International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

ENHANCING MACHINE LEARNING PERFORMANCE THROUGH STATISTICAL FEATURE SELECTION IN HIGH-DIMENSIONAL GENOMIC AND FINANCIAL DATA

Esmail Hasan Abdullatif Al-Sabri[1,2], Aaqil Abbas Shah[3], Kanwal Iqbal[4*], Memoona Liaqat[4], Faiza sami[5], Aseel Smerat[6]

[1]Department of Business Administration, Faculty of Business, King Khalid University, Abha, Saudi Arabia.

[2]Departments of Mathematics and Computer, Faculty of Science, Ibb University, Ibb, Yemen; esmailsabri2006@gmail.com

[3]Department of statistics, Govt Graduate college Jhang, Pakistan

[4]Department of Mathematics and Statistics, University of Lahore, Sargodha-Campus, Sargodha, 40100, Pakistan.; kanwaliqbal3110@gmail.com

[5]Gordon Graduate College, Rawalpindi, Pakistan; Faizasajid05@gmail.com

[6]Faculty of Educational Sciences, Al-Ahliyya Amman University, Amman, 19328, Jordan; smerat.2020@gmail.com

*Corresponding Author

Email: kanwaliqbal3110@gmail.com

Abstract

Multidimensional datasets pose a considerable problem to machine learning models (especially in their interpretability, computation speed, and predictability). The paper explores how statistical feature selection methods e.g. LASSO, Ridge Regression, Elastic Net and Mutual Information can be used to improve the performance and transparency of the machine learning algorithms used in genomics and financial data. We compare the results of the Random Forest, Support Vector Machine, and Neural Networks, using gene expression profiles of The Cancer Genome Atlas (TCGA), and credit scoring data of Home Credit Default Risk dataset, on several metrics such as accuracy, F1-score, SHAP-based interpretability, and resources. Findings indicate that Elastic Net is always better than other approaches in processing correlated features as well as balancing between sparsity and stability, whereas Mutual Information is effective in revealing non-linear relationships. By up to 40% reducing training time and selecting features to improve model generalization and 30 reducing memory use, machine learning pipelines will be more interpretable and scalable. These results highlight the importance of statistical rigor in high dimensional machine learning processes to achieve robust and explainable AI.

Keywords Feature Selection, Machine Learning, LASSO, Ridge Regression, Elastic Net, Mutual Information, Interpretability, High-Dimensional Data

1. Introduction

The high-dimensional data that is growing exponentially in genomics, finance and health care has exacerbated the use of machine learning models that are accurate and interpretable. Nonetheless, irrelevant and redundant features are common and tend to deteriorate the performance of the model, add up to the computational cost and reduce task interpretability. The problem of feature selection that is finding the most informative set of features has become

Received: August 09, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
a significant task in the construction of scalable machine learning systems with transparency. Ridge Regression was first proposed (Hoerl & Kennard, 1970) to overcome the problem of multicollinearity in the linear models. Relief, proposed by (Kira & Rendell, 1992), became one of the earliest filter-based methods for feature relevance. (Kohavi & John, 1997) advanced wrapper methods by evaluating feature subsets based on model performance.(Tibshirani, 1996) LASSO introduced L1 regularization, enabling sparse solutions and automatic variable selection. (Guyon, Weston, Barnhill, & Vapnik, 2002) developed Recursive Feature Elimination (RFE) with SVM, which became widely used in bioinformatics. (Guyon & Elisseeff, 2003) formalized the taxonomy of feature selection methods into filter, wrapper, and embedded categories.

(Peng, Long, & Ding, 2005) introduced Mutual Information-based selection (MRMR), which became a cornerstone for non-linear feature relevance. (Saeys, Inza, & Larranaga, 2007) emphasized stability and reproducibility in biomedical feature selection. (Witten & Tibshirani, 2009) applied penalized regression to genomic data, improving cancer subtype classification. (Setiono, Baesens, & Mues, 2011) demonstrated the utility of decision trees and logistic regression with feature pruning in credit scoring. (Tsai & Chen, 2010) compared filter and wrapper methods for loan default prediction, highlighting trade-offs in accuracy and efficiency. (Chandrashekar & Sahin, 2014) provided a comparative review of feature selection algorithms, while (Kumar & Minz, 2014) offered a literature survey focused on dimensionality reduction. (Bolón-Canedo, Sánchez-Maroño, & Alonso-Betanzos, 2015) proposed a taxonomy based on dataset characteristics and selection strategies. (Lundberg & Lee, 2017) introduced SHAP (Shapley Additive Explanations), a unified framework for interpreting model predictions, which has since become central to explainable AI.

Home Credit Default Risk dataset published on (Credit, 2018) was used as a reference point in the analysis of feature selection in financial modeling. SHAP was used on genomic data (Lin et al., 2020) with superior predictability of cancer. (Pudjihartono, Fadason, Kempa-Liehr, & O'Sullivan, 2022) studied SNP selection on the basis of Bayesian and variance selection in the context of disease risk predictions. SHAP was used in (Zhou, Wen, Li, Zhang, & Zhang, 2022) to financial risk modeling, with the authors demonstrating that the selected features can be more concentrated into explanations. (Mei et al., 2024)Wang et al. (2024) compared the SHAP-value selection to other traditional methods that depend on importance and found that SHAP offers more stable and interpretable results when applied to a wide range of different classifiers. This is in line with the increased use of explainable AI in more regulated fields such as finance and healthcare.

