal microvascular signs and risk of stroke: the Multi-Ethnic Study of Atherosclerosis (MESA)." In: *Stroke; Journal of the American Heart Association* 43.12 (2012), pp. 1984–1992.

[112] Mohammad Ali Keyvanrad and Mohammad Mehdi Homayounpour. *A brief survey on deep belief networks and introducing a new object oriented toolbox (DeeBNet)*. Tech. rep. 2014, 27 pages.

[113] Samina Khalid, Tehmina Khalil, and Nasreen Shamila. "A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning". In: *Science and Information Conference (SAI)*. 2014, pp. 372–378.

[114] Maryam Mahsal Khan, Stephan K Chalup, and Alexandre Mendes. "Parkinson's Disease Data Classification Using Evolvable Wavelet Neural Networks". In: *Artificial Life and Computational Intelligence: Second Australasian Conference, ACALCI 2016, Canberra, ACT, Australia, February 2-5, 2016, Proceedings*. Cham: Springer International Publishing, 2016, pp. 113–124.

[115] Jaekwon Kim, Ungu Kang, and Youngho Lee. "Statistics and Deep Belief Network-Based Cardiovascular Risk Prediction". In: *Healthc Inform Res* 23.3 (2017), pp. 169–175.

[116] Kenji Kira and Larry A Rendell. "A practical approach to feature selection". In: *Proceedings of the ninth international workshop on Machine learning*. 1992, pp. 249–256.

[117] Ronald Klein et al. "Changes in Retinal Vessel Diameter and Incidence and Progression of Diabetic Retinopathy". In: *Arch Ophthalmol* 130.6 (2012), pp. 749–755.

[118] V Kolmogorov and R Zabin. "What energy functions can be minimized via graph cuts?" In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26.2 (2004), pp. 147–159.

[119] Igor Kononenko. "Machine learning for medical diagnosis: history, state of the art and perspective". In: *Artificial Intelligence in Medicine* 23.1 (2001), pp. 89 –109.

[120] Kornelia Kotseva et al. "EUROASPIRE III: a survey on the lifestyle, risk factors and use of cardioprotective drug therapies in coronary patients from 22 European countries". In: *European Journal of Cardiovascular Prevention & Rehabilitation* 16.2 (2009), pp. 121–137.

[121] Amit Kumar et al. "A new hybrid feature selection approach using feature association map for supervised and unsupervised classification". In: *Expert Systems with Applications* 88 (2017), pp. 81–94.

[122] Anit Kumar. "Encoding schemes in genetic algorithm". In: *International Journal of Advanced Research in IT and Engineering* 2.3 (2013), pp. 1–7.

[123] TW Lam, W MA, and R Luo. "Restricted Boltzmann Machine and its Potential to Better Predict Cancer Survival". In: *Biomedical Journal of Scientific and Technical Research* (2018).

[124] IM Lanzl et al. "How are healthy vessels getting old?" In: *Dtsch Med Wochenschr* 131 (2006), pp. 180–181.

[125] H. Larochelle et al. "Learning Algorithms for the Classification Restricted Boltzmann Machine". In: *J. Mach. Learn. Res.* 13.1 (2012), pp. 643–669.

[126] Hugo Larochelle and Yoshua Bengio. "Classification using discriminative restricted Boltzmann machines." In: *ICML*. Vol. 307. ACM International Conference Proceeding Series. ACM, 2008, pp. 536–543.

[127] J Laurikkala and M Juhola. "Nearest neighbour classification with heterogeneous proximity functions." eng. In: *Studies in Health Technology and Informatics* 77 (2000), pp. 753–757.

[128] D. D. Lewis. "Feature Selection and Feature Extraction for Text Categorization". In: *Proceedings of the Workshop on Speech and Natural Language*. HLT '91. Harriman, New York: Association for Computational Linguistics, 1992, pp. 212–217.

[129] Chunfeng Lian et al. "Robust Cancer Treatment Outcome Prediction Dealing with Small-Sized and Imbalanced Data from FDG-PET Images". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II*. Ed. by Sebastien Ourselin et al. Cham: Springer International Publishing, 2016, pp. 61–69.

[130] M. Lichman. *UCI Machine Learning Repository*. 2013. URL: http://archive.ics.uci.edu/m.

