Adaptive Elastic Net on High-Dimensional Sparse Data with Multicollinearity: Application to Lipomatous Tumor Classification

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

[1]*Department of Orthopaedic Surgery, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok 10700, Thailand*

[2]*Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Pathum Thani 12120, Thailand*

Abstract: Predictive models can experience instabilities because of the combination of high-dimensional sparse data and multicollinearity problems. The adaptive Least Absolute Shrinkage and Selection Operator (adaptive Lasso) and adaptive elastic net were developed using the adaptive weight on penalty term. These adaptive weights are related to the power order of the estimators. Therefore, we concentrate on the power of adaptive weight on these penalty functions. This study purposed to compare the performances of the power of the adaptive Lasso and adaptive elastic net methods under high-dimensional sparse data with multicollinearity. Moreover, we compared the performances of the ridge, Lasso, elastic net, adaptive Lasso, and adaptive elastic net in terms of the mean of the predicted mean squared error (MPMSE) for the simulation study and the classification accuracy for a real-data application. The results of the simulation and the real-data application showed that the square root of the adaptive elastic net performed best on high-dimensional sparse data with multicollinearity.

Keywords: Diagnostic classification, High-dimensional sparse data, Machine-learning, Multicollinearity, Penalized logistic regression, Penalty function.

INTRODUCTION

Lipomatous tumors are a group of the most common soft-tissue tumors, which can be benign lipomas or low/high-grade liposarcomas [1]. Lipomas are benign adipocytic tumors (accounting for approximately one-third of soft tissue tumors), which can be treated conservatively with observation only. Surgical excision is unnecessary except possibly for a symptomatic patient. Atypical lipomatous tumors/well-differentiated liposarcomas (ALTs/WDLSs) are adipocytic malignancies, accounting for 40-45% of all liposarcomas [2]. They recur locally or dedifferentiate to high-grade sarcoma, but rarely metastasize. Regarding a diagnostic system before surgery, although magnetic resonance imaging (MRI) is the most useful diagnostic tool for lipomatous soft-tissue tumors, identifying intramuscular (IM) lipomas that are deep-seated, larger than 5 cm, and symptomatic can be challenging because of their resemblance to ALTs/WDLSs [2]. With tumors representing a diagnostic dilemma from MRI images alone, biopsy is regarded as the reference standard. However, biopsy sampling errors can occur [3]. To alleviate diagnostic uncertainty following MRI, creating a diagnostic system before surgery is beneficial for treatment planning help determine the urgency of surgery, and prevents unnecessary treatment.

Currently, developments in machine-learning algorithms (ML) using MRI-based radiomic features have revolutionized health sciences in diagnostics. Due to a large number of radiomic features, we desire predictive models that can deliver precise outcomes to help decision-making. Logistic regression models are widely applied in data analysis [4, 5] and machine-learning communities [6, 7]. Regarding binary logistic regression coefficients, maximum likelihood estimation (MLE) is a widely used approach to estimate coefficients in the model [8-10]. However, the hybrid of the high-dimensional data and multicollinearity can lead to model over-fitting [11, 12] and can inflate the variance of the maximum likelihood estimators in the model [8, 9]. Consequently, the MLE used for coefficient estimation in logistic regression is unstable for building a classification model [13]. To remedy this problem, the penalized approach can be employed in the logistic regression model [14, 15]. Presently, the elastic net is one of the popular methods for penalty function [16]. Previous studies focused on developing an adaptive weight for the penalty term [17, 18]. However, no studies have compared the performances of penalized logistic regression, focusing on the power of adaptive weight on the adaptive elastic net under high-dimensional sparse data with multicollinearity.

*Address correspondence to these authors at the Department of Orthopaedic Surgery, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok 10700, Thailand; E-mail: chandhanarat.c@gmail.com

Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Pathum Thani 12120, Thailand;
E-mail: monthira@mathstat.sci.tu.ac.th

E-ISSN: 1929-6029/24

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 performance of the power of adaptive weight on the penalized methods under high-dimensional sparse data with multicollinearity in a simulation study. Along with this, the classification performance of these approaches was compared on a lipomatous tumor data application.

MATERIALS AND METHODS

Logistic Regression

The logistic regression approach is widely used in medical diagnostic classifications. With the binary outcome variable, a dependent variable Y_i has a Bernoulli distribution with the parameter $\pi_i = e^{x_i\beta} \Big/ \left(1+e^{x_i\beta}\right)$ where $\underset{\sim}{x}_i = \left(1, x_{i1}, x_{i2}, ..., x_{ip}\right)$ represents a vector of independent variables for the i^{th} observation, $i = 1, 2, 3, ..., n$ and a vector composed of logistic regression coefficients is $\underset{\sim}{\beta} = (\beta_0,\ \beta_1,\ \beta_2, ..., \beta_p)^T$ when p is the number of independent variables and n is the sample size. The binary logistic regression model can be written as:

$$Y_i = \pi_i + \varepsilon_i,\ i = 1, 2, 3, ..., n \tag{1}$$

where ε_i is the random error, which is assumed to follow a distribution with a mean of zero and variance of $\pi_i(1-\pi_i)$.

The transformation of π_i is a central of the model (also called the logit function), which can be determined as follows:

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \underset{\sim}{x}_i\underset{\sim}{\beta} = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j \quad , \quad i = 1, 2, 3, ..., n \text{ and } j = 1, 2, 3, ..., p. \tag{2}$$

The log-likelihood function for the set of observations $\left(y_i, \underset{\sim}{x}_i\right)$ can be written as:

$$\ell(\underset{\sim}{\beta}) = \sum_{i=1}^{n}\left[y_i \ln(\pi_i) + (1-y_i)\ln(1-\pi_i)\right]$$

$$= \sum_{i=1}^{n}\left[y_i\left(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j\right) - \ln\left(1+\exp\left(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j\right)\right)\right]. \tag{3}$$

The estimated parameter of equation (3) can be determined using the maximum likelihood estimation (MLE), which is as follows:

$$\hat{\underset{\sim}{\beta}}_{MLE} = \arg\max_{\underset{\sim}{\beta}}\left(\sum_{i=1}^{n}\left[y_i \ln(\pi_i) + (1-y_i)\ln(1-\pi_i)\right]\right) \tag{4}$$

where $\hat{\underset{\sim}{\beta}}_{MLE}$ is a $(p+1)\times 1$ vector of the maximum likelihood estimators. However, this approach has some limitations about high-dimensional data and multicollinearity. Consequently, the penalized approach is applied as an alternative to the MLE.