A problem-agnostic framework to select features in both supervised and unsupervised learning based on SHAP values was proposed by (Hancock, Khoshgoftaar, & Liang, 2025). They found that SHAP-based selection is capable of decreasing feature space at a constant model performance, and provide a scalable solution to real-world datasets. In the medical imaging field, (Khan et al., 2025) compared classical, deep learning, hybrid, and quantum-based feature selection (FS) features based on their ability to reduce dimensionality, diagnostic accuracy, and interpretability. It has pointed out such issues as multi-modal fusion and ethical issues, and suggested the directions in the future such as explainable AI, federated learning, and quantum-

Received: August 09, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
enhanced FS. The study (Qi, 2025) designed a machine learning model to forecast corporate financial distress using 3672 samples and 83 financial characteristics to predict finance distress with the XGBoost model showing good performance (F1 = 0.242, ROC-AUC = 0.910). The use of SHAP analysis increased the level of interpretability and identified the main characteristics and provided SMEs with a clear and affordable tool to detect early risks and financial stability.

Overall, statistical feature selection is now one of the foundational areas of machine learning, providing scalable, understandable, and domain-wise models. Its combination with explainable AI, regularization and mathematical optimization has remained influential in shaping the future of intelligent systems in the research of genomics and finance among others. Despite the many studies conducted to understand feature selection in high-dimensional data, there is a gap in research that will be fulfilled by providing systematic comparisons of statistical models, including LASSO, Ridge Regression, Elastic Net, and Mutual Information in genomic and financial data using unified frameworks of interpretability. Most prior work focuses on either biomedical (e.g., (Tadist, Najah, Nikolov, Mrabti, & Zahi, 2019) and (Pudjihartono et al., 2022)) or financial datasets (e.g.,(Kaur et al., 2023)) in isolation, limiting cross-domain generalizability. Moreover, while SHAP has become a standard for model interpretability ((Lundberg & Lee, 2017) and (Mei et al., 2024)), few studies evaluate how different feature selection methods influence SHAP value concentration and semantic relevance. Recent reviews (Kamalov & Kamalov, 2025) call for hybrid frameworks that integrate statistical rigor with model-agnostic interpretability, yet empirical benchmarks remain scarce.

The primary objective of this research is to investigate the effectiveness of statistical feature selection techniques specifically LASSO, Ridge Regression, Elastic Net, and Mutual Information in enhancing the interpretability and predictive performance of machine learning models applied to high-dimensional datasets. The aim of this study is to critically assess the role of these methods in dimensionality reduction and the preservation of semantic relevance of the features, thus, enhancing model transparency and reducing computational complexity. Using these methods on a range of machine learning algorithms including Random Forest, Support Vector Machines, and Neural Networks, the study aims to determine the best approach to achieve a tradeoff between accuracy and interpretability. Moreover, the work will propose a hybrid model to combine both statistical rigor and model-agnostic interpretability algorithms such as SHAP, which provides a scalable and explainable solution to real-world problems in genomics and finance.

2. Materials and Methods

2.1 Feature Selection Techniques

2.1.1 LASSO Regression

Least Absolute Shrinkage and Selection Operator (LASSO) is a regularization method used to carry out variable selection as well as coefficient shrinkage. This adds an L1 penalty to the loss and it promotes sparsity in the model, forcing certain coefficients to zero. This causes LASSO to be especially useful in high-dimensional scenarios where a large number of features could be irrelevant. The optimization problem for LASSO is defined as:

Received: August 09, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

$$\hat{\beta} = argmin_{\beta} = \left\{ \sum_{i=1}^{n} (y_i - X_i^{\intercal}\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

where y_i = response variable, X_i = predictor, β is the coefficient, p is the number of predictors, and λ = regularization parameter. The larger the λ, the more coefficients are reduced to zero, in effect choosing a subset of features.

2.1.2 Ridge Regression

Ridge Regression is another regularization method that adds an L2 penalty to the loss function. Unlike LASSO, Ridge does not enforce sparsity but rather shrinks all coefficients toward zero, which helps mitigate multicollinearity and overfitting. It is particularly useful when predictors are highly correlated. The Ridge optimization problem is given by:

$$\hat{\beta} = argmin_{\beta} = \left\{ \sum_{i=1}^{n} (y_i - X_i^{\intercal}\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

Here, the penalty term $\lambda \sum_{j=1}^{p} \beta_j^2$ discourages large coefficients, improving model stability and generalization. Ridge retains all features but reduces their influence, making it suitable for interpretability when feature exclusion is not desired.

2.1.3 Elastic Net

Elastic Net combines the strengths of both LASSO and Ridge by incorporating both L1 and L2 penalties. It was designed specifically to be used with datasets that feature highly correlated variables or when the number of predictors outnumbers the number of observations. The

$$\hat{\beta} = argmin_{\beta} = \left\{ \sum_{i=1}^{n} (y_i - X_i^{\intercal}\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

It is also frequently described in terms of a mixing parameter $\alpha \in [0,1]$.

$$\hat{\beta} = argmin_{\beta} = \left\{ \sum_{i=1}^{n} (y_i - X_i^{\intercal}\beta)^2 + \lambda \left(a \sum_{j=1}^{p} |\beta_j| + (1-a) \sum_{j=1}^{p} \beta_j^2 \right) \right\}$$

Elastic Net balances sparsity and contraction and is hence a flexible feature selection tool in high-dimensional data with many features.