[131] Gerald Liew et al. "Retinal Vascular Imaging: A New Tool in Microvascular Disease Research". In: *Circulation: Cardiovascular Imaging* 1.2 (2008), pp. 156–161.

[132] Laurence S Lim et al. "Dynamic Responses in Retinal Vessel Caliber With Flicker Light Stimulation and Risk of Diabetic Retinopathy and Its Progression." In: *Investigative Ophthalmology & Visual Science* 58.5 (2017), pp. 2449–2455.

[133] M. Lin, K. Tang, and X. Yao. "Dynamic Sampling Approach to Training Neural Networks for Multiclass Imbalance Classification". In: *IEEE Transactions on Neural Networks and Learning Systems* 24.4 (2013), pp. 647–660.

[134] Jinghua Liu et al. "Feature selection based on quality of information". In: *Neurocomputing* 225 (2017), pp. 11–22.

[135] Y. Liu et al. "Understanding of Internal Clustering Validation Measures". In: *2010 IEEE ICDM*, pp. 911–916.

[136] Wei-Yin Loh and Yu-Shan Shih. "Split selection methods for classification trees". In: *Statistica sinica* 7 (1997), pp. 815–840.

[137] K Madasamy and M Ramaswami. "Data Imbalance and Classifiers: Impact and Solutions from a Big Data Perspective". In: *International Journal of Computational Intelligence Research* 13 (2017), pp. 2267–2281.

[138] Jay Magidson. *SPSS for Windows CHAID release 6.0*. SPSS Incorporated, 1993.

[139] Syed S Mahmood et al. "The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective". In: *The Lancet* 383.9921 (2014), pp. 999 –1008.

[140] David M Maslove, Tanya Podchiyska, and Henry J Lowe. *Discretization of continuous features in clinical datasets*. eng. 2013.

[141] Frank J Massey. "The Kolmogorov-Smirnov Test for Goodness of Fit". In: *Journal of the American statistical Association* 46.253 (1951), pp. 68–78.

[142] *Global atlas on cardiovascular disease prevention and control*. Tech. rep. Geneva, 2011, p. 164.

[143] Alireza Moayedikia et al. "Feature selection for high dimensional imbalanced class data using harmony search". In: *Engineering Applications of Artificial Intelligence* 57 (2017), pp. 38–49.

[144] K G M Moons et al. "Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker". In: *Heart* 98.9 (2012), pp. 683–690.

[145] Parham Moradi and Mozhgan Gholampour. "A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy". In: *Applied Soft Computing* 43 (2016), pp. 117–130.

[146] Stephanie Mroczkowska et al. "Coexistence of macro- and micro-vascular abnormalities in newly diagnosed normal tension glaucoma patients." In: *Acta ophthalmologica* 90.7 (2012), pp. 553–559.

[147] Edgar. Nagel, Walthard Vilser, and Ines Lanz. "Age, Blood Pressure, and Vessel Diameter as Factors Influencing the Arterial Retinal Flicker Response". In: *Investigative Ophthalmology & Visual Science* 45.5 (2004), pp. 1486–1492.

[148] T Thanh Nguyen et al. "Correlation of light-flicker-induced retinal vasodilation and retinal vascular caliber measurements in diabetes." In: *Investigative ophthalmology & visual science* 50.12 (2009), pp. 5609–5613.

[149] Menelaos Pavlou et al. "How to develop a more accurate risk prediction model when there are few events". In: *BMJ* 351 (2015).

[150] H. Peng, F. Long, and C. Ding. "Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27.8 (2005), pp. 1226–1238.

[151] H Peng, Fulmi Long, and C Ding. "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27.8 (2005), pp. 1226–1238.

[152] M Pfaff et al. "Prediction of cardiovascular risk in hemodialysis patients by data mining". In: *Methods Inf Med* 43.1 (2004), pp. 106–113.

[153] Vili Podgorelec et al. "Decision Trees: An Overview and Their Use in Medicine". In: *Journal of Medical Systems* 26.5 (2002), pp. 445–463.

[154] R. Poplin et al. "Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning". In: *Nature Biomedical Engineering* 2.3 (2018), pp. 158–164.

[155] *Prevention of cardiovascular disease : guidelines for assessment and management of cardiovascular risk*. Geneva: World Health Organization, 2007. ISBN: 9789241547178.

