


The Multicollinearity Effect on the Performance of Machine Learning Algorithms: Case Examples in Healthcare Modelling

*1 Hasan YILDIRIM

*1 Corresponding Author, Department of Mathematics, Karamanoğlu Mehmetbey University, Türkiye, hasanyildirim@kmu.edu.tr 

Abstract

The data extracted from various fields inherently consists of extremely correlated measurements in parallel with the exponential increase in the size of the data that need to be interpreted owing to the technological advances. This problem, called the multicollinearity, influences the performance of both statistical and machine learning algorithms. Statistical models proposed as a potential remedy to this problem have not been sufficiently evaluated in the literature. Therefore, a comprehensive comparison of statistical and machine learning models is required for addressing the multicollinearity problem. Statistical models (including Ridge, Liu, Lasso and Elastic Net regression) and the eight most important machine learning algorithms (including Cart, Knn, Mlp, MARS, Cubist, Svm, Bagging and XGBoost) are comprehensively compared by using two different healthcare datasets (including Body Fat and Cancer) having multicollinearity problem. The performance of the models is assessed through cross validation methods via root mean square error, mean absolute error and r-squared criteria. The results of the study revealed that statistical models outperformed machine learning models in terms of root mean square error, mean absolute error and r-squared criteria in both training and testing performance. Particularly the Liu regression often achieved better relative performance (up to 7.60% to 46.08% for Body Fat data set and up to 1.55% to 21.53% for Cancer data set on training performance and up to 1.56% to 38.08% for Body Fat data set and up to 3.50% to 23.29% for Cancer data set on testing performance) among regression methods as well as compared to machine algorithms. Liu regression is mostly disregarded in the machine learning literature, but since it outperforms the most powerful and widely used machine learning algorithms, it appears to be a promising tool in almost all fields, especially for regression-based studies including data with multicollinearity problem.

Keywords: Machine learning; Multicollinearity; Feature selection; Collinearity; Artificial intelligence

1. INTRODUCTION

Mathematical modeling mainly involves the processes of prediction and inference using a set of explanatory variables (i.e. attributes) that are considered to have an effect on a particular (i.e. target) variable. Facilitated by technological advances, the data collection process has significantly increased the scale of the variables. There have emerged highly correlated measurements that are assessed in almost every field, especially in areas such as health, marketing and finance [1]. In big databases containing thousands of variables, it is inevitable that complex patterns of relationships between variables will be discovered. The relationship is considered reasonable to a certain extent, but if it is extreme, a phenomenon known in the statistical literature as multicollinearity (i.e. collinearity) arises [2].

The multicollinearity problem stands out as a problem that is encountered quite frequently in the increasing data size with the ease of data collection in real life problems but is generally under-emphasized [3]. However, mathematically,

this problem causes both statistical and machine learning models to often yield inaccurate inferences and poor predictions (i.e. generalization ability).

The approaches to multicollinearity problem have differed in the statistics and machine learning literatures. In the statistics literature, the focus has been on variable selection by stepwise methods or theoretically modification of the classical ordinary least squares (OLS) estimator (like ridge, Liu estimator) by adding a penalty term to it [3]. In the field of machine learning, particularly in the field of artificial neural networks, models have been proposed with the assumption that they are not affected by multicollinearity due to the complex architecture [3, 4]. In the literature, the ridge estimator has received considerable attention in comparisons, while its alternative, the Liu estimator, has been relatively ignored. Therefore, there is a lack of a comprehensive comparison between machine learning methods and Liu regression.

In this study, we have compared the widely known models in the statistics literature (including Liu regression) with the most important machine learning models in the multicollinearity problem and aimed to contribute the following insights to the literature: (i) In addition to the widely known ridge, lasso and elastic net regression models, the Liu regression model is also considered in this study, (ii) it has been shown that statistical models can provide more effective results than complex machine learning models, (iii) the problem of multicollinearity has been demonstrated to be a problem that should not be ignored and can severely affect the performance of even the most powerful models.

The general layout of the study is as follows:

Section 2 presents the related studies on the subject. The problem of multicollinearity and possible diagnostic approaches are discussed in Section 3. The details about the models used in the study are explained in Section 4. The modelling process and experimental settings are covered in Section 5. The benchmarking results are reported in Section 6. A summary of the conclusions from the study is outlined in Section 7.

2. RELATED WORKS

In the context of multicollinearity, one of the first contributions in the statistical literature was made by James and Stein [5, 6], who proposed the Stein estimator based on equal shrinkage of the coefficients in the classical OLS model. Although this study is the foundation of shrinkage estimators, alternative estimators were required due to equal proportion shrinkage and the inability to handle coefficients with opposite sign. The most noteworthy contribution to this issue was presented by Hoerl and Kennard [7] by proposing the ridge estimator, which is based on shrinking towards zero instead of excluding correlated variables from the model by adding a penalty term to the classical OLS estimator. Since the choice of k in the Ridge estimator is quadratic and complex and that the Stein estimator shrinks all coefficients equally, there have been some disadvantages. Therefore, Liu [8] proposed the Liu estimator, which combines the Ridge and Stein estimator which is like Ridge but includes a penalty term in linear form as well as the Stein estimator properties. Stein, ridge and Liu estimators address the multicollinearity problem to a certain extent by shrinking the coefficients, but they do not have the ability to perform variable selection. The Least Absolute Shrinkage Selection Operator (Lasso), an alternative estimator that utilizes this capability, was proposed by Tibshirani [9]. The Lasso estimator can directly shrink variables to zero instead of shrinking them individually by keeping them in the model and thus making variable selection. To deal with the shortcomings of Lasso in the case of high dimensional and severe collinearity, Zou and Hastie [10] proposed a new estimator called elastic net, which is based on a process that incorporates both Ridge and Lasso simultaneously.

The studies in the field of machine learning have mainly developed within the framework of models based on artificial neural networks. Garg and Tai [4] proposed a model called FA-ANN based on factor analysis and artificial neural networks to deal with multicollinearity. Li and Niu [11]

introduced a new model called R-ELM for multicollinearity by incorporating ridge regression into the algorithm of an extreme learning machine which is a kind of a feed-forward neural network. Panduro and Torsen [12] suggested a two-stage model consisting of principal component analysis and stepwise regression models to overcome the problem of multicollinearity. Dumancas and Bello [13] compared correlated lipid profile data with twelve different machine learning methods (including ridge, lasso, elastic net, extreme gradient boosting, support vector machines etc.). Kilinc et al. [14] conducted a simulation study comparing genetic algorithm and multivariate adaptive splines models as variable selection methods in the presence of multicollinearity. A novel approach of feature selection based on the idea of feature filters has been carried out by Katrutsa and Strijov [15], enabling feature selection without regard to the prediction model. A CNN-based approach has been proposed by Hoseinzade and Haratizadeh [16] to model the correlation between various features in stock market data. Kim et al. [17] presented a combination of principal component analysis and artificial neural networks for correlated and high dimensional data. Obite et al. [18] have compared artificial neural networks and classical least squares models by using real and simulated datasets. Hua [19] proposed a approach of efficient data preprocessing with undersampling and embedded feature selection to address the imbalance of traffic samples and derive the leading features of incoming flows. Qaraad [20] introduced a hybrid optimization model for Cancer Classification to regularize and select the most informative subset of variables in a high-dimensional domain. Bi et al. [21] proposed a heterogeneous phoneme identification system including partial least squares and support vector machines to improve the diagnostic tasks for phoneme pronunciation for correlation data. Abubakar et al. [22] performed a simulation study and compared multiple regression, ridge regression, stepwise regression and partial least squares regression methods on a multicorrelated data. Mahadi et al. [23] introduced a new and efficient technique utilizing the recursive least squares (RLS) algorithms with a time-varying regularization parameter to ensure robustness and improve performance. Kaneko [24] presented a new criterion, called cross-validated permutation feature importance, to assess the feature importance ability of a machine learning model, particularly in the presence of multicollinearity issues. Genç [25] proposed a new regularized extreme learning machine (ELM) algorithm, square-root lasso ELM (SQRTL-ELM), to deal with the shortcomings of the extreme learning machine, including the instability, weak generalizability, and overfitting in the case of multicollinearity.