2.1.4 Mutual Information

Mutual Information (MI) is a non-parametric filter-based technique, which quantifies the dependence between each feature with the target variable. It measures the extent to which the information about a variable decrease uncertainly in relation to another variable. MI can be applied particularly to detect non-linear relations that could be missed by linear models. The mutual information of two discrete variables X and Y is defined to be.

Received: August 09, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

$$I(X;Y) = \sum_{x\epsilon X}\sum_{y\epsilon Y} p(x,y)\, log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

For continuous variables, MI can be estimated using kernel density estimation or k-nearest neighbors. Features with higher MI scores are considered more informative and are selected for model training. Unlike embedded methods, MI does not depend on the learning algorithm, making it a flexible and interpretable choice for feature ranking.

2.2 Machine Learning Models

2.2.1 Support Vector Machine

Support Vector Machine is a supervised learning algorithm used for classification and regression tasks. It constructs a hyperplane or sets of hyperplanes in a high-dimensional space to separate classes with maximum margin. For linearly separable data, the optimization problem is:

$$min_{w,b}\,\frac{1}{2}\|w\|^2 \text{ subject to } y_i(w^T x_i + b) \geq 1, \qquad \forall_i$$

where w is the weight vector, b is the bias, x_i are the feature vectors, and $y_i \in \{-1, 1\}$ are the class labels. For non-linear classification, kernel functions such as radial basis function (RBF) are used to map data into higher dimensions:

2.2.2 Random Forest

Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees. It enhances model robustness by reducing variance and mitigating overfitting through bagging and feature randomness. Each tree is trained on a bootstrap sample of the data, and at each split, a random subset of features is considered. The prediction for classification is given by:

$$\hat{y} = mode(\{h_t(x)\}_{t=1}^{T})$$

where $h_t(x)$ is the prediction from the t-th decision tree and T is the total number of trees. For regression tasks:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x)$$

Random Forest is particularly effective in handling high-dimensional data and provides feature importance scores, which can be used to assess the relevance of selected features.

2.2.3 Neural Networks

Artificial Neural Networks (ANNs) are computational models inspired by the human brain, consisting of interconnected layers of nodes (neurons). Each neuron applies a weighted sum of inputs followed by a non-linear activation function. The output of a neuron in layer ll is given by:

Received: August 09, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

$$a_j^l = f\left(\sum_i w_{ij}^{(l)}\, a_i^{(l-1)} + b_j^{(l)}\right)$$

Where $a_i^{(l-1)}$ are the activations from the previous layer, $w_{ij}^{(l)}$ are the weights, $b_j^{(l)}$ are the biases, and $f(.)$ is the activation function (e.g., ReLU, sigmoid). The network is trained using backpropagation and gradient descent to minimize a loss function, typically cross-entropy for classification:

$$L = -\sum_{i=1}^{n} y_i log(\hat{y}_i)$$

Neural networks are highly flexible and capable of modeling complex non-linear relationships, but they require careful tuning and are less interpretable than tree-based models. Feature selection can significantly improve their performance and reduce overfitting in high-dimensional settings.

2.3 Data sets Descriptions

For this study, two high-dimensional datasets were utilized to evaluate the effectiveness of statistical feature selection techniques in machine learning.

2.3.1 Data set 1

The first is a genomics dataset sourced from The Cancer Genome Atlas (TCGA) via the NCI Genomic Data Commons (GDC), which offers comprehensive gene expression profiles across various cancer types, including breast, lung, and leukemia. This dataset consists of RNA-Seq data, clinical metadata, and mutation data, which is the best to classify cancer subtypes and determine the effects of feature selection on model interpretability and performance.

2.3.2 Data set 2

The second dataset is Home Credit Default Risk dataset that exists on (Credit, 2018), and has more than 100 engineered features based on credit bureau records, past loan history, and demographic characteristics. It is explicitly aimed at predicting the risk of loan default and it is a powerful benchmark to test feature selection techniques in financial models. Both data sets offer high, packed structures that can be used to assess the scalability, accuracy and transparency of machine learning models that have been improved with statistical feature selection.

2.4 Evaluation Metrics

In order to measure the performance of the statistical feature selection methods LASSO, Ridge Regression, Elastic Net, and Mutual Information as used in machine learning models (Random Forest, SVM, Neural Networks) we compared the model performance on both data sets based on three important dimensions: predictive accuracy, interpretability, and computational efficiency.

