DEVELOPMENT OF UNSUPERVISED FEATURE SELECTION METHODS FOR HIGH DIMENSIONAL BIOMEDICAL DATA IN REGRESSION DOMAIN

by

Ferdi SARAC

A thesis submitted in partial fulfillment of the requirements of Northumbria University for the degree of Doctor of Philosophy

2017

Declaration of Authorship

I declare that the work contained in this thesis has not been submitted for any other award and that it is all my own work. I also confirm that this work fully acknowledges opinions, ideas and contributions from the work of others. Any ethical clearance for the research presented in this thesis has been approved. Approval has been sought and granted by the Faculty Ethics Committee / University Ethics Committee / external committee 12.09.2015

FERDI SARAC

Signed: _____

Date: _____

"Whether you think you can or whether you think you can't, you are right!"

Henry Ford

Abstract

In line with technological developments, there is almost no limit to collect data of high dimension in various fields including bioinformatics. In most cases, these high dimensional datasets contain many irrelevant or noisy features which need to be filtered out to find a small but biologically meaningful set of attributes. Although there have been various attempts to select predictive feature sets from high dimensional data in classification and clustering, there have only been limited attempts to do this for regression problems. Since supervised feature selection methods tend to identify noisy features in addition to discriminative variables, unsupervised feature selection methods (USFSMs) are generally regarded as more unbiased approaches. The aim of this thesis is, therefore, to provide (i) a comprehensive overview of feature selection methods for regression problems where feature selection methods are shown along with their types, references, sources, and code repositories (ii) a taxonomy of feature selection methods for regression problems to assist researchers to select appropriate feature selection methods for their research (iii) a deep learning based unsupervised feature selection framework, DFSFR (iv) a K-means based unsupervised feature selection method, KBFS. To the best of our knowledge, DFSFR is the first deep learning based method to be designed particularly for regression tasks. In addition, a hybrid USFSM, DKBFS, is proposed which combines KBFS and DFSFR to select discriminative features from very high dimensional data. The proposed frameworks are compared with the state-of-the-art USFSMs, including Multi Cluster Feature Selection (MCFS), Embedded Unsupervised Feature Selection (EUFS), Infinite Feature Selection (InFS), Spectral Regression Feature Selection (SPFS), Laplacian Score Feature Selection (LapFS), and Term Variance Feature Selection (TV) along with the entire feature sets as well as the methods used in previous studies. To evaluate the effectiveness of proposed methods, four different case studies are considered: (i) a low dimensional RV144 vaccine dataset; (ii) three different high dimensional peptide binding affinity datasets; (iii) a very high dimensional GSE44763 dataset; (iv) a very high dimensional GSE40279 dataset. Experimental results from these data sets are used to validate the effectiveness of the proposed methods. Compared to state-of-the-art feature selection methods, the proposed methods achieve improvements in prediction accuracy of as much as 9% for the RV144 Vaccine dataset, 75% for the peptide binding affinity datasets, 3% for the GSE44763 dataset, and 55% for the GSE40279 dataset.

Acknowledgements

The present work is generated and developed during 3 years as a graduate student in a warm, friendly research and study environment at Northumbria University, Newcastle. The author would like to thank his principal supervisor Dr. Huseyin Seker, for leading me to success and for sharing his knowledge and time like a friend throughout the author's 3 years of PhD study. The author would also like to thank his second supervisor Prof. Ahmed Bouridane for his invaluable support, assistance and guidance.

Contents

List of Figures

List of Tables

Abbreviations

USFSMs	**U**n**S**upervised **F**eature **S**election **M**ethods
SVR	**S**upport **V**ector **R**egression
MSVR	**M**ulti **S**upport **V**ector **R**egression
PPI	**P**rotein **P**rotein **I**nteragtion
CV	**C**ross **V**alidation
RNA	**R**ibonucleic **A**cid
DNA	**D**eoxyribonucleic **A**cid
BMI	**B**ody **M**ass **I**ndex
HIV	**H**uman **I**mmunodeficiency **V**irus
CpG	**C**ytosine **P**hosphate **G**uanine
MISO	**M**ulti **I**nput **S**ingle **O**utput
RBM	**R**estricted **B**olzman **M**achine

List of Publications

1. *Ferdi Sarac, Huseyin Seker, and Ahmed Bouridane. Exploration of unsupervised feature selection methods to predict chronological age of individuals by utilising cpg dinucleotics from whole blood. In 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2017.*

2. *Ferdi Sarac, Volkan Uslan, Huseyin Seker, and Ahmed Bouridane. A supervised feature selection framework in relation to prediction of antibody feature-function activity relationships in RV144 vaccines. In Systems Man and Cybernetics (SMC 2016) Conference of the IEEE, 2016.*

3. *Ferdi Sarac, Volkan Uslan, Huseyin Seker, and Ahmed Bouridane. Exploration of unsupervised feature selection methods in relation to the prediction of cytokine release effect correlated to antibody features in rv144 vaccines. In Bioinformatics and Bioengineering (BIBE), 2015 IEEE 15th International Conference on, pages 1-4. IEEE, 2015.*

4. *Ferdi Sarac, Volkan Uslan, Huseyin Seker, and Ahmed Bouridane. Comparison of unsupervised feature selection methods for high-dimensional regression problems in prediction of peptide binding affinity. In 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 8173-8176. IEEE, 2015.*

5. *F. Sarac, V. Uslan, H. Seker, A. Bouridane, Unsupervised selection of RV144 hiv vaccine-induced antibody features correlated to natural killer cell-mediated cytotoxic reactions, in: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2016.*

6. *Ferdi Sarac and Huseyin Seker. An instance selection framework for mining data streams to predict antibody-feature function relationships on rv144 hiv vaccine recipients. In Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on, pages 33563361, 2016.*

7. *Ferdi Sarac, Huseyin Seker, and Ahmed Bouridane. A Support Vector Regression Based Model for the Quantitative Prediction of Age and Body Mass Index by using Epigenetic Information from Peripheral Blood. International Conference on Cloud and Big Data Computing (ICCBDC 2017) of the ACM.*

Chapter 1

Introduction

1.1 Motivation

In line with the technological developments, there is almost no foreseeable limit to the collection of data of high dimension in fields, such as bioinformatics, computer vision, machine learning. Therefore, there is a pressing need to be able to deal with high dimensional data. Over the last three decades, the dimensionality of data associated with various scientific fields has dramatically increased. The growth trend in the feature and sample size in UCI Machine Learning Repository from mid 80s to 2012 are shown 1.1(a) and 1.1(b), respectively. [1]. It is clear that there is a need for not only the organisation, distribution and storage of higher volumes of data, but also for identifying and understanding important information from them through the use of machine learning tools to automatically analyse the content of large volumes of data.

One of the aforementioned domains is bioinformatics, where high dimensional biomedical data needs to be processed. There are various types of biomedical data, including peptide binding affinities and epigenetic biomarkers that contain a large number of features. For example, there are over 512 billion peptides for each major histocompatibility complex (MHC) molecule [7]. Biological experiments with such large volumes of biomedical data is often impractical, costly and time consuming.

Machine learning methods have become one of the preferred approaches to the analysis of high dimensional biomedical data. However, the handling of high

dimensional data poses many challenges to most existing machine learning algorithms. One of the considerable challenges is curse of dimensionality which states that if the number of features increases, the number of data samples required to train learning algorithm exponentially increases to achieve the same level of performance for classification, regression, and clustering tasks.

Another important challenge when dealing with high dimensional data is that such data does not only contain relevant features, but also a significant number of irrelevant and redundant features which usually deteriorate learning performance, increase computational cost, and lead to overfitting. Relevant features are the ones that contain important information which can be used to solve a prediction problem. Redundant features encompass critical information which has been already provided by another feature, and therefore, these features do not provide additional useful information for the predictive model [8]. Irrelevant features are those that have no valuable information; hence, their presence reduce the learning performance of predictive models. Consequently, there is a need to remove redundant and irrelevant features from high dimensional data in order to increase the prediction performance of a model and to reduce computational time.

In order to overcome the aforementioned problems, dimensionality reduction, which is one of the most effective tools to address those challenges, can be used. Dimensionality reduction methods can be divided into two main categories: feature selection and feature extraction.

Feature extraction reduces the dimensionality of the data and constructs new input data with no physical meaning, and these methods include Locally Linear Embedding (LLE) [9], Neighbourhood Preserving Embedding (NPE) [10], and kernel PCA [11]. On the other hand, feature selection builds a subset of relevant attributes without changing the original semantics of the data. Preserving the original semantics of data is vital, especially in biomedical domain. In addition, feature selection reduces execution time and improves the accuracy of prediction which are preferred in many real-world applications [12].

It is profoundly beneficial to remove irrelevant and redundant features prior to learning, particularly if the number of attributes are significantly greater than the number of samples, as is the usual case in biomedical data. Moreover, feature selection methods generate a subset of relevant features in biomedical data so

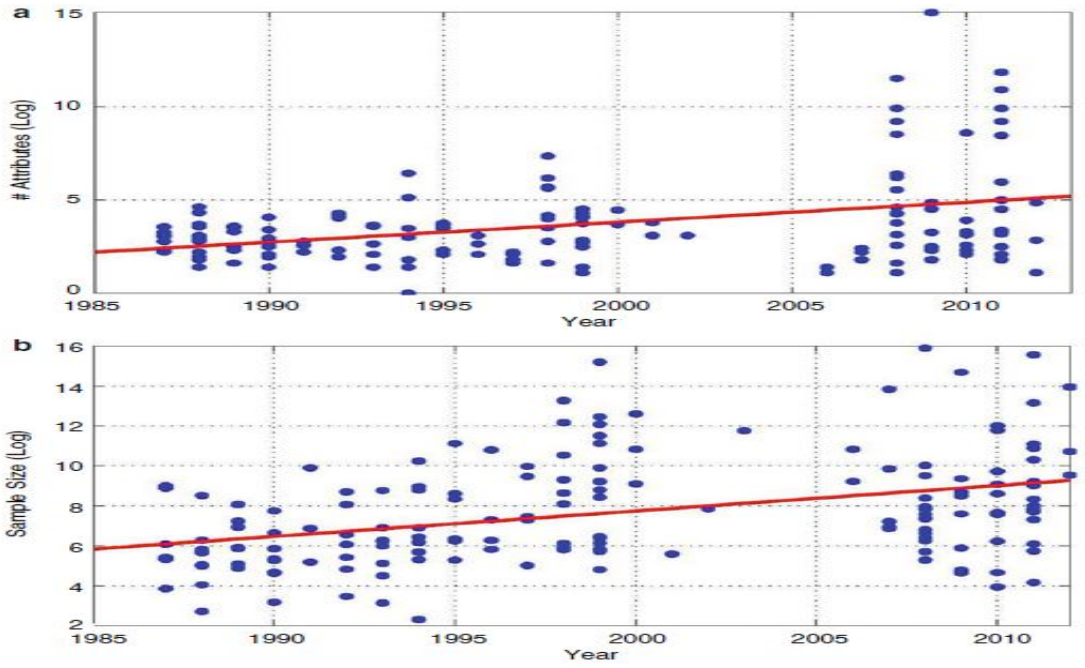

FIGURE 1.1: Growth trend in UCI Machine Learning Repository [1].

that those features can be further analysed in biology laboratories to discover new insights in the field.

Another important challenge of dealing with high dimensional data is that label (output) information is generally not available, owing to the high cost of manual labelling [13]. Therefore, unsupervised feature methods are needed to deal with unlabelled high dimensional data.

In line with the technological developments, data has been generated; however, floating point data is much more in agenda. For example, a decade ago, the problem of peptide binding was to predict whether peptide binds or not. However, current technological developments have lead researchers to predict bindings of peptides quantitatively. Consequently, this study focuses on unsupervised feature selection particularly for regression problems.

1.2 Aims and Objectives of This Thesis

The thesis focuses primarily on feature selection problems with extremely high dimensional data in regression domain.

- Developing an unsupervised feature selection method that is capable of dealing with high dimensionality of data, identifying discriminative features and removing redundant, noisy and irrelevant ones.
- Achieving better prediction and generalisation performance than the existing methods.

In order to achieve the project aim ther following objectives have been set:

- There have been various attempts to select predictive feature sets from high dimensional data sets in classification and clustering; however, only limited attempts have been made to do this for regression problems. Therefore, one of the goals of this study is to develop a feature selection method designed particularly for regression problems in order to fill this gap in the literature.
- Deep learning has been shown to be capable of representing data at multiple levels of abstraction. It is able to derive discriminative features, resulting in enhanced accuracy. Although various feature selection methods have been proposed in the current literature, no deep learning based feature selection method exists specifically for regression tasks.
- Most real world data is unlabelled; therefore, unsupervised feature selection methods are needed since supervised methods can not be applied to unlabelled data. Furthermore, supervised methods tend to identify noisy features as well as relevant ones, yet unsupervised methods do not intend to select features that can act as noise. Therefore, supervised feature selection can be considered as a biased approach whereas unsupervised feature selection can be regarded as unbiased [14].
- Researchers have mainly paid attention to single-output regression analysis so far [15]. However, multi-output regression is crucial, especially in the analysis of biomedical data.
- Although plenty of reviews of feature selection methods can be found in the literature for classification and clustering, no review of feature selection methods specifically for regression tasks has yet been published.

As mentioned above, researchers have paid more attention to feature selection for classification rather than for regression. In order to justify that a literature

search using the keywords “feature selection classification” and “feature selection regression” has been conducted of publications listed at PubMED, Scopus and Web of Science (It is worth noting that searching with different keywords, such as feature selection and classification or feature selection for classification have produced almost the same results, therefore, the number of studies found using different versions of keywords are consistent). The numbers of publications per year for the feature selection for regression and feature selection for classification between 2011 and 2016 are shown in Figs. 1.2-1.5. As shown in Fig. 1.2, feature selection for classification studies are more than three times those for regression according to PubMed. Fig. 1.3 shows that feature selection for classification studies are approximately 6 times the number of studies for feature selection for regression studies according to Scopus. Fig. 1.4 illustrates a comparison of published feature selection studies for classification and regression on Web of Science, which suggests that there are approximately five times as many feature selection for classification studies than feature selection for regression studies. As shown in Fig. 1.5, the literature search indicates that there have been feature selection for classification studies approximately five times as many studies as feature selection for regression studies from 2011 to 2016. Thus, it is concluded that feature selection for regression is understudied.

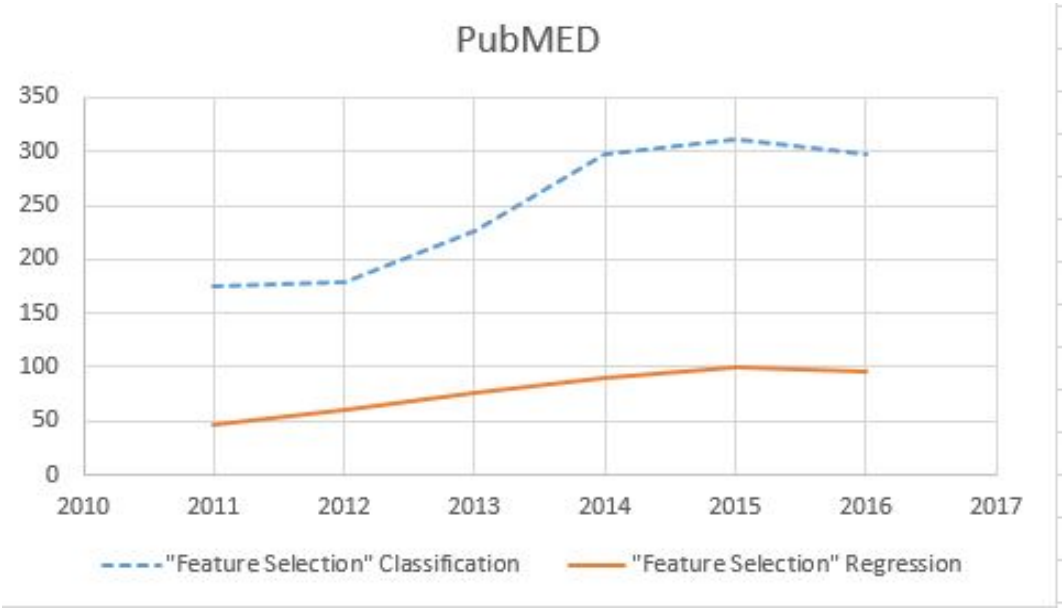

FIGURE 1.2: A Comparison of Published Feature Selection Studies for Classification and Regression on PubMed.

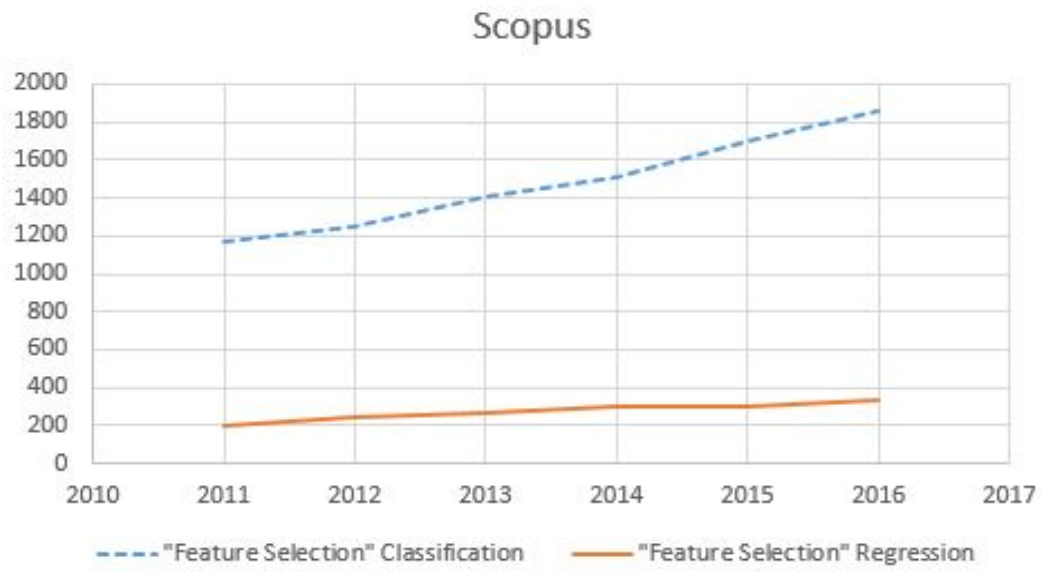

FIGURE 1.3: A Comparison of Published Feature Selection Studies for Classification and Regression on Scopus.

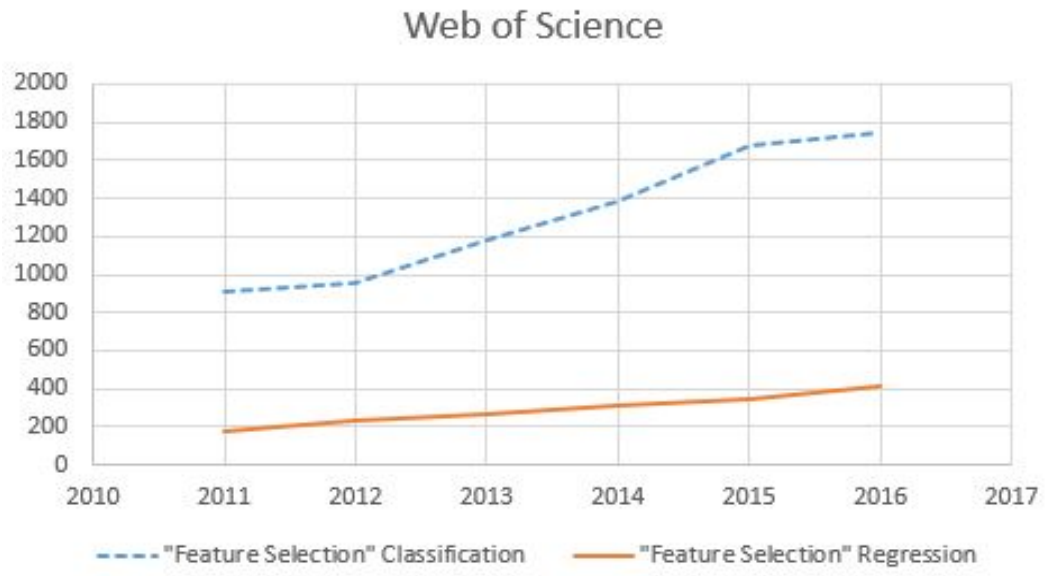

FIGURE 1.4: A Comparison of Published Feature Selection Studies for Classification and Regression on Web of Science.

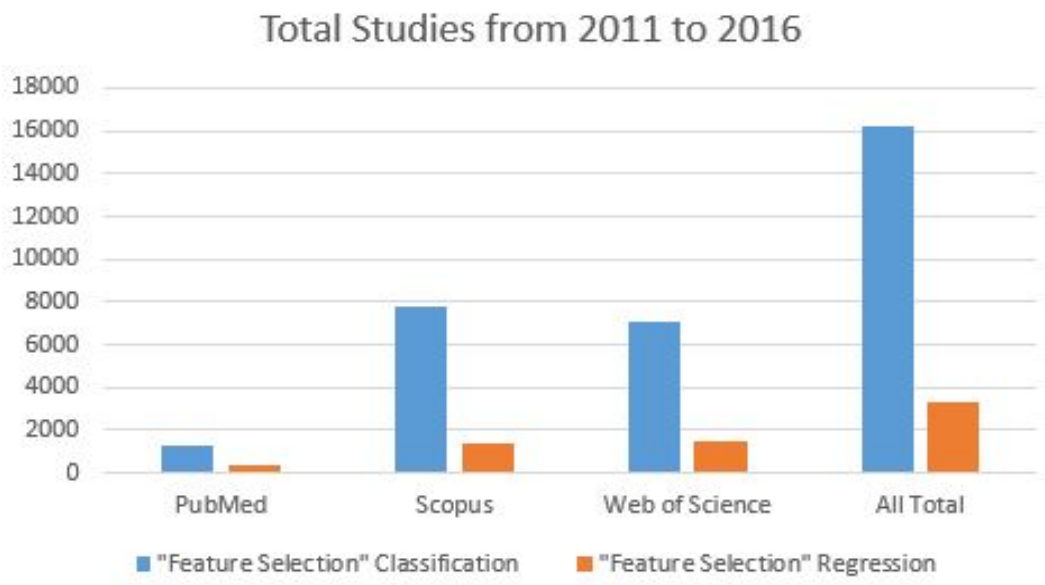

FIGURE 1.5: A Comparison of Total Number of Published Feature Selection Studies for Classification and Regression on PubMed, Scopus, Web of Science from 2011 to 2016.

In this thesis, therefore, a novel deep learning based unsupervised feature selection framework, a K-means based unsupervised framework, and a hybrid method for regression problems are provided to overcome the aforementioned problems and to fill the research gap in the literature. In addition, the proposed deep learning based unsupervised framework is capable of handling both multi input-single output (MISO) and multi input-multi output (MIMO) prediction.

By proposing these frameworks, the intention is not only to obtain better generalisation and performance than with existing unsupervised feature selection methods, but also to be able to identify a small subset of relevant features from biomedical data which can be further analysed in real biology labs. The ultimate goal is to be able to identify biologically relevant features from biomedical data, such as the identification of age-related biomarkers from the whole blood of individuals in order to contribute to society.

1.3 Contributions of the Thesis

In accordance with the objectives of this study, the contributions of this thesis are as follows:

- A comprehensive review of existing feature selection methods, which can be used for regression tasks, is provided.
- A taxonomy of existing feature selection methods for regression tasks is offered.
- Feature selection methods are developed that can be applicable to unlabelled data.
- A K-means based unsupervised feature selection framework for high dimensional data is proposed particularly for the regression domain, which achieves better performance (in terms of higher accuracy with fewer features) than existing feature selection methods. (Published work ([16]) and another work is under review [17]).
- A deep learning based unsupervised feature selection method is designed that can be applied specifically for regression tasks.
- Multi input-multi output regression analysis is applied so that associations among target variables can be revealed. (This work is under review [18] [17]).
- A hybrid unsupervised feature selection method is proposed which combines the proposed K-means and deep learning based frameworks.

1.4 Thesis Overview

The thesis is organised as follows:

Chapter 1 introduces the problems of dealing with high dimensional data, indicates the importance of feature selection, and establishes the goals of this thesis. The main contributions of this study are also summarised in this chapter.

Chapter 2 discusses the challenges of dealing with high dimensional data, such as the curse of dimensionality and overfitting. Feature selection and feature extraction are defined and their advantages and disadvantages are presented. A comprehensive review of existing feature selection algorithms for regression tasks is conducted, and a taxonomy of existing unsupervised feature selection methods particularly for regression problems is provided.

Chapter 3 describes the regression models, which are exploited in this study to perform both single input-multi output and multi input-multi output regression. The evaluation metrics that are used to analyse and compare the effectiveness of unsupervised feature selection methods are presented, and the RV144 vaccine, peptide binding affinity, GSE44763 and GSE40279 data sets that are exploited in this research to evaluate the performance of proposed frameworks are described.

Chapter 4 describes the K-means algorithm, presents its basic properties and the shortcomings of existing K-means based feature selection methods. The proposed K-means based unsupervised feature selection framework, which is called as KBFS is then introduced. Finally, the results of the application of the proposed method compared to state-of-the-art unsupervised feature selection techniques over the RV144 vaccine, peptide binding affinities, GSE44763 and GSE40279 data sets are presented.

Chapter 5 identifies research gaps in the literature and describes deep belief network (DBN) which is a type of deep neural network used in this research. The proposed deep learning based unsupervised feature selection framework for regression tasks is presented which is called DFSFR. A new hybrid model, which combines the proposed KBFS and DFSFR methods, is also proposed in this chapter. The proposed hybrid method is named DKBFS. Finally, experimental results are presented to show effectiveness of proposed methods.

Chapter 6 presents discussions of the performance of feature selection methods which are reviewed in detail. The robustness of unsupervised feature selection methods for the RV144 vaccine, peptide binding affinity, GSE44763, and GSE40279 data sets is shown, and a general discussion and interpretation of the research findings of this study is provided.

Chapter 7 concludes the thesis and suggests possible topics for future research.

Chapter 2

Review on Feature Selection Methods

This chapter is devoted to reviewing existing feature selection methods. The challenges of dealing with high dimensional data is reviewed first, then dimensionality reduction, feature selection and feature extraction will be described. In the following section, existing feature selection methods for regression problems will be presented as well as a taxonomy of feature selection methods for regression problems. Finally, a list of those methods along with their types, sources and code availability will be presented. This taxonomy is provided to assist researchers to select the appropriate feature selection method for their research.

2.1 Challenges of Dealing With High Dimensional Data

High dimensional data has become very common in various domains, such as social media, biostatistics, bioinformatics, computational biology, etc. High dimensional data poses many challenges to most of the existing machine learning and data mining algorithms. One of the considerable challenge is the curse of dimensionality which is presented in following section. In addition, high dimensional data requires large storage and high computational cost for data analytics.

Real world data usually contains irrelevant and redundant features which are generally not beneficial to discriminate samples from different classes or clusters

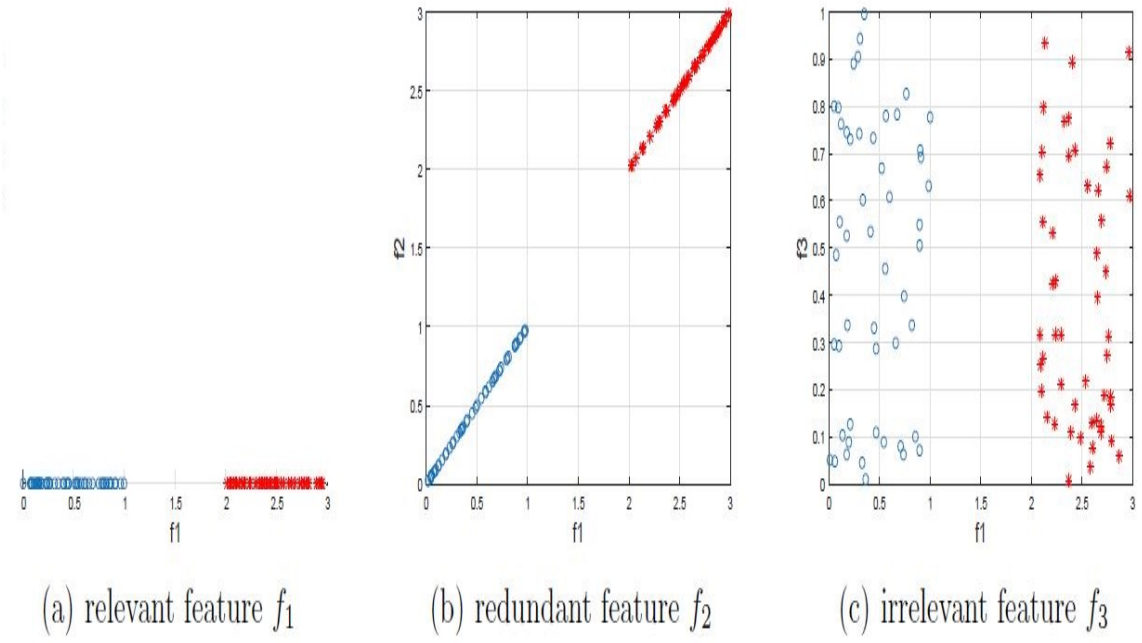

FIGURE 2.1: Relevant, Redundant and Irrelevant Features [2].

[2]. In fact, those features generally deteriorate learning performance and increase the computational cost. Hence, removing those features is usually beneficial for the learning model. In Fig. 2.1 [2], relevant, irrelevant and redundant features are demonstrated. In Fig. 2.1(a), a relevant feature, $f1$, is shown. Notice that f1 is a relevant feature as it discriminates two clusters. As shown in Fig. 2.1(b), if f1 and f2 are considered together, f2 is redundant because f2 is highly correlated to f1. In Fig. 2.1(c), f3 is an irrelevant feature since it is not able to separate two clusters. Consequently, learning performance will not be affected if f2 and f3 are removed.

Another important challenge of dealing with high dimensional data is overfitting. If a data set contains a huge number of features and relatively small number of samples, learning model is prone to overfitting which might negatively affect learning performance of the model [2].

2.2 Curse of Dimensionality

The curse of dimensionality is first introduced by Bellman [19] in order to specify that if the number of features increases, the amount of data to be generalised

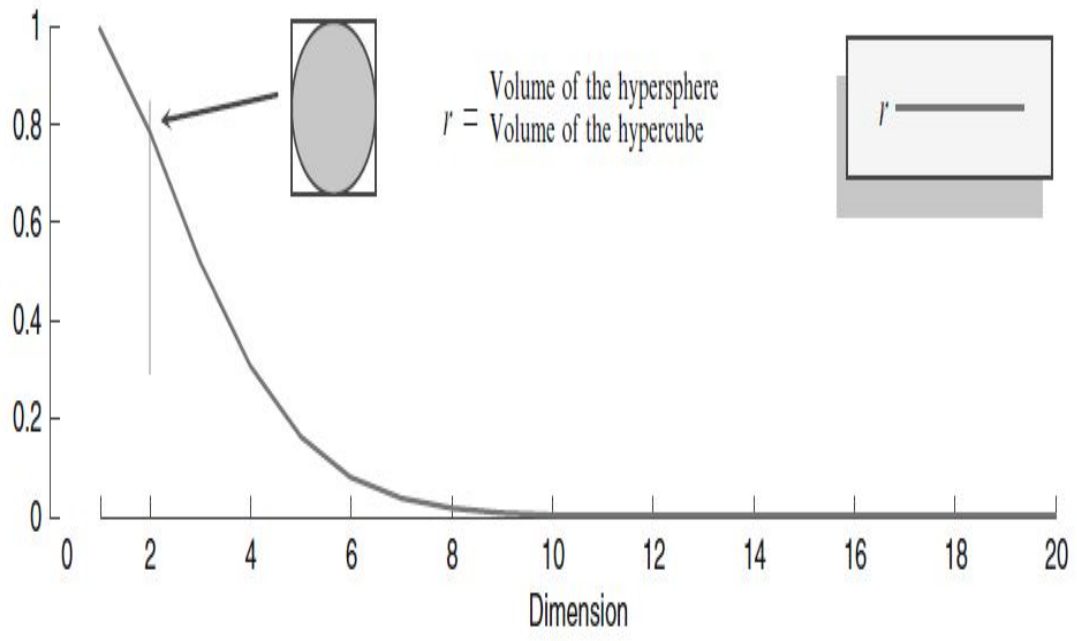

FIGURE 2.2: The Ratio of the Volume of The Hypersphere Enclosed by the Unit Hypercube [3].

is increases exponentially to achieve the same level of performance for classification, regression, and clustering [1]. In other words, exponential increase in volume results in adding extra dimensions to Euclidean space [3]. Thanks to this exponential growth, the volume of space increases which causes high sparsity in data. On the other hand, this sparseness is not uniformly distributed over the search space. In order to show that the size of unit hypersphere can be compared with the size of unit cube as shown in Fig. 2.2 [3]. As the dimensionality increases, the volume of hypersphere gets closer to zero whereas the volume of surrounding hypercube remains constant; furthermore, nearly entire high dimensional space is quite far away from the centre. Consequently, if the dimensionality goes to infinity, the ratio of difference between maximum (d_{max}) and minimum (d_{min}) euclidean distance from sample to centroid and the minimum distance (d_{min}) goes to zero:

$$lim_{d\to\infty}\frac{d_{max}-d_{min}}{d_{min}}\to 0 \tag{2.1}$$

Therefore, the data become more sparse as dimensionality increases. In order to overcome aforementioned problems, dimensionality reduction methods, such as

feature selection or feature extraction can be used. Next section presents feature selection and feature extraction.

2.3 Feature Selection and Feature Extraction

Feature selection and feature extraction are both effective dimensionality reduction techniques and they are able to improve performance, reduce the computational complexity and the cost, and decrease the requirements for the storage of the data [20]. In contrast to feature extraction, feature selection techniques do not change the original semantics of the variables, actually, it eliminates redundant or irrelevant features to identify meaningful smaller subset of the variables [21]. Furthermore, feature extraction generates a sequence of new features without knowing their physical meanings [2]. This is quite dangerous and it may cause calamitous results if it is utilised on biomedical data since preserving intrinsic information of biomedical data is extremely important. On the other hand, feature selection identifies a subset of relevant attributes by preserving actual meanings of original features. Therefore, feature selection does not change original semantics of the attributes, indeed, it increases feature readability and interpretability [22].

Feature selection methods are generally designed for three different strategies: filter [23] [24] [25], wrapper [26] [27] [28] [29] and embedded selection [30]. Filter subset selection is performed independent from the prediction algorithm. Filter methods are computationally fast; however, they do not take learning algorithms into account which generally results in lower prediction performance [31]. Unlike filters, wrapper methods require a pre-determined learning algorithm and utilise the dependency between features and prediction algorithm to select a subset of features. Consequently, the prediction performance of wrappers is better than filters, however, they are costly to compute and inefficient for dealing with high dimensional data [32] [33]. Embedded methods exploit the advantages of filter and wrapper methods, thereby, they learn the prediction algorithm and select features, simultaneously. Embedded methods are still dependent to induction algorithms, yet they are more computationally efficient than wrappers.

Feature evaluation process of filter based methods can be univariate and multivariate. Univariate filters rank features independently according to their importance whereas multivariate filters evaluate each feature with respect to the other features [34] [35]. Therefore, multivariate feature selection methods are able to handle feature redundancy [36]. Three different feature selection strategies, which are filter, wrapper and embedded, are summarised in Table 2.1.

Method	Advantages	Disadvantages
Filter	• Fast, Scalable, Independent from learning algorithm, The lowest computational cost, Good generalisation ability	• No interaction with prediction algorithm
Wrapper	• Simple, Interacts with learning algorithm, Captures feature dependencies Good prediction performance	• Computationally expensive, Dependent to learning algorithm, Risk of overfitting
Embedded	• Interaction with induction algorithm, Capture feature dependencies, Lower computational cost than wrappers	• Feature Selection is dependent on learning algorithm

TABLE 2.1: The Advantages and Disadvantages of Different Feature Selection Strategies

Based on the availability of information and problem definition in prediction, feature selection methods can be divided into two main categories: unsupervised and supervised feature selection. In the supervised feature selection scenario, features are selected according to their correlation with outputs (e.g., class labels). In case of unsupervised feature selection, only data inputs are used to select relevant features where the output information (e.g., class label) is not available or taken into account. As output information is not used for the feature selection,

carrying out the unsupervised feature selection is more challenging. Furthermore, supervised feature selection methods tend to identify relevant features as well as noisy ones whereas unsupervised feature selection methods do not tend to identify features that can act as noise [16].

In the following sections, existing feature selection methods for regression problems will be presented. A taxonomy of the existing methods is also presented to assist researchers to select an appropriate feature selection method for their research. To the best of our knowledge, this is the first study that provides a comprehensive review of feature selection methods particularly for regression problems.

2.4 Supervised Feature Selection

In this section supervised feature selection methods for regression problems are presented. Although there have been various attempts to select predictive feature sets from high-dimensional data sets in classification and clustering, there is a limited attempt to study it in regression problems as demonstrated in Figs 1.2-1.5 where the number of studies in PubMed, Scopus and Web of Science on the feature selection in regression domains are found to be significantly different than those in classification ones. Therefore, feature selection methods for regression problems are presented in this section, yet most of these methods have not been used for regression problems.

2.4.1 Filter Methods

In this subsection supervised filter feature selection methods are presented.

2.4.1.1 Correlation Based Feature Selection (CFS)

Correlation based feature selection (CFS) [25] is a filter feature selection algorithm that aims to minimise internal correlation of selected variables and maximise the dependence between the selected variables and target. Briefly, it uses a correlation based heuristic to rank the features. The CFS does not only evaluate feature-feature correlations, but also measures input-output correlations.

If a feature is highly correlated to another feature, it is considered irrelevant. However, if a feature is strongly correlated with the target, it is determined as relevant [24]. The CFS estimates correlation between features and the target, r_{xy}, by solving the following formula:

$$r_{xy} = \frac{\sum xy}{n\sigma_x\sigma_y} \tag{2.2}$$

where X and Y are the features and the target variable respectively, σ_x is the standard deviation of the x, σ_y is standard deviation of the y, and n represents the number of samples. CFS ranks feature subsets rather than scoring each feature individually; therefore, CFS is a multivariate feature selection method.

CFS has been applied only in data sets with low dimension for regression tasks and it is observed that their performances varied from one data set to another and they generally produced average performance in various domains [25] [37].

2.4.1.2 Minimum Redundancy Maximum Relevance (mRmR)

Minimum redundancy maximum relevance (mRmR) [38] is a filter-based and supervised feature selection algorithm that selects features which are mutually far away from each other, yet they are highly correlated to the target variable. The idea of minimum redundancy is to select features that are considerable dissimilar. The idea of maximum relevance is to maximise the total relevance of all features. The minimum redundancy can be calculated as:

$$W = \frac{1}{|S|^2}\sum_{i,j} c(i,j) \tag{2.3}$$

and the maximum relevance can be found by solving the following formula:

$$V_F = \frac{1}{|S|}\sum_{i\epsilon S} F(i,h) \tag{2.4}$$

where S is the set of features, $|S|$ is the number of features in S, $c(i;j)$ is the correlation between features i and j, h is the target, and $F(i,h)$ is the F-statistic.

mRmR method is one of the few feature selection methods that can be applied in both classification and regression tasks. The literature appears to suggest

that it usually yields reasonably better performance on high dimensional data sets where the number of features are dramatically greater than the number of samples [39] [40].

2.4.1.3 Canonical Correlation Analysis (CCA)

Canonical Correlation Analysis (CCA) [41] describes the relationship between two multivariate sets of variables. The CCA constructs a subset of features according to the correlation between input and output variables. In order to calculate the correlation between U_i and V_j, the covariance between these two variables, $cov(U_i, V_j)$, is divided by the square root of the product of the variances:

$$Correlation = \frac{(cov(U_i, V_j))}{\sqrt{var(U_i, V_j)}} \tag{2.5}$$

The canonical correlation is a particular type of this correlation. Thus, i-th canonical variate pair is the correlation between U_i ,V_i and it can be calculated from the following formula:

$$p_i^* = \frac{(cov(U_i, V_j))}{\sqrt{var(U_i, V_j)}} \tag{2.6}$$

where U_is are a set of linear combinations for X, and V_js are a set of linear combinations for Y, cov is co-variance, p is correlation and var represents the variance.

