EMBEDDED FEATURE SELECTION METHODS WITH HIGH DIMENSIONALITY FOR ELASTIC NET AND LOGISTIC REGRESSION MODELS

AIEDH MRISI ALHARTHI

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy

Faculty of Science
Universiti Teknologi Malaysia

OCTOBER 2022

DEDICATION

This thesis is dedicated to my late father and my kindest mother.

ACKNOWLEDGEMENT

All praise and thanks to Allah S.W.T for granting me vast strength and perfect blessing to complete my Ph.D. study.

First and foremost, I would like to express my sincere appreciation to my main thesis supervisor, Professor Ts. Dr. Muhammad Hisyam Lee, for his immense encouragement, guidance, critics, and friendship. Indeed, under his supervision, it was an exciting Ph.D. journey for me, and as a result, the lessons I have gained along the journey will undoubtedly serve me well for the rest of my life. Furthermore, I am also very thankful to my co-supervisor, Professor Dr. Zakariya Yahya Algamal, for his guidance, advice, and motivation. Without their continued support and interest, this thesis would not have been the same as presented here.

I would like to thank all who helped me directly or indirectly in bringing this thesis to light. I send my gratitude to all of them. I am also indebted to Universiti Teknologi Malaysia (UTM) for providing the required facilities for the Ph.D. program and to Taif University, KSA, for funding my Ph.D. research.

Last but not least, a special thanks to my tender mother and all of my dearest family members, who have always done their best to support me throughout my life. Without their encouragement, I would not have been able to achieve everything I have today. They are and will always be a source of strength and courage for me.

ABSTRACT

Feature selection and classification in high-dimensional data is a challenging problem in scientific research such as biology, medicine, and finance. In such data, highly correlated features and missing data often exist. Therefore, selecting informative features and adequate handling of missing values are significant to find an optimal model in terms of interpretability and prediction accuracy. In recent years, embedded feature selection methods, including penalized regression, have attracted many statisticians since these methods often obtain model estimates with higher prediction accuracy. Nevertheless, most penalized methods lack the consistency of feature selection, encouragement of grouping effects, and handling missing values when dealing with high-dimensional data. Hence, this study aims to improve the process of feature selection and handling of missing values by proposing several improvements in the penalized high-dimensional approaches. An alternative initial weight was introduced in the adaptive least absolute shrinkage and selection operator (LASSO) to improve the feature selection performance. Then, an initial ratio and adjusted variance weights inside the L_1-norm penalty of the adaptive elastic net are proposed to encourage the grouping effect. Furthermore, imputation penalized logistic regression with the adaptive LASSO approach was proposed to enhance the handling of missing values in high-dimensional data. Simulation studies with varying numbers of predictor variables, sample sizes, correlation coefficients, and the proportion of missing values were performed to evaluate the effectiveness of the proposed methods. The proposed adaptive LASSO methods were also compared with LASSO and other versions of adaptive LASSO methods, while the proposed adaptive elastic net methods were compared with the existing elastic net and adaptive elastic net methods. The proposed methods were also applied to a chemometrics dataset and eight gene expression microarray datasets in which the number of genes (features) is more than the sample size. The results indicated that the proposed methods outperform their competitors in selecting the most relevant features and achieving higher classification accuracy, sensitivity, and specificity values. It also reduces dimensionality and selects the most helpful features for cancer classification, resulting in optimal models that concurrently perform feature selection and patient classification. On the other hand, the proposed adaptive elastic net method is shown superior to the other methods in terms of encouraging the group effect. In conclusion, this study shows that the proposed methods are appropriate for gene expression data classification and other high-dimensional data classification analyses.

ABSTRAK

Pemilihan ciri dan klasifikasi di dalam data berdimensi tinggi adalah permasalahan yang mencabar dalam penyelidikan saintifik seperti biologi, perubatan, dan kewangan. Dalam data begini, seringkali wujud ciri data yang berkorelasi tinggi dan data hilang. Oleh itu, pemilihan ciri berinformatif dan keupayaan menangani masalah nilai hilang adalah signifikan untuk mendapatkan model yang optima dari segi pentafsiran dan ketepatan ramalan. Beberapa tahun kebelakangan ini, kaedah pemilihan ciri terbenam, termasuklah regresi terhukum telah menarik minat ramai ahli statistik kerana kaedah ini sering memperolehi penganggaran model dengan kejituan yang lebih tinggi. Walau bagaimanapun, kebanyakan kaedah terhukum kurang menepati pemilihan ciri yang konsisten, tidak mempertimbangkan kesan kelompok dan pengendalian data hilang apabila melibatkan data berdimensi tinggi. Maka, matlamat kajian ini ialah menambahbaik proses bagi pemilihan ciri dan pengendalian nilai hilang dengan mencadangkan beberapa penambahbaikan di dalam dimensi tinggi terhukum. Satu pemberat awal alternatif telah diperkenalkan di pengecutan mutlak terkecil mudah suai dan pemilihan operator (LASSO) bagi menambahbaik prestasi pemilihan ciri. Kemudian, satu nisbah awal dan varians pemberat dilaraskan hukum L_1-norm elastik jaring mudah suai telah dicadangkan untuk menggalakkan kesan pengumpulan. Tambahan pula, imputasi regresi logistik terhukum dengan pendekatan LASSO mudah suai telah dicadangkan untuk meningkatkan pengendalian nilai hilang di dalam data berdimensi tinggi. Kajian simulasi dengan nombor pembolehubah peramal, saiz sampel, pekali korelasi, dan perkadaran nilai hilang yang berbeza-beza dilakukan untuk menilai keberkesanan kaedah yang dicadangkan. Kaedah LASSO mudah suai yang dicadangkan turut dibandingkan dengan LASSO dan kaedah LASSO mudah suai versi lain, manakala kaedah elastik jaring mudah suai yang dicadangkan dibandingkan dengan elastik jaring dan elastik jaring mudah suai yang sedia ada. Kaedah-kaedah yang dicadangkan turut diaplikasikan kepada satu set data kemometrik dan lapan set data mikrotatasusunan ekpresi gen yang mana bilangan gen (ciri) lebih daripada saiz sampel. Keputusan menunjukkan kaedah yang dicadangkan mengatasi prestasi pesaing-pesaingnya dalam memilih ciri yang paling relevan dan mencapai nilai klasifikasi kejituan, sensitiviti dan kekhususan yang lebih tinggi. Ianya turut mengurangkan dimensi dan memilih ciri yang paling berguna bagi klasifikasi kanser, menghasilkan model optima yang dapat melakukan pemilihan ciri dan klasifikasi pesakit secara serentak. Selain itu, kaedah elastik jaring mudah suai yang dicadangkan ditunjukkan lebih baik daripada kaedah lain daripada segi penggalakkan kesan pengumpulan. Kesimpulannya, kajian ini menunjukkan bahawa kaedah-kaedah yang dicadangkan adalah sesuai untuk klasifikasi data ekspresi gen dan analisis klasifikasi data berdimensi tinggi yang lain.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACC	-	Accuracy
AEN	-	Adaptive Elastic Net
AIC	-	Akaike Information Criteria
ALASSO	-	Adaptive Least absolute shrinkage and selection
AUC	-	Area under the Curve
Aut	-	Autism
BIC	-	Bayesian Information Criteria
Bip	-	Bipolar disorder
BSS	-	Between-groups sum of squares
CA	-	Classification Accuracy
CBP	-	Correlation-Based Penalty
CBEP	-	Correlation Based Elastic Penalty
CDA	-	coordinate decent algorithm
CV	-	Cross-Validation
DLBCL	-	Diffuse Large B-cell Lymphoma
EN	-	Elastic Net
ERR	-	Error Rate
FP	-	False Positive
FN	-	False Negative
GLM	-	Generalized Linear Model
GM	-	Geometric Mean
LASSO	-	Least absolute shrinkage and selection operator
LogiR	-	Logistic Regression
MAR	-	Missing at Random
MCAR	-	Missing Completely at Random
MI	-	Multiple Imputation

MR	-	Misclassification Rate
MLE	-	Maximum Likelihood Estimates
MS	-	Model Size
MSE	-	Mean Squared Error
NMAR	-	Not Missing at Random
OLS	-	Ordinary Least Squares
1-DWM	-	One-dimensional weighted Mahalanobis distance
PLM	-	Penalized Likelihood Method
PLogiR	-	Penalized Logistic Regression
PLR	-	Penalized Linear Regression
QSAR	-	Quantitative Structure-Activity Relationship
RSS	-	Residual Sum of Squares
SCAD	-	Smoothly Clipped Absolute Deviation
Sco	-	Sarcoma
SEN	-	Sensitivity
SPE	-	Specificity
TP	-	True Positive
TPR	-	True Positive Rate
TN	-	True Negative
TNR	-	True Negative Rate
WDBC	-	Breast Cancer Wisconsin (Diagnostic)
WSS	-	Within-groups sum of squares

LIST OF SYMBOLS

L_1	-	L_1-norm
L_2	-	L_2-norm
λ	-	Tuning parameter of penalized method for cross validation
σ^2	-	Variance
σ^2_{wj}	-	Weighted variance
$g(.)$	-	Penalty term
x	-	Predictor variables (features)
$\mathbf{X}$	-	Data matrix
n	-	Number of observations (sample size)
p	-	Number of predictor variable (features)
y	-	Response variable
j^{th}	-	A feature number
i^{th}	-	A observation number
$(.)^T$	-	Transpose
β	-	Regression coefficients
ϵ	-	vector of random errors
$\mathbf{I}$	-	Identity matrix
L	-	Maximum values of the likelihood function for the model
$\ell(.)$	-	Log-likelihood function
$\pi(x_i)$	-	Conditional probability of y equal to 1 given x
ω	-	An initial weight
γ	-	positive constant
k	-	Number of folds of cross-validation
$P(.)$	-	Probability

CHAPTER 1

INTRODUCTION

In this introduction, the study emphasizes the need to improve feature selection and classification methods to face the challenges imposed by the high dimensionality of the data, where some classification methods may not be applicable for analyzing data directly. Section 1.1 provides a background of the study, which affirms the recently developed methods and techniques that can be used to deal with high-dimensional data. In Section 1.2, this research states the problem of the study focusing on the emerging new methods in generalized linear models with high-dimensional data. Then, the study states in Section 1.3 some scientific research questions that were answered in this study. The last four sections of this chapter were devoted to the objectives, significance, scope, and limitations of the study.