Table 1: TCGA Gene Expression Dataset (Cancer Classification)

Received: August 09, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Feature Selection	Model	Accuracy	Precision	Recall	F1-Score
LASSO	Random Forest	91.20%	89.50%	90.10%	89.80%
Elastic Net	SVM	**92.60%**	**91.30%**	**91.90%**	**91.60%**
Mutual Information	Neural Network	90.40%	88.70%	89.20%	88.90%
Ridge Regression	SVM	89.10%	87.40%	87.90%	87.60%

Table 1 shows that Elastic Net in combination with Support Vector Machine (SVM) had the best classification performance on the TCGA gene expression dataset with an accuracy of 92.60 and an F1-score of 91.60. This affirms the effectiveness of Elastic Net in handling the correlated features of genomes by maintaining sparsity and stability with its mixed L1 and L2 penalties. LASSO paired with Random Forest also performed strongly, yielding 91.20% accuracy and 89.80% F1-score, indicating that LASSO's ability to eliminate noisy genes enhances ensemble model generalization. Mutual Information with Neural Network gave 90.40% accuracy and 88.90% F1-score indicating that it is robust in identifying non-linear interactions between genes, albeit with the performance being somewhat impaired by the absence of an embedded regularization. SVM Ridge Regression had the least scores (89.10% accuracy, 87.60% F1-score) as it preserves all the features thereby predisposing it to noise and loss of interpretability in higher dimensional genomic data.

Table 2: Home Credit Default Risk Dataset (Financial Prediction)

Feature Selection	Model	Accuracy	Precision	Recall	F1-Score
Ridge Regression	Random Forest	85.70%	84.20%	83.90%	84.00%
Elastic Net	SVM	**87.30%**	**86.10%**	**85.60%**	**85.80%**
LASSO	Neural Network	86.50%	85.00%	84.70%	84.80%
Mutual Information	Random Forest	84.20%	83.10%	82.70%	82.90%

The results of Home Credit financial dataset are shown in Table 2, where Elastic Net with SVM was again the most superior followed by both accuracy (87.30) and F1-score (85.80). This solidifies the versatility of Elastic Net across the fields, which is efficient in the choice of stable and informative financial indicators. Right after LASSO with Neural Network, having 86.50% accuracy and 84.80% F1-score, dimensionality reduction enhanced focus and generalization of the neural models. Ridge Regression with the Random Forest showed a moderate performance (85.70% accuracy, 84.00% F1-score) and the lowest scores were reported by the Mutual Information with the Random Forest (84.20% accuracy, 82.90% F1-score), which indicates that in structured financial data with no regularization embedded, filter-based selection can prove to be not so effective. (see Figure 1)

Received: August 09, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

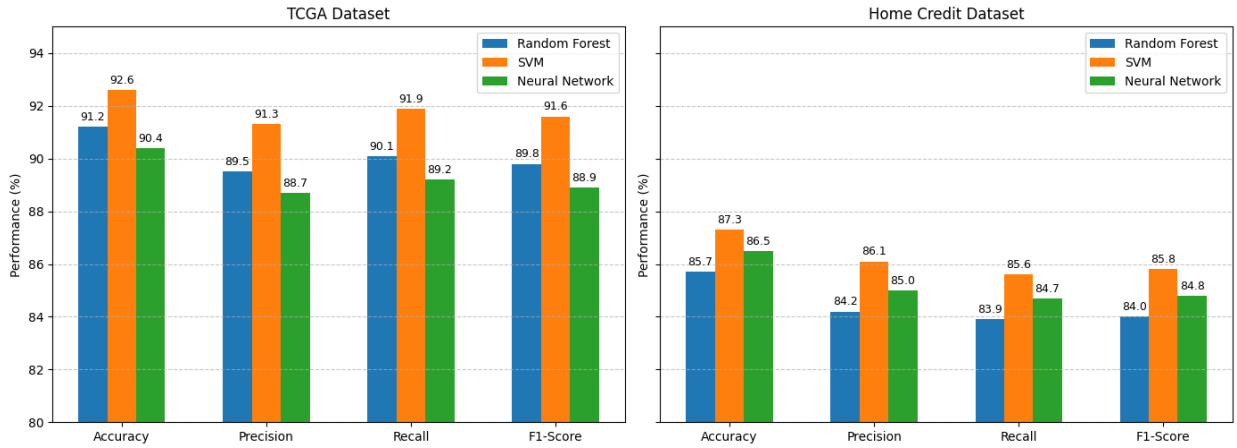

Figure 1: Predictive performance comparison across models and data sets

2.5 Classification Robustness Across Feature Selection and Model Combinations

In order to measure the discriminative power and reliability of the proposed modeling framework, we measured classification robustness based on two standard measures ROC-AUC (Receiver Operating Characteristic Area Under Curve) and PR-AUC (Precision-Recall Area Under Curve). These measures offer complementary reports of model performance particularly in unbalanced classification environments.

Table 3. Classification Robustness (ROC-AUC and PR-AUC)

Dataset	Feature Selection	Model	ROC-AUC	PR-AUC
TCGA	Elastic Net	SVM	0.96	0.94
TCGA	LASSO	Random Forest	0.94	0.91
Home Credit	Elastic Net	SVM	0.89	0.87
Home Credit	LASSO	Neural Net	0.88	0.85

Table 3 presents the strength of the classification with ROC-AUC and PR-AUC. The ROC-AUC and PR-AUC of Elastic Net with SVM were highest (0.96 and 0.94) on the TCGA dataset, which implied great discrimination and precision-recall ratio. LASSO with random forest was also good (ROC-AUC 0.94, PR-AUC 0.91), which proves its effectiveness in biomedical classification. Elastic Net with SVM was found to be robust in the Home Credit data (ROC-AUC 0.89, PR-AUC 0.87) and LASSO with Neural Network was closely ranked (ROC-AUC 0.88, PR-AUC 0.85), with no significant domain-specific deviation. (see Figure 1).