3. OVERVIEW OF THE MULTICOLLINEARITY

Multicollinearity refers to the near-linear dependencies among the explanatory (i.e. attribute) variables in a regression task. The reasons of this issue can be given as: (i) Data collection method, (ii) Constraints on model or population, (iii) Model identification errors and (iv) An over-defined model [26].

3.1. The Consequences of Multicollinearity

Multicollinearity can cause many serious problems, both theoretical and practical views [26, 27]. Montgomery et al. presented these problem as follows.

- i. The regression model yields coefficients with larger variances and covariances.
- ii. The absolute value of the coefficients can be obtained as larger.
- iii. Although there are exceptions, the model can often produce poorer predictions.

3.2. Multicollinearity Diagnostics

In the literature, various approaches have been proposed to detect the multicollinearity. These approaches are mainly based on the information of the data structure. This information is extracted by calculating the $X'X$ matrix which essentially represents the correlation matrix. The off-diagonal elements of $X'X$ matrix can provide us useful insights about the correlation level of explanatory variables. The common approaches based on this matrix are given as follows [26, 28]:

3.2.1. Calculation the Correlation Matrix

For a given X data matrix, the correlation matrix is obtained as follows:

$$r = X'X, \quad r_{ij} \neq 0 \quad (1)$$

where r_{ij} represents to the off-diagonal elements of the matrix. If r_{ij} exceeds a threshold (usually 0.70), this means that the corresponding variables have highly correlated each other. It may be insufficient to define multicollinearity due to treating variables as pair.

3.2.2. Variance Inflation Factor

Variance inflation factor (VIF) is based on the inverse of $X'X$ matrix. The off-diagonal elements of this inverse matrix provide a more useful and powerful information about the multicollinearity level and calculated as follows:

$$VIF_j = C_{ij} = (1 - R_j^2)^{-1} \quad (2)$$

where R_j^2 is the coefficient of determination calculated via the regression of x_j over the remain $p - 1$ variables. As the value of VIF depending on each variable increases, the severity of the multicollinearity increases. As a common practice, VIF values exceeding 5 or 10 provide strong evidence of a poor model in terms of generalization and estimation abilities [26].

3.2.3. Eigenvalues Analysis

The eigenvalues analysis is an alternative and beneficial approach to VIF or correlation-based approaches. It is mainly based on the decomposition the $X'X$ matrix into two-

different matrix including the eigenvalues and eigenvectors. This decomposition is defined as:

$$X'X = TAT' \quad (3)$$

where $A_{p \times p}$ is the diagonal matrix, whose diagonal elements correspond to the eigenvalues ($\lambda_i, i = 1, 2, \dots, p$) of $X'X$ matrix and $T_{p \times p}$ is the orthogonal matrix whose columns correspond to the eigenvectors of $X'X$ matrix. The presence of small-valued eigenvalues may be evidence of the existence of multicollinearity between the columns of the data. Instead of focusing each eigenvalue, condition number (CN) which is basically a representation of the spread of eigenvalues is commonly used and calculated as follows:

$$CN = \frac{\lambda_{max}}{\lambda_{min}} \quad (4)$$

where λ_{max} and λ_{min} correspond to the maximum and minimum eigenvalues obtained via TAT' eigen decomposition, respectively. For further details on this decomposition, please refer to Strang [29].

As a common practice, where a CN exceeding 1000 provides evidence for the existence of a severe multicollinearity, $100 < CN < 1000$ shows strong multicollinearity among the columns (i.e. variables) of data matrix [26].

3.3. Solutions to Multicollinearity

Various methods have been proposed in the literature to deal with the multicollinearity. Although the first recommended approach is to collect additional data, this may not always be possible due to the economic reasons or being impossible. The second is to use alternative approaches (like ridge, liu, lasso and elastic net regression) that do not rely on the calculation of least squares. Third, it is to redefine the model by creating new or groups of variables depending on the multi-correlated variables [26]. Finally, various pre-processing methods including centering, scaling, normalization, and standardization are applied as more common approaches for multicollinearity or other problems (like outliers) in the field of machine learning. For this study, we will focus on alternative models (like ridge, liu, lasso and elastic net regression).

4. THE OVERVIEW OF MODELS

4.1. Regression Models

Regression analysis is one of the major areas in machine learning and has been widely used for different learning tasks in various disciplines due to some superior properties like simplicity, interpretability and easy integrability. In a classical linear regression model can be expressed in a matrix notation as

$$y = X\beta + \varepsilon, \quad i = 1, 2, \dots, n \quad (5)$$

where y is an $(n \times 1)$ vector of the response variable, X is an $(n \times p)$ matrix of explanatory variables, β is a $(p \times 1)$

vector of coefficients to be estimated and $\boldsymbol{\varepsilon}$ is a $(n \times 1)$ vector of random errors.

The $\hat{\boldsymbol{\beta}}$ via OLS estimator can be obtained by using simple algebra as

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (6)$$

The ridge estimator, which is the most well-known and used in machine learning and data-oriented studies as an alternative to the OLS estimator, is proposed by Hoerl and Kennard [7] is defined as follows:

$$\hat{\boldsymbol{\beta}}_R = (\mathbf{X}^T \mathbf{X} + k \mathbf{I}_p)^{-1} \mathbf{X}^T \mathbf{y} \quad , \quad k \geq 0 \quad (7)$$

where k is called as ridge tuning parameter and \mathbf{I}_p shows the identity matrix of dimension p . Ridge estimator deals with the multicollinearity problem by adding a small positive term (k) to the diagonal elements of $\mathbf{X}^T \mathbf{X}$ matrix. For a positive optimal k value, the ridge estimator may provide better results than ordinary least squares. In short, ridge estimator improves OLS in terms of prediction accuracy and stability of coefficients for a certain amount of increasing on the bias.

Liu estimator was proposed by Liu [8] as an alternative to the ridge-type estimator to deal with multicollinearity. Although the idea behind the Liu estimator is similar in terms of shrinking the estimated with a small constant (i.e. Liu tuning parameter), the form of Liu tuning parameter in Liu estimator has a linear form, unlike the non-linear form in ridge estimator. The result of this situation is to be able to calculate easier and faster the Liu tuning parameter than the ridge tuning parameter. Another advantage of the Liu estimator over the ridge estimator is to be able to select the appropriate tuning parameter. The general form of the Liu estimator is given as

$$\hat{\boldsymbol{\beta}}_{Liu}^{(d)} = (\mathbf{X}^T \mathbf{X} + \mathbf{I}_p)^{-1} (\mathbf{X}^T \mathbf{y} + d \hat{\boldsymbol{\beta}}) \quad , \quad 0 < d < 1 \quad (8)$$

where d refers to the Liu tuning parameter and $\hat{\boldsymbol{\beta}}$ is the OLS estimator.