[156] Purushottam, Kanak Saxena, and Richa Sharma. "Efficient Heart Disease Prediction System". In: *Procedia Computer Science* 85 (2016). International Conference on Computational Modelling and Security (CMS 2016), pp. 962 –969.

[157] J R Quinlan. "Induction of Decision Trees". In: *Machine Learning.* 1.1 (1986), pp. 81–106.

[158] Arvind Raghu et al. "Implications of Cardiovascular Disease Risk Assessment Using the WHO/ISH Risk Prediction Charts in Rural India". In: *PLOS ONE* 10.8 (Aug. 2015), pp. 1–13.

[159] Juan F Ramirez-Villegas et al. "Heart Rate Variability Dynamics for the Prognosis of Cardiovascular Risk". In: *PLoS One* 6.2 (2011).

[160] V M Rao and V N Sastry. "Unsupervised feature ranking based on representation entropy". In: *2012 1st International Conference on Recent Advances in Information Technology (RAIT)*. 2012, pp. 421–425.

[161] Yuri Rodrigues et al. "Wrappers Feature Selection in Alzheimer's Biomarkers Using kNN and SMOTE Oversampling". In: *Trends in Applied and Computational Mathematics* 18.1 (2017), p. 15.

[162] Kai Rothaus, Xiaoyi Jiang, and Paul Rhiem. "Separation of the retinal vascular graph in arteries and veins based upon structural knowledge". In: *Image and Vision Computing* 27.7 (2009), pp. 864–875.

[163] Patrick Royston. "Polynomial Regression". In: *Wiley StatsRef: Statistics Reference Online*. American Cancer Society, 2014.

[164] Andrea M. Ruggiero. "Comparison of a traditional and a multilevel Cox Proportional Hazards model". PhD thesis. 2001, p. 42.

[165] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Third. Series in Artificial Intelligence. Prentice Hall, 2010.

[166] Astrid Schneider, Gerhard Hommel, and Maria Blettner. "Linear Regression Analysis". In: *Dtsch Arztebl International* 107.44 (2010), pp. 776–782.

[167] Manjeevan Seera and Chee Peng Lim. "A hybrid intelligent system for medical data classification". In: *Expert Systems With Applications* 41.5 (2014), pp. 2239–2249.

[168] S.B. Seidelmann et al. "Retinal vessel calibers in predicting long-term cardiovascular outcomes: the atherosclerosis risk in communities study". In: *Circulation* (2016), pp. 1328–1338.

[169] B U Seifertl and W Vilser. "Retinal Vessel Analyzer (RVA)–design and function." In: *Biomedizinische Technik. Biomedical engineering* 47.2 (2002), pp. 678–681.

[170] S Seshadri. *Retinal Vascular Function and Cardiovascular Risk Factors*. Aston University, 2015. URL: https://books.google.co.uk/books?id=7CrovgEACAAJ.

[171] Swathi Seshadri, Aniko Ekart, and Doina Gherghel. "Ageing effect on flicker-induced diameter changes in retinal microvessels of healthy individuals". In: *Acta ophthalmologica* 94.1 (2016), pp. 35–42.

[172] Swathi Seshadri et al. "Retinal vascular function in asymptomatic individuals with a positive family history of cardiovascular disease". In: *Acta Ophthalmologica* 96.8 (2018), e956–e962.

[173] Wenqian Shang et al. "An Improved kNN Algorithm – Fuzzy kNN". In: *Computational Intelligence and Security: International Conference, CIS 2005, Xi'an, China, December 15-19, 2005, Proceedings Part I*. Ed. by Yue Hao et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 741–746.

[174] A. Shukla, H. M. Pandey, and D. Mehrotra. "Comparative review of selection techniques in genetic algorithm". In: *2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE)*. IEEE. 2015, pp. 515–519.

[175] Nigel C. Smeeton. "Early history of the kappa statistic". In: *Biometrics* 41 (1985), pp. 795–795.

[176] Jyoti Soni et al. "Article: Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction". In: *International Journal of Computer Applications* 17.8 (2011), pp. 43–48.

[177] Nitish Srivastava and Ruslan Salakhutdinov. "Multimodal Learning with Deep Boltzmann Machines". In: *J. Mach. Learn. Res.* 15.1 (Jan. 2014), pp. 2949–2980.

