Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v18n2p264

Imputation of Missing Values with Adaptive Elastic Net for Gene Selection in High-dimensional Data

By Alharthi

15 October 2025

Electronic Journal of Applied Statistical Analysis
Vol. 18, Issue 02, October 2025, 264-281
DOI: 10.1285/i20705948v18n2p264

Imputation of Missing Values with Adaptive Elastic Net for Gene Selection in High-dimensional Data

Aiedh Mrisi Alharthi*[a]

[a]*Department of Mathematics, Turabah University College, Taif University , P.O. Box 11099, Taif 21944, Saudi Arabia*

15 October 2025

Missing data is a problem that often arises in a variety of real-world systems. The performance of classification algorithms operating on these systems would suffer as a result. Effective imputation approaches abound to tackle this issue in case of missing data with low dimensions. In addition, one of the most common methods for concurrently doing variable selection and coefficient estimation in high-dimensional data is the penalized regression technique. However, one of the most significant problems associated with high-dimensional data is that it often includes an enormous quantity of missing data, which means that conventional imputation methods may not adequately address it. This paper proposes the imputation of missing values with the adaptive elastic net as an extension of penalized techniques to enhance gene selection and impute missing values in high-dimensional data. The effectiveness of the proposed method is evaluated by applying it to high-dimensional datasets that are taken from real-world situations with varying numbers of features, sample sizes, and percentages of missing datasets. A comparison is made between the proposed approach and various imputation-penalized methods that are currently in use for high-dimensional data. The findings of the comparison experiments reveal that the proposed technique is superior to its rivals since it achieves a better value for classification accuracy, sensitivity, and specificity than its competitors.

keywords: Missing values, Imputations, Adaptive elastic net, Logistic regression, High-dimensional data.

*Corresponding author: amharthi@tu.edu.sa

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

1 Introduction

Missing data is a widespread problem that arises in most scientific studies, such as biological, epidemiology, and social research. According to diverse sources in the datasets, this missingness has many causes. These involve a lack of measurements, data unavailability, survey non-response, inadequate information, missing files, etc. (Jiang et al., 2020; Rácz and Gere, 2025). Therefore, many statistical methods need the use of entire datasets. Analyses that do not adequately handle missing data might lead to incorrect estimations and inferences (Deng et al., 2016). Thus, several statistical techniques may be used to deal with the issue of missing data. Jiang et al. (2020) claimed that disregarding the observation of missing data is a simple solution to the problem of missing values. Since a few observations have missing data, there is typically no substantial concern. On the contrary, removing many observations with missing data leads to a significant reduction in the amount of data available (Khan and Hoque, 2020; Zhang, 2015). Additionally, it brings about a detrimental effect on the statistical strength and effectiveness of the data (Kwak and Kim, 2017). Imputation creates full data without omitting the missing instances for analysis by filling in the missing data with some suitable values. Therefore, to impute missing values, ad-hoc techniques such as mean replacement, maximum likelihood methods, single imputation, and multiple imputations (MI) may be utilized (Zahid and Heumann, 2019). Because of this, effective imputation strategies are necessary in order to solve the issue of missing data.

The availability of high-dimensional data is an extra issue that often emerges in many scientific research fields, such as ecological sciences, economics, sociological surveys, genetics, machine learning, and health (Zahid et al., 2020). Variable selection is a crucial stage in the analysis of high-dimensional data. In the past decade, there has been a notable increase in the use of gene selection strategies in biological datasets. These datasets frequently include a higher number of features compared to the amount of samples, which may lead to overfitting and negatively affect the learning process. Moreover, from a scientific and epistemological perspective, only a limited set of genes has significant implications and are intricately linked to the corresponding disease (Li et al., 2019). It is possible to think of these issues as a machine learning feature selection problem; hence, finding informative genes is an effective method for addressing these challenges. Over the preceding ten years, there has been significant development in the field of variable selection approaches. Among these approaches are those that are subject to penalties. It is used for the purpose of selecting features and classifying them. The logistic regression has recently received considerable attention. Within this framework, penalized approaches use a particular kind of penalty term to carry out feature selection and classification concurrently. A wide variety of logistic regression models may be used, each with its own set of penalties. The "Least Absolute Shrinkage and Selection Operator" penalty, which is often referred to as "LASSO" is one of these potential penalties (Tibshirani, 1996). According to Tibshirani, LASSO is based on the $L_1 - norm$. The so-called "Smoothly Clipped Absolute Deviation" (SCAD) is yet another disciplinary approach that Fan and Li (2001) had proposed. approaches such as the elastic net (EN) (Zou and Hastie, 2005), the adaptive $L_1 - norm$ (Zou, 2006), and the adaptive elastic

net (AEN) approach are also considered to be additional penalties (Ghosh, 2011; Zou and Zhang, 2009).

It is common practice to use penalized regression to simultaneously carry out variable selection and coefficient estimations in high-dimensional data. Nevertheless, one of the most serious challenges of high-dimensional data is that it often contains a large quantity of missing data. Previous studies have shown that the majority of microarray datasets are incomplete to varying degrees, with the percentages ranging from fifty percent to ninety-five percent each (Chen et al., 2016). With major advancements in techniques, multiple imputation (MI) (Little and Rubin, 2019; Rubin, 1996) and software Su et al. (2011); van Buuren and Groothuis-Oudshoorn (2011) have gained popularity as a method for addressing missing data.