1.1 Research Background

As data collection technology evolves over the last few years, high-dimensional data are becoming increasingly available such as genetic, genomic, biological, social, economic, and chemometric data. In these kinds of data, the number of predictor variables (feature) is hugely larger than the sample size; this is called high dimensionality. For example, in the genomic studies, tens of thousands of genes could be involved in a study, while the number of participants in that study (sample size) is less than 100 persons or so (Adragni, 2015; Yang *et al.*, 2018; Manhrawy *et al.*, 2021). Also, high dimensional data appears in chemometrics when modeling "quantitative structure activity relationship" (QSAR), where the number of molecular descriptors surpasses the number of compounds (Al-Fakih *et al.*, 2019). This represents a challenge to the statisticians and researchers as the use of traditional statistical methods and techniques to analyze the high dimensional data is impossible (Algamal and Mohammad Ali, 2017).

Any statistical study involving high-dimensional data seeks to find a statistical model that can be used to classify the variables and make predictions. When dealing with high dimensional data, the number of columns by far exceeds the number of rows in the design matrix representing this data. As a result, the matrix is not invertible (singular) (Algamal *et al.*, 2017; Filzmoser *et al.*, 2012; Fu and Xu, 2012). That is, the linear equation used to find the coefficients matrix has no solution. Moreover, in any linear model relating the response variable to the predictor variables, the prediction error increases as the predictor variables increase. This makes the traditional statistical models, including the generalized linear models (GLM), inapplicable and inappropriate. The statistical studies concerning high dimensional data suffer from model overfitting, estimation instability, prediction, and interpretation (Pourahmadi, 2013).

To overcome the problems associated with processing and tackling high dimensional data, statisticians and researchers have recently been developing new methods to deal with high dimensional data. One vital technique is (predictor, explanatory, or feature) variable selection, which plays an essential role in statistical modeling when dealing with high dimensional data. The aim of the variable selection technique is to choose as small as a possible subset of the relevant variables from a large set of predictor variables. That is, this technique is considered a classifier. It classifies the predictor variables according to their relevance to the problem that the statistical study seeks to address clearly. The selection process improves the statistical model in the sense of accuracy and interpretability. Consequently, it decreases the effect of multicollinearity and prevents overfitting (Liu *et al.*, 2018; Fan and Lv, 2010).

Traditional subset selection methods such as backward elimination, forward selection, and stepwise selection methods often perform poorly in the sense of both variable selection and coefficients estimation in linear models, especially in high-dimensional data, when multicollinearity is present. Furthermore, these traditional methods computationally become more expensive in the high dimensional case. For example, backward elimination fails because it starts with all predictor variables. On the contrary, both forward selection and stepwise selection start with a model consisting of a single predictor variable, which computationally make them more expensive as the potentially time-consuming fitting has to be performed many times (Rish and

Grabarnik, 2014). Therefore, due to the high dimensionality of the data, the classical variable selection methods (such as backward elimination, forward selection, stepwise selection, Akaike information criterion (AIC), Bayesian information criterion (BIC), and others) are impractical, inefficient, and time-consuming (Bühlmann and van de Geer, 2011; Chen and Chen, 2012).

Consequently, over the last decades, researchers have developed a variety of feature selection techniques. These techniques are divided into three groups. The first group is filter approaches. It includes the most common feature selection techniques, in which each feature is evaluated individually, irrespective of how well it performs in the group. The second group is wrapper approaches. It evaluates the feature group selection process using a variety of algorithms. Despite wrapper techniques, such as "forward feature selection" and "backward feature elimination," being more effective in feature selection than filter methods, wrapper methods are computationally very expensive. The embedded methods are the third group, which incorporates the benefits of both the filter and wrapper groups. It contains penalization techniques that can model and select features simultaneously (Agrawal *et al.*, 2021; Li *et al.*, 2020; Liu *et al.*, 2018).

Recently, statisticians have set a flexible framework of penalized methods that have proven to be practical, efficient, and accurate when dealing with high-dimensional data. In these methods, a penalty term is added to the statistical model with the aim of reducing high dimensionality by selecting a small subset of the vast set of predictor variables. One advantage of these methods is to reduce the complexity of the statistical model and provide criteria for variable selection and classification. Associated with these penalizing methods constraints that are based on L_1-norm, L_2-norm, or both L_1 and L_2 norms of the model coefficients. These constraints force the coefficients of the irrelevant variables to shrink to zero. The amount of penalty term provides a tradeoff between the variance and the bias of the selected statistical model. As this amount increases, the size of the selected subset of predictor variables decreases and vice versa. On the other hand, the minor penalty leads to selecting more predictor variables with low bias but significant variance. In contrast, a high penalty leads to choosing a small number of predictor variables with more significant bias but lower variance. Therefore, the suitable choice for the amount of the penalty term controls prediction accuracy and makes the model interpretable (Casella *et al.*, 2013; Doerken *et al.*, 2019).

Therefore, the ridge regression was introduced by Hoerl and Kennard (1970) is used to overcome the multicollinearity problem produced by the linear regression model. It uses L_2-norm penalty to shrink the regression coefficients towards zero, but it never makes them equal to zero. It is one of the most common penalizing methods. Ridge regression adds the L_2-norm based penalty to the residual sum of squares. As a result, it reduces the variance of the parameter estimators, which gives better properties in both estimation and prediction. Although as a tradeoff tool, the estimated parameters are biased and have some limitations, such as it is not capable of performing the variable selection. Therefore, it produces uninterpretable statistical models.

Another commonly used penalizing method is "Least absolute shrinkage and selection operator" (LASSO), which was proposed by Tibshirani (1996). It uses L_1-norm penalty to shrink the coefficients of some predictor variables to zero. Therefore, it is an efficient classifier and variable selection method. However, despite the advantage of LASSO of being a good variable selection tool, it has some limitations and shortcomings. First, it cannot select more variables than the number of observations because of the nature of the convex optimization problem. This seems to be a limiting feature for a variable selection method (Zou and Hastie, 2005). Second, in the presence of multicollinearity, LASSO does not encourage group selection. That is, it selects only one variable from the group and does not care which one is selected (Zou and Hastie, 2005). The third shortcoming is that LASSO does not enjoy the oracle properties, which refer to the consistency of LASSO as an estimator and the ability of LASSO to select the exact right features whose coefficients are not equal to zero. In other words, using the language of Fan and Li (2001), a penalty term is called enjoy oracle properties when it can identify the right subset model (consistent variable selection), and it has an asymptotic normal distribution.

The limitations of the LASSO and ridge methods motivated statisticians to improve them and develop new methods. For example, Zou and Hastie (2005) introduced the elastic net method, where the penalty is based on a linear combination of L_1 and L_2 norms. Tutz and Ulbricht (2009) introduced a correlation-based penalty method as an alternative to the elastic net method. Although this correlation-based penalty has the advantages of the ridge method of dealing with grouped variables, it does not perform variable selection. El Anbari and Mkhadri (2013) claimed that if the absolute correlation between predictor variables is less than 0.95, the elastic net may be slightly less reliable. In addition, the elastic net does not incorporate the information about the data into the L_2-norm during the computation. This motivated the two authors to use the correlation-based penalty instead of the L_2-norm and the L_1-norm penalties. Consequently, they proposed to use the correlation-based penalty instead of the L_1 and L_2-norms. In fact, they needed to amend the L_2-norm in the elastic net instead of replacing it with the correlation-based penalty. The reason is that the correlation-based penalty gives wrong estimates when the correlation between variables is perfect. Moreover, this amendment uses the same algorithm that is used in computing the elastic net model, which may be helpful in reducing the time of computation.

As far as the oracle properties are concerned, Fan and Li (2001) showed that LASSO does not have the oracle properties because of the inconsistency it has in variable selection. As a result, the identification of the true model cannot be guaranteed. Furthermore, the efficiency of its estimators is less than that of the oracle. To address this issue, Zou (2006) introduced the adaptive LASSO (ALASSO) method, which penalizes various coefficients in the L_1-norm penalty term with different weights. He proved that if the small (large) weights are chosen to penalize the coefficients of the important (unimportant) predictor variables, then the ALASSO model becomes consistent. For the initial weight, Zou (2006) used the ordinary least squares (OLS) estimates inside the L_1-norm penalty term, but in the presence of multicollinearity, he used the ridge regression estimates.

However, in high dimensional data, the OLS and the maximum likelihood estimates (MLE) are not available, and, therefore, the ALASSO is no longer applicable. Hence, some researchers used the LASSO estimates as an initial weight (Lian, 2012). Furthermore, the ALASSO method cannot handle the situation of multicollinearity and cannot select more variables than the number of observations. Consequently, Zou and

Zhang (2009) proposed the adaptive elastic net (AEN) method by replacing the L_1-norm penalty with the ALASSO penalty. He employed the elastic net estimates as an initial weight. In fact, in both low and high-dimensional data, using LASSO estimates and elastic net estimates as initial weights in the ALASSO and AEN, respectively, may not be appropriate. This is because both the LASSO and the elastic net are inconsistent in selected variables. For these reasons, this study proposes appropriate alternative initial weights in the case of dealing with high dimensional data.

