

Mathematical Methods in Feature Selection: A Review

Firuz Kamalov ^{1,*}, Hana Sulieman ², Ayman Alzaatreh ², Maher Emarly ², Hasna Chamlal ³
and Murodbek Safaraliev ⁴

¹ Department of Electrical Engineering, Canadian University Dubai, Dubai P.O. Box 117781, United Arab Emirates

² Department of Mathematics and Statistics, American University of Sharjah, Sharjah P.O. Box 26666, United Arab Emirates; hsulieman@aus.edu (H.S.); aalzaatreh@aus.edu (A.A.); memarly@aus.edu (M.E.)

³ Computer Science and Systems Laboratory (LIS), Faculty of Sciences Ain Chock, Hassan II University of Casablanca, Casablanca 20360, Morocco; hasna.chamlal@univh2c.ma

⁴ Ural Power Engineering Institute, Ural Federal University, Yekaterinburg 620002, Russia; murodbek_03@mail.ru

* Correspondence: firuz@tud.ac.ae

Abstract: Feature selection is essential in machine learning and data science. Recently, there has been a growing effort to apply various mathematical methods to construct novel feature selection algorithms. In this study, we present a comprehensive state-of-the-art review of such techniques. We propose a new mathematical framework-based taxonomy to group the existing literature and provide an analysis of the research in each category from a mathematical perspective. The key frameworks discussed include variance-based methods, regularization methods, and Bayesian methods. By analyzing the strengths and limitations of each technique, we provide insights into their applicability across various domains. The review concludes with emerging trends and future research directions for mathematical methods in feature selection.

Keywords: feature selection; mathematical methods; machine learning; variance; Bayesian; regularization; classification; regression

MSC: 62R07; 62F07; 62J05; 62J07; 68T99



Academic Editor: Nicolae Herisanu

Received: 15 February 2025

Revised: 14 March 2025

Accepted: 17 March 2025

Published: 18 March 2025

Citation: Kamalov, F.; Sulieman H.; Alzaatreh, A.; Emarly, M.; Chamlal, H.; Safaraliev, M. Mathematical Methods in Feature Selection: A Review.

Mathematics **2025**, *13*, 996. <https://doi.org/10.3390/math13060996>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

1.1. Motivation

In the era of big data, datasets have grown exponentially in both size and complexity, posing significant challenges to data analysis and machine learning. As a result, feature selection has become a key preprocessing step that aims to reduce the dimensionality of data by selecting a subset of relevant features that contribute most to the predictive performance of models. It enhances model accuracy and interpretability, as well as reducing computational costs and mitigating the risk of overfitting [1–3].

Traditional feature selection techniques often rely on heuristic approaches or domain expertise. However, modern feature selection methods have increasingly leveraged rigorous mathematical theories and optimization techniques based on linear algebra, information theory, statistics, sparse representation, and optimization algorithms. Mathematical approaches provide a structured and quantifiable foundation for feature selection, enabling more robust and scalable solutions across diverse application domains.

Despite the extensive body of literature on feature selection, including numerous review articles, there is no comprehensive analysis specifically focused on the mathematical techniques underpinning feature selection. Given the increasing significance of math-

ematical methods in this domain, there is a pressing need for a dedicated review that systematically examines these approaches.

1.2. Related Work

Feature selection is widely recognized as a critical process for data preprocessing in machine learning, prompting multiple survey articles that examine the topic from various perspectives. For instance, Dokeroglu et al. [4] offer a comprehensive survey of metaheuristic-based feature selection algorithms, assessing advances in exploration and exploitation operators, parameter tuning, and fitness evaluation functions. Their work emphasizes the rapid proliferation of new metaheuristic methods and explores current challenges and potential research directions in that domain. Similarly, Dhal and Azad [5] provide an extensive review of feature selection in machine learning, focusing on structured and unstructured datasets and discussing the impact of high-dimensionality, noise, and storage complexity on the feature selection process. They compile a broad taxonomy of existing methods and applications, including commonly used datasets and performance metrics across different ML fields.

Pudjihartono et al. [6] shift the perspective to disease risk prediction, illustrating how feature selection methods can help identify the most informative features (e.g., SNPs) in high-dimensional genetic data. By addressing the “curse of dimensionality”, their review highlights the benefits of feature selection in improving the generalizability of machine learning models, albeit at the cost of handling large, noisy datasets. In contrast, swarm intelligence-based feature selection is the focal point of a survey by Rostami et al. [7], who evaluate emerging swarm intelligence algorithms and discuss their strengths and weaknesses in terms of computational complexity, convergence properties, and practical performance.

Although these prior works collectively cover a broad spectrum of approaches—ranging from metaheuristics to domain-specific implementations—they tend to concentrate on high-level algorithmic designs, application-specific issues, or comparative evaluations of heuristic techniques. Consequently, there remains a gap in the literature concerning the underlying mathematical frameworks that unify these diverse feature selection methods. The present review aims to bridge this gap by examining feature selection methods through a mathematical lens, focusing on three principal frameworks—variance-based, regularization-based, and Bayesian. By discussing the theoretical underpinnings, strengths, and potential limitations of each framework, this review intends to complement and extend existing surveys, offering a more in-depth exploration of the core mathematical principles that govern feature selection.

1.3. Methodology

To ensure that the reviewed works reflect leading-edge research and foundational advancements, we employed a systematic literature search strategy. First, major academic databases (Scopus and Google Scholar) were queried with keywords such as “feature selection”, “mathematical methods”, “sparse modeling”, and “Bayesian feature selection.” The initial search was restricted to publications from the last ten years to capture recent progress, but earlier seminal papers were included where necessary to provide historical context.

Next, the studies were screened for relevance based on their abstracts and titles, and only those discussing mathematically grounded approaches such as theoretical proofs, rigorously formulated optimization algorithms, or well-established statistical or Bayesian frameworks, were retained. Additional filtering was conducted by examining citation counts, the reputation of the publishing venue, and whether the work included a clear articulation of the mathematical principles involved. This vetting process resulted in a targeted corpus of articles spanning various mathematical paradigms. Ultimately, we

synthesized these findings to propose a new taxonomy and to highlight the most promising directions for future research in mathematical feature selection.

1.4. Contributions

This paper presents a comprehensive review of modern mathematical approaches to feature selection. We classify these techniques based on their mathematical foundations, including variance-based methods, regularization techniques, and Bayesian approaches. Each method is described in detail, highlighting its key strengths and limitations. Additionally, we discuss practical applications for each technique, providing researchers with insights to select the most suitable feature selection method for their specific needs.

The key contributions of the paper are the following:

1. We categorize feature selection methods based on their mathematical foundation.
2. We provide technical backgrounds of the main techniques and discuss their strengths and weaknesses.
3. We offer guidance for researchers to select appropriate feature selection methods based on specific application domains.

The remainder of this paper is organized as follows: Section 2 provides an overview of mathematical feature selection methods, which lays the foundation for the subsequent sections. Section 3 introduces variance-based approaches, while Section 4 examines regularization-based techniques. Section 5 delves into Bayesian feature selection methods. Section 6 presents related methods, including hybrid strategies and graph-theoretical approaches for feature selection. Section 7 discusses current challenges and highlights potential research directions. Finally, Section 8 concludes this paper with closing remarks.

2. Overview of Mathematical Feature Selection Methods

Mathematical approaches to feature selection can be broadly categorized according to the underlying theoretical frameworks used to rank or eliminate predictors. In particular, three families of methods have been extensively studied in the literature: *variance-based*, *regularization-based*, and *Bayesian* methods (Figure 1). These approaches have distinct mathematical formulations but share a common objective of identifying informative features to improve model accuracy, interpretability, and computational efficiency. In this section, we provide an overview of these principal mathematical frameworks, serving as a roadmap for the subsequent sections of this paper.

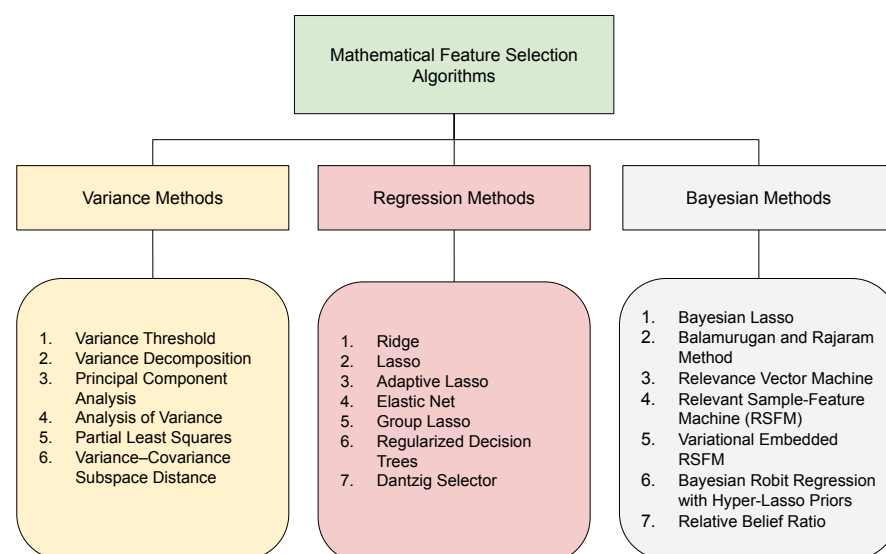


Figure 1. Framework for mathematical approaches to feature selection.

2.1. Variance-Based Feature Selection

Variance-based methods rely on measures of dispersion to evaluate the usefulness of features. Their fundamental premise is that a feature with higher variance (or that better captures the variance–covariance structure of the dataset) is more likely to be informative, especially when the target variable is unknown or only weakly informative. Simple approaches, such as removing features below a variance threshold, can efficiently filter out near-constant or degenerate predictors. More sophisticated techniques decompose the variance of a response variable among the predictors to isolate their individual and joint contributions. Although variance-based methods can be fast and intuitive, they often do not consider the target variable, unless combined with supervised extensions (e.g., ANOVA-based criteria). Section 3 discusses key variance-driven methods, including principal component analysis (PCA), ANOVA, and Sobol variance decomposition.

2.2. Regularization-Based Feature Selection

Regularization methods embed feature selection within the model-building process by adding penalty terms to the model’s objective function. The penalties—often based on norms such as L_1 (Lasso) or a combination of L_1 and L_2 (elastic net)—shrink the coefficients of less informative features to zero, thus automatically yielding a sparse model. This *embedded* nature makes regularization particularly attractive in high-dimensional scenarios where the number of features may vastly exceed the number of samples. While classical regularization frameworks focus on linear models, more recent developments explore penalization in tree-based methods, support vector machines, and other nonlinear models. Notably, each regularization scheme offers a distinct bias–variance trade-off and various assumptions regarding the data (e.g., correlated predictors or group structures). Section 4 presents widely used regularization-based approaches along with their mathematical formulations, strengths, and computational challenges.

2.3. Bayesian-Based Feature Selection

Bayesian feature selection approaches build on probability theory and Bayes’ theorem to incorporate prior information and uncertainty quantification directly into the modeling process. By treating model parameters and even the regularization hyperparameters as random variables, these techniques can better capture posterior distributions and deliver probabilistic statements about feature importance. Sparsity-inducing priors, such as Laplace or heavy-tailed distributions, facilitate the elimination of irrelevant features while preserving the crucial ones. Extensions range from Bayesian Lasso to more advanced kernels and hierarchical models (e.g., Relevance Vector Machine or Bayesian robust regression). One of the key advantages of Bayesian approaches is their interpretability in terms of posterior probabilities, though this benefit often comes at the cost of more intensive computations (e.g., Markov chain Monte Carlo sampling). Detailed discussions of these methods are provided in Section 5.