From equation (3), we can be written in a form of the penalized approach as follows:

$$\ell^*(\underset{\sim}{\beta}) = -\ell(\underset{\sim}{\beta}) + P_\lambda(\underset{\sim}{\beta}) \tag{5}$$

where $P_\lambda(\underset{\sim}{\beta})$ is the penalty function and λ is the tuning parameter.

Penalized Logistic Regression

The aim of penalized logistic regression is to determine logistic regression coefficients when the data are high correlated and high dimensional. Penalized logistic regression coefficient is defined as follows:

$$\hat{\underset{\sim}{\beta}}_{PLR} = \arg\min_{\underset{\sim}{\beta}}\left(-\left\{\sum_{i=1}^{n}\left[\begin{array}{l} y_i \ln(\pi_i) + \\ (1-y_i)\ln(1-\pi_i)\end{array}\right]\right\} + P_\lambda(\underset{\sim}{\beta})\right);\ \lambda \geq 0 \tag{6}$$

where $\hat{\underset{\sim}{\beta}}_{PLR} = (\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2, ..., \hat{\beta}_p)^T$. $P_\lambda(\underset{\sim}{\beta})$ and λ are a penalty term and the tuning parameter, respectively. In the case of $\lambda = 0$, $\hat{\underset{\sim}{\beta}}_{PLR} = \hat{\underset{\sim}{\beta}}_{MLE}$. Regrading selecting λ, cross-validation is commonly used to evaluate the optimal value of this parameter. Currently, the elastic net and adaptive elastic net are popular approaches for cancer classification [19, 20], which are described below.

Elastic Net

Elastic net was developed by Zou and Hastie [16], which combines the properties of Lasso and ridge regression. This method comprises the parts of the ℓ_1 -norm and ℓ_2 -norm penalties, which is defined in two steps. First, the naive elastic net estimators are determined as follows:

$$\hat{\underset{\sim}{\beta}}_{\text{Nelastic}} = \arg\min_{\underset{\sim}{\beta}}\left(-\left\{\sum_{i=1}^{n}\left[\begin{array}{l} y_i \ln(\pi_i) + \\ (1-y_i)\ln(1-\pi_i)\end{array}\right]\right\} + \lambda_1\sum_{j=1}^{p}\left|\beta_j\right| + \lambda_2\sum_{j=1}^{p}\beta_j^2\right) \tag{7}$$

where $\lambda_1, \lambda_2 \geq 0$; $\lambda = \lambda_1 + \lambda_2$; and $\alpha = \lambda_2 \big/ (\lambda_1 + \lambda_2)$ when $\alpha \in [0, 1)$.

Then, the estimation of $\underset{\sim}{\beta}$ using the elastic net penalty is determined as

$$\hat{\underset{\sim}{\beta}}_{\text{elasticnet}} = (1+\lambda_2)\hat{\underset{\sim}{\beta}}_{\text{Nelastic}} . \quad (8)$$

Regarding shrinkage of $\hat{\underset{\sim}{\beta}}$, parameters λ_1 and λ_2 control the shrinkage of $\hat{\underset{\sim}{\beta}}$ using cross-validation strategy [21].

Adaptive Elastic Net

The adaptive elastic net method is a hybrid of adaptive Lasso and ridge regression [18]. Consequently, it enjoys oracle properties and has out-performance to the elastic net method. The adaptive elastic net penalty is as follows:

$$P^{Aelastic}_{\lambda_1,\lambda_2}(\underset{\sim}{\beta}) = \lambda_1 \sum_{j=1}^{p} \hat{w}_j \left|\beta_j\right| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \quad (9)$$

where $\hat{\underset{\sim}{w}}_j = \left|\left(\hat{\underset{\sim}{\beta}}_{elasticnet}\right)_j\right|^{-\gamma}$; $\gamma > 0$. γ is the power of adaptive weight.

Hence, the estimation of $\underset{\sim}{\beta}$ using adaptive elastic net can be determined as follows:

$$\hat{\underset{\sim}{\beta}}_{Aelastic} = \arg\min_{\underset{\sim}{\beta}} \left(-\left\{ \sum_{i=1}^{n} \begin{bmatrix} y_i \ln(\pi_i) + \\ (1-y_i)\ln(1-\pi_i) \end{bmatrix} \right\} + P^{Aelastic}_{\lambda_1,\lambda_2}(\underset{\sim}{\beta}) \right). \quad (10)$$

The tuning parameters (λ_1 and λ_2) control the shrinkage of $\hat{\underset{\sim}{\beta}}$ by using Bayesian information criterion cross-validation approach.

Monte Carlo Simulation

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

1. High-dimensional sparse data [22]. Given $p > n$. Under the sparsity assumption on the true coefficients ($\underset{\sim}{\beta}$), we defined that the number of significant predictors equaled q, and $q < p$. $\underset{\sim}{x}_i = \left(\underset{\sim}{x}_{iA}, \underset{\sim}{x}_{iB}\right)$ when $\underset{\sim}{x}_{iA} = \left(x_{i1}, x_{i2}, x_{i3}, ..., x_{iq}\right)^T \in \mathbb{R}^q$ and $\underset{\sim}{x}_{iB} = \left(x_{i(q+1)}, x_{i(q+2)}, x_{i(q+3)}, ..., x_{ip}\right)^T \in \mathbb{R}^{p-q}$. Therefore, $\underset{\sim}{X} = \left(\underset{\sim}{x}_A, \underset{\sim}{x}_B\right)^T \in \mathbb{R}^{n\times p}$ is the matrix of all independent variables when $\underset{\sim}{x}_A = \left(\underset{\sim}{x}_{1A}, ..., \underset{\sim}{x}_{nA}\right)^T \in \mathbb{R}^{n\times q}$ and $\underset{\sim}{x}_B = \left(\underset{\sim}{x}_{iB}, ..., \underset{\sim}{x}_{nB}\right)^T \in \mathbb{R}^{n\times(p-q)}$.