In view of that, high-dimensional data frequently comprises a substantial amount of missing data, making it challenging to use conventional imputation methods appropriately. According to previous research, most microarray datasets are incomplete to varying degrees, ranging from 50% to 90% (Chen *et al.*, 2016; Wang *et al.*, 2021). In addition, missing values are present in 45 % of the datasets in the University of California Irvine (UCI) repository (Tran *et al.*, 2016), which is one of the most common data stores for benchmarking machine learning problems (Asuncion and Newman, 2007). Missing data is increasingly being handled with the use of multiple imputation (MI) Rubin (1996); Little and Rubin (2019), which has seen major advancements in techniques and software (van Buuren and Groothuis-Oudshoorn, 2011; Su *et al.*, 2011). However, MI approaches may not work correctly in high-dimensional data (Zahid and Heumann, 2019; Zhao and Long, 2016). For such cases, penalized regression approaches have drawn a lot of attention in recent literature, including LASSO, to perform simultaneous parameter estimation and feature selection (Deng *et al.*, 2016). However, LASSO has some limitations, which are stated above. Against this backdrop, this study proposes adaptive LASSO with imputation penalized logistic regression for being more appropriate in such a case as an extension of the penalized methods to improve the performance and impute missing values.

1.2 Problem Statements

Penalized methods play an essential role in the feature selection and classification of high-dimensional data. One commonly used method is LASSO, which has some shortcomings. First, it cannot select more predictor variables than the number of observations. Second, in the presence of multicollinearity, the LASSO selects one variable from a highly correlated group of variables and leaves the others. Third, the LASSO lacks the oracle properties. As a result, it is an inconsistent feature selection tool. Moreover, the elastic net penalty method is considered the most frequent penalized method that overcomes the first two shortcomings of LASSO. Unfortunately, it outperforms LASSO only when there are highly correlated predictor variables. However, a high correlation among predictor variables may not exist in many situations. This is considered one of the drawbacks of the elastic net. Besides, although a correlation-based penalty was proposed instead of using L_2-norm penalty in elastic net, it no longer gives an accurate estimation when the correlation among variables is perfect.

The limitations mentioned above of LASSO and elastic net motivated statisticians to use adaptive LASSO and elastic net in order to overcome the problems of the LASSO and elastic net methods. Adaptive LASSO basically uses the OLS estimates as initial weights. However, this is no more valid in high dimensional data. Despite several statisticians used the LASSO estimates as initial weights. On the other hand, adaptive elastic net uses elastic net estimates as initial weights. However, this is not an appropriate choice for the initial weights because both LASSO and elastic net lack the oracle properties. Furthermore, both of these methods do not consider the weights for all the features in any implementation. In addition, one of the most vital issues with high-dimensional data is that it often contains large quantities of missing data that common multiple imputation approaches may not work correctly.

Therefore, the search for effective adaptive penalizing methods in high dimensional data has become a necessity in order to improve some penalizing methods so that they can effectively select features in order to achieve high prediction, classification accuracy, stability and consistency, and the ability to adequately deal with different situations of high-dimensional data including missing values and grouping effect.

1.3 Research Questions

In light of the problem statements, the following questions were tackled in this study.

(a) How to construct adaptive penalizing methods that improve the prediction accuracy for high dimensional data?

(b) How to propose adaptive penalized methods that work on on the grouping effect?

(c) How to propose an imputation method that can handle missing values in high-dimensional data?

(d) How to evaluate the performance of proposed adaptive penalizing methods?

1.4 Research Objectives

The research objectives are as follows:

(a) To improve the adaptive LASSO by using alternative initial weights for logistic regression models with high-dimensional data.

(b) To construct an adaptive elastic net by employing new initial weights inside the L_1-norm to encourage the grouping effect in high-dimensional data.

(c) To propose an imputation method for penalized logistic regression with adaptive LASSO.

(d) To evaluate the performance of proposed methods using simulation studies and real-world data.

1.5 Significance of the Study

Improving effective penalizing methods is essential to deal with high-dimensional data to guarantee high performance in prediction, handling of missing values, and classification in terms of accuracy and consistency. Therefore, these methods have been a major concern to many statisticians and researchers. This study thus focused on improving penalizing methods to achieve such desired unique advantages of high-performance accuracy, stability, and consistency. It is known that every technique has its strengths and limitations; hence the need for adaptive penalizing methods become necessary. The results of the proposed penalizing methods improved the accuracy of prediction, classification, and feature selection, compared to other existing penalizing methods.

Furthermore, the finding of this study can benefit to early diagnosis of patients with cancer that machine learning approaches play an important role in classification, analysis, and prediction in medical science today. The importance of curing patients and safe lives with early detection ability justifies the need for more effective, regularization (penalizing) approaches that can concurrently perform both model and feature selections. For researchers in other fields, the study can help them to unscrew the potential use of the proposed methods that various researchers were not able to explore.

1.6 Scope of the Study

This study concentrated on improving the process of feature selection, prediction accuracy, and handling of missing values through the use of alternative initial weights in adaptive LASSO and adaptive elastic net for high dimensional data. Simulation studies with varying numbers of predictor variables, sample sizes, correlation coefficients, and the proportion of missing values were performed to evaluate the effectiveness of the proposed methods. In addition, real-world data was used to assess the proposed penalizing methods. The major parts of the real dataset used are real-world datasets obtained from the medical discipline like gene expression microarray of different cancer types in which the number of genes is often much more than the sample size. Other datasets are from chemometrics when modeling "quantitative structure-activity relationship", where the number of molecular descriptors surpasses the number of compounds. Deep comparative studies were conducted to compare the proposed penalizing likelihood methods with other existing related methods. All of the simulation studies and real-world applications are implemented using the R programming language.

1.7 Limitations of the Study

There may be three possible limitations in this study. First, although algorithms of proposed methods implement well for logistic regression models, they need improvement in order to use for other regression models. In addition, proposed penalized methods cannot apply to imbalanced data refers to those types of datasets where the target class has an uneven distribution of observations, i.e., one class label has a very high number of observations, and the other has a very low number of observations. Furthermore, the present study concentrated on dealing with three different rates of missing values, namely 10%, 20%, and 30% in high-dimensional data. Therefore, the performance of the proposed imputation penalized method did not be investigated when the proportion of missing values is more than 30% in such data.

1.8 Organization of Thesis

Following this introductory chapter of the study. This thesis is organized as the following. Chapter 2 presents a review of the past literature on penalized likelihood approaches. The research methodology is covered in Chapter 3. It began by explaining the penalized linear regression and extended linear model methods. It also went over the statistical properties of the penalized approaches that were used. It was then followed by a detailed presentation of the proposed penalizing approaches and evaluation metrics used. In chapter 4, the performance of each proposed method is evaluated through simulation studies and real-world applications of the logistic regression models. The findings and discussion for the effectiveness of the proposed methods are also presented. Chapter 5 ends the thesis with a summary and future directions of study in this area.

REFERENCES

Adragni, K. P. (2015). Independent screening in high-dimensional exponential family predictors' space. *Journal of Applied Statistics*. 42(2), 347–359. ISSN 13600532. doi:10.1080/02664763.2014.949640.

Agrawal, P., Abutarboush, H. F., Ganesh, T. and Mohamed, A. W. (2021). Metaheuristic Algorithms on Feature Selection: A Survey of One Decade of Research (2009-2019). *IEEE Access*. 9, 26766–26791. ISSN 2169-3536. doi:10.1109/ACCESS.2021.3056407.

Al-batah, M. S. (2019). Ranked Features Selection with MSBRG Algorithm and Rules Classifiers for Cervical Cancer. *International Journal of Online and Biomedical Engineering (iJOE)*. 15(12), 4. ISSN 2626-8493. doi:10.3991/ijoe.v15i12.10803.

Al-Fakih, A., Algamal, Z., Lee, M., Aziz, M. and Ali, H. (2019). QSAR classification model for diverse series of antifungal agents based on improved binary differential search algorithm. *SAR and QSAR in Environmental Research*. 30(2), 131–143. ISSN 1062-936X. doi:10.1080/1062936X.2019.1568298.

Al-Fakih, A. M., Algamal, Z. Y., Lee, M. H. and Aziz, M. (2018). A penalized quantitative structure–property relationship study on melting point of energetic carbocyclic nitroaromatic compounds using adaptive bridge penalty. *SAR and QSAR in Environmental Research*. 29(5), 339–353. ISSN 1062-936X. doi:10.1080/1062936X.2018.1439531.

Al-Khateeb, S. F. M. (2019). Tuning paramater selection in penalized logistic regression with application in cancer. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*. 18(36), 11–22. ISSN 1305-7820. Retrievable at `https://dergipark.org.tr/en/pub/ticaretfbd/issue/55971/644095`.

Al-Thanoon, N. A., Qasim, O. S. and Algamal, Z. Y. (2018). Tuning parameter estimation in SCAD-support vector machine using firefly algorithm with application in gene selection and cancer classification. *Computers in Biology and Medicine*. 103(October), 262–268. ISSN 18790534. doi:10.1016/j.compbiomed.2018.10.034.

Algamal, Z. Y. (2017). Classification of gene expression autism data based on adaptive penalized logistic regression. *Electronic Journal of Applied Statistical Analysis*. 10(2), 561–571. ISSN 20705948. doi:10.1285/i20705948v10n2p561.

Algamal, Z. Y., Alhamzawi, R. and Mohammad Ali, H. T. (2018). Gene selection for microarray gene expression classification using Bayesian Lasso quantile regression. *Computers in Biology and Medicine*. 97(April), 145–152. ISSN 18790534. doi: 10.1016/j.compbiomed.2018.04.018.