On the other hand, it is essential to note that MI approaches may not be as successful when dealing with high-dimensional data, where the number of variables (p) in the imputation model is much more than the sample size (n) (Deng et al., 2016; Brini and van den Heuvel and, 2024). When the number of variables in the imputation model exceeds the sample size, which is frequently referred to as ($p > n$ or $p \gg n$), the problem becomes more significant, and traditional likelihood estimates are no longer attainable. The low number of variables poses a constraint on the usage of sequential regression imputation in this scenario (Zahid and Heumann, 2019; Zhao and Long, 2016).

Furthermore, Chen and Wang (2013) proposed a new multiple imputation LASSO (MI-LASSO) approach as an extension of the LASSO approach to improving variable selection on multiply imputed data. In this manner, group penalties are applied to the estimated regression coefficients of a single variable across all imputed datasets in the MI-LASSO approach to ensure consistent variable selection across various datasets. Deng et al. (2016) presented and evaluated the use of penalized regression to handle the high dimensionality of imputation models. Besides, as an extension of MI-LASSO and to tackle the issue of variable selection in longitudinal data with missing values, Geronimi and Saporta (2017) introduced a multiple imputation regularized generalized estimating equations method in which multiple imputation is used to deal with missing data, and a group LASSO penalty is used to select variables.

In addition, Zahid and Heumann (2019) proposed a multiple imputation approach, which is abbreviated by "mispr," to overcome the issue of many features with missing data. Therefore, they used sequential regularized regression models, where each variable with missing data is considered to have a specific distributional form and is imputed using its own ridge penalty imputation model. In addition, for the selection of each potential predictor, Zahid et al. (2020) proposed using the magnitude of the parameter estimates overall imputed datasets. Thus, an absolute value restriction is placed on the sum of these estimations to help determine whether the predictor should be included in the model or not. In other words, the heuristic techniques determined whether to include a predictor in the model based on a threshold established on its frequency in the parsimonious models derived from the imputed datasets. They prioritized the absolute magnitudes of the parameter estimations in the M parsimonious models rather than the frequency of occurrence.

Moreover, Wang et al. (2021) improved a penalized learning-based imputation method

to enhance the use of the local structure of microarray data. Hence, they introduced their approach to estimating the missing entries of a target gene with its neighbours, considering the elastic net penalty's concurrent variable selection and grouping effect.Zahid et al. (2021) proposed constructing a robust imputation model by including as many probable candidate variables as possible to achieve consistent and unbiased results utilizing a semi-compatible imputation model. They used the $L_1 - norm$ to accommodate the maximum number of features in the imputation model, and the $L_2 - norm$ penalty was used to fit the resultant model. Recently, there has been growing interest in applying the imputation methods in high-dimensional data to address missing values and perform feature selection simultaneously. Such as Lee and Park (2024) proposed a new technique that estimates the conditional independence structure among variables before the imputation process, Liang et al. (2024) incorporated a novel multiple imputation approach based on matrix completion, Zhang and Kim (2024) introduced an algorithm that used the horseshoe shrinkage prior, and a compositional data imputation method based on the adaptive group LASSO was presented by (Tian et al., 2025).

In situations when high-dimensional data is present, the methods and software packages that are now available for MI may perform much less effectively. Therefore, our proposal involves using the adaptive elastic net to impute missing values in high-dimensional data as an extension of penalized approaches. This approach aims to improve gene selection and imputation of missing values. In order to achieve this, the initial weight employed is the one-dimensional weighted Mahalanobis distance (1-DWM) inside the $L_1 - norm$ in adaptive elastic net regression. This enables the filling in of missing data for every feature. The suggested technique, which is called imputation adaptive elastic net regression (IAENR), is evaluated in comparison to various imputation methods currently in use for high-dimensional data. As a consequence of reaching higher levels of sensitivity and specificity as well as classification accuracy, the IAENR is able to surpass its rivals, as shown by the comparative testing findings. The remaining elements of this article are organized in the way that is described below. Section 2 is devoted to explaining the technique and the materials that were used. In Section 3, the results of the experimental study that was conducted to evaluate the effectiveness of the IAENR method compared to other penalized approaches are provided and discussed. The conclusion of this study is then presented in Section 4.

2 Methods

2.1 Missing Values Imputation

The missing values are one of the most common phenomena in practical research. Analyzing a dataset using conventional statistical methods requires that the complete dataset be analyzed without any missing data. An inappropriate statistical conclusion might be made if relevant data is removed from a study. However, by substituting probable values for the omitted ones, the imputation method completes the set of data without omitting any of the analytically valuable information. According to Little and Rubin (2019); Rubin (1996), three primary cases of missing data mechanisms exist: In the first

case, missing completely at random (MCAR). This is a situation in which missingness does not depend on either observed or missing data. The second case is Missing at Random (MAR), where missingness only depends on observed data but not on missing data, i.e., $P(missing/complete\ data) = P(missing/observed\ data)$. Missing not at random (MNAR) is the third case, where missingness depends on unobserved data, i.e., $P(missing/complete\ data) \neq P(missing/observed\ data)$.

