science, engineering and health studies

Research Article

Science, Engineering and Health Studies
https://li01.tci-thaijo.org/index.php/sehs
ISSN (Online): 2630-0087

Adaptive Lasso sparse logistic regression on high-dimensional data with multicollinearity

Narumol Sudjai[1], Monthira Duangsaphon[2], and Chandhanarat Chandhanayingyong[1]*

[1] Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok 10700, Thailand
[2] Faculty of Science and Technology, Thammasat University, Pathum Thani 12120, Thailand

***Corresponding author**:*
Chandhanarat
Chandhanayingyong
chandhanarat.c@gmail.com

***Received**: 24 July 2023*
***Revised**: 26 September 2024*
***Accepted**: 15 October 2024*
***Published**: 23 September 2025*

***Citation**:*
Sudjai, N., Duangsaphon, M., & Chandhanayingyong, C. (2025). Adaptive Lasso sparse logistic regression on high-dimensional data with multicollinearity. Science, Engineering and Health Studies, 19, 25020002.

ABSTRACT

A combination of high-dimensional sparse data and multicollinearity problems can lead to instabilities in a predictive model when applied to a new data set. The least absolute shrinkage and selection operator (Lasso) is widely employed in machine-learning algorithm for variable selection and parameter estimations. Although this method is computationally feasible for high-dimensional data, it has some drawbacks. Thus, the adaptive Lasso was developed using the adaptive weight on penalty function. This adaptive weight is related to the power order of the estimators. Hence, we focus on the power of adaptive weight on two penalty functions: adaptive Lasso and adaptive elastic net. This study aimed to compare the performances of the power of the adaptive Lasso and adaptive elastic net methods under high-dimensional sparse data with multicollinearity. Moreover, the performances of four penalized methods were compared: Lasso, elastic net, adaptive Lasso, and adaptive elastic net. They were compared using the mean of the predicted mean squared error for the simulation study and the classification accuracy for a real-data application. The results showed that the higher-order of the adaptive Lasso method performed best on very high-dimensional sparse data with multicollinearity when the initial weight was determined using a ridge estimator. However, in the case of high-dimensional sparse data with multicollinearity, the square root of the adaptive Lasso together with the initial weight using Lasso was the best option.

Keywords: high-dimensional data; machine learning; multicollinearity; penalized logistic regression; penalty function

1. INTRODUCTION

Advances in technology have resulted in computers being able to store vast amounts of data effectively. With such large volumes of data, tools are desired that can extract useful information. Particularly needed are predictive models that can provide accurate results to help decision-making. Logistic regression models are widely employed in data analysis (Makalic & Schmidt, 2010; Sudjai & Duangsaphon, 2020) and machine learning communities (Sudjai et al., 2023a, 2023b). In the case of a binary outcome variable, the classical method used to determine coefficients in the logistic regression model is maximum likelihood estimation (MLE). However, the MLE is only appropriate when the data is large enough and has no multicollinearity (Hosmer et al., 2013; Kleinbaum & Klein, 2010; Senaviratna & Cooray, 2021). One of the challenges in model building is high-dimensional data, which can lead to model overfitting (Brimacombe, 2014). Another challenge in model building is the presence of

https://doi.org/10.69598/sehs.19.25020002

multicollinearity (Belsley et al., 1980), which can inflate the variance of the MLEs in the logistic regression model (Hosmer et al., 2013; Kleinbaum & Klein, 2010). Consequently, the MLE used for coefficient estimation in logistic regression is unstable and inappropriate for building a classification model (Kastrin & Peterlin, 2010).

In order to remedy these two problems, the penalized method can be employed in the logistic regression model. This method can reduce variance in parameter estimation and help alleviate model overfitting (Hosseinnataj et al., 2019; Pavlou et al., 2016). Presently, the popular methods for penalty function are ridge regression, Lasso, and elastic net (Hoerl & Kennard, 1970; Tibshirani, 1996; Zou & Hastie, 2005). However, the performance of each method is not the same for each data item. Thus, several previous studies focused on developing an adaptive weight for the penalty function (Zou, 2006; Zou & Zhang, 2009).

However, no studies have compared the performances of penalized methods in logistic regression, focusing on the power of adaptive weight on the penalty function under high-dimensional sparse data with multicollinearity. Therefore, this study focused on the power of adaptive weight on the adaptive Lasso and adaptive elastic net methods. The aim was to compare the performances of the power of the adaptive Lasso and adaptive elastic net methods under high-dimensional sparse data with multicollinearity. Additionally, the performances of four penalized methods (Lasso, elastic net, adaptive Lasso, and adaptive elastic net) were compared on simulation study and a real-data application.

2. MATERIALS AND METHODS

Binary logistic regression was employed to evaluate the logistic regression coefficient, where a dependent variable (Y_i) is a dichotomous variable, i.e. 1 = positive class or 0 = negative class. This dependent variable has a Bernoulli distribution ($Y_i \sim Bernoulli(\pi_i)$). Hence, $y_i \in \{0,1\}$, is a $n \times 1$ vector where n is the sample size. X is a $n \times p$ data matrix of p independent variables and $\utilde{x}_i$ is a $1 \times p$ vector of the independent variables for the i^{th} row of X. Therefore, the binary logistic regression model is as in Equation 1.

$$\pi_i = \frac{exp\{\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j\}}{1 + exp\{\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j\}}, i = 1,2,3,\ldots,n \text{ and } j = 1,2,3,\ldots,p \qquad (1)$$

where π_i represents a probability that an observation is in a specified category of the dichotomous variable. $\pi_i = P(Y_i = 1|\utilde{x}_i)$ is the conditional probability that $y_i = 1$, given $\utilde{x}_i$. For $y_i = 0$, the conditional probability that $y_i = 0$, given $\utilde{x}_i$, can be presented as $1 - \pi_i = P(Y_i = 0|\utilde{x}_i)$.

Logistic regression is the logit transformation, which is given in Equation 2.

$$ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j \qquad (2)$$

where $\utilde{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)^T$ is a vector composed of logistic regression coefficients. β_0 is the intercept. β_j is a $p \times 1$ unknown coefficient vector. The left term in Equation 2 is the logit function.

The classical method used for parameter estimation in the model is the MLE, determined in Equation 3.

$$\hat{\utilde{\beta}}_{MLE} = arg \max_{\utilde{\beta}} \left(\sum_{i=1}^{n}[y_i\, ln(\pi_i) + (1 - y_i)\, ln(1 - \pi_i)]\right) \qquad (3)$$

Effect of multicollinearity on the MLE

When the number of independent variables far exceeds the sample size (referred to as high-dimensional data), the common phenomenon of multicollinearity arises (Belsley et al., 1980; Brimacombe, 2014). This leads to inflation in variances of the MLE. Consequently, the obtained estimators are unstable and cannot reflect the actual effects of the independent variables (Urgan & Tez, 2008).

Problem of model interpretation

In cases where $p \gg n$, the obtained predictive model may be complex, making it more difficult to interpret. A complex predictive model can be important for understanding complex processes. However, as the model becomes increasingly large and convoluted, serious problems can arise. The model may be overfitting (Brimacombe, 2014), and it may not be well identified in classification.

Therefore, variable selection procedures are crucial to alleviate the above problems. With the MLE, the stepwise method is widely used for the automatic selection of significant predictors. The Wald statistic is used for hypothesis testing in logistic regression. However, this testing cannot be used when high-dimensional data are used because the development of the Wald statistic is based on maximum likelihood estimators (Hosmer et al., 2013; Kleinbaum & Klein, 2010). Consequently, penalized logistic regression is used as an alternative method to the MLE.