Algamal, Z. Y. and Lee, M. H. (2015a). Adjusted Adaptive LASSO in High-dimensional Poisson Regression Model. *Modern Applied Science*. 9(4), 170–177. ISSN 1913-1852. doi:10.5539/mas.v9n4p170.

Algamal, Z. Y. and Lee, M. H. (2015b). Applying Penalized Binary Logistic Regression with Correlation Based Elastic Net for Variables Selection. *Journal of Modern Applied Statistical Methods*. 14(1), 168–179. ISSN 1538-9472. doi:10.22237/jmasm/1430453640.

Algamal, Z. Y. and Lee, M. H. (2015c). High Dimensional Logistic Regression Model using Adjusted Elastic Net Penalty. *Pakistan Journal of Statistics and Operation Research*. 11(4), 667. ISSN 2220-5810. doi:10.18187/pjsor.v11i4.990.

Algamal, Z. Y. and Lee, M. H. (2015d). Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification. *Expert Systems with Applications*. 42(23), 9326–9332. ISSN 09574174. doi:10.1016/j.eswa.2015.08.016.

Algamal, Z. Y. and Lee, M. H. (2015e). Penalized Poisson Regression Model using adaptive modified Elastic Net Penalty. *Electronic Journal of Applied Statistical Analysis*. 8(2), 236–245. ISSN 20705948. doi:10.1285/i20705948v8n2p236.

Algamal, Z. Y. and Lee, M. H. (2015f). Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification. *Computers in Biology and Medicine*. 67, 136–145. ISSN 00104825. doi:10.1016/j.compbiomed.2015.10.008.

Algamal, Z. Y. and Lee, M. H. (2017a). A new adaptive L_1-norm for optimal descriptor selection of high-dimensional QSAR classification model for anti-hepatitis C virus activity of thiourea derivatives. *SAR and QSAR in Environmental Research*. 28(1), 75–90. ISSN 1062-936X. doi:10.1080/1062936X.2017.1278618.

Algamal, Z. Y. and Lee, M. H. (2017b). A novel molecular descriptor selection method in QSAR classification model based on weighted penalized logistic regression. *Journal of Chemometrics*. 31(10), e2915. ISSN 08869383. doi:10.1002/cem.2915.

Algamal, Z. Y. and Lee, M. H. (2019). A two-stage sparse logistic regression for optimal gene selection in high-dimensional microarray data classification. *Advances in Data Analysis and Classification*. 13(3), 753–771. ISSN 1862-5347. doi:10.1007/s11634-018-0334-1.

Algamal, Z. Y., Lee, M. H., Al-Fakih, A. M. and Aziz, M. (2015). High-dimensional QSAR prediction of anticancer potency of imidazo[4,5-b]pyridine derivatives using adjusted adaptive LASSO. *Journal of Chemometrics*. 29(10), 547–556. ISSN 08869383. doi:10.1002/cem.2741.

Algamal, Z. Y., Lee, M. H., Al-Fakih, A. M. and Aziz, M. (2016). High-dimensional QSAR modelling using penalized linear regression model with $L_{1/2}$-norm. *SAR and QSAR in Environmental Research*. 27(9), 703–719. ISSN 1062-936X. doi:10.1080/1062936X.2016.1228696.

Algamal, Z. Y., Lee, M. H., Al-Fakih, A. M. and Aziz, M. (2017). High-dimensional QSAR classification model for anti-hepatitis C virus activity of thiourea derivatives based on the sparse logistic regression model with a bridge penalty. *Journal of Chemometrics*. 31(6), e2889. ISSN 08869383. doi:10.1002/cem.2889.

Algamal, Z. Y. and Mohammad Ali, H. T. (2017). An efficient gene selection method for high-dimensional microarray data based on sparse logistic regression. *Electronic Journal of Applied Statistical Analysis*. 10(1), 242–256. ISSN 20705948. doi:10.1285/i20705948v10n1p242.

Alon, U., Barka, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*. 96(12), 6745–6750. ISSN 00278424. doi:10.1073/pnas.96.12.6745.

Alquier, P. and Hebiri, M. (2011). Generalization of L_1 constraints for high dimensional regression problems. *Statistics and Probability Letters*. 81(12), 1760–1765. ISSN 01677152. doi:10.1016/j.spl.2011.07.011.

Androulakis, E., Koukouvinos, C. and Mylona, K. (2011). Tuning Parameter Estimation in Penalized Least Squares Methodology. *Communications in Statistics - Simulation and Computation*. 40(9), 1444–1457. ISSN 0361-0918. doi:10.1080/03610918.2011.575507.

Andu, Y., Lee, M. H. and Algamal, Z. Y. (2020). Variable selection of yearly high dimension stock market price using ordered homogenous pursuit lasso. In *AIP Conference Proceedings*, vol. 2266. AIP Publishing LLC. ISBN 0735420297, 090012. doi:10.1063/5.0019161.

Andu, Y., Lee, M. H. and Algamal, Z. Y. (2021). Adaptive Elastic Net with Distance Correlation on the Grouping Effect and Robust of High Dimensional Stock Market Price. *Sains Malaysiana*. 50(9), 2755–2764. ISSN 01266039. doi:10.17576/jsm-2021-5009-21.

Arisoy, M., Temiz-Arpaci, O., Yildiz, I., Kaynak-Onurdag, F., Aki, E., Yalcin, I. and Abbasoglu, U. (2008). Synthesis, antimicrobial activity and QSAR studies of 2, 5-disubstituted benzoxazoles. *SAR and QSAR in Environmental Research*. 19(5-6), 589–612. ISSN 1062-936X.

Asar, Y. and Genç, A. (2016). New shrinkage parameters for the Liu-type logistic estimators. *Communications in Statistics: Simulation and Computation*. 45(3), 1094–1103. ISSN 15324141. doi:10.1080/03610918.2014.995815.

Asuncion, A. and Newman, D. (2007). *UCI machine learning repository*.

Bahassine, S., Madani, A., Al-Sarem, M. and Kissi, M. (2020). Feature selection using an improved Chi-square for Arabic text classification. *Journal of King Saud University - Computer and Information Sciences*. 32(2), 225–231. ISSN 22131248. doi:10.1016/j.jksuci.2018.05.010.

Basu, A., Ghosh, A., Jaenada, M. and Pardo, L. (2021). Robust adaptive Lasso in high-dimensional logistic regression with an application to genomic classification of cancer patients. *arXiv preprint arXiv:2109.03028*.

Becker, N., Toedt, G., Lichter, P. and Benner, A. (2011). Elastic SCAD as a novel penalization method for SVM classification tasks in high-dimensional data. *BMC Bioinformatics*. 12(1), 138. ISSN 14712105. doi:10.1186/1471-2105-12-138.

Bertsimas, D. and Van Parys, B. (2020). Sparse high-dimensional regression: Exact scalable algorithms and phase transitions. *The Annals of Statistics*. 48(1), 300–323. ISSN 0090-5364. doi:10.1214/18-AOS1804.

Bielza, C., Robles, V. and Larrañaga, P. (2011a). Regularized logistic regression without a penalty term: An application to cancer classification with microarray data. *Expert Systems with Applications*. 38(5), 5110–5118. ISSN 09574174. doi: 10.1016/j.eswa.2010.09.140.

Bielza, C., Robles, V. and Larrañaga, P. (2011b). Regularized logistic regression without a penalty term: An application to cancer classification with microarray data. *Expert Systems with Applications*. 38(5), 5110–5118. ISSN 09574174. doi: 10.1016/j.eswa.2010.09.140.

Bouchlaghem, Y., Akhiat, Y. and Amjad, S. (2022). Feature Selection: A Review and Comparative Study. *E3S Web of Conferences*. 351, 01046. ISSN 2267-1242. doi:10.1051/e3sconf/202235101046.

Boulesteix, A.-L., De Bin, R., Jiang, X. and Fuchs, M. (2017). IPF-LASSO: Integrative L_1-Penalized Regression with Penalty Factors for Prediction Based on Multi-Omics Data. *Computational and Mathematical Methods in Medicine*. 2017, 7691937. ISSN 1748-670X. doi:10.1155/2017/7691937.

Bühlmann, P., Rütimann, P., van de Geer, S. and Zhang, C.-H. (2013). Correlated variables in regression: Clustering and sparse estimation. *Journal of Statistical Planning and Inference*. 143(11), 1835–1858. ISSN 03783758. doi:10.1016/j.jspi.2013.05.019.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Springer Series in Statistics. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-20191-2. doi:10.1007/978-3-642-20192-9.

Bunea, F. (2008a). Consistent selection via the Lasso for high dimensional approximating regression models. In *Pushing the Limits of Contemporary Statistics:*

Contributions in Honor of Jayanta K. Ghosh. (pp. 122–137). Institute of Mathematical Statistics. ISBN 0940600757. doi:10.1214/074921708000000101.

Bunea, F. (2008b). Honest variable selection in linear and logistic regression models via L_1 and $L_1 + L_2$ penalization. *Electronic Journal of Statistics*. 2(none), 1153–1194. ISSN 1935-7524. doi:10.1214/08-EJS287.

Caner, M. and Fan, Q. (2015). Hybrid generalized empirical likelihood estimators: Instrument selection with adaptive lasso. *Journal of Econometrics*. 187(1), 256–274. ISSN 18726895. doi:10.1016/j.jeconom.2015.01.007.

Caner, M. and Zhang, H. H. (2014). Adaptive elastic net for generalized methods of moments. *Journal of Business & Economic Statistics*. 32(1), 30–47. ISSN 0735-0015.

Casella, G., Fienberg, S., Olkin, I., James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An introduction to statistical learning*. vol. 112. Springer. ISBN 9780387781884.