As far as single imputation techniques are concerned, they provide a defined value for a dataset's missing actual value. The computational cost of this approach is cheaper. Hence, the researchers have developed many other methods of single imputation. The most crucial process is studying other responses and choosing the most meaningful one. single imputations may be computed by taking the mean, median, and mode of the variable's observed entries to impute missing values (Khan and Hoque, 2020). In a single imputation, values that are based on an assumption are considered as real values. However, uncertainty is not taken into account in single imputation procedures. Furthermore, these values may be subject to standard errors. Consequently, the findings might be biased Holman and Glas (2005). Other approaches, such as those based on machine learning, may also be used for single imputation (Pelckmans et al., 2005).

In view of that, many simulation models used by multiple imputation (MI) techniques generated different results for the imputation of one missing value. In these procedures, imputed data is used to provide a wide variety of possible outcomes. However, although MI techniques are more complex than single imputation, they do not suffer from the problem of bias values. The MI for missing data may be summed up into three main phases. First, it is necessary to perform imputation in order to produce M-independent imputed values that match missing data. The second phase is to perform the desired analysis on each of the M by utilizing standard, complete data methods. In the third phase, the study's findings are integrated using Rubin's guidelines (Rubin, 2004). to obtain a single set of parameter estimations. Therefore, previous research has introduced packages in the R programming language to implement MI approaches more effectively. "Multivariate Imputation via Chained Equations" (commonly known as "mice") is one of the most popular packages (van Buuren and Groothuis-Oudshoorn, 2011). In addition, the packages "mi" (Su et al., 2011) and "Amelia" (Honaker et al., 2011) are among the others.

2.2 Regularized Logistic Regression Model

The statistical classification method known as logistic regression is used to make predictions about the value of a categorical response variable. This method has just two possible values, which are represented by the numbers 0 and 1. Logistic regression is an effective approach when dealing with low-dimensional data. However, the logistic regression method may become ineffective when dealing with high-dimensional data, such as gene expression data sets. This is because the prediction accuracy and computing efficiency of the technique are poor. The phenomenon known as overfitting, which takes place when the number of predictors is greater than the number of observed values, is another problem that hampers the use of logistic regression (Casella et al., 2013). It is

possible to carry out gene selection and classification concurrently by using logistic regression with a penalty, which is employed in several classification applications. In order to comply with the regularization technique, this model is penalized, and its coefficients are reduced (Li et al., 2020). Recently, penalized regression algorithms have become more popular due to their improved prediction accuracy and more efficient processing capabilities.

To better understand the concept, we suppose that $\boldsymbol{X} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ is a matrix representing a set of data. In this matrix, each column represents a gene, and each row represents a sample. In this example, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ represents the i^{th} input sample, $x_{i,j}$ is the expression value of the j^{th} feature in the i^{th} sample, and $\boldsymbol{y} = (y_1, ..., y_n)^T$ is the response vector. Here, y_i is the corresponding classification label, which can take on the values 0 or 1. For logistic regression, the following is the definition of the class posterior probability:

$$p(y_i = 1|x_{ij}) = \pi(x_i) = \frac{\exp(x_j^T \beta_j)}{1 + \exp(x_j^T \beta_j)}, \quad i = 1, 2, \ldots, p \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a vector of the dimension p of the unknown coefficients. Following this, we may get the estimator $\hat{\boldsymbol{\beta}}$ by minimizing the log-likelihood function:

$$\ell(\beta, y_i) = -\sum_{i=1}^{n} \{y_i \log \pi(x_i) + (1 - y_i) \log(1 - \pi(x_i)\} \tag{2}$$

A strong discriminative tool (feature selection) is the classification technique of logistic regression, which is used to classify data. On the other hand, since the design matrix is not invertible, the logistic regression method is not an effective classification strategy when the dataset in question is of a large dimension. The consequence of this is that it is unable to provide reliable estimates of the regression coefficient. Furthermore, overfitting happens when datasets have a high dimensionality, in which the number of genes(features) exceeds dramatically the sample size. It is also possible that multicollinearity will have an effect on its estimators (Algamal, 2017; Manhrawy et al., 2021).

From a statistical point of view, further variables that are not connected to the classification may produce noise and diminish the efficacy of the classification. Statistical analysts often use feature selection strategies that are able to get rid of characteristics that are unnecessary or redundant in order to improve the accuracy of categorization. In addition to the logistic regression, there are additional classification techniques that may be used. One of these ways is the regularized logistic regression (RLR), which is employed to decrease high dimensionality and improve classification accuracy (El Guide et al., 2020). Despite the fact that regularization techniques are often used for high-dimensional data, Doerken et al. (2019) stated that they are equally capable of being utilized for low-dimensional data.

In the process of regularized logistic regression, the log-likelihood function can be modified by incorporating a positive penalty component. This modification causes some coefficients to become zero, which results in the production of a sparse solution. By

including a penalty term in the equation, RLR is able to punish a logistic model that has an excessive number of features. Consequently, when the coefficients are limited, the coefficients of characteristics that are not as significant either get very close to zero or exactly zero. Another name for this method is penalization. Following is an explanation of how the approach is put up.