2.4. Connections and Comparisons

Although each of the above frameworks—variance-based, regularization-based, and Bayesian—arises from distinct mathematical theories, they are not mutually exclusive. For instance, variance-based screening can serve as an initial filter before applying a Lasso model, thereby reducing dimensionality while controlling computational overhead. Similarly, Bayesian analogs of regularization methods, such as Bayesian Lasso, combine the embedding of sparse penalties with posterior inference. Each class of methods balances bias–variance trade-offs differently and offers unique insights into the data.

3. Variance-Based Feature Selection

Variance measures the dispersion of the data. It is an important concept in statistics and has been widely used as a foundation for many feature selection algorithms. There exists a range of variance-based methods starting from the simple variance threshold method to the sophisticated Sobol sensitivity index. These methods leverage variance to identify informative, non-redundant, and discriminative features, enabling more efficient and effective data representation for downstream machine learning tasks.

In this section, we discuss six major variance-based methods. We describe each method together with its mathematical formulation. In addition, we discuss their strengths and limitations. A summary of the variance-based methods is provided in Table 1. A more detailed discussion of each method is provided in the following subsections.

Table 1. Summary of variance-based feature selection methods.

Method	Strengths	Limitations	Supervision	Use
Variance Threshold [8]	Simple and fast to compute.	Ignores the target. Threshold choice is heuristic.	Unsupervised	Classification, Regression
Variance Decomposition (Sobol) [9]	Captures nonlinear interactions. Relates directly to output variance.	Computationally intensive. Requires a specific model. Suitable mainly for simulation contexts.	Supervised	Regression
PCA [10]	Effective for high-dimensional data. Captures maximum variance directions and can reduce noise.	No target usage. Limited interpretability.	Unsupervised	Classification, Regression
ANOVA [8]	Relies on the F-statistic and proven under certain conditions.	Assumes normality and equal variances. Univariate and does not capture feature interactions.	Supervised	Classification
Partial Least Squares [11]	Incorporates correlation with the target. Handles collinearity well in high-dimensional data.	Produces latent factors that reduce interpretability. Choosing the number of components is often heuristic.	Supervised	Regression
VCSD [12]	Unsupervised approach that preserves the covariance structure. Attempts to minimize the difference between subspaces.	The optimization can be combinatorial and may require heuristic search. Can be challenging for large problems.	Unsupervised	Classification, Regression

3.1. Variance Threshold

Variance threshold is a filter approach that discards features whose variances are below a chosen cutoff. It relies on the intuition that features with little or no variance are unlikely to offer much predictive power. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be the data with n samples and d

features. Denote X_{ij} as the value of feature j in sample i . The sample mean of feature j is $\bar{x}_j = \frac{1}{n} \sum_{i=1}^n X_{ij}$, and its sample variance is

$$\sigma_j^2 = \frac{1}{n-1} \sum_{i=1}^n (X_{ij} - \bar{x}_j)^2. \quad (1)$$

Let θ be a pre-defined threshold. Then, the feature is removed, if $\sigma_j^2 < \theta$. The threshold is usually chosen as a small positive value to remove near-constant features. The approach is computationally cheap and easy to interpret [13]. However, it overlooks the possibility that some low-variance features may be important for specific predictive tasks. It is common to use this method as an initial data-cleaning step to handle degenerate or trivial features. Variance threshold is an unsupervised method as it does not take into account the values of the target variable. As such, it may miss relevant interactions between the dependent and independent variables.

The variance threshold method can be useful both as a standalone method and in combination with more sophisticated methods [14]. A study of 14 filter methods for high-dimensional gene expression data found the variance threshold to achieve the optimal results [8]. The variance threshold method was used in conjunction with Boruta to select optimal features for IDS [15].

3.2. Variance Decomposition

Sobol indices are based on the decomposition of the variance of the target variable in terms of the input variables. It is often applied in engineering- and simulation-based studies. Sobol decomposition is a sophisticated method that can capture complex interactions between multiple features and the target variable. There are two types of Sobol indices used in feature selection: the first-order index and the total-order index.

The method requires a model

$$Y = f(X_1, X_2, \dots, X_d),$$

where each X_i is an input feature. The output variance

$$V = \text{Var}[Y]$$

is decomposed according to how each feature or combination of features contributes to it. The Hoeffding—Sobol decomposition is given by

$$f(\mathbf{X}) = f_0 + \sum_{i=1}^d f_i(X_i) + \sum_{1 \leq i < j \leq d} f_{ij}(X_i, X_j) + \dots, \quad (2)$$

with each term representing partial variances. The first-order Sobol index for feature i is

$$S_i = \frac{V_i}{V}, \quad (3)$$

where V_i is the variance in the output that is exclusively attributable to X_i . The total-order Sobol index is given by

$$S_{T_i} = \frac{V_i + \sum_{i \in u} V_u}{V}, \quad (4)$$

where V_u is the variance of Y attributed exclusively to the subset u of the feature set. The total-order index captures all variance contributions of X_i alone and in combination with other variables. Low indices indicate that a feature can be removed with little impact on output variance.

The Sobol index approach can discover nonlinear interactions that simpler correlation-based measures miss [9]. However, it is computationally expensive because the model must be evaluated multiple times under different sampling schemes. Additionally, the method requires a specific model which reduces the generalizability of the results.

It is best suited for regression problems with a definable model of how inputs affect an output quantity. It has been used in a variety scenarios including gene expression [16], tunneling-induced settlement prediction [17], drilling [18], intrusion detection [19], and others.

3.3. Principal Component Analysis

PCA is an unsupervised dimensionality reduction method that projects data onto orthogonal directions of greatest variance. Each direction, known as a principal component, is formed as a linear combination of the original features. While PCA is originally employed as a feature extraction technique, it can also be used to select the optimal features [10,20]. Many studies employ 2–3 principal components to visualize and analyze the data [10].

PCA is calculated by diagonalizing the covariance matrix. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be mean-centered so that each feature column has an average of zero. Its covariance matrix is

$$\Sigma = \frac{1}{n-1} \mathbf{X}^\top \mathbf{X}. \quad (5)$$

The eigenvalue decomposition of Σ yields

$$\Sigma \mathbf{V} = \mathbf{V} \Lambda, \quad (6)$$

where $\mathbf{V}$ contains eigenvectors and Λ the corresponding eigenvalues. The eigenvalues represent the amount of variance captured by each principal component. The top k components can be selected to reduce the dimensionality while retaining most of the variance.

PCA does not use a target variable, so it may discard low-variance features that are actually predictive in a supervised setting. Despite this, it is widely used for noise removal and data compression, especially in large-scale tasks. It produces new latent variables that combine original features, which can complicate interpretability.

Although PCA produces linear combinations of features, the weights of principal components can be used to measure feature importances [21,22]. PCA-based feature selection algorithms have been used in a variety applications, including hyperspectral remote sensing [23], medical diagnosis [24], gene expression [25], and others.

3.4. Analysis of Variance

ANOVA-based feature selection typically applies to classification tasks. Each feature is tested to see how well it separates the classes by comparing its between-class variance to its within-class variance [8]. Suppose there are k classes, with n_m samples in class m . The mean of feature j in class m is $\bar{x}_{j,m}$, and the overall mean is $\bar{x}_j$. The between-class sum of squares is

$$\text{SSB}_j = \sum_{m=1}^k n_m (\bar{x}_{j,m} - \bar{x}_j)^2, \quad (7)$$

and the within-class sum of squares is

$$\text{SSW}_j = \sum_{m=1}^k (n_m - 1) s_{j,m}^2. \quad (8)$$

The F-statistic for feature j is

$$F_j = \frac{SSB_j / (k - 1)}{SSW_j / (n - k)}. \quad (9)$$

A higher F-statistic indicates greater discriminative power. ANOVA is straightforward to implement and relies on established statistical theory. However, it assumes normality and homogeneity of variance across classes, which may not always hold in real-world data. It also considers each feature individually and does not capture interactions.

ANOVA-based feature selection has shown robust results in multiple studies [5,26]. It has been applied in various fields, including computer vision [27,28], protein identification [29], genomic data [30], and others. There also exist several extensions of ANOVA that aim to improve feature selection in specific domains. Differential variance was proposed in [31] to improve feature selection for cancer detection. An extension of ANOVA for imbalanced data was proposed in [32].

3.5. Partial Least Squares

PLS is a supervised technique that, like PCA, projects data into a lower-dimensional latent space, but uses the relationship with the response variable to guide the extraction of factors. After centering and scaling, PLS finds weight vectors that maximize the covariance between projections

$$\mathbf{t} = \mathbf{X}\mathbf{w}, \quad \mathbf{u} = \mathbf{Y}\mathbf{c}. \quad (10)$$

The data matrices are then deflated to remove the explained variation, and the process repeats to extract multiple components. By aligning variance in $\mathbf{X}$ with variance in $\mathbf{Y}$, PLS can handle strongly collinear predictors. Its drawback is that the resulting components are linear combinations of the original predictors, which can make interpretation difficult. Selecting the correct number of components often relies on cross-validation or domain expertise.

PLS is popular in high-dimensional regression contexts [11,33]. It can be further adapted for classification under variations such as discriminant analysis (PLS-DA), path modeling (PLS-PM) [34,35], the kernel technique (KPLS) [36], and others.

3.6. Variance–Covariance Subspace Distance-Based Feature Selection (VCSD)

Most recently, variance–covariance subspace distance-based (VCSD) feature selection was proposed in [12]. VCSD is an unsupervised approach designed to select a subset of features whose spanned subspace best approximates the entire data matrix in terms of its variance–covariance structure. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be the original data and let I be a set of feature indices with $|I| = k$. The submatrix $\mathbf{X}_I$ includes only the features in I . VCSDFS seeks to minimize

$$\|\mathbf{X}\mathbf{X}^\top - \mathbf{X}_I\mathbf{X}_I^\top\|_F^2, \quad (11)$$

where the Frobenius norm $\|\cdot\|_F$ measures the difference in the covariance-like structure of the full data versus the selected subset. The idea is to preserve as much of the original subspace geometry as possible. This objective can be combinatorially large, so approximate or greedy methods may be employed for practical data sizes. It serves as a valuable tool for situations where no target labels exist but the user wishes to maintain global variance–covariance patterns.

3.7. Discussion

The above methods reveal different ways of leveraging variance in feature selection. Variance threshold and PCA are unsupervised methods that do not take into account the target variable. The simplest approach is variance threshold, which can quickly eliminate

constant or near-constant features but ignores any target or deeper data structure, while PCA reduces dimensionality by retaining directions of maximal variance. Supervised methods include ANOVA, Sobol indices, and PLS. ANOVA can rank features by their ability to distinguish classes, though it makes strong statistical assumptions. Partial Least Squares aligns predictor variance with a response variable, making it suitable for high-dimensional regression, albeit with interpretability trade-offs. Variance Decomposition (Sobol indices) can capture complex nonlinear interactions, but demands a callable model and can be computationally intensive. More recently, VCSD was introduced, which preserves the global variance–covariance subspace structure of the data, though it may require complex optimization and is best approached with suitable heuristics for large datasets.