The least absolute shrinkage and selection operator (Lasso) is proposed by Tibshirani [9] in order to obtain a more predictive and sparse solution than OLS and ridge by carrying out variable selection. Lasso estimator is defined as

$$\hat{\boldsymbol{\beta}}_{Lasso} = \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 \quad (9)$$

$$\text{subject to} \quad \sum_{j=1}^p |\beta_j| \leq t$$

where t corresponds to the bound on the sum of the absolute values of coefficients and corresponding the upper limit of maximum size for expanding.

$\hat{\boldsymbol{\beta}}_{Lasso}$ can be also written in Lagrangian form as:

$$\hat{\boldsymbol{\beta}}_{Lasso} = \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j| \quad (10)$$

where λ is the Lagrangian multiplier.

Zou and Hastie [10] proposed the elastic net as a regularization and variable selection method. In the elastic net, the superiorities of both ridge and Lasso methods have been used in a unified model. Thus, an effective variable selection process can be carried out by considering the grouping effect (the relationships between variables). The naive elastic net estimator proposed by Zou and Hastie [8] is defined on a standardized data set as follows:

$$\hat{\boldsymbol{\beta}}_{ENet} = \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2 \quad (11)$$

where λ_1 and λ_2 are the non-negative constants corresponding to the size of the L_1 norm of the coefficients and the size of L_2 norm of the coefficients, respectively. The solution can be written as a constrained form of the optimization problem as

$$\hat{\boldsymbol{\beta}}_{ENet} = \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 \quad (12)$$

$$\text{subject to} \quad [(1 - \alpha) \sum_{j=1}^p |\beta_j| + \alpha \sum_{j=1}^p \beta_j^2] \leq t$$

where $\alpha \in [0,1]$ and determines the closeness of the solution of the ridge or lasso estimators. For $\alpha = 1$, the solution is equivalent to the ridge solution, and with $\alpha = 0$, the solution is reduced to the lasso solution.

4.2. Machine Learning Models

The machine learning models evaluated in this study can be divided into three subgroups: (i) Tree-based (CART, Bagging, Random Forests, Extreme Gradient Boosting), (ii) Kernel-based (Support Vector Machines), (iii) Instance-based (KNN), (iv) Splines-based (Cubist, MARS) and (v) Neural Networks-Based (Multilayer Perceptron). This section presents the main characteristics of each of these algorithms.

4.3. K-Nearest Neighbor (KNN)

K-Nearest Neighbors [30, 31] regression is a popular non-parametric supervised machine learning approach used for predicting continuous target variable based on the similarity of data points in a feature (i.e. attributes) space. It is based on identifying the k nearest neighbors to a given point and averaging the values of these neighbors to determine the prediction value of that point. This procedure can inherently be adapted easily to both classification and regression tasks. KNN regression requires the selection of the distance metric and the hyperparameter k (number of neighbors) which are generally found via cross-validation techniques. The local neighborhood of data points in the feature space is considered by the algorithm in an effort to minimize prediction error.

4.4. Support Vector Machines (SVM)

Support Vector Machines [32, 33] regression is a robust and flexible machine learning technique and has a goal to discover a hyperplane that maximizes the distance between the data points and the regression line. It performs this maximization step with observations called support vectors,

which represent a very small subset of observations. SVM assumes a hyperplane as follows:

$$f(x) = \langle w, x \rangle + b \quad (13)$$

where $f(x)$ corresponds to the target, w for the weight vector determining the direction of the hyperplane, x for the feature vector and b is the bias term used for the location of the hyperplane. The objective function is defined as:

$$\min_{w,b,\epsilon} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \epsilon_i \quad (14)$$

with subject to the constraints:

$$y_i - \langle w, x_i \rangle - b \leq \epsilon_i, \quad \epsilon_i \geq 0, \quad i = 1, \dots, n \quad (15)$$

The trade-off between maximizing the margin and minimizing the error is determined by the hyperparameter C . SVM also gives the capability to explore possible non-linear relationships by transforming the data into a higher dimensional space with various functions called kernel functions (linear, polynomial, radial basis etc.).

4.5. Classification and Regression Trees (CART)

The Classification and Regression Trees [34] algorithm is well-known and fundamental machine learning algorithm that is a member of the decision tree family. In the context of regression, CART builds a binary tree structure where each internal node indicates an evaluation based on a particular attribute and each leaf node provides a prediction for the target variable.

The objective of CART regression is to split the feature space in such a way that the resulting tree minimizes the sum of squared differences between the predicted and actual target values. This recursive partitioning procedure is carried out until predefined stopping requirements have been met, such as a maximum tree depth or a specified number of data points in each leaf node.

4.6. Bagging

Bootstrap aggregation [35], often known as bagging, is an ensemble learning technique that seeks to enhance the prediction performance and robustness of regression models by the combination of numerous base models. Using bootstrapped subsets of the original training data, a set of separate and independently trained regression models—often decision trees (i.e. CART) or linear regressors—is built in bagging regression. These subsets are generated through random sampling with replacement, allowing certain data points to be included more than once while leaving out others. The predictions of these basis models are then combined to provide the final prediction. Each base model is trained on one specific subset.

4.7. Random Forests (RF)

The concept of bagging is extended to decision trees by the efficient ensemble learning technique known as Random

Forests [36]. Random Forests employs a group of decision tree regressors in the context of regression to generate accurate predictions. Similar to Bagging, multiple decision trees are built independently on bootstrapped subsets of the training data. Different from the bagging, only a random subset of attributes is taken into consideration for splitting at each split node of a tree. The final regression prediction is calculated by averaging the predictions of all the individual and decorrelated trees.

The strength of random forests regression algorithm appears in its ability to combine the interpretability of decision trees with the predictive power of ensemble learning. Each decision tree in the ensemble discovers a certain amount of the patterns in the data, and the predictions from every single tree collectively generate a more precise and stable prediction for the target variable.

4.8. Extreme Gradient Boosting (XGBoost)

The XGBoost [37] is a fundamental and state-of-the-art machine learning algorithms which can be used for both regression and classification tasks. It builds an ensemble of decision trees in a sequential manner (referring to the boosting), with each new tree being built to address the errors of the previous trees. Gradient descent optimization is employed to minimize the specific loss function (such as mean squared error) in each tree.

In order to improve model performance and training effectiveness, XGBoost incorporates a variety of novel strategies, such as a regularized objective function weighted quantile sketching and optimal feature splitting. With the comprehensive flexibility, XGBoost allows over hyperparameters, practitioners can customize the model's performance to achieve more versatile, robust and scalable results.

4.9. Cubist

The Cubist [38-40] algorithm for regression is a cutting-edge and effective algorithm that does exceptionally in capturing complex non-linear relationships in data while providing interpretability models. It integrates components of rule-based modeling and regression trees to produce a hybrid ensemble of regression model. It utilizes an innovative approach through the development of several models (including linear regressions, decision trees or rule-based learners), each of which focuses on various aspects of the structure of the data. In this way, Cubist stands out as a practical tool due to its interpretability and accurate predictions.