2.4.1.4 Maximum Likelihood Feature Selection (MLFS)

Maximum Likelihood Feature Selection (MLFS) [42] is a filter, multivariate and supervised feature selection method that prioritises variables based on input-target dependency measure. It utilises Maximum Likelihood Mutual Information (MLMI) [43] in order to measure the dependency between predictors and the target. MLMI is an estimator of mutual information which depends on density estimation. MLMI directly models the density ratio, $w(x, y)$ by [42]:

$$w(x, y) = \frac{P_{xy}(x, y)}{P_x(x)P_y(y)} \tag{2.7}$$

where $P_{xy}(x, y)$, is the joint density of X and Y, $P_x(y)$, $P_y(x)$ are densities of X and Y respectively. MLFS can be exploited for both classification and regression problems.

2.4.1.5 Least Squares Feature Selection (LSFS)

Least Squares Feature Selection (LSFS) [44] is quite similar to the MLFS method. Unlike MLFS, the LSFS evaluates the dependency between features and the target via squared loss mutual information (LSMI) [45]. LSMI directly estimates the density ratio, $r(x, y)$, by:

$$r(x, y) = \frac{P(x, y)}{P(x)P(y)} \tag{2.8}$$

where x_i and y_i, $(i = 1, 2, \ldots, n)$ are given a set of paired samples (x_i, y_i) which are drawn independently from the joint distribution with a density of $p(x, y)$.

MLFS is a supervised, multivariate filter, and information-based feature selection method that can be utilised for both classification and regression tasks.

2.4.1.6 Distance Measure Based Conditional Mutual Information (CMI_{DIST})

Distance Measure Based Conditional Mutual Information (CMI_{DIST}) [15] is a supervised and filter feature selection method that can be used to perform both single and multi-output regression tasks. It applies information based techniques to determine the importance of the features. Conditional Mutual Information is exploited in order to find the clusters in a data set. The CMI_{DIST} selects a feature that produces the highest mutual information with respect to the target variable.

CMI_{DIST} has been shown to produce good performance for small dimensional data sets, particularly if the number of samples are greater than number of features [15].

2.4.1.7 Selection via Intersection Method (SEVIM)

Selection via Intersection Method (SEVIM) [46] is a supervised and filter feature selection method. In SEVIM, features are selected based on the intersection of Maximum R^2, F score and p-values of the variables. In deed, incremental maximum R^2 technique is exploited, and in order to rank features the intersection of maximum R^2, F score and p values of the features are considered. The maximum R^2 can be formulated as [46]:

$$a = logL(M)logL(0) \tag{2.9}$$

$$b = \frac{logL(0)}{n} \tag{2.10}$$

$$Q = 1 - e^{\frac{2a}{n}} \tag{2.11}$$

$$R^2 = \frac{Q}{1 - e^{2b}} \tag{2.12}$$

where n is the number of features, $logL(M)$ is the maximised logarithmic likelihood and $logL(0)$ refers to the logarithmic likelihood of null model which contains only intercept term.

While finding a subset of features with highest R^2 is in progress, the F score of the subsets and their related p values are also calculated. Briefly, let $X = (x_1, x_2, \ldots, x_n)$ denotes a data matrix where $x_i \in R^d$ is the feature descriptor of the $i - th$ sample. SEVIM selects a feature, x_i, if $x_i \in F \cap P \cap R$.

SEVIM has been shown to produce good results for data sets where the number of features are greater than number of samples SEVIM.

2.4.2 Wrappers

The goal of the wrapper feature selection is to achieve maximum accuracy with the minimum number of discriminative features. Wrapper methods embeds the model hypothesis search within feature subset search. The wrapper approaches

of feature selection attempt to identify the minimum discriminative features in order to achieve a high prediction accuracy [47]. Since wrappers interact with the learning algorithm, their prediction performance is better than filters [48]. On the other hand, wrappers are computationally very expensive, and thereby they are under the risk of overfitting.

2.4.2.1 Sequential Forward Selection (SFS)

Sequential Forward Selection (SFS) is a supervised and wrapper feature selection method that starts from an empty set and gradually adds features one at a time until no further improvement of evaluation function value is possible [49]. When an attribute is added to the current set, the SFS puts the attribute to the learning structure that generalises the best. Once an attribute is added to the learning structure, the SFS cannot remove it. The aim of the evaluation function is to minimise the mean square error for prediction. A common pitfall of the SFS is that it may not contain inter-dependent attributes because it adds variables one at a time [50]. The SFS is more applicable to small data sets [51]. The pseudo code for the SFS algorithm is presented in Algorithm 4.

Algorithm 1 Sequential Forward Selection Algorithm

1: **procedure**
2: *Start with the empty set* $Y_0 = \emptyset$;
3: *Select the next best feature* $x^+ = arg_{x \notin Y_k} max J(Y_k + x)$
4: *Update* $Y_{k+1} = Y_k + x^+; \quad k = k + 1$
5: *go to 2*

SFS method has generally been applied to low dimensional data sets for regression tasks and it produced good results [50]. As mentioned earlier, SFS is more applicable to small data sets. SFS is a widely utilised feature selection algorithm thanks to its simplicity and speed [52].

2.4.2.2 Sequential Backward Selection (SBS)

Sequential Backward Selection (SBS) and SFS can be considered as antipodes. In contrast to SFS, the SBS is initialised with entire set of attributes, and it

updates the feature set by removing the feature which least reduces the value of the objective function. The pseudo code for SBS is presented in Algorithm 5.

Algorithm 2 Sequential Backward Selection Algorithm

1: **procedure**
2: *Start with the entire set* $Y_0 = X$;
3: *Remove the worst feature* $x^- = arg_{x \in Y_k} max J(Y_k - x)$
4: *Update* $Y_{k+1} = Y_k - x^-$; $k = k + 1$
5: *go to 2*

Since SBS starts with the whole set of features, thereby, its early evaluations are comparatively expensive [53]. The primary disadvantage of SBS is that once a feature is removed, it will never be re-evaluated [54]. The SBS spends most of its time for visiting a large subset; therefore, SBS can be exploited when the optimal feature subset contains a large number of attributes.

In [50], SBS was applied to a number of different data sets, but generally with low dimension and it generally produced better results than SFS. However, the number of features were at most 14 on those data sets.

2.4.2.3 Sequential Floating Selection (SFLS)

The SFS and SBS work on one direction either adding or removing an attribute at a time. Sequential Floating Selection (SFLS) works on both directions either adding or removing variables or eliminating added variables, and thereby the SFLS enhances the reliability of the final feature subset. There are two different types of SFLS methods: Sequential Floating Forward Selection (SFFS) and Sequential Floating Backward Selection (SFBS). The SFFS is initiated with the empty set as the SFS does; however, after each forward step, the SFFS performs backward steps until the objective function increases. On the other hand, the SFBS is initialised by the full set and after each backward step, the SFBS caries out forward steps as long as the objective function increases. The F is a statistical parameter which can be used to judge whether the models including different feature subsets are sequentially generated or not. The F parameter can

be calculated from the following formula [55]:

$$F = \frac{MSM}{MSE} = \frac{\frac{\sum_i (\widehat{y}_i - \overline{y}_i)^2}{q-1}}{\frac{(\sum_i \widehat{y}_i - \overline{y}_i)^2}{n-q}} \tag{2.13}$$

where i is the number of samples, y is the target, $\overline{y}$ is the mean of the target, $\widehat{y}$ is the predicted target, n is the number of features, q is the number of selected features, and MSM and MSE are mean of squares for model and mean of squares for error, respectively.

2.4.2.4 Bi-Directional Search

The goal of the Bi-directional Search algorithm is to ensure that the SFS and SBS converge toward the same solution. Therefore, features selected by the SFS should not be removed by the SBS, and the features removed by SBS should not be added by SFS. The pseudo code for BDS is illustrated in Algorithm 6.

Algorithm 3 Bi-Directional Search Selection Algorithm

1: **procedure**
2: *Start SFS with the empty set* $Y_F = \emptyset$;
3: *Start with the entire set* $Y_B = X$;
4: *Select the best feature*
5: $x^+ = \underset{x \notin Y_{F_k}, x \in Y_{B_k}}{argmin} [J(Y_{F_k} + x)$
6: $Y_{F_{k+1}} = Y_{F_k} + x^+$
7: Remove the worst feature
8: $x^- = \underset{x \notin Y_{F_{k+1}}, x \in Y_{B_k}}{argmax} [J(Y_{B_k} - x)$
9: $Y_{Bk+1} = Y_{B_k} - x^-; \quad k = k + 1$
10: *go to 2*

2.4.2.5 Feature Selection by Computing Statistical Scores (FeaLect)

FeaLect [56] is a feature selection method that statistically sorts features to prioritise them. It generates a number of samples from training data, and then determines the best relevance ordering of the features for each sample. At the end, it combines those to select maximally relevant features. Basically, FeaLect

selects a random subset B. Then selects k-features, in which by applying Least Absolute Shrinkage and Selection Operator (LASSO) method. If a feature belongs to subset B, then the value of the feature is $1/k$ otherwise the value of the feature is zero. This process is repeated 100 times and average values of features are calculated. LASSO, which is presented in the next subsection, can select relevant features as well as irrelevant ones, especially if the number of training instances goes to infinity [56]. The FeaLect is a wrapper feature selection algorithm that overcomes this problem by statistically scoring each feature to accomplish a robust feature selection [57].

2.4.3 Embedded Methods

In this subsection embedded supervised feature selection methods are presented. The objective function of embedded methods is to optimise the performance of a learning algorithm.

2.4.3.1 Least Absolute Shrinkage and Selection Operator (LASSO)

Least Absolute Shrinkage and Selection Operator (LASSO) [58] is a regression analysis method which changes coefficient estimation and makes some of them zero in order to perform feature selection. LASSO exploits l_1 norm regularisation for least square linear regression, and it attempts to minimise the following objective function:

$$LASSO = \underset{\beta}{argmin} \|y - \beta X\|_2^2 + \lambda \|\beta\|_1 \tag{2.14}$$

where the response random variable $Y \in R$ is dependent on a d-dimensional covariate $X \in R^d$ and the training data $D = (x_i, y_i)_1^n$ is independently and identically sampled from a fixed joint distribution P_{XY}, and λ is a regularisation parameter. The l_1 norm regularisation shrinks most of the coefficients toward zero; in other words, it performs feature selection [59]. The l_1 norm can be defined as the sum of the absolute values of components of the vector which can be calculated from:

$$\|\beta\|_1 = \sum_{i=1}^{n} |\beta_i| \tag{2.15}$$

The LASSO method is commonly used for genomics [60] [61].

LASSO is an embedded and supervised feature selection method. Even though LASSO is extremely useful for small n (n is number of samples), and large p (p is number of features) problems, it can select at most n features [60].

2.4.3.2 High-Dimensional Feature Selection by Feature-Wise Kernelized Lasso (HSIC LASSO)

The LASSO assumes that a linear correlation between features and the target exists. High-Dimensional Feature Selection by Feature-Wise Kernelized Lasso (HSIC LASSO) [62] can be considered as a non-linear form of LASSO. The HSIC LASSO attempts to solve the following optimisation problem [63]:

$$\begin{aligned} &HSIC_{LASSO} = \min_{\alpha \in R^d} \frac{1}{2}\|\overline{L} - \sum_{k=1}^{d} \alpha_k \overline{K}^{(k)}\|_{Frob}^2 + \lambda\|\alpha\|_1 \\ &subject\ to\ \alpha_1, \alpha_2, \ldots, \alpha_d \geq 0 \end{aligned} \tag{2.16}$$

where d represents the number of features, $\|.\|_{Frob}$ is the Frobenius norm, $\overline{K}^{(k)} = \Gamma K \Gamma$, $\overline{L} = \Gamma L \Gamma$ are centred Gram matrices, $K_{i,j}^k = K(x_{k,i}, x_{k,j})$ and $L_{i,j} = L(y_i, y_j)$ are Gram matrices, $K(x, x')$ and $L(y, y')$ are kernel functions, $\Gamma = I_n - \frac{1}{n} 1_n 1_n^T$ is the centring matrix, I_n is the n-dimensional identity matrix, λ is a regularisation parameter, α is regression coefficient vector, and 1_n is the n-dimensional vector with all ones. HSIC LASSO is a sparse based, embedded and supervised feature selection method. It is utilised to select features from high dimensional data sets to perform non-linear regression tasks.

In [64], LASSO is compared with 8 different feature selection method including LSMI and mRmR. It achieved the second-best performance over 23 low dimensional data sets (the highest number of features for a data set were 617). This results appears to suggest that LASSO works well on low dimensional data.

2.4.3.3 Least Angle Regression for Feature Selection (LARS)

Least Angle Regression for Feature Selection (LARS) [61] is similar to LASSO and it can be considered as a stepwise variant of LASSO. The LARS selects a predictor, x_i, from a data set which has the largest absolute correlation with the target (y), and then it performs simple linear regression of y on x_i. Other predictors which are orthogonal to the x_i are selected, and then a linear model is constructed by exploiting the selected predictors. The LARS performs linear regression with l_1 norm regularisation; therefore, most of the coefficients are shrunk toward zero, and thereby feature selection is accomplished.

The initial coefficients are zero ($\beta = 0$). The LARS increases these coefficients so that their features have the highest correlation with the output variable in each iteration till all coefficients become non-zero.

2.4.3.4 GUIDE

GUIDE [65] is a regression tree algorithm which aims to provide an unbiased feature selection using the Chi-Squared test. The GUIDE starts with the selection of the most important feature by exploiting the Chi-Square statistic. If none of the feature is considered significant; then, linear combinations of two features are determined and the most significant feature is decided by using Bonferroni [66] corrections. If the most significant variable is still not found, then interaction tests between pairs of features with Bonferronni corrections are performed. If the most important feature is still not found, then the feature with lowest p-value is selected at the beginning stage (Chi-Square test). Then, the split points on the most important feature, which decreases miss-prediction error, are found. Splitting continues until pre-defined number of observations exceed cases of a node [67]. GUIDE is a statistical based, embedded, and supervised feature selection algorithm which can be exploited for both classification and regression tasks.

2.4.3.5 Minimum Redundancy Spectral Feature Selection (MRSF)

SPEC (which is presented in section 2.3.1.4) ignores feature relevance, therefore, it cannot handle feature redundancy. MRSF [68] can be considered as an extension of SPEC where features are jointly evaluated to identify feature relevance.

The MRSF is a sparse learning based, embedded, and supervised feature selection method that evaluates a set features jointly and eliminates redundant ones. The MRSF attempts to solve the following optimisation problem:

$$\underset{W}{argmin}\|W'X - Y\|_2^2 + \lambda\|W\|_{2,1} \tag{2.17}$$

where $W \in R^{d*q}$ is a projection matrix, ϵ is a predefined parameter, and $Y \in R^{n*q}$ is embedding of the input data (by eigen decomposition) $X \in R^{d*n}$.

In [68], MRFS produced better results than HSIC and mRmR over six different data sets for classification tasks. These benchmarks have at most 11340 features. The performance of MRFS for regression tasks needs to be investigated.

2.4.3.6 Elastic Net (EN)

The LASSO penalises l_1 norm regularisation to shrink many coefficients to exactly '0'; therefore, LASSO can be utilised for feature selection. However, LASSO tends to select only one of the highly correlated features, which may not always be the best choice [69]. In order to select features with high correlations, Zhu and Hastie proposed Elastic Net (EN) [70] which uses both l_1 and l_2 norm regularisation given by:

$$penalty(w) = \sum_{i=1}^{n} |w_i|^{\gamma} + (\sum_{i=1}^{n} w_i^2)^{\lambda} \tag{2.18}$$

where $0 \leq \gamma \leq 1$ and $\lambda \geq 1$ are individual tuning parameters. The EN is a sparsity-based feature selection method that performs feature selection and regression, simultaneously.

In [15], EN and CMI_{DIST} methods are compared. EN produced better results than CMI_{DIST} if the number of features are less than 50. This result suggests that EN is suitable for very low dimensional data sets.

2.4.4 Remarks on the Use of Supervised Feature Selection

In this chapter, existing feature selection methods for regression problems are reviewed. In this section, unsupervised and supervised feature selection methods are compared and their advantages and disadvantages are presented. In Table 2.2, supervised and unsupervised feature selection methods are compared, in addition, their advantages and disadvantages are listed along with their references.

2.5 Unsupervised Feature Selection Methods

In previous section supervised feature selection methods for regression problems are presented. In this section, unsupervised feature selection methods for regression problems are presented.

2.5.1 Filter Methods

This subsection presents unsupervised filter feature selection methods for regression tasks.

2.5.1.1 Term Variance (TV)

Term Variance (TV) [82] is an unsupervised and univariate filter feature selection method that ranks features according to their variance. TV can be formulated as:

$$TV_i = var(x_i) = \frac{1}{n}\sum_{j=1}^{n}(x_{ij} - \overline{x_i}) \tag{2.19}$$

where $\overline{x_i}$ is sample mean of x_i. Even though TV is a simple method, it is computationally faster. Therefore, it can be applied to very high dimensional data.

TABLE 2.2: A Comparison of Supervised and Unsupervised Feature Selection Methods

Feature Selection		
Supervised	**Unsupervised**	**Ref**
The output weights are trained by solving a regularised least squares problem	The output weights are obtained by solving a generalised eigen value problem	[71]
A relevance of a feature is measured according to its correlation with label information	The relevance of a feature is measured according to its ability in preserving some data characteristics (e.g. variance)	[72]
Background knowledge of data is available	Background knowledge of data is not available	[72]
Traces for learning activities and strategies	Can be applied on any data set comprising traces of activities	[73]
Complexity is low, requires one iteration	Complexity is high requires several iterations	[73]
Requires human (expert) intervention to obtain labels	Does not require human (expert) intervention to obtain true labels	[13]
Limited data available	Adequate data available	[74]
Impractical	Practical	[74]
Not easily applicable for crowd sourcing	Easily applicable for crowd sourcing	[74]
Present labels	Abcent labels	[75] [76]
Applicable for classification usually	Applicable for classification, regression and clustering	[77]
Effective for selecting discriminative features	Effective for clustering features	[78]
Higher accuracy	Less accuracy	[79]
More reliable performance	Less reliable performance	[79]
Ignore correlation between different features	Ignore correlation between features and labels	[80]
Less challenging when applied to high dimensional data	More challenging when applied to high dimensional data	[81]
Time consuming and costly	Computational time greatly reduced	[77]
Difficult to apply for text classification, fault diagnosis, and information retrieval	Easily applicable for information retrieval, fault diagnosis, and text classification	[71]

2.5.1.2 Infinite Feature Selection (InFS)

Infinite Feature Selection (InFS) [4] is a filter and unsupervised feature selection method. In InFS, each feature is represented with a node in a graph and features are selected according to their centrality score. All possible subsets of features are considered as paths on a graph and each feature is ranked. The pseudo code for infinite feature selection method is illustrated in Fig. 2.3.

Input: $F = \{f^{(1)}, ..., f^{(n)}\}$, α
Output: $\check{s}$ energy scores, for each feature
  Building the graph
  for $i = 1 : n$ **do**
    for $j = 1 : n$ **do**
      $\sigma_{ij} = max(std(f^{(i)}), std(f^{(j)}))$
      $c_{ij} = 1 - |Spearman(f^{(i)}, f^{(j)})|$
      $A(i,j) = \alpha\sigma_{ij} + (1-\alpha)c_{ij}$
    end for
  end for
  Letting paths tend to infinite
  $r = \frac{0.9}{\rho(A)}$
  $\check{S} = (\mathbf{I} - rA)^{-1} - \mathbf{I}$
  $\check{s} = \check{S}\,\mathbf{e}$
  return $\check{s}$

FIGURE 2.3: The Pseudo Code for InFS Algorithm [4].

2.5.1.3 Laplacian Score Feature Selection (LapFS)

Laplacian Score Feature Selection (LapFS) [83] is a graph based, unsupervised and univariate filter feature selection algorithm that ranks features according to their locality preserving power. In Laplacian Score, features are evaluated independently; therefore, the LapFS algorithm cannot handle feature redundancy [84]. LapFS utilises pairwise similarities between features which are calculated using the heat kernel. Laplacian score of a feature, f_i, can be calculated from the following formula:

$$Lap(f_i) = \frac{\widetilde{f}_i' L f_i'}{\widetilde{f}_i' D f_i'} \tag{2.20}$$

where $f_i = f_i - \frac{f_i' D 1}{1' D 1} 1$, $1 = [1, 1, ..., 1]'$, D is degree or diagonal matrix defined as $D(i,i) = \sum_{j=1}^{n} S(i,j)$, S is affinity matrix $S(i,j) = \frac{e^{-\|x_i - x_j\|^2}}{t}$ and the Laplacian matrix (L) is $L = D - S$. Keep in mind that constructing a Laplacian graph

is computationally expensive, especially, if the number of features are extremely large.

2.5.1.4 Spectral Regression Feature Selection (SPEC)

Spectral Regression Feature Selection (SPEC) [85] can be considered as an extension of LapFS. LapFS is an unsupervised feature selection method which exploits data variance and separability to assess feature relevance [86]. The goal of the SPEC is to investigate some intrinsic properties of both supervised and unsupervised feature selection and to develop a unified framework which is built on spectral graph theory. Likewise LapFS, SPEC cannot handle feature redundancy because it evaluates each feature independently. Therefore, in SPEC, the correlation between features is not taken into account. SPEC exploits the Radial Basis Function (RBF) in order to calculate the similarity, s_{ij}, between two points x_i and x_j by:

$$S_{ij} = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}} \tag{2.21}$$

where the graph G is constructed from S, and the projection matrix (W) is constructed from graph G, and the degree matrix (D) is a diagonal matrix that can be calculated from $D_{ii} = \sum_{i=1}^{n} W_{ij}$. Given W, and D, the Laplacian Matrix (L) can be expressed as:

$$L = D - W; \quad L = D^{-1/2} L D^{-1/2} \tag{2.22}$$

As far as the feature selection for classification problems is concerned, SPFS has been shown to be an average method compared to others [87]. On the other hand, SPEC has shown its effectiveness for regression tasks in several studies [88] [89].

2.5.1.5 Trace Ratio Criterion for Feature Selection

Trace ratio feature selection [90] individually ranks features according to their scores which are computed in trace ratio norm. Two affinity matrices are defined by trace ratio criterion: S_w and S_b. They represent within class similarity, and

between class data similarity respectively. Their corresponding graph Laplacian and diagonal matrices can be calculated from the following formula:

$$\begin{aligned} D_w(i,i) &= \sum_{j=1}^{n} S_w(i,j) \\ D_b(i,i) &= \sum_{j=1}^{n} S_b(i,j) \end{aligned} \tag{2.23}$$

and

$$\begin{aligned} L_w &= D_w - S_w \\ L_b &= D_b - S_b \end{aligned}$$

where k is the number of features to be selected, $W = [W_{i1}, W_{i2}, \ldots, W_{ik}] \in R^{d*k}$ is the selection indicator matrix such that only ith element of w_{ij} is 1 and the others are 0. The trace ratio criterion of best selection matrix, W, can be calculated from [91]:

$$Trace\_ratio(W) = \underset{W}{argmax} = \frac{tr(W'X'L_bXW)}{tr(W'X'L_wXW)} \tag{2.24}$$

Trace Ratio is a similarity based, supervised and filter feature selection method that can be utilised for both classification (including multi-class classification) and regression tasks [2].

2.5.1.6 KCEN

KCEN [14] is a K-means clustering based unsupervised feature selection method where the number of clusters equals the number of selected features.

Given a data set $X = x_1, \ldots, x_j, \ldots, x_n$ in which $x_j = (x_{j1}, \ldots, x_{jd})^T \in R^d$, K-Means algorithm attempts to find K clusters of X, $C = C_1, \ldots, C_j, \ldots, C_k$, such that

$$\begin{aligned} &C_i \neq \emptyset, i = 1, \ldots, k \\ &\cup_{i=1}^{k} C_i = X \\ &C_i \cap C_j = \emptyset, \quad i, j = 1, \ldots, k \quad and \quad i \neq j \end{aligned} \tag{2.25}$$

where k is a user-defined integer. It is shown above that a pattern can only be allowed to belong one cluster. After determination of the cluster centroids, a feature which is the closest to the cluster centroid is selected as a representative feature for the cluster. Therefore, the number of clusters determines the number of selected features in KCEN algorithm. KCEN is a univariate filter, unsupervised, and statistical-based feature selection method.

KCEN method is effective and simple it produced comparable results on different high dimensional data sets [88].

Recently, there is no wrapper unsupervised feature selection method proposed in the literature Therefore, next sub-section presents embedded methods.

2.5.2 Embedded Methods

This subsection presents unsupervised embedded feature selection methods for regression problems.

2.5.2.1 Multi-Cluster Feature Selection (MCFS)

Multi-Cluster Feature Selection (MCFS) [92] is an unsupervised and embedded feature selection algorithm that selects a set of features by utilising spectral regression and l_1 norm regularisation. The correlation between features are evaluated using spectral analysis. MCFS consists of three main steps. First step is spectral clustering that is utilised to disclose cluster structure of the input data. Second step is sparse coefficient learning, and the final step is feature selection. MCFS exploits the eigen vectors of the graph Laplacian to appropriately cluster samples in an unsupervised manner. In order to create a graph of samples and to reveal local structure of a data, the k-nearest neighbour (KNN) method is exploited, and thereby a similarity matrix is gained. The Heat kernel affinity or similarity matrix, S_{ij} can be expressed as:

$$S_{ij} = e^{-\frac{\|x_i - x_j\|^2}{\sigma}} \tag{2.26}$$

where x_i and x_j are connected samples in KNN graph, and σ is a pre-defined parameter. Laplacian matrix is calculated from $L = D - S$ where D is a diagonal matrix. MCFS utilises l_1 norm and spectral regression and minimises the following function:

$$L_z = \lambda D_z \tag{2.27}$$

where $Z = [z_1, z_2, \ldots, z_t]$ denotes eigen vectors and t is a predefined parameter. A subset of relevant features can be found by minimising the following function:

$$\min_{w_i} \|xw_i - e_i\|_2^2 + \alpha\|w_i\|_1 \tag{2.28}$$

where w_i is the feature coefficient vector for the i-th embedding. MCFS solves sparse regression problems and gets t sparse feature coefficients, $W = [W_1, W_2, \ldots, W_t]$, and each coefficient corresponds to one embedding of data. MCFS ranks features based on their score which can be calculated from:

$$MCFS\_score(j) = \max_i |W_{i,j}| \tag{2.29}$$

where $W_{i,j}$ is the j-th element of vector W.

2.5.2.2 Unsupervised Discriminative Feature Selection (UDFS)

UDFS [93] is a sparse learning based, embedded, and unsupervised feature selection method that jointly utilises local discriminative information and feature correlations to select features. UDFS attempts to solve the following objective function:

$$\min_{W'W=I} tr(W'XLX'W) + \beta\|W\|_{21} \tag{2.30}$$

where I is the identity matrix, W is a projection matrix, $L = D^{-1/2}(D-S)D^{-1/2}$, X is the input data, tr is trace, $\|\|_{21}$ indicates l_{21} norm regularisation and $W^{'}$ is the transpose matrix of W.

2.5.2.3 Non Negative Discriminative Feature Selection (NDFS)

NDFS [94] is an unsupervised and embedded feature selection algorithm that performs feature selection and spectral clustering, simultaneously. Similar to the UDFS, NDFS exploits l_{2-1} norm regularisation to eliminate irrelevant features. NDFS aims to solve the following objective function:

$$\begin{aligned} &\underset{G,W}{min} \quad tr(G^T LG) + \beta\|XW - G\|_F^2 + \alpha\|W\|_{2,1} \\ &subject \quad to \quad GG^T = I_n, G \geq 0 \end{aligned} \tag{2.31}$$

where α and β are parameters, G is the weight cluster indicator matrix, X is the input data, and $L = D^{-1/2}(D - S)D^{-1/2}$ and S is the similarity matrix that can be computed from the Equation (2.32).

2.5.2.4 Robust Unsupervised Feature Selection (RUFS)

RUFS [95] is an unsupervised, sparse learning based and embedded feature selection algorithm that selects discriminative features by jointly performing robust feature selection and robust clustering. RUFS attempts to solve the following objective function:

$$\begin{aligned} &\underset{F,G,W}{min}\|X - GF\|_{2,1} + vTr[G^T LG] + \\ &\alpha\|XW - G\|_{2,1} + \beta\|W\|_{2,1} \\ &subject \quad to \quad G \in R_+^{nxc}, G = Y(Y^T Y)^{-1/2}, F \in R_+^{cxd} \end{aligned} \tag{2.32}$$

where v, α, $\beta \in R_+$ are user-defined parameters, G is the weight cluster indicator matrix, which represents pseudo class labels, X is the given input data, W is the projection matrix, L is the Laplacian matrix (which is presented in section 2.3.1.3), and F is the cluster centres in the original whole feature space.

2.5.2.5 Joint Embedding Learning and Sparse Regression (JELSR)

JELSR [87] is an unsupervised, sparse learning based and embedded feature selection technique that joins embedding learning with sparse regression to perform feature selection [87]. The method is quite similar to Multi Cluster Feature Selection (MCFS) and Minimum Redundancy Feature Selection (MRSF) methods; however, JELSR provides a new technique by applying local minimal approximation weights and l_{21} norm regularisation. JELSR attempts to solve the following optimisation function:

$$\begin{aligned} &\min_{WY} Tr(YLY') + \beta\|W'X - Y\|_2^2 + \alpha\|W\|_{21} \\ &subject\ to\ YY' = I \end{aligned} \tag{2.33}$$

where Y is the low dimension representation of the input, X, and W is the projection matrix, Tr is trace, and α and β are parameters.

2.5.2.6 Unsupervised Feature Selection with Adaptive Structure Learning (FSASL)

Unsupervised Feature Selection with Adaptive Structure Learning (FSASL) [96] is a sparse learning based, embedded and unsupervised feature selection method that jointly performs feature selection and structural learning. Unlike other embedded feature selection methods, such as MCFS, NDFS and JELSR, FSASL exploits the output of feature selection to feed into structure learning procedure in order to accomplish better structure learning. FSASL attempts to solve the following optimisation problem:

$$\begin{aligned} &\min_{W,S,P}(\|W'X - W'XS\|^2 + \alpha\|S\|_1) + (\sum_{i,j}^{n}\|W'x_i - W'x_j\|^2 P_{ij} + \mu P_{i,j}^2) + \gamma\|W\|_{21} \\ &subject\ to \quad S_{ii} = 0; \quad P1_n = 1_n; \quad P \geq 0; \quad W'XX'W = I \end{aligned} \tag{2.34}$$

where $W \in R^{d*c}$ is transformation matrix, γ is the regularisation parameter, $S \in R^{n*n}$ is the optimal sparse combination weight matrix which can be obtained from following function:

$$\begin{aligned} &\min_{S} \sum_{i=1}^{n} \|x_i - xS_i\|^2 + \alpha \|S_i\|_1 \\ &subject\ to\ Sii = 0 \end{aligned} \tag{2.35}$$

where α is utilised to balance sparsity and reconstruction error. $P_{ij} \in R^{n*n}$ is probabilistic neighborhood matrix and it can be calculated from the following formula:

$$\begin{aligned} &\min_{P} \sum_{i,j}^{n} \|x_i - x_j\|_2^2 P_{ij} + \mu P_{ij}^2 \\ &subject\ to\ P1_n = 1_n; \quad P \geq 0 \end{aligned} \tag{2.36}$$

where μ is the regularisation parameter and 1_n is the n-dimensional vector with all ones.

2.5.2.7 Embedded Unsupervised Feature Selection(EUFS)

NDFS, RUFS, and MCFS use clustering algorithms to disclose discriminative information from a data, and generate the cluster labels. They select features using the labels as if the selection method is supervised. Unlike these methods, EUFS [78] embeds feature selection into a clustering algorithm via sparse learning without transformation. EUFS aims to solve the following optimisation problem:

$$\begin{aligned} &\min_{U,V} \|X - UV^T\|_{2,1} + \alpha \|V\|_{2,1} + \beta Tr(U^T L U) \\ &subject \quad to \quad U^T U = I, U \geq 0 \end{aligned} \tag{2.37}$$

where l_{12} norm is applied to the cost function in order to decrease the impact of outliers and noise, α and β are user-defined parameters in which $\alpha, \beta \geq 0$, U

is the cluster indicator, V is latent feature matrix, Tr is trace, I is the identity matrix and L is the Laplacian matrix (as presented in section 2.3.1.3).

Even if EUFS is an embedded feature selection method, it is computationally inexpensive; therefore, it can easily be applied to high dimensional or ultra high dimensional data.

2.5.2.8 Unsupervised Feature Selection Using Feature Similarity (FSFS)

FSFS [97] is a similarity based, unsupervised and filter feature selection method that groups features into clusters using pairwise similarities between features, and then, selects the most representative feature from each cluster [98]. The FSFS exploits feature dependency/similarity to eliminate redundant features. The Maximal information compression index similarity measure [99] is used for clustering. In [97], the author used the well-known correlation coefficient:

$$\rho(X_i, X_j) = \frac{1/n \sum_{k=1}^{n}(X_{ik} - \overline{X_i})(X_{jk} - \overline{X_j})}{\sqrt{Var(X_i)Var(X_j)}} \tag{2.38}$$

where $\rho(X_i, X_j) = 1$ means strongly correlated and $\rho(X_i, X_j) = 0$ means uncorrelated. The author proposed maximal information compression index, MICI, that can be computed from the following formula:

$$2\lambda(X_i, X_j) = Var(X_i) + Var(X_j) - \sqrt{(Var(X_i) + Var(X_j))^2 - 4Var(X_i)Var(X_j)(1 - \rho(X_i, X_j))} \tag{2.39}$$

where λ is a parameter, $\rho(X_i, X_j)$ is the correlation coefficient which is aforementioned, Var represents the variance, X is the input data, and n is the number of samples.

2.6 A Taxonomy of Feature Selection Methods for Regression

In this section, a taxonomy of feature selection methods for regression problems is provided.

In section 2.1, feature selection methods are categorised as filters, wrappers and embedded methods. Filter methods are sub-categorised as univariate and multivariate methods. Furthermore, based on the availability of information and problem definition in prediction, feature selection methods can be also divided into two main categories: supervised and unsupervised.

In this section, a taxonomy of feature selection methods for regression problems is presented. Feature selection methods are not only categorised based on their types, but also they are categorised based on their intramural learning style, such as information based, similarity based, statistical based, or sparse learning based feature selection methods. To the best of our knowledge, this is the first comprehensive taxonomy for feature selection methods particularly in regression domain. This taxonomy of feature selection methods for regression problems is shown in Fig. 2.4.

In addition to providing a taxonomy, a comprehensive overview of feature selection methods for regression problems is also provided where feature selection methods are shown along with their types, references, sources, and code repositories. This comprehensive overview of feature selection methods for regression tasks is presented in Table 2.3.

2.7 Summary

In this chapter, an in-depth literature review of feature selection methods for regression problems has been proposed. There are three different types of feature selection methods: filter, wrapper and embedded methods. Filter methods are computationally faster, yet they do not interact with the prediction algorithm. Wrapper methods are computationally expensive; however, they produce better prediction performance than filters since they interact with a prediction algorithm. Likewise wrapper methods, embedded methods are also dependent on a learning algorithm, and therefore they produce better prediction performance than filters. On the other hand, embedded methods are computationally less expensive than wrappers, and more expensive than filters. Filter selection methods can be sub-divided into univariate and multivariate filters. Univariate filters assess the importance of each feature individually whereas multivariate filters determine this in the context of other features. Based on the availability

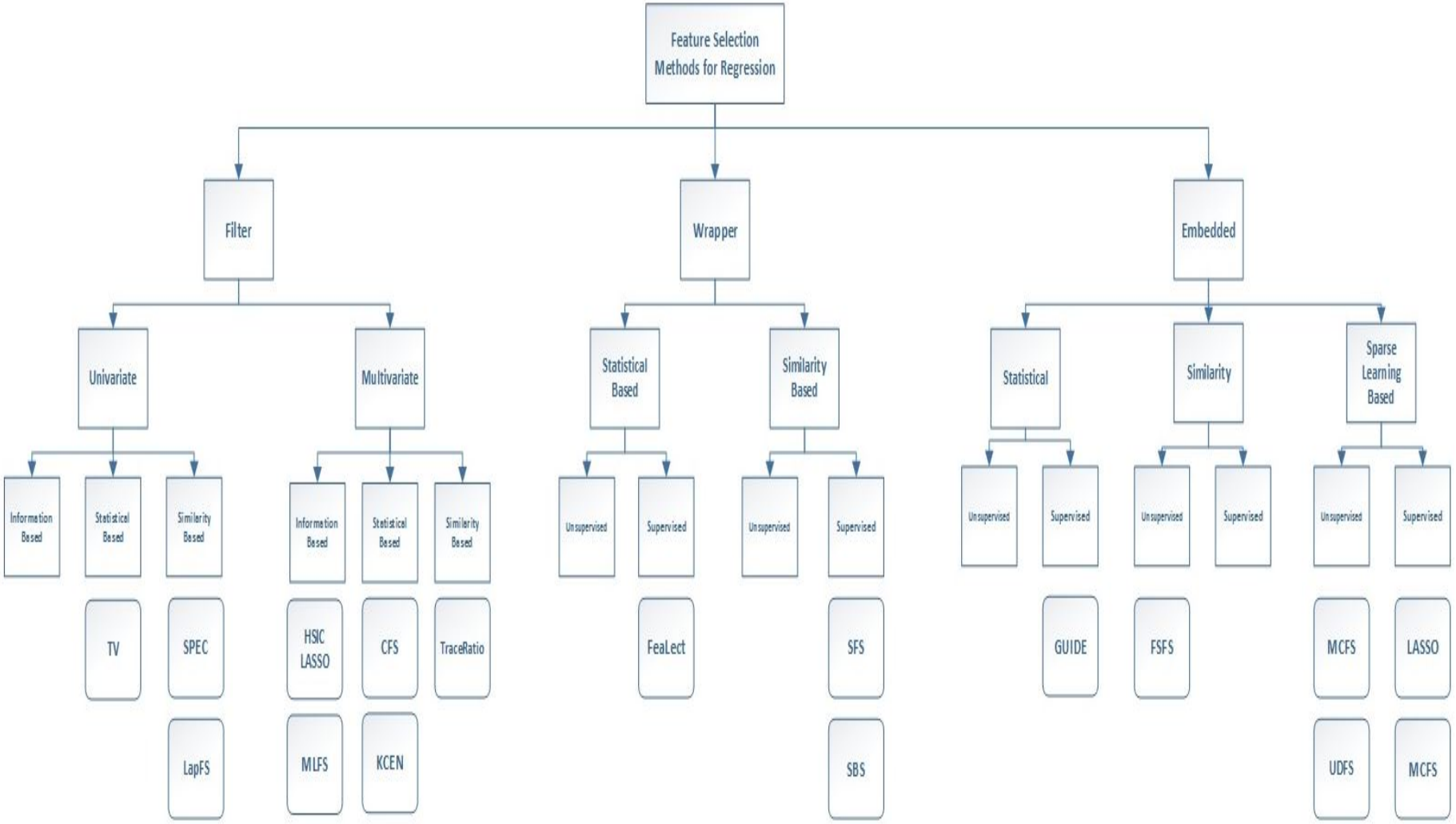

FIGURE 2.4: A Taxonomy Feature Selection Methods for Regression Problems

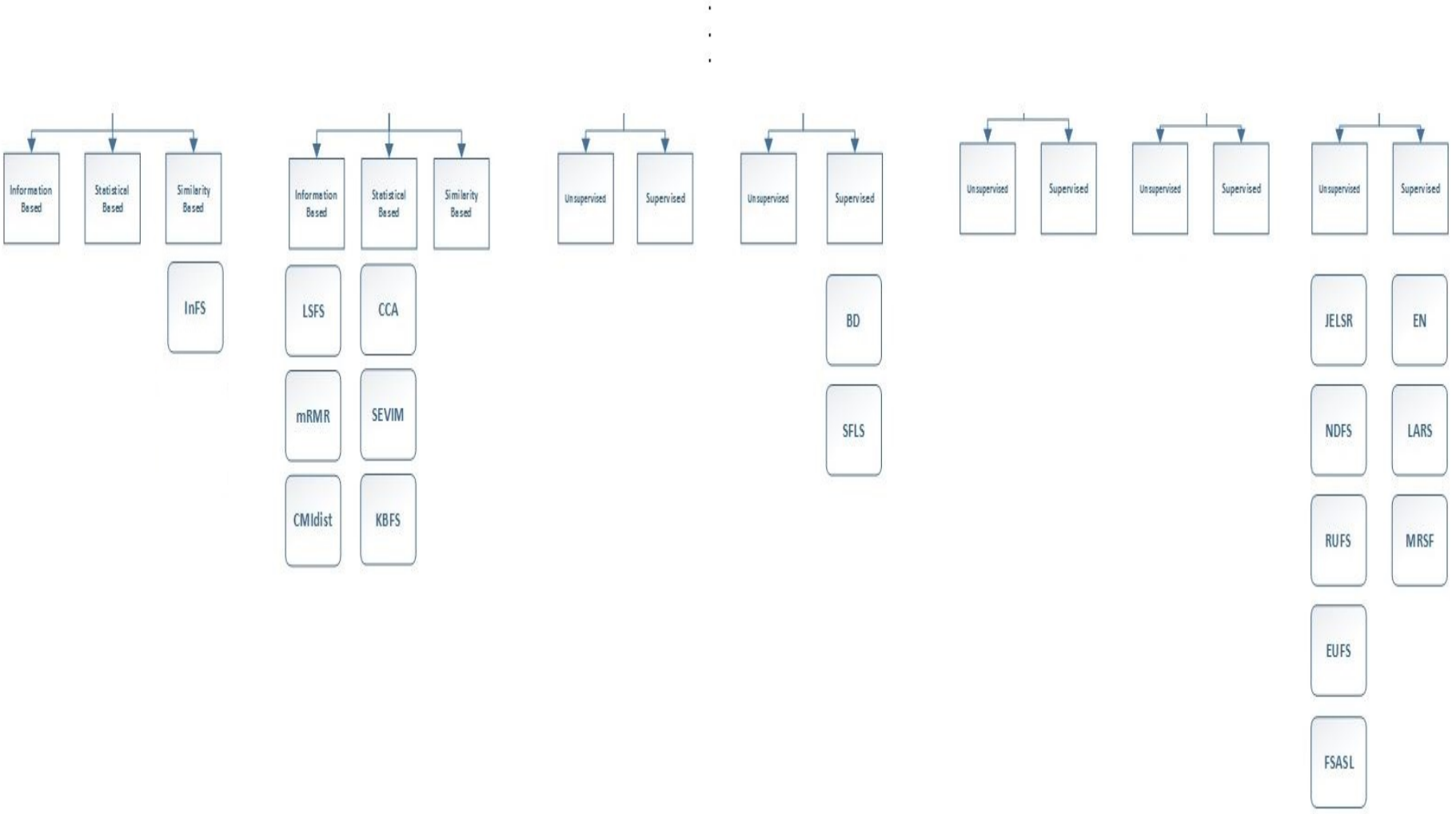

Figure 2.4 A Taxonomy Feature Selection Methods for Regression Problems (continued)

TABLE 2.3: A List of Feature Selection Methods for Regression Problems

Method	Type	SubType	Class	Reference	Code
SEVIM	Supervised	Embedded	Statistical	[46]	Matlab
MCFS	Both	Embedded	Sparse Learning	[92]	Matlab
KCEN	Unsupervised	Filter	Statistical	[14]	Matlab
EUFS	Unsupervised	Embedded	Sparse Learning	[78]	Matlab
RUFS	Unsupervised	Embedded	Sparse Learning	[95]	Matlab
UDFS	Unsupervised	Embedded	Sparse Learning	[93]	Matlab
MLFS	Supervised	Filter	Information	[42]	Matlab
NDFS	Unsupervised	Embedded	Sparse Learning	[94]	Matlab
FSASL	Unsupervised	Embedded	Sparse Learning	[96]	Matlab
JELSR	Unsupervised	Embedded	Sparse Learning	[87]	Matlab
SPEC	Both	Filter	Similarity	[85]	Matlab
FSFS	Unsupervised	Embedded	Similarity	[97]	Matlab
LapFS	Unsupervised	Filter	Similarity	[83]	Matlab
KBFS	Unsupervised	Filter	Statistical	[100]	Matlab
InFS	Unsupervised	Filter	Similarity	[4]	Matlab
LASSO	Supervised	Embedded	Sparse Learning	[58]	Matlab
LARS	Supervised	Embedded	Sparse Learning	[61]	Matlab
LSFS	Supervised	Filter	Information	[44]	Matlab
SFS	Supervised	Wrapper	Similarity	[101]	Matlab
SBS	Supervised	Wrapper	Similarity	[101]	Matlab
BD	Supervised	Wrapper	Similarity	[101]	Matlab
SFLS	Supervised	Wrapper	Similarity	[101]	Matlab
MRSF	Supervised	Embedded	Sparse Learning	[68]	Matlab
Trace Ratio	Supervised	Filter	Similarity	[90]	Matlab
EN	Supervised	Embedded	Sparse Learning	[70]	Matlab
CMI_{DIST}	Supervised	Filter	Information	[15]	Matlab
FeaLect	Supervised	Wrapper	Statistical	[56]	R
CFS	Supervised	Filter	Statistical	[?]	Matlab
TV	Unsupervised	Filter	Statistical	[82]	Matlab
CCA	Supervised	Filter	Statistical	[41]	Matlab
GUIDE	Supervised	Embedded	Statistical	[65]	Matlab
HSIC LASSO	Supervised	Filter	Information	[62]	Matlab
mRmR	Supervised	Filter	Information	[38]	Matlab

of information and problem definition in prediction, feature selection methods can be divided into two main categories: supervised and unsupervised feature selection. Supervised methods attempt to identify relevant features as well as noisy ones; on the other hand, unsupervised methods do not tend to select features which can act as noise. Consequently, compared to the supervised feature selection, unsupervised feature selection can be considered as a more unbiased approach. There have been a number of feature selection algorithms provided in the literature. They are used generally for classification, regression and clustering. Compared to the methods discussed for the classification, the literature appears to suggest that there is a lack of studies in regression-based problems for feature selection, in particular, unsupervised feature selection methods. In addition to providing a literature review of feature selection methods, a taxonomy of them, specifically for regression problems is also provided. In this taxonomy, feature selection methods are not only categorised according to their types, but also classified based on their intrinsic learning approaches.