In many practical settings, a combination of these methods is employed to capitalize on their respective advantages. Researchers might first remove constant features using a threshold, then reduce redundancy by examining correlations or decompositions, and finally refine the feature set with a target-aware method like PLS or ANOVA. The choice of strategy depends on whether the task is supervised or unsupervised, on the size of the dataset, and on the nature of the domain from which the data originate. The methods outlined here represent fundamental approaches that highlight the central role of variance-related criteria in feature selection.

4. Regularization-Based Feature Selection

In the embedded feature selection framework, regularization or sparse learning is incorporated during model training to constrain feature weights in order to identify the most important and informative features for the best prediction performance. Thus, finding the best feature subset and constructing the model are combined into one inseparable learning process. Penalized regression represents the most common form of regularization models, where a penalty term is added to the model's loss function (*residual sum of squares*) to constrain the magnitude of the coefficients associated with the features [37]. Features with zero or very small coefficients can be eliminated from the model. This not only improves the stability and accuracy of the learning algorithm, but also prevents model overfitting. There are various types of penalty terms and tuning parameters which define different penalized regression approaches, such as Ridge regression, Lasso regression, and elastic net regression. Each approach has its own unique properties and advantages when it comes to feature selection, model building, and prediction performance.

In this section, we discuss seven major regularization methods used in feature selection. We describe each method together with its mathematical formulation. In addition, we discuss their strengths and limitations. Table 2 depicts a summary of the considered methods. A more detailed discussion of each method is provided in the following subsections.

In the following subsections, we provide a more detailed discussion of each method. Without loss of generality, we will only consider linear models in which the training of the target variable Y can be based on a linear combination of X .

Table 2. Summary of variance-based feature selection methods.

Method	Strengths	Limitations	Supervision	Use
Ridge [37]	Handles multicollinearity between features. More stable due to its continuous shrinkage.	Reduces model interpretability. Produces estimation bias, Sensitive to feature scaling.	Supervised	Classification, Regression
Lasso [38]	Produces more sparse models. Prevents overfitting.	Performs poorly when relationship is nonlinear or highly complex. Inconsistent in presence of feature multicollinearity.	Supervised	Classification, Regression
Adaptive Lasso [39]	Flexibility in penalty application. Improves feature selection stability and performance.	Computationally more intensive.	Supervised	Classification, Regression
Elastic Net [40]	Combines strengths of Ridge and Lasso. Highly robust for correlated features.	Computationally more intensive due to having more parameters to tune. Risk of overfitting.	Supervised	Classification, Regression
Group Lasso [41]	Balances between individual feature selection and group level selection.	Computationally more intensive due to dual regularization. Requires predetermined groups of features.	Supervised	Regression
Regularized Decision Trees [42]	Improves model generalization to unseen data. Captures nonlinear relationships between feature and target variable.	Highly sensitive to changes in training data. Risk of over-regularization.	Supervised Unsupervised	Classification, Regression
Dantzig selector [43]	Promotes sparsity. Performs well when the number of features exceeds number of instances.	Can be computationally unstable. Limited interpretability of estimated coefficients.	Supervised	Regression

4.1. Ridge Regression

Ridge regression [37,44,45] is a powerful regularization technique used for creating small models in the presence of a large number of features. Ridge regression carries out L_2 regularization by adding a penalty for the sum of squares of coefficients to the optimization target. That is, for $\mathbf{X} \in \mathbb{R}^{n \times d}$ and a given λ , the Ridge estimator provides values of the coefficients β that minimize the following:

$$\hat{\beta}(\text{Ridge}) = \arg \min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_2^2 \quad (12)$$

where

$$\|\beta\|_2^2 = \sum_{j=1}^d \beta_j^2$$

where $\mathbf{Y} = (y_1, \dots, y_n)^T$ is the target variable and λ is the regularization parameter controlling the strength of the penalty.

In Ridge regression, the penalty term is regularized such that if the coefficients β take large values, the optimization function is penalized. It allows for better model performance through a bias—variance trade-off, hence reducing multicollinearity in the features and providing a more stable selection process. Ridge shrinks coefficients of less important features towards zero without eliminating them. This makes it more beneficial when all features are believed to contribute to the target variable, albeit producing higher estimation bias [46]. Due to its sensitivity to feature scaling, standardizing or normalizing the features is often essential before applying Ridge regression.

4.2. Least Absolute Shrinkage and Selection Operator (Lasso)

Similar to Ridge regression, the Lasso method [37,38,45] employs a continuous shrinkage procedure in which it applies L_1 regularization by adding the norm of the coefficients vector to the optimization objective function. It reduces the values of some of the coefficients to zero and features with non-zero coefficients are used for the model construction, resulting in more sparse models. The norm of the coefficients is generally constrained to be smaller than a predetermined value (upper bound). The Lasso estimate is the solution to the following optimization problem [37]:

$$\hat{\beta}(\text{LASSO}) = \arg \min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|_1 \quad (13)$$

where

$$\|\beta\|_1 = \sum_{j=1}^d |\beta_j|$$

If λ is big enough, certain coefficients will reach a value of zero. The advantage of the Lasso method is that it can provide better prediction accuracy, because shrinking and removing the coefficients can reduce variance without a substantial increase in the estimation bias. The method may, however, not perform well in applications where the number of features d exceeds the number of samples n due to its inability to select all relevant features. It can also struggle when features are highly correlated, often selecting only one feature from a group of correlated features. Ref. [47] generalized the Lasso estimation and introduced Relaxed Lasso (rLasso), which includes an additional tuning parameter that balances the fully regularized Lasso and the unpenalized optimization problem. The additional relaxation parameter allows for less shrinkage on selected and important features, leading to enhanced model accuracy and higher feature selection consistency.

4.3. Adaptive Lasso

The Lasso method is straightforward to apply, but it can be inconsistent in handling feature selection, in addition to having an inherent bias even under finite parameter conditions, as noted by the authors in [48]. To address these limitations, the authors in [39] proposed the following adaptive Lasso penalization problem:

$$\hat{\beta}(\text{AdapLasso}) = \arg \min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \sum_{j=1}^d \omega_j |\beta_j| \quad (14)$$

where the adaptive weight vector $(\omega_1, \omega_2, \dots, \omega_d)^T$ is a known quantity used to guide the L_1 penalty of the model coefficients. The authors in [49] demonstrated that when the weights $(\omega_1, \omega_2, \dots, \omega_d)^T$ are data-dependent, the feature selection process become more stable and consistent. Furthermore, in problems where $d \gg n$, [50] demonstrated that adaptive Lasso improves the stability of the feature selection under partial orthogonality of features, where the features with zero coefficients are weakly correlated with the features with non-zero coefficients. The effectiveness of the method, however, heavily relies on the initial estimates of the adaptive weights. Poor choices in the initial estimates can lead to poor performance of the feature selection process, in addition to being more computationally intensive than standard Lasso.

4.4. Elastic Net

In a manner similar to a stretchable fishing net, elastic net selects all relevant features and performs continuous shrinkage [40,51]. Unlike Lasso, which only selects individual features while disregarding the relationships that may exist between features, elastic net regularization selects groups of correlated variables. It combines the advantages of both Lasso and Ridge by setting certain coefficients to exactly zero, promoting sparsity, while integrating the L_2 penalty to shrink in the direction of zero to stabilize the estimation. The optimization problem of the elastic net is given by [40]:

$$\hat{\beta}(\text{Elastic Net}) = \arg \min_{\beta} \|Y - X\beta\|_2^2 + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1 \quad (15)$$

where $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are two regularization parameters. Elastic net allows for the adjustment of the mixing parameter defined by the ratio of L_1 to L_2 penalties, providing flexibility to adapt to different datasets and modeling scenarios. It eliminates the restriction on the number of selected features (Lasso) and stabilizes the selection from highly correlated features by adding a quadratic component to the penalty (Ridge). A strategy to efficiently apply the elastic net algorithm is shown in [40]. Ref. [51] conducted exhaustive numerical experiments on real datasets and concluded that, in terms of prediction accuracy, the elastic net always outperforms Lasso. Following the work of [47], the authors in [52] proposed Relaxed adaptive Lasso for more accurate identification of relevant features and higher prediction accuracy, particularly in datasets with more features than instances.

4.5. Group Lasso

In machine learning, integrating problem-specific assumptions into the learning process can lead to greater precision accuracy. For applications with naturally grouped features, such as different levels of a categorical variable or variables within the same domain, using group Lasso is more appropriate [41]. Group Lasso extends Lasso optimization and provides a structured approach to the selection and elimination of predefined groups of features. Suppose that features are divided into G disjoint groups $(g_1, g_2, \dots, g_G)$ where each group g_i constitutes a number of related features. Ref. [41] proposed the following group Lasso L_2 minimization problem:

$$\hat{\beta}(\text{Group Lasso}) = \arg \min_{\beta} \left\| Y - \sum_{i=1}^G X^{(g_i)} \beta^{(g_i)} \right\|_2^2 + \lambda \sum_{i=1}^G \left\| \mathbf{w}_i \beta^{(g_i)} \right\|_2 \quad (16)$$

where $X^{(g_i)}$ is the submatrix of X with columns corresponding to the features in group g_i , $\beta^{(g_i)}$ is the coefficient vector of that group, and $\mathbf{w}_i$ is some penalty weight vector computed based on the size of group g_i . Group Lasso can handle multicollinearity between features such that if a group consists of correlated features, the procedure either selects

or eliminates them together. The authors in [53] extended the group Lasso to allow also for the selection of individual features within the selected groups. The resulting extension is called sparse group Lasso and is mathematically formulated by the following convex optimization problem:

$$\hat{\beta}(\text{Sparse Lasso}) = \arg \min_{\beta} \left\| \mathbf{Y} - \sum_{i=1}^G X^{(g_i)} \beta^{(g_i)} \right\|_2^2 + \lambda_2 \sum_{i=1}^G \left\| \mathbf{w}_i \beta^{(g_i)} \right\|_2 + \lambda_1 \|\beta\|_1 \quad (17)$$

where λ_1 controls the strength of the L_1 Lasso penalty (individual feature selection) and λ_2 controls the strength of the group-Lasso L_2 penalty (group feature selection). Sparse Group Lasso is a powerful approach, balancing between individual and group feature selections. Despite the computational challenges of the techniques, it is still popular in application domains such as genomics, finance, neuroscience, and other fields where features have natural group structures [54].

4.6. Regularized Decision Trees

Regularized decision trees represent a generalization of the traditional decision tree algorithms, aiming to improve their performance in feature selection and mitigate overfitting. They provide an effective way to select important features by incorporating penalties on the complexity of the tree, resulting in the selection of the most relevant features while ignoring noise and irrelevant features. Several penalizations of the decision tree have been developed including tree depth, node complexity, and feature importance scores [42]. In this subsection, we focus on the regression tree where the prediction at each leaf node is typically the average of the target variable in a given region R_m and the splitting criterion is based on the maximum variance reduction between the regions defined by the split, thereby minimizing the prediction error defined by the following loss impurity function:

$$\mathcal{L}(\mathcal{T}) = \sum_{m=1}^M \sum_{i \in R_m} (y_i - \bar{y}_m)^2 \quad (18)$$

where $\bar{y}$ is the average of the target in region R_m , $m = 1, 2, \dots, M$. The mathematical formulation of the tree regularization is given by the following optimization problem:

$$\min_{\mathcal{T}} f = \mathcal{L}(\mathcal{T}) + \lambda \Omega(\mathcal{T}) \quad (19)$$

where $\Omega(\mathcal{T})$ is a tree complexity penalty function and λ is a regularization parameter balancing between minimum prediction error and tree complexity [55].