2. All independent variables are correlated by using the Toeplitz correlation structure, which is as following [23].

$$\sum_k = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{k-1} \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{k-2} \\ \rho^2 & \rho & 1 & \rho & \cdots & \rho^{k-3} \\ \rho^3 & \rho^2 & \rho & 1 & \cdots & \rho^{k-4} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{k-1} & \rho^{k-2} & \rho^{k-3} & \rho^{k-4} & \cdots & 1 \end{pmatrix}_{k\times k} \quad (11)$$

where number of the independent variables (k) represents a positive integer and $0 \le \rho \le 1$.

The Monte Carlo simulations were constructed using 50 and 100 independent variables. The sample size equaled 30 and 40. We generated the independent variables from the multivariate normal distribution with a mean of zero and covariance $\sum$ ($\underset{\sim}{X} \sim N(0, \sum)$). The dependent variables were generated from the Bernoulli distribution with parameter π_i. The degree of correlation (ρ) was set to 0.75, 0.85, and 0.95. The number of significant predictors (q) equaled 15. The logistic regression coefficients were set the constant values as $\underset{\sim}{\beta}$. Subsequently, we split the data into two subsets (80% of learning dataset, and 20% of testing dataset). The simulation study compared the performances of the ridge, Lasso, elastic net, adaptive Lasso, and adaptive elastic net using the predicted mean square errors (PMSE). The estimated PMSE was calculated as following:

$$\text{PMSE} = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n} . \quad (12)$$

where y_i and $\hat{y}_i$ were the i^{th} actual and predicted values of the dependent variables, respectively. We used a 10-fold cross-validation strategy to estimate the optimal value of the tuning parameter (λ) [14, 16, 21]. To achieve a stationary result, the experiment was repeated 1000 times. Hence, the MPMSE was determined from the average of 1000 estimates of $PMSE_j$.

$$\text{MPMSE} = \frac{1}{1000} \sum_{j=1}^{1000} PMSE_j . \quad (13)$$

The penalized method providing the lowest MPMSE was regarded as the best option. Figure **1** represents the flowchart of the simulation procedure for this study. Moreover, in Figure **2** shows the workflow diagram of the machine-learning procedure for the real-data application. The classification accuracy of the methods was evaluated as following:

$$\text{Accuracy (\%)} = \frac{TP + TN}{TP + FP + FN + TN} \times 100 . \quad (14)$$

Where the true positive (*TP*) shows that the prediction is correct. A true negative (*TN*) indicates that the prediction is correct. False positive (*FP*) shows that the prediction is wrong (type I error). A false negative (*FN*) presents that the prediction is wrong (type II error).

Software

All simulations and analyses were performed using R version 4.3.2 (R Foundation for Statistical Computing, Vienna, Austria). We used the package 'glmnet' to fit models using all above penalized approaches. Regarding support vector machine (SVM), the package 'e1071' was used to build the models.