Chapter 3

Regression Methods, Data Sets and Statistical Validation

In this chapter, the prediction methods, data sets and metrics for statistical validation are presented.

3.1 Prediction Methods

In this thesis, support vector-based models are used to evaluate the prediction performances of unsupervised feature selection methods since they have produced impressive generalisation and performance in wide variety of bioinformatics applications [102] [16]. Multi input-single output (MISO) and multi input-multi output (MIMO) prediction tasks are performed using Support Vector Regression (SVR) and multi support vector regression (MSVR) respectively.

3.1.1 Support Vector Regression

Support Vector Regression (SVR) aims to find a model function f(x) that shows the relationship between the features and the target. In SVR, the ϵ-intensive loss function is used [88]. In Fig. 3.1, the one-dimensional linear regression function with an epsilon intensive band is shown.

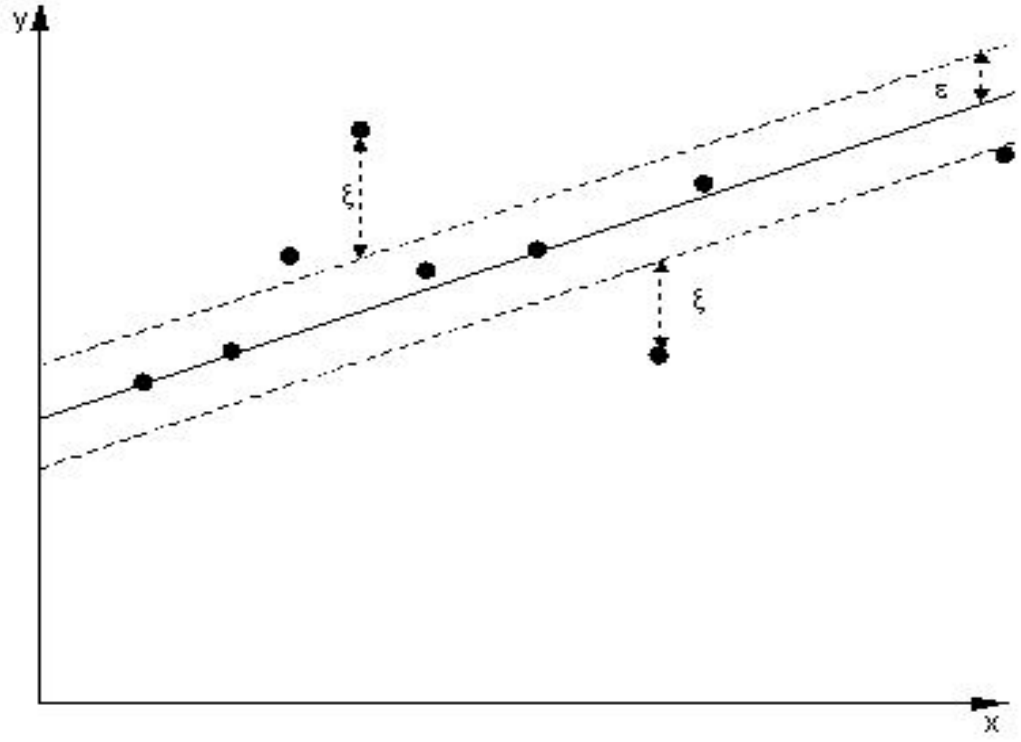

FIGURE 3.1: One Dimensional Linear Regression with Epsilon Intensive Band adapted from [5]

.

A margin tolerance (ϵ) is set to identify features to be normalised. Any residue of a regression less than ϵ is considered as noisy or meaningless. Only features out side of the ϵ-region are penalised, as given by:

$$C \sum_{i}^{n} \xi_i^k \tag{3.1}$$

where k is a positive integer and ξ is the orthogonal distance away from the ϵ-region. The regression function, f(x), is defined as:

$$f(x) =< w, x > \ + \ b \tag{3.2}$$

where w stands for a weight vector, and b is the bias. By minimising the weight vector and fixing the margin, the optimisation problem can be defined as:

$$\min_{\xi,\xi^*,w} \frac{1}{2}\|w\|^2 + C \sum \xi^2 + \xi^{2*} \tag{3.3}$$

This is subject to:

$$\begin{aligned} & y_i - < w, x_i > -b \leq \epsilon + \xi_i \\ & < w, x_i > b - y_i \leq \epsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0 \end{aligned} \tag{3.4}$$

where

$$X = (x_1, x_2, \ldots, x_n) \tag{3.5}$$

is the model, and

$$Y = (y_1, y_2, \ldots, y_n) \tag{3.6}$$

is the target.

As a certain number of training instances are selected as support vectors, the weighted sum of these support vectors is then obtained to develop a regression model. In this study, the SVR part of the process is implemented using the LIBSVM library [103].

3.1.2 Multi Support Vector Regression (MSVR)

Researchers have mainly paid attention to single-output regression analysis [15]. However, multi-output regression is crucial, especially in the analysis of biomedical data. The purpose of multi-output regression is to achieve a mapping of an input feature space into a multi-dimensional output space [104]. In this study, multi-output support vector regression (MSVR) [105] is exploited to perform the multi-output regression tasks. MSVR not only considers relationships among features, but also examines interrelationships among output variables.

The purpose of the uni-dimensional regression estimation problem is to find a model function which maps inputs ($x \in R^d$) to an observable output ($y \in R$). On the other hand, the multi-dimensional regression estimation problem aims to find a model function $\theta(x)$ that maps input variables ($x \in R^d$) to an observable vector output ($y \in R^t$) in which w^j and b^j (j=1,...,t) regressors need to be found for every target variable. Therefore, it attempts to solve the following function:

$$\min_{w^j, b^j, \xi_i} \sum_{j=1}^{k} \|w^j\|^2 + C \sum_{i=1}^{m} \xi \tag{3.7}$$

subject to

$$\begin{aligned} &\|y_i - w\theta(x_i) - b\| \leq \epsilon + \xi_i \qquad \forall i = 1, ..., m \\ &\xi_i \geq 0 \qquad\qquad\qquad\qquad\quad \forall i = 1, ..., m \end{aligned} \tag{3.8}$$

where $w = [w^1, ...w^k]^T$ and $b = [b^1, ...b^k]^T$ are k-dimensional linear regressors in t-dimensional Hilbert space, and m is the number of samples.

3.2 Data Sets

In this study, in order to evaluate the performance of proposed methods, four different case studies are considered: (i) a low dimensional RV144 vaccine data set; (ii) high dimensional peptide binding affinity data sets, which contain three different tasks; (iii) a very high dimensional GSE44763 data set; and (iv) a very high dimensional GSE40279 data set are exploited. The problem statement for all these data sets and their characteristics are presented.

3.2.1 RV144 HIV Vaccine

3.2.1.1 Problem Statement

Antibodies are specialised Y-shaped glycoproteins (gp) that are produced by plasma cells to defend against intruders that cause infection. Antibodies are crucial for the immune system since they play a role in protecting against foreign substances or antigens. Antibodies consist of two antigen-binding fragments: fragment antigen-binding (Fab) and fragment crystallizable (Fc). Fab regions are the arms of the antibodies called immunoglobulin G (IgG) which are responsible for the identification of infected cells [106]. On the other hand, Fc regions stimulate the innate immune system to neutralise antigens.

Antigens that exist in vaccines stimulate immune system response by instructing B-cells in order to produce antibodies which are responsible for protection. Vaccine-induced immunity effectors, or antibodies, are important defenders

against antigens, including HIV viruses. Vaccination provides active protection since it trains the immune system to recognise antigens. Then, the immune system produces specific antibodies to fight against the antigens. The function of antibodies is to recognise and bind to antigens. This detection process begins when antibodies recognise a small region on the surface of an antigen called the epitope [107]. Vaccine-mediated antibodies are important defenders against intruders including Human Immunodeficiency Virus (HIV) [108]. HIV attacks and destroys the immune system; indeed, it causes depletion of CD4-positive lymphocytes. The RNA of HIV has only nine genes that contain the code necessary to produce structural enzymes [109]. HIV poses a number of immunological threats to the human immune system due to its extensive genetic diversity. Furthermore, HIV is capable of developing countermeasures to avoid the effect of antibodies. HIV can prevent itself from being detected by the immune system thanks to its reverse transcription ability. This ability enables HIV to mutate approximately $3x10^5$ per nucleotide base [110].

Therefore, producing an effective vaccine which can elicit antibodies to block HIV is vital to neutralise the virus. Novel vaccine strategies are required to overcome the aforementioned challenges posed by HIV. Increasing the knowledge of associations between virus and immune system would ultimately result in producing an effective vaccine; an example is RV144. Functional antibodies are considered to be HIV inhibitors [111]. These inhibitory antibodies are capable of binding to virions, reducing their movement across mucus and mediating a variety of Fc receptor-mediated anti-HIV-1 activities, such as Antibody Dependent Cellular Cytotoxicity (ADCC) [112] [113]. ADCC-mediated antibodies can eradicate HIV infected CD4 cells [114] and block the transmission of HIV within 24 hours after viral entry [115]. HIV-1 transmissions commonly take place on mucosal surfaces; hence, mucosa is an excellent region to bind and engulf the virus. Antibody activities in mucosal tissues are shown in Fig. 3.2 [6].

Vaccination is a provider of active immunity since it stimulates the immune system to produce antibodies which fight against a virus. Interestingly, specific antibodies provide protection against specific antigens [116]. Moreover, the amounts of antibodies that are produced by the immune system are statistically related to the protection given, since antibodies will be needed for the subsequent attacks from antigens [14]. The functional characteristics of antibodies are also crucial for HIV protection; therefore, the identification of specific antibodies

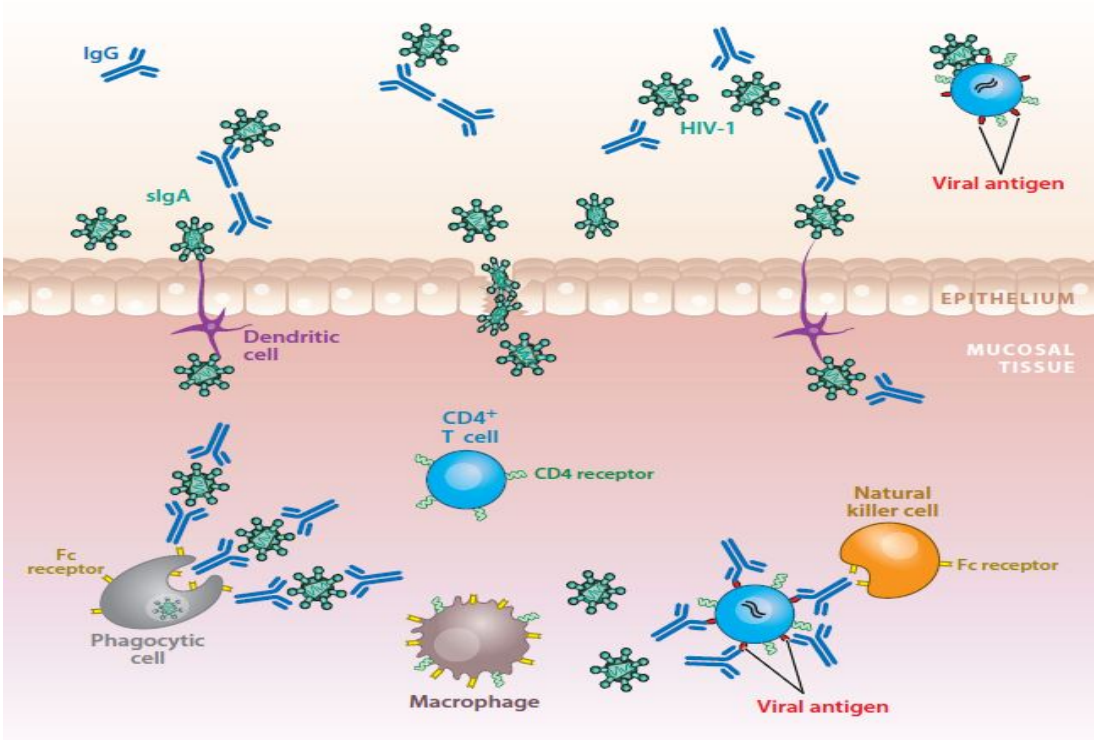

FIGURE 3.2: Antibody Activities on Mucosal Tissues [6]

that mediate effector functions to neutralise HIV is essential for producing an effective HIV vaccine. Antibodies can also collaborate with other functions to provide prevention against viruses [117] [118]. Therefore, the identification of discriminative antibody features is crucial in producing prevention against HIV.

3.2.1.2 The Data Set

The RV144 data set provided in [119] is utilised in this study to model their antibody feature-function relationships and to disclose HIV-specific antibodies. This data set contains 100 plasma samples (20 of them are placebo and 80 of them are vaccine injected) obtained from the individuals who participated in the RV144 vaccine trial at week 26. Three different cell-mediated assays are used in this thesis: Antibody Dependent Cellular Phagocytosis (ADCP), Antibody Dependent Cellular Cytotoxicity (ADCC), and Natural Killer cell Cytokine release.

ADCC antibodies are capable of identifying infected cells, and these antibodies are involved in the binding of epitopes of HIV-infected cells [120]. Cytotoxicity activities are mediated by Natural Killer (NK) cells which can directly kill virally infected cells by adhering to them. Cytokine release activity includes the NK

cell surface expression of CD107a and the quantitative detection of MIP-1-β and IFN-γ levels produced inside the cell. An antigen microsphere based liquid array is applied to determine antibodies (gp41, gp140, p24, gp120, and V1V2) and subclasses (IgG1, IgG2, IgG3, IgG4).

HIV-1 vaccine injection has been shown to be less effective due to the intrinsic variability of the virus. The identification of distinctive antibodies which correlate with protection against HIV-1 infection, along with increasing the knowledge of associations between immune mechanisms and HIV, would ultimately result in the development an effective vaccines against HIV. In this study, three different cell-mediated assays, ADCC, ADCP, and Cytokine release, are used. The purposes of exploiting the RV144 data set are: (i) to differentiate functional antibodies; (ii) to identify the relationships between the human immune system and the HIV virus; and (iii) to test the effectiveness of the DFSFR framework for the given data set. The aim of this study is, therefore, to predict functional relationships between antibody features and their functional activities in RV144 vaccine recipients. Each data sample has twenty antibody features that consist of features related to IgG subclass and antigen specificity.

3.2.2 Peptide Binding Affinity

3.2.2.1 Problem Statement

The understanding of interactions among proteins is an essential domain of research in systems biology, with applications in protein engineering and drug design. Proteins are bio-polymers that consist of chains of amino acid residues. Proteins play fundamental roles in cellular functions. For example, approximately, 50% of the cell dry weight of the human body is protein [121].

An amino acid is a small molecule that consists of an amine (NH_2) and carboxyl (COOH) functional groups with an organic substituent, called the R-group which determines the unique characteristics of each amino acid. A peptide is short linear chain of an amino acid sequence which contains two or more amino acids linked by peptide bonds. Polypeptides consist of a series of amino acid units and residues linked by peptide bonds [122]. A protein is a biological macromolecule composed of one or more polypeptides. Consequently, the building blocks of both peptide and protein molecules are amino acids, and thereby peptides can

be exploited as a secondary structure of proteins to modulate protein-protein interactions [123] [124] [125].

Protein-protein interactions (PPIs) play a role in mediating signal transactions, sensing the environment, triggering immunological responses, and monitoring gene expression [126]. Furthermore, PPIs play a crucial role in the progress of human diseases such as viral infections. Therefore, increasing knowledge of the underlying principles of PPIs can ultimately result in the revealing intrinsic biochemistry of different diseases and the development of drug design [127]. However, the empirical determination of PPIs is not amendable, and thereby, to increase the understanding of PPIs, computational methods such as classification, regression and feature selection can be used. The purpose of this study is, therefore, to predict binding affinity values for peptides using amino acid descriptors. Amino acid descriptors quantitatively describe the physicochemical properties of the peptides [128]. Affinity refers to the strength of binding. The difficulty of the peptide binding affinity prediction problems when building a prediction model is that the number of features is very large (in this study, around 5000) whereas the number of peptides in the training data set is relatively small (in this study <150).

3.2.2.2 The Data Sets

In this study, three different high-dimensional peptide data sets provided at the Comparative Evaluation of Prediction Algorithms CoEPrA modelling competition [129] are used in order to further improve the predictivity of the affinity of peptides and, in particular, to test the predictive capability of the proposed DFSFR framework for the given data sets. Each data set contains training and test data sets and physicochemical descriptors have been provided for each small peptide for both training and test data sets. Each amino acid in a peptide is described by 643 descriptors. Tasks 1 and 3 contain nona-peptides that have a total of 5787 descriptors (=643x9) whereas Task 2 consists of octa-peptides that are characterised using a total of 5144 descriptors (=643x8). The characteristics of the peptide binding affinity data sets are given in Table 3.1. A more detailed description of these data sets is provided in Appendix A.

TABLE 3.1: General Characteristics of the CoEPrA Data sets Used for the Prediction of Peptide Binding Affinity

Datasets	Number of Peptide Sequences		Number of Peptide Sequence Descriptors
	Training	Testing	
Task 1	89	88	5787
Task 2	76	76	5144
Task 3	133	133	5787

The number of amino acid occurrences in training and testing peptide affinity data sets for each experiment are shown in Tables 3.2-3.7. In other words, these tables demonstrate the distribution of amino acids which positioned at the peptide locations for each of the training and testing data sets of related tasks. Physico-chemical descriptors are provided for each peptide for both training and testing data sets. Each amino acid is described by 643 descriptors. For example, Proline (P) contributes greatly in Task 1 data set at locations 4 and 6, and Valine (V) strongly contributes in Task 1 data at location 9. Leucine (L) contributes weakly in the Task 1 training data set at location 2; nevertheless, it strongly contributes at position 2 in the Task 1 testing data. Therefore, prediction for Task 1 is quite difficult. In the Task 2 data set, Leucine (L), Isoleucine (I), Phenylalanine (F), Serine (S), Asparaigne (N), Glycine (G), Glutomic Acid (E), Threonine (T) amino acids appear approximately 60 times at their separate respective locations. Leucine (L) and Valine (V) make a considerable contributions to the Task 3 model at locations 2 and 9 respectively.

TABLE 3.2: Amino acid occurrences in Training Data Set for Task 1

	Location								
Amino Acid	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Alanine	1	2	2	0	0	0	1	2	14
Arginine	5	0	0	0	0	0	0	0	0
Asparagine	1	0	6	1	0	1	1	11	0
Aspartic Acid	0	0	29	4	0	2	1	2	1
Cysteine	1	1	2	1	0	1	1	2	0
Glutamine	0	0	1	10	4	2	2	3	0
Glutamic Acid	0	0	0	0	0	0	2	3	0
Glycine	3	0	1	6	16	1	1	1	2
Histidine	1	1	3	1	1	0	8	1	1
Isoleucine	3	2	3	0	4	1	2	1	5
Leucine	3	6	5	2	10	1	1	4	6
Lysine	2	0	1	2	0	0	0	0	1
Methionine	1	4	4	0	1	1	0	0	0
Phenylalanine	9	1	13	1	33	2	11	0	1
Proline	1	1	0	52	1	50	14	4	1
Serine	2	0	3	4	1	3	4	12	1
Threonine	0	7	1	3	5	6	1	39	3
Tryptophan	0	0	12	0	1	0	1	2	1
Tyrosine	2	1	3	0	3	14	1	1	1
Valine	3	1	0	2	9	4	37	1	51

TABLE 3.3: Amino acid occurrences in Testing Data Set for Task 1

	Location								
Amino Acid	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Alanine	3	0	4	1	1	1	5	2	13
Arginine	4	0	0	3	3	1	0	1	0
Asparagine	2	1	3	1	0	3	0	5	1
Aspartic acid	0	1	25	8	2	0	1	5	0
Cysteine	0	1	1	0	1	2	1	2	2
Glutamine	0	2	0	11	0	1	0	2	1
Glutamic acid	0	0	2	3	2	0	1	5	1
Glycine	3	1	3	1	16	2	1	4	0
Histidine	2	0	1	1	6	1	11	2	0
Isoleucine	29	4	2	1	6	4	3	4	6
Leucine	3	65	6	0	8	2	6	4	16
Lysine	2	0	3	0	0	0	0	0	0
Methionine	1	3	1	0	0	1	3	1	1
Phenylalanine	8	0	17	1	24	5	8	2	0
Proline	0	0	2	45	2	46	10	1	0
Serine	4	1	2	4	1	2	3	8	0
Threonine	3	5	2	4	0	3	2	39	1
Tryptophan	2	1	10	2	2	0	0	1	0
Tyrosine	19	0	3	1	5	10	1	0	0
Valine	3	3	1	1	9	4	32	0	46

TABLE 3.4: Amino acid occurrences in Training Data Set for Task 2

	Location							
Amino Acid	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Alanine	1	0	1	1	1	0	0	1
Arginine	0	0	1	1	0	1	0	1
Asparagine	2	0	1	0	2	66	1	9
Aspartic	1	1	1	1	0	1	2	1
Cysteine	0	0	0	0	0	0	1	0
Glutamine	2	1	1	0	0	0	1	1
Glutamic	0	67	0	1	0	1	2	0
Glycine	1	2	1	1	65	2	0	1
Histidine	1	0	1	0	0	0	1	1
Isoleucine	1	1	1	1	1	0	1	57
Leucine	1	1	2	1	1	0	64	0
Lysine	1	1	1	1	0	3	1	0
Methionine	1	0	0	0	0	0	0	1
Phenylalanine	60	1	2	1	1	0	1	0
Proline	1	0	0	1	1	0	1	0
Serine	1	0	63	1	1	0	0	1
Threonine	0	1	0	61	0	0	0	0
Tryptophan	1	0	0	1	1	1	0	1
Tyrosine	0	0	0	1	0	0	0	0
Valine	1	0	0	2	2	1	0	1

TABLE 3.5: Amino acid occurrences in Testing Data Set for Task 2

	Location							
Amino Acid	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Alanine	1	4	0	0	1	1	1	0
Arginine	1	0	0	0	1	0	1	0
Asparagine	0	1	0	1	0	59	0	10
Aspartic	1	0	0	0	1	0	0	0
Cysteine	0	0	0	0	0	0	0	0
Glutamine	0	0	0	1	1	1	0	0
Glutamic	1	62	1	0	1	1	0	0
Glycine	0	0	0	1	63	1	1	0
Histidine	1	1	1	2	1	1	0	0
Isoleucine	0	0	2	0	0	1	1	55
Leucine	1	1	0	1	0	2	64	2
Lysine	0	0	1	1	1	0	0	1
Methionine	0	1	1	1	1	2	1	0
Phenylalanine	68	0	2	0	1	2	0	1
Proline	0	1	1	2	0	1	1	1
Serine	0	1	63	1	2	1	1	0
Threonine	1	1	1	64	1	1	1	1
Tryptophan	0	1	1	1	0	0	1	0
Tyrosine	1	1	1	0	1	1	1	1
Valine	0	1	1	0	0	1	2	4

TABLE 3.6: Amino acid occurrences in Training Data Set for Task 3

	Location								
Amino Acid	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Alanine	10	3	15	6	16	14	17	12	22
Arginine	5	0	1	8	3	4	3	1	0
Asparagine	2	0	4	6	3	4	3	0	0
Aspartic	1	0	10	9	5	3	0	5	0
Cysteine	2	1	2	1	1	2	2	4	1
Glutamine	1	0	1	13	2	4	4	1	0
Glutamic	0	0	2	4	4	3	3	6	0
Glycine	10	0	10	15	19	9	1	9	0
Histidine	1	0	2	2	5	1	2	4	0
Isoleucine	14	13	6	4	5	6	11	5	15
Leucine	17	88	22	10	15	16	16	29	33
Lysine	2	0	0	6	1	1	0	1	0
Methionine	5	10	7	1	2	6	2	3	0
Phenylalanine	16	0	7	4	10	6	19	11	0
Proline	1	0	4	20	5	26	8	5	0
Serine	13	0	9	9	1	5	7	16	0
Threonine	5	9	5	8	6	8	6	12	2
Tryptophan	4	0	8	3	4	2	1	2	0
Tyrosine	19	0	12	1	5	1	7	4	0
Valine	5	9	6	3	21	12	21	3	60

TABLE 3.7: Amino acid occurrences in Testing Data Set for Task 3

	Location								
Amino Acid	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Alanine	17	6	17	8	17	6	16	19	27
Arginine	7	0	0	3	3	0	1	1	1
Asparagine	2	0	1	1	2	5	4	2	0
Aspartic	2	0	8	7	11	2	3	0	0
Cysteine	0	0	2	5	1	4	3	4	0
Glutamine	3	1	2	17	3	7	4	3	0
Glutamic	0	0	4	4	2	1	0	3	0
Glycine	10	0	4	23	21	8	3	9	0
Histidine	5	0	3	3	6	2	1	5	0
Isoleucine	16	4	6	1	4	5	4	6	14
Leucine	15	87	21	9	15	26	17	22	34
Lysine	4	0	2	5	1	1	3	1	0
Methionine	3	15	8	1	1	3	3	1	2
Phenylalanine	13	0	9	3	8	5	18	3	0
Proline	0	1	3	9	1	24	11	6	0
Serine	4	0	7	12	6	4	8	20	0
Threonine	1	7	4	6	4	11	8	13	2
Tryptophan	3	0	6	0	3	1	4	5	0
Tyrosine	16	0	18	3	4	5	3	2	0
Valine	12	12	8	13	20	13	19	8	53

3.2.3 Age and Obesity Prediction (The GSE44763 Data Set)

3.2.3.1 Problem Statement

The prediction of human age from epigenetic information can be used to identify human remains for forensic analysis, chronological age, and age-related diseases. Aging and obesity contribute to fatal diseases, including cancers and circulatory and respiratory disease. Recent studies have proven that CpG dinucleotides are associated with both aging and obesity [130].

The degree of methylation at CpG sites is linearly correlated with aging, which indicates that CpG dinucleotides are appropriate biomarkers to predict the chronological age of individuals [131]. Consequently, the identification of age-related CpG biomarkers is crucial in the prediction of chronological age.

Circulatory disease, cancers, and respiratory disease are three of the main causes of mortality [132] [133]. Obesity can increase the risk of these three fatal disease types as well as other diseases such as diabetes and depression. According to the World Health Organization (WHO), there were over 600 million obese people worldwide in 2014 [134]. Most of the time, the risks associated with obesity-related diseases also increase with aging [130]. Even though obesity has a heritable component, whole genome association studies have provided only a few genetic polymorphisms which are associated with obesity. Some genetic variants, such as LEP, LEPR, and POMC, contribute to obesity; however, these variants do not fully explain the heritability of obesity. Indeed, they only specify a portion of the heritability of obesity $(40 - 70\%)$ [135]. Therefore, other variants such as epigenetic changes, which are potentially heritable changes in gene expression, must be considered. Some studies indicate that the epigenetic profile can be used to differentiate between low and high responders to calorie or caloric restriction [133].

The most common epigenetic mark is DNA methylation, which can be related to obesity [136] [137]. Several studies have proven the association between DNA methylation and obesity [138] [130].

The goal of this study is, therefore, to reveal relationships among CpG dinucleotides, aging and obesity. In other words, the purpose of this study is to

TABLE 3.8: A description of participants in the lean and obese group

	Obese	Lean
Subjects	24	22
Age (years)	57 (42-70)	55 (41-69)
Weight (kg)	92 (78-108)	60 (40-75)
BMI (kg/m2)	35 (30-42)	22 (16-25)

disclose specific CpG biomarkers that are related to aging and obesity. However, in the GSE44763 data set, there are approximately 28000 CpG biomarkers (features) and 46 samples. It is clear that building a predictive model is problematic when the number of samples is profoundly less than the number of features.

3.2.3.2 The GSE44763 Data Set

The GSE44763 data set provided in [130] is utilised to model the associations among CpG biomarkers (features), chronological age and obesity. This data set contains 27482 Cytosine-phosphate-Guanine (CpG) biomarkers from the peripheral blood of 46 adult female donors (samples). There are 24 obese and 22 lean subjects. A person is considered to be obese if their BMI is greater than or equal to 30 kg/m^2 and a subject with less than 25 kg/m^2 is considered lean. A description of the participants in the lean and obese groups is shown in Table 3.8.

In this study, Ilumina average beta values are utilised as numerical data where the Beta-value is the ratio of the methylated probe intensity and the overall intensity (sum of methylated and unmethylated probe intensities). The Beta value for an ith investigated CpG island is determined as follows [139]:

$$Beta_i = \frac{max(y_{i,methy}, 0)}{max(y_{i,unmethy}, 0) + max(y_{i,unmethy}, 0) + \alpha} \tag{3.9}$$

where $y_{i,methy}$ and $y_{i,unmethy}$ are the intensities measured by the ith methylated and unmethylated probes respectively, and α is a constant offset which is added to the denominator in order to regularise the Beta value if unmethylated and methylated probe intensities are low. The default value of α is 100.

3.2.4 Age Prediction (The GSE40279 Data Set)

3.2.4.1 Problem Statement

The identification of the age of individuals from epigenetic biomarkers can reveal vital information for criminal investigation, disease prevention, and the extension of life. Changes in DNA methylation are strongly associated with chronological age and the process of disease development. Changes in DNA methylation are also one of the most important indicators of biological aging [140] [141] [142]. DNA methylation can be utilised to precisely predict the chronological age of individuals from blood samples [143]. It has recently been shown that the aging process is highly related to changes in DNA methylation patterns. Furthermore, DNA methylation marks have been associated with age-related diseases such as Alzheimer's disease, metabolic disease, and cancer [144].

The purpose of this study is to disclose associations between CpG biomarkers and chronological age. The difficulty of revealing important information from CpG biomarkers is that the numbers of CpG biomarkers are very large (in this study, approximately 500.000) while the number of samples are relatively small (in this study around 700).

3.2.4.2 The GSE40279 Data Set

The GSE40279 data set provided in [145] is utilised to model the relationship between CpG biomarkers (features) and chronological age. This data set contains 473034 Cytosine-phosphate-Guanine (CpG) biomarkers (features) from the whole blood of 656 donors (samples) aged 19 to 101.

A pre-processing step is applied to the GSE40279 data set to map the data into lower dimensional space so that it can be exploited by feature selection methods. First, the standard deviation of each sample, which refers to the amount of variation in the data samples, is calculated. A standard deviation of a data sample can be equal to zero, if and only if, the values of all of the samples are identical. If all of the sample for a feature are identical, then the feature is not a discriminative one. Therefore, a pre-prosessing step is performed to eliminate the features which have the lowest variation in the data samples. As a result, approximately four out of five of the features are eliminated in this pre-processing

step, and only 90000 CpG biomarkers (features) are exploited to perform feature selection.

A general overview of the characteristics of all data sets which are exploited in this study is presented in Table 3.9. The GSE40279 data set can be determined as high dimensional as far as classification is concerned; however, in the regression domain, the GSE40279 data set can be considered as ultra-high dimensional.

TABLE 3.9: A General Overview of all of the Data Sets Used in this Study

Datasets	Number of		Sections in which the Results are Provided	Description
	Features	Samples		
RV144	20	100	LD	
Task 1	5787	177	HD	
Task 2	5144	152	HD	
Task 3	5787	256	HD	
GSE44763	27482	46	Very HD	
GSE40279	473034	656	Ultra HD	

LD:Low dimensional, HD:High dimensional

3.3 Statistical Validation and Performance Evaluation Metrics

In this section, the model validation technique which is used to test the effectiveness of the proposed approaches, is presented. Then, statistical evaluation metrics which are exploited to assess the capability of the predictive models are presented.

3.3.1 Statistical Validation of the Results

Various error estimation and validation techniques are provided in the literature. In this study, in order to evaluate the effectiveness of the predictive models for unseen samples, the most common and popular error estimation method [146], cross validation (CV), is utilised.

The cross validation method splits the data into two sets: training and testing. The training part is used to train a model, and the testing set is exploited for evaluation. One of the advantages of CV is that it efficiently produces unbiased error estimate because its process is repeated for different samples drawn from a population; therefore, the average error estimates will approximate the expected error for the designed regressors across all possible equal-sized samples [147].

In this study, k-fold cross validation is used to evaluate the performance of the predictive models where k is an integer. Therefore, the set of size $\frac{k-1}{k}$ samples are exploited for training and the other set of size $\frac{1}{k}$ samples are used for testing. The error rate of CV, called E, can be considered as the average error rate on $\frac{1}{k}$ testing samples, called E_i. E can be expressed as:

$$E = \frac{1}{k}\sum_{i=1}^{k} E_i \tag{3.10}$$

3.3.2 Performance Evaluation Metrics

In this section, the performance evaluation metrics which are used to evaluate the effectiveness of unsupervised feature selection methods are presented.

3.3.2.1 Root Mean Square Error (RMSE)

The Root Mean Square Error (RMSE) [148] has been utilised as a standard statistical metric to evaluate the performance of models in different research areas [149]. It provides a complete picture of the distribution of error. The RMSE can be expressed as:

$$RMSE = \sqrt{\frac{\sum_i^n (y_i - y_i')^2}{n}} \tag{3.11}$$

where n is the number of samples, and y_i and $y_i^{'}$ are the expected and predicted output respectively.

3.3.2.2 Pearson Correlation Coefficient (PCC)

The Pearson Correlation Coefficient (PCC) is an evaluation metric that is utilised to assess the performance of predictive models. The PCC evaluates the strength of the relationship between two variables. It can be calculated as:

$$PCC = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\sqrt{\Sigma x_i^2 - (\Sigma {x_i}^2)}\sqrt{\Sigma y_i^2 - (\Sigma {y_i}^2)}} \tag{3.12}$$

where x and y are vales the two quantitative variables and PCC indicates the linear association between them. A value of PCC that is equal to 1 indicates a perfect linear correlation.

3.3.2.3 Theil's U Statistics

Theils U statistics [150] is an accuracy measure that evaluates the prediction performance of a model. It can be calculated using the following formula:

$$U = \frac{RMSE}{\sqrt{1/n\sum_{i}^{n} y_i^2}} \times \frac{1}{\sqrt{1/n\sum_{i}^{n} {y_i'}^2}} \tag{3.13}$$

where y and y' are actual and corresponding forecasted values respectively. The RMSE is calculated by using Eq.5.11. A value of U which is closer to 0 indicates greater prediction performance.

3.3.2.4 Mean Absolute Deviation (MAD)

The Mean Absolute Deviation, MAD, is an average estimator of the absolute error of the predictive model. The MAD can be calculated from the following formula:

$$MAD = \frac{\sum_i^n |y_i - y_i^{'}|}{n} \tag{3.14}$$

where y_i is the actual and y_i' is the predicted value and n represents the number of samples.

3.3.2.5 Mean Absolute Percentage Error (MAPE)

Mean Absolute Percentage Error, MAPE, estimates the average of absolute percentage error of the predictive model. The MAPE is formulated as:

$$MAPE = \frac{1}{n}\sum_{i}^{n}\frac{|y_i - y_i^{'}|}{|y_i|} * 100 \tag{3.15}$$

where y_i is the actual and y_i' is the predicted value and n represents the number of samples.