Penalized logistic regression analysis

The aim of penalized logistic regression is to determine logistic regression coefficients when the data are highly correlated and highly dimensional. The penalized logistic regression coefficient is defined in Equation 4.

$$\hat{\utilde{\beta}}_{PLR} = arg \min_{\utilde{\beta}} \left(-\left\{\sum_{i=1}^{n}[y_i\, ln(\pi_i) + (1 - y_i)\, ln(1 - \pi_i)]\right\} + P_\lambda(\utilde{\beta})\right); \lambda \geq 0 \qquad (4)$$

where $P_\lambda(\utilde{\beta})$ is a penalty function term, and λ is the tuning parameter. In the case where λ equals zero, $\hat{\utilde{\beta}}_{PLR} = \hat{\utilde{\beta}}_{MLE}$. Regrading selecting λ, cross-validation is commonly used to evaluate the optimal value of this parameter.

2.1 Ridge regression

Ridge regression was originally designed to remedy the multicollinearity problem in a linear regression model. This penalty function method was proposed by Hoerl and Kennard (1970). Schaefer et al. (1984) proposed a modified ridge estimator, and it was subsequently applied by Le Cessie and Van Houwelingen (1992). The method of Le Cessie and Van Houwelingen significantly reduces the variance of $\hat{\utilde{\beta}}$. The ridge regression penalty (ℓ_2-norm penalty) is defined by Equation 5.

$$P_\lambda^{ridge}(\utilde{\beta}) = \lambda \sum_{j=1}^{p} \beta_j^2 \qquad (5)$$

Hence, the estimation of $\underset{\sim}{\beta}$ using the ridge regression penalty is defined in Equation 6.

$$\underset{\sim}{\hat{\beta}}_{ridge} = \arg\min_{\underset{\sim}{\beta}} \left(-\{\textstyle\sum_{i=1}^{n}[y_i \, ln(\pi_i) + (1-y_i)\, ln(1-\pi_i)]\} + P_{\lambda}^{ridge}(\underset{\sim}{\beta})\right) ; \lambda \geq 0 \qquad (6)$$

In the logistic regression model, the ridge regression penalty (also called the shrinkage penalty) penalizes the model by shrinking the coefficients toward zero. In the case of $\lambda = 0$, there is no shrinkage. With an increase in the value of λ ($\lambda \to \infty$), the magnitudes of the coefficients will tend to decrease but will not equal zero. This method has good performance when the data are high dimensional and the independent variables are collinear. However, an obvious disadvantage of this method is the lack of a variable selection property because it includes all independent variables in the final model. Therefore, the obtained model may be difficult to interpret when there is a large number of independent variables (James et al., 2013).

2.2 Lasso

Lasso was proposed by Tibshirani in 1996. This method avoids the disadvantage of ridge regression (the inability to reduce the number of independent variables in the final model). The concept of Lasso is similar to ridge regression in that the coefficient estimates are shrunk toward zero. The Lasso penalty (ℓ_1-norm penalty) is determined in Equation 7.

$$P_{\lambda}^{lasso}(\underset{\sim}{\beta}) = \lambda \textstyle\sum_{j=1}^{p}|\beta_j| \qquad (7)$$

Thus, the estimation of $\underset{\sim}{\beta}$ using the Lasso penalty is presented in Equation 8.

$$\underset{\sim}{\hat{\beta}}_{lasso} = \arg\min_{\underset{\sim}{\beta}} \left(-\{\textstyle\sum_{i=1}^{n}[y_i \, ln(\pi_i) + (1-y_i)\, ln(1-\pi_i)]\} + P_{\lambda}^{lasso}(\underset{\sim}{\beta})\right) \quad ; \lambda > 0 \qquad (8)$$

The tuning parameter λ controls the shrinkage of $\underset{\sim}{\hat{\beta}}$ by using cross-validation method (Efron et al., 2004; Hastie et al., 2009). When this tuning parameter is sufficiently large, the Lasso penalty has the effect of shrinking some coefficient estimates to exactly zero. This means that Lasso can perform variable selection. Consequently, the model obtained from Lasso is easier to interpret than that from ridge regression (James et al., 2013). Although Lasso is computationally feasible for high-dimensional data, it has some drawbacks. First, if $p \gg n$, Lasso selects at most n independent variables. Moreover, if there is a group of variables among which the pairwise correlations are very high, Lasso tends to select only one independent variable from the whole group and does not care which one is selected (Zou & Hastie, 2005). Finally, Lasso does not have oracle properties (Fan & Li, 2001; Zou, 2006).

2.3 Elastic net

Elastic net, proposed by Zou and Hastie in 2005, combines the properties of Lasso and ridge regression. The elastic net penalty includes the parts of the ℓ_1-norm and ℓ_2-norm penalties, which are defined in two steps. In the first step, the naive elastic net estimators are determined as in Equation 9.

$$\underset{\sim}{\hat{\beta}}_{Nelastic} = \arg\min_{\underset{\sim}{\beta}} \left(-\{\textstyle\sum_{i=1}^{n}[y_i \, ln(\pi_i) + (1-y_i)\, ln(1-\pi_i)]\} + \lambda_1 \sum_{j=1}^{p}|\beta_j| + \lambda_2 \sum_{j=1}^{p}\beta_j^2\right) \qquad (9)$$

where the penalty parameters (λ_1, λ_2) are more than or equal to zero. $\lambda = \lambda_1 + \lambda_2$ and $\alpha = \frac{\lambda_2}{(\lambda_1+\lambda_2)}$ when $\alpha \in [0,1)$.

The estimation of $\underset{\sim}{\beta}$ using the elastic net penalty is given in Equation 10.

$$\underset{\sim}{\hat{\beta}}_{elasticnet} = (1+\lambda_2)\underset{\sim}{\hat{\beta}}_{Nelastic} \qquad (10)$$

For λ_1 and λ_2, these parameters are used to control the shrinkage of $\underset{\sim}{\hat{\beta}}$ with cross-validation strategy (Hastie et al., 2009). Although the elastic net has superior performance to Lasso, it also lacks oracle properties (Zou & Zhang, 2009).

2.4 Adaptive Lasso

One important reason for Lasso may be instability due to the lack of oracle properties (Fan & Li, 2001). To overcome this disadvantage, Zou (2006) proposed the adaptive Lasso in 2006. The concept of the adaptive Lasso is a different weight for each parameter in the ℓ_1-norm penalty. The adaptive Lasso penalty is defined in Equation 11.

$$P_{\lambda}^{Alasso}(\underset{\sim}{\beta}) = \lambda \textstyle\sum_{j=1}^{p} w_j|\beta_j| \qquad (11)$$

Therefore, the estimation of $\underset{\sim}{\beta}$ using the adaptive Lasso penalty is given in Equation 12.

$$\underset{\sim}{\hat{\beta}}_{Alasso} = \arg\min_{\underset{\sim}{\beta}} \left(-\{\textstyle\sum_{i=1}^{n}[y_i \, ln(\pi_i) + (1-y_i)\, ln(1-\pi_i)]\} + P_{\lambda}^{Alasso}(\underset{\sim}{\beta})\right) \qquad (12)$$

where $w = (w_1, w_2, \ldots, w_p)^T$ is a vector composed of weight vector. $w_j = |\hat{\beta}_j|^{-\gamma}$; $\gamma > 0$ and γ is the power of the adaptive weight. It can be seen that w_j depends on the root n-consistent initial values of $\hat{\beta}_j$. The initial weight can be determined by using the MLE, ridge regression, or Lasso method (Pavlou et al., 2016; Zou, 2006). If $w_j = |(\hat{\beta}_{ridge})_j|^{-\gamma}$, $(\hat{\beta}_{ridge})_j$ is obtained from Equation 6. For $w_j = |(\hat{\beta}_{lasso})_j|^{-\gamma}$, $(\hat{\beta}_{lasso})_j$ is obtained from Equation 8. This weighted method is used to reduce the selection bias by assigning a smaller weight to large coefficients and a higher weight to small coefficients. Consequently, the adaptive Lasso can truly enjoy oracle properties (Zou, 2006). For λ and γ, these parameters are used as 2-dimensional cross-validation to tune the adaptive Lasso.