Common tree regularization techniques include the following:

- Tree Depth limits the tree depth, controlling for excessive feature interactions. The penalty function is $\Omega(\mathcal{T}) = |\mathcal{T}|$, where $|\mathcal{T}|$ denotes the number of leaf nodes. A deeper tree has more splits, and each additional split increases the likelihood of overfitting. High values of λ force more shallow depth, selection of the most informative features, and less model complexity [56].
- Feature importance penalizes features with low importance. The complexity penalty function is defined by

$$\Omega(\mathcal{T}) = \sum_{j=1}^d w_j \sum_{t \in \mathcal{T}_j} \Delta I(t)$$

where $\Delta I(t)$ is the reduction in impurity, which is typically the variance of the target variable, at node t . Larger $\Delta I(t)$ implies more informative feature. With this approach, features with insignificant contribution are eliminated [42].

- L1 (Lasso-like) regularization on leaf predictions that shrinks some leaf values to zero, leading to the elimination of weak features. This technique is inspired by Lasso penalization, as $\Omega(\mathcal{T}) = \sum_t |w_t|$, w_t is the weight of leaf t [57].
- L2 (Ridge-like) regularization controls for feature dependencies and uses $\Omega(\mathcal{T}) = \sum_t w_t^2$ [57].
- Minimum samples per split regularization force at least the minimum number of samples at each leaf node in order to prevent the tree from creating leaves that represent only a small number of data points, which could lead to overfitting and the selection of unimportant features. It uses the penalty function $\Omega(\mathcal{T}) = \mathbf{I}(n_{leaf} < \tau)$, where τ is the minimum number of samples per split [58].
- Cost complexity pruning ensures the removal of parts of the tree that do not provide power in predicting the target variable. The technique introduces a penalty based on the number of terminal nodes in a tree, i.e., $|\mathcal{T}|$ [59].

Combining the above regularization techniques, the overall objective for a regularized regression tree can be constructed with multiple penalization terms, controlling the complexity of the tree in various ways.

Decision trees are inherently easy to interpret, and with regularization, their simplicity is enhanced and their complexity is reduced. Regularization also enhances the ability of the decision trees to capture nonlinear relationships between features and target variables and their robustness to outliers. There exists a risk of over-regularization with regression decision trees, leading to the loss of important features [56].

4.7. Dantzig Selector

Introduced by [43], Dantzig selector is designed for high-dimensional regression problems, particularly those with a lower number of features than instances. It is closely related to LASSO but employs a different approach for feature selection. Features selected by Dantzig selector have coefficients $\hat{\beta}$ that satisfy the following l_1 regularization problem:

$$\hat{\beta} = \arg \min_{\beta} \begin{cases} \min_{\beta} \|\beta\|_1 \\ \text{s.t. } \|X^T(Y - X\beta)\|_{\infty} \leq (1 + s^{-1}) \sqrt{2 \log(d)} \sigma \end{cases} \quad (20)$$

where s is a positive scalar and σ is the standard deviation of the prediction errors. Dantzig selector encourages sparsity in the estimated coefficients and can handle multicollinear features. It is designed to handle noise in the data effectively and can produce a consistent feature selection process. However, its limitations, such as computational complexity and assumptions about linear relationships between features and target variables, necessitate careful consideration when applying it.

4.8. Discussion

Regularization strategies have rapidly become an essential part of machine learning in order to improve the performance of the learning algorithm by selecting a subset of relevant features while preventing model overfitting. They are effective in extremely high-dimensional problems, particularly those where the number of features significantly exceeds the number of instances. The strategies discussed in this section are focused on the supervised learning of linear models. Regularization has been broadly investigated in other learning models, such as L_1 -logistic regression [60–62], L_1 -SVM [63], and Hybrid Huberized SVM (HHSVM), where the characteristics of L_1 and L_2 norms are utilized simultaneously with SVM [64] in addition to several L_1 -regularized linear discriminant analysis (LDA) methods [65,66] that been developed for classification problems in high-dimensional datasets.

Although regularized approaches have several advantages, there are also challenges associated with these techniques. They often require strong model assumptions that may not hold in real-world situations [55]. Furthermore, regularized feature selection methods suffer from several computational challenges that can affect their efficiency and effectiveness. In addition to the significant computational overhead required by many feature selection techniques in high dimensionality, regularized methods also suffer from convergence issues and local minima, particularly for non-convex optimization problems [67]. Researchers have discussed other computational challenges of regularization methods, including scalability challenges due to the iterative optimization and memory-intensive requirements for large datasets [68]; model underfitting due to over-regularization and inappropriate selection of regularization parameter values λ [40]. Tuning the value of λ to find the optimal level of regularization can actually be computationally expensive. Cross-validation is often used to determine the best value of λ , further increasing the computational load, particularly for large datasets or when performing the grid search over a wide range of values.

Numerous approaches have been developed in the literature to mitigate or reduce the challenges associated with the regularized strategies for feature selection. Prominent approaches include feature screening to eliminate irrelevant features before model training [69–71] and hybrid feature selection that provides a robust framework for effectively combining multiple feature selection techniques to leverage the strengths of each technique in finding the optimal and most informative feature subset efficiently and quickly [6,72,73].

5. Bayesian-Based Feature Selection

Bayes' theorem provides a principled framework for updating probabilities based on observed evidence. It forms the foundation of probabilistic reasoning and has been extensively employed in the development of feature selection algorithms, and is expressed as follows:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \quad (21)$$

where $P(A|B)$ represents the posterior probability of event A given evidence B , $P(B|A)$ is the likelihood, $P(A)$ is the prior probability of A , and $P(B)$ is the evidence or marginal likelihood. A diverse range of Bayes-based methods has been proposed, spanning from basic Bayesian approaches like Lasso to advanced Bayesian variable selection techniques as relative belief ratio. These methods look for the probabilistic dependencies between features and target variables to identify the most informative and discriminative features, enabling more efficient and effective representation of the data for machine learning applications.

Table 3 presents a summary of the Bayesian-based feature selection methods. A comprehensive discussion of each method, including its mathematical formulation, strengths, and limitations, is provided in the following subsections. This section covers seven key Bayesian feature selection techniques, offering insights into their functionality and practical applications.

Table 3. Summary of Bayesian-based feature selection methods.

Method	Strengths	Limitations	Supervision	Use
Bayesian LASSO [74]	Incorporates prior knowledge and provides uncertainty quantification. Handles collinearity effectively.	Computationally expensive due to MCMC sampling. Sensitive to prior choice.	Supervised	Regression
Balamurugan and Rajaram Method [75]	Identifies and removes redundant features efficiently. Improves classification accuracy. Provides probabilistic predictions with automatic sparsity. Handles nonlinear relationships well.	Computational overhead for large feature sets. Sensitive to threshold choice.	Supervised	Classification
Relevance Vector Machine [76]	Enhances interpretability and sparsity. Automatic relevance determination of features and samples. Efficient approximation of posterior distributions. Suitable for small datasets with high-dimensional features.	Computationally intensive. Requires careful kernel selection.	Supervised	Regression, Classification
Relevant Sample-Feature Machine [77]		Computational complexity with large datasets. Performance depends on kernel choice.	Supervised	Regression, Classification
Variational Embedded RSFM [78]		Computationally demanding optimization. Performance dependent on variational assumptions.	Supervised	Regression, Classification
Bayesian Robit Regression with Hyper-LASSO Priors [79]	Handles grouping structures effectively. Provides robust sparsity control for high-dimensional data. Ranks features effectively by updating prior beliefs. Provides a principled Bayesian measure for feature importance.	Requires MCMC sampling, increasing computational burden. Sensitive to hyperparameter tuning.	Supervised	Classification
Relative Belief Ratio Feature Selection [80]		Computationally intensive. Requires careful prior elicitation.	Supervised	Classification

5.1. Bayesian Lasso

The Bayesian Lasso is based on the original Lasso method presented in Section 4.2. It was introduced by Park and Casella [74] to extend the classical Lasso by incorporating a fully Bayesian framework for feature selection and regularization in high-dimensional regression problems. In this method, a Laplace prior is placed on the regression coefficients

to encourage sparsity while incorporating uncertainty in the parameters through posterior inference. The Laplace prior for the j -th coefficient is formulated as

$$p(\beta_j|\lambda) \propto \exp(-\lambda|\beta_j|), \quad (22)$$

where λ denotes the regularization parameter that controls the level of sparsity. Unlike classical Lasso, in Bayesian Lasso, λ is treated as a random variable and is typically assigned a hyperprior, such as a Gamma distribution, to facilitate data-driven estimation of the regularization parameter. The posterior distribution of the coefficients is obtained by combining the Laplace prior with the likelihood function of the data. Given a response vector $\mathbf{y}$ and a feature matrix $\mathbf{X}$, the posterior distribution is expressed as

$$p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \lambda) \propto p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) \prod_{j=1}^p p(\beta_j|\lambda). \quad (23)$$

Inference in the Bayesian Lasso is typically conducted using Markov Chain Monte Carlo (MCMC) techniques, such as Gibbs sampling, which facilitate the estimation of the full posterior distribution of the coefficients. This allows for both feature selection and the quantification of uncertainty associated with the selected features.

The Bayesian LASSO offers several advantages, including its ability to incorporate prior knowledge through the Bayesian framework, yielding improved sparsity and uncertainty quantification compared to classical LASSO. However, the computational complexity of MCMC sampling can be a limitation, especially for large-scale datasets. Additionally, the choice of the hyperprior for λ may influence the selection of features, necessitating careful prior specification to achieve optimal performance.

5.2. Balamurugan and Rajaram Method

Balamurugan and Rajaram [75] introduced a Bayes' theorem-based feature selection method to address challenges associated with high-dimensional datasets. In this method, feature reduction is achieved by leveraging Bayes' theorem to compute conditional probabilities, facilitating the identification and elimination of redundant features. The primary objective of the method is to enhance classification accuracy and computational efficiency by retaining only the most informative features.

The methodology involves defining an initial set of features, represented as $A = \{a_1, a_2, \dots, a_n\}$. For every possible pair of features (a_i, a_j) , the conditional probabilities $P(a_i, a_j|c)$ are computed concerning the class attribute c . Dependencies between features are identified by evaluating these conditional probabilities against a predefined threshold. Features exhibiting strong dependencies are subsequently removed to minimize redundancy while preserving the most relevant features for classification tasks.

This method effectively applies Bayesian principles to feature selection by modeling dependencies among features and the target variable. The integration of Bayes' theorem ensures a systematic approach to dimensionality reduction, leading to improved classification performance in high-dimensional datasets. However, the method's reliance on pairwise feature evaluation may introduce computational overhead, particularly in cases involving a large number of features. Despite this limitation, the approach remains a valuable tool for feature selection, offering improved interpretability and data efficiency.

5.3. Relevance Vector Machine

The Relevance Vector Machine (RVM), introduced by Tipping [76], has been developed as a sparse Bayesian learning framework for regression and classification tasks. This method extends the support vector machine (SVM) by incorporating Bayesian principles

to achieve sparsity while providing probabilistic outputs. The RVM identifies a subset of training samples, known as relevance vectors, that contribute significantly to the predictive model, while the weights associated with irrelevant features are driven to zero through the Bayesian inference process. This characteristic allows for the automatic selection of informative features, making the RVM a robust and efficient tool for feature selection, particularly in high-dimensional datasets.