4.10. The Multivariate Adaptive Regression Splines (MARS)

MARS [41] is based on generating the piecewise-linear models via the elements of linear regression and decision trees. By splitting the input space into pieces and fitting linear models within each piece, MARS models have the ability to capture non-linear relationships. The approach can adaptively build the complexity of the model according to the data by selecting appropriate features and generating

basis functions via a forward and backward stepwise procedure. These features enhance the algorithm's ability to deal with noisy outliers and outliers and to capture interactions between attributes.

4.11. Multilayer Perceptron (MLP)

The MLP [42] algorithm is an artificial neural network architecture extending the principles of feedforward neural networks to capture complex relationships between input features and target variable. Each layer is in capable of handling and modifying the input data, and it is constructed up of several interconnected layers of artificial neurons. In order to minimize the difference between the predicted and actual target values, the network learns the ability to adjust its internal parameters, such as weights and biases, during training.

5. EXPERIMENTEL PROCESS AND SETTINGS

5.1. Data Description

In the study, two different datasets (Body Fat and Cancer) are utilized for the model comparison. The Body Fat set

originally shared by Johnson [43] and downloaded a commonly used database [44] for machine learning studies. The data set includes thirteen anthropometric measurements as explanatory variables and body fat percentage as the response variable belong to 252 individuals. Cancer data was retrieved from a publicly available database [45] to estimate the percentage of mortality due to cancer based on ten different variables belonging to 3047 individuals.

The datasets are investigated for multicollinearity by using the diagnostic methods presented in Section 3 and results are given in Tables 1-2. According to the results of Body Fat data, it can be said that there is a problem of multicollinearity in the data due to the presence of variables (weight, abdomen and hip) below the tolerance value of 0.1 and above the VIF value of 5. In addition, the condition value (CN:527.95) corresponds to a strong level of multicollinearity. Likewise, a similar interpretation can be drawn for the Cancer data as the condition number is 1265.84 and the VIF values calculated for some variables (such as avgAnnCount, popEst2015, PercentMarried) is greater than 5. The correlation analysis results given in Figures 1-2 also support the findings that there are high relationships between the variables.

Table 1. The results of multi-collinearity diagnostics of the Body Fat data

Variables	Tolerance	VIF	Symbol	Eigenvalue	Condition Index	CN
Age	0.4444	2.2505	λ_1	0.0732	13.7779	527.9409
Weight	0.0298	33.5093	λ_2	0.0206	25.9354	
Height	0.5972	1.6746	λ_3	0.0038	60.3083	
Neck	0.2312	4.3245	λ_4	0.0031	66.8449	
Chest	0.1057	9.4609	λ_5	0.0026	73.0770	
Abdomen	0.0850	11.7671	λ_6	0.0020	82.5060	
Hip	0.0676	14.7965	λ_7	0.0016	93.7259	
Thigh	0.1286	7.7779	λ_8	0.0012	107.3558	
Knee	0.2168	4.6121	λ_9	0.0008	134.3121	
Ankle	0.5241	1.9080	λ_{10}	0.0006	148.5825	
Biceps	0.2763	3.6197	λ_{11}	0.0006	155.0220	
Forearm	0.4561	2.1925	λ_{12}	0.0005	172.3598	
Wrist	0.2961	3.3775	λ_{13}	0.0001	316.5741	

Table 2. The results of multi-collinearity diagnostics of the Cancer data

Variables	Tolerance	VIF	Symbol	Eigenvalue	Condition Index	CN
avgAnnCount	0.129	7.73	λ_1	1.6456	2.2701	1265.8461
incidenceRate	0.935	1.07	λ_2	0.4700	4.2478	
medianIncome	0.276	3.63	λ_3	0.1823	6.8209	
popEst2015	0.130	7.71	λ_4	0.0973	9.3374	
povertyPercent	0.208	4.81	λ_5	0.0623	11.6623	
MedianAge	0.993	1.01	λ_6	0.0262	17.9836	
AvgHouseholdSize	0.809	1.24	λ_7	0.0204	20.4101	
PercentMarried	0.161	6.20	λ_8	0.0124	26.1204	
PctMarriedHouseholds	0.186	5.39	λ_9	0.0022	61.5450	
BirthRate	0.944	1.06	λ_{10}	0.0013	81.0914	

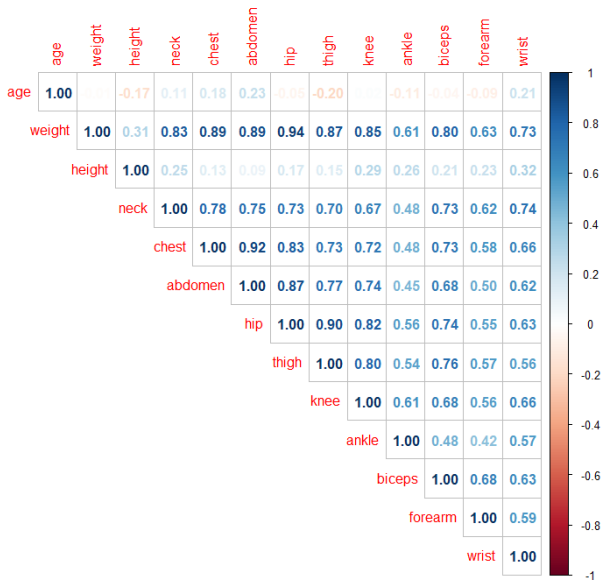


Figure 1. The correlation values between attributes for Body Fat data

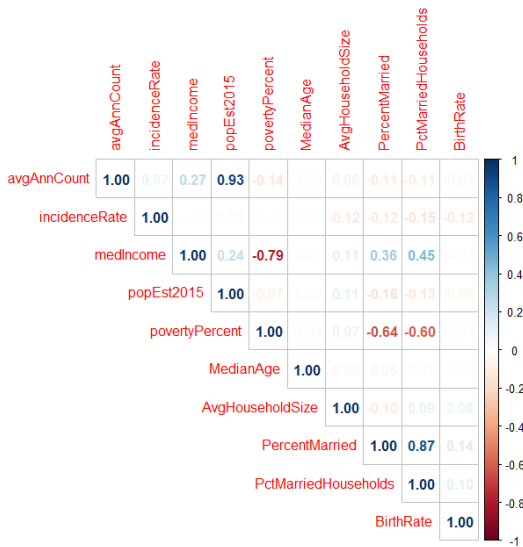


Figure 2. The correlation values between attributes for Cancer data

5.2. Performance Metrics

In regressional studies of machine learning, the most common performance metrics can be given as (i) Root mean squared error, (ii) Mean absolute error and (iii) R square (the coefficient of determination). Each of this metric is calculated based on the difference between the target value (t_i) and the predicted value (y_i) by the model as follows:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (t_i - y_i)^2} \tag{16}$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |t_i - y_i| \tag{17}$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (t_i - y_i)^2}{\sum_{i=1}^n (t_i - \bar{t})^2} \tag{18}$$

Among these metrics, while lower value of RMSE or MAE indicates a better model, higher values of R^2 provides a more explainable model. Due to the nature of the mathematical definition, the range of RMSE or MAE is $[0, \infty]$ and $[0, 1]$ for R^2 .