2.5 Adaptive elastic net

The adaptive elastic net method proposed by Zou and Zhang (2009) is a hybrid of adaptive Lasso and ridge regression. Consequently, it enjoys oracle properties and has superior performance to the elastic net method. The adaptive elastic net penalty is given in Equation 13.

$$P_{\lambda_1,\lambda_2}^{Aelastic}(\underset{\sim}{\beta}) = \lambda_1 \textstyle\sum_{j=1}^{p} \widehat{w}_j|\beta_j| + \lambda_2 \sum_{j=1}^{p}\beta_j^2 \qquad (13)$$

where $\widehat{w}_j = \left|(\hat{\beta}_{elasticnet})_j\right|^{-\gamma}$; $\gamma > 0$.

Thus, the estimation of $\underset{\sim}{\beta}$ using adaptive elastic net is presented in Equation 14.

$$\hat{\underset{\sim}{\beta}}_{Aelastic} = arg\min_{\underset{\sim}{\beta}} \left(-\{\textstyle\sum_{i=1}^{n}[y_i\, ln(\pi_i) + (1-y_i)\, ln(1-\pi_i)]\} + P_{\lambda_1,\lambda_2}^{Aelastic}(\underset{\sim}{\beta})\right) \tag{14}$$

The tuning parameter controls the shrinkage of $\hat{\underset{\sim}{\beta}}$ by using the Bayesian information criterion cross-validation method.

2.6 Monte Carlo simulation

The important factors affecting the accuracy of a predictive/classification model are the number of predictors (p), the sample size (n), and high correlation among predictors. In this simulation study, two conditions were considered:

1) High-dimensional sparse data (Cherkassky & Mulier, 2007). For $p > n$ and under the sparsity assumption on the true coefficients ($\underset{\sim}{\beta}$), that the number of significant predictors defined is equal to q, and given $q < p$. $\underset{\sim}{x}_i = (\underset{\sim}{x}_{iA}, \underset{\sim}{x}_{iB})$ with $\underset{\sim}{x}_{iA} = (x_{i1}, x_{i2}, x_{i3}, \ldots, x_{iq})^T \in \mathbb{R}^q$. Along with $\underset{\sim}{x}_{iB} = (x_{i(q+1)}, x_{i(q+2)}, x_{i(q+3)}, \ldots, x_{ip})^T \in \mathbb{R}^{p-q}$. Thus, $\underset{\sim}{X} = (\underset{\sim}{x}_A, \underset{\sim}{x}_B)^T \in \mathbb{R}^{n\times p}$ is the matrix of all independent variables where $\underset{\sim}{x}_A = (\underset{\sim}{x}_{iA}, \ldots, \underset{\sim}{x}_{nA})^T \in \mathbb{R}^{n\times q}$ and $\underset{\sim}{x}_B = (\underset{\sim}{x}_{iB}, \ldots, \underset{\sim}{x}_{nB})^T \in \mathbb{R}^{n\times(p-q)}$.

2) All independent variables are correlated by using the Toeplitz correlation structure, given in Equation 15 (Hardin et al., 2013).

$$\Sigma_k = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{k-1} \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{k-2} \\ \rho^2 & \rho & 1 & \rho & \cdots & \rho^{k-3} \\ \rho^3 & \rho^2 & \rho & 1 & \cdots & \rho^{k-4} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{k-1} & \rho^{k-2} & \rho^{k-3} & \rho^{k-4} & \cdots & 1 \end{pmatrix}_{k\times k} \tag{15}$$

where k denote the number of independent variables, which is a positive integer and $0 \le \rho \le 1$.

The Monte Carlo simulations were done using 50 and 1000 independent variables (p). The sample size (n) equaled 30 and 40. The dependent variables were generated from the Bernoulli distribution with parameter π_i. The independent variables were generated from the multivariate normal distribution with a mean of zero and covariance $\sum$ ($\underset{\sim}{X} \sim N(0, \sum)$). The degree of correlation (ρ) was set to 0.1, 0.3, 0.5, 0.75, 0.85, and 0.95. Interpretation of ρ is as follows: negligible correlation ($0.00 < \rho < 0.30$); low positive correlation ($0.30 \le \rho < 0.50$); moderate positive correlation ($0.50 \le \rho < 0.70$); high positive correlation ($0.70 \le \rho < 0.90$); and very high positive correlation ($0.90 \le \rho < 1.00$) (Mukaka, 2012). The number of significant predictors (q) equaled 15. The logistic regression coefficients were set the constant values as $\underset{\sim}{\beta}$. After that, the data was split into two subsets (the learning data set, 80% and the testing data set, 20%). The simulation study compared the performances of the four penalized methods (Lasso, elastic net, adaptive Lasso, and adaptive elastic net) using the predicted mean square errors (PMSE). The estimated PMSE was evaluated as per Equation 16.

$$\mathrm{PMSE} = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n} \tag{16}$$

where y_i and $\hat{y}_i$ were the i^{th} actual and predicted values of the dependent variables, respectively. The optimal value of the tuning parameter (λ) was found using a 10-fold cross-validation strategy (Hastie et al., 2009; Pavlou et al., 2016; Zou & Hastie, 2005). The experiment was repeated 1000 times to obtain a stationary result. Thus, the MPMSE was evaluated from the average of 1,000 estimates of PMSE$_j$ using Equation 17.

$$\mathrm{MPMSE} = \frac{1}{1000}\sum_{j=1}^{1000} \mathrm{PMSE_j}. \tag{17}$$

The penalized methods providing the lowest MPMSE were considered the best option. The flowchart of the simulation procedure is shown in Figure 1.

For the real-data application, the workflow diagram of the machine-learning procedure with the four penalized methods is shown in Figure 2. The classification accuracy of each method was determined as per Equation 18.

$$\text{Accuracy (\%)} = \frac{\mathrm{TP + TN}}{\mathrm{TP + FP + FN + TN}} \times 100. \tag{18}$$

where the true positive (TP) / true negative (TN) presents that the prediction is correct. False positive (FP) represents that the prediction is wrong (also called a type I error). A false negative (FN) shows that the prediction is wrong (also called a type II error).

3. RESULTS AND DISCUSSION

3.1 Simulation study

Table 1 lists the MPMSE values for the four methods for different ρ when p = 50 and 1000, and n = 30 and 40. The MPMSE values of the all methods increased when ρ was increased while holding p and n fixed. For an increase in n, the MPMSE values of all methods decreased. The MPMSE values of the adaptive Lasso method were less than those of the Lasso, elastic net, and adaptive elastic net methods.

With high-dimensional sparse data (p = 50, n = 30, and n = 40, and for different ρ), the performance of the adaptive Lasso method depended on the power of the adaptive weight (γ) and the initial weight used. In the case of ρ = 0.1, 0.3, and 0.5, the MPMSE values of the adaptive Lasso method with $w_j = \left|(\hat{\beta}_{ridge})_j\right|^{-1}$ were less than those for the other methods. For ρ = 0.75, 0.85, and 0.95, the inflation of the MPMSE values for $w_j = \left|(\hat{\beta}_{lasso})_j\right|^{-0.5}$ was the smallest compared with the other methods.