Cateni, S., Colla, V. and Vannucci, M. (2022). Improving the Stability of the Variable Selection with Small Datasets in Classification and Regression Tasks. *Neural Processing Letters*. ISSN 1370-4621. doi:10.1007/s11063-022-10916-4.

Chatterjee, A., Gupta, S. and Lahiri, S. (2015). On the residual empirical process based on the ALASSO in high dimensions and its functional oracle property. *Journal of Econometrics*. 186(2), 317–324. ISSN 03044076. doi:10.1016/j.jeconom.2015.02.012.

Chen, J. and Chen, Z. (2012). Extended BIC for small-n-large-P sparse GLM. *Statistica Sinica*. 22(2), 555–574. ISSN 10170405. doi:10.5705/ss.2010.216.

Chen, Q. and Wang, S. (2013). Variable selection for multiply-imputed data with application to dioxin exposure study. *Statistics in Medicine*. 32(21), 3646–3659. ISSN 02776715. doi:10.1002/sim.5783.

Chen, S.-B., Zhang, Y.-M., Ding, C. H., Zhang, J. and Luo, B. (2019). Extended adaptive Lasso for multi-class and multi-label feature selection. *Knowledge-Based Systems*. 173, 28–36. ISSN 09507051. doi:10.1016/j.knosys.2019.02.021.

Chen, Y., Wang, A., Ding, H., Que, X., Li, Y., An, N. and Jiang, L. (2016). A global learning with local preservation method for microarray data imputation. *Computers* *in Biology and Medicine*. 77, 76–89. ISSN 00104825. doi:10.1016/j.compbiomed.2016.08.005.

De Mol, C., De Vito, E. and Rosasco, L. (2009). Elastic-net regularization in learning theory. *Journal of Complexity*. 25(2), 201–230. ISSN 0885064X. doi:10.1016/j.jco.2009.01.002.

Deng, Y., Chang, C., Ido, M. S. and Long, Q. (2016). Multiple Imputation for General Missing Data Patterns in the Presence of High-dimensional Data. *Scientific Reports*. 6(1), 21689. ISSN 2045-2322. doi:10.1038/srep21689.

Detwiller, K. Y., Fernando, N. T., Segal, N. H., Ryeom, S. W., D'Amore, P. A. and Yoon, S. S. (2005). Analysis of Hypoxia-Related Gene Expression in Sarcomas and Effect of Hypoxia on RNA Interference of Vascular Endothelial Cell Growth Factor A. *Cancer Research*. 65(13), 5881–5889. ISSN 0008-5472. doi:10.1158/0008-5472.CAN-04-4078.

Di Santo, R., Tafi, A., Costi, R., Botta, M., Artico, M., Corelli, F., Forte, M., Caporuscio, F., Angiolella, L. and Palamara, A. T. (2005). Antifungal agents. 11. N-substituted derivatives of 1-[(aryl) (4-aryl-1 H-pyrrol-3-yl) methyl]-1 H-imidazole: synthesis, anti-candida activity, and QSAR studies. *Journal of medicinal chemistry*. 48(16), 5140–5153. ISSN 0022-2623.

Doerken, S., Avalos, M., Lagarde, E. and Schumacher, M. (2019). Penalized logistic regression with low prevalence exposures beyond high dimensional settings. *PLOS ONE*. 14(5), e0217057. ISSN 1932-6203. doi:10.1371/journal.pone.0217057.

Dong, Y., Song, L., Wang, M. and Xu, Y. (2014). Combined-penalized likelihood estimations with a diverging number of parameters. *Journal of Applied Statistics*. 41(6), 1274–1285. ISSN 0266-4763. doi:10.1080/02664763.2013.868415.

Dudoit, S., Fridlyand, J. and Speed, T. P. (2002). Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *Journal of the American Statistical Association*. 97(457), 77–87. ISSN 0162-1459. doi:10.1198/016214502753479248.

Eid, H. F., Hassanien, A. E., Kim, T.-h. and Banerjee, S. (2013). Linear Correlation-Based Feature Selection for Network Intrusion Detection Model. In *Communications* *in Computer and Information Science*. (pp. 240–248). vol. 381 CCIS. Springer. ISBN 9783642405969. doi:10.1007/978-3-642-40597-6_21.

El Anbari, M. and Mkhadri, A. (2013). The Adaptive Gril Estimator with a Diverging Number of Parameters. *Communications in Statistics - Theory and Methods*. 42(14), 2634–2660. ISSN 0361-0926. doi:10.1080/03610926.2011.615438.

Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*. 96(456), 1348–1360. ISSN 0162-1459. doi:10.1198/016214501753382273.

Fan, J. and Lv, J. (2010). A Selective Overview of Variable Selection in High Dimensional Feature Space. *Statistica Sinica*. 20(1), 101–148. ISSN 1017-0405.

Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 75(3), 531–552. ISSN 13697412. doi:10.1111/rssb.12001.

Fang, K., Wang, X., Zhang, S., Zhu, J. and Ma, S. (2015). Bi-level variable selection via adaptive sparse group Lasso. *Journal of Statistical Computation and Simulation*. 85(13), 2750–2760. ISSN 0094-9655. doi:10.1080/00949655.2014.938241.

Ferris, M. C. and Mangasarian, O. L. (1995). Breast Cancer Diagnosis via Linear Programming. *IEEE Computational Science and Engineering*. 2(3), 70. ISSN 1070-9924. doi:10.1109/MCSE.1995.414885.

Filzmoser, P., Gschwandtner, M. and Todorov, V. (2012). Review of sparse methods in regression and classification with application to chemometrics. *Journal of Chemometrics*. 26(3-4), 42–51. ISSN 08869383. doi:10.1002/cem.1418.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*. 33(1), 1–22. doi:https://www.ncbi.nlm.nih.gov/pubmed/20808728.

Fu, G.-H. and Xu, Q.-S. (2012). Grouping Variable Selection by Weight Fused Elastic Net for Multi-Collinear Data. *Communications in Statistics - Simulation and Computation*. 41(2), 205–221. ISSN 0361-0918. doi:10.1080/03610918.2011.579369.

Fu, G.-H., Zhang, W.-M., Dai, L. and Fu, Y.-Z. (2014). Group Variable Selection with Oracle Property by Weight-Fused Adaptive Elastic Net Model for Strongly Correlated

Data. *Communications in Statistics - Simulation and Computation*. 43(10), 2468–2481. ISSN 0361-0918. doi:10.1080/03610918.2012.752841.

Georgiev, V. (2000). Membrane Transporters and Antifungal Drug Resistance. *Current Drug Targets*. 1(3), 261–284. ISSN 13894501. doi:10.2174/1389450003349209.

Georgopapadakou, N. H. (1998). Antifungals: mechanism of action and resistance, established and novel drugs. *Current Opinion in Microbiology*. 1(5), 547–557. ISSN 13695274. doi:10.1016/S1369-5274(98)80087-8.

Geronimi, J. and Saporta, G. (2017). Variable selection for multiply-imputed data with penalized generalized estimating equations. *Computational Statistics & Data Analysis*. 110, 103–114. ISSN 01679473. doi:10.1016/j.csda.2017.01.001.

Ghosh, S. (2011). On the grouped selection and model complexity of the adaptive elastic net. *Statistics and Computing*. 21(3), 451–462. ISSN 0960-3174. doi: 10.1007/s11222-010-9181-4.

Hamim, M., El Moudden, I., D Pant, M., Moutachaouik, H. and Hain, M. (2021). A Hybrid Gene Selection Strategy Based on Fisher and Ant Colony Optimization Algorithm for Breast Cancer Classification. *International Journal of Online and Biomedical Engineering (iJOE)*. 17(02), 148. ISSN 2626-8493. doi:10.3991/ijoe.v17i02.19889.

Hamraz, M., Khan, Z., Khan, D. M., Gul, N., Ali, A. and Aldahmani, S. (2022). Gene Selection in Binary Classification Problems Within Functional Genomics Experiments via Robust Fisher Score. *IEEE Access*. 10, 51682–51692. ISSN 2169-3536. doi:10.1109/ACCESS.2022.3172281.

Han, X., Fang, E. X. and Tang, C. Y. (2021). Pre-processing with Orthogonal Decompositions for High-dimensional Explanatory Variables. *arXiv preprint arXiv:2106.09071*.

Hasan, M. K., Alam, M. A., Roy, S., Dutta, A., Jawad, M. T. and Das, S. (2021). Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021). *Informatics in Medicine Unlocked*. 27, 100799. ISSN 23529148. doi:10.1016/j.imu.2021.100799.

Hashemi, S. H. B., Karimi, S. and Mahboobi, H. (2014). Lifestyle changes for prevention of breast cancer. *Electronic Physician*. 6(3), 894–905. ISSN 2008-5842. doi:10.14661/2014.894-905.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*. 12(1), 55–67. ISSN 0040-1706. doi: 10.1080/00401706.1970.10488634.

Holman, R. and Glas, C. A. W. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*. 58(1), 1–17. ISSN 00071102. doi:10.1111/j.2044-8317.2005.tb00312.x.

Honaker, J., King, G. and Blackwell, M. (2011). Amelia II: A Program for Missing Data. *Journal of Statistical Software*. 45(7), 1–47. ISSN 1548-7660. doi:10.18637/jss.v045.i07.

Huang, H., Gao, Y., Zhang, H. and Li, B. (2021). Weighted Lasso estimates for sparse logistic regression: non-asymptotic properties with measurement errors. *Acta Mathematica Scientia*. 41(1), 207–230. ISSN 0252-9602. doi:10.1007/s10473-021-0112-6.

Huang, H.-H. H., Liu, X.-Y. Y. and Liang, Y. (2016). Feature Selection and Cancer Classification via Sparse Logistic Regression with the Hybrid $L_{1/2+2}$ Regularization. *PLOS ONE*. 11(5), e0149675. ISSN 1932-6203. doi:10.1371/journal.pone.0149675.