The prediction function of the RVM is expressed as

$$y(\mathbf{x}) = \sum_{i=1}^N w_i k(\mathbf{x}, \mathbf{x}_i) + w_0, \quad (24)$$

where $y(\mathbf{x})$ represents the predicted output for input $\mathbf{x}$, $k(\mathbf{x}, \mathbf{x}_i)$ is the kernel function that measures the similarity between input $\mathbf{x}$ and training sample $\mathbf{x}_i$, w_i denotes the weight associated with the i -th basis function, and w_0 is the bias term. The kernel function $k(\cdot, \cdot)$ allows the RVM to capture nonlinear relationships by mapping the inputs into a higher-dimensional feature space.

A key feature of the RVM is its hierarchical Bayesian framework, where a zero-mean Gaussian prior is placed on each weight w_i , with its variance controlled by a hyperparameter α_i . This prior is defined as

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=1}^N \mathcal{N}(w_i|0, \alpha_i^{-1}), \quad (25)$$

where $\mathcal{N}(\cdot)$ represents a Gaussian distribution, and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_N)$ denotes the hyperparameters governing the sparsity. During training, many of the α_i values tend to infinity, which forces the corresponding weights w_i to zero, effectively pruning irrelevant features.

The Bayesian nature of the RVM offers several advantages, including robustness against overfitting and improved interpretability by highlighting the most informative features. Despite its advantages, the RVM's reliance on iterative optimization and the need to estimate hyperparameters can introduce computational complexity, especially when applied to large datasets. Nonetheless, the RVM has been successfully employed in various feature selection tasks, demonstrating its effectiveness in improving model performance and enhancing computational efficiency.

5.4. Relevant Sample-Feature Machine

Relevant Sample-Feature Machine (RSFM) was introduced by Mohsenzadeh et al. [77] in 2013 as a feature selection method based on sparse Bayesian machine learning. It extends the Relevance Vector Machine (RVM) by incorporating a sparse Bayesian framework and Gaussian priors to achieve feature selection. RSFM operates as a sparse kernel-based learning model, where the feature selection process is guided by Bayesian inference to enhance sparsity and interpretability.

The output of the RSFM is predicted using the kernel function $f(\mathbf{x}|\mathbf{w})$, which is formulated as

$$f(\mathbf{x}|\mathbf{w}) = w_0 + \sum_{m=1}^M w_m k(x, x_n), \quad (26)$$

where $k(x, x_n)$ represents the kernel function measuring the similarity between input x and sample x_n , and $\mathbf{w} = (w_0, w_1, \dots, w_m)^T$ denotes the weight vector associated with the basis functions. By exploiting the sparsity-inducing properties of Bayesian priors, irrelevant features are assigned weights close to zero, effectively pruning redundant information.

The RSFM method provides several advantages, including enhanced sparsity, improved model interpretability, and automatic relevance determination of both samples and

features. The Bayesian approach allows the model to handle high-dimensional datasets efficiently while mitigating the risk of overfitting. However, the computational complexity associated with inference, particularly in large-scale applications, can be a limitation. The choice of kernel function and prior distributions also influences the model's performance, necessitating careful tuning for optimal feature selection outcomes. Overall, the RSFM method represents a significant advancement in sparse Bayesian learning, offering an effective solution for feature selection in various domains, including biomedical data analysis and text classification.

5.5. Variational Embedded RSFM

The Variational Embedded RSFM (VRSFM) method was introduced by Mirzaei, Mohsenzadeh, and Sheikhzadeh [78] in 2017 as an enhancement of the RSFM framework. It employs a Bayesian model of RSFM with a focus on variational Bayesian approximation to improve feature selection for both classification and regression tasks. Prior Gaussian distributions are placed on the model parameters and their hyperparameters to enforce sparsity and facilitate robust feature selection.

In the VRSFM approach, the posterior distributions of the model parameters are approximated using variational Bayesian inference. Given an observation set $\mathbf{X} = \{x_1, x_2, \dots, x_n\}$ and corresponding responses $\mathbf{y}$, the posterior distribution of the weight vector $\mathbf{w}$ is approximated by minimizing the Kullback–Leibler (KL) divergence between the true posterior $p(\mathbf{w}|\mathbf{X}, \mathbf{y})$ and the variational distribution $q(\mathbf{w})$, expressed as follows:

$$\text{KL}(q(\mathbf{w}) \parallel p(\mathbf{w}|\mathbf{X}, \mathbf{y})) = \int q(\mathbf{w}) \log \frac{q(\mathbf{w})}{p(\mathbf{w}|\mathbf{X}, \mathbf{y})} d\mathbf{w}. \quad (27)$$

The variational inference approach provides an efficient approximation to the intractable posterior distributions, enabling the method to operate effectively even in scenarios with small-sized datasets. The optimization of the variational distribution parameters is conducted iteratively to minimize the KL divergence, ensuring that the approximated posterior closely resembles the true posterior.

By leveraging Bayesian principles, the VRSFM method balances model complexity and predictive performance while improving interpretability and robustness. This technique extends the applicability of RSFM, making it particularly suitable for feature selection in high-dimensional datasets with limited sample sizes. Despite its advantages, the performance of VRSFM is influenced by the choice of variational distributions and the complexity of the optimization process. The computational cost associated with iterative updates may be significant in large-scale applications. However, the method's ability to provide probabilistic feature selection and uncertainty quantification makes it a valuable tool in various domains, including biomedical data analysis and financial modeling.

5.6. Bayesian Robit Regression with Hyper-Lasso Priors

Bayesian Robit Regression with Hyper-LASSO priors (BayesHL) has been developed as a feature selection method tailored for high-dimensional datasets, particularly those exhibiting grouping structures. This approach employs a heavy-tailed Robit model combined with Hyper-LASSO priors to achieve robust feature selection, ensuring sparsity and the capability to uncover grouping structures without necessitating a pre-specified grouping index. BayesHL has been effectively applied in gene expression analysis to identify subsets of genes associated with disease outcomes, such as the 5-year survival of endometrial cancer patients [79].

In the Robit regression model, binary outcomes y_i are modeled by replacing the normal distribution in probit regression with a scaled Student's t-distribution. The model is formulated as follows:

$$y_i = I(z_i > 0), \quad z_i = \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \sim T(\alpha, \omega), \quad (28)$$

where $\mathbf{x}_i$ denotes the vector of features for the i -th observation, $\boldsymbol{\beta}$ represents the vector of regression coefficients, ϵ_i follows a scaled Student's t-distribution $T(\alpha, \omega)$ with degrees of freedom α , and $I(\cdot)$ is the indicator function.

To induce sparsity, a Cauchy prior is assigned to each regression coefficient β_j , expressed as

$$\beta_j \sim T(1, \omega_1), \quad (29)$$

where ω_1 is a small scale parameter ensuring sparse solutions. The heavy-tailed nature of the Cauchy prior allows for a few large coefficients (signals) while shrinking irrelevant ones (noise) toward zero.

The posterior distribution of the parameters is given by

$$p(\boldsymbol{\beta}, \boldsymbol{\lambda} | \mathbf{y}, \mathbf{X}) \propto p(\mathbf{y} | \mathbf{X}, \boldsymbol{\beta}) \prod_{j=1}^p p(\beta_j | \lambda_j) p(\lambda_j), \quad (30)$$

where λ_j represents the latent variance for each coefficient, modeled as

$$\lambda_j \sim \text{Inverse-Gamma} \left(\frac{\alpha}{2}, \frac{\alpha \omega_1^2}{2} \right). \quad (31)$$

The utilization of heavy-tailed priors in BayesHL enables more aggressive shrinkage of irrelevant features compared to traditional Lasso, while retaining significant predictors. This characteristic is particularly advantageous in high-dimensional settings with potential grouping structures among features. However, the computational complexity associated with fully Bayesian methods, including the need for Markov Chain Monte Carlo (MCMC) sampling, can be a limitation, especially with large-scale datasets. Despite this, BayesHL provides a flexible and powerful framework for feature selection, offering probabilistic interpretations and accommodating complex data structures.

5.7. Relative Belief Ratio

The Relative Belief Ratio (RBR) feature selection method [80] offers a Bayesian approach to feature selection in binary classification problems. This method evaluates the significance of each feature by quantifying the change in belief from prior to posterior distributions, thereby identifying and ranking features based on the relative belief strength and hence importance.

The RBR proposed by [81] is defined as the ratio of the posterior density to the prior density at a specific parameter value. For a parameter θ with prior distribution $\pi(\theta)$ and posterior distribution $\pi(\theta | x)$, given data x , the relative belief ratio is expressed as

$$RB(\theta | x) = \frac{\pi(\theta | x)}{\pi(\theta)}. \quad (32)$$

A value of $RB(\theta | x) > 1$ indicates evidence in favor of θ , while $RB(\theta | x) < 1$ suggests evidence against θ . This measure allows for a systematic comparison of features by assessing how the observed data update the prior beliefs about each feature's relevance.

To quantify the strength of the evidence provided by the RBR, a strength function $S(\theta | x)$ is defined as

$$S(\theta | x) = P(RB(\theta | x) \leq RB(\theta_0 | x) | x), \quad (33)$$

where θ_0 is a baseline parameter value. The strength function $S(\theta | x)$ represents the posterior probability that the relative belief ratio for θ is less than or equal to that for θ_0 . A lower strength value indicates stronger evidence in favor of θ , providing a calibrated measure of the evidence's robustness. In practical applications, the RBR method involves estimating the relative belief ratios for each feature and computing a strength score to rank the features. The strength score ℓ_j for feature V_j is calculated by integrating the relative belief ratios over the parameter space, providing a quantitative measure of each feature's importance. Features with lower strength scores are considered more significant.

The RBR method has been applied to both synthetic and real-world datasets, demonstrating its effectiveness in identifying significant features and achieving high classification accuracy in high-dimensional gene datasets, outperforming in general traditional feature selection techniques such as Information Gain and Symmetrical Uncertainty. However, the implementation of the RBR method requires careful elicitation of prior distributions, and can be computationally intensive due to the necessity of estimating posterior distributions. Despite these challenges, the RBR method provides a robust framework for feature selection, particularly suited for high-dimensional datasets where traditional methods may falter.

5.8. Discussion

The Bayesian-based methods discussed above illustrate a range of strategies for leveraging probabilistic reasoning in feature selection. These methods vary in complexity and application, from sparsity-inducing approaches like Bayesian Lasso to more advanced frameworks such as the RBR method. Bayesian Lasso builds on classical Lasso by incorporating prior distributions and providing uncertainty quantification, making it particularly useful in regression tasks where collinearity among features is present. However, its reliance on MCMC sampling increases computational demands, requiring careful tuning of hyperparameters.

More sophisticated methods, including RVM and its extensions, like RSFM and VRSFM, combine sparse Bayesian learning with kernel-based modeling to identify relevant features. These methods excel in high-dimensional settings with nonlinear relationships by enforcing sparsity through hierarchical priors. While these approaches offer flexibility and improved model interpretability, their computational overhead, particularly in large datasets, necessitates efficient optimization techniques and kernel selection strategies.