In the case of very high-dimensional sparse data (p = 1000, n = 30, n = 40, and for different ρ), the smallest MPMSE values were obtained from the adaptive Lasso method using $w_j = \left|(\hat{\beta}_{ridge})_j\right|^{-1}$ for ρ = 0.1, 0.3, and 0.5. For ρ = 0.75, 0.85, and 0.95, the MPMSE values for the higher-order of the adaptive Lasso method together with the initial weight using the ridge estimator were less than those for the other methods.

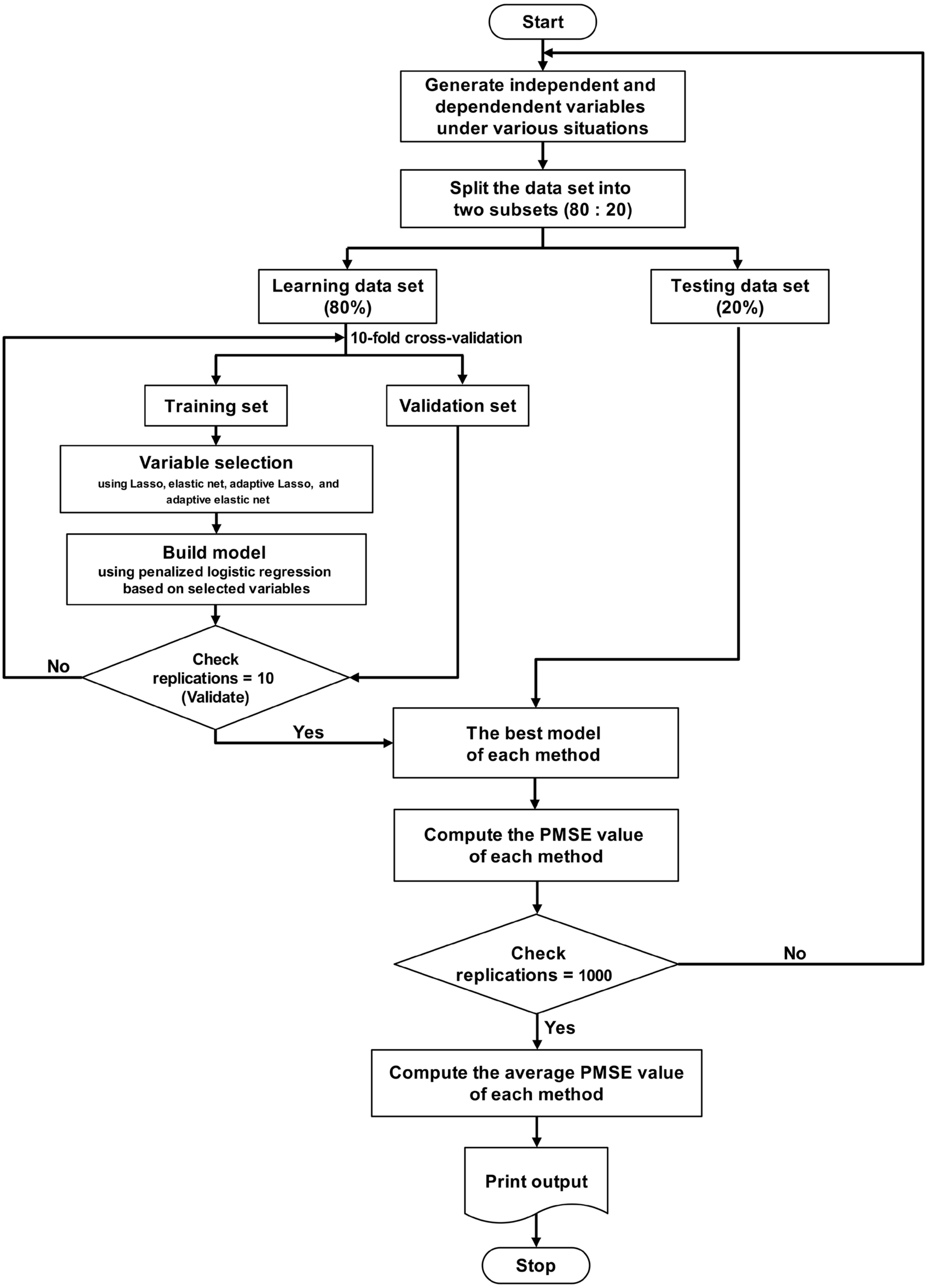
Start
Generate independent and dependendent variables under various situations
Split the data set into two subsets (80 : 20)
Learning data set (80%)
Testing data set (20%)
10-fold cross-validation
Training set
Validation set
Variable selection
using Lasso, elastic net, adaptive Lasso, and adaptive elastic net
Build model
using penalized logistic regression based on selected variables
Check replications = 10 (Validate)
No
Yes
The best model of each method
Compute the PMSE value of each method
Check replications = 1000
No
Yes
Compute the average PMSE value of each method
Print output
Stop

Figure 1. Flowchart of the simulation procedure

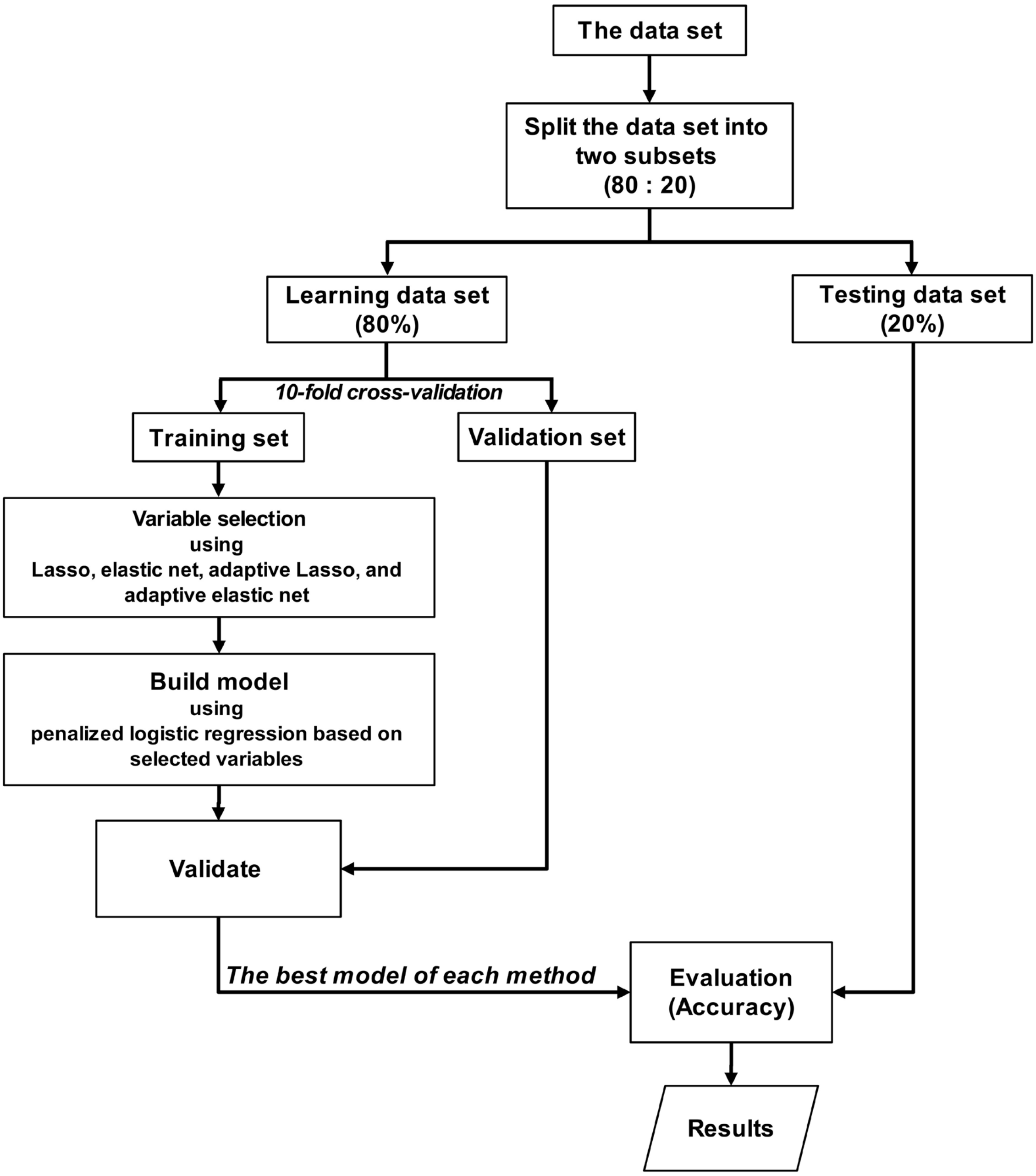
The data set
Split the data set into two subsets (80 : 20)
Learning data set (80%)
Testing data set (20%)
10-fold cross-validation
Training set
Validation set
Variable selection using Lasso, elastic net, adaptive Lasso, and adaptive elastic net
Build model using penalized logistic regression based on selected variables
Validate
The best model of each method
Evaluation (Accuracy)
Results