[178] Ewout W Steyerberg et al. "Assessing the performance of prediction models: a framework for traditional and novel measures". In: *Epidemiology* 21.1 (2010), pp. 128–138.

[179] X. Sun et al. "Selection of interdependent genes via dynamic relevance analysis for cancer diagnosis". In: *Journal of Biomedical Informatics* 46.2 (2013), pp. 252 –258.

[180] Y Sun. "Iterative RELIEF for Feature Weighting: Algorithms, Theories, and Applications". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.6 (2007), pp. 1035–1051.

[181] Wong T. et al. "Quantitative retinal venular caliber and risk of cardiovascular disease in older persons: The cardiovascular health study". In: *Archives of Internal Medicine* 166.21 (2006), pp. 2388–2394.

[182] W. Tang et al. "Classification for overlapping classes using optimized overlapping region detection and soft decision". In: *2010 13th International Conference on Information Fusion*. 2010, pp. 1–8.

[183] Eugenia Tedeschi-Reiner et al. "Relation of Atherosclerotic Changes in Retinal Arteries to the Extent of Coronary Artery Disease". In: *The American journal of cardiology* 96.8 (2005), pp. 1107–1109.

[184] Robert Tibshirani et al. "Diagnosis of multiple cancer types by shrunken centroids of gene expression". In: *Proceedings of the National Academy of Sciences* 99.10 (2002), pp. 6567–6572.

[185] D M Titterington and J Sedransk. "Imputation of missing values using density estimation". In: *Statistics & Probability Letters* 8.5 (1989), pp. 411–418.

[186] Yavuz Unal, Kemal Polat, and H Erdinc Kocer. "Classification of Vertebral Column Disorders and Lumbar Discs Disease using Attribute Weighting Algorithm with Mean Shift Clustering". In: *Measurement* 77 (2016), pp. 278–291.

[187] Michael Unser. "Ten good reasons for using spline wavelets". In: *Proc. SPIE, Wavelets Applications in Signal and Image Processing* 3169.5 (1997), pp. 422–431.

[188] N Vladimir Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995, p. 188.

[189] Roy Varshavsky et al. "Novel Unsupervised Feature Filtering of Biological Data". In: *Bioinformatics* 22.14 (2006), e507–e513.

[190] Eric Yuk Fai Wan et al. "Ten-year risk prediction models of complications and mortality of Chinese patients with diabetes mellitus in primary care in Hong Kong: a study protocol". In: *BMJ Open* 8.10 (2018).

[191] J. Wang, P. Neskovic, and L. N. Cooper. "Neighborhood size selection in the k-nearest-neighbor rule using statistical confidence". In: *Pattern Recognition* 39.3 (2006), pp. 417–423.

[192] Jie Jin Wang et al. "Retinal vessel diameter and cardiovascular mortality: pooled data analysis from two older populations." In: *European heart journal* 28.16 (2007), pp. 1984–1992.

[193] Jingcheng Wang et al. "Dimension reduction method of independent component analysis for process monitoring based on minimum mean square error". In: *Journal of Process Control* 22.2 (2012), pp. 477–487.

[194] Jingzhong Wang and Xia Li. "An improved KNN algorithm for text classification". In: *2010 International Conference on Information, Networking and Automation (ICINA)*. Vol. 2. 2010, pp. V2–436–V2–439.

[195] J.J. Wang et al. *Retinal vascular calibre and the risk of coronary heart disease-related death*. 2006.

[196] N. Wang, J. Melchior, and L. Wiskott. "An analysis of Gaussian-binary restricted Boltzmann machines for natural images." In: *ESANN*. 2012.

[197] N. Wang, J. Melchior, and L. Wiskott. "Gaussian-binary Restricted Boltzmann Machines on Modeling Natural Image Statistics". In: *PLOS ONE* 12 (3 2017).

[198] M Wasikowski and X w. Chen. "Combating the Small Sample Class Imbalance Problem Using Feature Selection". In: *IEEE Transactions on Knowledge and Data Engineering* 22.10 (2010), pp. 1388–1400.

[199] Geoffrey I Webb. "Lazy Learning". In: *Encyclopedia of Machine Learning*. Ed. by Claude Sammut and Geoffrey I Webb. Boston, MA: Springer US, 2010, pp. 571–572.