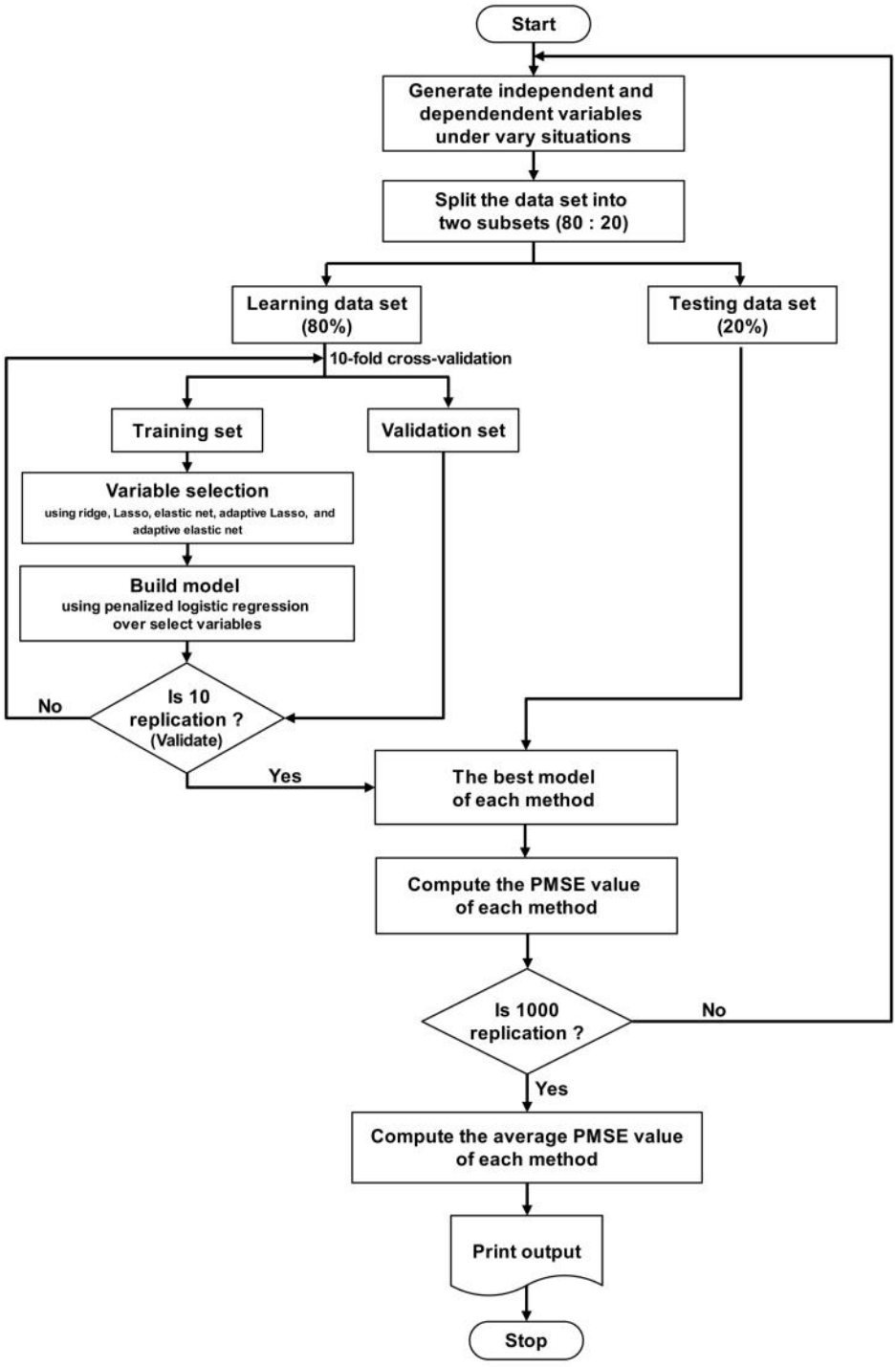

Figure 1: Flowchart of the simulation procedure.

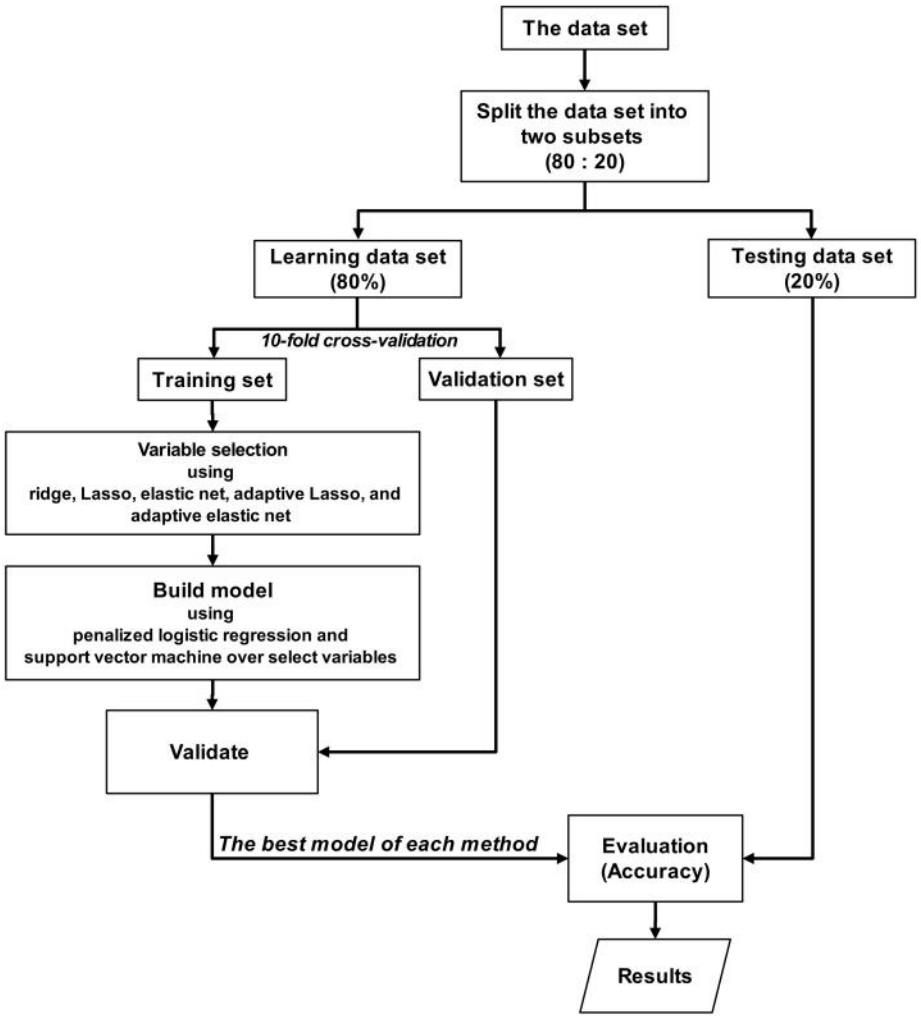

Figure 2: Workflow diagram of the machine-learning procedure.

RESULTS AND DISCUSSION

Simulation Study

Table **1** presents the MPMSE values for the penalized methods for different ρ when p = 50 and 100, and n = 30 and 40. We found 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.

With high-dimensional sparse data (p = 50, n = 30, and n = 40, and for different ρ), we found that the performance of the adaptive Lasso and adaptive elastic net methods depended on the power of the adaptive weight (γ). The MPMSE values of the adaptive elastic net with γ = 0.5 were less than those for the other methods.

In Table **2**, when p = 100, n = 30 and 40, and for different ρ, we found that the smallest MPMSE values were also obtained from the adaptive elastic net method with γ = 0.5.

From the simulated results in Tables **1** and **2**, it can be seen that the key factors influencing the MPMSE values were the power of adaptive weight on penalty term, the correlated independent variables (or correlation coefficient level), and the sample size. An increase in the correlation coefficient level lends to an increase in the MPMSE values for all methods when holding p and n fixed. The worst case was obtained when the correlation coefficient level was very high (ρ = 0.95). Additionally, an increase in the sample size affects a decrease in the MPMSE values for all methods, while holding ρ and p fixed.

Real-Data Applications

In this section, we compared the performances of the penalized methods on a real-data set, which is given below.