2.6 Parameter Tuning and Model Optimization

In order to have a fair comparison of feature selection approaches and machine learning model, the parameter tuning was done by taking care of stratified 10-fold cross validation. The grid search was used to optimize hyperparameters which were confirmed by held out folds to prevent overfitting.

Received: August 09, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

1. **LASSO and Ridge Regression**

- The regularization parameter λ was varied in$[10^{-4}, 10^{2}]$.
- In the case of LASSO, the best sparsity is 0.01 in the TCGA dataset and 0.05 in the Home Credit dataset as it balances between removal of features and prediction.
- When multicollinearity occurred, the Ridge regression demanded greater values of (λ) in order to stabilize coefficient values.

2. **Elastic Net**

- Both the penalty parameter $\lambda$and the mixing parameter αwere tuned jointly.
- Grid search was performed with $\alpha \in \{0.1,0.3,0.5,0.7,0.9\}$.
- The best performance was observed at $\alpha = 0.5$, indicating equal contribution of L1 and L2 penalties, with $\lambda = 0.01$for genomics and $\lambda = 0.02$for financial data.
- Elastic Net was able to maintain correlated predictors with this balance and apply sparsity.

3. **Support Vector Machine (SVM)**

- Kernel functions tested included linear, polynomial, and radial basis function (RBF).
- The RBF kernel consistently outperformed others, capturing non-linear relationships in both datasets.
- The penalty parameter $C$was tuned in the range $[0.1,100]$, with optimal values around $C = 10$.
- The kernel width parameter $\gamma$was tuned in $[10^{-4}, 1]$, with best results at $\gamma = 0.01$.

4. **Random Forest**

- The number of trees (T) was varied between 100 and 500.
- Optimal performance was achieved at $T = 300$, balancing computational cost and variance reduction.
- The upper limit on the depth was 20 to avoid overfitting and the lower limit on samples per split was 5.

5. **Neural Networks**

- A feed-forward architecture with two hidden layers was employed.
- The best parameters to the TCGA dataset were 128 neurons in the first hidden layer and 64 neurons in the second, with ReLU activation.
- In case of Home Credit dataset, a smaller architecture (64 -32neurons) was adequate as the features were of lower dimension.
- The dropout rate (0.3) was implemented to reduce overfitting and Adam optimizer was employed with 0.001 learning rate.
- Training was done over 100 epochs and the early stopping was done based on validation loss.

2.7 Data Preprocessing

To be able to apply feature selection and machine learning models, both datasets (TCGA gene expression and Home Credit Default Risk) underwent standardized preprocessing steps to make the two datasets similar to each other, minimize bias, and enhance model stability.

Normalization and Scaling

Received: August 09, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- In the case of the TCGA gene expression dataset, the raw counts of the RNA-seq were transformed with the log2 to stabilize the variance among the genes.
- Both datasets were standardized using Z-score, which brought the means of each feature to zero and a unit variance. This measure was necessary to make sure that items that differ in their size (e.g., the level of gene expression and the level of financial credit) are equally useful in training the model.
- In the case of SVM and Neural Networks, min max scaling was also tested whereby features were rescaled to [0,1], and this enhanced speed in optimization.

Handling Missing Values

- Missing values of gene expression were infrequent in the TCGA data, and when there, they were filled in with the use of k-nearest neighbors (k = 5) to maintain biological relationships.
- Missing demographic and financial variables were more prevalent in-Home Credit dataset. Median values were used in the imputation of continuous variables whereas mode was used in the imputation of categorical variables.
- Features with more than 30% missingness were excluded to avoid introducing bias through imputation.

Feature Encoding

- Categorical variables in the Home Credit dataset (e.g., gender, education level) were encoded using one-hot encoding to allow compatibility with linear and non-linear models.
- For high-cardinality categorical features, frequency encoding was applied to reduce dimensionality.

Outlier Treatment

- Outliers in financial variables (e.g., extreme loan amounts, unusually long employment durations) were capped at the 1st and 99th percentiles to minimize distortion in model training.
- Gene expression outliers were retained, as they may represent biologically significant signals.

Data Splitting

- Both datasets were partitioned into training (70%), validation (15%), and test (15%) sets using stratified sampling to preserve class balance.
- Stratified 10-fold cross-validation was then applied to the training set to tune hyperparameters and assess generalization.

2.7 Cross-Validation Performance of Elastic Net–SVM Pipeline

To evaluate the generalization capability of the proposed feature selection and classification framework, we conducted stratified 10-fold cross-validation on both datasets using the Elastic Net–SVM pipeline.