The regularized log-likelihood equation can be defined as

$$\text{RLR} = -\ell(\boldsymbol{\beta}, y_i) + \lambda g(\boldsymbol{\beta}) \tag{3}$$

here, $\ell(\boldsymbol{\beta}, y_i)$ represents the log-likelihood defined by Equation (2), $g(\boldsymbol{\beta})$ represents the penalization term, and $\lambda > 0$ represents a regulatory factor used to adjust the penalty amount. Following this, the RLR of Equation (3) is minimized with respect to λ to determine the coefficient estimations. According to Friedman et al. (2010), the intention of the penalty is to reduce the variances of the estimates and to make them biased, which ultimately leads to an improvement in the accuracy of the forecast. Classification and feature selection in high-dimensional datasets are two common applications for these penalizing approaches, which belong to a family of embedded selection methods (Liang et al., 2013).

Before solving the RLR minimization problem, let the response vector $\boldsymbol{y}$ be centered, and the columns of $\boldsymbol{X}$ (features) are usually standardized so that $\sum_{i=1}^{n} y_i = 0$, $\sum_{i=1}^{n} x_{ij} = 0$ and $(n^{-1}) \sum_{i=1}^{n} x_{ij}^2 = 1$ for $j \in \{1, 2, ..., p\}$. Standardization sets the intercept term β_0 to zero. $\boldsymbol{\beta}$ is estimated using LASSO (L1-norm penalization) as follows:

$$\hat{\beta}_{LASSO} = \arg\min_{\beta} \left[-\sum_{i=1}^{n} \{y_i \log \pi(x_i) + (1 - y_i) \log(1 - \pi(x_i))\} + \lambda \sum_{j=1}^{p} |\beta_j| \right] \tag{4}$$

here, λ represents the tuning parameter. When the value of $\lambda = 0$, Equation (4) reduces to the standard minimal likelihood estimator. Given that $\lambda \to \infty$, penalization enforces that all features must have a value of zero.

The ALASSO approach was first presented by Zou (2006) with the intention of resolving the overestimation issue that was present in the LASSO algorithm. This was achieved by substituting the L1 penalty with a weighted penalty, as stated by (Bühlmann and van de Geer, 2011). Through the process of assigning different weights to various coefficients, Zou made modifications to the L1-penalty mechanism. Various shrinkage approaches, including Ridge, LASSO, and others, might be used to determine the allocated weights. Given the above definition, the penalized logistic model that is connected with ALASSO is as follows:

$$\begin{aligned} \hat{\beta}_{LASSO} = \arg\min_{\beta} \Bigg[& \sum_{i=1}^{n} \{y_i \log \pi(x_i) + (1 - y_i) \log(1 - \pi(x_i))\} \\ & + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{\left(|\hat{\beta}_j^{initial}|\right)^{\gamma}} \Bigg] \end{aligned} \tag{5}$$

where, $\lambda, \gamma \geq 0$ and $\hat{\beta}_j^{initial}$ is an initial estimate for each β_j estimated using the LASSO technique or other shrinkage techniques. Here we set $\gamma = 1$ for simplicity.

In the process of gene selection, the EN is an additional significant penalized strategy that is used. In an effort to compensate for the shortcomings of LASSO, Zou and Hastie (2005) came up with this solution. It is possible to target genes that have a strong correlation and choose related genes all at once by using EN, which combines L2 and L1 norms. Using the EN penalty as a basis, the following equation may be used to calculate RLR:

$$\hat{\beta}_{EN} = \arg\min_{\beta} \left[-\sum_{i=1}^{n} \{y_i \log \pi(x_i) + (1 - y_i) \log(1 - \pi(x_i))\} + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right] \tag{6}$$

Equation (6) shows that the EN estimator relies on two regulatory parameters, λ_1 and λ_2, which are assumed to have only non-negative values. Equation (6) provides a solution using the RLR method.

In the same way that the EN technique may create the grouping effect, other penalized regression methods, such as the AEN methods presented by Ghosh (2011); Zou and Zhang (2009), which provided two AEN estimators, can also successfully do this. The adaptive weight was included into the L1-norm penalty that was contained inside the EN. The adaptive weights of the two separate AEN techniques are distinct from one another. Through the use of the EN estimator, Zou and Zhang (2009) managed to develop the adaptive weight. However, Ghosh (2011) constructed the adaptive weight by using the least-squares estimator. For λ_2 fixed, the RLR that is calculated using the AEN of β is as follows:

$$\hat{\beta}_{AEN} = \arg\min_{\beta} \left[-\sum_{i=1}^{n} \{y_i \log \pi(x_i) + (1 - y_i) \log(1 - \pi(x_i))\} + \lambda_1 \sum_{j=1}^{p} w_j |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right] \tag{7}$$

where $w_j = (|\hat{\beta}|)^{-\gamma}$, $j = 1, 2, ..., p$ represents the adjusted weight generated by the initial estimator $\hat{\beta}$. Here γ denotes a non-negative constant. A procedure known as coordinate descent is used to solve equations (4)-(7) (Friedman et al., 2010).

2.3 Proposed Improvement

Data analytics performance might be negatively impacted by the presence of missing data. This might lead to an erroneous forecast if the missing data is not imputed

correctly. The effective management of missing values is becoming more important in the present era of big data, which is characterized by the tremendous quantity of data that is generated every second and the fact that data utilization is a primary issue for stakeholders. Also, this study is motivated by the notion that the $L_1 - norm$ penalty may be used in RLR to apply the RLR technique to high-dimensional data sets. This is the driving force behind this. As a consequence of the fact that the $L_1 - norm$ is not consistent with feature selection, this method may lead to the selection of genes that are irrelevant and might be considered redundant (Algamal et al., 2018). That is another reason why this research is being conducted. When it comes to big coefficients, PLR estimates that are based on the $L_1 - norm$ penalty may be skewed due to the fact that they are subject to greater penalties.