Table 1: Mean of the Predicted Mean Square Errors (MPMSE) Values for Different Penalized Methods When $p = 50$

n	ρ	Ridge	Lasso	Adaptive Lasso			Elastic net	Adaptive elastic net		
				$\gamma = 0.5$	$\gamma = 1$	$\gamma = 2$		$\gamma = 0.5$	$\gamma = 1$	$\gamma = 2$
30	0.75	0.227	0.190	0.185	0.184	0.187	0.192	0.183*	0.187	0.191
	0.85	0.228	0.193	0.191	0.190	0.192	0.196	0.189*	0.193	0.196
	0.95	0.230	0.195	0.198	0.197	0.199	0.198	0.194*	0.195	0.197
40	0.75	0.224	0.187	0.182	0.183	0.183	0.191	0.180*	0.182	0.185
	0.85	0.227	0.191	0.190	0.188	0.191	0.194	0.185*	0.188	0.192
	0.95	0.229	0.193	0.197	0.195	0.199	0.197	0.193*	0.193	0.196

Lasso, Least Absolute Shrinkage and Selection Operator; * The method providing the smallest MPMSE.

Table 2: Mean of the Predicted Mean Square Errors (MPMSE) Values for Different Penalized Methods When $p = 100$

n	ρ	Ridge	Lasso	Adaptive Lasso			Elastic net	Adaptive elastic net		
				$\gamma = 0.5$	$\gamma = 1$	$\gamma = 2$		$\gamma = 0.5$	$\gamma = 1$	$\gamma = 2$
30	0.75	0.221	0.187	0.184	0.182	0.185	0.192	0.180*	0.187	0.191
	0.85	0.222	0.190	0.189	0.188	0.192	0.193	0.187*	0.193	0.196
	0.95	0.225	0.196	0.196	0.195	0.199	0.198	0.193*	0.195	0.197
40	0.75	0.218	0.184	0.183	0.181	0.184	0.187	0.180*	0.182	0.185
	0.85	0.220	0.187	0.185	0.184	0.191	0.189	0.182*	0.188	0.192
	0.95	0.223	0.194	0.195	0.194	0.198	0.197	0.193*	0.193	0.196

Lasso, Least Absolute Shrinkage and Selection Operator; * The method providing the smallest MPMSE.

Table 3: Percentage Accuracy of Classification of the Machine-Learning Algorithms for Discriminating between Intramuscular Lipomas and Atypical Lipomatous Tumors/Well Differentiated Liposarcomas

Variable selection method	Ridge	Lasso	Adaptive Lasso			Elastic net	Adaptive elastic net		
			$\gamma = 0.5$	$\gamma = 1$	$\gamma = 2$		$\gamma = 0.5$	$\gamma = 1$	$\gamma = 2$
Classifier with penalized logistic regression	82.0	87.4	90.0	91.2	88.0	86.5	93.0*	91.6	90.1
Classifier with support vector machine	87.0	89.0	90.8	92.0	89.5	87.0	92.5*	91.5	89.9

Lasso, Least Absolute Shrinkage and Selection Operator; * The machine-learning algorithms providing the highest accuracy.

Lipomatous soft-tissue tumor data set was obtained from 40 patients (20 IM lipomas and 20 ALTs/WDLSs), which had been treated at our institution between 2010 and 2020. The patients were diagnosed using their final pathological findings, and underwent MRI scans and total excision surgery. For our case study, the binary outcome of interest was an IM lipoma (benign tumor) or an ALT/WDLS (malignant tumor). The predictors of interest were 50 radiomic features as continuous variables; these features were extracted from preoperative T1-weighted MRI (Appendix A). We can see that the number of predictors/independent variables is large compared with the number of observations. It is clear that the high-dimensional problem was presented in this data set. Regarding Figure **3**, the correlation matrix presents different shades and the Pearson correlation coefficient values. The light shade (or the Pearson correlation coefficient value is close to zero) represents that the predictors have a low correlation, whereas the dark shade (or the Pearson correlation coefficient value is close to 1 or -1) presents a high correlation among predictors. Moreover, the variance inflation factor (or VIF) values for almost predictors were over 10, which indicates a problematic amount of collinearity. It is obvious that the multicollinearity problem was to occur in this data set.