Bayesian methods tailored for complex data structures, such as Bayesian Robit Regression with BayesHL and the Balamurugan and Rajaram method, offer solutions for scenarios involving dependencies or grouping structures among features. BayesHL effectively handles high-dimensional datasets with group structures by employing heavy-tailed priors, whereas the Balamurugan and Rajaram method focuses on reducing redundancy through pairwise dependency evaluation. Both methods enhance interpretability and classification performance, but may face scalability issues in large datasets.

The RBR method offers a unique Bayesian measure for feature importance by updating prior beliefs based on observed evidence. This method is particularly suited for high-dimensional datasets with limited sample sizes, as it ranks features using a strength score derived from the relative belief ratio. Although the RBR method provides a principled framework for feature selection, it requires careful prior elicitation and is computationally intensive due to the need for posterior estimation.

In practical applications, the combination of multiple Bayesian methods can be advantageous. For example, Bayesian Lasso can be used as an initial filtering step to eliminate irrelevant features, followed by kernel-based methods such as RVM or RSFM for refining the feature set. Advanced techniques like BayesHL or RBR can further prioritize domain-specific features or handle grouping structures. The choice of methods depends on the specific task, data characteristics, and computational resources. Collectively, these Bayesian-based approaches emphasize the importance of probabilistic modeling in feature selection, particularly in addressing uncertainty, dependencies, and high-dimensional feature spaces.

6. Related Methods

In addition to the three major themes discussed above, there exist other approaches related to mathematical feature selection. While not purely mathematical, these approaches employ ideas from mathematics to build selection algorithms.

6.1. Hybrid Feature Selection

Hybrid feature selection methods have recently gained popularity due to increasing data dimensionality in fields like image processing, bioinformatics, and natural language processing. These methods combine multiple techniques, leveraging their strengths to handle complex datasets, improve algorithm performance, and reduce computational costs. The most common hybrid approach integrates filter and wrapper methods, where a computationally efficient filter initially removes irrelevant features, followed by a wrapper method that iteratively selects the optimal feature subset based on model performance. Additionally, ensemble schemes combining multiple feature selection methods in parallel can further enhance accuracy and stability [6].

This subsection provides a description of some of the most recently published hybrid feature selection frameworks. The authors in [82] introduced a two-step hybrid feature selection method: the first step utilizes Fuzzy Joint Mutual Information to effectively decrease the initial search space and exclude irrelevant features, while the second wrapper-based step uses the Binary Cheetah Optimizer Algorithm for further refining the feature subset through the K-Nearest Neighbors (KNN) classifier. Validated on 23 medical datasets and 14 high-dimensional microarray datasets, the proposed hybrid feature selection process achieved higher accuracy in 78.26% of datasets and reduced the feature size by 84.79%, making it suitable for real-world, high-dimensional datasets and decision-making in medical data analysis. In [83], a filter-wrapper feature selection framework is designed for cancer gene selection and classification. The initial filter stage uses a t-test or an F-test to remove noisy genes. The selected genes are then refined through a multi-objective optimization based on Pareto optimality criteria, framed within the support vector machine classifier. The robustness and effectiveness of the proposed framework in producing high prediction accuracy with a minimal gene subset was evaluated using both simulated and published high-dimensional microarray datasets. The authors of [84] proposed a hybrid feature selection framework for high-dimensional datasets with low sample sizes (HDLSSs) in which gradual permutation filtering is used first to rank features based on their permutation importance score and to eliminate irrelevant features. After the filtering step, a heuristic hybrid search combining forward search, consolation match, and backward elimination is used to remove redundancy and identify the optimal feature set. Experimental evaluations showed that the proposed method effectively selects a small number of important features while achieving high prediction performance. [85] developed a hybrid method that merges a metaheuristic optimization algorithm called quantum particle swarm optimization (QPSO) with principal component analysis (PCA). QPSO optimizes feature selection by simulating quantum particle behaviors, while PCA reduces dimensionality and

addresses multicollinearity. Applied to software defect prediction, this approach improved the accuracy of the artificial neural network model and provided a more interpretable feature subset. In the context of regression problems in high-dimensional datasets, a filter-embedded hybrid feature selection methodology was proposed in [73]. In the filtering phase, the nonparametric Kendall's tau correlation coefficient is used to reduce the feature space by removing features with negative correlations. In the embedded method phase, the elastic net regularized regression model is trained based on the pre-selected features in the preceding phase, resulting in a fewer non-zero coefficients and leveraging the grouping effect, which naturally groups highly correlated variables.

The authors of [86] designed a nested ensemble filter-wrapper hybrid feature selection method consisting of two stages. In the filtering stage tree-based ensemble consisting of random forest and extra tree classifiers are used to evaluate individual feature importances and the top 20 features are fed to the second stage in which random forest-based sequential backward selection is applied to eliminate the remaining redundant and correlated features. Extensive empirical testing showed that the proposed methodology significantly outperforms the benchmarks algorithms especially on multi-class datasets. A multi-filter hybrid method was developed in [87] for cancer classification. Three filter methods are employed to rank the features. The top candidates from each filter are merged, and a genetic algorithm-based wrapper is then applied to select an optimal feature subset. This multi-filter approach not only reduces redundancy, but also improves classification accuracy compared to using a single filter.

6.2. Graph Theory-Based Feature Selection

Graph theory-based feature selection methods have gained prominence due to their ability to model complex feature relationships effectively. By representing features as nodes and defining edges based on similarity measures, these methods facilitate a structured analysis that enhances the selection process [88,89].

A fundamental approach in graph-based feature selection involves constructing an undirected graph, where each feature corresponds to a node, and the edges represent dependencies or interactions between them. In cancer classification, for instance, Zhou et al. [88] proposed a method utilizing the maximum mutual information coefficient to establish relationships among genes. The approach identifies core features within the graph, ensuring the selection of an informative subset that minimizes redundancy while preserving classification accuracy. The methodology has demonstrated improvements in predictive performance compared to traditional selection techniques.

Graph-theoretic approaches have also been applied in supervised learning settings. Rostami et al. [89] introduced a multi-objective optimization framework for selecting relevant genes in microarray data. The method employs centrality measures and community detection to identify genes that exhibit strong dependencies with class labels while maintaining diversity among selected features. The optimization framework effectively balances feature relevance and redundancy, leading to more robust and interpretable selections.

In unsupervised scenarios, where class labels are unavailable, feature selection relies on preserving the intrinsic structure of the data. A common strategy involves constructing a similarity graph among features and selecting those that best maintain the structural integrity of the graph. Tang et al. [90] proposed an advanced technique that leverages multiple graph fusion, integrating different similarity graphs to enhance robustness. By learning feature weights dynamically, this approach selects features that best capture the overall data distribution, thereby improving clustering and classification performance.

Recent research has further expanded on these approaches by integrating hybrid strategies and semi-supervised learning techniques. Banerjee et al. [91] introduced a hybrid

feature selection method that utilizes graph centrality measures to improve classification outcomes. Similarly, Li et al. [92] provided a comprehensive survey exploring feature selection techniques with limited labeled data, highlighting the role of semi-supervised approaches in enhancing graph-based feature selection. Fu et al. [93] proposed adaptive graph convolutional networks that dynamically refine feature selection in a semi-supervised setting, further demonstrating the adaptability of graph-based methodologies across different learning paradigms.

6.3. Correlation-Based Feature Selection

Recent studies have increasingly emphasized correlation information as a key factor in effective feature selection. The focus arises from the multifaceted relationships that can exist among features and labels, particularly in high-dimensional or multi-label contexts. By exploiting both global and local dependencies, these methods aim to balance between capturing robust predictive signals and mitigating redundancy.

Faraji et al. [94] propose a multi-label feature selection model that accounts for both global and local label correlations. Their approach represents features and labels in a shared latent space, thereby capturing hidden relationships across multiple labels. Employing an $L_{2,1}$ -norm regularization in their objective function, they present an alternating optimization algorithm that yields sparse coefficients relevant for multi-label classification.

Addressing the complexities of linear and nonlinear dependencies, Gong et al. [95] introduce a filter-based feature selection algorithm that combines the Pearson correlation coefficient with normalized mutual information. Their proposed evaluation function, CONMI, assesses both the relevance of features to class variables and the redundancy among selected features.

Fan et al. [96] highlight the importance of capturing both label correlations and feature redundancy in multi-label scenarios. Their proposed scheme—Learning Correlation Information for Multi-Label Feature Selection (LCIFS)—jointly models label correlations via an adaptive spectral graph and reduces redundant features by enforcing an $\ell_{2,p}$ -norm regularization. Experiments on twelve realistic multi-label datasets, including text, image, and audio domains, illustrate that LCIFS achieves robust results compared to nine competitive methods.

7. Challenges and Future Directions

One key challenge in mathematical feature selection involves handling higher-order interactions among predictors while preserving interpretability. Traditional methods often prioritize marginal or pairwise relationships, yet many real-world domains such as genomics or climate science demand a deeper understanding of complex dependencies. Designing algorithms that capture intricate structural associations among features, while retaining transparency for end-users and domain experts, thus presents a compelling avenue for future research.

A second area of interest encompasses non-convex and robust selection approaches that address issues such as outliers, heavy-tailed distributions, or adversarial data contamination. Although L_1 -based regularization methods often dominate regularization-based feature selection, alternative penalty functions (e.g., SCAD, MCP) and robust cost formulations can potentially enhance performance, especially under noisy conditions. The effective integration of these techniques with existing feature selection pipelines would address data integrity concerns without compromising scalability or computational feasibility.

Emerging data paradigms also highlight the need for online and streaming feature selection procedures. In such scenarios, data arrive continuously rather than in fixed batches, making static, offline methods insufficient or computationally expensive. New al-

gorithms must update feature importance adaptively, possibly by maintaining approximate variance–covariance structures in real time or by employing sequential Bayesian inference. Such advances would open opportunities for real-time analytics, a rapidly growing domain in the face of expanding datasets and evolving streaming technologies.

Another notable challenge in evaluating feature selection algorithms is the scarcity of datasets with pre-established relevant features. In this regard, synthetic datasets can play a pivotal role in advancing feature selection methodologies, as they enable researchers to rigorously assess algorithmic performance under controlled conditions and validate new approaches with known ground truths [97,98].

The integration of mathematical feature selection principles in deep learning architectures constitutes another promising direction. Sparse regularization layers (e.g., group-Lasso) or hierarchical Bayesian priors can be embedded directly into neural networks to enhance interpretability and reduce computational overhead. Moreover, many application areas would benefit from domain-informed Bayesian priors that anchor feature selection in specialized scientific knowledge. Whether derived from expert input, experimental constraints, or biological pathways, carefully formulated priors can produce both robust and contextually meaningful results. Finally, ensuring that these methods scale gracefully remains a persistent challenge, given the ever-increasing size and complexity of modern datasets. Parallelization strategies, more efficient sampling procedures, and approximate inference techniques that preserve theoretical guarantees may help accommodate real-world computational constraints, allowing the theoretical benefits of mathematical feature selection to materialize in practical, large-scale environments.

8. Conclusions

Mathematical methods form the backbone of many state-of-the-art feature selection techniques, offering theoretical rigor and improved interpretability in both supervised and unsupervised contexts. By systematically examining variance-based, regularization-based, and Bayesian approaches, this review highlights the diverse mathematical paradigms that guide the identification and retention of informative features.