The "one-dimensional weighted Mahalanobis distance" (1-DWM) was used by Peng et al. (2013) as a criterion of gene effectiveness. This was done in order to extend the influence of individual genes to the combined impact of multigene, which can be expressed as

$$J(x_{.j}) = \frac{(\bar{x}_{1j} - \bar{x}_{2j})^2}{\sigma_{wj}^2}, \tag{8}$$

where x_j is a column vector, denoting the feature j across samples, and $\sigma_{wj}^2 = w_{1j} \; . \; \sigma_{1j}^2 + w_{2j} \; . \; \sigma_{2j}^2$, indicates the weighted variance of the feature j, σ_{kj}^2 represents the variance of feature j in class k , w_k is the prior probability or weight of class k , where k in this paper is 2; i.e., we have exactly two classes and $w_1 = w_2 = 0.5$.

As a result, the purpose of this research is to propose the imputation of missing values with the adaptive elastic net as an extension of penalized techniques to enhance gene selection and impute missing values in high-dimensional data. Imputing missing values for every gene and using the 1-DWM as an initial weight inside the L1-norm is the approach that is used in order to accomplish this. The strategy that has been described works to enhance feature selection in high-dimensional data while also addressing missing values.

The j^{th} component of the $p-$dimensional vector of features can be expressed as

$$w_j = \frac{1}{|J(x_{.j})|}, \quad j = 1, 2, ..., p, \tag{9}$$

where $J(x_{.j})$ is the weight for every feature j that is defined as Equation (8).

The "naniar" package in R is used to impute missing values in the technique that has been presented. Furthermore, the suggested weight in this work offers a comparatively big amount of weight to the feature with a low ratio value while providing a small amount of weight to the feature with a high ratio value. This is done to reduce inconsistencies in the selection of features. Once the weights of the features have been assigned appropriately, the IAENR is able to choose related features with a high degree of reliability. A representation of the implementation method for the IAENR approach may be seen in Figure 1. In order to guarantee the existence of a global maximum point for the solution, the IAENR equation is convex, which guarantees its existence.

Finding the IAENR solution may be accomplished via the use of the coordinate descent approach.

$$\hat{\beta}_{IAENR} = \arg\min_{\beta} \left[-\sum_{i=1}^{n} \{y_i \log \pi(x_i) + (1-y_i)\log(1-\pi(x_i))\} + \lambda_1 \sum_{j=1}^{p} w_j|\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right] \tag{10}$$

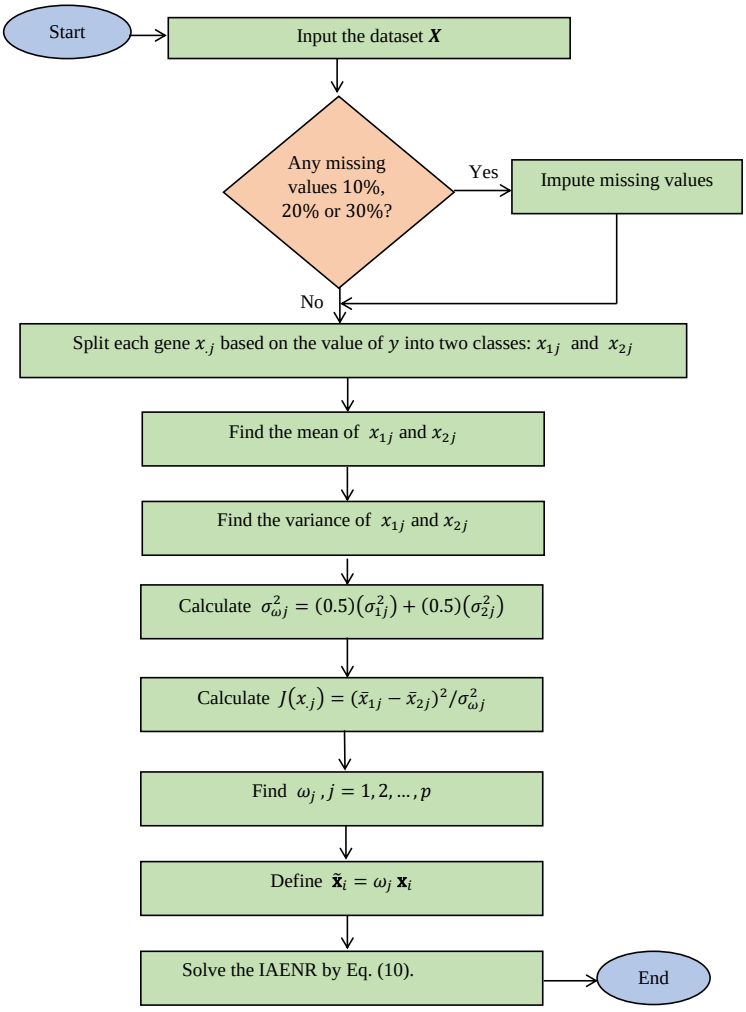

Figure 1: Flowchart of IAENR

In a genomic study with missing gene expression values, applying AEN after imputation might fail to capture the true relationships between genes if the imputation method does not consider the adaptive penalization of AEN. Combining both steps can guide the imputation by the feature importance weights from AEN, leading to more accurate and biologically meaningful results. In summary, combining AEN with missing value imputation is preferable to sequential application because it ensures better integration of both tasks, reduces biases, improves model performance, and is more robust in high-dimensional and sparse data settings.