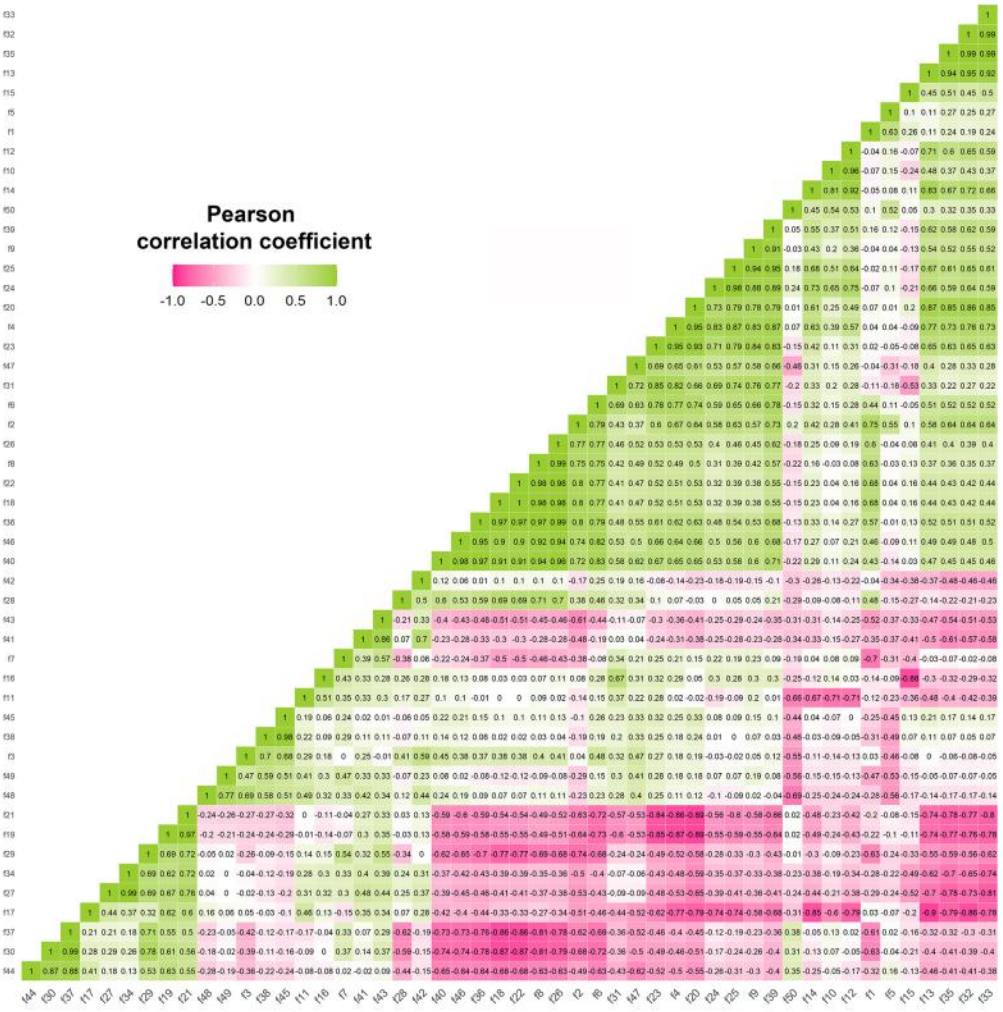

Figure 3: Correlation matrix of fifty radiomic features in forty patients.

To remedy the two above problems, we applied the penalized approach (i.e., ridge, Lasso, elastic net, adaptive Lasso, and adaptive elastic net) on this case study.

In Table **3**, we appraised the classification performances of the penalized methods in distinguishing between IM lipomas and ALTs/WDLSs. We can see that the highest accuracy values were obtained from the adaptive elastic net method with $w_j = \left|\hat{\beta}_j^{Aelastic}\right|^{-0.5}$, while the lowest accuracy values were obtained from the ridge method.

From the results of the real-data applications in Table **3**, it is apparent that the adaptive elastic net method with $w_j = \left|\hat{\beta}_j^{Aelastic}\right|^{-0.5}$ 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.

CONCLUSION

We propose the use of the adaptive elastic net method with γ = 0.5 for classification on high-dimensional data with multicollinearity. Both the simulation study and the real-data application, it is clear that the classification performance of the adaptive elastic net logistic regression model depends on the power of adaptive weight on this penalty term. In practice, if the penalty technique is appropriate, it results the classification model that has good performance.

AUTHOR CONTRIBUTIONS

Conceptualization, N.S., M.D., C.C.; methodology, N.S., M.D., C.C.; software, N.S., M.D.; validation, N.S., M.D., C.C.; formal analysis, N.S., M.D., C.C.; investigation, N.S., C.C.; resources, C.C.; data curation, N.S., C.C.; writing—original draft preparation, N.S.; writing—review and editing, N.S., M.D., C.C.; project administration, N.S.; funding acquisition, C.C.

All authors have read and agreed to the published version of the manuscript.

FUNDING

This research was funded by the Siriraj Foundation Fund for advanced sarcoma research, grant number D004146.

INSTITUTIONAL REVIEW BOADR STATEMENT

The lipomatous soft-tissue tumor data set in this study was retrieved after approval from the Ethics Committee of the Faculty of Medicine Siriraj Hospital, Mahidol University (approval number MU-MOU CoA 874/2020).

ACKNOWLEDGEMENTS

We thank Supani Duangkaew for her research coordination and Pakorn Yodprom for physics consultation and magnetic resonance image retrieval.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

APPENDIX A

The radiomic features of interest in our case study comprised 7 first-order features, 17 gray-level co-occurrence matrix (GLCM), 13 gray-level run length matrix (GLRLM), 10 gray-level size zone matrix (GLSZM), and 3 shape-based 3D (Table **A1**). The first-order features were the simplest statistical descriptors that explained the distribution of their gray-level values in MRI images. Texture-based features (i.e., GLCM, GLRLM, and GLSZM) describe the spatial arrangement of gray-level pixels in a neighborhood on the images (e.g., homogeneity/heterogeneity/ fineness/coarseness of region of interest (ROI) on the images). Additionally, shape-based 3D features explain the geometric properties of ROI.

The details of formula and definition for these features were explained according to PyRadiomics' documentation. (https://pyradiomics.readthedocs.io/en/latest/features.html, accessed on 9 March 2024) (Table **A2**).

Table A1: List of Fifty Radiomic Features

First-order features	Texture-based features			Shape-based features
	GLCM	GLRLM	GLSZM	
firstorder_10Percentile (f1)	glcm_Autocorrelation (f8)	glrlm_GrayLevelVariance (f25)	glszm_GrayLevelNonUniformity (f38)	shape_MajorAxisLength (f48)
firstorder_90Percentile (f2)	glcm_ClusterTendency (f9)	glrlm_HighGrayLevelRunEmphasis (f26)	glszm_GrayLevelVariance (f39)	shape_SurfaceArea (f49)
firstorder_Energy (f3)	glcm_Contrast (f10)	glrlm_LongRunEmphasis (f27)	glszm_HighGrayLevelZoneEmphasis (f40)	shape_SurfaceVolumeRatio (f50)
firstorder_Entropy (f4)	glcm_Correlation (f11)	glrlm_LongRunHighGrayLevelEmphasis (f28)	glszm_LargeAreaEmphasis (f41)	
firstorder_Minimum (f5)	glcm_DifferenceAverage (f12)	glrlm_LongRunLowGrayLevelEmphasis (f29)	glszm_LargeAreaHighGrayLevelEmphasis (f42)	
firstorder_Maximum (f6)	glcm_DifferenceEntropy (f13)	glrlm_LowGrayLevelRunEmphasis (f30)	glszm_LargeAreaLowGrayLevelEmphasis (f43)	
firstorder_Skewness (f7)	glcm_DifferenceVariance (f14)	glrlm_RunEntropy (f31)	glszm_LowGrayLevelZoneEmphasis (f44)	
	glcm_Imc1 (f15)	glrlm_RunLengthNonUniformityNormalized (f32)	glszm_SizeZoneNonUniformity (f45)	
	glcm_Imc2 (f16)	glrlm_RunPercentage (f33)	glszm_SmallAreaHighGrayLevelEmphasis (f46)	
	glcm_InverseVariance (f17)	glrlm_RunVariance (f34)	glszm_ZoneEntropy (f47)	
	glcm_JointAverage (f18)	glrlm_ShortRunEmphasis (f35)		
	glcm_JointEnergy (f19)	glrlm_ShortRunHighGrayLevelEmphasis (f36)		
	glcm_JointEntropy (f20)	glrlm_ShortRunLowGrayLevelEmphasis (f37)		
	glcm_MaximumProbability (f21)			
	glcm_SumAverage (f22)			
	glcm_SumEntropy (f23)			
	glcm_SumSquares (f24)			

GLCM, gray-level co-occurrence matrix; GLRLM, gray-level run length matrix; GLSZM, gray-level size zone matrix.