3.3.2.6 Coefficient of Determination (q^2)

The Coefficient of Determination ($q2$) is a statistical metric based on the proportion of variability in a data set. If the value of q^2 is close to 1, is means that a model has been successfully constructed; on the other hand, negative q^2 values suggest that a model ineffectively approximates the predicted values [151]. The q^2 metric can be calculated from the following formula:

$$q^2 = 1 - \frac{\sum_{i}^{n}(y_i - y_i^{'})^2}{\sum_{i}^{n}(y_i - \overline{y})^2} \tag{3.16}$$

where y and $y^{'}$ are actual and corresponding forecasted values respectively, n is the number of samples and $\overline{y}$ is the mean of all actual values in the prediction data set.

3.3.2.7 Mean Square Error (MSE)

The Mean Square Error (MSE) represents the average of predictive model estimation errors, therefore, it measures the prediction performance of the model. The MSE can be expressed as:

$$MSE = \frac{\sum_{i}^{n}(y_i - y_i^{'})^2}{n} \tag{3.17}$$

where n is the number of samples, and y_i and $y_i^{'}$ are the expected and the predicted values respectively. The MSE can also be calculated from the RMSE since RMSE = $\sqrt{\text{MSE}}$.

3.4 Summary

In this chapter, the prediction methods, data sets, and statistical validation and performance evaluation techniques which are used in this study to evaluate the performance of the proposed methods are presented. The MISO and MIMO regression tasks are performed using SVR and MSVR, respectively. The effectiveness of unsupervised feature selection methods, including the proposed methods, are tested with a total of six different data sets. The RV144 Vaccine data set consists of 100 plasma samples where 20 of which are placebo and 80 are vaccine-injected samples. Each data sample has twenty antibody features that consist of features related to IgG subclass and antigen specificity. The goal of exploiting this data set is to reveal the relationships between antibody features and their effector functions. The peptide binding affinity data sets consist of three different tasks where Tasks 1 and 3 contain nona-peptides which have a total of 5787 amino acid descriptors and Task 2 consists of octa-peptides with a total of 5144 amino acid descriptors. The goal of using this data set is to predict peptide binding affinity values by using the given amino acid descriptors. The GSE40279 data set contains 473034 CpG biomarkers (features) from the whole blood of 656 individuals (samples) aged 11 to 101. The goal of utilising this data set is to disclose age-related CpG dinucleotides (features) and reveal the associations between CpG dinucleotides (features) and chronological age. In this study, k-fold cross validation technique is utilised for model error estimation. In addition, eight different evaluation metrics, namely RMSE, MSE, MAPE, MAD, q^2, U, and PCC are exploited to assess prediction performances of the predictive models.

Chapter 4

K-Means Based Unsupervised Feature Selection

In this chapter, a K-means based unsupervised feature selection framework for regression problems is proposed. First, the K-means algorithm is described along with its advantages and disadvantages. Then, the proposed K-Means based unsupervised feature selection framework for particularly regression problems is presented. Next, existing K-means based feature selection methods are reviewed. Final section presents the results of the application of the proposed method compared to the state-of-the-art unsupervised feature selection techniques as well as the baseline (entire feature set) with the RV144 Vaccine, peptide binding affinity, GSE44763, and GSE40279 data sets.

4.1 Introduction

Clustering can be defined as a way to group data naturally. The K-means [152] is a classic unsupervised learning algorithm that aims to find user-defined number of clusters which are represented by centroids. K means algorithm is practical, simple and typically fast [153]. The process of the K-means algorithm consists of the following steps:

(i) A centroid is defined for each cluster; thus, a total of k centroids are defined.

(ii) Each data point is assigned to the closest centroid.

(iii) Centroid positions are recomputed.

(iv) Steps (ii) and (iii) are repeated until no more moves are possible for the centroids.

4.2 K-Means Based Unsupervised Feature Selection Method (KBFS)

Before describing the proposed K-means based unsupervised feature selection method, the K-means algorithm is explained in detail. Its advantages and limitations are presented, and then the proposed framework is described.

A K-means algorithm for two clusters is illustrated in Fig. 4.1. In Fig. 4.1(a), two centroids are randomly placed, and in Fig4.1(b), a hyperline is generated to differentiate between the points on the left-hand side of the hyperline which belong to the red centroid, and the points on the right-hand side which are assigned to the yellow centroid which is shown in Fig. 4.1(d). The positions of the centroids are recomputed by taking the mean of all data points belonging to the same cluster (either the yellow or the red). The same procedure is repeated in Fig. 4.1(e), Fig. 4.1(f), Fig. 4.1(g), Fig. 4.1(h), and Fig. 4.1(i) until the objective function has converged.

The purpose of the K-means algorithm is to classify or to group data into a set of clusters. Grouping or classifying data is extremely useful for classification purposes. However, the K-means algorithm is generally not effective for regression problems. In this study, the K-means algorithm is modified to perform feature selection particularly for regression tasks.

The K-means algorithm is a partitional clustering algorithm that attempts to find k partitions of a given data, where k is a user-defined integer. Therefore:

Given a data set $X = x_1, \ldots, x_j, \ldots, x_n$ in which $x_j = (x_{j1}, \ldots, x_{jd})^T \in R^d$, K-Means attempts to find K clusters of X, $C = C_1, \ldots, C_j, \ldots, C_k$, such that

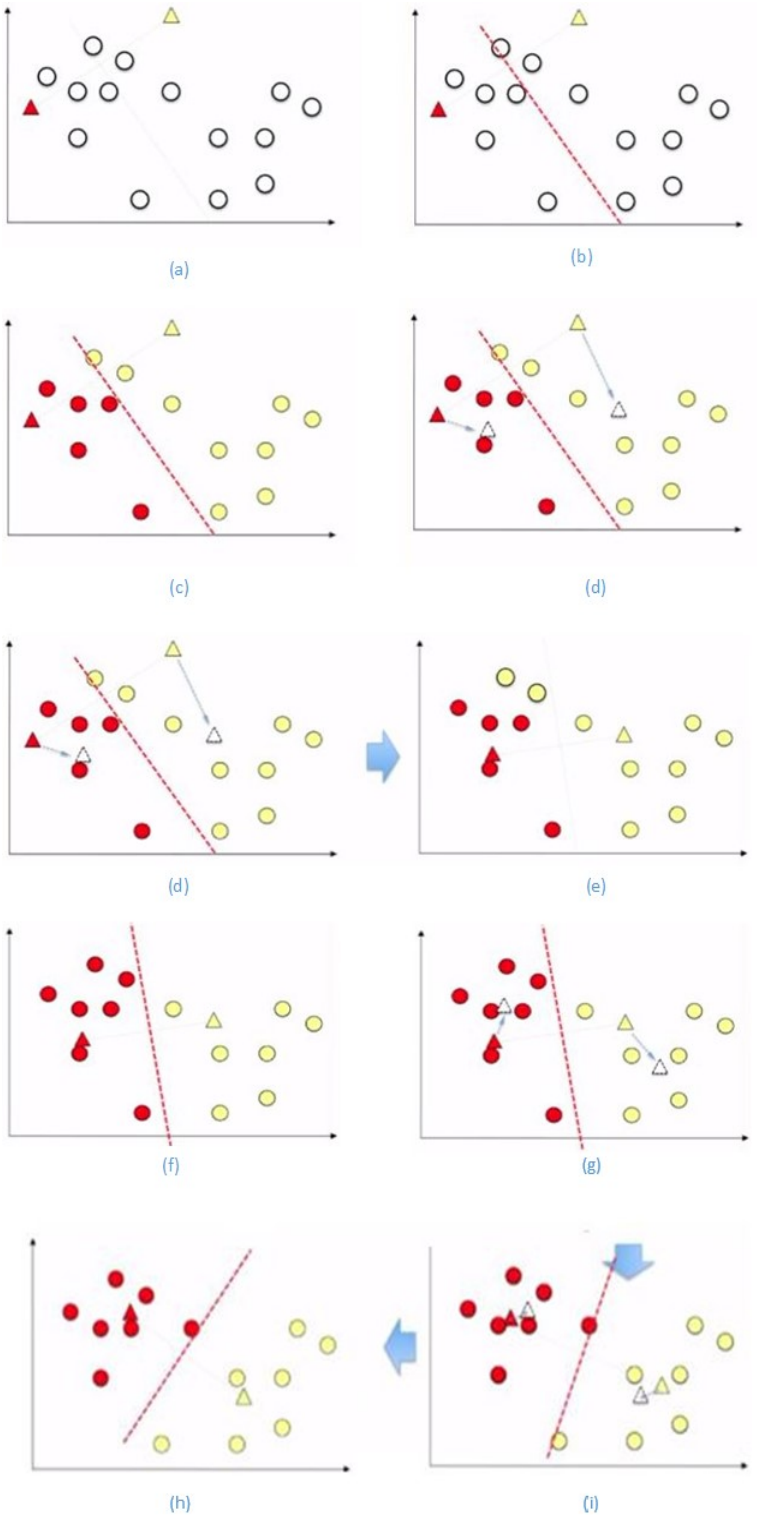

FIGURE 4.1: Basic K-Means Algorithm for Clustering purpose.

$$
\begin{aligned}
&C_i \neq \emptyset, i = 1, \ldots, k \\
&\cup_{i=1}^{k} C_i = X \\
&C_i \cap C_j = \emptyset, \quad i, j = 1, \ldots, k \quad and \quad i \neq j
\end{aligned}
\tag{4.1}
$$

where k is a user-defined integer. It is shown above that a pattern can only be allowed to belong one cluster.

There is no established method found to determine optimum number of clusters and to initialise the centroids [154] [155]. One of the most popular methods for the initialisation of centroids is to run algorithm with random initial centers [154]. A random sample of data points can be also be selected as an initial centroid [156]. There are plenty of K-means algorithms provided in the literature. In [157], the authors provide an algorithm, called CCIA, to initialise the centroid data points, and in another paper [158], a method is proposed which performs clustering without pre-defining the exact number of clusters, in [159], a starting point for the kth cluster centre is calculated by minimising an auxiliary cluster function, in [160], the author used different min-max distance measure to determine the distance between a data point and its cluster centroid. However, all of these K-means based methods are used for clustering purposes.

There are various advantages and disadvantages of using K-means algorithm. The advantages are that:

- The K-means is one of the most popular partitioning clustering algorithms thanks to its superior scalability [161].
- It is a simple, practical and efficient algorithm [160], in addition, it is generally very fast [153].
- The K-means is also very effective for processing high dimensional data [160].

The disadvantages include that:

- If there are outliers where points are far away from the cluster centroid in comparison to other points in that cluster, they can seriously harm the results.

- The K-means clusters data points according to their Euclidean distance to the centre points, and thereby it does not consider the different densities of each cluster. Consequently, each cluster has to consist of the same number of data points [161].
- The K-means algorithm produces different results for different user-defined numbers of clusters (k) [162].
- K-Means randomly initialises the centroids and different values of the initial centroids would produce different results.

In order to exploit the advantages of K-means algorithm and mitigate some of its disadvantages, a K-means based unsupervised feature selection framework for regression problems is proposed. Since the K-means is very effective for dealing with high dimensional data, and because most existing feature selection algorithms are not suitable to directly apply to high dimensional data, a novel K-means based unsupervised feature selection is needed.

At the starting point, a simple K-means based unsupervised feature selection algorithm is proposed [88] to deal with high dimensional data in regression domain. In our earlier study [88], K-means clustering algorithm is utilised for the quantitative prediction of peptide binding affinities being one of the most challenging post-genome regression problems of very high-dimension compared to extremely small size of samples. The clustering algorithm is used to partition the features into a number of clusters. The feature that is the closest to the cluster centre is then selected to represent the cluster. Therefore, the number of clusters determines the number of selected features. This basic K-means algorithm has produced better results than some of the state-of-the-art unsupervised feature selection methods for the peptide binding affinity prediction in [88]. This algorithm was named KCEN, but in [88], it did not produce the best results for the prediction of peptide binding affinities. Therefore, it needs to be further improved so that it might produce better prediction results.

The proposed K-means based unsupervised feature selection method, KBFS, begins by transposing the data so that features become instances and samples become features. Then, the data is divided into k-clusters where k is a user defined integer. As mentioned above, the K-means algorithm ranks features based on their distances to centroids, and it generally utilises Euclidean or squared

Euclidean distance measure. In KBFS, the centroids are identified via the K-means algorithm, however, instead of using one centroid points, three centroid points are utilised in the final stage. The distances of all features to the all centroids are calculated and the closest two features to a centroid are selected as other centroids. In other words, three centre points are identified based on their distance to the centre of each cluster. Then, in order to calculate the distances among features and centroids, the most commonly used metric, which is euclidean distance, is utilised. Euclidean distance can be calculated by:

$$J = \sum_{j=1}^{K} \sum_{i=1}^{n} \|x_i - C_j\|^2 \tag{4.2}$$

where x_is $i = 1, \ldots, n$ are a set of features to be partitioned to K clusters and C_js $j = 1, \ldots, K$ are the centroid points.

One of the problems of clustering algorithms is that the results of clustering can be profoundly affected by differences in scale among the dimension from which the distances are computed [160]. Therefore, the proposed algorithm performs normalisation process as the initial step of the clustering process to deal with this problem.

Normalisation is the process of scaling the inputs so that the values of inputs lie between set limits. This enables numerical calculations to be performed rapidly and easily [163]. Therefore, the proposed feature selection framework starts by normalising the raw data set. Normalisation of the input features can be achieved by:

$$x' = a + \frac{(x - x_{min}) * (b - a)}{(x_{max} - x_{min})} \tag{4.3}$$

where x is the original value of the input, and $x^{'}$ is normalised value. The a and b are the arbitrary points which present the limits of the values. In this study, input data is normalised into the range [0, 1] (a=0 and b=1).

In KBFS, three centroid points are exploited for each cluster and features are ranked based on their absolute distance values to those centroids. A feature with the lowest distance to the any of three centroid points in a cluster is considered as the most important one. In KBFS distance measure is calculated by:

$$
\begin{aligned}
J_{i1} &= \sum_{j=1}^{K} \sum_{i=1}^{n} \|x_i - C_{j1}\|^2 \\
J_{i2} &= \sum_{j=1}^{K} \sum_{i=1}^{n} \|x_i - C_{j2}\|^2 \\
J_{i3} &= \sum_{j=1}^{K} \sum_{i=1}^{n} \|x_i - C_{j3}\|^2
\end{aligned}
\tag{4.4}
$$

The weight of a feature is then calculated by:

$$
WX_i = \frac{1}{min(J_{i1}, J_{i2}, J_{i3})} \tag{4.5}
$$

The purposes of identifying three centre points are minimising the randomisation error, dealing with outliers and getting a handle on upcoming features. In K-means algorithm, the distances between two features is not influenced by upcoming features [153]. On the other hand, in KBFS, ranking error for upcoming features is minimised since three centroids are used to calculate feature weights. In KBFS, even though euclidean distance measure is utilised, at the final stage, a feature can be considered to belong to another cluster according to its distance measure to the centroids.

As mentioned above, the K-means method randomly initialises the centroids and this might profoundly affect the clustering results. Therefore, the process of KBFS is repeated 100 times to minimise the randomisation error. At the end, the mean of the distances between the centroids and the features are calculated in order to rank features. Therefore,

$$
\frac{1}{WX_i} = \frac{1}{p} \sum_{t=1}^{p} min(\sum_{j=1}^{K} \sum_{i=1}^{n} \|x_i - C_{j1}\|^2, \sum_{j=1}^{K} \sum_{i=1}^{n} \|x_i - C_{j2}\|^2, \sum_{j=1}^{K} \sum_{i=1}^{n} \|x_i - C_{j3}\|^2) \tag{4.6}
$$

where $p = 1, 2, \ldots, 100$, C represents clusters, $x_i s$ are features where $i = 1, 2, \ldots, n$, K is the number of clusters, WX_i is the weight of $i - th$ feature, and $C_j s$ are centroids.

Support Vector Regression (SVR) has been shown to be a powerful prediction method, and it generally yields better predictive model with higher generalisation ability [46]. Therefore, in order to evaluate the performance of the proposed K-means based unsupervised feature selection method, SVR, which is presented in Chapter 3, is exploited. The robustness of the proposed KBFS framework is compared with that of the state-of-the-art unsupervised feature selection methods over different high dimensional data sets, such as GSE40279 which contains more than 450k features, and a relatively small sample size(656). The prediction results of unsupervised feature selection methods, including KBFS, are presented in next section.

A complete flowchart of the proposed unsupervised feature selection framework is presented in Fig. 4.2. As mentioned above, the proposed framework begins by normalising the input data so that the values of the input data stay between set limits. Then, the input data is transposed so that features become samples and samples become variables. The transposition of data enables the predictive model to cluster features rather than instances and to use the features as part of the feature selection process. Therefore, feature-feature dissimilarities are revealed. Then, the transposed data is used by KBFS method to determine the weights of each feature, and features are ranked based on their weights. The ranked features are then forwarded to the regression model, which uses SVR to generate a model and exploits evaluation metrics to evaluate the performance of the predictive model. Finally, prediction results are generated as the final output of the proposed framework.

The pseudo code of KBFS unsupervised feature selection algorithm is shown below:

Input: dxn data matrix A(d features n samples), number of clusters (K),

for i=1:100 Randomly initialise centroids
Generate K cluster centroids randomly within the
range of the data or select K objects randomly as
initial cluster centroids. Let the centroids be C1, C2,..,CK
Compute distance of all features to these initial centroids
Identify 2 more centroids (features) which are closest to initial centroids
Calculate the distance measure by exploiting Equation 4.4
Calculate the weight of features by using Equation 4.5

Calculate the final weights of the features by Equation 4.6
end

Calcuate the mean of the final weights of features to decide their weights.

Output: The cluster indices of each point, the distance of each feature to each centroid, the final weights of each feature.

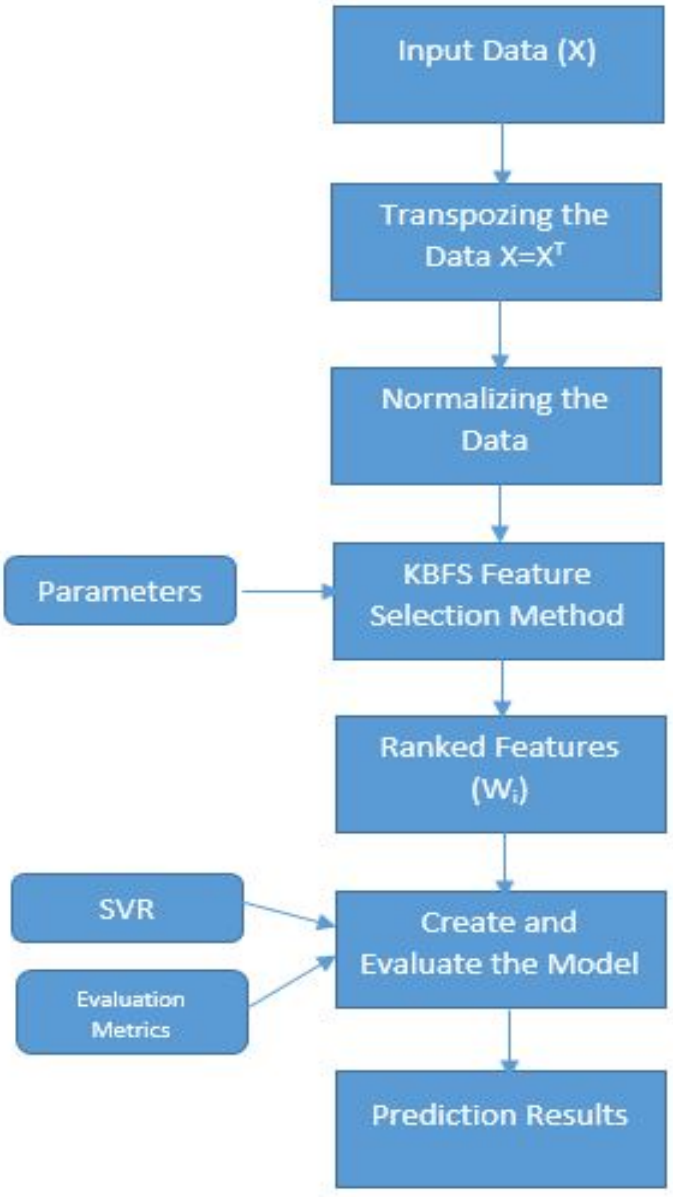

FIGURE 4.2: The Flowchart of The Proposed KBFS Framework.

The advantages and disadvantages of the proposed unsupervised feature selection framework are listed below. The advantages are that:

- The framework is also capable of dealing with upcoming features since it utilises three centroid points including features.
- The proposed method can better deal with outliers than the K-means method since it exploits three centroids, rather than utilising only one centroid which is even not a feature.
- By applying the normalisation process as a initial step of clustering, the proposed framework performs numerical calculations rapidly and easily, therefore, the proposed method is computationally fast.
- Since the clustering is repeated 100 hundred times, the proposed method produces more robust and reliable results than the K-means method.

Disadvantages:

- Since the proposed framework repeats clustering algorithm 100 times, it might be slower than K-Means algorithm.
- The number of clusters, K, is user defined and different numbers of clusters might produce different results.

4.3 Remarks on previous K-Means Based Feature Selection Methods

A number of K-Means based feature selection methods have been provided in literature [164] [165] [166] [100]. The first K-means based feature weighting method to be introduced was the SYNCLUS algorithm [164]. The SYNCLUS method begins by assigning a set of initial weights to variables; then, the K-means algorithm is utilised to partition the data into clusters. SYNCLUS assesses a new set of optimal weights via a weighted mean-square cost function. These two stages are iterated till an optimal set of weights is gained. The SYNCLUS method is computationally expensive, and therefore it cannot be applied to high dimensional data [166]. In [165], an entropy weighting K-means algorithm called EWKM was proposed for subspace clustering. It jointly minimises the within-cluster distribution and maximises the negative weight entropy in the clustering process. In

EWKM, the weight of each feature in each cluster is determined by including the weight entropy in the K-means objective function. Then, features are selected based on their weights. In [100], another feature weighting K-means based algorithm was proposed which uses a generalised Fisher ratio that minimises the ratio of the average of within-cluster distribution over between-cluster distribution. Among several candidate clusters, the one with the minimal Fisher ratio is selected as the ultimate cluster. This method decides the final weights of each feature from a pre-defined set of weights that cannot be guaranteed as optimal weights. In another study [166], a K-means based weighting algorithm called W-k-Means is proposed. W-k-Means decides the weight of a feature according to its variance in within-cluster distance. However, the W-k-Means algorithm randomly initialises the weight of each feature, however, those weights may not be guaranteed to provide an optimal solution.

All of the variations of K-means based feature selection methods use centroids, which are not features, to determine the weights of features. In this case, feature-feature correlation; in other words, multivariate feature selection is missing. In KBFS, features are utilised as centroid points; therefore, multivariate feature selection is accomplished by calculating feature-feature dissimilarity. Furthermore, since the K-means randomly identifies the centroid points, the clustering results can dramatically change for each run. KBFS overcomes this problem by repeating K-means for 100 times; therefore, it produces more robust results than existing methods. What is more? KBFS does not assign initial weights to the variables, instead, it determines the weight of each feature by solving the equation (4.6).

4.4 Results

This section presents the results of the application of the proposed KBFS method compared to the state-of-the-art unsupervised feature selection techniques as well as the baseline (entire feature set) with the RV144 Vaccine, peptide binding affinity, GSE44763, and GSE40279 data sets. These data sets are presented in Chapter 3.

4.4.1 Results for RV144 Vaccine Data Set

The RV144 data set provided in [119] is used in this study to model the antibody feature-function relationship. This data set contains 100 plasma samples (20 of which are placebo and 80 are vaccine-injected) obtained from the individuals participating in the RV144 vaccine trial at week 26. Three different cell-mediated assays are used: Antibody Dependent Cellular Phagocytosis; Antibody Dependent Cellular Cytotoxicity; and Natural Killer Cell Cytokine Release activities. The accuracy results for the proposed KBFS framework are compared with those presented in a previous study [119], and are also compared with results from four different state-of-the-art unsupervised feature selection methods, namely MCFS, InFS, LapFS, and SPFS, along with the entire feature set. In this study, the PCC and RMSE metrics are used so as to analyse the performance of unsupervised feature selection algorithms. The PCC metric is used to be able to perform a consistent comparison with the previous study [119]. The RMSE measure is exploited to compare the performance of the predictive models for performing MISO and MIMO regression tasks. SVR and MSVR are utilised to perform MISO and MIMO regression tasks respectively.

The SVR-based predictive models for the regression tasks are constructed using feature selection methods (filtered feature set). Their performance is then evaluated using a five-fold cross validation method. The RV144 data set is divided into two sets of samples. Four out of five samples, with a total of 64 samples, are utilised for training and the rest (16 samples) for testing purposes. This process is repeated 200 times by randomly creating subsets of the samples for the five-fold cross validation in order to avoid a bias towards and to assess the effect of randomisation in the cross validation. At the end, the mean performance and its corresponding standard deviation (std) values are obtained for each of the predictive models.

The prediction performance of unsupervised feature selection methods on three cell-mediated assays are summarised in Tables 4.1-4.3. Table 4.1 shows the PCC and RMSE results of predictive models for Natural Killer cell Cytokine release activities. The predictive models aim to estimate the level of cytokine release in order to understand its functionality for protection. The results suggest that KBFS outperforms state-of-the-art methods with 0.52 PCC using 16 features. SPEC yields the second-best result, at 0.51 PCC, with 16 antibody features.

TABLE 4.1: Comparison of Unsupervised Feature Selection Methods for the Antibody Features and Natural Killer Cell Cytokine Release Activity Relationship.

Metrics	PCC	RMSE
KBFS (16)	$\mathbf{0.52 \pm 0.17}$	1.95 ± 0.71
MCFS (16)	0.49 ± 0.17	$\mathbf{1.93 \pm 0.67}$
Laplacian (16)	0.49 ± 0.18	1.94 ± 0.70
SPEC (16)	0.51 ± 0.17	2.05 ± 0.68
InFS (18)	0.49 ± 0.17	2.04 ± 0.74
Baseline (20)	0.49 ± 0.17	2.04 ± 0.7

TABLE 4.2: Comparison of Unsupervised Feature Selection Methods for the Antibody Features and Cellular Cytotoxic Activity Relationship.

Metrics	PCC	RMSE
KBFS(11)	$\mathbf{0.43 \pm 0.19}$	5.43 ± 0.99
MCFS (18)	0.39 ± 0.18	5.42 ± 0.97
Laplacian (12)	0.39 ± 0.18	$\mathbf{5.42 \pm 0.93}$
SPEC (18)	0.41 ± 0.18	5.44 ± 0.92
InFS (14)	0.40 ± 0.17	5.48 ± 0.98
Baseline(20)	0.38 ± 0.18	5.6 ± 0.98

Other methods produce average results. Interestingly, RMSE results of unsupervised methods are profoundly different than PCC results of them. For example, MCFS shares the worst performance with InFS, LapFS and the baseline for PCC metric; on the other hand, it produces the best results for RMSE metric.

The prediction results of unsupervised predictive models for ADCC activities are presented in Table 4.2. KBFS again achieves the best PCC result yielding 0.42 using only 10 antibody feature. InFS produces the second-best result with 0.40 PCC utilising 14 antibody features. Other methods produce average results.

Table 4.3 presents the prediction results of USFSMs for ADCP activities. As can be clearly seen in the table, KBFS filtered predictive model outperforms the predictive models implemented with the complete feature set, InFS and SPEC. On the other hand, KBFS, Laplacian Score and MCFS produce the same PCC results with 12, 3 and 17 antibody features respectively. It is observed that the RMSE results of the predictive models are slightly different from their PCC results. KBFS produces the best RMSE results for ADCP assay, at 30.8; on the other hand, MCFS yields the best result for the Cytokine assay giving 1.93 RMSE.

TABLE 4.3: Comparison of Unsupervised Feature Selection Methods for the Antibody Features and Cellular Phagocytosis Activity Relationship.

Metrics	PCC	RMSE
KBFS(12)	0.65 ± 0.17	30.8 ± 3.87
MCFS (17)	0.65 ± 0.14	31.9 ± 3.86
Laplacian (3)	0.65 ± 0.15	31.3 ± 3.81
SPEC (18)	0.61 ± 0.14	32.7 ± 4.03
InFS (18)	0.64 ± 0.15	32.2 ± 3.97
Baseline(20)	0.61 ± 0.15	33.1 ± 3.62

TABLE 4.4: A Comparison of the Results with the Previous Study for the Antibody Features and Cellular Phagocytosis Activity Relationship.

Regression	PCC
Lars [119]	0.61±0.15
GP [119]	0.53±0.16
SVR [119]	0.56±0.19
KBFS	**0.65 ±0.17**

TABLE 4.5: A Comparison of the Results with the Previous Study for the Antibody Features and Cellular Cytotoxic Activity Relationship.

Regression	PCC
Lars [119]	0.42±0.18
GP [119]	0.24±0.21
SVR [119]	0.14±0.24
KBFS	**0.43±0.19**

The prediction results of the proposed method are also compared with those of the previous study [119] where the same data set by using the same cross validation method is utilised (5-fold with 200 replicates) and comparison results are shown in Tables 4.4-4.6. The results appear to suggest that DFSFR has a better quantitative accuracy than the predictive models constructed using Lars, GP and SVR as presented in the previous study for ADCC and ADCP assays, at 0.43 and 0.65 PCC respectively. In particular, the proposed approach yields as much as 1.16x and 3x better outcomes than the results of SVR for the ADCP and ADCC assays respectively. KBFS has slightly lower quantitative performance as compared to the predictive model for the Cytokine assay constructed using SVR as presented in the previous study. However, it still has better quantitative performance than the Lars and GP predictive models for the Cytokine assay.

Overall, the proposed KBFS framework generally achieves the best performance on all cell-mediated assays, which thereby verifies that it is is able to select

TABLE 4.6: A Comparison of the Results with Previous Study for the Antibody Features and Natural Killer Cell Cytokine Release Activity Relationship.

Regression	PCC
Lars [119]	0.51±0.21
GP [119]	0.46±0.24
SVR [119]	**0.55± 0.15**
KBFS	0.52±0.17

informative antibody features.

4.4.1.1 Results for Multi-Input-Single-Output (MISO) and Multi-Input-Multi-Output (MIMO) Regression

A comparison of prediction results of predictive models for MISO and MIMO regression tasks is shown in Table 4.7. The results of SVR-based models for the Cytokine, ADCC, and ADCP assays are listed in Table 4.7. The average RMSE results of SVR-based models are calculated by taking the mean of the RMSE results for each assay. The results suggest that the MSVR-based predictive model outperforms the SVR-based predictive model, which indicates that some correlations exist amongst the target variables. In the previous study [119], these MIMO regression correlations are not taken into account. Analysing dependencies between antibody features as well as response variables (the functional activities of antibody features) may ultimately result in producing an effective vaccine so that HIV or AIDS may be conquered.

TABLE 4.7: A comparison of Unsupervised Prediction Results for SVR and MSVR for Anticipating Antibody Feature-Function Relationship.

Metrics	KBFS	MCFS	LapFS	SPEC	InFS
SVR (Cytokine)	1.95 ± 0.71	1.93 ± 0.67	1.94 ± 0.70	2.05 ± 0.68	2.04 ± 0.74
SVR (ADCC)	5.42 ± 0.99	5.42 ± 0.97	5.42 ± 0.93	5.44 ± 0.92	5.48 ± 0.98
SVR (ADCP)	30.8 ± 3.87	31.9 ± 3.86	31.3 ± 3.81	32.7 ± 4.03	32.2 ± 3.97
SVR (Average)	12.72 ± 1.85	13.08 ± 1.85	12.88 ± 1.97	13.83 ± 1.85	13.24 ± 1.83
MSVR	11.72 ± 1.95	13.01 ± 1.77	12.07 ± 1.70	12.30 ± 1.68	12.83 ± 1.74

4.4.2 Results for Peptide Binding Affinity Data Sets

Three different high dimensional peptide data sets, provided in the CoEPrA modelling competition [129], are used. Each data set consists of training and test data sets, therefore, there is no need for cross validation. Tasks 1 and 3 contain nona-peptides that contain a total of 5787 amino acid descriptors. Task 1 consists of 89 training and 88 testing samples, whereas Task 3 has 133 training and 133 testing instances. Task 2 consists of octa-peptides with a total of 5144 amino acid descriptors. It contains 76 training and 76 testing samples.

The prediction performance of the proposed KBFS framework for Tasks 1, 2 and 3 are compared with five different USFSMs, namely MCFS, KCEN, EUFS, LapFS and SPFS, along with the entire feature set (baseline). The prediction performance of unsupervised feature selection methods over different tasks is summarised in Tables 4.8-4.10. In order to investigate the robustness of the USFSMs, their default parameters are exploited. The number of selected features is initially 50 and then incremented by 50 to form feature sets of {50, 100, ..., 250, 300}. Table 4.10 demonstrates the performance of the USFSMs for Task 1. The results suggest that KBFS produces the best results for MAD, MSE, RMSE, MAPE, U and q^2 metrics with 100 selected features. The second best results are achieved by SPEC with 300 features. Other methods produce average results.

A comparative analysis of USFSMs for Task 2 is shown in Table 4.9. The results of the experiment with the Task 2 data set confirm that KBFS generally yields the best results for all metrics, yielding 0.28 MAD, 0.17 MSE, 0.41 RMSE, 4.05 MAPE, 0.008 U and $0.70q^2$ with 300 features. SPFS produces the second-best results, with 0.28 MAD, 0.17 MSE, 0.41 RMSE, 3.9 MAPE, 0.007 U, and 0.68 q^2 with 300 features. The results for SPFS are very similar to those for KBFS, but the latter achieves the best results using 200 features while SPFS produces the second-best results with 300 features. Other USFSMs produce average results.

TABLE 4.8: Regression Results of the Unsupervised Feature Selection Methods for Task 1

Metrics	MAD	MSE	RMSE	MAPE	U	q2
KBFS(300)	**0.48±0.03**	**0.34**±0.05	**0.59±0.04**	**9.65±0.8**	**0.017±0.03**	**0.61**±0.7
EUFS(300)	0.49±0.0	0.53±0.0	0.73±0.0	9.86±0.0	0.024±0.0	0.46±0.0
KCEN(200)	0,51±0,04	0.48±0.07	0.69±0.05	10.1±0.76	0.023±0.0	0.44±0.82
MCFS(50)	0,57±0.0	0.54±0.0	0.74±0.0	11.4±0.0	0.025±0.0	0.37±0.0
LapFS(300)	0,58±0.0	0.61±0.0	0.78±0.0	11.3±0.0	0.027±0.0	0.30±0.0
SPFS(300)	0,50±0.0	0.37±0.0	0.61±0.0	9.7±0.0	0.020±0.0	0.57±0.0
Baseline	1,07±0.0	1.82±0.0	1.35±0.0	21±0.00	0.043±0.0	-1.0±0.0

A comparative analysis of USFSMs for Task 3 is shown in Table 4.10. The proposed approach clearly generates the best results, yielding 0.58 MAD, 0.52 MSE, 0.72 RMSE, 8.59 MAPE, 0.19 q^2 and 0.014 U. The results for EUFS and LapFS are similar. They both produce 0.6 MAD, 0.58 MSE, 0.76 RMSE; however, EUFS yields 9 MAPE, 0.079 q^2 and 0.014 U whereas LapFS achieves 8.6 MAPE, 0.081 q^2 and 0.015 U.

Given these analyses, all the results present a clear message that the SVR-based predictive model with all the features fails. This outcome suggests the necessity of feature selection. It is also observed that the performance of the USFSMs is relatively sensitive to the number of selected features. The number of selected features is provided in parenthesis located just next to the USFSM results in the tables.

TABLE 4.9: Regression Results of the Unsupervised Feature Selection Methods for Task 2

Metrics	MAD	MSE	RMSE	MAPE	U	q2
KBFS(200)	**0.28±0.03**	**0.17±0.04**	**0.41±0.02**	**4.05 ±0.4**	**0.008±0.002**	**0.7±0.04**
EUFS(100)	0.39±0.0	0.43±0.0	0.65 ±0.0	5.98 ±0.0	0.011±0.0	0.2±0.0
KCEN(200)	0.35 ±0.0	0.27±0.0	0.52±0.0	5±0.0	0.023±0.0	0.49±0.0
MCFS(300)	0.32 ±0.0	0.2±0.0	0.45±0.0	4.6 ±0.0	0.009±0.0	0.62±0.0
LapFS(300)	0.35 ±0.0	0.29 ±0.0	0.54±0.0	5.1 ±0.0	0.009±0.0	0.45±0.0
SPFS(300)	0.28 ±0.0	0.17±0.0	0.41±0.0	3.9±0.0	0.007±0.0	0.69±0.0
Baseline	0.29±0.0	0.16±0.0	0.4±0.0	4.02±0.0	0.007±0.0	0.7±0.0

TABLE 4.10: Regression Results of the Unsupervised Feature Selection Methods for Task 3

Metrics	MAD	MSE	RMSE	MAPE	U	q2
KBFS(150)	**0.58**±0.02	**0.52**±0.03	**0.72**±0.03	**8.59**±0.8	**0.014**±0.001	**0.19**±0.01
EUFS(150)	0.61 ± 0.0	0.58 ± 0.0	0.76 ± 0.0	9 ± 0.0	0.014 ± 0.0	0.07 ± 0.0
KCEN(300)	0.66 ± 0.0	0.67 ± 0.0	0.81 ± 0.0	9.7 ± 0.0	0.016 ± 0.0	−0.06 ± 0.0
MCFS(50)	0.7 ± 0.0	0.76 ± 0.0	0.87 ± 0.0	10.1 ± 0.0	0.017 ± 0.0	−0.20 ± 0.0
LapFS(50)	0.6 ± 0.0	0.58 ± 0.0	0.76 ± 0.0	8.6 ± 0.0	0.015 ± 0.0	0.08 ± 0.0
SPFS(300)	0.67 ± 0.0	0.75 ± 0.0	0.86 ± 0.0	9.9 ± 0.0	0.017 ± 0.0	−0.18 ± 0.0
Baseline	1.17 ± 0.0	2.51 ± 0.0	1.58 ± 0.0	17 ± 0.0	0.031 ± 0.0	−2.97 ± 0.0

4.4.3 Results for the GSE44763 Data Set

The GSE44763 data set [130] is utilised to model the associations among CpG biomarkers (features), chronological age and obesity. This data set contains 27482 Cytosine-phosphate-Guanine (CpG) biomarkers from the peripheral blood of 46 adult female donors (samples). There are 24 obese subjects and 22 lean subjects. In this study, a subject is considered obese if his/her BMI is greater than or equal to 30, and a subject is considered as lean if his/her BMI is less than 25. In order to investigate the robustness of the USFSMs their default parameters are used. The number of selected features is initially 50 and then incremented by 50 to form feature sets of $\{50, 100, \ldots, 250, 500\}$.

The performance of the proposed KBFS method is compared with the state-of-the-art USFSMs, including EUFS, InFS, LapFS, and SPFS along with the entire feature set. In order to evaluate the robustness of USFSMs, support vector-based methods are used since their effectiveness has been proven and they provide better generalisation and performance in a wide range of bioinformatics applications [102] [14]. To observe the results for these methods using different metrics, three different metrics are used to assess the quality of the USFSMs, which are Mean Absolute Deviation (MAD), Root Mean Squared Error (RMSE) and Theils U-statistics (U). The RMSE metric is utilised to calculate prediction errors for both MISO and MIMO regression tasks. SVR and MSVR are exploited to perform MISO and MIMO regression tasks, respectively. The prediction results of the predictive models are calculated and averaged with the five-fold cross validation method. Therefore, four out of five samples are used for training and the rest of the samples are utilised for testing purposes. The five-fold cross validation is repeated 200 times in order to gain more unbiased results. Then, the mean performance and its corresponding standard deviation (std) values are obtained for each of the predictive models.

In this study, Ilumina average beta values are utilised as numerical data where the Beta-value is the ratio of the methylated probe intensity and the overall intensity (sum of methylated and unmethylated probe intensities). Beta value for an ith investigated CpG island is determined as follows [139]:

$$Beta_i = \frac{max(y_{i,methy}, 0)}{max(y_{i,unmethy}, 0) + max(y_{i,unmethy}, 0) + \alpha} \tag{4.7}$$

TABLE 4.11: The Performances of USFSMs for Prediction of Chronological Age

Metrics	MAD	RMSE	U
KBFS(500)	8.02±1.53	9.31±1.63	$\frac{303}{100000}\pm\frac{52}{100000}$
EUFS(500)	8.24±1.91	9.51±1.41	$\frac{308}{100000}\pm\frac{42}{100000}$
InFS(500)	8.14±1.40	9.39±1.33	$\frac{306}{100000}\pm\frac{40}{100000}$
SPFS(350)	8.21±1.46	9.51±1.35	$\frac{309}{100000}\pm\frac{40}{100000}$
LapFS(150)	8.32±1.45	9.58±1.37	$\frac{311}{100000}\pm\frac{40}{100000}$
Baseline	8.12 ± 1.43	9.41 ± 1.37	$\frac{307}{100000}\pm\frac{41}{100000}$

where $y_{i,methy}$ and $y_{i,unmethy}$ are the intensities measured by the ith methylated and unmethylated probes respectively, and α is a constant offset which is added to the denominator in order to regularise the Beta value if unmethylated and methylated probe intensities are low. The default value of α is 100.