Figure 2. Workflow diagram of the machine-learning procedure

From the simulated results in Table 1, Figure 4a, and Figure 4c, it can be seen that the important factors influencing the MPMSE values were the correlated independent variables (i.e., ρ), the power of adaptive weight on penalty function, the initial weight, and the sample size. An increase in the correlation coefficient level leads to an increase in the MPMSE values for all methods when p and n are held constant. The worst case was obtained when the correlation coefficient level was very high (ρ = 0.95). In the case of the power of the adaptive weight on the ℓ_1-norm penalty, choosing the higher-order of the adaptive Lasso method using the initial weight with the ridge estimator produced the smallest MPMSE value for very high-dimensional sparse data with multicollinearity. However, for high-dimensional sparse data with multicollinearity, $w_j = \left|\left(\hat{\beta}_{lasso}\right)_j\right|^{-0.5}$ is preferred. We can see that an increase in the sample size (n) decreases the MPMSE values for all methods, while holding ρ and p fixed.

3.2 Real-data applications

In this section, the performances of 4 penalized methods were compared on two real-data sets with high-dimensional sparse data with multicollinearity.

First, a soft-tissue tumor data set was obtained from 40 patients (20 intramuscular lipomas and 20 well-differentiated liposarcomas). Between 2010 and 2020, the patients were diagnosed using their final pathological findings, and underwent preoperative magnetic resonance imaging (MRI) scans and total excision surgery. For our case study, the binary outcome of interest was an intramuscular lipoma or a well-differentiated liposarcoma.

The predictors of interest were 50 texture features as quantitative data that were extracted from preoperative T1-weighted MRI (Supplementary: Table S1). These features explain the spatial arrangement of gray-level pixels in a neighborhood on the MRI images such as fineness, coarseness, homogeneity, and heterogeneity. Regarding the detail of definition and formula for these texture features, they can be described according to PyRadiomics' documentation (Supplementary: Table S2).

Another data set (i.e., leukemia data set) from Golub et al. (1999) with gene expression monitoring data (via DNA microarray) was used to classify patients with acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). This leukemia data set was obtained from 72 patients. The data set contains measurements corresponding to ALL and AML specimens from bone marrow and peripheral blood, and it involves 7129 leukemia genes. For our case study, 40 patient samples were randomly selected from the data set. The outcome of interest was AML/ALL as a dichotomous variable. The predictors of interest comprised 1000 leukemia genes as continuous variables, which represented a subset of the 7129 genes.

Table 1. MPMSE values for different penalized methods

p	n	ρ	**Lasso**	**Elastic net**	**Adaptive Lasso**						**Adaptive elastic net**		
					$w_j = \left\|(\hat{\beta}_{ridge})_j\right\|^{-\gamma}$			$w_j = \left\|(\hat{\beta}_{lasso})_j\right\|^{-\gamma}$			$w_j = \left\|(\hat{\beta}_{elasticnet})_j\right\|^{-\gamma}$		
					$\gamma = 0.5$	$\gamma = 1$	$\gamma = 2$	$\gamma = 0.5$	$\gamma = 1$	$\gamma = 2$	$\gamma = 0.5$	$\gamma = 1$	$\gamma = 2$
50	30	0.10	0.1767	0.1797	0.1656	0.1618*	0.1627	0.1665	0.1691	0.1740	0.1682	0.1706	0.1746
		0.30	0.1772	0.1812	0.1679	0.1657*	0.1658	0.1679	0.1706	0.1742	0.1710	0.1735	0.1773
		0.50	0.1833	0.1878	0.1728	0.1708*	0.1721	0.1738	0.1767	0.1816	0.1754	0.1777	0.1822
		0.75	0.1888	0.1918	0.1851	0.1842	0.1852	0.1828*	0.1841	0.1870	0.1859	0.1875	0.1910
		0.85	0.1924	0.1980	0.1909	0.1907	0.1913	0.1880*	0.1894	0.1927	0.1909	0.1930	0.1961
		0.95	0.1957	0.1984	0.1951	0.1951	0.1978	0.1918*	0.1933	0.1960	0.1946	0.1959	0.1977
	40	0.10	0.1665	0.1710	0.1589	0.1565*	0.1573	0.1572	0.1606	0.1646	0.1579	0.1609	0.1663
		0.30	0.1741	0.1743	0.1636	0.1623*	0.1639	0.1648	0.1668	0.1715	0.1636	0.1672	0.1718
		0.50	0.1747	0.1759	0.1667	0.1666*	0.1675	0.1675	0.1702	0.1742	0.1671	0.1698	0.1743
		0.75	0.1845	0.1883	0.1821	0.1824	0.1830	0.1778*	0.1795	0.1841	0.1802	0.1825	0.1857
		0.85	0.1870	0.1908	0.1887	0.1893	0.1917	0.1825*	0.1857	0.1892	0.1855	0.1882	0.1925
		0.95	0.1943	0.1953	0.1973	0.1985	0.2030	0.1896*	0.1917	0.1948	0.1930	0.1938	0.1969
1000	30	0.10	0.1807	0.1837	0.1555	0.1522*	0.1546	0.1689	0.1708	0.1742	0.1744	0.1761	0.1776
		0.30	0.1845	0.1871	0.1595	0.1574*	0.1579	0.1734	0.1752	0.1775	0.1764	0.1774	0.1794
		0.50	0.1851	0.1899	0.1636	0.1594*	0.1601	0.1760	0.1767	0.1787	0.1793	0.1809	0.1833
		0.75	0.1890	0.1902	0.1680	0.1625	0.1620*	0.1779	0.1799	0.1817	0.1811	0.1831	0.1848
		0.85	0.1898	0.1937	0.1685	0.1660	0.1656*	0.1793	0.1811	0.1839	0.1840	0.1858	0.1877
		0.95	0.1929	0.1956	0.1798	0.1753	0.1749*	0.1835	0.1861	0.1876	0.1861	0.1879	0.1902
	40	0.10	0.1679	0.1718	0.1447	0.1399*	0.1421	0.1548	0.1566	0.1594	0.1561	0.1583	0.1617
		0.30	0.1780	0.1837	0.1556	0.1519*	0.1524	0.1645	0.1664	0.1699	0.1665	0.1697	0.1723
		0.50	0.1815	0.1858	0.1598	0.1545*	0.1586	0.1722	0.1735	0.1766	0.1727	0.1743	0.1805
		0.75	0.1851	0.1885	0.1649	0.1596	0.1590*	0.1752	0.1765	0.1793	0.1759	0.1777	0.1809
		0.85	0.1893	0.1921	0.1678	0.1642	0.1637*	0.1770	0.1790	0.1824	0.1786	0.1810	0.1844
		0.95	0.1913	0.1930	0.1762	0.1737	0.1733*	0.1813	0.1834	0.1868	0.1820	0.1851	0.1879