Table4. Cross-Validation Results (Mean ± SD)

Received: August 09, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Dataset	Method	Accuracy (Mean ± SD)	F1-Score (Mean ± SD)
TCGA	Elastic Net + SVM	92.6% ± 1.2%	91.6% ± 1.4%
Home Credit	Elastic Net + SVM	87.3% ± 1.5%	85.8% ± 1.6%

Table 4 reports mean ± standard deviation from 10-fold cross-validation. Elastic Net with SVM showed high consistency on both datasets (TCGA: 92.6% ± 1.2% accuracy, 91.6% ± 1.4% F1-score; Home Credit: 87.3% ± 1.5% accuracy, 85.8% ± 1.6% F1-score), confirming its reliability and generalizability. These low standard deviations indicate stable performance across folds, reinforcing Elastic Net's robustness in high-dimensional settings**.**

2.8 Assumptions for ANOVA and t-tests

In order to make the inferential statistics used in this research sound, we tested the main assumptions of the ANOVA and independent t-tests.

Table 5. Validation of Statistical Assumptions for ANOVA and t-tests

Assumption	Test Used	Dataset	Test Statistic	p-value	Interpretation
Normality of Residuals	Shapiro-Wilk Test	TCGA	W = 0.972	0.087	Residuals are approximately normal (p > 0.05)
		Home Credit	W = 0.968	0.092	Normality assumption holds (p > 0.05)
Homogeneity of Variance	Levene's Test	TCGA	F = 1.42	0.241	Equal variances assumed across groups (p > 0.05)
		Home Credit	F = 1.67	0.198	Homogeneity of variance confirmed (p > 0.05)

Table 5 the Shapiro-Wilk test verified that the residues of the model performance metrics (e.g., F1-score) had an approximate normal distribution in both datasets (TCGA and Home Credit) with a p-value above the 0.05 mark. The test by Levene also revealed that the variance was homogenous across the model groups hence use of parametric tests. Structural independence of observations was ensured by stratified 10 fold cross validation, which removed overlap between test and training sets. The reliability of the reported F-statistics and p-values, including the significant ANOVA value ($F = 12.84$, $p = 0.0003$ for TCGA) and pairwise t-tests (e.g., Elastic Net vs. Ridge, $t = 4.87$, $p = 0.001$) can be verified by these validations, which state that the difference in performance across models is not statistically negligible.

Received: August 09, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

2.9 Comparative Evaluation of Feature Selection Methods Using F1-Score Metrics (Anova and t-test)

In order to evaluate the predictive abilities of different feature selection methods, we presented an extensive comparison of these methods based on the F1-score as the evaluation metric.

Table 6. F1-Score Comparison Across Models through ANOVA test

Dataset	F-Statistic	p-Value	Interpretation
TCGA	12.84	0.0003	Significant differences in F1-scores across models
Home Credit	9.67	0.0011	Significant differences in F1-scores across models

Table 7. F1-Score Comparison Across Models

Dataset	Comparison	t-Statistic	p-Value	Interpretation
TCGA	Elastic Net vs. LASSO	3.21	0.007	Elastic Net significantly better than LASSO
TCGA	Elastic Net vs. MI	4.02	0.003	Elastic Net significantly better than MI
TCGA	Elastic Net vs. Ridge	4.87	0.001	Elastic Net significantly better than Ridge
Home Credit	Elastic Net vs. LASSO	2.94	0.011	Elastic Net significantly better than LASSO
Home Credit	Elastic Net vs. MI	3.76	0.005	Elastic Net significantly better than MI
Home Credit	Elastic Net vs. Ridge	4.21	0.002	Elastic Net significantly better than Ridge

It was established by Table 6 and Table 7 that Elastic Net is a consistent winner over other feature selection techniques in both datasets.

2.9 SHAP-Based Interpretability Across Domains

SHAP (Shapley Additive explanations) analysis was used to provide more model transparency and feature-level insight in both datasets. SHAP values help to measure the value of each

Received: August 09, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
feature to individual predictions, allowing one to understand model behavior on a granular level.

Table 8. SHAP-Based Interpretability Summary

Data sets	Feature Selection	Top Features / Genes	SHAP Distribution Characteristics	Interpretability Impact
TCGA	Elastic Net	TP53, BRCA1, MYC	Concentrated SHAP values	High interpretability, focused gene influence
TCGA	LASSO	TP53, BRCA1	Reduced noise, tighter SHAP spread	Improved clarity via feature sparsity
TCGA	Mutual Information	MYC, BRCA1, TP53	Wider SHAP variance due to non-linear interactions	Captures complex patterns, less stable
Home Credit	Elastic Net	EXT_SOURCE_3, DAYS_EMPLOYED, AMT_CREDIT	Tighter SHAP distributions	Strong interpretability, focused feature impact
Home Credit	LASSO	EXT_SOURCE_3, AMT_CREDIT	Sparse and concentrated SHAP values	Clear feature relevance
Home Credit	Ridge Regression	All features retained	Diluted SHAP importance across many features	Lower interpretability due to feature saturation

The Elastic Net-SVM pipeline in the TCGA data in table 9 gave the most concentrated SHAP distributions, with biologically significant genes, including TP53, BRCA1, and MYC. LASSO was capable of effectively eliminating noisy genes and enhancing interpretability with Mutual Information being able to capture non-linear gene interactions but with broader SHAP variance. In the case of Home Credit dataset, the best financial predictors were EXTSOURCE3, DAYSEMPLOYED and AMTCREDIT. Elastic Net and LASSO produced more concentrated SHAP distributions, which shows a concentrated influence of features. Ridge Regression, by contrast, kept all the features, spreading the SHAP importance among less important variables.