5.3. Preprocessing and Parameter Tuning

In the study, the data were preprocessed before performing an analysis. Initially, the data was centered to smooth out the high variability in the data structure. The data set is split into approximately seventy five percent as training data and twenty-five percent as test data. Statistical and machine learning models were built on the training data and the generalization performance was measured on the test data. During the training of the models, five-times ten-fold cross validation technique was used. The training performance of the models was calculated separately for each of the cross-validation data and the overall average was computed. A grid search space of twenty parameters was utilized to determine the model parameters. The details of parameters are given in Table 3. Each possible combination of parameters was trained by cross validating the models and the models with the best parameters were extracted. The performance (RMSE, MAE and R-Square values) of each model on the test data was calculated using the optimum parameter values.

Table 3. The ranges of parameters corresponding to each model

Model	Range of Parameters
CART	cost-complexity: [-10, -1] min_n: [2, 40]
Cubist	committees: [1, 100] neighbors: [0, 9]
Elastic Net	penalty: [-10, 0]
KNN	neighbors: [1, 15] dist_power: [0.1, 2]
Lasso	penalty: [-10, 0]
Liu	d: [0, 1]
MARS	prod_degree: [1, 2]
MLP	hidden_units: [1, 10] penalty: [-10, 0] epochs: [10, 1000]
RF	mtry: [1, 13] min_n: [2, 40]
Ridge	penalty: [-10, 0]
SVM (Poly)	cost: [-10, 5] degree: [1, 3]
SVM (Radial)	cost: [-10, 5] rbf_sigma: [-10, 0]
XGBoost	trees: [1, 2000] min_n: [2, 40] tree_depth: [1, 15] learn_rate: [-3, -0.5] loss_reduction: [-10, 1.5] sample_size: [0.1, 1]

Table 4. Performance comparisons of models corresponding to the train data set

Model	Data	RMSE	Mae	R ²
Bagging (CART)	Body Fat	4.672	3.956	0.683
	Cancer	8.176	6.144	0.458
CART	Body Fat	5.343	4.357	0.643
	Cancer	8.987	6.807	0.370
Cubist	Body Fat	4.525	3.795	0.763
	Cancer	8.095	6.118	0.468
Elastic Net	Body Fat	4.446	3.694	0.773
	Cancer	8.360	6.388	0.444
KNN	Body Fat	5.072	4.169	0.689
	Cancer	8.427	6.360	0.427
Lasso	Body Fat	4.451	3.677	0.773
	Cancer	8.374	6.409	0.431
Liu	Body Fat	4.111	3.398	0.818
	Cancer	7.825	6.007	0.512
MARS	Body Fat	4.423	3.506	0.715
	Cancer	8.138	6.093	0.463
MLP	Body Fat	6.005	4.814	0.500
	Cancer	9.971	7.634	0.182
RF	Body Fat	4.640	3.846	0.735
	Cancer	8.010	6.018	0.482
Ridge	Body Fat	4.668	3.756	0.686
	Cancer	8.373	6.413	0.431
SVM (Poly)	Body Fat	4.599	3.832	0.760
	Cancer	8.252	6.159	0.447
SVM (Radial)	Body Fat	4.668	3.664	0.699
	Cancer	8.467	6.458	0.429
XGBoost	Body Fat	4.736	3.901	0.725
	Cancer	7.949	5.966	0.488

Table 5. Performance comparisons of models corresponding to the testing data set

Model	Data	RMSE	Mae	R ²
Bagging (CART)	Body Fat	4.690	3.861	0.730
	Cancer	8.372	6.227	0.437
CART	Body Fat	5.855	4.942	0.547
	Cancer	9.037	6.892	0.363
Cubist	Body Fat	4.708	3.834	0.680
	Cancer	8.530	6.328	0.421
Elastic Net	Body Fat	4.530	3.569	0.701
	Cancer	9.000	6.668	0.364
KNN	Body Fat	5.071	4.251	0.634
	Cancer	8.611	6.501	0.406
Lasso	Body Fat	4.511	3.547	0.704
	Cancer	8.999	6.670	0.363
Liu	Body Fat	4.442	3.581	0.789
	Cancer	7.968	6.174	0.432
MARS	Body Fat	4.895	3.969	0.722
	Cancer	8.586	6.340	0.417
MLP	Body Fat	6.134	4.956	0.545
	Cancer	10.386	7.875	0.135
RF	Body Fat	4.6325	3.912	0.689
	Cancer	8.257	6.182	0.452
Ridge	Body Fat	4.689	3.816	0.750
	Cancer	8.971	6.664	0.363
SVM (Poly)	Body Fat	4.708	3.849	0.680
	Cancer	8.808	6.667	0.380
SVM (Radial)	Body Fat	4.782	3.939	0.725
	Cancer	9.160	6.513	0.342
XGBoost	Body Fat	4.9579	4.006	0.661
	Cancer	8.340	6.188	0.443

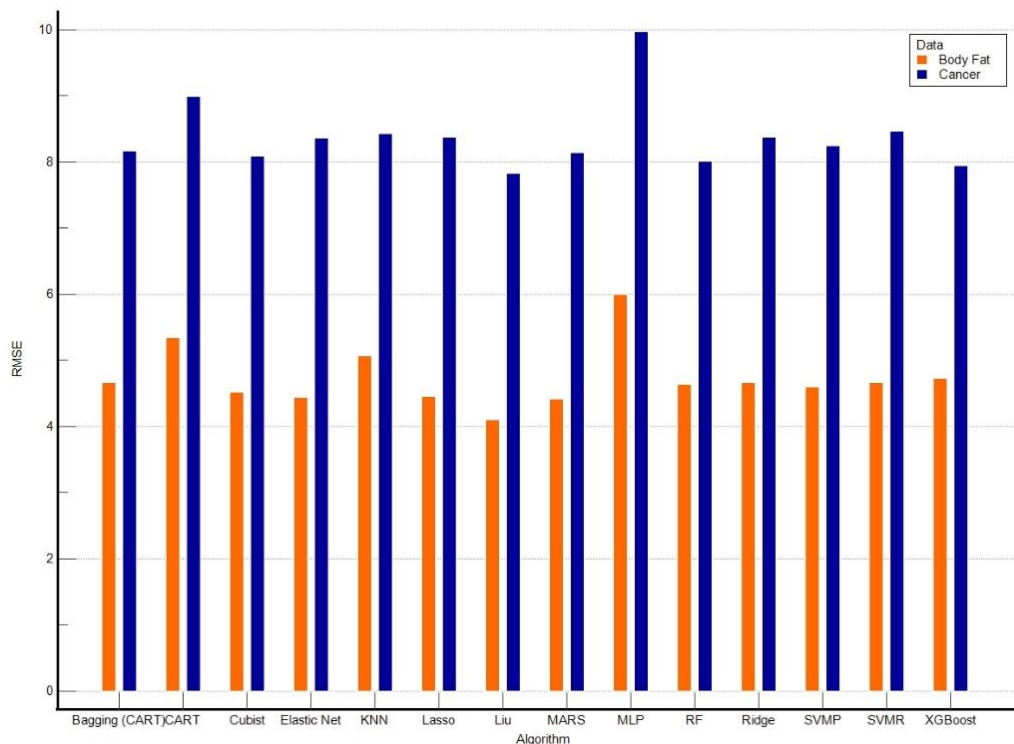


Figure 3. The visual representation of testing RMSE results for Body Fat and Cancer data sets