Note: * The penalized methods providing the lowest MPMSE

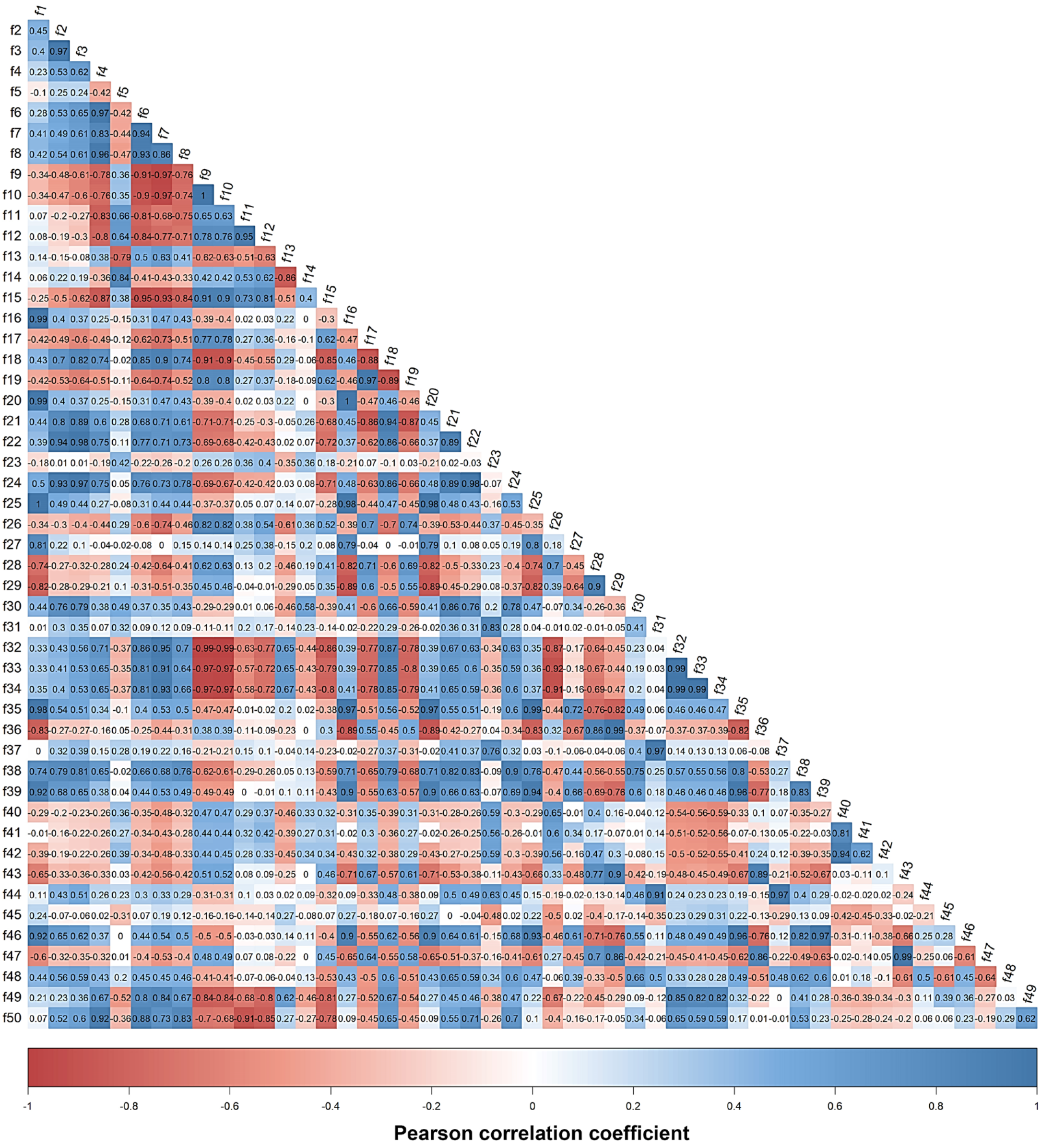

Figure 3. Correlation matrix of 50 texture features in 40 patients

Regarding Figure 3, the correlation matrix presents different Pearson correlation coefficient values and shades. The light shades denote that the predictors have a low correlation, whereas the dark shades present a high correlation among predictors. It is apparent that the multicollinearity problem occurred in this sample data set.

Table 2, shows the classification performances of the four penalized methods in distinguishing between intramuscular lipomas and well-differentiated liposarcomas. The highest accuracy values were obtained from the adaptive Lasso method with $w_j = \left|(\hat{\beta}_{lasso})_j\right|^{-0.5}$, while the lowest accuracy values were obtained from the elastic net method. For Table 3, the higher-order of the adaptive Lasso method together with the initial weight using the ridge estimator showed the best performance for differentiating between AML and ALL.

Table 2. Accuracy of machine-learning algorithms for distinguishing between intramuscular lipomas and well-differentiated liposarcomas for 50 texture features in 40 patients

	Lasso	Elastic net	Adaptive Lasso						Adaptive elastic net		
			$w_j = \left|(\hat{\beta}_{ridge})_j\right|^{-\gamma}$			$w_j = \left|(\hat{\beta}_{lasso})_j\right|^{-\gamma}$			$w_j = \left|(\hat{\beta}_{elasticnet})_j\right|^{-\gamma}$		
			$\gamma = 0.5$	$\gamma = 1$	$\gamma = 2$	$\gamma = 0.5$	$\gamma = 1$	$\gamma = 2$	$\gamma = 0.5$	$\gamma = 1$	$\gamma = 2$
Accuracy (%)	66.7	58.3	66.7	62.7	40.5	87.5*	71.0	58.3	70.8	70.7	54.2

Note: * The penalized methods providing the highest accuracy

Table 3. Accuracy of machine-learning algorithms for differentiating between acute myeloid leukemia and acute lymphoblastic leukemia for 1000 leukemia genes in 40 patients

	Lasso	Elastic net	Adaptive Lasso						Adaptive elastic net		
			$w_j = \left\|(\hat{\beta}_{ridge})_j\right\|^{-\gamma}$			$w_j = \left\|(\hat{\beta}_{lasso})_j\right\|^{-\gamma}$			$w_j = \left\|(\hat{\beta}_{elasticnet})_j\right\|^{-\gamma}$		
			$\gamma = 0.5$	$\gamma = 1$	$\gamma = 2$	$\gamma = 0.5$	$\gamma = 1$	$\gamma = 2$	$\gamma = 0.5$	$\gamma = 1$	$\gamma = 2$
Accuracy (%)	85.0	82.5	87.0	88.0	90.0*	85.0	77.5	73.0	85.0	83.0	80.0