2.4 Evaluation Metrics

To assess the effectiveness of the proposed technique, IAENR, this subsection makes use of three criteria that are often employed for evaluating predictive models, notably in the context of healthcare (Tharwat, 2021). The classification accuracy (CA), sensitivity (SEN), and specificity (SPE) measures are included in these metrics. These metrics are defined as

$$CA = \frac{TN + TP}{FP + TP + TN + FN} \times 100\% \tag{11}$$

$$SEN = \frac{TP}{FN + TP} \times 100\% \tag{12}$$

$$SPE = \frac{TN}{TN + FP} \times 100\% \tag{13}$$

where TP, FP, TN, and FN are True Positive, False Positive, True Negative, and False Negative, respectively. When the values of the assessment criteria are greater, it indicates that the classification performance is better.

2.5 Dataset Description

An evaluation of the effectiveness of the suggested approach, known as IAENR, is carried out by applying it to two gene expression datasets that include varying quantities of genes and observations. The general public may access these datasets, which have been extensively used by several academics in the past. In the first place, there is the data set on colon cancer, which has a total of 6500 genes and 62 people, four of whom have malignant tumors and twenty-two of whom have noncancerous tissues. The Affymetrix oligonucleotide array technology plays a role in the acquisition of this information. In this particular data set, only two thousand gene expressions were used, and they were arranged according to the greatest minimal intensity across all of the samples (Alon et al., 1999). Prostate cancer is the subject of the second data collection. It has a total of 12600 genes. In the sample, 52 individuals have malignant prostate tumors, and 50 additional patients have tumor tissues that are not cancerous (Singh et al., 2002).

3 Findings and Discussion

In this part, the data mentioned before were taken into consideration to explain the behavior of various strategies concerning the selection of variables when there is missing data. In the course of comparison trials with Elastic Net and AElastic Net, it was shown that the suggested approach, which is known as IAENR, is successful. First, we went through the process of using these approaches to complete the data without missing values. To facilitate an objective comparison, we randomly divided each data into two parts: a training dataset, which included seventy percent of the samples, and a test dataset, which contained thirty percent of the samples. With the use of the "glmnet" package in R, the 10-fold cross-validation (CV) was used to the training dataset one hundred times in order to determine the ideal value of λ . On the other hand, the procedure that was followed in order to assess the methods that had missing values is as follows. In this research, we begin by utilizing the "missForest" package of the programming language R to seed missing values in our datasets with various rates (10%, 20%, and 30%). This is done under the assumption that there is no missing data in the response variable. As a second step, we imputed the values that were missing by utilizing the "naniar" package that is included in the R programming language. In the third step, we carried out the imputing of data as full data using penalized approaches. In both the training and testing datasets, the average number of genes that were picked, as well as the averaged classification accuracy (CA), sensitivity (SEN), and specificity (SPE), are provided in Tables 1 and 2, respectively. Evaluations of the EN and AEN's performance were also carried out with the goal of making comparisons.

As Tables 1 and 2 demonstrate, the suggested approach, IAENR, chose genes that were higher than those picked by the EN and AEN in both the colon and prostate datasets, which had various levels of missing values. For instance, in the colon data set that had 10% of the data missing, the IAENR picked 22 genes, but the EN and AEN genes selected 17 and 19 genes, respectively. On the other hand, with regard to both sets of data, we discovered that EN gives the least amount of chosen genes, which means that IAENR encouraged the grouping effect.

As an additional point, we note in Tables 1 and 2 that the average classification accuracy, sensitivity, and specificity of the IAENR in both the training and testing sets are much higher than those of the EN and AEN. This is the case in both of the data sets that were used in this investigation. When it comes to the Colon data, which contains 20% missing data, for instance, the classification accuracy of IAENR in the training set is 94.61%, which is higher than the accuracy of EN (89.00%) and AEN (91.75%). In addition, the sensitivity of the IAENR is 87.50% which is higher than the sensitivity of the EN and AEN, which are 81.59% and 83.56%, respectively, in the case of prostate data that is missing thirty percent of the data. Within the testing sets of the colon and prostate datasets, which include varying percentages of missing values, the same observation may be drawn to the same conclusion.