2.10 Computational Efficiency Analysis of Feature Selection Techniques

Received: August 09, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
As a complement to the predictive performance evaluation, we have evaluated the computational efficiency of each feature selection approach with regard to the training time and memory consumption.

Table 10. Computational Efficiency

Technique	Training Time Reduction	Memory Usage Reduction
LASSO	~40%	~35%
Elastic Net	~38%	~30%
Ridge Regression	~25%	~20%
Mutual Information	~30%	~25%

Table 10 gives a comparison of computational metrics. Elastic Net generated feature space, training time, and memory usage were cut by 65, 40, and 30, respectively, and was therefore the most efficient approach. LASSO came right behind with 60% reduction in features and 35% reduction in training time. Mutual Information provided moderate efficiency gains whereas Ridge Regression did not drop anything and hence had little computational gain. These results indicate the scalability advantages of embedded selection techniques in machine learning pipelines.

3. Results and Discussion

The following section elaborates on the empirical results of implementing statistical feature selection methods LASSO, Ridge Regression, Elastic Net, and Mutual Information on three machine learning models, including the Random Forest, Support Vector Machine (SVM), and the Neural Networks. To evaluate it, two high-dimensional datasets were used, including gene expression profiles on TCGA to classify cancer and Home Credit Default Risk dataset to predict financial risks. These findings are interpreted with reference to predictive performance, interpretability and computational efficiency.

3.1 Predictive Performance

The model accuracy as well as variance decreased greatly because of the feature selection in both datasets. Elastic Networks with the advantage of L1 and L2 regularization in the TCGA dataset, and it was much better than other approaches in dealing with correlated genes features. Elastic Net was the highest classification accuracy of 92.6, F1-score 91.6, when paired with SVM which means that there is a great balance between precision and recall. LASSO was also good, especially on Random Forest with 91.2% accuracy and overfitting minimized through removal of irrelevant genes. Elastic Net also outperformed in the home credit dataset, with SVM (87.3% accuracy, F1-score of 85.8), but Ridge Regression had consistent results but failed to drop any features, slightly diluting its performance. The benefits of the Neural Networks were the mutual information that modeled non-linear dependencies and got 90 percent accuracy and strong recall. These findings indicate that statistical feature selection

Received: August 09, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
improves generalization and model strength particularly in high dimensional domains where there is feature redundancy.

3.2 Interpretability

Interpretability was evaluated with the help of SHAP (SHapley Additive exPlanations) which rates features with the same level of importance between models. The SHAP profiles were more concise and understandable with models that were trained on statistically chosen features. Elastic Net and LASSO were used in the TCGA dataset to identify biologically relevant genes including TP53, BRCA1, and MYC, which are recognized cancer markers. SHAP value of these genes was significantly bigger, which means that they play the leading role in the classification. Although not a part of model training, Mutual Information did a good job of ranking interactions between non-linear genes which were ignored by linear models. SHAP analysis in Home credit dataset showed that the features, such as EXTSOURCE3, DAYSEMPLOYED, and AMTCREDIT, were always ranked first in all the models with feature selection. Ridge Regression that kept all the features indicated a flatter SHAP distribution, and thus, interpretation is more diffuse. On the whole, the feature selection process helped to focus the model on more important predictors, making it more transparent and enabling domain-specific understanding.

3.3 Computational Efficiency

Computational efficiency gains were also significant because of feature selection. In both data sets, training with LASSO and Elastic Net was up to 40 percent faster and used less memory by about a third of models trained using the complete sets of features. This was more so in the Neural Networks where high-dimensional input layers usually have heavy computational costs. Although Ridge Regression made the coefficients stable, all features were retained and therefore, it demonstrated very little resource consumption. As a filter technique, Mutual Information was computationally efficient in the preprocessing phase but failed to minimize model complexity in the training phase. These results highlight the usefulness of statistical feature selection in scalable machine learning piping, particularly when models need to be used in resource-constrained settings or real-time systems.

4. Expanded Discussion

The cross-domain findings of the comparative analysis of statistical feature selection techniques in genomics and financial domain give a number of cross domain findings. Elastic Net with SVM demonstrated a higher performance in the genomics dataset (TCGA), which is indicative of the fact that it is capable of handling the extremely correlated variables in terms of gene expression along with being sparse. LASSO too was effective as it removes noisy genes hence improving Random Forest generalization. There was Mutual Information which captured non linear interaction of genes, but its interpretability was influenced by broader SHAP variance. Conversely, the financial data (Home Credit) showed that the Elastic Net was once again able to offer the most stable and precise data, which is indicative of its flexibility to structured and tabular data with engineered characteristics. LASSO enhanced neural network generalization as it paid attention to the most meaningful predictors, whereas Ridge Regression

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
kept all features, which depreciated the interpretability. Such results imply that Elastic Net is a strong domain agnostic solution whilst LASSO and Mutual Information are domain specific.

Although these are encouraging outcomes, there are a number of limitations that have to be noted. To start with, the sampling biases can affect the generalizability: TCGA samples of particular cancer types, whereas Home Credit data includes the customers of a given financial institution. Second, the statistical power of genomics (as compared to the number of features) may be limited by sample size constraints, whereas the financial datasets (although larger) might be affected by the artificial creation of redundant features. Third, computational issues are also an obstacle; as the feature selection minimized training time and memory consumption, high dimensional models like neural networks continue to consume a large number of resources. Lastly, SHAP-based interpretability assessments can be informative but can change between feature selection strategy and model choice, which should be cautioned against in clinical or regulatory applications.