Note: * The penalized methods providing the highest accuracy

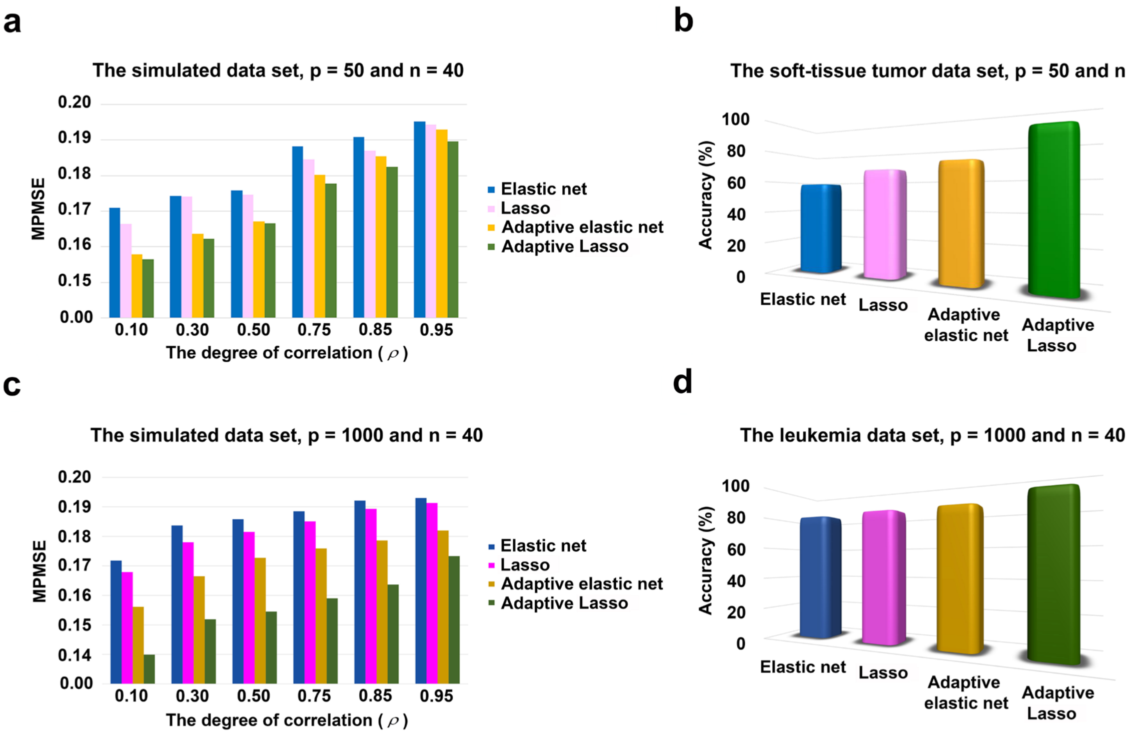

Figure 4. The mean of the predicted mean square errors (mpmse) values and accuracies of the four penalized methods

From the results of the real-data applications in Tables 2 and 3 as well as Figure 4b and Figure 4d, it is obvious that the adaptive Lasso method showed a better performance than the other methods for classification on the high-dimensional sparse data with multicollinearity. This finding corresponds to the results of the simulation study.

4. CONCLUSION

We propose the use of the adaptive Lasso method for classification in binary outcome on high-dimensional sparse data with multicollinearity as follows:

1) In the case of high-dimensional sparse data, it should be considered using the first power of the method and ridge estimator as the initial weight when there is a low or moderate correlation between the independent variables. On the other hand, if the independent variables are highly correlated, the initial weight should be evaluated using the Lasso estimator and γ = 0.5.

2) For very high-dimensional sparse data with multicollinearity, the higher-order of the adaptive Lasso method should be employed, using ridge estimator as the initial weight.

From the simulation study and the real-data applications, it is clear that the predictive performance of the penalized logistic regression model depends on the penalty function. In practice, if the penalty function method is appropriate, it constructs models that have good performance. Table 4 compares the appropriateness and limitations of each penalized method in the penalized logistic regression model.

ACKNOWLEDGMENTS

The authors thank Supani Duangkaew for her research coordination and Pakorn Yodprom for physics consultation and magnetic resonance image retrieval. Additionally, the authors gratefully acknowledge the financial support provided by the Siriraj Foundation Fund for advanced sarcoma research, grant number D004146.

Table 4. Comparison the appropriateness and limitations of each penalized method

Method	Appropriateness of application	Limitation
Ridge	- All independent variables relate to the dependent variable. - Multicollinearity exists.	- Lacks selection of variables
Lasso	- The independent variables show a low to moderate level of collinearity.	- When n is greater than p and the independent variables exhibit high collinearity, Lasso is dominated by ridge regression. - If the number of variables p is much greater than n ($p \gg n$), Lasso will only select up to n variables before reaching saturation. - When independent variables in the data set have a high pairwise correlation, Lasso chooses only one or a few variables from a group of correlated ones, without considering which one is chosen. - Lacks oracle properties
Elastic net	- Multicollinearity is present.	- Lacks oracle properties
Adaptive Lasso	- The independent variables exhibit a high level of correlation.	
Adaptive elastic net	- The independent variables are highly correlated.	

REFERENCES

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons.

Brimacombe, M. (2014). High-dimensional data and linear models: A review. *Open Access Medical Statistics*, *4*, 17–27. https://doi.org/10.2147/OAMS.S56499

Cherkassky, V., & Mulier, F. (2007). *Learning from data: Concepts, theory, and methods* (2nd ed.). John Wiley & Sons.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, *32*(2), 407–499. https://doi.org/10.1214/009053604000000067

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360. https://doi.org/10.1198/016214501753382273

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., & Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, *286*(5439), 531–537. https://doi.org/10.1126/science.286.5439.531

Hardin, J., Garcia, S. R., & Golan, D. (2013). A method for generating realistic correlation matrices. *The Annals of Applied Statistics*, *7*(3), 1733–1762. https://doi.org/10.1214/13-AOAS638

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Hosmer, D. W., Jr., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). John Wiley & Sons.

Hosseinnataj, A., Bahrampour, A., Baneshi, M., Zolala, F., Nikbakht, R., Torabi, M., & Mazidi Sharaf Abadi, F. (2019). Penalized Lasso methods in health data: Application to trauma and influenza data of Kerman. *Journal of Kerman University of Medical Sciences*, *26*(6), 440–449. https://doi.org/10.22062/jkmu.2019.89573

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. Springer.

Kastrin, A., & Peterlin, B. (2010). Rasch-based high-dimensionality data reduction and class prediction with applications to microarray gene expression data. *Expert Systems with Applications*, *37*(7), 5178–5185. https://doi.org/10.1016/j.eswa.2009.12.074

Kleinbaum, D. G., & Klein, M. (2010). *Logistic regression: A self-learning text* (3rd ed.). Springer.

Le Cessie, S., & Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *41*(1), 191–201. https://doi.org/10.2307/2347628

Makalic, E., & Schmidt, D. F. (2010). Review of modern logistic regression methods with application to small and medium sample size problems. In J. Li (Ed.), *AI 2010: Advances in artificial intelligence* (pp. 213–222). Springer. https://doi.org/10.1007/978-3-642-17432-2_22

Mukaka, M. M. (2012). Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, *24*(3), 69–71. https://pubmed.ncbi.nlm.nih.gov/23638278/

Pavlou, M., Ambler, G., Seaman, S., De Iorio, M., & Omar, R. Z. (2016). Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Statistics in Medicine*, *35*(7), 1159–1177. https://doi.org/10.1002/sim.6782

Schaefer, R. L., Roi, L. D., & Wolfe, R. A. (1984). A ridge logistic estimator. *Communications in Statistics - Theory and Methods*, *13*(1), 99–113. https://doi.org/10.1080/03610928408828664

Senaviratna, N. A. M. R., & Cooray, T. M. J. A. (2021). Multicollinearity in binary logistic regression model. *Theory and Practice of Mathematics and Computer Science*, *8*, 11–19.