Variance-based methods, which rely on measures of dispersion, generally excel at fast dimensionality reduction, but may be less apt at capturing task-specific relationships between predictors and target variables. Regularization-based algorithms circumvent the issue by embedding feature selection directly in the model training phase, ensuring that sparsity naturally emerges through penalized optimization. Bayesian methods, by contrast, leverage prior distributions and posterior inference to quantify and incorporate uncertainty in a flexible, robust manner.

Although these techniques are each built on distinct theoretical frameworks, their integration can yield synergistic benefits. For example, a preliminary variance-based filter can reduce dimensionality before more computationally intensive Bayesian or regularization-based methods refine the selection process. The multi-stage strategy exemplifies how theoretical assumptions and practical considerations jointly drive successful feature selection outcomes.

Ultimately, the choice of a feature selection method hinges on domain requirements, data characteristics, and computational resources. From capturing nonlinear interactions to employing robust priors informed by scientific expertise, the continued expansion of mathematical feature selection promises both heightened predictive performance and deeper insights into high-dimensional data.

Author Contributions: Methodology, F.K., H.S. and A.A.; Formal analysis, F.K., H.S. and A.A.; Investigation, F.K., H.S., A.A., M.E., H.C. and M.S.; Writing—original draft, F.K., H.S., A.A., M.E.,

H.C. and M.S.; Writing—review and editing, F.K.; Visualization, F.K.; Supervision, F.K., H.S. and A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Mitra, R.; Ali, E.; Varam, D.; Sulieman, H.; Kamalov, F. Variable Selection in Data Analysis: A Synthetic Data Toolkit. *Mathematics* **2024**, *12*, 570. [\[CrossRef\]](#)
2. Abu Khurma, R.; Aljarah, I.; Sharieh, A.; Abd Elaziz, M.; Damaševičius, R.; Krilavičius, T. A review of the modification strategies of the nature inspired algorithms for feature selection problem. *Mathematics* **2022**, *10*, 464. [\[CrossRef\]](#)
3. Olaniran, O.R.; Sikiru, A.O.; Allohobi, J.; Alharbi, A.A.; Alharbi, N.M. Hybrid Random Feature Selection and Recurrent Neural Network for Diabetes Prediction. *Mathematics* **2025**, *13*, 628. [\[CrossRef\]](#)
4. Dokeroglu, T.; Deniz, A.; Kiziloğlu, H.E. A comprehensive survey on recent metaheuristics for feature selection. *Neurocomputing* **2022**, *494*, 269–296. [\[CrossRef\]](#)
5. Dhal, P.; Azad, C. A comprehensive survey on feature selection in the various fields of machine learning. *Appl. Intell.* **2022**, *52*, 4543–4581. [\[CrossRef\]](#)
6. Pudjihartono, N.; Fadason, T.; Kempa-Liehr, A.W.; O’Sullivan, J.M. A review of feature selection methods for machine learning-based disease risk prediction. *Front. Bioinform.* **2022**, *2*, 927312. [\[CrossRef\]](#)
7. Rostami, M.; Berahmand, K.; Nasiri, E.; Forouzandeh, S. Review of swarm intelligence-based feature selection methods. *Eng. Appl. Artif. Intell.* **2021**, *100*, 104210. [\[CrossRef\]](#)
8. Bommert, A.; Welchowski, T.; Schmid, M.; Rahnenführer, J. Benchmark of filter methods for feature selection in high-dimensional gene expression survival data. *Briefings Bioinform.* **2022**, *23*, bbab354. [\[CrossRef\]](#)
9. Kamalov, F. Orthogonal variance decomposition based feature selection. *Expert Syst. Appl.* **2021**, *182*, 115191. [\[CrossRef\]](#)
10. Jolliffe, I.T.; Cadima, J. Principal component analysis: A review and recent developments. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2016**, *374*, 20150202. [\[CrossRef\]](#)
11. Jiang, Q.; Yan, X.; Yi, H.; Gao, F. Data-driven batch-end quality modeling and monitoring based on optimized sparse partial least squares. *IEEE Trans. Ind. Electron.* **2019**, *67*, 4098–4107. [\[CrossRef\]](#)
12. Karami, S.; Saber-Movahed, F.; Tiwari, P.; Marttinen, P.; Vahdati, S. Unsupervised feature selection based on variance–covariance subspace distance. *Neural Netw.* **2023**, *166*, 188–203. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Ambarwati, Y.S.; Uyun, S. Feature selection on magelang duck egg candling image using variance threshold method. In Proceedings of the 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 10 December 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 694–699.
14. Chen, C.W.; Tsai, Y.H.; Chang, F.R.; Lin, W.C. Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results. *Expert Syst.* **2020**, *37*, e12553. [\[CrossRef\]](#)
15. Fida, M.A.F.A.; Ahmad, T.; Ntahobari, M. Variance threshold as early screening to Boruta feature selection for intrusion detection system. In Proceedings of the 2021 13th International Conference on Information & Communication Technology and System (ICTS), Surabaya, Indonesia, 20–21 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 46–50.
16. Zouhri, W.; Homri, L.; Dantan, J.Y. Handling the impact of feature uncertainties on SVM: A robust approach based on Sobol sensitivity analysis. *Expert Syst. Appl.* **2022**, *189*, 115691. [\[CrossRef\]](#)
17. Zhang, P. A novel feature selection method based on global sensitivity analysis with application in machine learning-based prediction model. *Appl. Soft Comput.* **2019**, *85*, 105859. [\[CrossRef\]](#)
18. Tariq, S.; Dan, S. Sensitivity Analysis and Feature Selection for Drilling-Oriented Models. *J. Energy Resour. Technol.* **2023**, *145*, 123201. [\[CrossRef\]](#)
19. Kamalov, F.; Moussa, S.; El Khatib, Z.; Mnaouer, A.B. Orthogonal variance-based feature selection for intrusion detection systems. In Proceedings of the 2021 International Symposium on Networks, Computers and Communications (ISNCC), Dubai, United Arab Emirates, 31 October–2 November 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–5.
20. Rao, R.S.; Dewangan, S.; Mishra, A.; Gupta, M. A study of dealing class imbalance problem with machine learning methods for code smell severity detection using PCA-based feature selection technique. *Sci. Rep.* **2023**, *13*, 16245. [\[CrossRef\]](#)
21. Adhao, R.; Pachghare, V. Feature selection using principal component analysis and genetic algorithm. *J. Discret. Math. Sci. Cryptogr.* **2020**, *23*, 595–602. [\[CrossRef\]](#)
22. Omuya, E.O.; Okeyo, G.O.; Kimwele, M.W. Feature selection for classification using principal component analysis and information gain. *Expert Syst. Appl.* **2021**, *174*, 114765. [\[CrossRef\]](#)

23. Uddin, M.P.; Mamun, M.A.; Hossain, M.A. PCA-based feature reduction for hyperspectral remote sensing image classification. *IETE Tech. Rev.* **2021**, *38*, 377–396. [\[CrossRef\]](#)
24. Gopal, V.N.; Al-Turjman, F.; Kumar, R.; Anand, L.; Rajesh, M. Feature selection and classification in breast cancer prediction using IoT and machine learning. *Measurement* **2021**, *178*, 109442. [\[CrossRef\]](#)
25. Townes, F.W.; Hicks, S.C.; Aryee, M.J.; Irizarry, R.A. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol.* **2019**, *20*, 1–16. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Song, Y.; Zhang, J.; Zhang, Y.D.; Hou, Y.; Yan, X.; Wang, Y.; Zhou, M.; Yao, Y.F.; Yang, G. FeAture Explorer (FAE): A tool for developing and comparing radiomics models. *PLoS ONE* **2020**, *15*, e0237587. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Kaddar, B.; Fizazi, H.; Mansouri, D.E.K. Convolutional neural network features selection based on analysis of variance. In Proceedings of the 2019 International Conference on Theoretical and Applicative Aspects of Computer Science (ICTAACS), Skikda, Algeria, 15–16 December 2019; IEEE: Piscataway, NJ, USA, 2019; Volume 1, pp. 1–5.
28. Singha, J.; Roy, A.; Laskar, R.H. Dynamic hand gesture recognition using vision-based approach for human–computer interaction. *Neural Comput. Appl.* **2018**, *29*, 1129–1141. [\[CrossRef\]](#)
29. Tan, J.X.; Li, S.H.; Zhang, Z.M.; Chen, C.X.; Chen, W.; Tang, H.; Lin, H. Identification of hormone binding proteins based on machine learning methods. *Math. Biosci. Eng.* **2019**, *16*, 2466–2480. [\[CrossRef\]](#)
30. Tadist, K.; Najah, S.; Nikolov, N.S.; Mrabti, F.; Zahi, A. Feature selection methods and genomic big data: A systematic review. *J. Big Data* **2019**, *6*, 1–24. [\[CrossRef\]](#)
31. Roberts, A.G.; Catchpole, D.R.; Kennedy, P.J. Variance-based feature selection for classification of cancer subtypes using gene expression data. In Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–8.
32. Ebeunuwa, S.H.; Sharif, M.S.; Alazab, M.; Al-Nemrat, A. Variance ranking attributes selection techniques for binary classification problem in imbalance data. *IEEE Access* **2019**, *7*, 24649–24666. [\[CrossRef\]](#)
33. Chun, H.; Keleş, S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2010**, *72*, 3–25. [\[CrossRef\]](#)
34. Lee, L.C.; Liong, C.Y.; Jemain, A.A. Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: A review of contemporary practice strategies and knowledge gaps. *Analyst* **2018**, *143*, 3526–3539. [\[CrossRef\]](#)
35. Yuan, Z.; Zhang, L.; Wang, D.; Jiang, J.; Harrington, P.D.; Mao, J.; Zhang, Q.; Li, P. Detection of flaxseed oil multiple adulteration by near-infrared spectroscopy and nonlinear one class partial least squares discriminant analysis. *LWT* **2020**, *125*, 109247. [\[CrossRef\]](#)
36. Talukdar, U.; Hazarika, S.M.; Gan, J.Q. A kernel partial least square based feature selection method. *Pattern Recognit.* **2018**, *83*, 91–106. [\[CrossRef\]](#)
37. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1996**, *58*, 267–288. [\[CrossRef\]](#)
38. Fonti, V.; Belitser, E. Feature selection using LASSO. *VU Amst. Res. Pap. Bus. Anal.* **2017**, *30*, 1–25.
39. Zou, H. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **2006**, *101*, 1418–1429. [\[CrossRef\]](#)
40. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2005**, *67*, 301–320. [\[CrossRef\]](#)
41. Yuan, M.; Lin, Y. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2007**, *68*, 49–67. [\[CrossRef\]](#)
42. Deng, H.; Runger, G. Feature selection via regularized trees. In Proceedings of the 2012 International Joint Conference on Neural Networks (IJCNN), Brisbane, Australia, 10–15 June 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 1–8.
43. Candès, E.; Tao, T. The Dantzig selector: Statistical estimation when p is much larger than n . *Ann. Stat.* **2007**, *35*, 2313–2351.
44. Liu, H.; Motoda, H. *Computational Methods of Feature Selection*; CRC Press: Boca Raton, FL, USA, 2007.
45. Bouchlaghem, Y.; Akhiat, Y.; Amjad, S. Feature Selection: A Review and Comparative Study. *E3S Web Conf.* **2022**, *351*, 01046. [\[CrossRef\]](#)
46. Fan, J.; Li, R. Statistical challenges with high dimensionality: Feature selection in knowledge discovery. *arXiv* **2006**, arXiv:math/0602133.
47. Meinshausen, N. Relaxed Lasso. *Comput. Stat. Data Anal.* **2007**, *52*, 374–393. [\[CrossRef\]](#)
48. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [\[CrossRef\]](#)
49. Zhang, R.; Zhao, T.; Lu, Y.; Xu, X. Relaxed adaptive lasso and its asymptotic results. *Symmetry* **2022**, *14*, 1422. [\[CrossRef\]](#)
50. Huang, J.; Ma, S.; Zhang, C. Adaptive lasso for sparse high-dimensional regression models. *Stat. Sin.* **2008**, *18*, 1603.
51. Wang, W.; Liang, J.; Liu, R.; Song, Y.; Zhang, M. A robust variable selection method for sparse online regression via the elastic net penalty. *Mathematics* **2022**, *10*, 2985. [\[CrossRef\]](#)
52. Sudjai, N.; Duangsaphon, M.; Chandhanayingyong, C. Relaxed Adaptive Lasso for Classification on High-Dimensional Sparse Data with Multicollinearity. *Int. J. Stat. Med Res.* **2023**, *12*, 97–108. [\[CrossRef\]](#)
53. Simon, N.; Friedman, J.; Hastie, T.; Tibshirani, R. A sparse-group lasso. *J. Comput. Graph. Stat.* **2013**, *22*, 231–245. [\[CrossRef\]](#)