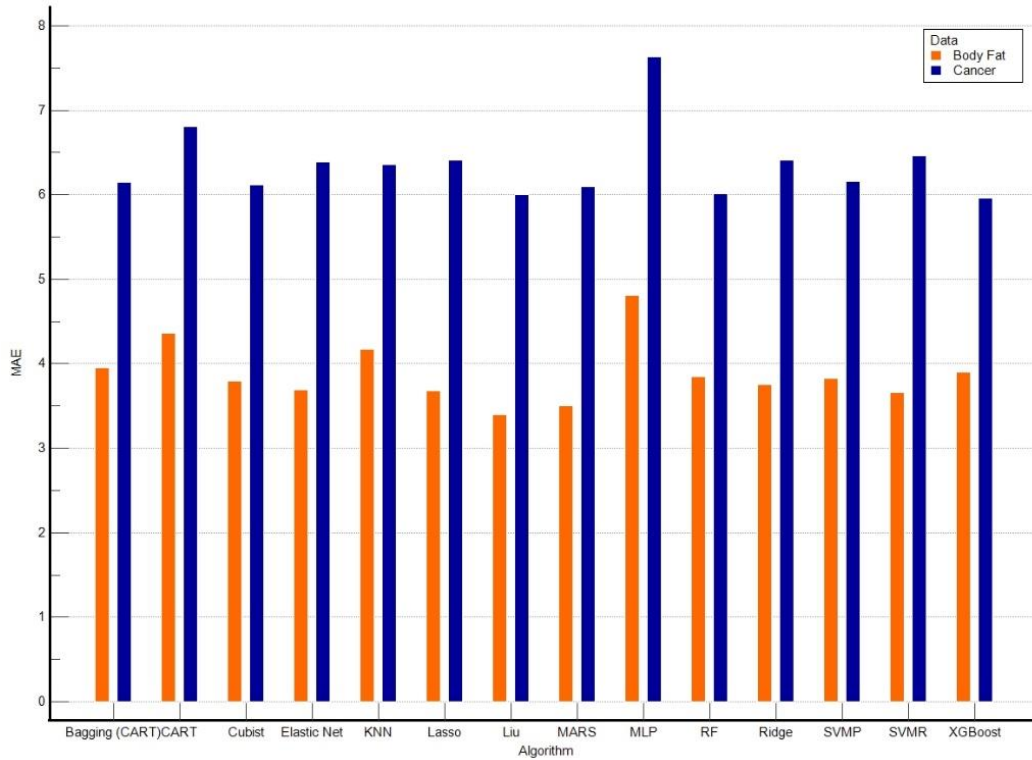


Figure 4. The visual representation of testing MAE results for Body Fat and Cancer data sets

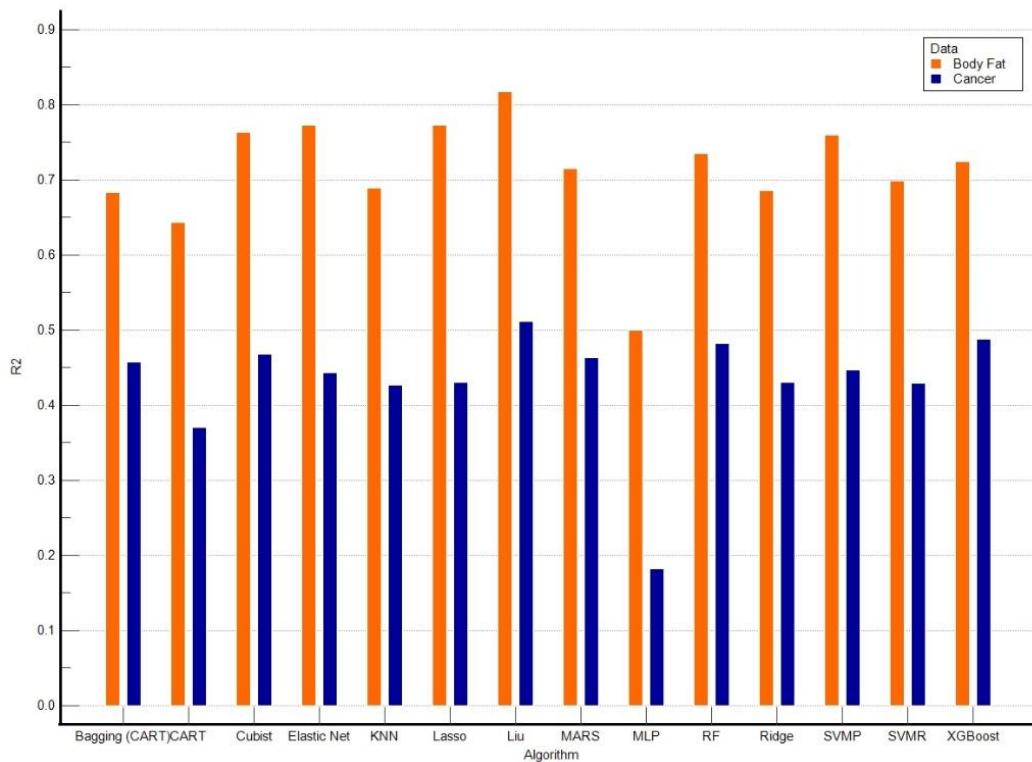


Figure 5. The visual representation of testing R-squared results for Body Fat and Cancer data sets

6. RESULTS

This section presents the results of thirteen different models including statistical and machine learning models. The comparison results of these models for two datasets are given in Table 4 for training data and Table 5 for test data. A visual representation of the performance values corresponding to the test data is given in Figures 3-5.

When the training performances of the models are evaluated for Body Fat data, the Liu regression model corresponding to the optimal parameter found as $d = 0.78$ (RMSE=4.111, MAE=3.398 and R-square=0.818) performed the best in each of the three performance criteria for this type of data set. In general, statistical models tend to perform better than machine learning models. The MLP, KNN and CART models showed the poorest performances, respectively.

Regarding the training performances for cancer data, Liu regression stands out in the RMSE (7.825) and R-square (0.512) criteria with $d = 0.86$ parameter value, while XGBoost is the best algorithm in the MAE (5.966) criterion. MLP and CART methods perform the weakest.

According to the test results, regression model was found to be the best model based on RMSE (4.442) and R-square (0.789) values, while Lasso regression model in the MAE (3.547) criterion. Liu, Lasso and Elastic Net regression models performed relatively close to each other but better than the remaining algorithms. Similar to the training performance, CART, KNN and MLP models showed weaker performance on the test data. Liu regression is superior to other algorithms in RMSE (7.968) and MAE (6.174) criteria in cancer data, whereas Random Forest algorithm has more generalizable performance with respect to R-square (0.452) value only. The SVM (with Radial kernel) and MLP algorithms were found to be the relatively weakest algorithms in the test performances.