The prediction performance of USFSMs are summarised in Tables 4.11-4.13. Table 4.11 shows the robustness of USFSMs for the prediction of chronological age. The results suggest that the proposed method outperforms the state-of-the-art unsupervised feature selection methods. KBFS produces the best results yielding 8.02 RMSE, 9.31 MAD and 0.003 U with 500 features and InFS yields the second-best results, with 8.14 MAD, 9.39 RMSE, and 0.003 U. Other feature selection methods produce average results. Interestingly, all of the USFSMs produce similar U results; however, the results for different metrics are consistent. For example, DKBFS yields the best results for all different metrics.

Surprisingly, the complete feature set (baseline) produces 8.12 MAD, 9.41 RMSE and 0.00307 U, and thereby yields better results than LapFS, SPFS, and EUFS. This outcome implies that most of the CpG biomarkers are related to aging. It is also observed that the performance of the USFSMs is relatively sensitive to the number of selected features. The number of selected features are shown in parenthesis in the tables.

A comparison of USFSMs for BMI prediction is shown in Table 4.12. The outcomes of the experiments clearly emphasise that the proposed KBFS outperforms state-of-the-art USFSMs. KBFS produces the best results with 500 features for

TABLE 4.12: The Performances of USFSMs for the Prediction of BMI

Metrics	MAD	RMSE	U
KBFS (500)	6.90±1.55	7.44±1.48	$\frac{87}{10000}\pm\frac{19}{10000}$
EUFS (400)	6.93±1.09	7.75±1.17	$\frac{9}{1000}\pm\frac{17}{10000}$
InFS(400)	6.93±1.09	7.75±1.16	$\frac{9}{1000}\pm\frac{17}{10000}$
SPFS (450)	6.99±1.52	7.52±1.51	$\frac{89}{10000}\pm\frac{21}{10000}$
LapFS (450)	6.98±1.56	7.5±1.55	$\frac{88}{10000}\pm\frac{2}{1000}$
Baseline	7.04 ± 1.62	7.62 ± 1.59	$\frac{89}{10000}\pm\frac{21}{10000}$

MAD, RMSE and U yielding 6.90, 7.44, and 0.0087 respectively. Other feature selection methods produce average results.

TABLE 4.13: The Performances of USFSMs for MSVR and SVR

Metrics	MSVR	SVR
KBFS(500)	8.45 ± 0.95	8.37 ± 1.55
LapFS(150)	8.69 ± 0.85	8.54 ± 1.46
SPFS(350)	9.4 ± 0.83	8.51 ± 1.43
InFS(400)	9.05 ± 0.81	8.57 ± 1.24
EUFS(400)	8.45 ± 0.77	8.63 ± 1.29

4.4.3.1 Results for Multi Input-Single Output (MISO) and Multi Input-Multi Output (MIMO) Regression

In this study, in addition to MISO regression, MIMO regression is performed to examine whether or not there is a relationship between age and obesity based on CpG biomarkers. A comparison of MISO and MIMO regression results is presented in Table 4.13. The results suggest that there is no strong correlation between obesity and aging based on the selected CpG dinucleotides (features).

Therefore, most of the age-related CpG islands are not related to obesity. Interestingly, only the MSVR result for EUFS are better than its result for SVR. This result appears to suggest that some of the CpG biomarkers which are selected by EUFS are related to both aging and obesity.

4.4.4 Results for the GSE40279 Data Set

The GSE40279 data set provided in [145] is used to model the relationship between CpG biomarkers and chronological age. This data set contains 473034 CpG biomarkers (features) from the whole blood of 656 donors (samples) aged 19 to 101.

A pre-processing step is applied to map the data into lower dimensional space so that feature selection methods can be applied to the data set. First, the standard deviations of the samples, which refer to the amount of variation in data samples, are calculated. The standard deviation of a sample can only be zero if, and only if, the samples are identical. If a feature is identical in all samples, then the feature is not discriminative. Therefore, before applying feature selection, the features which have the lowest variation in the data are eliminated. As a result, approximately four out of five of the features are eliminated in this pre-processing step.

Then, unsupervised feature selection methods are applied to identify discriminative CpG biomarkers (features). The number of selected features starts from 900 in order that a subset of features contains at least 1% of the entire feature set. A set of 90000 features is assessed using six different USFSMs along with the entire feature set.

The performance of the proposed KBFS method with the GSE40279 data set is compared with that of state-of-the-art unsupervised feature selection methods, including EUFS, LapFS, and Term Variance (TV) along with the entire feature set.

Support vector based models [167] are exploited to assess the quantitative prediction performances of unsupervised feature selection methods since they have achieved superior generalisation and performance in a large variety of bioinformatics applications [102] [16]. Support vector based predictive models for regression tasks are constructed using USFSMs (filtered feature set) and the complete

TABLE 4.14: A Comparison of USFSMs for The Prediction of Chronological Ages of Individuals using CpG Dinucleotides

Metrics	MAPE	q2	U	MAD
KBFS(900)	13.93 ± 0.91	0.09 ± 0.02	0.0032 ± 0.0002	11.08 ± 0.78
EUFS(9000)	14.47 ± 0.69	0.003 ± 0.009	0.0034 ± 0.0001	11.69 ± 0.61
LapFS(6300)	14.43 ± 0.84	0.01 ± 0.01	0.0034 ± 0.0002	11.63 ± 0.73
TV(900)	14.16 ± 0.90	0.06 ± 0.02	0.0033 ± 0.0002	11.34 ± 0.79
Baseline(90000)	14.61 ± 0.88	0.003 ± 0.01	0.0034 ± 0.0002	11.86 ± 0.79

feature set. As there is no separate training and test data sets 8-fold cross validation is used to evaluate the performance of the predictive models. The cross validation is repeated 50 times by randomly creating subsets of the instances for the 8-fold cross validation to avoid bias towards and alleviate the impact of the random split. The means and standard deviations of the metrics are calculated over these 50 runs and presented in Table 4.14. The number of selected features for each predictive model is shown in parenthesis in the tables.

The results appear to suggest that the proposed method yields better results than those of other USFSMs over different metrics. KBFS yields 13.93 MAPE, 0.09 q^2, 25.5 MAD and 0.0032 U with only 900 features. TV produces 14.16 MAPE, 0.06 q^2, 0.0033 U, and 11.34 MAD with 900 features and outperforms LapFS and EUFS. LapFS and EUFS which produce average results.

Given this analysis, all of the results present a clear message that the SVR-based predictive model with all of the features fails. This outcome suggests the necessity of feature selection. It also proves that the majority of CpG biomarkers are not related to the determination of an individual's chronological age.

4.4.4.1 An Aggressive Research of Features from GSE40279 Data Set

The experimental results of the experiment conducted with the GSE4079 data set suggest that the proposed DFSFR, KBFS, DKBFS frameworks produce better results than other USFSMs for all different metrics. However, even though the number of features are drastically reduced from 473034 to 900, the number

of CpG biomarkers (features) are still too high to be easily analysed in real biology laboratories. In this case, an aggressive research study is been conducted with different subset of CpG dinucleotides selected by KBFS. The purpose of this aggressive process is to obtain the minimum number of dinucleotides which represent the whole data set with the same or higher accuracy so that they can be further analysed in real biology labs.

The number of features used starts from 1 and is then incremented by 1 until 900 is reached to make an aggressive reduction of the selected CpG biomarkers. As shown in Table 4.15, the final predictive model of KBFS yields 10.69 MAD, 0.0031 U, 0.11 q2 error rate for age prediction with only 41 dinucleotides, which corresponds to only 0.00867% of the entire dinucleotide range. These 41 dinucleotides are listed in Table 4.16.

The smallest subset is found by KBFS (41 features) and it achieves better performance than existing feature selection methods. Those features are listed in Table 4.16. Further research can be carried out to investigate the 41 CpG biomarkers listed in Table 4.16 in biological laboratories to establish their biological relevance.

TABLE 4.15: Detailed Assessment of CpG Dinucleotides Using the Proposed KBFS framework

Method	MAPE	q2	U	MAD
KBFS (41)	$\mathbf{13.65 \pm 0.77}$	$\mathbf{0.11 \pm 0.03}$	$\mathbf{0.0031 \pm 0.0002}$	$\mathbf{10.69 \pm 0.59}$
KBFS (900)	13.93 ± 0.91	0.09 ± 0.02	0.0032 ± 0.0002	11.08 ± 0.78

Table 4.16: List of 41 CpG Dinucleotides

cg13869341	cg17149495	cg15174812	cg05662829
cg12045430	cg16162899	cg11422233	cg16047670
cg14008030	cg17866181	cg17501828	cg14057946
cg00381604	cg17308840	cg03344490	cg01070250
cg20826792	cg15394630	cg10037654	cg07264491
cg20253340	cg22802167	cg27534567	cg18761878
cg03130891	cg24159721	cg05001044	cg08858441
cg24335620	cg08477687	cg00645010	cg23917638
cg21870274	cg24669183	cg21996134	cg00168193
cg00034556	cg15560884	cg22394869	cg05597748
cg03348902			

4.5 Summary

In this chapter, a novel K-means based unsupervised feature selection framework, KBFS, is proposed. The advantages of the proposed method compared to existing K-means based methods are that it takes advantages of utilising features as centroids to determine feature-feature dissimilarity and to rank features, it produces more robust results by reducing randomisation error, and is also able to deal with upcoming features since KBFS is capable of updating feature weights. The disadvantages of the proposed framework include that the number of clusters, k, is still defined by the user and different numbers of clusters would produce different prediction results. The proposed framework might also be slower than the K-means method since it repeats the clustering algorithm 100 times. Experimental results with different high dimensional data sets, which are presented in previous section, have shown that the proposed framework produces better results than the state-of-the-art unsupervised feature selection methods with fewer

features. There are a number of K-means based feature selection algorithms provided in the literature, however, they are generally utilised for classification or clustering. On the other hand, the literature appears to suggest that there is a lack of studies in regression based problems for K-Means based feature selection.

Experimental studies conducted on the RV144 Vaccine, peptide binding affinity, GSE44763 and GSE40279 data sets to show the effectiveness of the proposed KBFS method its results are compared with those of state-of-the-art feature selection methods as well as with those of previous studies. The RV144 vaccine data set consists of 20 antibody features and 100 plasma samples that are obtained from the individuals participating in the RV144 vaccine trial week 26. Three different cell-mediated assays are used: Antibody Dependent Cellular Phagocytosis, Antibody Dependent Cellular Cytotoxicity and Natural Killer Cell Cytokine Release activities. The goal of exploiting the RV144 data set is to reveal antibody features that take action against HIV; in other words, to disclose the relationship between antibody features and their effector functions. In the previous study [119], only the MISO regression task was considered; however, MIMO regression was not taken into account. On the other hand, in this study, in addition to performing MISO regression analysis, MIMO regression analysis is applied so that associations among target variables can be revealed. The results of the experiments conducted with the RV144 data set indicate that there are not only correlations among variables, but also there are some correlations among the target variables. The accuracy results of the proposed KBFS approach indicate that it generally outperforms state-of-the-art USFSMs as well as the previous study [119]. From the experimental results for the RV144 Vaccine data set, it can be concluded that the proposed DFSFR framework can reveal discriminative antibody features that fight against HIV.

In this study, three different peptide binding affinity data sets are exploited. Tasks 1 and 3 contain nona-peptides that contain a total of 5787 amino acid descriptors and 89 samples. Task 1 consists of 89 training and 88 testing samples, whereas Task 3 includes 133 training and 133 testing instances. Task 2 consists of octa-peptides that have a total of 5144 amino acid descriptors. It has 76 training and 76 testing samples. Each descriptor contains 643 amino acids. The goal of exploiting the peptide binding affinity data set is to predict peptide binding affinity values by using amino acid descriptors, since these descriptors quantitatively describe the physicochemical properties of the peptides [128]. Affinity

refers to the strength of binding or interaction. PPIs play a role in mediating signal transactions, sensing the environment, triggering immunological responses, and monitoring gene expression [126]. Furthermore, PPIs play a crucial role in the progression of human diseases such as viral infections. Therefore, increasing knowledge of the underlying principles of PPIs can ultimately result in disclosing the intrinsic biochemistry of different diseases, and thereby the development of drug design [127]. The proposed KBFS framework generally outperforms the state-of-the-art USFSMs for all different tasks.

The GSE44763 data set contains 27482 CpG biomarkers (features) from the peripheral blood of 46 adult female donors (samples). There are 24 obese subjects and 22 lean subjects. The goal of exploiting this data set is to reveal the associations among CpG biomarkers, and the chronological age and BMI of individuals. The proposed KBFS framework outperforms the other USFSMs for both age and BMI prediction. The experimental results suggest that the proposed framework can reveal age and obesity-related CpG biomarkers (features) from the given data. In addition to performing MISO regression analysis, MIMO regression analysis is also performed. From the experimental results it can be concluded that no strong correlation exists between obesity and chronological age.

The GSE40279 data set consists of 473034 Cytosine-phosphate-Guanine (CpG) biomarkers (features) from whole blood of 656 donors (samples) aged 19 to 101. The goal of exploiting this data set is to reveal the relationship between CpG dinucleotides and the chronological age of individuals from the given data. A pre-processing step is applied to the GSE40279 data set so that the features with the lowest variation in the sample are eliminated, and thereby the number of features is reduced from 473034 to 90000. Then, USFSMs are applied to identify discriminative CpG biomarkers (features). The number of selected features starts from 900 in order that a subset of features contains at least 1% of the features. A set of 90000 features is assessed by utilising four different USFSMs along with the entire feature set. The proposed KBFS framework produces better results than other USFSMs.

The promising experimental results have led to further investigations of the three different sets of 900 CpG biomarkers which are selected by DKBFS, DFSFR, and KBFS. An aggressive assessment is made of those CpG dinucleotides. Here, the number of features used starts from 1 and then incremented by 1 until 900 is

reached. KBFS produces the best results and outperforms existing USFSMs with only 41 features.

Chapter 5

Deep Learning Based Feature Selection for Regression (DFSFR)

In this chapter, the proposed DFSFR method is presented. First, the concept of a deep neural network is introduced. The advantages of exploiting deep learning based methods are then considered. Then, existing deep learning based feature selection methods are briefly discussed, in the Background section. Finally, a novel deep learning based feature selection framework, particularly useful for regression problems, is proposed.

5.1 Introduction

Deep neural networks are constructed around a deep architecture where there are many hidden layers. Non-linear operations are performed in each layer, which transforms the representation at one level into representation at a more abstract level on the input data to learn very complex functions under study [168]. One of the advantages of deep learning is that the layers of features are not designed by human intervention: instead, they are learned from data, by exploiting a general-purpose learning procedure.

Deep learning methods are capable of handling the following problems [169]:

- Deep learning architectures are able to learn complex and highly varying functions where the number of variations are profoundly greater than the number of training samples.
- Deep learning methods are capable of learning with little human intervention (input).
- Deep learning architectures can learn from a large set of examples, and the computational complexity is almost linearly associated with the number of samples.
- Deep learning provides for robust unsupervised learning, that is capable of computing most of the statistical structure in the observed data.

Deep learning has been shown to be capable of representing data at multiple levels of abstraction. It is able to derive discriminative features resulting in enhanced accuracy [170]. Although most of the applications of the deep learning concept are in this direction, there is a recent study where this concept is primarily adapted to the refinement of the features extracted by using deep learning in the classification domain [170]. However, although this concept has been shown to be a powerful learning approach, it has not been explored for feature selection from naturally-collected feature sets in a regression domain. Therefore, a novel unsupervised feature selection method has been developed by adapting a deep learning concept in the regression domain. To the best of our knowledge, the proposed feature selection framework is the first, unsupervised, deep belief network based feature selection algorithm to perform regression tasks.

5.2 Background

In line with the technological developments, the terminology of deep learning has gained more attention as the deep learning-based architectures have been shown to be able to tackle more complex systems and to better learn data representations in an unsupervised manner . The deep learning architectures are a special case of artificial neural networks but with quite large number of layers and neurons in each layer. Therefore, in order to avoid classic artificial neural networks, deep learning terminology has been preferred instead. Therefore, this

study has taken account deep learning terminology. A literature search using the keywords " 'deep learning' and 'feature selection' " on both Web of Science and Scopus yielded a total of 75 studies as of 3 December 2016. They can be divided into 3 main categories;

(i) There has been a wide range of studies on general deep learning architectures for feature learning for characterisation and classification of objects, but there does not seem to have been any exploration of the feature selection concept.

(ii) Hybrid models where deep learning is first employed for feature learning and extraction. A feature selection method is then used to select more relevant features from the feature set derived by the deep learning architecture for classification purposes. This approach is expected to further refine the deep learning-based feature set. For example, sparse group LASSO and multi-modal deep neural networks were utilised for image classification in which the LASSO-based feature selection is adapted [171]. Random Forest method is another method used along with the deep learning architecture to train input data and rank features [172] in which types of credit risks were predicted and the number of features was reduced by 21%. Based on the cross-validation assessment, the next best features are selected according to their median score, average score and standard deviation of features. In [173] Stacked Denoising Auto Encoder and t-test are exploited to identify non-linear information in morphological features for pulmonary nodule classification in CT scans and their method achieved 2.1 % better accuracy than that of original raw features. The features with the highest p-values above a desired threshold ($p > 0.001$) are then eliminated. In [170] Deep Belief Network (DBN), a feature selection method (e.g., t-test, relief-f) and unsupervised active learning are used to select genes/MiRNAs, and then, Support Vector Machine (SVM) and Random Forest are used for cancer diagnosis. This method achieved better classification results than classical feature selection methods in hepatocellular carcinoma (HCC) by 9%, lung cancer by 6% and breast cancer by around 10%.

(iii) Feature selection embedded in to deep learning architecture is an approach where the deep learning method is used to identify relevant features. This proposes another feature selection technique based on the deep learning

concept. Due to its complexity and novelty, there have appeared only three main algorithms in the literature, but they are only for classification purposes [174] [175] [176]. In [174], feature selection is carried out at input level of deep learning structure to select features for multi-class data. This feature selection method is called Deep Feature Selection, and has recently been applied in the supervised prediction of active positions of cis-Regulatory regions [177]. In [175], the Deep Belief Network (DBN) and supervised fine-tuning are utilised to select temporal ultrasound features to detect prostate tissues, and then, a Support Vector Machine (SVM) with Radial Basis Kernel (RBF) is used for the detection of prostate cancer. In [176], an iterative feature learning algorithm is developed by using Deep Belief Network for the classification of remote sensing scenes.

As presented, the literature review appears to suggest that the deep learning approach is very popular for feature extraction and learning, particularly in image and video processing applications, but is still at a very early stage of feature selection, mainly for classification tasks. The literature review also reveals the fact that there is no deep learning based feature selection explored or developed for regression analysis. Therefore, to the best of our knowledge, the proposed method is the first of its kind in which a deep learning based feature selection method in regression domain is developed and presented.

5.3 Deep Learning Based Feature Selection for Regression (DFSFR)

DBN has generally been regarded as one of the best known deep leaning models [178]. It has proven its ability to discover better discriminative features and; consequently, to improve accuracy [170]. Furthermore, DBN has been shown outstanding performances on visual object recognition and image denoising [179]. However, the idea of DBN for feature selection for regression has not been applied yet. The novel unsupervised feature selection framework, DFSFR, utilises deep belief network to select discriminative antibody features and then applies SVR to perform regression task. Therefore, DFSFR is a multi-level feature selection framework that incorporates deep learning and SVR in order to select most discriminative features from high dimensional data. The proposed unsupervised

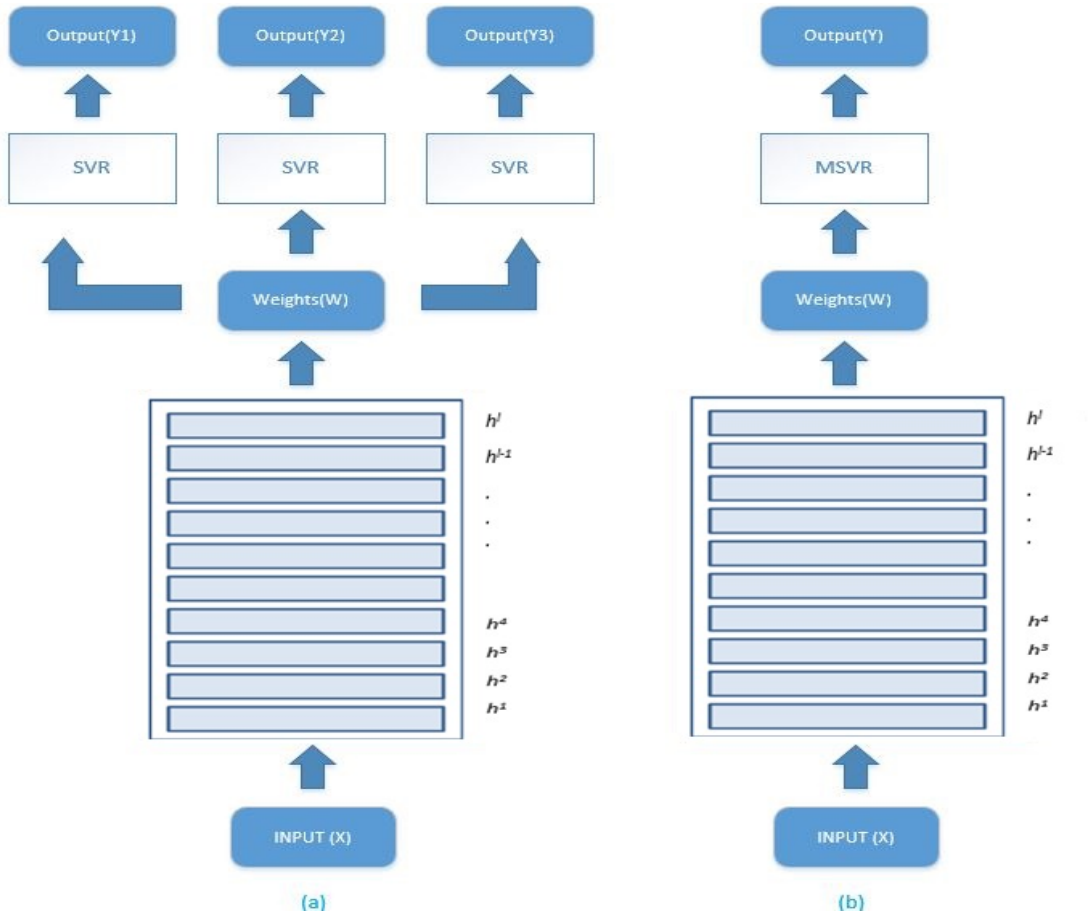

FIGURE 5.1: DFSFR Framework (a) multi-output (b) single-output. h represents hidden neurons.

feature selection framework, DFSFR, is demonstrated in Fig. 5.1. DFSFR takes input variables and feeds them into the deep belief network, then, DBN uses the weights provided from hidden nodes to produce weights for features. Next, features are prioritised according to their weights. Then SVR takes ranked features to generate a predictive model and produce estimated output variables. Finally evaluation metrics are exploited to assess effectiveness of proposed method.

DBN incorporates simple learning modules: Restricted Boltzman Machines (RBMs), which consist of visible and hidden layers that represent features. These hidden and visible layers are connected by symmetrical weights. Input layer is represented by h^0 and last hidden layer, h^l, computes the output by utilising the output of previous layer $h^l - 1$. Therefore, output can be calculated from the following formula [169]:

$$h^l = \varphi(b^l + W^l + h^{l-1}) \tag{5.1}$$

where b^l a vector of offsets, W^k a matrix of weights, and φ is the activation function. The output layer is appropriate to make predictions. For quantitative prediction or regression tasks the output is:

$$h^l = \alpha_{0k} + \alpha_k \varphi(b_i^l W_i^l h^{l-1}) \tag{5.2}$$

where W_i^l is the ith row of W^l, α_{0k} is the bias, and α_k represents a set of weights between the last and next to last layers. The probability of visible and hidden neuron vectors for DBN can be calculated by:

$$P(v, h^1, ..., h^l) = P(h^{l-1}, h^l)(\prod_{k=0}^{l-2} P(h^k | h^{k+1}) \tag{5.3}$$

where $P(h^{k-1}|h^k)$ is a conditional probability for the visible units conditioned on the hidden units of the RBM at level k, v is vector of visible units, and $P(h^l - 1, h^l)$ represents joint distribution in the top level which is RBM. A general representation of a DBN with an input and l hidden neurons is demonstrated in Fig. 5.2. The last two layers comprise an RBM. Weight updates for a single RBM are performed with a gradient descent or accent; the difference is the sign which is plus or minus, utilised to perform update.

$$\Delta W_{ij}(t+1) = W_{ij}(t) + \epsilon \frac{\partial logp(v)}{\partial W_{ij}} \tag{5.4}$$

where $p(v)$ is probability of a visible vector, ϵ is a parameter with a small value , and $\frac{\partial logp(v)}{\partial W_{ij}}$ is the gradient which can also be calculated as [180]:

$$\epsilon \frac{\partial logp(v)}{\partial W_{ij}} = \epsilon < x_i, h_j >_{data} - < v_i, h_j >_{model} \tag{5.5}$$

where $<>_p$ represents averages with respect to distribution p.

In RBM, weight updates are defined by Equation (5.4). The probability of visible vector, $p(v)$, can be calculated from:

$$p(v) = \frac{1}{Z} \sum_h e^{-E(v,h)} \tag{5.6}$$

where Z is the partition function and $E(v,h)$ is the energy function assigned to the state of the network. Therefore, probability of each pair of hidden and visible vectors can be defined as:

$$p(v,h;\theta) = \frac{1}{Z(\theta)} \sum_{h} e^{-E(v,h;\theta)} \tag{5.7}$$

where $Z(\theta) = \sum_h \sum_v (-E(v,h;\theta))$ and the energy function is:

$$E(v,h) = a^T h - b^T v - v^T w h = -\sum_i b_i v_i - \sum_j a_j h_j - \sum_{i,j} w_{ij} v_i h_j \tag{5.8}$$

where a_i and b_i are bias of visible inputs, v_i and hidden variables, h_j, respectively and w_{ij} are weights between units of layers. By utilising Equation (5.8) Equation (5.7) can be rewritten as [181]:

$$\begin{aligned} p(v,\theta) &= \sum_h \frac{e^{-E(v,h;\theta)}}{\sum_{v,h} e^{-E(v,h;\theta)}} \\ &= \frac{1}{Z(\theta)} \sum_h exp(v^T w h + b^T v + a^T h) \\ &= \frac{1}{Z(\theta)} exp(b^T w) \prod_j^F \sum_{h_j \in 0,1} exp(a_j h_j + \sum_1^D w_{ij} v_i h j) \\ &= \frac{1}{Z(\theta)} exp(b^T w) \prod_j^F (1 + exp(a_j + \sum_1^D w_{ij} v_i)) \end{aligned} \tag{5.9}$$

By utilising the energy function the following equations can be defined:

$$p(v|h;\theta) = \prod P(v_i|h) \quad and \quad P(v_i = 1|h) = \varphi(b_j + \sum h_i w_{ij}) \tag{5.10}$$

$$p(h|v;\theta) = \prod P(h_j|v) \quad and \quad P(h_j = 1|v) = \varphi(a_j + \sum v_j w_{ij}) \tag{5.11}$$

where φ is the sigmoid function which can be calculated from:

$$\varphi(x) = \frac{1}{1 + exp(-x)} \tag{5.12}$$

However, the energy function is not applicable for regression tasks where continuous data is used. Therefore, RBM needs to be modified in order to deal with regression tasks. The energy function can be revised by replacing binary inputs with linear units with independent Gaussian noise, so that RBM can handle continuous-valued data [182]. This method is called as Gaussian-Bernoulli Restricted Boltzmann Machines (GBRBMs) [183]. The energy function for real-valued data can be calculated from:

$$E(v, h; \theta) = \sum_{i} \frac{(v_i - b_i)^2}{2\sigma_i^2} - \sum_{j} a_j h_j - \sum_{i,j} \frac{w_{ij} v_i h_j}{\sigma_i} \tag{5.13}$$

where $\theta = \{W, a, b, \sigma^2\}$ is a vector and σ_i is the variance of visible or input variable v_i.

After modifying RBM, DBN is capable of handling real-valued data. The proposed model takes given data as input to DBN, and DBN generates RBM weight matrix, W, of dimension (number of hidden units, number of inputs). Then, DFSFR assigns feature weights, G, according to following formula:

$$G_j = \frac{\sum_{i=1}^{d} W_i}{h} \tag{5.14}$$

where d is number of features, h is number of hidden neurons, and W represents a weight vector. Finally, SVR or MSVR takes the vector G, performs regression and calculates the prediction performance of the model by utilising evaluation metrics, e.g., RMSE.

5.4 A Hybrid Unsupervised Feature Selection Method (DKBFS)

In Chapter 4, the KBFS framework is presented. In this chapter, a novel deep learning based unsupervised feature selection method, DFSFR, is proposed. Experimental results, which are shown in next section, conclude that the proposed methods produced better results than the state of the art unsupervised feature selection methods. The KBFS method is utilised for the GSE44763 and GSE40279 data sets, which are considered to be ultra high dimensional. However, since GSE44763 and GSE40279 data sets are considered to be ultra high dimensional,

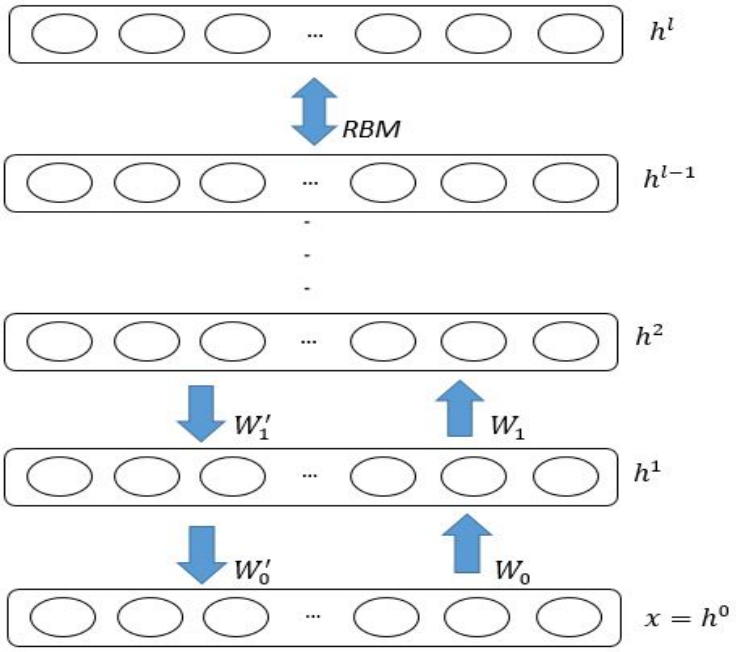

FIGURE 5.2: General Representation of DBN.
(The top two layers constitute an RBM. Ws represent weights between units of layers and $W's$ are the transpose of Ws).

KBFS is utilised as a pre-processing step of DFSFR method. Therefore, a hybrid method that combines both KBFS and DFSFR is proposed and abbreviated as DKBFS, is generated. This hybrid method has achieved the best results on GSE40279 data set and produced the second best result on GSE44763 data set (the best result is achieved by DFSFR). The experimental results are conducted on the GSE44763 and GSE40279 data set and presented in next section.

DKBFS integrates KBFS and DFSFR methods where KBFS is used as a pre-filtering step for KBFS. User defined number of features are eliminated by using KBFS and selected features are exploited as input variables for DFSFR. Then, DFSFR generates the weights of features. Weighted features are then used as input variables of SVR to construct a predictive model.

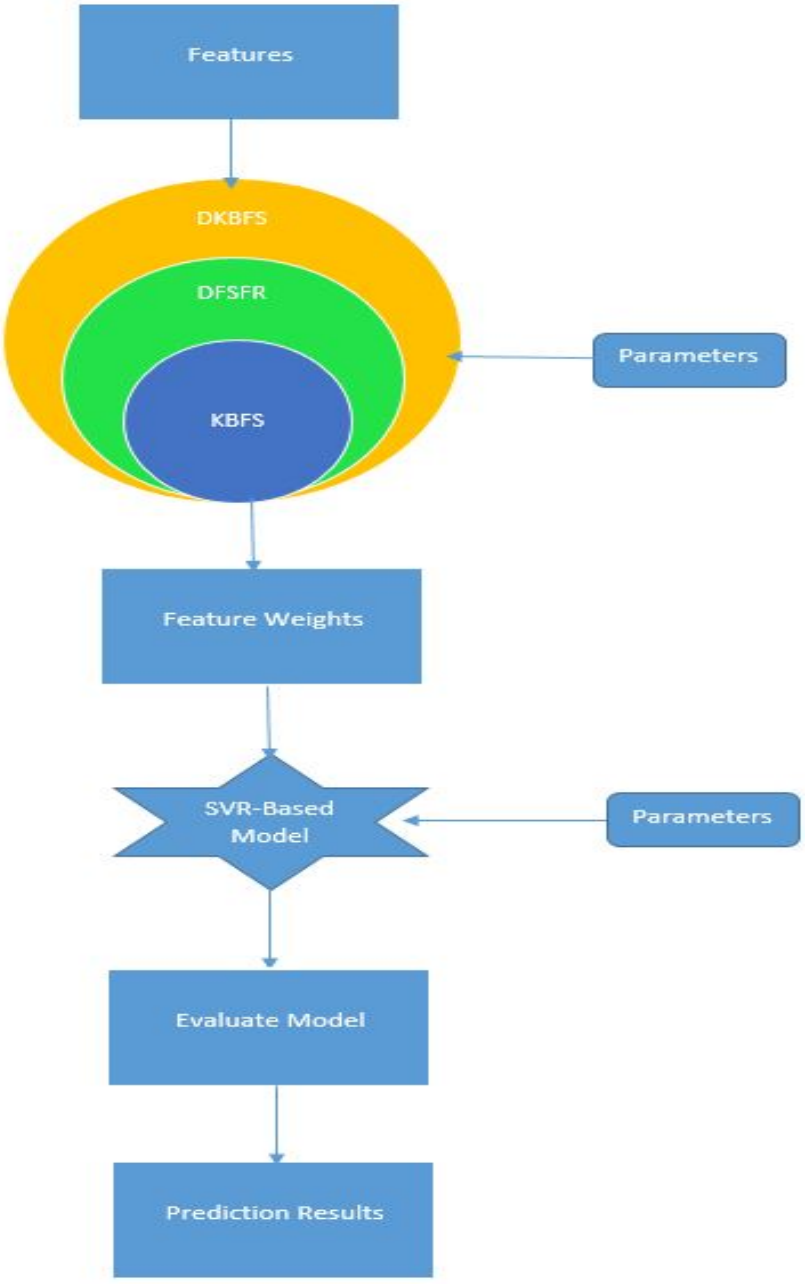

FIGURE 5.3: The Flowchart of DKBFS.
(DKBFS is a hybrid unsupervised feature selection method where KBFS is embedded into DFSFR)

5.5 Results

This chapter presents the results of the application of the proposed DFSFR framework compared to the state-of-the-art unsupervised feature selection techniques as well as the baseline (entire feature set) with the RV144 Vaccine, peptide binding affinity, GSE44763, and GSE40279 data sets. DKBFS method is compared with the state-of-the-art unsupervised feature selection methods for GSE44763 and GSE4027 data sets because this method is developed to deal only with ultra high dimensional data.

5.5.1 Results for RV144 Vaccine Data Set

As mentioned in chapter 3, the RV144 data set provided in [119] is used in this study to model the antibody feature-function relationship. This data set contains 100 plasma samples (20 of which are placebo and 80 are vaccine-injected) obtained from the individuals participating in the RV144 vaccine trial at week 26. Three different cell-mediated assays are used: Antibody Dependent Cellular Phagocytosis; Antibody Dependent Cellular Cytotoxicity; and Natural Killer Cell Cytokine Release activities. The accuracy results for the proposed DFSFR framework are compared with those presented in a previous study [119], and are also compared with results from four different state-of-the-art unsupervised feature selection methods, namely MCFS, InFS, LapFS, and SPFS, along with the entire feature set. In this study, the PCC and RMSE metrics are used so as to analyse the performance of unsupervised feature selection algorithms. The PCC metric is used to be able to perform a consistent comparison with the previous study [119]. The RMSE measure is exploited to compare the performance of the predictive models for performing MISO and MIMO regression tasks. SVR and MSVR are utilised to perform MISO and MIMO regression tasks respectively.

The SVR-based predictive models for the regression tasks are constructed using feature selection methods (filtered feature set). Their performance is then evaluated using a five-fold cross validation method. The RV144 data set is divided into two sets of samples. Four out of five samples, with a total of 64 samples, are utilised for training and the rest (16 samples) for testing purposes. This process is repeated 200 times by randomly creating subsets of the samples for the five-fold cross validation in order to avoid a bias towards and to assess the effect

of randomisation in the cross validation. At the end, the mean performance and its corresponding standard deviation (std) values are obtained for each of the predictive models.

The prediction performance of unsupervised feature selection methods on three cell-mediated assays are summarised in Tables 5.1-5.3. Table 5.1 shows the PCC and RMSE results of predictive models for Natural Killer cell Cytokine release activities. The predictive models aim to estimate the level of cytokine release in order to understand its functionality for protection. The results suggest that DFSFR outperforms state-of-the-art methods with 0.54 PCC using 16 features. SPEC yields the second-best result, at 0.51 PCC, with 16 antibody features. Other methods produce average results.

The prediction results of unsupervised predictive models for ADCC activities are presented in Table 5.2. DFSFR again achieves the best PCC result yielding 0.48 using only 1 antibody feature. InFS produces the second-best result with 0.40 PCC utilising 14 antibody features. Other methods produce average results.

Table 5.3 presents the prediction results of USFSMs for ADCP activities. As can be clearly seen in the table, the predictive models that have used the other USFSMs yielded poorer results than the DFSFR filtered predictive model. Moreover, the DFSFR filtered predictive model outperforms the predictive models implemented with the complete feature set. DFSFR achieves the best prediction accuracy, yielding 0.66 PCC with 13 antibody features. Laplacian Score and MCFS produce the same PCC results with 3 and 17 antibody features respectively. It is observed that the RMSE results of the predictive models are slightly different from their PCC results. DFSFR produces the best RMSE results for ADCP and ADCC assays, at 27.8 and 5.42, respectively; on the other hand, MCFS yields the best result for the Cytokine assay giving 1.93 RMSE.

The prediction results of the proposed method are also compared with those of the previous study [119] where the same data set by using the same cross validation method is utilised (5-fold with 200 replicates) in order to carry out realistic and consistent comparison. The results are shown in Tables 5.4-5.6. The results appear to suggest that DFSFR has a better quantitative accuracy than the predictive models constructed using Lars, GP and SVR as presented in the previous study for ADCC and ADCP assays, at 0.48 and 0.66 PCC respectively. In particular, the proposed approach yields as much as 1.17x and 3.4x better

TABLE 5.1: Comparison of Unsupervised Feature Selection Methods for the Antibody Features and Natural Killer Cell Cytokine Release Activity Relationship.

Metrics	PCC	RMSE
DFSFR (16)	$\mathbf{0.54 \pm 0.15}$	1.96 ± 0.69
MCFS (16)	0.49 ± 0.17	$\mathbf{1.93 \pm 0.67}$
Laplacian (16)	0.49 ± 0.18	1.94 ± 0.70
KBFS (16)	0.52 ± 0.17	1.95 ± 0.71
SPEC (16)	0.51 ± 0.17	2.05 ± 0.68
InFS (18)	0.49 ± 0.17	2.04 ± 0.74
Baseline (20)	0.49 ± 0.17	2.04 ± 0.7

TABLE 5.2: Comparison of Unsupervised Feature Selection Methods for the Antibody Features and Cellular Cytotoxic Activity Relationship.

Metrics	PCC	RMSE
DFSFR (1)	$\mathbf{0.48 \pm 0.17}$	5.42 ± 0.87
MCFS (18)	0.39 ± 0.18	5.42 ± 0.97
Laplacian (12)	0.39 ± 0.18	5.42 ± 0.93
KBFS(10)	0.39 ± 0.19	5.47 ± 0.99
SPEC (18)	0.41 ± 0.18	5.44 ± 0.92
InFS (14)	0.40 ± 0.17	5.48 ± 0.98
Baseline(20)	0.38 ± 0.18	5.6 ± 0.98

TABLE 5.3: Comparison of Unsupervised Feature Selection Methods for the Antibody Features and Cellular Phagocytosis Activity Relationship.