Table A2: Example of the Formula and Definition of the Radiomic Features

Feature	Feature class name	Feature name	Formula	Definition
Histogram-based	First order: *When X is set of voxel intensity within a segmented region of interest (ROI).* $X = \{x_1, x_2, x_3, ..., x_{N_p}\}$	firstorder_Minimum	Minimum = min (X)	The minimum value of X
		firstorder_Energy	$\text{Energy} = \sum_{i=1}^{N_p}(x_i + c)^2$ where c is optimal value which shifts the intensities to prevent negative values in X .	Measures the magnitude of voxel values in an image.
		firstorder_Skewness	$\text{Skewness} = \frac{\frac{1}{N_p}\sum_{i=1}^{N_p}(x_i - \bar{x})^3}{\left(\sqrt{\frac{1}{N_p}\sum_{i=1}^{N_p}(x_i - \bar{x})^2}\right)^3}$	Measures the asymmetry of the distribution of voxel intensity within a segmented ROIs. - Negative skewness indicates that the curve is extended towards the left side. (mean < median < mode) - Skewness = 0, which means that the curve is a normal distribution. - Positive skewness means that the curve is extended towards the right side. (mode < median < mean)
Texture-based	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* $(\approx 2.2 \times 10^{-16})$. $p_{x_i} = \sum_{j=1}^{N_g} p(i,j)$ *be the marginal row probabilities.* $p_{y_j} = \sum_{i=1}^{N_g} p(i,j)$ *be the marginal column probabilities.* μ_x *is the mean gray level intensity of* p_x *and defined as* $\mu_x = \sum_{i=1}^{N_g}\left[\left(p_{x_i}\right)(i)\right] \cdot \mu_y$ *is the mean gray level intensity of* p_y *and defined as* $\mu_y = \sum_{j=1}^{N_g}\left[\left(p_{y_j}\right)(j)\right]$.	glcm_Autocorrelation	$\text{Autocorrelation} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)ij$	Measures the magnitude of the fineness and coarseness of texture.
		glcm_ClusterTendency	$\text{Cluster tendency} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\left(i + j - \mu_x - \mu_y\right)^2 p(i,j)$	Measures groupings of voxels with similar gray-level values.
		glcm_Correlation	$\text{Correlation} = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)ij - \mu_x\mu_y}{\sigma_x(i)\sigma_y(j)}$	Correlation is a value between 0 (uncorrelated) and 1 (perfectly correlated) showing the linear dependency of gray-level values to their respective voxels in the GLCM.
	Gray-level run length matrix (GLRLM): *Where* N_g *is the number of discreet intensity*	glrlm_LowGrayLevelRunEmphasis (LGLRE)	$\text{LGLRE} = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{P(i,j\mid\theta)}{i^2}}{N_r(\theta)}$	Measures the distribution of low gray-level values, with higher value indicating a greater concentration of low gray-level values in the image.

	values in the image N_r is the number of discreet run lengths in the image. $P(i,j)$ is the run-length matrix for an arbitrary direction θ, when $i=1,2,3,...,N_g$ and $j=1,2,3,...,N_r$. ε is an arbitrarily small positive number $(\approx 2.2\times10^{-16})$.	glrlm_RunEntropy (RE)	$\mathrm{RE} = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j\|\theta)\log_2\left(p(i,j\|\theta)+\varepsilon\right)$	Measures the uncertainty/randomness in the distribution of run lengths and gray levels. A higher value indicates more heterogeneity in the texture patterns.
		glrlm_ShortRunEmphasis (SRE)	$\mathrm{SRE} = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{P(i,j\|\theta)}{j^2}}{N_r(\theta)}$	Measures the distribution of short run lengths, with a greater value indicative of shorter run lengths and more fine textural textures.
	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 $i=1,2,3,...,N_g$ and $j=1,2,3,...,N_s$. ε is an arbitrarily small positive number ($\approx 2.2\times10^{-16}$).*	glszm_GrayLevelNonUniformity (GLN)	$\mathrm{GLN} = \frac{\sum_{i=1}^{N_g}\left(\sum_{j=1}^{N_s}P(i,j)\right)^2}{N_z}$	Measures the variability of gray-level intensity values in the image, with a lower value indicating more homogeneity in intensity values.
		glszm_SizeZoneNonUniformity (SZN)	$\mathrm{SZN} = \frac{\sum_{j=1}^{N_s}\left(\sum_{i=1}^{N_g}P(i,j)\right)^2}{N_z}$	Measures the variability of size zone volumes in the image, with a lower value indicating more homogeneity in size zone volumes.
		glszm_ZoneEntropy (ZE)	$\mathrm{ZE} = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}p(i,j)\log_2\left(p(i,j)+\varepsilon\right)$	Measures the uncertainty/randomness in the distribution of zone sizes and gray levels. A higher value indicates more heterogeneneity in the texture patterns.
Shape-based	Shape features (3D)	shape_SurfaceVolumeRatio	*Surface area to volume ratio* = $\frac{A}{V}$, where $A=\sum_{i=1}^{N_f}\left(\frac{1}{2}\left\|a_ib_i\times a_ic_i\right\|\right)$. a_ib_i and a_ic_i are edges of the i^{th} triangle in the mesh, formed by vertices a_i, b_i and c_i. V is shape_MeshVolume feature (i.e., the mesh volume in mm^3 of the segmented ROI). $V=\sum_{i=1}^{N_f}\left(\frac{(Oa_i\cdot(Ob_i\times Oc_i))}{6}\right)$, where N_f is the number of faces (triangles) defining the Mesh. For each face i in the mesh, defined by points a_i, b_i and c_i, the (signed) volume V_f of the tetrahedron defined by that face and the origin of the image (O) is calculated.	In the case of a lower value, it indicates that there is a more compact (sphere-like) shape.