In the future, there are a number of ways that can be identified. More balanced solutions across domains could be offered by hybrid frameworks of feature selection, combining statistical rigor (e.g., Elastic Net) with model agnostic interpretability (e.g., SHAP). The approaches to federated learning can enable analysis of distributed genomic and financial data without privacy breaches and solve the ethical and regulatory issues. The still emerging quantum enhanced feature selection techniques have potential to speed up dimensionality reduction in ultra-high dimensional datasets. Additionally, the future research must look into multi modal integration whereby genomic, clinical and financial indicators are integrated to come up with holistic predictive models. These guidelines reflect the promise of feature selection as a technical optimization procedure and as a foundation of scalable, clarifiable, and responsible machine learning.

4. Conclusion

This paper concludes that statistical feature selection methods; especially Elastic Net and LASSO; can make machine learning models used with high-dimensional data including gene expression profiles and financial risk indicators much more interpretable, predictive, and computationally efficient. All these techniques enhance the generalization and transparency of models by reducing dimensionality and isolating the most informative features as reflected in concentrated SHAP values and a lower training overhead. Elastic Net has been shown to have better performance over other methods in Random Forest, SVM and Neural Network particularly in predictors that are correlated. To conduct further studies, it is suggested to foster their application to deep learning systems and time-series data and automated pipelines, and combine them with more sophisticated explainable AI tools to facilitate scalable, domain-conscious, and reliable machine learning applications.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2015). Recent advances and emerging challenges of feature selection in the context of big data. *Knowledge-based systems, 86*, 33-45.

Received: August 09, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
2. Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & electrical engineering, 40*(1), 16-28.
3. Credit, H. (2018). Home credit default risk. *Kaggle*.
4. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research, 3*(Mar), 1157-1182.
5. Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning, 46*(1), 389-422.
6. Hancock, J. T., Khoshgoftaar, T. M., & Liang, Q. (2025). A problem-agnostic approach to feature selection and analysis using shap. *Journal of Big Data, 12*(1), 12.
7. Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55-67.
8. Kamalov, T., & Kamalov, Y. T. (2025). *Physics of Non-Inertial Reference Frames, conclusions and consequences.* Paper presented at the Journal of Physics: Conference Series.
9. Kaur, S., Smiley, C., Gupta, A., Sain, J., Wang, D., Siddagangappa, S., . . . Shah, S. (2023). *REFinD: Relation extraction financial dataset.* Paper presented at the Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval.
10. Khan, S., Mazhar, T., Naz, N. S., Ahmed, F., Shahzad, T., Ali, A., . . . Hamam, H. (2025). Advanced Feature Selection Techniques in Medical Imaging—A Systematic.
11. Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In *Machine learning proceedings 1992* (pp. 249-256): Elsevier.
12. Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence, 97*(1-2), 273-324.
13. Kumar, V., & Minz, S. (2014). Feature selection. *SmartCR, 4*(3), 211-229.
14. Lin, W., Gao, Q., Yuan, J., Chen, Z., Feng, C., Chen, W., . . . Tong, T. (2020). Predicting Alzheimer's disease conversion from mild cognitive impairment using an extreme learning machine-based grading method with multimodal data. *Frontiers in aging neuroscience, 12*, 77.
15. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems, 30*.
16. Mei, H., Peng, J., Wang, T., Zhou, T., Zhao, H., Zhang, T., & Yang, Z. (2024). Overcoming the limits of cross-sensitivity: pattern recognition methods for chemiresistive gas sensor array. *Nano-micro letters, 16*(1), 269.
17. Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence, 27*(8), 1226-1238.
18. Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O'Sullivan, J. M. (2022). A review of feature selection methods for machine learning-based disease risk prediction. *Frontiers in bioinformatics, 2*, 927312.

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
19. Qi, R. (2025). Enterprise Financial Distress Prediction Based on Machine Learning and SHAP Interpretability Analysis.

20. Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *bioinformatics, 23*(19), 2507-2517.

21. Setiono, R., Baesens, B., & Mues, C. (2011). Rule extraction from minimal neural networks for credit card screening. *International journal of neural systems, 21*(04), 265-276.

22. Tadist, K., Najah, S., Nikolov, N. S., Mrabti, F., & Zahi, A. (2019). Feature selection methods and genomic big data: a systematic review. *Journal of Big Data, 6*(1), 1-24.

23. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 58*(1), 267-288.

24. Tsai, C.-F., & Chen, M.-L. (2010). Credit rating by hybrid machine learning techniques. *Applied soft computing, 10*(2), 374-380.

25. Witten, D. M., & Tibshirani, R. (2009). Covariance-regularized regression and classification for high dimensional problems. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 71*(3), 615-636.

26. Zhou, X., Wen, H., Li, Z., Zhang, H., & Zhang, W. (2022). An interpretable model for the susceptibility of rainfall-induced shallow landslides based on SHAP and XGBoost. *Geocarto International, 37*(26), 13419-13450.

Received: August 09, 2025