Sudjai, N., & Duangsaphon, M. (2020). Liu-type logistic regression coefficient estimation with multicollinearity using the bootstrapping method. *Science, Engineering and Health Studies*, *14*(3), 203–214. https://doi.org/10.14456/sehs.2020.19

Sudjai, N., Siriwanarangsun, P., Lektrakul, N., Saiviroonporn, P., Maungsomboon, S., Phimolsarnti, R., Asavamongkolkul, A., & Chandhanayingyong, C. (2023a). Robustness of radiomic features: Two-dimensional versus three-dimensional MRI-based feature reproducibility in lipomatous soft-tissue tumors. *Diagnostics*, *13*(2), Article 258. https://doi.org/10.3390/diagnostics13020258

Sudjai, N., Siriwanarangsun, P., Lektrakul, N., Saiviroonporn, P., Maungsomboon, S., Phimolsarnti, R., Asavamongkolkul, A., & Chandhanayingyong, C. (2023b). Tumor-to-bone distance and radiomic features on MRI distinguish intramuscular lipomas from well-differentiated liposarcomas. *Journal of Orthopaedic Surgery and Research*, *18*(1), Article 255. https://doi.org/10.1186/s13018-023-03718-4

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Urgan, N. N., & Tez, M. (2008). Liu estimator in logistic regression when the data are collinear. In *International Conference on Continuous Optimization and Knowledge-Based Technologies, Lithuania, Selected Papers* (pp. 323–327). Vilnius.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429. https://doi.org/10.1198/016214506000000735

Zou, H., & 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.

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

SUPPLEMENTARY

Table S1. List of 50 texture features

Gray-level co-occurrence matrix (GLCM)	Gray level dependence matrix (GLDM)	Gray-level run length matrix (GLRLM)	Gray-level size zone matrix (GLSZM)	Neighbouring gray tone difference matrix (NGTDM)
glcm_Autocorrelation (f1)	gldm_GrayLevelNonUniformity (f23)	glrlm_GrayLevelVariance (f24)	glszm_GrayLevelNonUniformity (f37)	ngtdm_Contrast (f50)
glcm_ClusterProminence (f2)		glrlm_HighGrayLevelRunEmphasis (f25)	glszm_GrayLevelVariance (f38)	
glcm_ClusterTendency (f3)		glrlm_LongRunEmphasis (f26)	glszm_HighGrayLevelZoneEmphasis (f39)	
glcm_Contrast (f4)		glrlm_LongRunHighGrayLevelEmphasis (f27)	glszm_LargeAreaEmphasis (f40)	
glcm_Correlation (f5)		glrlm_LongRunLowGrayLevelEmphasis (f28)	glszm_LargeAreaHighGrayLevelEmphasis (f41)	
glcm_DifferenceAverage (f6)		glrlm_LowGrayLevelRunEmphasis (f29)	glszm_LargeAreaLowGrayLevelEmphasis (f42)	
glcm_DifferenceEntropy (f7)		glrlm_RunEntropy (f30)	glszm_LowGrayLevelZoneEmphasis (f43)	
glcm_DifferenceVariance (f8)		glrlm_RunLengthNonUniformity (f31)	glszm_SizeZoneNonUniformity (f44)	
glcm_Id (f9)		glrlm_RunLengthNonUniformityNormalized (f32)	glszm_SmallAreaEmphasis (f45)	
glcm_Idm (f10)		glrlm_RunPercentage (f33)	glszm_SmallAreaHighGrayLevelEmphasis (f46)	
glcm_Idmn (f11)		glrlm_ShortRunEmphasis (f34)	glszm_SmallAreaLowGrayLevelEmphasis (f47)	
glcm_Idn (f12)		glrlm_ShortRunHighGrayLevelEmphasis (f35)	glszm_ZoneEntropy (f48)	
glcm_Imc1 (f13)		glrlm_ShortRunLowGrayLevelEmphasis (f36)	glszm_ZonePercentage (f49)	
glcm_Imc2 (f14)				
glcm_InverseVariance (f15)				
glcm_JointAverage (f16)				
glcm_JointEnergy (f17)				
glcm_JointEntropy (f18)				
glcm_MaximumProbability (f19)				
glcm_SumAverage (f20)				
glcm_SumEntropy (f21)				
glcm_SumSquares (f22)				

Table S2. Example of the formula and definition of the texture features

Feature class name	Feature name	Formula	Definition
Gray level co-occurrence matrix (GLCM): *Where is the normalized co-occurrence matrix* $\left[p(i,j)=\frac{P(i,j)}{\sum P(i,j)}\right]$*.* $P(i,j)$ *is the co-occurrence matrix for an arbitrary* δ *and* θ*.* N_g *is the number of discrete intensity levels in the image.* ε *is an arbitrarily small positive number* $(\approx 2.2\times 10^{-16})$*.*	glcm_JointEnergy	Joint energy = $\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\big(p(i,j)\big)^2$	Measures homogeneous patterns in the image. Greater energy values indicate that there are more instances of intensity value pairs in the image that neighbour each other at higher frequencies.
	glcm_JointEntropy	Joint entropy = $-\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)\, log_2(p(i,j)+\varepsilon)$	Measures the randomness/variability in neighbourhood intensity values.
Gray-level run-length matrix (GLRLM): *Where* N_g *is the number of discreet intensity values in the image.* N_r *is the number discreet run lengths in the image.* $P(i,j)$ *is the run-length matrix for an arbitrary direction* θ*, when* $i = 1,2,3,\ldots,N_g$ *and* $j = 1,2,3,\ldots,N_r$*.*	glrlm_RunLengthNonUniformity (RLN)	$RLN=\frac{\sum_{j=1}^{N_r}\left(\sum_{i=1}^{N_g}P(i,j\|\theta)\right)^2}{N_r(\theta)}$	Measures the similarity of run lengths throughout the image. Lower values indicate that there are more homogeneity among run lengths in the image.
	glrlm_LowGrayLevelRunEmphasis (LGLRE)	$\mathrm{LGLRE}=\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{P(i,j\|\theta)}{i^2}}{N_r(\theta)}$	Measures the distribution of low gray-level values. Higher values indicate that there are a greater concentration of low gray-level values in the image.
Gray level size zone matrix (GLSZM): *Where* N_g *is the number of discreet intensity values in the image.* N_s *is the number of discreet zone sizes in the image.* N_p *is the number of voxels in the image.* N_z *is the number of zones in the ROI, which is equal to* $\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}P(i,j)$ *and* $1\le N_z\le N_p$*.* $P(i,j)$ *is the size zone matrix, when* $i = 1,2,3,\ldots,N_g$ *and* $j = 1,2,3,\ldots,N_s$*.*	glszm_SmallAreaEmphasis (SAE)	$\mathrm{SAE}=\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\frac{P(i,j)}{j^2}}{N_z}$	Measures the distribution of small size zones. Greater values indicate that there are more smaller size zones and more fine textures.
	glszm_LargeAreaEmphasis (LAE)	$\mathrm{LAE}=\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}P(i,j)j^2}{N_z}$	Measures the distribution of large area size zones. Greater values indicate that there are more larger size zones and more coarse textures.