Metrics	PCC	RMSE
DFSFR (13)	$\mathbf{0.66 \pm 0.14}$	$\mathbf{27.8 \pm 3.65}$
MCFS (17)	0.65 ± 0.14	31.9 ± 3.86
Laplacian (3)	0.65 ± 0.15	31.3 ± 3.81
KBFS(12)	0.65 ± 0.17	0.30 ± 3.84
SPEC (18)	0.61 ± 0.14	32.7 ± 4.03
InFS (18)	0.64 ± 0.15	32.2 ± 3.97
Baseline(20)	0.61 ± 0.15	33.1 ± 3.62

outcomes than the results of SVR for the ADCP and ADCC assays respectively. DFSFR has slightly lower quantitative performance as compared to the predictive model for the Cytokine assay constructed using SVR as presented in the previous study. However, it still has better quantitative performance than the Lars and GP predictive models for the Cytokine assay. Furthermore, DFSFR produces the best result with the least standard deviation (0.14) that means that proposed method produces more stable results than those of previous study.

Overall, the proposed DFSFR framework achieves the best performance on all

TABLE 5.4: A Comparison of the Results with the Previous Study for the Antibody Features and Cellular Phagocytosis Activity Relationship.

Regression	PCC
Lars [119]	0.61±0.15
GP [119]	0.53±0.16
SVR [119]	0.56±0.19
DFSFR	**0.66 ±0.14**

TABLE 5.5: A Comparison of the Results with the Previous Study for the Antibody Features and Cellular Cytotoxic Activity Relationship.

Regression	PCC
Lars [119]	0.42±0.18
GP [119]	0.24±0.21
SVR [119]	0.14±0.24
DFSFR	**0.48±0.17**

TABLE 5.6: A Comparison of the Results with Previous Study for the Antibody Features and Natural Killer Cell Cytokine Release Activity Relationship.

Regression	PCC
Lars [119]	0.51±0.21
GP [119]	0.46±0.24
SVR [119]	**0.55± 0.15**
DFSFR	0.54±0.15

cell-mediated assays, which thereby verifies that it is is able to select informative antibody features. In order to develop an effective HIV vaccine, specific antibodies which fight against HIV should be identified.

5.5.1.1 Results for Multi-Input-Single-Output (MISO) and Multi-Input-Multi-Output (MIMO) Regression

A comparison of prediction results of predictive models for MISO and MIMO regression tasks is shown in Table 5.7. The results of SVR-based models for the Cytokine, ADCC, and ADCP assays are listed in Table 5.7. The average RMSE results of SVR-based models are calculated by taking the mean of the RMSE results for each assay. The results suggest that the MSVR-based predictive model outperforms the SVR-based predictive model, which indicates that some correlations exist amongst the target variables. In the previous study [119], these MIMO regression correlations are not taken into account. Analysing dependencies between antibody features as well as response variables (the functional activities

of antibody features) may ultimately result in producing an effective vaccine so that HIV or AIDS may be conquered.

TABLE 5.7: A comparison of Unsupervised Prediction Results for SVR and MSVR for Anticipating Antibody Feature-Function Relationship.

Metrics	DFSFR	MCFS	LapFS	SPEC	InFS
SVR (Cytokine)	1.96 ± 0.69	1.93 ± 0.67	1.94 ± 0.70	2.05 ± 0.68	2.04 ± 0.74
SVR (ADCC)	5.42 ± 0.87	5.42 ± 0.97	5.42 ± 0.93	5.44 ± 0.92	5.48 ± 0.98
SVR (ADCP)	27.8 ± 3.65	31.9 ± 3.86	31.3 ± 3.81	32.7 ± 4.03	32.2 ± 3.97
SVR (Average)	11.72 ± 1.78	13.08 ± 1.85	12.88 ± 1.97	13.83 ± 1.85	13.24 ± 1.83
MSVR	10.42 ± 1.65	13.01 ± 1.77	12.07 ± 1.70	12.30 ± 1.68	12.83 ± 1.74

5.5.1.2 Additional Results and Discussion

A summary of the results of predictive models for three cell-mediated assays is presented showing comparative analyses of USFSMs for results for the Cytokine, ADCC and ADCP assays in Figs 5.4, 5.5, and 5.6, respectively. They share 5 antibody features, namely, IgG2.gp41, IgG3.gp140, IgG3.p24, IgG4.p24 and IgG4.p120 in their filtered sets.

In this study, antibody features which are mutually selected by unsupervised feature selection methods are also examined. The antibody features which are commonly selected by unsupervised feature selection methods for the Cytokine, ADCC, and ADCP assays are shown in Table 5.8. There is only one common feature selected by USFSMs for each of the ADCC and ADCP assays, which are IgG1.gp41 and IgG3.p24 respectively. On the other hand, USFSMs share seven antibody features for the Cytokine assay, namely, IgG1.p2, IgG3.p24, IgG4.gp41, IgG4.gp41, IgG4.gp140, IgG4.p24, IgG4.gp120, IgG3.V1V2. The antibody feature IgG3.p24 is selected by unsupervised feature selection methods for the Cytokine and ADCP cell-mediated assays. Interestingly, none of the individual antibody feature is selected in all assays. Each effector function performs different tasks to fight against antigens, and specific antibodies provide specific protection against specific antigens. This might be one reason why no single antibody feature is selected by all USFSMs. Another reason for this might be that the DFSFR method achieves the best performance for the ADCC assay by utilising only one antibody feature. If any unsupervised method does not select

this feature, then there will be no universally selected feature for the Cytokine assay.

The best subset of antibody features, which provides the best predictive performance for all cell-mediated assays, is identified in this study. These antibody features are listed in Table 5.9. A distribution of antibody features based on their importance is provided in Fig. 5.7. Antibody features are given scores ranging from 5-100 based on their importance. The set of antibody features with values greater than 20 constitutes the best feature subset.

TABLE 5.8: Selected Mutual Features for Unsupervised Learning.

Cytokine	ADCC	ADCP
IgG1.p24	IgG1.gp41	IgG3.p24
IgG3.p24	-	-
IgG4.gp41	-	-
IgG4.gp140	-	-
IgG4.p24	-	-
IgG4.gp120	-	-
IgG3.V1V2	-	-

(There is only one common feature selected by unsupervised feature selection methods for ADCP and ADCC assays. Seven different antibody features are mutually selected by unsupervised methods for cytokine assay).

TABLE 5.9: The Best Subset of Features for all of the Cell-Mediated Assays.

Antibody Features
IgG1.gp41
IgG3.gp140
IgG4.gp120
IgG3.p24
IgG3.V1V2
IgG2.V1V2
IgG4.p24
IgG2.gp140
IgG2.gp41
IgG4.gp41
IgG4.gp140
IgG2.p24
IgG1.gp120
IgG1.p24
IgG1.gp140
IgG4.V1V2

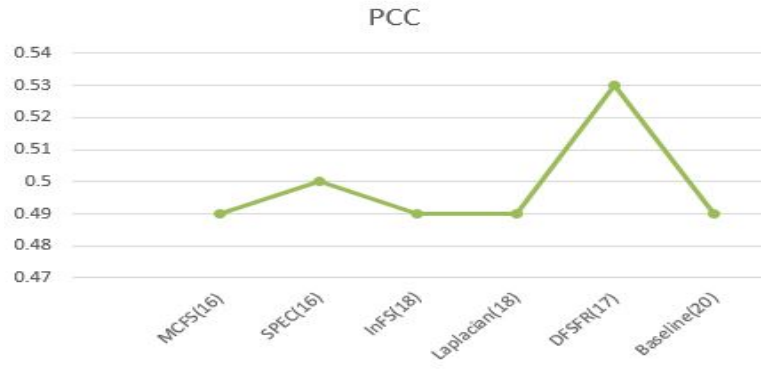

FIGURE 5.4: Selected Number of Features and Their Corresponding PCC Results for the Cytokine Assay

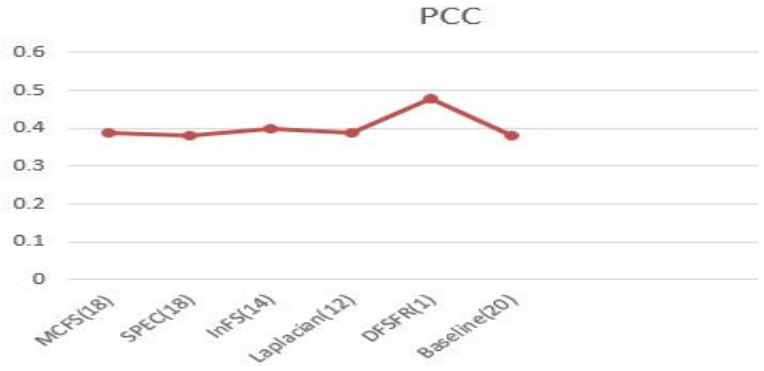

FIGURE 5.5: Selected Number of Features and Their Corresponding PCC Results for the ADCC Assay

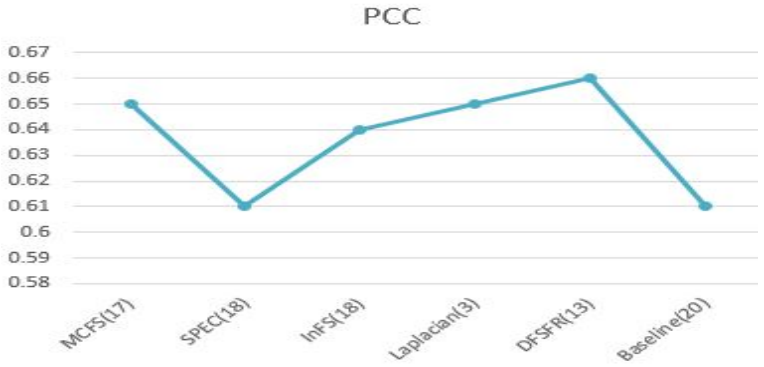

FIGURE 5.6: Selected Number of Features and Their Corresponding PCC Results for ADCP Assay

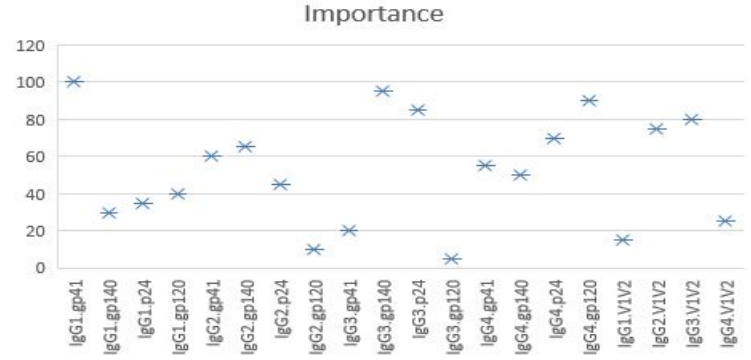

FIGURE 5.7: Distribution of Antibody Features Based on Their Importance

5.5.2 Results for Peptide Binding Affinity Data Sets

As mentioned in chapter 3, three different high dimensional peptide data sets, provided in the CoEPrA modelling competition [129], are used. Each data set consists of training and test data sets, therefore, there is no need for cross validation. Tasks 1 and 3 contain nona-peptides that contain a total of 5787 amino acid descriptors. Task 1 consists of 89 training and 88 testing samples, whereas Task 3 has 133 training and 133 testing instances. Task 2 consists of octa-peptides with a total of 5144 amino acid descriptors. It contains 76 training and 76 testing samples.

The prediction performance of the proposed DFSFR framework for Tasks 1, 2 and 3 are compared with five different USFSMs, namely MCFS, KCEN, EUFS, LapFS and SPFS, along with the entire feature set (baseline). The prediction performance of unsupervised feature selection methods over different tasks is summarised in Tables 5.10-5.12. In order to investigate the robustness of the USFSMs, their default parameters are exploited. The number of selected features is initially 50 and then incremented by 50 to form feature sets of {50, 100, ..., 250, 300}. Table 5.10 demonstrates the performance of the USFSMs for Task 1. The results suggest that DFSFR produces the best results for MAD, MSE, RMSE, MAPE, U and q^2 metrics with 100 selected features. The second best results are achieved by SPEC with 300 features. Other methods produce average results.

A comparative analysis of USFSMs for Task 2 is shown in Table 5.11. The results of the experiment with the Task 2 data set confirm that DFSFR achieves the best results for all metrics, yielding 0.27 MAD, 0.16 MSE, 0.39 RMSE, 3.8 MAPE, 0.006 U and $0.71q^2$ with 250 features. SPFS produces the second-best results, with 0.28 MAD, 0.17 MSE, 0.41 RMSE, 3.9 MAPE, 0.007 U, and 0.68 q^2 with 300 features. The results for SPFS are very similar to those for DFSFR, but the latter achieves the best results using 250 features while SPFS produces the second-best results with 300 features. Other USFSMs produce average results.

TABLE 5.10: Regression Results of the Unsupervised Feature Selection Methods for Task 1

Metrics	MAD	MSE	RMSE	MAPE	U	q2
DFSFR(100)	**0.39±0.0**	**0.27±0.0**	**0.52±0.0**	**7.84±0.0**	**0.017±0.0**	**0.69±0.0**
EUFS(300)	0.49±0.0	0.53±0.0	0.73±0.0	9.86±0.0	0.024±0.0	0.46±0.0
KCEN(200)	0,51±0,04	0.48±0.07	0.69±0.05	10.1±0.76	0.023±0.0	0.44±0.82
MCFS(50)	0,57±0.0	0.54±0.0	0.74±0.0	11.4±0.0	0.025±0.0	0.37±0.0
LapFS(300)	0,58±0.0	0.61±0.0	0.78±0.0	11.3±0.0	0.027±0.0	0.30±0.0
SPFS(300)	0,50±0.0	0.37±0.0	0.61±0.0	9.7±0.0	0.020±0.0	0.57±0.0
Baseline	1,07±0.0	1.82±0.0	1.35±0.0	21±0.00	0.043±0.0	-1.0±0.0

A comparative analysis of USFSMs for Task 3 is shown in Table 5.12. The proposed approach clearly generates the best results, yielding 0.54 MAD, 0.48 MSE, 0.69 RMSE, 7.96 MAPE, 0.24 q^2 and 0.013 U. The results for EUFS and LapFS are similar. They both produce 0.6 MAD, 0.58 MSE, 0.76 RMSE; however, EUFS yields 9 MAPE, 0.079 q^2 and 0.014 U whereas LapFS achieves 8.6 MAPE, 0.081 q^2 and 0.015 U.

Given these analyses, all the results present a clear message that the SVR-based predictive model with all the features fails. This outcome suggests the necessity of feature selection. It is also observed that the performance of the USFSMs is relatively sensitive to the number of selected features. The number of selected features is provided in parenthesis located just next to the USFSM results in the tables.

One of the most important observations is the consistency of the results over six different metrics. For example, the proposed DFSFR method produces the best results on all different tasks over different metrics. These results indicate that the performance of USFMs does not seem to differ that much.

The results of proposed DFSFR framework are also compared with those of our earlier study [88] which was conducted on the same peptide data sets. Tables

5.13-5.15 show the prediction results of both the proposed method and our earlier study for Tasks 1-3.

TABLE 5.11: Regression Results of the Unsupervised Feature Selection Methods for Task 2

Metrics	MAD	MSE	RMSE	MAPE	U	q2
DFSFR(250)	**0.27±0.0**	**0.16±0.0**	**0.39±0.0**	**3.8 ±0.0**	**0.006±0.0**	**0.71±0.0**
EUFS(100)	0.39±0.0	0.43±0.0	0.65 ±0.0	5.98 ±0.0	0.011±0.0	0.2±0.0
KCEN(200)	0.35 ±0.0	0.27±0.0	0.52±0.0	5±0.0	0.023±0.0	0.49±0.0
MCFS(300)	0.32 ±0.0	0.2±0.0	0.45±0.0	4.6 ±0.0	0.009±0.0	0.62±0.0
LapFS(300)	0.35 ±0.0	00.29 ±0.0	0.54±0.0	5.1 ±0.0	0.009±0.0	0.45±0.0
SPFS(300)	0.28 ±0.0	0.17±0.0	0.41±0.0	3.9±0.0	0.007±0.0	0.69±0.0
Baseline	0.29±0.0	0.16±0.0	0.4±0.0	4.02±0.0	0.007±0.0	0.7±0.0

TABLE 5.12: Regression Results of the Unsupervised Feature Selection Methods for Task 3

Metrics	MAD	MSE	RMSE	MAPE	U	q2
DFSFR(200)	**0.54**±0.0	**0.48**±0.0	**0.69**±0.0	**7.96**±0.0	**0.013**±0.0	**0.24**±0.0
EUFS(150)	0.61 ± 0.0	0.58 ± 0.0	0.76 ± 0.0	9 ± 0.0	0.014 ± 0.0	0.07 ± 0.0
KCEN(300)	0.66 ± 0.0	0.67 ± 0.0	0.81 ± 0.0	9.7 ± 0.0	0.016 ± 0.0	−0.06 ± 0.0
MCFS(50)	0.7 ± 0.0	0.76 ± 0.0	0.87 ± 0.0	10.1 ± 0.0	0.017 ± 0.0	−0.20 ± 0.0
LapFS(50)	0.6 ± 0.0	0.58 ± 0.0	0.76 ± 0.0	8.6 ± 0.0	0.015 ± 0.0	0.08 ± 0.0
SPFS(300)	0.67 ± 0.0	0.75 ± 0.0	0.86 ± 0.0	9.9 ± 0.0	0.017 ± 0.0	−0.18 ± 0.0
Baseline	1.17 ± 0.0	2.51 ± 0.0	1.58 ± 0.0	17 ± 0.0	0.031 ± 0.0	−2.97 ± 0.0

Table 5.13: Regression Results of DFSFR and the Previous Study for Task 1

Metrics	(DFSFR (100)/Previous (300))
MAD	0.39/0.50
MSE	0.27/0.37
RMSE	0.52/0.61
MAPE	7.84/9.7
q2	0.693/0.575
U	0.017/0.02

(The prediction performance of the DFSFR and a previous study. Number of selected features are shown in parenthesis just next to the feature selection method.)

Table 5.14: Regression Results of the Proposed DKBFS Method and the Previous Study for Task 2

Metrics	(DFSFR (250)/Previous (300))
MAD	0.27/0.28
MSE	0.16/0.17
RMSE	0.39/0.41
MAPE	3.8/3.9
q2	0.71/0.69
U	0.006/0.007

(The prediction performance of the DFSFR and the previous study. Number of selected features are shown in parenthesis just next to the feature selection method.)

TABLE 5.15: Regression Results of DKBFS and Previous Study for Task 3

Metrics	(DFSFR (200)/Previous (50))
MAD	0.54/0.60
MSE	0.48/0.58
RMSE	0.54/0.6
MAPE	7.96/8.6
q2	0.24/0.081
U	0.013/0.017

It can be seen from the Tables 5.13-5.15 that the proposed method produces better results than those of the previous study for all different tasks over all of the different metrics.

5.5.3 Results for the GSE44763 Data Set

As mentioned in chapter 3, the GSE44763 data set [130] is utilised to model the associations among CpG biomarkers (features), chronological age and obesity. This data set contains 27482 Cytosine-phosphate-Guanine (CpG) biomarkers from the peripheral blood of 46 adult female donors (samples). There are 24 obese subjects and 22 lean subjects. In this study, a subject is considered obese if his/her BMI is greater than or equal to 30, and a subject is considered as lean if his/her BMI is less than 25. In order to investigate the robustness of the USFSMs their default parameters are used. The number of selected features is initially 50 and then incremented by 50 to form feature sets of {50, 100, ..., 250, 500}.

The performance of the proposed DFSFR and DKBFS methods are compared with the state-of-the-art USFSMs, including EUFS, InFS, LapFS, and SPFS along with the entire feature set. In order to evaluate the robustness of USFSMs, support vector-based methods are used since their effectiveness has been proven and they provide better generalisation and performance in a wide range of bioinformatics applications [102] [14]. To observe the results for these methods using different metrics, three different metrics are used to assess the quality of the USFSMs, which are Mean Absolute Deviation (MAD), Root Mean Squared

Error (RMSE) and Theils U-statistics (U). The RMSE metric is utilised to calculate prediction errors for both MISO and MIMO regression tasks. SVR and MSVR are exploited to perform MISO and MIMO regression tasks, respectively. The prediction results of the predictive models are calculated and averaged with the five-fold cross validation method. Therefore, four out of five samples are used for training and the rest of the samples are utilised for testing purposes. The five-fold cross validation is repeated 200 times in order to gain more unbiased results. Then, the mean performance and its corresponding standard deviation (std) values are obtained for each of the predictive models.

In this study, Ilumina average beta values are utilised as numerical data where the Beta-value is the ratio of the methylated probe intensity and the overall intensity (sum of methylated and unmethylated probe intensities). Beta value for an ith investigated CpG island is determined as follows [139]:

$$Beta_i = \frac{max(y_{i,methy}, 0)}{max(y_{i,unmethy}, 0) + max(y_{i,unmethy}, 0) + \alpha} \tag{5.15}$$

where $y_{i,methy}$ and $y_{i,unmethy}$ are the intensities measured by the ith methylated and unmethylated probes respectively, and α is a constant offset which is added to the denominator in order to regularise the Beta value if unmethylated and methylated probe intensities are low. The default value of α is 100.

The prediction performance of USFSMs are summarised in Tables 5.16-5.18. Table 5.16 shows the robustness of USFSMs for the prediction of chronological age. The results suggest that the proposed DFSFR and DKBFS methods outperform the state-of-the-art unsupervised feature selection methods. DKBFS produces the best results yielding 7.81 MAD, 9.14 RMSE and 0.003 U with only 50 features. DFSFR achieves the second-best results, yielding 7.97 MAD, 9.17 RMSE and 0.003 U with 450 features. From these experimental results it can be concluded that the proposed DFSFR and DKBFS frameworks are able to disclose age-related CpG biomarkers (features) from the given data. Table 5.16 also indicates that the proposed DKBFS framework outperforms state-of-the-art USFSMs as well as DFSFR method. Other feature selection methods produce average results. Interestingly, all of the USFSMs produce similar U results; however, the results for different metrics are consistent. For example, DKBFS yields the best results for all different metrics.

TABLE 5.16: The Performances of USFSMs for Prediction of Chronological Age

Metrics	MAD	RMSE	U
DFSFR(450)	**7.97±1.39**	**9.17±1.36**	$\mathbf{\frac{3}{1000}\pm\frac{4}{10000}}$
DKBFS(50)	**7.81±1.29**	**9.14±1.24**	$\mathbf{\frac{3}{1000}\pm\frac{39}{100000}}$
EUFS(500)	8.24±1.91	9.51±1.41	$\frac{308}{100000}\pm\frac{42}{100000}$
InFS(500)	8.14±1.40	9.39±1.33	$\frac{306}{100000}\pm\frac{40}{100000}$
SPFS(350)	8.21±1.46	9.51±1.35	$\frac{309}{100000}\pm\frac{40}{100000}$
LapFS(150)	8.32±1.45	9.58±1.37	$\frac{311}{100000}\pm\frac{40}{100000}$
Baseline	8.12 ± 1.43	9.41 ± 1.37	$\frac{307}{100000}\pm\frac{41}{100000}$

TABLE 5.17: The Performances of USFSMs for the Prediction of BMI

Metrics	MAD	RMSE	U
DFSFR(200)	**6.43±1.04**	**7.23±1.06**	$\frac{85}{10000}\pm\frac{13}{10000}$
DKBFS (150)	6.58±1.05	7.34±1.15	$\mathbf{\frac{8}{1000}\pm\frac{1}{1000}}$
EUFS (400)	6.93±1.09	7.75±1.17	$\frac{9}{1000}\pm\frac{17}{10000}$
InFS(400)	6.93±1.09	7.75±1.16	$\frac{9}{1000}\pm\frac{17}{10000}$
SPFS (450)	6.99±1.52	7.52±1.51	$\frac{89}{10000}\pm\frac{21}{10000}$
LapFS (450)	6.98±1.56	7.5±1.55	$\frac{88}{10000}\pm\frac{2}{1000}$
Baseline	7.04 ± 1.62	7.62 ± 1.59	$\frac{89}{10000}\pm\frac{21}{10000}$

Surprisingly, the complete feature set (baseline) produces 8.12 MAD, 9.41 RMSE and 0.00307 U, and thereby yields better results than LapFS, SPFS, and EUFS. This outcome implies that most of the CpG biomarkers are related to aging. It is also observed that the performance of the USFSMs is relatively sensitive to the number of selected features. The number of selected features are shown in parenthesis in the tables.

A comparison of USFSMs for BMI prediction is shown in Table 5.17. The outcomes of the experiments clearly emphasise that the proposed DFSFR and DKBFS methods outperform state-of-the-art USFSMs. DFSFR produces the best results for MAD and RMSE yielding 6.43 and 7.23, respectively. However,

it produces the second-best result for U which is 0.0085. DKBFS achieves the second-best results for RMSE and MAD, which are 7.43 and 6.58 respectively. On the other hand, it achieves the best U results, yielding 0.008. DFSFR and DKBFS produce the best results by using 200 and 150 CpG biomarkers (features) respectively. Other feature selection methods produce average results.

It is observed that the results using different metrics are generally consistent. For example, DFSFR produces the best results and DKBFS achieves the second-best results with the RMSE and MAD metrics. Nevertheless, the results of the USFSMs for age prediction are slightly different than those for BMI prediction. For example, the baseline produces better results than EUFS, InFS, SPFS and LapFS for age prediction; however, it yields the worst results for BMI prediction. Furthermore, DFSFR achieves the best results for BMI prediction, but on the other hand produces the second-best results for age prediction. These results appear to suggest that, if a data set is multi-targeted, then USFSMs might produce different results for different targets especially if there is no correlation between the targets.

5.5.3.1 Results for Multi Input-Single Output (MISO) and Multi Input-Multi Output (MIMO) Regression

In this study, in addition to MISO regression, MIMO regression is performed to examine whether or not there is a relationship between age and obesity based on CpG biomarkers. A comparison of MISO and MIMO regression results is presented in Table 5.18. The results suggest that there is no strong correlation between obesity and aging based on the selected CpG dinucleotides (features). Therefore, most of the age-related CpG islands are not related to obesity. Interestingly, only the MSVR result for EUFS are better than its result for SVR. This result appears to suggest that some of the CpG biomarkers which are selected by EUFS are related to both aging and obesity.

TABLE 5.18: The Performances of USFSMs for MSVR and SVR

Metrics	MSVR	SVR
DFSFR(50)	8.55 ± 0.75	8.2 ± 1.21
DKBFS(150)	8.7 ± 1.85	8.24 ± 1.25
LapFS(150)	8.69 ± 0.85	8.54 ± 1.46
SPFS(350)	9.4 ± 0.83	8.51 ± 1.43
InFS(400)	9.05 ± 0.81	8.57 ± 1.24
EUFS(400)	8.45 ± 0.77	8.63 ± 1.29

5.5.4 Results for the GSE40279 Data Set

As mentioned in chapter 3, the GSE40279 data set provided in [145] is used to model the relationship between CpG biomarkers and chronological age. This data set contains 473034 CpG biomarkers (features) from the whole blood of 656 donors (samples) aged 19 to 101.

A pre-processing step is applied to map the data into lower dimensional space so that feature selection methods can be applied to the data set. First, the standard deviations of the samples, which refer to the amount of variation in data samples, are calculated. The standard deviation of a sample can only be zero if, and only if, the samples are identical. If a feature is identical in all samples, then the feature is not discriminative. Therefore, before applying feature selection, the features which have the lowest variation in the data are eliminated. As a result, approximately four out of five of the features are eliminated in this pre-processing step.

Then, unsupervised feature selection methods are applied to identify discriminative CpG biomarkers (features). The number of selected features starts from 900 in order that a subset of features contains at least 1% of the entire feature set. A set of 90000 features is assessed using six different USFSMs along with the entire feature set.

TABLE 5.19: A Comparison of USFSMs for The Prediction of Chronological Ages of Individuals using CpG Dinucleotides

Metrics	MAPE	q2	U	MAD
DKBFS(900)	**13.27±0.9**	**0.59±0.03**	**0.002±0.0001**	**7.3±0.58**
DFSFR(900)	13.41 ± 0.9	0.4 ± 0.04	0.0027 ± 0.0002	8.91 ± 0.66
EUFS(9000)	14.47 ± 0.69	0.003 ± 0.009	0.0034 ± 0.0001	11.69 ± 0.61
LapFS(6300)	14.43 ± 0.84	0.01 ± 0.01	0.0034 ± 0.0002	11.63 ± 0.73
TV(900)	14.16 ± 0.90	0.06 ± 0.02	0.0033 ± 0.0002	11.34 ± 0.79
Baseline(90000)	14.61 ± 0.88	0.003 ± 0.01	0.0034 ± 0.0002	11.86 ± 0.79

The performance of the proposed DFSFR and DKBFS methods with the GSE40279 data set is compared with that of state-of-the-art unsupervised feature selection methods, including EUFS, LapFS, and Term Variance (TV) along with the entire feature set.

Support vector based models [167] are exploited to assess the quantitative prediction performances of unsupervised feature selection methods since they have achieved superior generalisation and performance in a large variety of bioinformatics applications [102] [16]. Support vector based predictive models for regression tasks are constructed using USFSMs (filtered feature set) and the complete feature set. As there is no separate training and test data sets 8-fold cross validation is used to evaluate the performance of the predictive models. The cross validation is repeated 50 times by randomly creating subsets of the instances for the 8-fold cross validation to avoid bias towards and alleviate the impact of the random split. The means and standard deviations of the metrics are calculated over these 50 runs and presented in Table 5.19. The number of selected features for each predictive model is shown in parenthesis in the tables.

The results appear to suggest that the proposed DFSFR and DKBFS methods achieve better results than those of other USFSMs over different metrics. DKBFS produces the best results achieving 13.27 MAPE, 0.59 q2, 0.002 U and 7.3 MAD with only 900 CpG biomarkers (features). DFSFR achieves the second-best results, yielding 13.41 MAPE, 0.4 q2, 0.0027 U and 8.91 MAD with 900 CpG dinucleotides. TV produces 14.16 MAPE, 0.06 q^2, 0.0033 U, and 11.34 MAD

with 900 features and outperforms LapFS and EUFS. LapFS and EUFS which produce average results.

Another important observation is that the results over four different metrics are consistent. For example, DKBFS produces the best results and DFSFR yields the second best results for all different metrics.

Given this analysis, all of the results present a clear message that the SVR-based predictive model with all of the features fails. This outcome suggests the necessity of feature selection. It also proves that the majority of CpG biomarkers are not related to the determination of an individual's chronological age.

5.5.4.1 An Aggressive Research of Features from GSE40279 Data Set

The experimental results of the experiment conducted with the GSE4079 data set suggest that the proposed DFSFR, and DKBFS frameworks produce better results than other USFSMs for all different metrics. However, even though the number of features are drastically reduced from 473034 to 900, the number of CpG biomarkers (features) are still too high to be easily analysed in real biology laboratories. In this case, an aggressive research study is been conducted with three different subsets of CpG dinucleotides selected by DKBFS and DFSFR. The purpose of this aggressive process is to obtain the minimum number of dinucleotides which represent the whole data set with the same or higher accuracy so that they can be further analysed in real biology labs.

The number of features used starts from 1 and is then incremented by 1 until 900 is reached to make an aggressive reduction of the selected CpG biomarkers. As shown in Table 5.20, DFSFR achieves 0.57 q2, 0.0022 U and 7.41 MAD with 501 CpG dinucleotides (features), and DKBFS yields 0.61 q2, 0.002 U and 7.2 MAD with 669 CpG biomarkers (features).

It is observed that the performance of the feature selection methods is readily affected by the number of selected features. The number of selected features is provided in parenthesis located just next to the USFSM results in the table. For example, DKBFS achieves the best performance by utilising MAD, U and q^2 metrics with a dimensionality of 669.

TABLE 5.20: Detailed Assessment of CpG Dinucleotides Using the Proposed KBFS framework

Method	MAPE	q2	U	MAD
DKBFS (669)	13.39 ± 1.46	0.61 ± 0.03	$\&0.002 \pm 0.00001$	7.2 ± 0.52
DKBFS (900)	13.27 ± 0.9	0.59 ± 0.03	0.002 ± 0.0001	7.3 ± 0.58
DFSFR (501)	13.58 ± 1.22	0.57 ± 0.03	0.0022 ± 0.0001	7.41 ± 0.38
DFSFR (900)	13.41 ± 0.9	0.4 ± 0.04	0.0027 ± 0.0002	8.91 ± 0.66

5.6 Summary

In this chapter, a deep learning based unsupervised feature selection method, DFSFR, and a hybrid method, DKBFS, for regression tasks are proposed. To the best of our knowledge, the proposed DFSFR method is the first deep learning based feature selection method which selects features at input level. The proposed framework is capable of handling both MISO and MIMO regression tasks. The DKBFS method is a hybrid method that embeds KBFS into DFSFR algorithm to rank features. The KBFS method is used as a pre-filtering step for DFSFR. The flowchart of the proposed DKBFS method is shown in Fig. 5.3. Experimental studies have been conducted on different data sets and the results are presented in this chapter. Experimental results are used to demonstrate the robustness of the proposed methods. This results suggest that the proposed DFSFR and DKBFS methods outperform the -state-of-the-art USFSMs over different data sets.

It is observed that the results using different metrics are generally consistent. For example, DFSFR produces the best results and KBFS achieves the second-best results with the RMSE and MAD metrics. Nevertheless, the results of the USFSMs for age prediction are slightly different than those for BMI prediction. For example, the baseline produces better results than EUFS, InFS, SPFS and LapFS for age prediction; however, it yields the worst results for BMI prediction. Furthermore, DFSFR achieves the best results for BMI prediction, but on the other hand produces the second-best results for age prediction. These results appear to suggest that, if a data set is multi-targeted, then USFSMs might produce different results for different targets especially if there is no correlation between the targets.

In this study, in addition to MISO regression, MIMO regression is performed to examine whether or not there is a relationship between age and obesity based on CpG biomarkers. A comparison of MISO and MIMO regression results is presented in Table 5.18. The results suggest that there is no strong correlation between obesity and aging based on the selected CpG dinucleotides (features). Therefore, most of the age-related CpG islands are not related to obesity. Interestingly, only the MSVR result for EUFS are better than its result for SVR. This result appears to suggest that some of the CpG biomarkers which are selected by EUFS are related to both aging and obesity.

This chapter presents experimental studies conducted on the RV144 Vaccine, peptide binding affinity, GSE44763 and GSE40279 data sets. In order to show the effectiveness of the proposed DFSFR, and DKBFS methods, their results are compared with those of state-of-the-art feature selection methods as well as with those of previous studies. The RV144 vaccine data set consists of 20 antibody features and 100 plasma samples that are obtained from the individuals participating in the RV144 vaccine trial week 26. Three different cell-mediated assays are used: Antibody Dependent Cellular Phagocytosis, Antibody Dependent Cellular Cytotoxicity and Natural Killer Cell Cytokine Release activities. The goal of exploiting the RV144 data set is to reveal antibody features that take action against HIV; in other words, to disclose the relationship between antibody features and their effector functions. In the previous study [119], only the MISO regression task was considered; however, MIMO regression was not taken into account. On the other hand, in this study, in addition to performing MISO regression analysis, MIMO regression analysis is applied so that associations among target variables can be revealed. The results of the experiments conducted with the RV144 data set indicate that there are not only correlations among variables, but also there are some correlations among the target variables.

In this study, three different peptide binding affinity data sets are exploited. Tasks 1 and 3 contain nona-peptides that contain a total of 5787 amino acid descriptors and 89 samples. Task 1 consists of 89 training and 88 testing samples, whereas Task 3 includes 133 training and 133 testing instances. Task 2 consists of octa-peptides that have a total of 5144 amino acid descriptors. It has 76 training and 76 testing samples. Each descriptor contains 643 amino acids. The goal of exploiting the peptide binding affinity data set is to predict peptide binding

affinity values by using amino acid descriptors, since these descriptors quantitatively describe the physicochemical properties of the peptides [128]. Affinity refers to the strength of binding or interaction. PPIs play a role in mediating signal transactions, sensing the environment, triggering immunological responses, and monitoring gene expression [126]. Furthermore, PPIs play a crucial role in the progression of human diseases such as viral infections. Therefore, increasing knowledge of the underlying principles of PPIs can ultimately result in disclosing the intrinsic biochemistry of different diseases, and thereby the development of drug design [127]. The proposed DFSFR framework outperforms the state-of-the-art USFSMs for all different tasks. In addition, the proposed DFSFR method dramatically reduces the number of features: for Task 1 from 5787 to 100; for Task 2 from 5144 to 250; and for Task 3 from 5787 to 200.

The GSE44763 data set contains 27482 CpG biomarkers (features) from the peripheral blood of 46 adult female donors (samples). There are 24 obese subjects and 22 lean subjects. The goal of exploiting this data set is to reveal the associations among CpG biomarkers, and the chronological age and BMI of individuals. The proposed DFSFR and DKBFS frameworks outperform the other USFSMs and reduce the number of features by as much as 99.45%. The experimental results suggest that the proposed frameworks can reveal age and obesity-related CpG biomarkers (features) from the given data. In addition to performing MISO regression analysis, MIMO regression analysis is also performed. From the experimental results it can be concluded that no strong correlation exists between obesity and chronological age.

The GSE40279 data set consists of 473034 Cytosine-phosphate-Guanine (CpG) biomarkers (features) from whole blood of 656 donors (samples) aged 19 to 101. The goal of exploiting this data set is to reveal the relationship between CpG dinucleotides and the chronological age of individuals from the given data. A pre-processing step is applied to the GSE40279 data set so that the features with the lowest variation in the sample are eliminated, and thereby the number of features is reduced from 473034 to 90000. Then, USFSMs are applied to identify discriminative CpG biomarkers (features). The number of selected features starts from 900 in order that a subset of features contains at least 1% of the features. A set of 90000 features is assessed by utilising four different USFSMs along with the entire feature set. The proposed DFSFR, DKBFS, and KBFS frameworks produce better results than other USFSMs.

A general overview of the characteristics of all data sets which are exploited in this study is presented in Table 5.21. The GSE40279 data set can be determined as high dimensional as far as classification is concerned; however, in the regression domain, the GSE40279 data set can be considered as ultra-high dimensional.

TABLE 5.21: A General Overview of all of the Data Sets Used in this Study

Datasets	Number of		Sources of the Data Sets	Description
	Features	Samples		
RV144	20	100	[119]	LD
Task 1	5787	177	[129]	HD
Task 2	5144	152	[129]	HD
Task 3	5787	256	[129]	HD
GSE44763	27482	46	[130]	Very HD
GSE40279	473034	656	[145]	Ultra HD

LD:Low dimensional, HD:High dimensional

Chapter 6

Discussion

In this chapter, the experimental results for the RV144 Vaccine, the peptide binding affinity, the GSE44763, the GSE40279 data sets are discussed. Then, the findings from those experiments are presented. In this section, methods are discussed based on their results for different high dimensional data sets since each data set contains different number of features (dimensionality).

6.1 Discussion of the Results for RV144 Data

In Chapter 4 and 5 experimental studies conducted on RV144 Vaccine data set are presented. This data set is used to test the predictive capability of the proposed DFSFR and KBFS models for the given data set and to provide better generalisation and performance compared to a recent study conducted on the RV144 data set [119]. This data set contains 20 antibody features and 100 plasma samples (80 samples are vaccine injected and 20 samples are placebo).

The goal of the study is to disclose associations among antibody features and their effector functions. The effector functions can be described as actions of the immune system to fight against HIV. Therefore, the identification of specific antibody features involved in fighting against HIV is crucial in neutralising the virus.

Experimental results conducted on RV144 Vaccine data set suggest that the proposed frameworks, DFSFR and KBFS, outperform state-of-the-art unsupervised feature selection methods as well as the method used in the previous paper on

the same data set. DFSFR has a better quantitative accuracy performance than the predictive models constructed using Lars, GP and SVR presented in the data set paper for ADCC, and ADCP assays. DFSFR has a little less quantitative performance as compared to predictive model for Cytokine assay constructed using SVR presented in the data set paper. However, it still has better quantitative performance than the Lars and GP predictive models for the Cytokine assay. By utilising DFSFR framework, number of features are reduced to 1 for ADCC assay, 13 for ADCP assay and 16 for Cytokine assay. However, in data set paper, the number of selected features are not indicated; instead, filtered set is mentioned without providing the number of selected features.