To go further into the IAENR's performance, statistical tests should be run to see whether the classification accuracy differences shown in Tables 1 and 2 are statistically significant. The data analysis in this research was done using the paired t-test. The

Table 1: The 100-times-averaged criterion for the Colon data for the training and testing sets

			Training set			Testing set		
Missing %	Methods	Genes	CA%	SEN%	SPE%	CA%	SEN%	SPE%
No missing	EN	17	93.13	92.50	94.64	83.53	85.43	86.50
	AEN	19	93.72	93.41	95.64	84.40	86.45	86.41
	IAENR	22	97.10	95.14	96.63	88.92	89.50	88.73
10%	EN	17	92.22	90.12	93.00	81.14	82.34	86.74
	AEN	19	93.34	90.72	95.00	84.11	83.41	87.01
	IAENR	22	96.10	95.00	96.00	87.53	89.82	88.83
20%	EN	16	89.00	87.00	89.50	81.00	80.10	75.55
	AEN	15	91.75	87.01	90.00	80.50	78.92	75.92
	IAENR	21	94.61	88.85	92.62	85.48	83.33	81.75
30%	EN	17	84.66	81.50	84.56	79.50	77.73	79.88
	AEN	18	86.64	84.60	89.44	82.22	79.24	84.82
	IAENR	22	91.00	89.25	91.33	88.34	83.23	86.76

results are shown in Tables 3 and 4. The "improvement" column represents the relative improvement in mean average accuracy that the suggested approach offers compared to the other methods. Furthermore, at the 5% level of significance, Tables 3 and 4 show that our suggested technique, IAENR, differs significantly from all competing approaches.

In order to emphasize the effectiveness of the IAENR, we conducted a comparison between the results achieved for the Colon dataset in terms of the number of chosen genes and CA and the results provided by Ref. (Alharthi et al., 2022) for imputations adaptive penalized logistic regression (IAPLR). Our approach identified a greater number of genes compared to the IAPLR technique. Specifically, our method identified 76 genes, while the IAPLR method only identified 12 genes. Significantly, IAENR has the capacity to identify a greater number of genes compared to the IAPLR technique, suggesting that the majority of these extra-picked genes were likely to be strongly connected. In addition, our technique demonstrated a superior classification accuracy (CA) of 97.10% in comparison to IAPLR's CA of 96.12%.

Overall, applying the IAENR to gene selection, improving classification accuracy, and handling missing values in high-dimensional data has been a fruitful endeavor. For both the training and testing datasets, the suggested technique was shown to have higher classification performance using metrics such as high CA, SEN, and SPE. If the suggested technique can meet all three of these criteria at the same time, it is nominated as a possible gene selection methodology. In addition, our adaptive penalized method,

Table 2: The 100-times-averaged criterion for the Prostate data for the training and testing sets

			Training set			Testing set		
Missing %	Methods	Genes	CA%	SEN%	SPE%	CA%	SEN%	SPE%
No missing	EN	19	90.00	92.77	92.88	80.50	83.43	86.22
	AEN	21	92.15	94.37	94.56	83.45	86.00	86.50
	IAENR	27	94.50	95.35	95.95	88.19	89.28	88.25
10%	EN	21	90.20	90.10	91.67	79.90	82.80	83.85
	AEN	23	91.52	93.30	94.00	82.48	84.75	85.14
	IAENR	26	93.75	95.00	94.80	88.49	88.60	90.50
20%	EN	19	86.42	87.23	87.72	79.00	79.65	81.53
	AEN	19	87.86	88.42	90.50	82.00	83.66	86.69
	IAENR	25	90.60	90.20	91.20	85.75	86.52	89.62
30%	EN	20	81.60	81.59	81.34	76.58	78.39	78.63
	AEN	22	84.44	83.56	83.58	80.77	79.28	80.40
	IAENR	24	88.33	87.50	87.25	83.42	83.87	82.97

Table 3: Statistically significant findings using a paired t-test on Colon datasets

		Training dataset		Testing dataset	
Missing%	Methods	Improvement	p-value	Improvement	p-value
No missing	EN	2.12%	0.0020 (*)	6.45%	0.0000 (*)
	AEN	1.47%	0.0043 (*)	5.36%	0.0002 (*)
10%	EN	4.2%	0.0024 (*)	7.88%	0.0000 (*)
	AEN	2.96%	0.0018 (*)	4.07%	0.0001 (*)
20%	EN	6.3%	0.0003 (*)	5.53%	0.0000 (*)
	AEN	3.12%	0.0021 (*)	6.19%	0.0001 (*)
30%	EN	7.49%	0.0001 (*)	11.12%	0.0001 (*)
	AEN	5.03%	0.0005 (*)	7.44%	0.0001 (*)

(*) significant at $\alpha = 0.05$

IAENR, outperforms competing methods in terms of classification accuracy. Here we also see that our method takes the gene weights into consideration.

Table 4: Statistically significant findings using a paired t-test on Prostate dataset

		Training dataset		Testing dataset	
Missing %	Other methods	Improvement	p-value	Improvement	p-value
No missing	EN	5.00%	0.0024(*)	9.55%	0.0002 (*)
	AEN	2.55%	0.0051(*)	5.68%	0.0001 (*)
10%	EN	3.94%	0.0022(*)	10.75%	0.0002 (*)
	AEN	2.44%	0.0037(*)	7.28%	0.0001 (*)
20%	EN	4.84%	0.0001(*)	8.54%	0.0000 (*)
	AEN	3.12%	0.0011(*)	4.57%	0.0002 (*)
30%	EN	8.25%	0.0020(*)	8.93%	0.0003 (*)
	AEN	4.61%	0.0001(*)	3.28%	0.0006 (*)

(*) significant at $\alpha = 0.05$

4 Conclusion

Improving data analytics relies mainly on the imputation of missing values. Finding an approach to missing data imputation that is applicable to every kind of dataset is challenging. Despite advancements in variable selection techniques and tools, missing data is common in extensive, complex studies and may make data analysis difficult. Our primary goals in writing this article were to develop an IAENR approach for dealing with missing values in high-dimensional data and to enhance the performance of penalized logistic regression models. We confirm that the proposed IAENR outperforms EN and AEN as a classification and gene selection process when dealing with missing data when comparing the results obtained from applying the IAENR method to two datasets (colon and prostate) with seeding the same datasets at different rates of missing values. This confirms that our technique is a gene selection and classification strategy with solid statistical support. On top of that, it works with related datasets.