Huang, J. and Fan, X. (2013). Reliably assessing prediction reliability for high dimensional QSAR data. *Molecular Diversity*. 17(1), 63–73. ISSN 1381-1991. doi:10.1007/s11030-012-9415-9.

Huang, J., Ma, S. and Zhang, C. H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica*. 18(4), 1603–1618. ISSN 10170405.

Hui, F. K. C., Warton, D. I. and Foster, S. D. (2015). Tuning Parameter Selection for the Adaptive Lasso Using ERIC. *Journal of the American Statistical Association*. 110(509), 262–269. ISSN 0162-1459. doi:10.1080/01621459.2014.951444.

Ijaz, M., Asghar, Z. and Gul, A. (2019). Ensemble of penalized logistic models for classification of high-dimensional data. *Communications in Statistics - Simulation* *and Computation*. 50(7), 2072–2088. ISSN 0361-0918. doi:10.1080/03610918.2019.1595647.

Jadhav, S., He, H. and Jenkins, K. (2018). Information gain directed genetic algorithm wrapper feature selection for credit rating. *Applied Soft Computing*. 69, 541–553. ISSN 15684946. doi:10.1016/j.asoc.2018.04.033.

Jia, J. and Rohe, K. (2015). Preconditioning the Lasso for sign consistency. *Electronic Journal of Statistics*. 9(1), 1150–1172. ISSN 1935-7524. doi:10.1214/15-EJS1029.

Jia, J. and Yu, B. (2010). On model selection consistency of the elastic net when $p \gg n$. *Statistica Sinica*. 20(2), 595–611. ISSN 10170405. doi:http://www.jstor.org/stable/24309012.

Jiang, W., Josse, J. and Lavielle, M. (2020). Logistic regression with missing covariates—Parameter estimation, model selection and prediction within a joint-modeling framework. *Computational Statistics & Data Analysis*. 145, 106907. ISSN 01679473. doi:10.1016/j.csda.2019.106907.

Jiang, Y., He, Y. and Zhang, H. (2016). Variable Selection With Prior Information for Generalized Linear Models via the Prior LASSO Method. *Journal of the American Statistical Association*. 111(513), 355–376. ISSN 0162-1459. doi:10.1080/01621459.2015.1008363.

Kadam, V. J., Jadhav, S. M. and Vijayakumar, K. (2019). Breast Cancer Diagnosis Using Feature Ensemble Learning Based on Stacked Sparse Autoencoders and Softmax Regression. *Journal of Medical Systems*. 43(8), 263. ISSN 0148-5598. doi:10.1007/s10916-019-1397-z.

Kahya, M. A. (2019). Classification enhancement of breast cancer histopathological image using penalized logistic regression. *Indonesian Journal of Electrical Engineering and Computer Science*. 13(1), 405. ISSN 2502-4760. doi:10.11591/ijeecs.v13.i1.pp405-410.

Kalivas, J. H., Héberger, K. and Andries, E. (2015). Sum of ranking differences (SRD) to ensemble multivariate calibration model merits for tuning parameter selection and comparing calibration methods. *Analytica Chimica Acta*. 869, 21–33. ISSN 18734324. doi:10.1016/j.aca.2014.12.056.

Kargi, I. A., Ismail, N. B. and Mohamad, I. B. (2021). Improved Lasso (ILASSO) for Gene Selection and Classification in High Dimensional DNA Microarray Data. *International Journal of Online and Biomedical Engineering (iJOE)*. 17(08), 91–102. ISSN 2626-8493. doi:10.3991/ijoe.v17i08.24601.

Khan, S. I. and Hoque, A. S. M. L. (2020). SICE: an improved missing data imputation technique. *Journal of Big Data*. 7(1), 37. ISSN 2196-1115. doi:10.1186/s40537-020-00313-w.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of statistics*. 28(5), 1356–1378. ISSN 0090-5364.

Kock, A. B. (2016). CONSISTENT AND CONSERVATIVE MODEL SELECTION WITH THE ADAPTIVE LASSO IN STATIONARY AND NONSTATIONARY AUTOREGRESSIONS. *Econometric Theory*. 32(1), 243–259. ISSN 0266-4666. doi:10.1017/S0266466615000304.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*. 97(1-2), 273–324. ISSN 00043702. doi:10.1016/S0004-3702(97)00043-X.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*. 13, 8–17. ISSN 20010370. doi:10.1016/j.csbj.2014.11.005.

Kwak, S. K. and Kim, J. H. (2017). Statistical data preparation: management of missing values and outliers. *Korean Journal of Anesthesiology*. 70(4), 407. ISSN 2005-6419. doi:10.4097/kjae.2017.70.4.407.

Latkowski, T. and Osowski, S. (2015a). Computerized system for recognition of autism on the basis of gene expression microarray data. *Computers in Biology and Medicine*. 56, 82–88. ISSN 00104825. doi:10.1016/j.compbiomed.2014.11.004.

Latkowski, T. and Osowski, S. (2015b). Data mining for feature selection in gene expression autism data. *Expert Systems with Applications*. 42(2), 864–872. ISSN 09574174. doi:10.1016/j.eswa.2014.08.043.

Leng, C., Lin, Y. and Wahba, G. (2006). A note on the Lasso and related procedures in model selection. *Statistica Sinica*. 16(4), 1273–1284. ISSN 10170405.

Li, J., Jia, Y. and Zhao, Z. (2013). Partly adaptive elastic net and its application to microarray classification. *Neural Computing and Applications*. 22(6), 1193–1200. ISSN 0941-0643. doi:10.1007/s00521-012-0885-6.

Li, N., Yang, H. and Yang, J. (2019). Nonnegative estimation and variable selection via adaptive elastic-net for high-dimensional data. *Communications in Statistics - Simulation and Computation*. 0(0), 1–17. ISSN 0361-0918. doi:10.1080/03610918.2019.1642484.

Li, X., Wang, Y. and Ruiz, R. (2020). A Survey on Sparse Learning Models for Feature Selection. *IEEE Transactions on Cybernetics*, 1–19. ISSN 2168-2267. doi:10.1109/TCYB.2020.2982445.

Lian, H. (2012). Variable selection in high-dimensional partly linear additive models. *Journal of Nonparametric Statistics*. 24(4), 825–839. ISSN 10485252. doi:10.1080/10485252.2012.701300.

Liang, F., Song, Q. and Yu, K. (2013a). Bayesian Subset Modeling for High-Dimensional Generalized Linear Models. *Journal of the American Statistical Association*. 108(502), 589–606. ISSN 0162-1459. doi:10.1080/01621459.2012.761942.

Liang, Y., Liu, C., Luan, X.-Z., Leung, K.-S., Chan, T.-M., Xu, Z.-B. and Zhang, H. (2013b). Sparse logistic regression with a $L_{1/2}$ penalty for gene selection in cancer classification. *BMC Bioinformatics*. 14(1), 198. ISSN 1471-2105. doi:10.1186/1471-2105-14-198.

Lin, Z., Xiang, Y. and Zhang, C. (2009). Adaptive Lasso in high-dimensional settings. *Journal of Nonparametric Statistics*. 21(6), 683–696. ISSN 10485252. doi:10.1080/10485250902984875.

Little, R. J. A. and Rubin, D. B. (2019). *Statistical analysis with missing data*. vol. 793. John Wiley & Sons. ISBN 0470526793.

Liu, C. and Wong, H. S. (2019). Structured Penalized Logistic Regression for Gene Selection in Gene Expression Data Analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 16(1), 312–321. ISSN 1545-5963. doi:10.1109/TCBB.2017.2767589.

Liu, X.-Y., Liang, Y., Wang, S., Yang, Z.-Y. and Ye, H.-S. (2018). A Hybrid Genetic Algorithm With Wrapper-Embedded Approaches for Feature Selection. *IEEE Access*. 6, 22863–22874. ISSN 2169-3536. doi:10.1109/ACCESS.2018.2818682.

Lu, W., Goldberg, Y. and Fine, J. P. (2012). On the robustness of the adaptive lasso to model misspecification. *Biometrika*. 99(3), 717–731. ISSN 0006-3444. doi: 10.1093/biomet/ass027.

Luo, S. and Chen, Z. (2014). Sequential Lasso cum EBIC for feature selection with ultra-high dimensional feature space. *Journal of the American Statistical Association*. 109(507), 1229–1240. ISSN 0162-1459.

Mahdi, G. J. M., Mohammed, N. J. and Al-Sharea, Z. I. (2021). Regression shrinkage and selection variables via an adaptive elastic net model. *Journal of Physics: Conference Series*. 1879(3), 032014. ISSN 1742-6588. doi: 10.1088/1742-6596/1879/3/032014.

Manhrawy, I. I., Qaraad, M. and El-Kafrawy, P. (2021). Hybrid feature selection model based on relief-based algorithms and regulizer algorithms for cancer classification. *Concurrency and Computation: Practice and Experience*. 33(17), 1–17. ISSN 1532-0626. doi:10.1002/cpe.6200.

Mao, G. (2015). Efficient Penalized Estimation for Linear Regression Model. *Communications in Statistics - Theory and Methods*. 44(7), 1436–1449. ISSN 0361-0926. doi:10.1080/03610926.2012.763094.

Medina Marrero, R., Marrero-Ponce, Y., Barigye, S. J., Echeverría Díaz, Y., Acevedo-Barrios, R., Casañola-Martín, G. M., García Bernal, M., Torrens, F. and Pérez-Giménez, F. (2015). QuBiLs-MAS method in early drug discovery and rational drug identification of antifungal agents. *SAR and QSAR in Environmental Research*. 26(11), 943–958. ISSN 1029046X. doi:10.1080/1062936X.2015.1104517.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*. 34(3), 1436–1462. ISSN 0090-5364. doi:10.1214/009053606000000281.