Experimental results conclude that the proposed unsupervised framework, DFSFR, achieves the best performance on all assays, which thus verifies that it is able to reveal discriminative antibody features that provide protection against HIV.

Furthermore, in previous study [119], only MISO regression is considered where correlations among output variables are not taken into account. In this study, in addition to the MISO regression, MIMO regression is performed to determine whether associations exist among target variables. By exploiting MIMO regression, the prediction performance of the predictive model is increased approximately by 12 percent. This concludes that there are not only associations among antibody features, but also there are associations among effector functions. Analysing dependencies between antibody features as well as response variables may ultimately result in producing an effective RV144 vaccine so that HIV or AIDS may be conquered.

There is only one common feature selected by unsupervised feature selection methods for ADCC and ADCP assays IgG1.gp41 and IgG3.p24, respectively. On the other hand, seven different antibody features are mutually selected by unsupervised methods for the Cytokine assay: IgG1.p24, IgG3.p24, IgG4.gp41, IgG4.gp140, IgG4.p24, IgG4.gp120, IgG3.V1V2. In this study, distribution of antibody features based on their importance is also provided so that the most important antibody features might be further analysed in real word biology laboratories.

6.2 Discussion of the Results for Peptide Binding Affinity Data Sets

The experimental studies conducted on three different high dimensional peptide binding affinity data sets are presented in chapter 4 and 5. These data sets generally contain over 5000 descriptors for each peptide and they are used to evaluate the prediction performance of the proposed DFSFR framework for the given data sets.

The purpose of the study is to predict peptide binding affinity values by using amino acid descriptors. As mentioned previously, affinity refers to the strength of binding or interaction. Identification of peptide binding affinity values is important due to the fact that protein-protein interactions (PPIs) play a role in mediating signal transactions, sensing the environment, triggering immunological responses, and monitoring gene expression [126].

The outcomes of the experiments clearly emphasise the strengths of DFSFR and KBFS compared with the state-of-the-art unsupervised feature selection methods as well as the approaches used in a previous study [88] which were conducted on the same peptide data sets. DFSFR produces better performance than the state-of-the-art feature selection methods and our earlier study [88] for all three tasks. Six different metrics, namely MAD, MSE, RMSE, MAPE, U, and q^2, are used to examine the robustness of USFSMs. DFSFR achieves the best performance on all different tasks over different metrics. Furthermore, DFSFR dramatically reduces the number of features for all tasks: for Task 1 from 5787 to 100; for Task 2 from 5144 to 250; and for Task 3 from 5787 to 200. Based on the RMSE metric, which is the most popular evaluation metric, the prediction error achieved in this study compared to the previous study [88] is decreased by approximately 15% for Task 1, 5% for Task 2, and 10% for Task 3.

6.3 Discussion of the Results for GSE44763 Data Set

The experimental studies conducted on the GSE44673 data set are presented in chapter 4 and 5. This data set contains 27482 Cytosine-phosphate-Guanine

(CpG) biomarkers from peripheral blood of 46 adult female donors. There are 24 obese subjects and 22 lean subjects. Predictive modelling of such data is one of the most challenging problems in many feature selection applications since the dimensionality of data is extremely high, while the sample size is very small. [184].

Circulatory disease, cancers, and respiratory disease are three main causes of mortality [132] [133]. Obesity can increase the risk of these three fatal diseases as well as other diseases, such as diabetes and depression. According to the World Health Organization (WHO) there were over 600 million obese people worldwide in 2014 [134]. Most of the time, the risks associated with obesity related diseases are also increased by aging [130]. Consequently, aging and obesity contribute to fatal diseases including cancers, circulatory and respiratory disease. The GSE44763 data set is used to test the performances of the proposed KBFS and DKBFS frameworks as well as to determine the CpG dinucleotides related to the age and obesity from the data.

The affirmative results show the effectiveness of the proposed DFSFR, KBFS and DKBFS frameworks. They are compared with the state-of-the-art unsupervised feature selection methods and three different metrics are used to examine the robustness of USFSMs. DKBFS, DFSFR and KBFS produced better prediction performance than the state-of-the-art feature selection methods. The best results are achieved by DKBFS which is a hybrid feature selection framework that combines DFSFR and KBFS. When compared with InFS which produces the best results among existing methods, DKBFS decreases the MAD of the predictive model approximately by 4% and the RMSE of the model is reduced by approximately 3% for the prediction of chronological age, and it reduces the MAD of the model by approximately 8% and the RMSE of the model is also reduced by approximately 7% percent for the prediction of BMI.

Experimental results on the GSE44763 data set conclude that the proposed DFSFR, KBFS and DKBFS methods are capable of handling high dimensional data and can reveal CpG dinucleotides (features) related to age and obesity.

There are two different outputs in the GSE44763 data, which are BMI and chronological age. Therefore, the GSE44763 data set is suitable for performing MIMO tasks. MIMO task is performed by using MSVR. The MVSR results

on different USFSMs over GSE44763 data set suggest that there is not a strong correlation between aging and obesity based on selected CpG biomarkers.

6.4 Discussion of Results for GSE40279 Data Set

The GSE44763 data set is used to disclose age-related CpG dinucleotides (features). The GSE40279 data set contains approximately 16 times more features than the GSE44763 data set; therefore, developing a predictive model using the GSE40279 data set is more difficult than building a predictive model using the GSE44763 data set. The GSE40279 data set contains 473034 CpG dinucleotides (features) from whole blood of 656 donors aged 19 to 101.

The purpose of this study is to reveal the associations between age and CpG biomarkers or to identify age-related CpG biomarkers from the GSE40279 data set. Age prediction of individuals from molecular biomarkers is crucial for forensics, disease prevention and the extension of life. Therefore, the GSE40279 data set is exploited to evaluate the performance of the proposed DFSFR, DKBFS and KBFS frameworks as well as to identify age-related CpG dinucleotides from the given data.

The experimental results suggest that the proposed DFSFR, DKBFS and KBFS methods produce better performance than the state-of-the-art unsupervised feature selection methods. Four different evaluation metrics are used to analyse the effectiveness of USFSMs. The DKBFS method achieves the best results for all different metrics and DFSFR produces the second best performance. One interesting observation is that USFSMs generally produced very similar U results including baseline. When compared with the other USFSMs, DKBFS decreased the U statistics results by approximately 38%, which indicates the outstanding performance of this DKBFS framework.

It is observed that the performance of the feature selection methods is easily effected by the number of selected features. The number of selected features are: 900 for DKBFS, KBFS and TV, 6300 for LapFS, and 9000 for EUFS. Therefore, DKBFS produces the best performance by exploiting the minimum number of CpG dinucleotides.

The experimental results suggest that our proposed methods produce better results than other USFSMs for all different metrics. However, even though number of features are extremely reduced from 473034 to 900, the number of biomarkers are considerably high to be analysed in real biology labs. In this case, an aggressive research on three different feature subset of 900 CpG dinucleotides which are selected by DKBFS, KBFS, and DFSFR. The goal of this aggressive research is to obtain minimum number of dinucleotides which represent the whole data set with the same or higher accuracy so that those dinucleotides can be further analysed in real biology labs. The number of features are started from 1 and incremented by 1 till 900 to make an aggressive research on selected CpG biomarkers. KBFS produces better very good results with 41 features corresponding to only 0.00867% of the entire features.

6.5 General Discussion and Findings

In this section, the experimental results on different data sets are discussed and the findings based on these results are presented.

Extensive experiments have been designed and conducted to objectively assess the proposed DFSFR, KBFS and DKBFS models. In order to evaluate the performance of the DFSFR framework, the RV144 Vaccine, the peptide binding affinity, the GSE44763 and the GSE40279 data sets are used. Then, the performance of the DFSFR method is compared with the state-of-the-art USFSMs as well as methods used in previous studies [119] [88]. The KBFS and DKBFS frameworks are tested by exploiting GSE44763 and GSE40279 data sets because these frameworks are developed for very high dimensional data. To the best of our knowledge, this is the first study that presents exploration and comprehensive comparison of USFSMs in very high dimensional regression problems, particularly in biomedical domain.

Experimental results conducted on different high dimensional data sets appear to suggest that deep learning based methods, DFSFR and DKBFS outperform the state-of-the-art USFSMs. This might be because deep learning based methods benefit from deep structures to model non linearity. Furthermore, DBN consists of multiple layers of RBM might reach more abstract concepts through layer-wise learning in order to discover the data structure. However, deep learning has not

been widely exploited for feature selection especially in bioinformatics field. Even though various feature selection methods have been proposed in the literature, however, no deep learning based feature selection method exists particularly for regression tasks. To the best of our knowledge DFSFR is the first deep learning based feature selection method particularly for regression problems.

In addition to MISO regression, MIMO regression is exploited to examine associations among effector functions of antibodies (ADCP, ADCC, Cytokine), and to identify relationships between aging and obesity from the given data. The experimental results show that there are some correlations among effector functions: ADCP, ADCC, and Cytokine; however, no strong correlation exists between aging and obesity based on selected CpG biomarkers (features).

RV144 Vaccine data set contains three cell-mediated assays which are target variables. Therefore, the data set can be exploited for both MISO and MIMO regression purposes. In previous study [119], MIMO regression was not taken into account. However, there might be not also feature-target or feature-feature associations, but there may be also associations among target variables. In order to observe that MSVR is used. Experimental results conducted on RV144 data set conclude that there are correlation among target variables because the results which were obtained by performing MSVR was slightly better than the results that were produced by single SVR.

Overall, feature selection is effective and necessary. The selected features can not only reduce computational cost, but also improve the prediction performance of a learning model.

It is also observed that selection of features from very high dimensional data sets in regression domain seems to have been understudied. Therefore, it is important to explore existing and new methods to be adapted and devised for such an important domain as new data sets are being generated, which require such quantitative assessments.

6.6 Discussion of SVR and MSVR

In this study, default parameters of USFSMs are used to evaluate the robustness of USFSMs. However, the effects of SVR parameters are investigated. There are

three primary parameters in SVR for regression tasks: C, γ and ϵ. C is the cost parameter which is used to avoid overfitting, ϵ refers to error tolerance where errors less than ϵ will be tolerated. γ sets the value of gamma in the kernel function. The best parameters are not known beforehand; therefore, to find optimal parameter sets, "grid search" method can be used. The goal is to select the best (C, ϵ, γ) parameter set so that the model can accurately predict unknown data. It is observed that using even numbers to determine the C parameter is practical (for example, C=2,4,6,8). For the γ parameter, the default parameter is $\gamma = 1/n$, where n is the number of selected features. However, it is observed that $\gamma = 0.1$ can be used as a starting point and then this can be increased progressively by 0.1 to find the best value of γ. It is also observed that the default value of the ϵ parameter ($\epsilon = 0.1$) is good. To perform a grid search, it can be incremented by 0.1 to identify the best ϵ parameter.

For MSVR there are crucial parameters, namely, C, σ, and ϵ. C is the regularisation parameter which regulates the trade off between minimising the error on the training data and minimising the norm of the weights. Optimisation problems of SVR and MSVR is provided on Chapter 3 which are Eq.(3.3) and Eq.(3.7), respectively. If C is too large; then, objective function will attempt to decrease w as much as possible so that the model function appropriately shows relationship between features and target. On the other hand, if C is too small, then, the model function will increase w that can ultimately be result in extremely large training error. Therefore, optimisation of C parameter is crucial. For high dimensional data sets, such as GSE44763, it is observed that if C is small (such as C=1), predictive model produces good results on MIMO regression. On the other hand, for low dimensional data sets, such as RV144 Vaccine, even numbers of C parameter is practical (for example, C=2,4,6,8).

6.7 Final Remarks

In this section, the strengths and weaknesses of unsupervised feature selection methods, which are exploited in this study, are presented.

It is observed that SPFS and LapFS usually produce similar results although SPFS generally achieves better results than LapFS. This might be because they both attempt to preserve the data similarity of the original features, however,

LapFS cannot handle feature redundancy. It is also observed that MCFS method performed well when the number of features is small (RV144 Vaccine data set); however, its performance declined as the dimensionality of data increases (peptide binding affinity data sets). In addition, MCFS is inefficient for application to very high dimensional data, such as the GSE40279 because this method employs the computation of a normalised Laplacian matrix, l_1 norm regularisation, and eigenvalue decomposition. LapFS computes a Laplacian matrix, and eigen value decomposition; however, it is still able to perform feature selection on the GSE40279 data set.

Another interesting point is that even though EUFS is an embedded method, and thus is computationally more expensive than filter methods, it is able to perform feature selection on ultra high dimensional GSE40279 data set.

Due to the extremely high run time and memory consumption, InFS, MCFS, SPFS could not be applied to the GSE40279 data set; instead, TV is used. Even though TV is a very simple unsupervised feature selection method, it produces good results on GSE40279 data set. Furthermore, because of its simplicity, it is the most computationally effective method compared to EUFS, LapFS, KBFS, and DKBFS.

The proposed KBFS method is a simple K-means based unsupervised method; however, it produces the second best results on ultra high dimensional GSE44763 and GSE40209 data sets. This might be due to the fact that unlike existing K-means based feature selection methods, which are capable of performing univariate feature selection, KBFS performs multivariate feature selection by exploiting feature-feature dissimilarity measure. It is observed that KBFS should be used to select features from very high dimensional data.

The proposed DFSFR method achieves the best results for the RV144 Vaccine, peptide binding affinity, the GSE44763 (for the prediction of BMI) data sets and it yields the second best results for the GSE44763 (for the prediction of chronological age), and the GSE40279 data sets. Therefore, DFSFR method can be utilised for low dimensional, high dimensional, very high dimensional and ultra high dimensional data.

The proposed DKBFS method produces the best results for the GSE44763 (prediction of chronological age), and GSE40279 data sets. Therefore, it is concluded that DKBFS method is useful when it is applied to extremely high dimensional

data. In summary, it is beneficial to exploit MCFS, LapFS and SPEC methods for low dimensional data sets. KBFS and DKBFS can be used for very high dimensional data sets. DFSFR method can be exploited for both low dimensional, and high dimensional data sets. The results of EUFS and InFS over different data sets are generally not consistent; thereby, the performances of these are highly dependent on data set.

Chapter 7

Conclusions and Future Works

This chapter concludes the research, and presents possible future works.

7.1 Conclusions

In line with the technological developments, there is almost no limit to collect data of high dimension in bioinformatics. These high dimensional data sets usually contain many redundant or noisy features which need to be filtered out to find a small but biologically meaningful set of attributes. Feature selection aims at identifying a subset of original features by eliminating redundant and noisy ones and this is an effective dimensionality reduction method that is widely used in machine learning and data mining. In fact, feature selection enables regressors to achieve better performance in terms of regression. There are mainly two different types of feature selection methods: unsupervised and supervised. Supervised feature selection methods can identify relevant features as well as noisy ones; however, unsupervised methods do not tend to identify features that can act as noise.

After conducting an intensive literature review, it is observed that selection of features from very high dimensional data sets in regression domain seems to have been understudied. The reason for this might be due to the fact that regression problems are more difficult than classification tasks [185].

In this study, a taxonomy of feature selection methods for regression problems is provided. To the best of our knowledge this is the first study that provides a

feature selection review as well as a taxonomy of feature selection methods for particularly regression tasks.

Two novel unsupervised feature selection frameworks are provided in this study, namely, KBFS and DFSFR. KBFS is a simple K-means based feature selection framework where features are selected according to a feature-feature dissimilarity measure. In K-means, one centroid point for each cluster is used, however, in KBFS, three centroids are exploited to determine weights of features. Indeed, the centroids of K-means are even not a feature. DFSFR is a deep learning based feature selection framework that selects features at the input level of DBN which is, to the best of our knowledge, the first deep learning based feature selection method in regression domain. This framework is capable of handling both multi-input single-output and multi-input multi-output regression tasks. A hybrid method, which combines DFSFR and KBFS, is also proposed and named as DKBFS. In DKBFS, KBFS is exploited as a pre-filtering method for DFSFR framework. Therefore, KBFS prioritises features according to their importance and identifies relevant features. Previously identified relevant features are then evaluated by DFSFR that attempts to decide an optimal feature subset. KBFS and DKBFS are proposed to deal with extremely high dimensional data.

To show the effectiveness of the proposed frameworks, experiments are conducted on different high dimensional biomedical data sets. Four different case studies are considered. In the first case study, the proposed methods are used to reveal the associations between antibody feature and their functional activities (ADCC, ADCP, NK Cell Cytokine Release) from the RV144 Vaccine data set. The purpose of this case study is to identify the most discriminative antibody features that fight against HIV.

In the second case study, proposed methods are applied to high dimensional peptide binding affinity data sets. Three different peptide binding affinity data sets are used. Each amino acid in the peptide sequences is then described by 643 physico-chemical descriptors. Tasks 1 and 3 contain nona-peptides that have a total of 5787 descriptors (=643x9) whereas Task 2 consists of octa-peptides that were characterised using a total of 5144 descriptors (=643x8). The goal of this study is to predict binding affinity values for peptides using amino acid descriptors. The purpose of this study is to predict affinity values of peptide binding since affinity refers the strength of binding.

In the third case study, very high dimensional GSE44763 data set, which consists of 27842 Cytosine-phosphate-Guanine (CpG) dinucleotides from peripheral blood of 46 adult female individuals, is exploited. There is a total of 46 subjects where the subjects are obese and 22 of them are lean. The aim of this study is to reveal age and obesity related CpG biomarkers from the given data.

In the fourth case study, ultra high dimensional GSE40279 data set which contains 473034 CpG biomarkers (features) from whole blood of 656 donors (samples) aged 19 to 101, is used. The goal of this study is to disclose the associations among CpG dinucleotides and aging from the given data.

The proposed methods obtain better or at least comparable results compared to other the state-of-the-art feature selection methods in the literature and it is shown that the proposed methods are robust and effective in identifying discriminative features from biomedical data.

In this thesis, in addition to providing novel feature selection frameworks, a comprehensive overview of feature selection methods for regression problems is also provided where feature selection methods are shown along with their types, references, sources, and code repositories. Finally, a taxonomy of feature selection methods for regression problems is proposed to assist researchers to select appropriate feature selection method for their research.

7.2 Contributions to the Literature

The main results and contributions of this research are briefly summarised as follows:

- The DFSFR method is proposed and applied to different high dimensional benchmarks: (i) RV144 Vaccine data set is used to disclose functional relationship between immune system and HIV (ii) Peptide binding affinity data sets are exploited to estimate binding affinity values of peptides from given data (iii) GSE44763 data set is used to reveal associations among CpG dinucleotides(features), and BMI and chronological age of individuals from given data (iv) GSE40279 data set is utilised to understand the relationships between chronological age and CpG dinucleotides from given

data. The results suggest that DFSFR yields an improvement in the prediction accuracy. As far as the literature is concerned, to the best of our knowledge, this novel deep learning based feature selection method in the regression domain, the first of its kind, has been shown to be better than other the state of the art methods by not only selecting smaller number of the features but also helping increase the predictive performances for both the single and multi-output regression models (journal article is under review [18]).

- The KBFS method is proposed and applied to very high dimensional GSE44763 and GSE40279 data sets. To the best of our knowledge, KBFS is the first K-means based unsupervised feature selection method that consider feature-feature dissimilarity measure to select features rather than ranking each feature individually. Therefore, KBFS can be determined as a multivariate filter selection method. Experimental results suggest that KBFS produces better predictive results than state of the art unsupervised feature selection methods (Published work [16] and a work is under review [17]).

- The DKBFS method, which combines DFSFR and KBFS, is proposed and applied to GSE44763 and GSE40279 data sets. In DKBFS, KBFS is exploited as a pre-filtering step. The results conclude that DKBFS achieves better prediction accuracy than state of the art unsupervised feature selection methods.

- A comprehensive overview of existing feature selection methods particularly for regression tasks. These methods are provided along with their types, references, sources, and code repositories. To the best of our knowledge, this review is first of its kind since there is no such review provided in the literature; therefore, this review will fill the research gap and assist researchers to select appropriate feature selection method for their research (under review).

- A taxonomy of exiting feature selection methods for regression problems is proposed which categorise feature selection methods according to their types, strategies, and intrinsic learning structure (This work is under review).

7.3 Future Works

This research study suggests new perspectives for a future work. The following suggestions could be explored as future works:

- This study addresses the problem of unsupervised feature selection from extremely high dimensional biomedical data. Data streams are rapidly and constantly growing. Analysis of rapidly changing data streams is quite difficult since the amount of data increases in timely manner [186]. We envision that current development of scientific research will soon lead to the need for development of feature selection methods which can learn from streams of data. Therefore, the research can be further extended by modifying DFSFR so that it would be able to process streams of constantly incoming data.

- In this study, in order to evaluate the robustness of unsupervised feature selection methods, default parameters of them have been utilised. Another direction of research might be to examine how the various parameters of USFSMs affect prediction results. Further research is now being geared towards further refinement of the feature selection and prediction methods by developing and fine-tuning the algorithms.

- In this study, multi-targeted GSE44763 and RV144 Vaccine data sets are exploited. However, more multi-targeted high dimensional regression data sets are required to test the effectiveness of the proposed frameworks for performing MIMO regression. Unfortunately, in some areas, such as bioinformatics the vast majority of data sets are single targeted; furthermore, a large number of data sets are not publicly available.

- In this research, SVR is utilised as a consecutive part of USFSMs; however, different types of regression techniques have been proposed in the literature, such as Gaussian Process Regression and Least Angle Regression, which can also be considered for exploitation to design the consecutive part of USFSMs.

- Another direction of future research might be revealing the biological relevance of selected antibody features, CpG dinucleotides, and amino acid descriptors, therefore, selected features can further be analysed in real biology labs.

Appendix A

CoEPrA Peptide Binding Affinity Data Sets

CoEPrA contains publicly available peptide binding affinity data sets. These data sets are used in the experimental studies of this thesis. The peptide binding affinity data sets are obtained from a modeling competition [129]. Each task has a separate training (Tables A.1-A.3) and test data set (Tables A.4-A.6). The columns correspond to peptide no, peptide residue, and expected real value of binding affinity.

Table A.1: List of peptides for CoEPrA Task 1 (Training).

No.	Peptide	Expected	No.	Peptide	Expected
1	ILDPFPVTD	2.94	46	IYDPFPVTV	5.41
2	ILDPFPVTY	3.19	47	YLSPGPVTA	5.44
3	ILDPFPVTH	3.6	48	LLFGYPVYV	5.45
4	SLHVGTQCA	3.79	49	YLFDGPVTA	5.5
5	HLLVGSSGL	3.91	50	ILDPFPVTT	5.54
6	NLQSLTNLL	3.96	51	RLWPLYPNV	5.57
7	SLNFMGYVI	4	52	YLFPGPVWA	5.59
8	ITSQVPFSV	4.06	53	YAIDLPVSV	5.63
9	VCMTVDSLV	4.2	54	YLFNGPVTV	5.65
10	LLMGTLGIV	4.21	55	ILDPFPVTF	5.67
11	ALIHHNTHL	4.3	56	YLWPGPVTV	5.7
12	MLDLQPETT	4.36	57	RLWPFYHNV	5.72
13	YVITTQHWL	4.39	58	YLAPGPVTA	5.74
14	ITFQVPFSV	4.42	59	IADPFPVTV	5.76
15	KTWGQYWQV	4.43	60	YLYPGPVTA	5.77
16	ITDQVPFSV	4.48	61	YLFPGPETA	5.81
17	LLAQFTSAI	4.51	62	ILDPFPVTP	5.82
18	VLHSFTDAI	4.54	63	FLWPFYPNV	5.89
19	ILDPFPVTK	4.59	64	FLDQVPFSV	5.98
20	YMNGTMSQV	4.67	65	FLWPFYHNV	5.99
21	ILDPFPVTW	4.71	66	ILWPLFHEV	6.03
22	FTDQVPFSV	4.76	67	ILWPLYPNV	6.06
23	KLHLYSHPI	4.77	68	ILDQVPFSV	6.09
24	ILDPFPVTS	4.78	69	ILNPFYPDV	6.11
25	YTDQVPFSV	4.8	70	FLWPLYPNV	6.14
26	IFDPFPVTV	4.89	71	FLNPFYPNV	6.16
27	CLTSTVQLV	4.93	72	FLNPIYHDV	6.16
28	YLWQYIFSV	4.94	73	YLFPGTVTA	6.16
29	IHDPFPVTV	4.96	74	YLCPGPVTA	6.18
30	RLMKQDFSV	4.97	75	YLFPPPVTV	6.19
31	VMGTLVALV	5.03	76	ILFPGPVTA	6.23
32	ILYQVPFSV	5.06	77	IIDPFPVTV	6.31
33	IPDPFPVTV	5.1	78	ILDPFPVTA	6.32
34	GLLGWSPQA	5.13	79	FLWPIYHNV	6.37
35	GLYSSTVPV	5.15	80	ILFPFVHSV	6.58
36	IISCTCPTV	5.17	81	ILDPFPVTG	6.66
37	FLCKQYLNL	5.21	82	YLFPFPITV	6.68
38	YLFPGPVTG	5.22	83	ILFPFPVEV	6.8
39	GTLGIVCPI	5.23	84	ILDDFPPTV	7.08
40	RLWPFYPNV	5.24	85	ILDPLPPTV	7.15
41	YLKPGPVTA	5.26	86	IMDPFPVTV	7.21
42	YLMPGPVTA	5.27	87	ILDPFPPPV	7.44
43	YMLDLQPET	5.28	88	ILDPFPITV	8.14
44	PLLPIFFCL	5.32	89	ILDPFPVTV	8.65
45	RLNPLYPNV	5.37			

Table A.2: List of peptides for CoEPrA Task 2 (Training).

No.	Peptide	Expected	No.	Peptide	Expected
1	FESTGNLD	5.01	39	FESTNNLI	7.748
2	FKSTGNLI	5.026	40	FDSTGNLI	7.814
3	FESTGNLR	5.232	41	FESTSNLI	7.821
4	FFSTGNLI	5.421	42	FESTWNLI	7.832
5	FESTGNLQ	5.687	43	FGSTGNLI	7.846
6	FESTGNLH	6	44	FESTGWLI	7.872
7	FESTGNLG	6.051	45	FESTINLI	7.887
8	FISTGNLI	6.329	46	FESDGNLI	7.89
9	QTFVVGCI	6.796	47	FESTLNLI	7.898
10	NEKSFKDI	6.91	48	FESTVNLI	7.912
11	FQSTGNLI	7.013	49	LEILNGEI	7.921
12	FLSTGNLI	7.088	50	FESTGKLI	7.927
13	FESTGNKI	7.159	51	DGLGGKLV	7.959
14	FESTGNLM	7.212	52	FESEGNLI	7.972
15	FESTGNDI	7.29	53	FESKGNLI	7.978
16	FESTGNLW	7.293	54	FEHTGNLN	7.982
17	KESTGNLI	7.308	55	FESWGNLI	7.989
18	FESTGNPI	7.41	56	FESTANLI	7.994
19	PESTGNLI	7.426	57	FEFTGNLN	8
20	FESTGNLA	7.455	58	FESTGVLI	8.023
21	FESTGNNI	7.521	59	FESAGNLI	8.031
22	FESTGNLS	7.525	60	FESPGNLI	8.042
23	FESTGNEI	7.541	61	FESTGNFI	8.044
24	VESTGNLI	7.545	62	FESTGNLI	8.046
25	FESTGNII	7.551	63	FESFGNLI	8.085
26	FESTGELI	7.593	64	FESRGNLI	8.095
27	HESTGNLI	7.607	65	FESYGNLI	8.099
28	FESTGNQI	7.612	66	FESTPNLI	8.141
29	AESTGNLI	7.624	67	FEATGNLN	8.178
30	SESTGNLI	7.641	68	FEDTGNLN	8.199
31	GESTGNLI	7.665	69	FEQTGNLN	8.217
32	FESTGDLI	7.683	70	FESTGRLI	8.222
33	IESTGNLI	7.715	71	FENTGNLN	8.224
34	MESTGNLI	7.716	72	FESVGNLI	8.23
35	QESTGNLI	7.727	73	FESIGNLI	8.239
36	NESTGNLI	7.736	74	FEGTGNLN	8.265
37	WESTGNLI	7.74	75	FERTGNLN	8.3
38	FESTGNHI	7.742	76	FELTGNLN	8.343

Table A.3: List of peptides for CoEPrA Task 3 (Training)

No.	Peptide	Expected	No.	Peptide	Expected
1	VVHFFKNIV	4.301	68	VLLDYQGML	7.095
2	VCMTVDSLV	5.146	69	LMIGTAAAV	7.102
3	LLGCAANWI	5.301	70	TVLRFVPPL	7.114
4	SAANDPIFV	5.342	71	NLGNLNVSI	7.119
5	TTAEEAAGI	5.38	72	ILHNGAYSL	7.127
6	LTVILGVLL	5.58	73	SIISAVVGI	7.159
7	LVSLLTFMI	5.716	74	VLAKDGTEV	7.174
8	QMTFHLFIA	5.778	75	YLEPGPVTI	7.187
9	ALPYWNFAT	5.82	76	FLYNRPLSV	7.212
10	FVTWHRYHL	5.869	77	FLWGPRALV	7.215
11	SLNFMGYVI	5.881	78	ILDQVPFSV	7.284
12	GIGILTVIL	6	79	ILSSLGLPV	7.301
13	IVMGNGTLV	6.001	80	LLFLGVVFL	7.301
14	SLSRFSWGA	6.041	81	YLVAYQATV	7.304
15	TVILGVLLL	6.072	82	YLEPGPVTV	7.342
16	WTDQVPFSV	6.145	83	ILSPFMPLL	7.347
17	AIAKAAAAV	6.176	84	YLSPGPVTA	7.383
18	ITSQVPFSV	6.196	85	IIDQVPFSV	7.398
19	ALAKAAAAI	6.211	86	YMNGTMSQV	7.398
20	GLGQVPLIV	6.301	87	FLCWGPFFL	7.415
21	LLSSNLSWL	6.342	88	LLFRFMRPL	7.447
22	SIIDPLIYA	6.342	89	ITWQVPFSV	7.457
23	YLVTRHADV	6.342	90	LLAVLYCLL	7.478
24	LIGNESFAL	6.38	91	GIRPYEILA	7.481
25	FLLPDAQSI	6.415	92	GLFLTTEAV	7.509
26	CLALSDLLV	6.447	93	YTYKWETFL	7.538
27	LLGRNSFEV	6.447	94	ALVGLFVLL	7.553
28	LLAVGATKV	6.477	95	SLDDYNHLV	7.583
29	MLLAVLYCL	6.478	96	FLLRWEQEI	7.592
30	AIYHPQQFV	6.504	97	SLLPAIVEL	7.62
31	ALAKAAAAL	6.511	98	YLSPGPVTV	7.642
32	FVNHRFTVV	6.523	99	GLIMVLSFL	7.658
33	WILRGTSFV	6.556	100	SLYADSPSV	7.658
34	TLDSQVMSL	6.58	101	RLLQETELV	7.682
35	GLYGAQYDV	6.602	102	IMDQVPFSV	7.719

36	MLASTLTDA	6.602	103	YLLPAIVHI	7.745
37	AIIDPLIYA	6.623	104	FLLLADARV	7.747
38	FLGGTPVCL	6.623	105	ALMDKSLHV	7.767
39	LMLPGMNGI	6.623	106	YLYPGPVTA	7.772
40	RLMIGTAAA	6.644	107	HMWNFISGI	7.818
41	LLFLLLADA	6.663	108	YLAPGPVTV	7.818
42	GTLGIVCPI	6.666	109	MLGTHTMEV	7.845
43	KLFPEVIDL	6.693	110	MTYAAPLFV	7.86
44	IAGGVMAVV	6.708	111	YLSQIAVLL	7.917
45	GLYRQWALA	6.733	112	YLMPGPVTV	7.932
46	MLQDMAILT	6.777	113	WLDQVPFSV	7.939
47	VILGVLLLI	6.785	114	SLYFGGICV	7.975
48	CLTSTVQLV	6.832	115	YLLALRYLA	8
49	ILLLCLIFL	6.845	116	SLLTFMIAA	8.027
50	DMWEHAFYL	6.879	117	GLMTAVYLV	8.051
51	ALTVVWLLV	6.893	118	FLLSLGIHL	8.053
52	LLPSLFLLL	6.903	119	FVVALIPLV	8.119
53	WMNRLIAFA	6.914	120	YLWPGPVTV	8.125
54	PLLPIFFCL	6.926	121	FLYGALRLA	8.149
55	ALAKAAAAA	6.947	122	LLLEAGALV	8.174
56	FLPWHRLFL	6.95	123	YLFPGPVTV	8.237
57	SLAGFVRML	6.954	124	ILFTFLHLA	8.268
58	TLGIVCPIC	6.964	125	RLPLVLPAV	8.292
59	KLTPLCVTL	6.991	126	YMDDVVLGV	8.301
60	LLCLIFLLV	6.996	127	GILTVILGV	8.342
61	RIWSWLLGA	7	128	NMVPFFPPV	8.403
62	SLLEIGEGV	7.009	129	FLYGAALLA	8.469
63	RLLDDTPEV	7.017	130	YLWPGPVTA	8.495
64	LLAGLVSLL	7.021	131	FLYGALALA	8.62
65	IAATYNFAV	7.032	132	FLDQVPFSV	8.658
66	YTDQVPFSV	7.066	133	ILWQVPFSV	8.77

Table A.4: List of peptides for CoEPrATask 1 (Testing)

1	YLFNGPVTA	5.8	45	IWDPFPVTV	5.13
2	IMDQVPFSV	5.71	46	YLFPGPSTA	5.69
3	RLLQETELV	4.83	47	KIFGSLAFL	4.4
4	HLESLFTAV	3.79	48	YLFPDPVTA	6.09
5	ILDPFPPTV	8.17	49	TLHEYMLDL	4.94
6	ILDPFPVTL	7.03	50	GILTVILGV	4.57
7	FLLSLGIHL	5.17	51	YLFPPPVTA	5.75
8	LQTTIHDII	3.9	52	RLWPIYHDV	5.55
9	IQDPFPVTV	6.05	53	SLDDYNHLV	5.27
10	VLLDYQGML	4.52	54	LLWFHISCL	4.13
11	FLWPIYHDV	6.16	55	VLIQRNPQL	5.06
12	TLGIVCPIC	4.68	56	YLFPGPMTA	5.98
13	YLFPGPVQA	6.14	57	HLYSHPIIL	5.41
14	FVTWHRYHL	4.21	58	WILRGTSFV	4.06
15	FLFPLPPEV	6.53	59	ILDPIPPTV	7.3
16	YLFPGPVTA	6.31	60	VTWHRYHLL	4.38
17	NLSWLSLDV	4.75	61	YLFPCPVTA	6.63
18	YLAPGPVTV	6	62	FLLTRILTI	4.95
19	ALPYWNFAT	4.66	63	IGDPFPVTV	3.92
20	ILDPFPVTE	3.13	64	MLGTHTMEV	5.37
21	ILDPFPVTQ	5.28	65	YLFPGVVTA	6.17
22	IDDPFPVTV	4.36	66	ILDPFPVTI	6.69
23	GLGQVPLIV	4.76	67	ILWPIYHNV	6.24
24	ALMPLYACI	5.08	68	YLEPGPVTL	5.41
25	GLSRYVARL	4.78	69	YLFPGPFTA	5.65
26	ILDDLPPTV	7.14	70	KLPQLCTEL	4.5
27	ILNPFYHNV	6.16	71	ILDPFPVTN	5.29
28	YLFDGPVTV	4.96	72	YLWDHFIEV	6.36
29	YLFQGPVTA	5.21	73	YLWQYIPSV	5.17
30	SLYADSPSV	5.24	74	ILKEPVHGV	5.59
31	YLNPGPVTA	5.53	75	ILKPLYHNV	5.25
32	RLWPIYHNV	5.77	76	ITAQVPFSV	4.43
33	RLNPFYHDV	4.24	77	YLFPGPFTV	5.81
34	FLKPFYHNV	5.73	78	YLFPGPMTV	5.85
35	ILDPFPVTM	6.13	79	TTAEEAAGI	3.39
36	IVDPFPVTV	6.21	80	FLFPGPVTA	6.18
37	LMAVVLASL	3.99	81	WLDQVPFSV	5.23
38	ITDPFPVTV	6.08	82	FLDDHFCTV	6.68
39	ILWQVPFSV	5.91	83	SVYDFFVWL	5.12
40	ITWQVPFSV	5.01	84	ILDPFPVTC	5.65
41	ICDPFPVTV	5.45	85	ILDPFPPEV	7.68
42	ALCRWGLLL	4.91	86	NMVPFFPPV	5.6
43	ILDDFPVTV	7.16	87	ISDPFPVTV	5.5
44	SIISAVVGI	4.47	88	INDPFPVTV	4.78

Table A.5: List of peptides for CoEPrA Task 2 (Testing)

1	YESTGNLI	7.74	39	FESTGHLI	7.997
2	FESTRNLI	7.679	40	FYSTGNLI	5.592
3	FESTGFLI	8.267	41	FPSTGNLI	8.113
4	FESTGTLI	7.922	42	DESTGNLI	7.712
5	FESTQNLI	7.819	43	FESQGNLI	8.094
6	FEKTGNLN	7.904	44	FESTKNLI	7.304
7	FEWTGNLN	8.225	45	FESTGNLL	7.737
8	FESTGQLI	7.92	46	FEVTGNLN	8.223
9	FASTGNLI	7.429	47	FLHPSMPV	7.149
10	FMSTGNLI	6.863	48	FESTMNLI	7.888
11	FESLGNLI	8.403	49	FEITGNLN	8.197
12	FNSTGNLI	6.244	50	FWSTGNLI	5.325
13	FESTGNSI	7.612	51	FEPTGNLN	8.043
14	RESTGNLI	7.544	52	FESTGNLN	7
15	FESTGPLI	8.302	53	FHSTGNLI	5.122
16	FESTDNLI	7.743	54	FEETGNLN	8.028
17	FESTGGLI	7.946	55	TESTGNLI	7.535
18	FTSTGNLI	7.547	56	FESTGNLK	5.01
19	FESTGNLT	7.293	57	FESTGSLI	7.992
20	FESTGNWI	7.974	58	FAFWAFVV	7.523
21	FESTGNLF	7.848	59	FESTGNRI	8.004
22	EESTGNLI	7.732	60	FESTGALI	7.964
23	FESTYNLI	7.46	61	LESTGNLI	7.716
24	FESTGNLP	5.919	62	FEYTGNLN	8.176
25	FESTGNGI	7.209	63	FEMTGNLN	8.222
26	FESTGILI	8.098	64	FESTGYLI	8.215
27	FESTGNVI	7.421	65	HAIHGLLV	7.319
28	FESTGMLI	7.979	66	FESTTNLI	7.821
29	FETTGNLN	8.232	67	FESTENLI	7.583
30	FESSGNLI	8.046	68	FAFPGELL	7.022
31	FESTGNLY	6.01	69	FESTGNLV	7.626
32	FESTHNLI	7.836	70	FESTGNYI	7.793
33	FESTGNTI	7.652	71	FESMGNLI	8.04
34	FESTGNAI	7.602	72	FESTGNMI	7.612
35	FVSTGNLI	7.216	73	FESHGNLI	8.248
36	FESTFNLI	7.895	74	FESTGLLI	8.079
37	FESNGNLI	7.88	75	FESGGNLI	7.985
38	AESKSVII	6.648	76	FSSTGNLI	7.718

Table A.6: List of peptides for CoEPrA Task 3 (Testing)

No.	Peptide	Expected	No.	Peptide	Expected
1	GLYSSTVPV	7.577	68	AMVGAVLTA	7.122
2	FTDQVPFSV	7.212	69	ITAQVPFSV	7.02
3	VLIQRNPQL	7.644	70	ILLSIARVV	6.342
4	LLWFHISCL	6.682	71	FLYGALLAA	8.201
5	FMGAGSKAV	6.2	72	ALMPLYACI	8
6	FVWLHYYSV	7.821	73	GLYYLTTEV	7.682
7	ALAKAAAAM	7.398	74	GLLGWSPQA	8.027
8	LLLCLIFLL	7.585	75	LLWQDPVPA	7.343
9	YAIDLPVSV	7.801	76	MLGNAPSVV	6.644
10	GLSRYVARL	7.174	77	SLADTNSLA	6.342
11	QVMSLHNLV	6.025	78	HLYSHPIIL	7.131
12	MMWYWGPSL	7.921	79	ALVLLMLPV	7.506
13	YLFPGPVTA	8.495	80	RMPAVTDLV	6.903
14	VLLPSLFLL	7.444	81	LLWSFQTSA	7.818
15	KIFGSLAFL	7.478	82	YLEPGPVTL	7.058
16	AVIGALLAV	7.747	83	ALAKAAAAV	6.597
17	ALLAGLVSL	7.117	84	YMLDLQPET	7.373
18	ALSTGLIHL	6.505	85	HLAVIGALL	6.986
19	YALTVVWLL	6.924	86	AMKADIQHV	6.777
20	YLDQVPFSV	8.638	87	RMFAANLGV	7.447
21	YVITTQHWL	6.877	88	IVGAETFYV	8.456
22	FLLTRILTI	8.073	89	LQTTIHDII	5.501
23	YMIMVKCWM	6.663	90	KLAGGVAVI	6.447
24	RLMKQDFSV	7.338	91	LLPLGYPFV	6.477
25	FLAGALLLA	6.223	92	ITFQVPFSV	7.179
26	FLEPGPVTA	6.898	93	GLYLSQIAV	7.017
27	LLAQFTSAI	7.301	94	LLVFACSAV	6.342
28	AVAKAAAAV	6.495	95	AMLQDMAIL	7.009
29	GLCFFGVAL	5.38	96	ILAGYGAGV	6.937
30	VIHAFQYVI	5.914	97	YLAPGPVTA	8.032
31	ILYQVPFSV	8.31	98	SLHVGTQCA	5.842
32	DLMGYIPLV	7.097	99	ILAQVPFSV	7.939
33	NLQSLTNLL	6	100	YLVSFGVWI	8.721
34	SVYVDAKLV	6.991	101	ALYGALLLA	8.143
35	RLLGSLNST	6.778	102	GLQDCTMLV	7.638
36	WLLIDTSNA	6.447	103	VLTALLAGL	7.086
37	KTWGQYWQV	7.957	104	FLYGALVLA	7.409
38	FLYGGLLLA	8.959	105	VLHSFTDAI	6.17

39	ITDQVPFSV	6.947	106	ILTVILGVL	6.419
40	FAFRDLCIV	6.963	107	ITMQVPFSV	7.398
41	YLYPGPVTV	8.051	108	LLFGYPVYV	7.886
42	WLSLLVPFV	8.164	109	HLESLFTAV	5.301
43	TLLVVMGTL	5.58	110	RLTEELNTI	6.06
44	LLDVPTAAV	7.77	111	VMGTLVALV	7.547
45	YLYVHSPAL	8.268	112	SVYDFFVWL	7.289
46	AMFQDPQER	5.74	113	YLMPGPVTA	8.367
47	VVLGVVFGI	7.845	114	ITYQVPFSV	7.48
48	MALLRLPLV	7.279	115	ILSQVPFSV	7.699
49	HLYQGCQVV	6.832	116	RLVSGLVGA	6.818
50	IISCTCPTV	6.58	117	LLLLGLWGL	7.658
51	DPKVKQWPL	6.176	118	NLYVSLLLL	7.114
52	QLFEDNYAL	7.764	119	RMYGVLPWI	7.538
53	LMAVVLASL	6.954	120	FVNHDFTVV	6.523
54	LLSCLGCKI	5.342	121	ALIHHNTHL	6.623
55	VVMGTLVAL	7.069	122	ALCRWGLLL	7
56	VALVGLFVL	5.079	123	GLVDFVKHI	6.663
57	LLACAVIHA	6.602	124	ILDEAYVMA	6.623
58	VLAGLLGNV	7.721	125	GLLGNVSTV	7.62
59	YLSEGDMAA	6.532	126	HLLVGSSGL	5.792
60	KILSVFFLA	8.301	127	ILMQVPFSV	8.125
61	IMPGQEAGL	7.188	128	VLVGGVLAA	6.732
62	FLYGALLLA	8.585	129	AAAKAAAAV	6.398
63	ALLSDWLPA	7.025	130	VLLLDVTPL	7.301
64	GLACHQLCA	6.38	131	YLDLALMSV	8.26
65	YMDDVVLGA	6.699	132	WLEPGPVTA	6.082
66	QLFHLCLII	6.886	133	LLVVMGTLV	5.869
67	FVDYNFTIV	6.62			

Appendix B

Learning in Restricted Bolzman Machines

This appendix explains the leaning mechanism of Restricted Bolzman Machines (RBM).