REFERENCES

[1] Johnson CN, Ha AS, Chen E, Davidson D. Lipomatous soft-tissue tumors. J Am Acad Orthop Surg 2018; 26(22): 779-88. https://doi.org/10.5435/JAAOS-D-17-00045

[2] Burusapat C, Wongprakob N, Wanichjaroen N, Pruksapong C, Satayasoontorn K. Atypical lipomatous tumor/well-differentiated liposarcoma with intramuscular lipoma-like component of the thigh. Case Rep Surg 2020; 2020: 8846932. https://doi.org/10.1155/2020/8846932

[3] Thavikulwat AC, Wu JS, Chen X, Anderson ME, Ward A, Kung J. Image-guided core needle biopsy of adipocytic tumors: diagnostic accuracy and concordance with final surgical pathology. AJR Am J Roentgenol 2021; 216(4): 997-1002. https://doi.org/10.2214/ajr.20.23080

[4] Makalic E, Schmidt DF. Review of modern logistic regression methods with application to small and medium sample size problems. In: Li J, ed. AI 2010: Advances in Artificial Intelligence. Berlin, Heidelberg: Springer 2011; 213-22.

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

[6] Sudjai N, Siriwanarangsun P, Lektrakul N, Saiviroonporn P, Maungsomboon S, Phimolsarnti R, *et al*. Tumor-to-bone distance and radiomic features on MRI distinguish intramuscular lipomas from well-differentiated liposarcomas. J Orthop Surg Res 2023; 18(1): 255. https://doi.org/10.1186/s13018-023-03718-4

[7] Sudjai N, Siriwanarangsun P, Lektrakul N, Saiviroonporn P, Maungsomboon S, Phimolsarnti R, *et al*. Robustness of radiomic features: two-dimensional versus three-dimensional MRI-based feature reproducibility in lipomatous soft-tissue tumors. Diagnostics 2023; 13(2): 258. https://doi.org/10.3390/diagnostics13020258

[8] Hosmer DW, Lemeshow SJ. Applied logistic regression. 3 ed. New Jersey: Wiley 2013.

[9] Kleinbaum DG, Klein M. Logistic regression: a self-learning text. 3rd ed. New York: Springer 2010.

[10] Senaviratna NAMR, Cooray TMJA. Multicollinearity in binary logistic regression model. In: Thapa N, editor. Theory and practice of mathematics and computer science. 1st ed. West Bengal: BP International 2021; pp. 11-9.

[11] Belsley DA, Kuh E, Welsch RE. Regression diagnostics: Identifying influential data and sources of collinearity. New York: John Wiley & Sons 1980.

[12] Brimacombe M. High-dimensional data and linear models: a review. Open Access Med Stat 2014; 4: 17-27. https://doi.org/10.2147/OAMS.S56499

[13] Kastrin A, Peterlin B. Rasch-based high-dimensionality data reduction and class prediction with applications to microarray gene expression data. Expert Syst Appl 2010; 37: 5178-85. https://doi.org/10.1016/j.eswa.2009.12.074

[14] Pavlou M, Ambler G, Seaman S, De Iorio M, Omar RZ. Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. Stat Med 2016; 35(7): 1159-77. https://doi.org/10.1002/sim.6782

[15] Hosseinnataj A, Bahrampour A, Baneshi M, Zolala F, Nikbakht R, Torabi M, *et al*. Penalized Lasso methods in health data: application to trauma and influenza data of Kerman. Journal of Kerman University of Medical Sciences 2019; 26(6): 440-9. https://doi.org/10.22062/jkmu.2019.89573

[16] Zou H, Hastie T. Regularization and variable selection via the elastic Net. J R Stat Soc Series B Stat Methodol 2005; 67(2): 301-20. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[17] Zou H. The adaptive Lasso and Its oracle properties. J Am Stat Assoc 2006; 101(476): 1418-29. https://doi.org/10.1198/016214506000000735

[18] Zou H, Zhang HH. On the adaptive elastic-net with a diverging number of parameters. Ann Stat 2009; 37(4): 1733-51. https://doi.org/10.1214/08-AOS625

[19] Kamalapathy PN, Ramkumar DB, Karhade AV, Kelly S, Raskin K, Schwab J, *et al*. Development of machine learning model algorithm for prediction of 5-year soft tissue myxoid liposarcoma survival. J Surg Oncol 2021; 123(7): 1610-7. https://doi.org/10.1002/jso.26398

[20] Kamalapathy PN, Gonzalez MR, de Groot TM, Ramkumar D, Raskin KA, Ashkani-Esfahani S, *et al*. Prediction of 5-year survival in soft tissue leiomyosarcoma using a machine learning model algorithm. J Surg Oncol 2023. https://doi.org/10.1002/jso.27514

[21] Hastie T, Tibshirani T, Friedman JB. The Elements of statistical learning: data mining inference and prediction. 2nd ed. Berlin/Heidelberg: Springer 2009.

[22] Cherkassky V, Mulier F. Learning from data: concepts, theory, and methods. 2nd ed. New Jersey: John Wiley and Sons 2006.

[23] Hardin J, Garcia SR, Golan D. A method for generating realistic correlation matrices. Ann Appl Stat 2013; 7(3): 1733-62, 30. https://doi.org/10.1214/13-AOAS638

Received on 12-02-2024 Accepted on 05-03-2024 Published on 29-03-2024

https://doi.org/10.6000/1929-6029.2024.13.04