[200] Gary M Weiss. "Mining with rarity: a unifying framework". In: *ACM Sigkdd Explorations Newsletter* 6.1 (2004), pp. 7–19.

[201] Stephen F. Weng et al. "Can machine-learning improve cardiovascular risk prediction using routine clinical data?" In: *PLOS ONE* 12.4 (Apr. 2017), pp. 1–14.

[202] Frank Wilcoxon. "Individual comparisons by ranking methods". In: *Biometrics bulletin* 1.6 (1945), pp. 80–83.

[203] *European Cardiovascular Disease Statistics 2017*. Tech. rep. Brussels, 2017, p. 192.

[204] I. H. Witten, E. Frank, and M. A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2011.

[205] Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[206] World Health Organisation - Europe Section. *Data and Statistics of Cardiovascular Diseases*. http://www.euro.who.int/en/health-topics/noncommunicable-diseases/cardiovascular-diseases/data-and-statistics, Last accessed on 2018-11-15. 2018.

[207] World Health Organisation - Global. *Cardiovascular diseases (CVDs)*. http://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds), Last accessed on 2018-11-15. 2017.

[208] Zhipeng Xie et al. "SNNB: A Selective Neighborhood Based Na{\"i}ve Bayes for Lazy Learning". In: *Advances in Knowledge Discovery and Data Mining: 6th Pacific-Asia Conference, PAKDD 2002 Taipei, Taiwan, May 6–8, 2002 Proceedings*. Ed. by Ming-Syan Chen, Philip S Yu, and Bing Liu. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 104–114.

[209] Haitao Xiong, Junjie Wu, and Lu Liu. "Classification with ClassOverlapping: A Systematic Study". In: *Proceedings of the 1st International Conference on E-Business Intelligence (ICEBI2010)*, Atlantis Press, 2010.

[210] Hongzeng Xu et al. "Development of a diagnosis model for coronary artery disease". In: *Indian Heart Journal* 69.5 (2017), pp. 634 –639.

[211] T Yamashita et al. "To Be Bernoulli or to Be Gaussian, for a Restricted Boltzmann Machine". In: *2014 22nd International Conference on Pattern Recognition*. 2014, pp. 1520–1525.

[212] Show-Jane Yen and Yue-Shi Lee. "Cluster-based under-sampling approaches for imbalanced data distributions". In: *Expert Systems with Applications* 36.3 (2009), pp. 5718–5727.

[213] Liuzhi Yin et al. "Feature selection for high-dimensional imbalanced data". In: *Neurocomputing* 105 (2013), pp. 3–11.

[214] Illhoi Yoo et al. "Data Mining in Healthcare and Biomedicine: A Survey of the Literature". In: *Journal of Medical Systems* 36.4 (2012), pp. 2431–2448.

[215] Alireza Yousefpour et al. "Feature Reduction Using Standard Deviation with Different Subsets Selection in Sentiment Analysis". In: *Intelligent Information and Database Systems*. Ed. by Ngoc Thanh Nguyen et al. Cham: Springer International Publishing, 2014, pp. 33–41.

[216] L Yu and H Liu. "Feature selection for high-dimensional data: A fast correlation-based filter solution". In: *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-* 20.2 (2003), p. 856.

[217] C Yun et al. "Feature Subset Selection Based on Bio-Inspired Algorithms". In: *Journal of Information Science and Engineering* 27.5 (2011), pp. 1667–1686.

[218] Ezzeddine Zagrouba, Siwar Ben Gamra, and Asma Najjar. "Model-based graph-cut method for automatic flower segmentation with spatial constraints". In: *Image and Vision Computing* 32.12 (2014), pp. 1007–1020.

[219] Huaxiang Zhang and Mingfang Li. "RWO-Sampling: A random walk over-sampling approach to imbalanced data classification". In: *Information Fusion* 20 (2014), pp. 99 –116.

[220] Qi Zhang et al. "Deep learning based classification of breast tumors with shear-wave elastography". In: *Ultrasonics* 72.Supplement C (2016), pp. 150–157.

[221] Zheng Zhao and Huan Liu. "Searching for Interacting Features". In: *Proceedings of the 20th International Joint Conference on Artifical Intelligence*. IJCAI'07. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2007, pp. 1156–1161.