Let X the input data and $P(X|\Theta)$ is the model to be learned, and Θ is a set of parameter which need to be estimated. Assume $S = x_1, x_2, \ldots, x_n$ is the data vector. Then the maximum likelihood can be calculated from the following formula:

$$logL(\Theta|S) = log\prod_{1}^{l} P(x_i|\Theta) = \prod_{1}^{l} logP(x_i|\Theta) \tag{B.1}$$

This is equivalent to minimising of the distance between Q underlying S and P, which are unknown distribution and the true distribution respectively, in relation to the Kullback-Leibler divergence [187]. Therefore,

$$KL(Q||P) = \sum_{x\in\Omega} Q(x)log\frac{Q(x)}{P(x)} = \sum_{x\in\Omega} Q(x)logQ(x) - \sum_{x\in\Omega} Q(x)logP(x) \tag{B.2}$$

and the update rule is:

$$Q^{t+1} = Q^(t) + \eta\frac{\partial}{\partial Q^t}(logL(Q^t|S)) - \lambda Q^t + \mu\delta Q^{t-1} = Q^t + \delta Q^t \tag{B.3}$$

where μ, Ω, and λ, are the learning, weight decay regularisation, momentum parameters, respectively. By exploting the Equation(4.6), the gradiant of log likelihood can be calculated from:

$$\begin{aligned}\frac{\partial logL(\Theta|v)}{\partial\Theta} &= \frac{\partial}{\partial\Theta}(log\sum_{h}e^{-E(v,h)}) - \frac{\partial}{\partial\Theta}(log\sum_{v,h}e^{-E(v,h)}) \\ &= -\sum_{h}P(h|v)\frac{\partial E(v,h)}{\partial\Theta} + \sum_{v,h}P(h|v)\frac{\partial E(v,h)}{\partial\Theta} \\ &= \left\langle\frac{\partial E(v,h)}{\partial\Theta}\right\rangle_{d} + \left\langle\frac{\partial E(v,h)}{\partial\Theta}\right\rangle_{m}\end{aligned} \tag{B.4}$$

where $\left\langle\frac{\partial E(v,h)}{\partial\Theta}\right\rangle_{d}$ and $\left\langle\frac{\partial E(v,h)}{\partial\Theta}\right\rangle_{m}$ expectations for the data and model distribution, respectively.

Bibliography

[1] Suhang Wang, Jiliang Tang, and Huan Liu. *Feature Selection*, pages 1–9. Springer US, Boston, MA, 2016. ISBN 978-1-4899-7502-7. doi: 10.1007/978-1-4899-7502-7_101-1. URL http://dx.doi.org/10.1007/978-1-4899-7502-7_101-1.

[2] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *arXiv preprint arXiv:1601.07996*, 2016.

[3] Eamonn Keogh and Abdullah Mueen. Curse of dimensionality. In *Encyclopedia of Machine Learning*, pages 257–258. Springer, 2011.

[4] Giorgio Roffo, Simone Melzi, and Marco Cristani. Infinite feature selection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4202–4210, 2015.

[5] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[6] John R Mascola and David C Montefiori. The role of antibodies in hiv vaccines. *Annual review of immunology*, 28:413–444, 2009.

[7] Webber WP Liao and Jonathan W Arthur. Predicting peptide binding to major histocompatibility complex molecules. *Autoimmunity reviews*, 10 (8):469–473, 2011.

[8] Muhammad Summair Raza and Usman Qamar. Understanding and using rough set based feature selection: Concepts, techniques and applications, 2017.

[9] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[10] Xiaofei He, Deng Cai, Shuicheng Yan, and Hong-Jiang Zhang. Neighborhood preserving embedding. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1208–1213. IEEE, 2005.

[11] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

[12] Zhihong Zhang. *Feature selection from higher order correlations*. PhD thesis, University of York, 2012.

[13] Mariam Kalakech, Philippe Biela, Ludovic Macaire, and Denis Hamad. Constraint scores for semi-supervised feature selection: A comparative study. *Pattern Recognition Letters*, 32(5):656–665, 2011.

[14] Ferdi Sarac, Volkan Uslan, Huseyin Seker, and Ahmed Bouridane. Exploration of unsupervised feature selection methods in relation to the prediction of cytokine release effect correlated to antibody features in rv144 vaccines. In *Bioinformatics and Bioengineering (BIBE), 2015 IEEE 15th International Conference on*, pages 1–4. IEEE, 2015.

[15] Pedro Latorre Carmona, José Martínez Sotoca, and Filiberto Pla. Filter-type variable selection based on information measures for regression tasks. *Entropy*, 14(2):323–343, 2012.

[16] Ferdi Sarac, Huseyin Seker, and Ahmed Bouridane. Exploration of unsupervised feature selection methods to predict chronological age of individuals by utilising cpg dinucleotics from whole blood. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2017.

[17] Ferdi Sarac, Huseyin Seker, and Ahmed Bouridane. Multi input multi output support vector based model for the quantitative prediction of age and body mass index by using epigenetic information from peripheral blood. In *International Conference on Cloud and Big Data Computing (ICCBDC 2017)*, 2017.

[18] Ferdi Sarac and Huseyin Seker. A deep learning-based unsupervised feature selection in multi-output regression domain. In *Nature*, 2017.

[19] R Bellman. Dynamic programming princeton university press princeton. *New Jersey Google Scholar*, 1957.

[20] Intisar Hussien, Sara Omer, Nour E Oweis, and Václav Snášel. Feature selection using semi discrete decomposition and singular value decompositions. In *Proceedings of the First International Scientific Conference Intelligent Information Technologies for Industry(IITI16)*, pages 87–97. Springer, 2016.

[21] Hui Yan. Sparsity preserving score for feature selection. In *Applied Informatics*, volume 2, pages 1–8. Springer, 2015.

[22] Aboul-Ella Hassanien, Ahmad Taher Azar, Vaclav Snasel, Janusz Kacprzyk, and Jemal H Abawajy. *Big data in complex systems: challenges and opportunities*, volume 9. Springer, 2015.

[23] Huan Liu, Rudy Setiono, et al. A probabilistic approach to feature selection-a filter solution. In *ICML*, volume 96, pages 319–327. Citeseer, 1996.

[24] Lei Yu and Huan Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *ICML*, volume 3, pages 856–863, 2003.

[25] MA Hall. Correlation based feature selection for discrete and numeric class machine learning. In *Proceedings of the Seventeenth International Conference on Machine Learning, Machine Learning*, pages 359–366, 2000.

[26] Rich Caruana and Dayne Freitag. Greedy attribute selection. In *ICML*, pages 28–36. Citeseer, 1994.

[27] Ron Kohavi and George H John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1):273–324, 1997.

[28] Alain Rakotomamonjy. Variable selection using svm-based criteria. *Journal of machine learning research*, 3(Mar):1357–1370, 2003.

[29] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[30] Chenping Hou, Feiping Nie, Dongyun Yi, and Yi Wu. Feature selection via joint embedding learning and sparse regression. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 1324, 2011.

[31] Pat Langley et al. Selection of relevant features in machine learning. In *Proceedings of the AAAI Fall symposium on relevance*, volume 184, pages 245–271, 1994.

[32] Artur J Ferreira and MáRio AT Figueiredo. Efficient feature selection filters for high-dimensional data. *Pattern Recognition Letters*, 33(13):1794–1804, 2012.

[33] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517, 2007.

[34] Alan Jović, Karla Brkić, and Nikola Bogunović. A review of feature selection methods with applications. In *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015 38th International Convention on*, pages 1200–1205. IEEE, 2015.

[35] Huan Liu and Hiroshi Motoda. *Computational methods of feature selection*. CRC Press, 2007.

[36] Charu C Aggarwal and Chandan K Reddy. *Data clustering: algorithms and applications*. CRC Press, 2013.

[37] Gauthier Doquire and Michel Verleysen. A graph laplacian based approach to semi-supervised feature selection for regression problems. *Neurocomputing*, 121:5–13, 2013.

[38] HC Peng, Chris Ding, and FH Long. Minimum redundancy-maximum relevance feature selection. pages 70–71, 2005.

[39] Hanyang Peng and Yong Fan. Direct l_ (2, p)-norm learning for feature selection. *arXiv preprint arXiv:1504.00430*, 2015.

[40] Chris Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(02):185–205, 2005.

[41] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28 (3/4):321–377, 1936.

[42] Taiji Suzuki, Masashi Sugiyama, Jun Sese, and Takafumi Kanamori. Approximating mutual information by maximum likelihood density ratio estimation. *FSDM*, 4:5–20, 2008.

[43] T. Suzuki, M. Sugiyama, and T. Tanaka. Mutual information approximation via maximum likelihood estimation of density ratio. In *2009 IEEE International Symposium on Information Theory*, pages 463–467, June 2009. doi: 10.1109/ISIT.2009.5205712.

[44] Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10(Jul):1391–1445, 2009.

[45] Taiji Suzuki, Masashi Sugiyama, Takafumi Kanamori, and Jun Sese. Mutual information estimation reveals global associations between stimuli and biological processes. *BMC bioinformatics*, 10(Suppl 1):S52, 2009.

[46] Ferdi Sarac, Volkan Uslan, Huseyin Seker, and Ahmed Bouridane. A supervised feature selection framework in relation to prediction of antibody feature-function activity relationships in RV144 vaccines. In *Systems Man and Cybernetics (SMC 2016) Conference of the IEEE*, 2016.

[47] Jinsong Leng, Craig Valli, and Leisa Armstrong. A wrapper-based feature selection for analysis of large data sets. 2010.

[48] Mineichi Kudo and Jack Sklansky. Comparison of algorithms that select features for pattern classifiers. *Pattern recognition*, 33(1):25–41, 2000.

[49] Iñaki Inza, Pedro Larrañaga, Rosa Blanco, and Antonio J Cerrolaza. Filter versus wrapper gene selection approaches in dna microarray domains. *Artificial intelligence in medicine*, 31(2):91–103, 2004.

[50] M Karagiannopoulos, D Anyfantis, SB Kotsiantis, and PE Pintelas. Feature selection for regression problems. *Proceedings of the 8th Hellenic European Research on Computer Mathematics & its Applications, Athens, Greece*, 2022, 2007.

[51] Jouni Pohjalainen, Okko Räsänen, and Serdar Kadioglu. Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits. *Computer Speech & Language*, 29(1):145–171, 2015.

[52] A Marcano-Cedeno, J Quintanilla-Domínguez, MG Cortina-Januchs, and D Andina. Feature selection using sequential forward selection and classification applying artificial metaplasticity neural network. In *IECON 2010-36th Annual Conference on IEEE Industrial Electronics Society*, pages 2845–2850. IEEE, 2010.

[53] Huan Liu and Hiroshi Motoda. *Feature Selection for Knowledge Discovery and Data Mining*. Springer Science & Business Media, 1998.

[54] P Clark and R Boswell. *Practical Machine Learning Tools and Techniques with Java Implementation*. Morgan Kaufmann Publisher, 2000.

[55] A Dash, DS Dhakre, and D Bhattacharya. Fitting of appropriate statistical model for study of growth and instability in cereal production of odisha. *Journal of Pharmacognosy and Phytochemistry*, 6(5):2495–2499, 2017.

[56] Habil Zare. Fealect: Feature selection by computing statistical scores, 2014.

[57] Minseok Seo and Sejong Oh. Cbfs: High performance feature selection algorithm based on feature clearness. *PloS one*, 7(7):e40419, 2012.

[58] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[59] Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(Nov):2541–2563, 2006.

[60] Debashis Ghosh and Arul M Chinnaiyan. Classification and selection of biomarkers in genomic data using lasso. *BioMed Research International*, 2005(2):147–154, 2005.

[61] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[62] Makoto Yamada, Wittawat Jitkrittum, Leonid Sigal, Eric P Xing, and Masashi Sugiyama. High-dimensional feature selection by feature-wise kernelized lasso. *Neural computation*, 26(1):185–207, 2014.

[63] Makoto Yamada, Avishek Saha, Hua Ouyang, Dawei Yin, and Yi Chang. N 3 lars: Minimum redundancy maximum relevance feature selection for large and high-dimensional data. *arXiv preprint arXiv:1411.2331*, 2014.

[64] Wittawat Jitkrittum, Hirotaka Hachiya, and Masashi Sugiyama. Feature selection via l1-penalized squared-loss mutual information. *IEICE TRANSACTIONS on Information and Systems*, 96(7):1513–1524, 2013.

[65] Wei-Yin Loh. Regression tress with unbiased variable selection and interaction detection. *Statistica Sinica*, pages 361–386, 2002.

[66] Carlo E Bonferroni. *Teoria statistica delle classi e calcolo delle probabilita*. Libreria internazionale Seeber, 1936.

[67] Youngjae Chang. *Robustifying regression and classification trees in the presence of irrelevant variables*. PhD thesis, Department of Statistics, University of Wisconsin, Madison., 2008.

[68] Zheng Zhao, Lei Wang, Huan Liu, et al. Efficient spectral feature selection with minimum redundancy. In *AAAI*, pages 673–678, 2010.

[69] Levi Waldron, Melania Pintilie, Ming-Sound Tsao, Frances A Shepherd, Curtis Huttenhower, and Igor Jurisica. Optimized application of penalized regression methods to diverse genomic data. *Bioinformatics*, 27(24):3399–3406, 2011.

[70] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[71] Gao Huang, Shiji Song, Jatinder ND Gupta, and Cheng Wu. Semi-supervised and unsupervised extreme learning machines. *IEEE Transactions on Cybernetics*, 44(12):2405–2417, 2014.

[72] Mohammed Hindawi, Haytham Elghazel, and Khalid Benabdeslem. Efficient semi-supervised feature selection by an ensemble approach. In *COPEM@ ECML/PKDD. International workshop on complex machine learning problems with ensemble methods*, pages 41–55, 2013.

[73] Ghazaleh Khodabandelou, Charlotte Hug, Rebecca Deneckere, and Camille Salinesi. Supervised vs. unsupervised learning for intentional process model discovery. In *Enterprise, Business-Process and Information Systems Modeling*, pages 215–229. Springer, 2014.

[74] Yang Gu, Yiqiang Chen, Junfa Liu, and Xinlong Jiang. Semi-supervised deep extreme learning machine for wi-fi based localization. *Neurocomputing*, 166:282–293, 2015.

[75] Shiping Wang, Witold Pedrycz, Qingxin Zhu, and William Zhu. Subspace learning for unsupervised feature selection via matrix factorization. *Pattern Recognition*, 48(1):10–19, 2015.

[76] Pengfei Zhu, Wangmeng Zuo, Lei Zhang, Qinghua Hu, and Simon CK Shiu. Unsupervised feature selection by regularized self-representation. *Pattern Recognition*, 48(2):438–446, 2015.

[77] Guangrong Li, Xiaohua Hu, Xiajiong Shen, Xin Chen, and Zhoujun Li. A novel unsupervised feature selection method for bioinformatics data sets through feature clustering. In *Granular Computing, 2008. GrC 2008. IEEE International Conference on*, pages 41–47. IEEE, 2008.

[78] Suhang Wang, Jiliang Tang, and Huan Liu. Embedded unsupervised feature selection. In *AAAI*, pages 470–476. Citeseer, 2015.

[79] Zhiqiang Zeng, Xiaodong Wang, Jian Zhang, and Qun Wu. Semi-supervised feature selection based on local discriminative information. *Neurocomputing*, 173:102–109, 2016.

[80] Xiaojun Chang, Feiping Nie, Yi Yang, and Heng Huang. A convex formulation for semi-supervised multi-label feature selection. In *AAAI*, pages 1171–1177, 2014.

[81] Jin Yao, Qi Mao, Steve Goodison, Volker Mai, and Yijun Sun. Feature selection for unsupervised learning through local learning. *Pattern Recognition Letters*, 53:100–107, 2015.

[82] Luying Liu, Jianchu Kang, Jing Yu, and Zhongliang Wang. A comparative study on unsupervised feature selection methods for text clustering. In *2005 International Conference on Natural Language Processing and Knowledge Engineering*, pages 597–601. IEEE, 2005.

[83] Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. In *Advances in Neural Information Processing Systems*, pages 507–14, 2005.

[84] Zheng Zhao, Fred Morstatter, Shashvata Sharma, Salem Alelyani, Aneeth Anand, and Huan Liu. Advancing feature selection research. *ASU feature selection repository*, pages 1–28, 2010.

[85] Zheng Zhao and Huan Liu. Spectral feature selection for supervised and unsupervised learning. In *Proceedings of the 24th international conference on Machine learning*, pages 1151–1157. ACM, 2007.

[86] Jennifer G Dy and Carla E Brodley. Feature selection for unsupervised learning. *Journal of machine learning research*, 5(Aug):845–889, 2004.

[87] Chenping Hou, Feiping Nie, Xuelong Li, Dongyun Yi, and Yi Wu. Joint embedding learning and sparse regression: A framework for unsupervised feature selection. *IEEE transactions on cybernetics*, 44(6):793–804, 2014.

[88] Ferdi Sarac, Volkan Uslan, Huseyin Seker, and Ahmed Bouridane. Comparison of unsupervised feature selection methods for high-dimensional regression problems in prediction of peptide binding affinity. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 8173–8176. IEEE, 2015.

[89] Ferdi Sarac, Volkan Uslan, Huseyin Seker, and Ahmed Bouridane. Unsupervised selection of rv144 hiv vaccine-induced antibody features correlated to natural killer cell-mediated cytotoxic reactions. In *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*, pages 3072–3075. IEEE, 2016.

[90] Feiping Nie, Shiming Xiang, Yangqing Jia, Changshui Zhang, and Shuicheng Yan. Trace ratio criterion for feature selection. In *AAAI*, volume 2, pages 671–676, 2008.

[91] Zheng Alan Zhao and Huan Liu. *Spectral feature selection for data mining*. CRC Press, 2011.

[92] Deng Cai, Chiyuan Zhang, and Xiaofei He. Unsupervised feature selection for multi-cluster data. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 333–342, 2010.

[93] Yi Yang, Heng Tao Shen, Zhigang Ma, Zi Huang, and Xiaofang Zhou. l2, 1-norm regularized discriminative feature selection for unsupervised learning.

In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 1589, 2011.

[94] Zechao Li, Yi Yang, Jing Liu, Xiaofang Zhou, Hanqing Lu, et al. Unsupervised feature selection using nonnegative spectral analysis. In *AAAI*, 2012.

[95] Mingjie Qian and Chengxiang Zhai. Robust unsupervised feature selection. In *IJCAI*. Citeseer, 2013.

[96] Liang Du and Yi-Dong Shen. Unsupervised feature selection with adaptive structure learning. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 209–218. ACM, 2015.

[97] Pabitra Mitra, CA Murthy, and Sankar K. Pal. Unsupervised feature selection using feature similarity. *IEEE transactions on pattern analysis and machine intelligence*, 24(3):301–312, 2002.

[98] Ahmed K Farahat, Ali Ghodsi, and Mohamed S Kamel. An efficient greedy method for unsupervised feature selection. In *2011 IEEE 11th International Conference on Data Mining*, pages 161–170. IEEE, 2011.

[99] Sankar K Pal and Pabitra Mitra. *Pattern recognition algorithms for data mining*. CRC press, 2004.

[100] Dharmendra S Modha and W Scott Spangler. Feature weighting in k-means clustering. *Machine learning*, 52(3):217–237, 2003.

[101] Alan Miller. *Subset selection in regression*. CRC Press, 2002.

[102] Wen Liu, Xiangshan Meng, Qiqi Xu, Darren Flower, and Tongbin Li. Quantitative prediction of mouse class I MHC peptide binding affinity using support vector machine regression (SVR) models. *BMC Bioinformatics*, 7 (1):182, 2006.

[103] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[104] Guangcan Liu, Zhouchen Lin, and Yong Yu. Multi-output regression on the output manifold. *Pattern Recognition*, 42(11):2737–2743, 2009.

[105] Devis Tuia, Jochem Verrelst, Luis Alonso, Fernando Pérez-Cruz, and Gustavo Camps-Valls. Multioutput support vector regression for remote sensing biophysical parameter estimation. *IEEE Geoscience and Remote Sensing Letters*, 8(4):804–808, 2011.

[106] Saheli Sadanand, Todd J Suscovich, and Galit Alter. Broadly neutralizing antibodies against hiv: New insights to inform vaccine design. *Annual review of medicine*, 67:185–200, 2016.

[107] Angela S Clem et al. Fundamentals of vaccine immunology. *Journal of global infectious diseases*, 3(1):73, 2011.

[108] Morgane Rolland, Paul T Edlefsen, Brendan B Larsen, Sodsai Tovanabutra, Eric Sanders-Buell, Tomer Hertz, Chris Carrico, Sergey Menis, Craig A Magaret, Hasan Ahmed, et al. Increased hiv-1 vaccine efficacy against viruses with genetic signatures in env v2. *Nature*, 490(7420):417–420, 2012.

[109] Hussein S Bagalb. *Cellular and Molecular Biological Studies of a Retroviral Induced Lymphoma, Transmitted via Breast Milk in a Mouse Model*. PhD thesis, University of Toledo, 2008.

[110] AS Perelson, P Essunger, and DD Ho. Dynamics of hiv-1 and cd4+ lymphocytes in vivo. *AIDS (London, England)*, 11:S17–24, 1996.

[111] Charlotta Nilsson, Said Aboud, Muhammad Bakari, Eligius F Lyamuya, Merlin L Robb, Mary A Marovich, Patricia Earl, Bernard Moss, Christina Ochsenbauer, Britta Wahren, et al. Potent functional antibody responses elicited by hiv-i dna priming and boosting with heterologous hiv-1 recombinant mva in healthy tanzanian adults. *PloS one*, 10(4):e0118486, 2015.

[112] Georgia D Tomaras and Barton F Haynes. Strategies for eliciting hiv-1 inhibitory antibodies. *Current Opinion in HIV and AIDS*, 5(5):421, 2010.

[113] Harriet L Robinson. Non-neutralizing antibodies in prevention of hiv infection. *Expert opinion on biological therapy*, 13(2):197–207, 2013.

[114] Yongjun Guan, Marzena Pazgier, Mohammad M Sajadi, Roberta Kamin-Lewis, Salma Al-Darmarki, Robin Flinko, Elena Lovo, Xueji Wu, James E Robinson, Michael S Seaman, et al. Diverse specificity and effector function among human antibodies to hiv-1 envelope glycoprotein epitopes exposed

by cd4 binding. *Proceedings of the National Academy of Sciences*, 110(1): E69–E78, 2013.

[115] Rasheed Ahmad, Sardar TAK Sindhu, Emil Toma, Richard Morisset, Jean Vincelette, Jose Menezes, and Ali Ahmad. Evidence for a correlation between antibody-dependent cellular cytotoxicity-mediating anti-hiv-1 antibodies and prognostic predictors of hiv infection. *Journal of clinical immunology*, 21(3):227–233, 2001.

[116] Margaret E Ackerman, Anne-Sophie Dugast, and Galit Alter. Emerging concepts on the role of innate immunity in the prevention and control of hiv infection. *Annual review of medicine*, 63:113–130, 2012.

[117] Margaret E Ackerman and Galit Alter. Opportunities to exploit non-neutralizing hiv-specific antibody activity. *Current HIV research*, 11(5): 365–377, 2013.

[118] Stanley A Plotkin. Correlates of protection induced by vaccination. *Clinical and Vaccine Immunology*, 17(7):1055–1065, 2010.

[119] Ickwon Choi, Amy W Chung, Todd J Suscovich, Supachai Rerks-Ngarm, Punnee Pitisuttithum, Sorachai Nitayaphan, Jaranit Kaewkungwal, Robert J O'Connell, Donald Francis, Merlin L Robb, et al. Machine learning methods enable predictive modeling of antibody feature: function relationships in rv144 vaccinees. *PLoS computational biology*, 11(4): e1004185, 2015.

[120] Wen Shi Lee, Matthew Sidney Parsons, Stephen John Kent, and Marit Lichtfuss. Can hiv-1-specific adcc assist the clearance of reactivated latently infected cells? *Frontiers in immunology*, 6, 2015.

[121] RA Freitas Jr. Human body chemical composition (section 3.1). *Nanomedicine: Basic Capabilities*, 1.

[122] J.M. Berg, J.L. Tymoczko, and L. Stryer. *Biochemistry, Fifth Edition*. W.H. Freeman, 2002. ISBN 9780716730514. URL https://books.google.co.uk/books?id=uDFqAAAAMAAJ.

[123] M Jesus Perez de Vega, Mercedes Martín-Martínez, and Rosario González-Muñiz. Modulation of protein-protein interactions by stabilizing/mimicking protein secondary structure elements. *Current topics in medicinal chemistry*, 7(1):33–62, 2007.

[124] Sébastien Giguère, Mario Marchand, François Laviolette, Alexandre Drouin, and Jacques Corbeil. Learning a peptide-protein binding affinity predictor with kernel ridge regression. *BMC bioinformatics*, 14(1):82, 2013.

[125] Luca Costantino and Daniela Barlocco. Privileged structures as leads in medicinal chemistry. *Current medicinal chemistry*, 13(1):65–85, 2006.

[126] Pavel P Kuksa, Martin Renqiang Min, Rishabh Dugar, and Mark Gerstein. High-order neural networks and kernel methods for peptide-mhc binding prediction. *Bioinformatics*, 31(22):3600–3607, 2015.

[127] Bruce Draper. *Large Margin Methods for Partner Specific Prediction of Interfaces in Protein Complexes*. PhD thesis, Colorado State University, 2014.

[128] Pingping Guan, Irini A Doytchinova, Valerie A Walshe, Persephone Borrow, and Darren R Flower. Analysis of peptide- protein binding using amino acid descriptors: Prediction and experimental verification for human histocompatibility complex hla-a* 0201. *Journal of medicinal chemistry*, 48(23):7418–7425, 2005.

[129] Ovidiu Ivanuic. Comparative evaluation of prediction algorithms (CoEPrA), 2006. URL http://www.coepra.org/.

[130] Markus Sällman Almén, Emil K Nilsson, Josefin A Jacobsson, Ineta Kalnina, Janis Klovins, Robert Fredriksson, and Helgi B Schiöth. Genome-wide analysis reveals dna methylation markers that vary with both age and obesity. *Gene*, 548(1):61–67, 2014.

[131] Yun Huang, Jing Yan, Jiayi Hou, Xiaodan Fu, Luyao Li, and Yiping Hou. Developing a dna methylation assay for human age prediction in blood and bloodstain. *Forensic Science International: Genetics*, 17:129–136, 2015.

[132] DW Haslam and WP James. Obesity. Lancet. 2005; 366; 1197-1209. *International Journal of Advancements in Research and Technology*, 1.

[133] Luigi Bouchard, Rémi Rabasa-Lhoret, May Faraj, Marie-Ève Lavoie, Jonathan Mill, Louis Pérusse, and Marie-Claude Vohl. Differential epigenomic and transcriptomic responses in subcutaneous adipose tissue between low and high responders to caloric restriction. *The American journal of clinical nutrition*, 91(2):309–320, 2010.

[134] Muhamad Hanafiah Juni. Obesity: A public health threats in developing countries. *International Journal of Public Health and Clinical Sciences*, 2 (2), 2015.

[135] Jill Waalen. The genetics of human obesity. *Translational Research*, 164 (4):293–301, 2014.

[136] Paul W Franks and Charlotte Ling. Epigenetics and obesity: the devil is in the details. *BMC medicine*, 8(1):88, 2010.

[137] Veronica Davé, Paul Yousefi, Karen Huen, Vitaly Volberg, and Nina Holland. Relationship between expression and methylation of obesity-related genes in children. *Mutagenesis*, page geu089, 2015.

[138] Shuo Wang, Jieyun Song, Yide Yang, Yining Zhang, Haijun Wang, and Jun Ma. Hif3a dna methylation is associated with childhood obesity and alt. *PloS one*, 10(12):e0145944, 2015.

[139] Pan Du, Xiao Zhang, Chiang-Ching Huang, Nadereh Jafari, Warren A Kibbe, Lifang Hou, and Simon M Lin. Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*, 11(1):587, 2010.

[140] Marc Jung and Gerd P Pfeifer. Aging and dna methylation. *BMC biology*, 13(1):7, 2015.

[141] Meaghan J Jones, Sarah J Goodman, and Michael S Kobor. Dna methylation and healthy human aging. *Aging cell*, 14(6):924–932, 2015.

[142] A Starnawska, Q Tan, A Lenart, M McGue, O Mors, AD Børglum, K Christensen, M Nyegaard, and L Christiansen. Blood dna methylation age is not associated with cognitive functioning in middle-aged monozygotic twins. *Neurobiology of Aging*, 50:60–63, 2017.

[143] Carola Ingrid Weidner, Qiong Lin, Carmen Maike Koch, Lewin Eisele, Fabian Beier, Patrick Ziegler, Dirk Olaf Bauerschlag, Karl-Heinz Jöckel, Raimund Erbel, Thomas Walter Mühleisen, et al. Aging of blood can be tracked by dna methylation changes at just three cpg sites. *Genome biology*, 15(2):R24, 2014.

[144] Bram Bekaert, Aubeline Kamalandua, Sara C Zapico, Wim Van de Voorde, and Ronny Decorte. Improved age determination of blood and teeth samples using a selected set of dna methylation markers. *Epigenetics*, 10 (10):922–930, 2015.

[145] Gregory Hannum, Justin Guinney, Ling Zhao, Li Zhang, Guy Hughes, SriniVas Sadda, Brandy Klotzle, Marina Bibikova, Jian-Bing Fan, Yuan Gao, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular cell*, 49(2):359–367, 2013.

[146] Jun Chin Ang, Andri Mirzal, Habibollah Haron, and Haza Nuzly Abdull Hamed. Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. *IEEE/ACM transactions on computational biology and bioinformatics*, 13(5):971–989, 2016.

[147] Ulisses Braga-Neto, Ronaldo Hashimoto, Edward R Dougherty, Danh V Nguyen, and Raymond J Carroll. Is cross-validation better than resubstitution for ranking genes? *Bioinformatics*, 20(2):253–258, 2004.

[148] Norman Levinson. The wiener (root mean square) error criterion in filter design and prediction. *Journal of Mathematics and Physics*, 25(1):261–278, 1946.

[149] Tianfeng Chai and Roland R Draxler. Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. *Geoscientific Model Development*, 7(3):1247–1250, 2014.

[150] Henri Theil. Applied economic forecasting. 1971.

[151] Volkan Uslan and Huseyin Seker. Quantitative prediction of peptide binding affinity by using hybrid fuzzy support vector regression. *Applied Soft Computing*, 43:210–221, 2016.

[152] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium*

on mathematical statistics and probability, volume 1, pages 281–297. Oakland, CA, USA., 1967.

[153] N Karthikeyani Visalakshi and J Suguna. K-means clustering using max-min distance measure. In *Fuzzy Information Processing Society, 2009. NAFIPS 2009. Annual Meeting of the North American*, pages 1–6. IEEE, 2009.

[154] Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678, 2005.

[155] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. Wiley, New York, 1973.

[156] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, 24(7):881–892, 2002.

[157] Shehroz S Khan and Amir Ahmad. Cluster center initialization algorithm for k-means clustering. *Pattern recognition letters*, 25(11):1293–1302, 2004.

[158] Krista Rizman Žalik. An efficient k-means clustering algorithm. *Pattern Recognition Letters*, 29(9):1385–1391, 2008.

[159] Adil M Bagirov. Modified global k-means algorithm for minimum sum-of-squares clustering problems. *Pattern Recognition*, 41(10):3192–3199, 2008.

[160] V Karthikeyani and J Suguna. K-means clustering using max-min distance measure. In *The 28th North American Fuzzy Information Processing Society Annual Conference (NAFIPS2009)*, 2009.

[161] Yordan P Raykov, Alexis Boukouvalas, Fahd Baig, and Max A Little. What to do when k-means clustering fails: a simple yet principled alternative algorithm. *PloS one*, 11(9):e0162259, 2016.

[162] Nameirakpam Dhanachandra, Khumanthem Manglem, and Yambem Jina Chanu. Image segmentation using k-means clustering algorithm and subtractive clustering algorithm. *Procedia Computer Science*, 54:764–771, 2015.

[163] Paulo JG Lisboa and Emmanuel C Ifeachor. *Artificial neural networks in biomedicine*. Springer Science & Business Media, 2000.

[164] Wayne S DeSarbo, J Douglas Carroll, Linda A Clark, and Paul E Green. Synthesized clustering: A method for amalgamating alternative clustering bases with differential weighting of variables. *Psychometrika*, 49(1):57–78, 1984.

[165] Liping Jing, Michael K Ng, and Joshua Zhexue Huang. An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. *IEEE Transactions on knowledge and data engineering*, 19(8), 2007.

[166] Joshua Zhexue Huang, Michael K Ng, Hongqiang Rong, and Zichen Li. Automated variable weighting in k-means type clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):657–668, 2005.

[167] Harris Drucker, C. J. C. Burges, Linda Kaufman, Alexander J. Smola, and Vladimir N. Vapnik. *Support Vector Regression Machines*, volume 9 of *Advances in Neural Information Processing Systems*. MIT Press, 1996.

[168] Mariette Awad and Rahul Khanna. *Deep Neural Networks*, pages 127–147. Apress, Berkeley, CA, 2015. ISBN 978-1-4302-5990-9. doi: 10.1007/978-1-4302-5990-9_7. URL http://dx.doi.org/10.1007/978-1-4302-5990-9_7.

[169] Yoshua Bengio et al. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

[170] Rania Ibrahim, Noha A Yousri, Mohamed A Ismail, and Nagwa M El-Makky. Multi-level gene/mirna feature selection using deep belief nets and active learning. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3957–3960. IEEE, 2014.

[171] Lei Zhao, Qinghua Hu, and Wenwu Wang. Heterogeneous feature selection with multi-modal deep neural networks and sparse group lasso. *IEEE Transactions on Multimedia*, 17(11):1936–1948, 2015.

[172] Ha Van-Sang and Nguyen Ha-Nam. Credit scoring with a feature selection approach based deep learning. In *MATEC Web of Conferences*, volume 54. EDP Sciences, 2016.

[173] Bum-Chae Kim, Yu Sub Sung, and Heung-Il Suk. Deep feature learning for pulmonary nodule classification in a lung ct. In *2016 4th International Winter Conference on Brain-Computer Interface (BCI)*, pages 1–3. IEEE, 2016.

[174] Yifeng Li, Chih-Yu Chen, and Wyeth W Wasserman. Deep feature selection: Theory and application to identify enhancers and promoters. In *International Conference on Research in Computational Molecular Biology*, pages 205–217. Springer, 2015.

[175] Shekoofeh Azizi, Farhad Imani, Bo Zhuang, Amir Tahmasebi, Jin Tae Kwak, Sheng Xu, Nishant Uniyal, Baris Turkbey, Peter Choyke, Peter Pinto, et al. Ultrasound-based detection of prostate cancer using automatic feature selection with deep belief networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 70–77. Springer, 2015.

[176] Qin Zou, Lihao Ni, Tong Zhang, and Qian Wang. Deep learning based feature selection for remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters*, 12(11):2321–2325, 2015.

[177] Yifeng Li, Wenqiang Shi, and Wyeth W Wasserman. Genome-wide prediction of cis-regulatory regions using supervised deep learning methods. *bioRxiv*, page 041616, 2016.

[178] Saikat Basu, Sangram Ganguly, Supratik Mukhopadhyay, Robert DiBiano, Manohar Karki, and Ramakrishna Nemani. Deepsat: a learning framework for satellite imagery. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 37. ACM, 2015.

[179] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[180] Mostafa A Salama, Aboul Ella Hassanien, and Aly A Fahmy. Deep belief network for clustering and classification of a continuous data. In *The 10th IEEE International Symposium on Signal Processing and Information Technology*, pages 473–477. IEEE, 2010.

[181] Ruslan Salakhutdinov. *Learning deep generative models*. PhD thesis, University of Toronto, 2009.

[182] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, 2009.

[183] Nan Wang, Jan Melchior, and Laurenz Wiskott. Gaussian-binary restricted boltzmann machines on modeling natural image statistics. *arXiv preprint arXiv:1401.5900*, 2014.

[184] Sarunas J Raudys, Anil K Jain, et al. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on pattern analysis and machine intelligence*, 13(3):252–264, 1991.

[185] Mohamed Medhat Gaber, Mihaela Cocea, Nirmalie Wiratunga, and Ayse Goker. *Advances in social media analysis*, volume 602. Springer, 2015.

[186] Ferdi Sarac and Huseyin Seker. An instance selection framework for mining data streams to predict antibody-feature function relationships on rv144 hiv vaccine recipients. In *Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on*, pages 003356–003361. IEEE, 2016.

[187] Asja Fischer and Christian Igel. Training restricted boltzmann machines: An introduction. *Pattern Recognition*, 47(1):25–39, 2014.