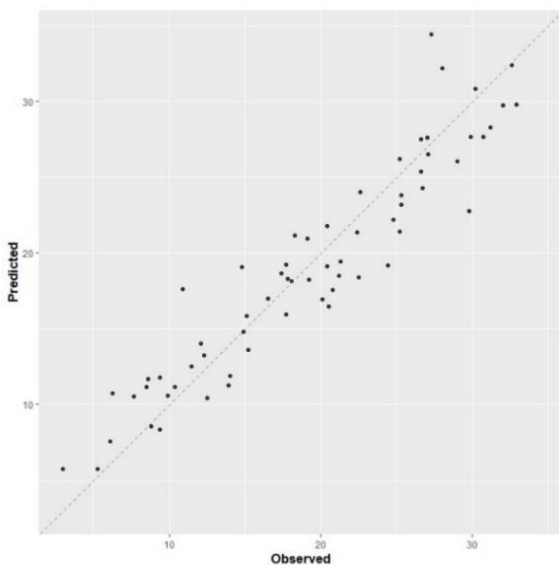


Figure 6. The scatter plot of observed and predicted value based on Liu regression testing results for Body Fat data

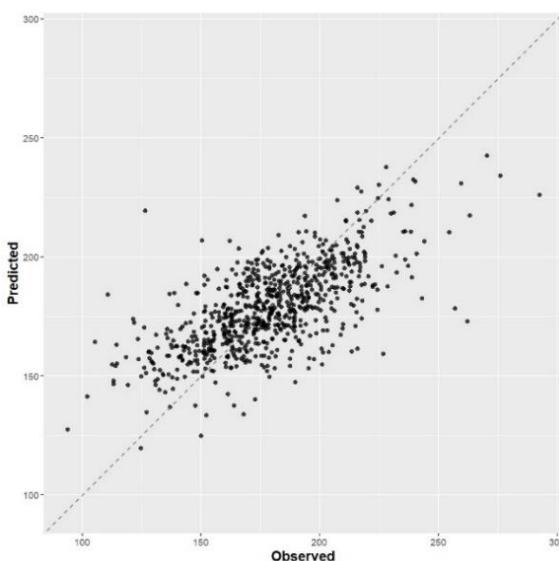


Figure 7. The scatter plot of observed and predicted value based on Liu regression testing results for Cancer data

The scatter plot generated to assess the fit between the predicted values of the Liu regression model and the actual response values is presented in Figure 6. According to this plot, it can be stated that the model predicts the actual response sufficiently well and provides values that are close to the actual values. The scatter plot provided in Figure 7 for cancer data confirms the similar interpretation and indicates that there is a strong fit between actual mortality rates and model predictions.

In the literature, it can be noted that the studies conducted on the body fat data are mainly along two directions: (i) Only statistics-oriented studies [46-47] and (ii) Studies based on a subset or different set of the data [48-51].

However, the results of the study reveal encouraging findings compared to the directly relevant studies in the literature. In the study conducted by Uçar et al. [52], our study produced better results compared to the comparison including artificial neural networks, support vector machines and decision trees algorithms (minimum RMSE=4.264 and $R^2=0.616$). Additionally, the proposed approach is superior (RMSE=4.6384, MAE=3.6974) in comparison to the study carried out by Shao [53], applying multiple linear regression, artificial neural networks, MARS and support vector machines algorithms on the same data (RMSE=4.6384, MAE=3.6974).

The application of machine learning algorithms in cancer research is widespread, with a focus on cancer prognosis and prediction, incidence rates and survival prediction. Carrizosa et al. [54] developed a novel tree based linear regression model on the identical data, only focusing on hierarchical categorical variables. This study is not directly comparable as it does not incorporate different algorithms, but the prominent algorithm types are similar to some alternative studies [55-58]. However, it can be said that Liu regression is a promising and powerful alternative for future studies as it is one of the first examples of Liu regression in cancer studies as far as we know [59-60].

7. CONCLUSION AND FUTURE WORKS

In this study, the problem of multicollinearity is addressed, and comparative results of statistical and machine learning models based on two different healthcare datasets are presented. The models were trained using the cross-validation method and their generalization and prediction performances were assessed on an independent test data set.

The results of the study show that statistical models outperform for the data set suffering from multicollinearity problem, particularly Liu regression, complex machine learning models in both training and testing performance. However, the study encounters two key limitations. Firstly, it is critical that the degree of multicollinearity (weak, strong or extreme) is correctly identified and taken into account in the comparison process. Secondly, the tuning parameters (penalty parameters) of statistical methodologies have the potential to affect model performances by choosing them more accurately through analytical approaches rather than searching within a certain range. A more extensive study

taking into account these two limitations may yield valuable results in the forthcoming studies.

Consequently, the choice of the appropriate method is critical, given that the problem of multicollinearity is widespread in real-life applications. Therefore, it can be concluded that statistical models are powerful tools with effective solutions to the problem of multicollinearity.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study has received no financial support.

REFERENCES

- [1] Ortiz, R., Contreras, M., & Mellado, C. (2023). Regression, multicollinearity and Markowitz. *Finance Research Letters*, 58, 104550.
- [2] Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica: Journal of the Econometric Society*, iii-115.
- [3] Chan, J. Y. L., Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z. W., & Chen, Y. L. (2022). Mitigating the multicollinearity problem and its machine learning approach: a review. *Mathematics*, 10(8), 1283.
- [4] A. Garg and K. Tai, 'Comparison of statistical and machine learning methods in modelling of data with multicollinearity', *IJMIC*, vol. 18, no. 4, p. 295, 2013, doi: 10.1504/IJMIC.2013.053535.
- [5] C. M. Stein, 'Multiple regression contributions to probability and statistics', *Essays in Honor of Harold Hotelling*, vol. 103, 1960.
- [6] C. M. Stein, 'Confidence sets for the mean of a multivariate normal distribution', *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 24, no. 2, pp. 265–285, 1962.
- [7] A. E. Hoerl and R. W. Kennard, 'Ridge Regression: Applications to Nonorthogonal Problems', *Technometrics*, vol. 12, no. 1, pp. 69–82, Feb. 1970, doi: 10.1080/00401706.1970.10488635.
- [8] L. Kejian, 'A new class of biased estimate in linear regression', *Communications in Statistics - Theory and Methods*, vol. 22, no. 2, pp. 393–402, Jan. 1993, doi: 10.1080/03610929308831027.
- [9] R. Tibshirani, 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, 1996.
- [10] H. Zou and T. Hastie, 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, 2005.
- [11] G. Li and P. Niu, 'An enhanced extreme learning machine based on ridge regression for regression', *Neural Computing and Applications*, vol. 22, pp. 803–810, 2013.
- [12] T. E. Panduro and B. J. Thorsen, 'Evaluating two model reduction approaches for large scale hedonic models sensitive to omitted variables and multicollinearity', *Letters in Spatial and Resource Sciences*, vol. 7, pp. 85–102, 2014.
- [13] G. G. Dumancas and G. Bello, 'Comparison of machine-learning techniques for handling multicollinearity in big data analytics and high-performance data mining', in *SC15: The International Conference for High Performance Computing Networking Storage and Analysis*, 2015, pp. 41–42.
- [14] B. Kilinc, B. Aşıkçıl, A. Erar, and B. Yazıcı, 'Variable selection with genetic algorithm and multivariate adaptive regression splines in the presence of multicollinearity', *International Journal of Advanced and Applied Sciences*, vol. 3, no. 12, 2016.
- [15] A. Katrutsa and V. Strijov, 'Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria', *Expert Systems with Applications*, vol. 76, pp. 1–11, Jun. 2017, doi: 10.1016/j.eswa.2017.01.048.
- [16] E. Hoseinzade and S. Haratizadeh, 'CNNpred: CNN-based stock market prediction using a diverse set of variables', *Expert Systems with Applications*, vol. 129, pp. 273–285, 2019.
- [17] J.-M. Kim, N. Wang, Y. Liu, and K. Park, 'Residual control chart for binary response with multicollinearity covariates by neural network model', *Symmetry*, vol. 12, no. 3, p. 381, 2020.
- [18] C. P. Obite, N. P. Olewuezi, G. U. Ugwuanyim, and D. C. Bartholomew, 'Multicollinearity Effect in Regression Analysis: A Feed Forward Artificial Neural Network Approach', *Asian Journal of Probability and Statistics*, pp. 22–33, Jan. 2020, doi: 10.9734/ajpas/2020/v6i130151.
- [19] Hua, Y. (2020, May). An efficient traffic classification scheme using embedded feature selection and lightgbm. In *2020 Information Communication Technologies Conference (ICTC)* (pp. 125-130). IEEE.
- [20] Qaraad, M., Amjad, S., Manhrawy, I. I., Fathi, H., Hassan, B. A., & El Kafrawy, P. (2021). A hybrid feature selection optimization model for high dimension data classification. *IEEE Access*, 9, 42884-42895.
- [21] Y. Bi, C. Li, Y. Benezeth, and F. Yang, 'Impacts of multicollinearity on CAPT modalities: An heterogeneous machine learning framework for computer-assisted French phoneme pronunciation training', *Plos one*, vol. 16, no. 10, p. e0257901, 2021.
- [22] A. Abubakar, U. F. Abbas, and K. E. Lasisi, 'Remedying Multicollinearity in Quantitative Analysis: A Simulation Studies', 2022.
- [23] Mahadi, M., Ballal, T., Moinuddin, M., & Al-Saggaf, U. M. (2022). A recursive least-squares with a time-varying regularization parameter. *Applied Sciences*, 12(4), 2077.