Mkhadri, A. and Ouhourane, M. (2015). A group VISA algorithm for variable selection. *Statistical Methods & Applications*. 24(1), 41–60. ISSN 1618-2510. doi:10.1007/s10260-014-0281-8.

Münch, M. M., Peeters, C. F. W., Van Der Vaart, A. W. and Van De Wiel, M. A. (2021). Adaptive group-regularized logistic elastic net regression. *Biostatistics*. 22(4), 723–737. ISSN 1465-4644. doi:10.1093/biostatistics/kxz062.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*. 135(3), 370. ISSN 00359238. doi:10.2307/2344614.

Pan, J. and Shang, J. (2018). Adaptive LASSO for linear mixed model selection via profile log-likelihood. *Communications in Statistics - Theory and Methods*. 47(8), 1882–1900. ISSN 0361-0926. doi:10.1080/03610926.2017.1332219.

Park, H., Sakaori, F. and Konishi, S. (2014). Robust sparse regression and tuning parameter selection via the efficient bootstrap information criteria. *Journal of Statistical Computation and Simulation*. 84(7), 1596–1607. ISSN 0094-9655. doi:10.1080/00949655.2012.755532.

Park, M. Y. and Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 69(4), 659–677. ISSN 1369-7412.

Pelckmans, K., De Brabanter, J., Suykens, J. and De Moor, B. (2005). Handling missing values in support vector machine classifiers. *Neural Networks*. 18(5-6), 684–692. ISSN 08936080. doi:10.1016/j.neunet.2005.06.025.

Peng, H., Fu, Y., Liu, J., Fang, X. and Jiang, C. (2013). Optimal gene subset selection using the modified SFFS algorithm for tumor classification. *Neural Computing and Applications*. 23(6), 1531–1538. ISSN 0941-0643. doi:10.1007/s00521-012-1148-2.

Potharaju, S. P. and Sreedevi, M. (2019). Distributed feature selection (DFS) strategy for microarray gene expression data to improve the classification performance. *Clinical Epidemiology and Global Health*. 7(2), 171–176. ISSN 22133984. doi:10.1016/j.cegh.2018.04.001.

Pötscher, B. M. and Schneider, U. (2009). On the distribution of the adaptive LASSO estimator. *Journal of Statistical Planning and Inference*. 139(8), 2775–2790. ISSN 03783758. doi:10.1016/j.jspi.2009.01.003.

Pourahmadi, M. (2013). *High-dimensional covariance estimation: with high-dimensional data*. vol. 882. John Wiley & Sons. ISBN 1118034295.

Qian, W. and Yang, Y. (2013). Model selection via standard error adjusted adaptive lasso. *Annals of the Institute of Statistical Mathematics*. 65(2), 295–318. ISSN 0020-3157. doi:10.1007/s10463-012-0370-0.

Qian, W., Yang, Y. and Zou, H. (2016). Tweedie's Compound Poisson Model With Grouped Elastic Net. *Journal of Computational and Graphical Statistics*. 25(2), 606–625. ISSN 1061-8600. doi:10.1080/10618600.2015.1005213.

Rencher, A. C. and Schaalje, G. B. (2008). *Linear models in statistics*. John Wiley & Sons. ISBN 0470192607.

Rhys, H. (2020). *Machine Learning with R, the tidyverse, and mlr*. Simon and Schuster. ISBN 1638350175.

Rish, I. and Grabarnik, G. (2014). *Sparse modeling: theory, algorithms, and applications*. CRC press. ISBN 1439828709.

Roberts, S. and Nowak, G. (2014). Stabilizing the lasso against cross-validation variability. *Computational Statistics & Data Analysis*. 70, 198–211. ISSN 01679473. doi:10.1016/j.csda.2013.09.008.

Rubin, D. B. (1996). Multiple Imputation after 18+ Years. *Journal of the American Statistical Association*. 91(434), 473–489. ISSN 0162-1459. doi:10.1080/01621459.1996.10476908.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*. vol. 81. John Wiley & Sons. ISBN 0471655740.

Ryan, M. M., Lockstone, H. E., Huffaker, S. J., Wayland, M. T., Webster, M. J. and Bahn, S. (2006). Gene expression analysis of bipolar disorder reveals downregulation of the ubiquitin cycle and alterations in synaptic genes. *Molecular Psychiatry*. 11(10), 965–978. ISSN 13594184. doi:10.1038/sj.mp.4001875.

Sampson, J. N., Chatterjee, N., Carroll, R. J. and Muller, S. (2013). Controlling the local false discovery rate in the adaptive Lasso. *Biostatistics*. 14(4), 653–666. ISSN 1465-4644. doi:10.1093/biostatistics/kxt008.

Shen, Q., Mei, Z. and Ye, B.-X. (2009). Simultaneous genes and training samples selection by modified particle swarm optimization for gene expression data classification. *Computers in Biology and Medicine*. 39(7), 646–649. ISSN 00104825. doi:10.1016/j.compbiomed.2009.04.008.

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S., Ray, T. S., Koval, M. A., Last, K. W., Norton, A., Lister, T. A., Mesirov, J., Neuberg, D. S., Lander, E. S., Aster, J. C. and Golub, T. R. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*. 8(1), 68–74. ISSN 1078-8956. doi:10.1038/nm0102-68.

Shortreed, S. M. and Ertefaie, A. (2017). Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*. 73(4), 1111–1122. ISSN 0006-341X. doi:10.1111/biom.12679.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R. and Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*. 1(2), 203–209. ISSN 15356108. doi:10.1016/S1535-6108(02)00030-2.

Singh, R. K. and M., D. S. K. (2019). Classification of Gene Expression Data using Efficient Feature Selection Technique and Resampling Method. *International Journal of Engineering and Advanced Technology*. 8(6), 406–414. ISSN 22498958. doi:10.35940/ijeat.E7816.088619.

Singla, R. K. and Bhat G, V. (2010). QSAR model for predicting the fungicidal action of 1,2,4-triazole derivatives against Candida albicans. *Journal of Enzyme Inhibition and Medicinal Chemistry*. 25(5), 696–701. ISSN 1475-6366. doi:10.3109/14756360903524296.

Sirimongkolkasem, T. and Drikvandi, R. (2019). On Regularisation Methods for Analysis of High Dimensional Data. *Annals of Data Science*. 6(4), 737–763. ISSN 2198-5804. doi:10.1007/s40745-019-00209-4.

Song, Q. and Liang, F. (2015). High-Dimensional Variable Selection With Reciprocal L_1-Regularization. *Journal of the American Statistical Association*. 110(512), 1607–1620. ISSN 0162-1459. doi:10.1080/01621459.2014.984812.

Su, Y.-S., Gelman, A. E., Hill, J. and Yajima, M. (2011). Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *Journal of Statistical Software*. 45(2), 1–31. doi:https://doi.org/10.7916/D8VQ3CD3.

Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics*. 17(1), 168–192. ISSN 2634-1964. doi:10.1016/j.aci.2018.08.003.

Tian, G.-L., Wang, M. and Song, L. (2014). Variable selection in the high-dimensional continuous generalized linear model with current status data. *Journal of Applied Statistics*. 41(3), 467–483. ISSN 0266-4763. doi:10.1080/02664763.2013.840271.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*. 58(1), 267–288. ISSN 00359246. doi:10.1111/j.2517-6161.1996.tb02080.x.

Tran, C. T., Zhang, M., Andreae, P. and Xue, B. (2016). A Wrapper Feature Selection Approach to Classification with Missing Data. In *European Conference on the Applications of Evolutionary Computation*. (pp. 685–700). Springer. doi:10.1007/978-3-319-31204-0_44.

Tran, Q. N. and Arabnia, H. R. (2016). *Emerging Trends in Applications and Infrastructures for Computational Biology, Bioinformatics, and Systems Biology*. Elsevier. ISBN 9780128042038. doi:10.1016/C2015-0-01779-8.

Tutz, G. and Ulbricht, J. (2009). Penalized regression with correlation-based penalty. *Statistics and Computing*. 19(3), 239–253. ISSN 0960-3174. doi:10.1007/s11222-008-9088-5.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice : Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*. 45(3), 1–67. ISSN 1548-7660. doi:10.18637/jss.v045.i03.

Vana, L., Hochreiter, R. and Hornik, K. (2016). Computing a journal meta-ranking using paired comparisons and adaptive lasso estimators. *Scientometrics*. 106(1), 229–251. ISSN 0138-9130. doi:10.1007/s11192-015-1772-6.

Venkatesh, B. and Anuradha, J. (2019). A Review of Feature Selection and Its Methods. *Cybernetics and Information Technologies*. 19(1), 3–26. ISSN 1314-4081. doi:10.2478/cait-2019-0001.

Vidaurre, D., Bielza, C. and Larrañaga, P. (2013). A Survey of L1 Regression. *International Statistical Review*. 81(3), 361–387. ISSN 03067734. doi:10.1111/insr.12023. Retrievable at `https://onlinelibrary.wiley.com/doi/10.1111/insr.12023`.

Waldmann, P., Ferenčaković, M., Mészáros, G., Khayatzadeh, N., Curik, I. and Sölkner, J. (2019). AUTALASSO: an automatic adaptive LASSO for genome-wide prediction. *BMC Bioinformatics*. 20(1), 167. ISSN 1471-2105. doi:10.1186/s12859-019-2743-3.