Acknowledgement

"The authors extend their appreciation to Taif University, Saudi Arabia, for supporting this work through project number (TU-DSPP-2025-03)."

References

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

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

Alharthi, A. M., Lee, M. H., and Algamal, Z. Y. (2022). Improving penalized logistic regression model with missing values in high-dimensional data. *International Journal of Online and Biomedical Engineering (iJOE)*, 18(02):pp. 40–54.

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

Brini, A. and van den Heuvel and, E. R. (2024). Missing data imputation with high-dimensional data. *The American Statistician*, 78(2):240–252.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Springer Series in Statistics. Springer Berlin Heidelberg, Berlin, Heidelberg.

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

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

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

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

Doerken, S., Avalos, M., Lagarde, E., and Schumacher, M. (2019). Penalized logistic regression with low prevalence exposures beyond high dimensional settings. *PLOS ONE*, 14(5):e0217057.

El Guide, M., Jbilou, K., Koukouvinos, C., and Lappa, A. (2020). Comparative study of L1 regularized logistic regression methods for variable selection. *Communications in Statistics - Simulation and Computation*, pages 1–16.

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

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1–22.

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

Ghosh, S. (2011). On the grouped selection and model complexity of the adaptive elastic net. *Statistics and Computing*, 21(3):451–462.

Holman, R. and Glas, C. A. W. (2005). Modelling non-ignorable missing-data mecha-

nisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, 58(1):1–17.

Honaker, J., King, G., and Blackwell, M. (2011). Amelia II: A Program for Missing Data. *Journal of Statistical Software*, 45(7):1–47.

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

Khan, S. I. and Hoque, A. S. M. L. (2020). SICE: an improved missing data imputation technique. *Journal of Big Data*, 7(1):37.

Kwak, S. K. and Kim, J. H. (2017). Statistical data preparation: management of missing values and outliers. *Korean Journal of Anesthesiology*, 70(4):407.

Lee, Y. and Park, S. (2024). High-dimensional missing data imputation via undirected graphical model. *Statistics and Computing*, 34(5):160.

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

Li, X., Wang, Y., and Ruiz, R. (2020). A Survey on Sparse Learning Models for Feature Selection. *IEEE Transactions on Cybernetics*, pages 1–19.

Liang, L., Zhuang, Y., and Philip, L. (2024). Variable selection for high-dimensional incomplete data. *Computational Statistics and Data Analysis*, 192:107877.

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

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

Manhrawy, I. I., Qaraad, M., and El-Kafrawy, P. (2021). Hybrid feature selection model based on relief-based algorithms and regularization algorithms for cancer classification. *Concurrency and Computation: Practice and Experience*, 33(17):1–17.

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

Peng, H., Fu, Y., Liu, J., Fang, X., and Jiang, C. (2013). Optimal gene subset selection using the modified SFFS algorithm for tumor classification. *Neural Computing and Applications*, 23(6):1531–1538.

Rubin, D. B. (1996). Multiple Imputation after 18+ Years. *Journal of the American Statistical Association*, 91(434):473–489.

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

Rácz, A. and Gere, A. (2025). Comparison of missing value imputation tools for machine learning models based on product development cases studies. *LWT*, 221:117585.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff,

P. W., Golub, T. R., and Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2):203–209.

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

Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics*, 17(1):168–192.

Tian, Y., Ali, M. K. M., Wu, L., and Li, T. (2025). Imputation method based on adaptive group lasso for high-dimensional compositional data with missing values. *Malaysian Journal of Fundamental and Applied Sciences*, 21(1):1551–1565.

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

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice : Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3):1–67.

Wang, A., Yang, J., and An, N. (2021). Regularized Sparse Modelling for Microarray Missing Value Estimation. *IEEE Access*, 9:16899–16913.

Zahid, F. M., Faisal, S., and Heumann, C. (2020). Variable selection techniques after multiple imputation in high-dimensional data. *Statistical Methods & Applications*, 29(3):553–580.

Zahid, F. M., Faisal, S., and Heumann, C. (2021). Multiple imputation with compatibility for high-dimensional data. *PLOS ONE*, 16(7):e0254112.

Zahid, F. M. and Heumann, C. (2019). Multiple imputation with sequential penalized regression. *Statistical Methods in Medical Research*, 28(5):1311–1327.

Zhang, Y. and Kim, S. (2024). Variable selection for high-dimensional incomplete data using horseshoe estimation with data augmentation. *Communications in Statistics - Theory and Methods*, 53(12):4235–4251.

Zhang, Z. (2015). Missing values in big data research: some basic skills. *Annals of translational medicine*, 3(21):323.

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

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

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

Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, 37(4):1733–1751.