- [24] Kaneko, H. (2023). Interpretation of Machine Learning Models for Data Sets with Many Features Using Feature Importance. *ACS omega*, 8(25), 23218-23225.
- [25] Genç, M. (2024). An Enhanced Extreme Learning Machine Based on Square-Root Lasso Method. *Neural Processing Letters*, 56(1), 5.
- [26] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2021.
- [27] D. A. Belsley, E. Kuh, and R. E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, 1st ed. in *Wiley Series in Probability and Statistics*. Wiley, 1980. doi: 10.1002/0471725153.
- [28] S. Weisberg, *Applied Linear Regression*, 1st ed. in *Wiley Series in Probability and Statistics*. Wiley, 2005. doi: 10.1002/0471704091.
- [29] Strang, G. (2022). *Introduction to linear algebra*. Wellesley-Cambridge Press.
- [30] E. Fix and J. L. Hodges, 'Discriminatory analysis. Nonparametric discrimination: Consistency properties', *International Statistical Review/Revue Internationale de Statistique*, vol. 57, no. 3, pp. 238–247, 1989.
- [31] N. S. Altman, 'An introduction to kernel and nearest-neighbor nonparametric regression', *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.
- [32] C. Cortes and V. Vapnik, 'Support-vector networks', *Machine learning*, vol. 20, pp. 273–297, 1995.
- [33] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik, 'Support vector regression machines', *Advances in neural information processing systems*, vol. 9, 1996.
- [34] B. Li, J. Friedman, R. Olshen, and C. Stone, 'Classification and regression trees (CART)', *Biometrics*, vol. 40, no. 3, pp. 358–361, 1984.
- [35] L. Breiman, 'Bagging predictors', *Machine learning*, vol. 24, pp. 123–140, 1996.
- [36] L. Breiman, 'Random forests', *Machine learning*, vol. 45, pp. 5–32, 2001.
- [37] T. Chen and C. Guestrin, 'Xgboost: A scalable tree boosting system', in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
- [38] J. R. Quinlan, 'Learning with continuous classes', in *5th Australian joint conference on artificial intelligence*, World Scientific, 1992, pp. 343–348.
- [39] J. R. Quinlan, 'Combining instance-based and model-based learning', in *Proceedings of the tenth international conference on machine learning*, 1993, pp. 236–243.
- [40] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.
- [41] J. H. Friedman, 'Multivariate adaptive regression splines', *The annals of statistics*, vol. 19, no. 1, pp. 1–67, 1991.
- [42] S. Haykin and N. Network, 'A comprehensive foundation', *Neural networks*, vol. 2, no. 2004, p. 41, 2004.
- [43] R. W. Johnson, 'Fitting percentage of body fat to simple body measurements', *Journal of Statistics Education*, vol. 4, no. 1, 1996.
- [44] 'Kaggle: Your Machine Learning and Data Science Community'. <https://www.kaggle.com/> (accessed Sep. 22, 2023).
- [45] 'Data World. <https://data.world/nrippner/cancer-trials> (accessed July. 18, 2024).
- [46] Frankenfield, D. C., Rowe, W. A., Cooney, R. N., Smith, J. S., & Becker, D. (2001). Limits of body mass index to detect obesity and predict body composition. *Nutrition*, 17(1), 26-30.
- [47] Fthenakis, Z. G., Balaska, D., & Zafirooulos, V. (2012). Uncovering the FUTREX-6100XL prediction equation for the percentage body fat. *Journal of medical engineering & technology*, 36(7), 351-357.
- [48] Deurenberg, P., Weststrate, J. A., & Seidell, J. C. (1991). Body mass index as a measure of body fatness: age-and sex-specific prediction formulas. *British journal of nutrition*, 65(2), 105-114.
- [49] Jackson, A. S., Stanforth, P. R., Gagnon, J., Rankinen, T., Leon, A. S., Rao, D. C., ... & Wilmore, J. H. (2002). The effect of sex, age and race on estimating percentage body fat from body mass index: The Heritage Family Study. *International journal of obesity*, 26(6), 789-796.
- [50] Meeuwssen, S., Horgan, G. W., & Elia, M. (2010). The relationship between BMI and percent body fat, measured by bioelectrical impedance, in a large adult sample is curvilinear and influenced by age and sex. *Clinical nutrition*, 29(5), 560-566.
- [51] Sung, H., & Mun, J. (2017). Development and cross-validation of equation for estimating percent body fat of Korean adults according to body mass index. *Journal of Obesity & Metabolic Syndrome*, 26(2), 122.
- [52] Uçar, M. K., Ucar, Z., Köksal, F., & Daldal, N. (2021). Estimation of body fat percentage using hybrid machine learning algorithms. *Measurement*, 167, 108173.
- [53] Shao, Y. E. (2014). Body fat percentage prediction using intelligent hybrid approaches. *The Scientific World Journal*, 2014.
- [54] Carrizosa, E., Mortensen, L. H., Morales, D. R., & Sillero-Denamiel, M. R. (2022). The tree based linear regression model for hierarchical categorical variables. *Expert Systems with Applications*, 203, 117423.
- [55] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13, 8-17.

- [56] Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, 2, 117693510600200030.
- [57] Parikh, R. B., Manz, C., Chivers, C., Regli, S. H., Braun, J., Draugelis, M. E., ... & O'Connor, N. R. (2019). Machine learning approaches to predict 6-month mortality among patients with cancer. *JAMA network open*, 2(10), e1915997-e1915997.
- [58] Zhu, W., Xie, L., Han, J., & Guo, X. (2020). The application of deep learning in cancer prognosis prediction. *Cancers*, 12(3), 603.
- [59] Yaqoob, A., Musheer Aziz, R., & verma, N. K. (2023). Applications and techniques of machine learning in cancer classification: A systematic review. *Human-Centric Intelligent Systems*, 3(4), 588-615.
- [60] Swanson, K., Wu, E., Zhang, A., Alizadeh, A. A., & Zou, J. (2023). From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. *Cell*, 186(8), 1772-1791.