Walsh, T. J., Gonzalez, C., Roilides, E., Mueller, B. U., Ali, N., Lewis, L. L., Whitcomb, T. O., Marshall, D. J. and Pizzo, P. A. (1995). Fungemia in Children Infected with the Human Immunodeficiency Virus: New Epidemiologic Patterns, Emerging Pathogens, and Improved Outcome with Antifungal Therapy. *Clinical Infectious Diseases*. 20(4), 900–906. ISSN 1058-4838. doi:10.1093/clinids/20.4.900.

Wang, H., Bian, C., Kong, L., An, Y., Du, Y. and Tian, J. (2021). A Novel Adaptive Parameter Search Elastic Net Method for Fluorescent Molecular Tomography. *IEEE Transactions on Medical Imaging*. 40(5), 1484–1498. ISSN 0278-0062. doi:10.1109/TMI.2021.3057704.

Wang, L., You, Y. and Lian, H. (2015). Convergence and sparsity of Lasso and group Lasso in high-dimensional generalized linear models. *Statistical Papers*. 56(3), 819–828. ISSN 0932-5026. doi:10.1007/s00362-014-0609-3.

Wang, M., Song, L. and Wang, X. (2010). Bridge estimation for generalized linear models with a diverging number of parameters. *Statistics & Probability Letters*. 80(21-22), 1584–1596. ISSN 01677152. doi:10.1016/j.spl.2010.06.012.

Wang, M. and Wang, X. (2014). Adaptive Lasso estimators for ultrahigh dimensional generalized linear models. *Statistics & Probability Letters*. 89(1), 41–50. ISSN 01677152. doi:10.1016/j.spl.2014.02.015.

Wang, S., Nan, B., Rosset, S. and Zhu, J. (2011). Random lasso. *The Annals of Applied Statistics*. 5(1), 468–485. ISSN 1932-6157. doi:10.1214/10-AOAS377.

Wang, T. and Zhu, L. (2011). Consistent tuning parameter selection in high dimensional sparse linear regression. *Journal of Multivariate Analysis*. 102(7), 1141–1151. ISSN 0047259X. doi:10.1016/j.jmva.2011.03.007.

Wang, X. and Wang, M. (2016). Variable selection for high-dimensional generalized linear models with the weighted elastic-net procedure. *Journal of Applied Statistics*. 43(5), 796–809. ISSN 0266-4763. doi:10.1080/02664763.2015.1078300.

Wang, Y., Li, X. and Ruiz, R. (2019a). Weighted General Group Lasso for Gene Selection in Cancer Classification. *IEEE Transactions on Cybernetics*. 49(8), 2860–2873. ISSN 2168-2267. doi:10.1109/TCYB.2018.2829811.

Wang, Y., Yang, X.-G. and Lu, Y. (2019b). Informative gene selection for microarray classification via adaptive elastic net with conditional mutual information. *Applied Mathematical Modelling*. 71, 286–297. ISSN 0307904X. doi:10.1016/j.apm.2019.01.044.

Wang, Z., Ma, S., Zappitelli, M., Parikh, C., Wang, C.-Y. and Devarajan, P. (2016). Penalized count data regression with application to hospital stay after pediatric cardiac surgery. *Statistical Methods in Medical Research*. 25(6), 2685–2703. ISSN 0962-2802. doi:10.1177/0962280214530608.

Wolberg, W. H., Street, W. and Mangasarian, O. (1994). Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. *Cancer Letters*. 77(2-3), 163–171. ISSN 03043835. doi:10.1016/0304-3835(94)90099-X.

Xing, J.-J., Liu, Y.-F., Li, Y.-Q., Gong, H. and Zhou, Y.-P. (2014). QSAR classification model for diverse series of antimicrobial agents using classification tree configured by modified particle swarm optimization. *Chemometrics and Intelligent Laboratory Systems*. 137, 82–90. ISSN 0169-7439.

Xu, D., Zhang, Z. and Wu, L. (2014). Variable selection in high-dimensional double generalized linear models. *Statistical Papers*. 55(2), 327–347. ISSN 0932-5026. doi:10.1007/s00362-012-0481-y.

Yang, Z., Liang, Y., Zhang, H., Chai, H., Zhang, B. and Peng, C. (2018). Robust Sparse Logistic Regression With the L_q ($0 < q < 1$) Regularization for Feature Selection Using Gene Expression Data. *IEEE Access*. 6, 68586–68595. doi:10.1109/ACCESS.2018.2880198.

Zahid, F. M., Faisal, S. and Heumann, C. (2020). Variable selection techniques after multiple imputation in high-dimensional data. *Statistical Methods & Applications*. 29(3), 553–580. ISSN 1618-2510. doi:10.1007/s10260-019-00493-7.

Zahid, F. M., Faisal, S. and Heumann, C. (2021). Multiple imputation with compatibility for high-dimensional data. *PLOS ONE*. 16(7), e0254112. ISSN 1932-6203. doi: 10.1371/journal.pone.0254112.

Zahid, F. M. and Heumann, C. (2019). Multiple imputation with sequential penalized regression. *Statistical Methods in Medical Research*. 28(5), 1311–1327. ISSN 0962-2802. doi:10.1177/0962280218755574.

Zeng, L. and Xie, J. (2014). Group variable selection via SCAD-L_2. *Statistics*. 48(1), 49–66. ISSN 0233-1888. doi:10.1080/02331888.2012.719513.

Zeng, P., Wei, Y., Zhao, Y., Liu, J., Liu, L., Zhang, R., Gou, J., Huang, S. and Chen, F. (2014). Variable selection approach for zero-inflated count data via adaptive lasso. *Journal of Applied Statistics*. 41(4), 879–894. ISSN 0266-4763. doi:10.1080/02664763.2013.858672.

Zhang, C. and Xiang, Y. (2016). On the oracle property of adaptive group Lasso in high-dimensional linear models. *Statistical Papers*. 57(1), 249–265. ISSN 09325026. doi:10.1007/s00362-015-0684-0.

Zhang, Z. (2015). Missing values in big data research: some basic skills. *Annals of translational medicine*. 3(21), 323. doi:10.3978/j.issn.2305-5839.2015.12.11.

Zhao, Y. and Long, Q. (2016). Multiple imputation in the presence of high-dimensional data. *Statistical Methods in Medical Research*. 25(5), 2021–2035. ISSN 0962-2802. doi:10.1177/0962280213511027.

Zhong, Y., Chalise, P. and He, J. (2020). Nested cross-validation with ensemble feature selection and classification model for high-dimensional biological data. *Communications in Statistics - Simulation and Computation*. 0(0), 1–18. ISSN 0361-0918. doi:10.1080/03610918.2020.1850790.

Zhou, D.-X. (2013). On grouping effect of elastic net. *Statistics & Probability Letters*. 83(9), 2108–2112. ISSN 01677152. doi:10.1016/j.spl.2013.05.014.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*. 101(476), 1418–1429. ISSN 0162-1459. doi:10.1198/016214506000000735.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 67(2), 301–320. ISSN 1369-7412. doi:10.1111/j.1467-9868.2005.00503.x.

Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*. 37(4), 1733–1751. ISSN 0090-5364. doi:10.1214/08-AOS625.

LIST OF PUBLICATIONS

Journal with Impact Factor

- **Aiedh Mrisi Alharthi**, Muhammad Hisyam Lee, Zakariya Yahya Algamal, and A.M. Al-Fakih (2020). Quantitative structure-activity relationship model for classifying the diverse series of anti-fungal agents using ratio weighted penalized logistic regression, *SAR and QSAR in Environmental Research*, 31:8, 571-583, DOI: 10.1080/1062936X.2020.1782467. **(Q2, IF:2.432)**.

Indexed Journal (SCOPUS)

1. **Aiedh Mrisi Alharthi**, Muhammad Hisyam Lee, and Zakariya Yahya Algamal (2020). Weighted L_1-norm Logistic Regression for Gene Selection of Microarray Gene Expression Classification. *International Journal on Advanced Science, Engineering and Information Technology*, 10(4), 1483. https://doi.org/10.18517/ijaseit.10.4.10907.
2. **Aiedh Mrisi Alharthi**, Muhammad Hisyam Lee, and Zakariya Yahya Algamal (2021). Gene selection and classification of microarray gene expression data based on a new adaptive L_1-norm elastic net penalty. *Informatics in Medicine Unlocked*, 24(April), 100622. https://doi.org/10.1016/j.imu.2021.100622.
3. **Aiedh Mrisi Alharthi**, Muhammad Hisyam Lee, and Zakariya Yahya Algamal (2021). Improving the Diagnosis of Breast Cancer Using Regularized Logistic Regression with Adaptive Elastic Net. *Universal Journal of Public Health*, 9(5), 317-323. DOI: 10.13189/ujph.2021.090514.
4. **Aiedh Mrisi Alharthi**, Muhammad Hisyam Lee, and Zakariya Yahya Algamal (2022). Improving penalized logistic regression model with missing values in high-dimensional data. *International Journal of Online and Biomedical Engineering (iJOE)*. 18(02), pp. 40–54. ISSN 2626-8493. doi: https://doi.org/10.3991/ijoe.v18i02.25047.

Submitted Manuscript

- **Aiedh Mrisi Alharthi**, Muhammad Hisyam Lee and Zakariya Yahya Algamal (2020). Improving the diagnosis of breast cancer based on a new weighted L_1-norm logistic regression method. *Pakistan Journal of Statistics and Operation Research*. Indexed Journal **(Indexed by SCOPUS)**.

Conference Presentations

- **Aiedh Mrisi Alharthi**, Muhammad Hisyam Lee and Zakariya Yahya Algamal. 18^{th} – 19^{th} August 2020. Penalized logistic regression with adaptive Elastic net for Improving the Diagnosis of Breast Cancer. 8^{th} International Graduate Conference on Engineering, Science and Humanities 2020 (IGCESH 2020), Universiti Teknologi Malaysia (UTM). Website: https://sps.utm.my/igcesh2020/ **(Presenter)**