54. Wang, H.; Leng, C. Adaptive Sparse Group Lasso for High-Dimensional Generalized Linear Models. *Stat. Sin.* **2020**, *30*, 347–368.
55. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: Berlin/Heidelberg, Germany, 2009.
56. Breiman, L.; Friedman, J.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*, 1st ed.; Chapman and Hall/CRC: Boca Raton, FL, USA, 1984. [\[CrossRef\]](#)
57. Zhang, T. Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Stat.* **2004**, *32*, 56–85. [\[CrossRef\]](#)
58. Zhang, G.; Gionis, A. Regularized impurity reduction: Accurate decision trees with complexity guarantees. *Data Min. Knowl. Discov.* **2023**, *37*, 434–475. [\[CrossRef\]](#)
59. Kiran, B.R.; Serra, J. Cost-Complexity Pruning of Random Forests. In *Mathematical Morphology and Its Applications to Signal and Image Processing (ISMM)*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 195–206.
60. Chintalapudi, N.; Angeloni, U.; Battineni, G.; Di Canio, M.; Marotta, C.; Rezza, G.; Sagaro, G.G.; Silenzi, A.; Amenta, F. LASSO Regression Modeling on Prediction of Medical Terms among Seafarers' Health Documents Using Tidy Text Mining. *Bioengineering* **2022**, *9*, 124. [\[CrossRef\]](#)
61. Ng, A.Y. Feature selection, L1 vs. L2 regularization, and rotational invariance. In Proceedings of the Twenty-First International Conference on Machine Learning, Banff, AB, Canada, 4–8 July 2004; p. 78.
62. Guo, S.; Guo, D.; Chen, L.; Jiang, Q. A l1-regularized feature selection method for local dimension reduction on microarray data. *Comput. Biol. Chem.* **2017**, *67*, 92–101. [\[CrossRef\]](#) [\[PubMed\]](#)
63. Zhu, J.; Rosset, S.; Tibshirani, R.; Hastie, T. 1-norm support vector machines. *Adv. Neural Inf. Process. Syst.* **2003**, *16*.
64. Wang, L.; Zhu, J.; Zou, H. Hybrid huberized support vector machines for microarray classification. In Proceedings of the 24th International Conference on Machine Learning (ICML), Corvallis, OR, USA, 20–24 June 2007; ACM: New York, NY, USA, 2007; pp. 983–990.
65. Li, C.N.; Shao, Y.H.; Deng, N.Y. Robust L1-norm two-dimensional linear discriminant analysis. *Neural Netw.* **2015**, *65*, 92–104. [\[CrossRef\]](#)
66. Wang, H.; Lu, X.; Hu, Z.; Zheng, W. Fisher discriminant analysis with L1-norm. *IEEE Trans. Cybern.* **2015**, *44*, 828–842. [\[CrossRef\]](#)
67. Hao, N.; Feng, Y.; Zhang, H.H. Model selection for high-dimensional quadratic regression via regularization. *J. Am. Stat. Assoc.* **2018**, *113*, 615–625. [\[CrossRef\]](#)
68. Fujiwara, Y.; Ida, Y.; Shiokawa, H.; Iwamura, S. Fast Lasso Algorithm via Selective Coordinate Descent. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16), Phoenix, AZ, USA, 12–17 February 2016.
69. Fan, J.; Lv, J. Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2008**, *70*, 849–911. [\[CrossRef\]](#)
70. Li, R.; Zhong, W.; Zhu, L. Feature screening via distance correlation learning. *J. Am. Stat. Assoc.* **2012**, *107*, 1129–1139. [\[CrossRef\]](#)
71. Barut, E.; Fan, J.; Verhasselt, A. Conditional Sure Independence Screening. *J. Am. Stat. Assoc.* **2016**, *111*, 1266–1277. [\[CrossRef\]](#)
72. Amini, F.; Hu, G. A two-layer feature selection method using genetic algorithm and elastic net. *Expert Syst. Appl.* **2021**, *166*, 114072. [\[CrossRef\]](#)
73. Chamlal, H.; Benzmane, A.; Ouaderhman, T. Elastic net-based high dimensional data selection for regression. *Expert Syst. Appl.* **2024**, *244*, 122958. [\[CrossRef\]](#)
74. Park, T.; Casella, G. The Bayesian Lasso. *J. Am. Stat. Assoc.* **2008**, *103*, 681–686. [\[CrossRef\]](#)
75. Balamurugan, S.A.A.; Rajaram, R. Effective and Efficient Feature Selection for Large-scale Data Using Bayes' Theorem. *Int. J. Autom. Comput.* **2009**, *6*, 62–71. [\[CrossRef\]](#)
76. Tipping, M.E. Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* **2001**, *1*, 211–244.
77. Mohsenzadeh, Y.; Sheikhzadeh, H.; Zadeh, B. Relevant Sample-Feature Machine based on sparse Bayesian learning. *J. Mach. Learn. Res.* **2013**, *58*, 1226–1238.
78. Mirzaei, S.; Mohsenzadeh, Y.; Sheikhzadeh, H. Variational Embedded RSFM: A Bayesian model for feature selection. *Pattern Recognit. Lett.* **2017**, *105*, 1–8. [\[CrossRef\]](#)
79. Jiang, L.; Greenwood, C.M.T.; Yao, W.; Li, L. Bayesian Hyper-LASSO Classification for Feature Selection with Application to Endometrial Cancer RNA-seq Data. *Sci. Rep.* **2020**, *10*, 9747. [\[CrossRef\]](#)
80. Alzaatreh, A.; Emarly, M.; Mitra, R.; Al-Labadi, L. Feature Selection in Binary Classification Problems Using Relative Belief Ratio. **2024**, submitted.
81. Evans, M. Measuring statistical evidence using relative belief. *Comput. Struct. Biotechnol. J.* **2016**, *14*, 91–96. [\[CrossRef\]](#)
82. Hegazy, A.E.; Hafiz, B.; Makhoul, M.A.; Salem, O.A.M. Optimizing medical data classification: Integrating hybrid fuzzy joint mutual information with binary Cheetah optimizer algorithm. *Clust. Comput* **2025**, *28*, 250. [\[CrossRef\]](#)
83. Banjoko, A.W.; Yahya, W.B.; Olaniran, O.R. A multi-objective optimization algorithm for gene selection and classification in cancer study. *Appl. Soft Comput.* **2025**, *172*, 112911. [\[CrossRef\]](#)

84. Shin, H.; Oh, S. An effective heuristic for developing hybrid feature selection in high dimensional and low sample size datasets. *BMC Bioinform.* **2024**, *25*, 390. [[CrossRef](#)] [[PubMed](#)]
85. Anju, A.J.; Judith, J.E. Hybrid feature selection method for predicting software defect. *J. Eng. Appl. Sci.* **2024**, *71*, 124. [[CrossRef](#)]
86. Kamalov, F.; Sulieman, H.; Moussa, S.; Reyes JA.; Safaraliev, M. Nested ensemble selection: An effective hybrid feature selection method. *Heliyon* **2023**, *9*, e19686. [[CrossRef](#)] [[PubMed](#)]
87. Ali, W.; Saeed, F. Hybrid Filter and Genetic Algorithm-Based Feature Selection for Improving Cancer Classification in High-Dimensional Microarray Data. *Processes* **2023**, *11*, 562. [[CrossRef](#)]
88. Zhou, K.; Yin, Z.; Gu, J.; Zeng, Z. A feature selection method based on graph theory for cancer classification. *Comb. Chem. High Throughput Screen.* **2024**, *27*, 650–660. [[CrossRef](#)]
89. Rostami, M.; Forouzandeh, S.; Berahmand, K.; Soltani, M.; Shahsavari, M.; Oussalah, M. Gene selection for microarray data classification via multi-objective graph theoretic-based method. *Artif. Intell. Med.* **2022**, *123*, 102228. [[CrossRef](#)]
90. Tang, C.; Zheng, X.; Zhang, W.; Liu, X.; Zhu, X.; Zhu, E. Unsupervised feature selection via multiple graph fusion and feature weight learning. *Sci. China Inf. Sci.* **2023**, *66*, 152101. [[CrossRef](#)]
91. Banerjee, A.; Goswami, S.; Kumar Das, A. A hybrid graph centrality-based feature selection approach for supervised learning. In Proceedings of the ICDMAI 2020: Data Management, Analytics and Innovation, Kolkata, India, 16–18 January 2026; Springer: Berlin/Heidelberg, Germany, 2021; Volume 2, pp. 401–419.
92. Li, G.; Yu, Z.; Yang, K.; Lin, M.; Chen, C.L.P. Exploring feature selection with limited labels: A comprehensive survey of semi-supervised and unsupervised approaches. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*, 6124–6144. [[CrossRef](#)]
93. Fu, S.; Wang, S.; Liu, W.; Liu, B.; Zhou, B.; You, X.; Peng, Q.; Jing, X.-Y. Adaptive graph convolutional collaboration networks for semi-supervised classification. *Inf. Sci.* **2022**, *611*, 262–276. [[CrossRef](#)]
94. Faraji, M.; Seyedi, S.A.; Tab, F.A.; Mahmoodi, R. Multi-label feature selection with global and local label correlation. *Expert Syst. Appl.* **2024**, *246*, 123198. [[CrossRef](#)]
95. Gong, H.; Li, Y.; Zhang, J.; Zhang, B.; Wang, X. A new filter feature selection algorithm for classification task by ensembling pearson correlation coefficient and mutual information. *Eng. Appl. Artif. Intell.* **2024**, *131*, 107865. [[CrossRef](#)]
96. Fan, Y.; Liu, J.; Tang, J.; Liu, P.; Lin, Y.; Du, Y. Learning correlation information for multi-label feature selection. *Pattern Recognit.* **2024**, *145*, 109899. [[CrossRef](#)]
97. Kamalov, F.; Sulieman, H.; Cherukuri, A.K. Synthetic data for feature selection. In *International Conference on Artificial Intelligence and Soft Computing*; Springer Nature: Cham, Switzerland, 2023; pp. 353–365.
98. Kamalov, F.; Elnaffar, S.; Sulieman, H.; Cherukuri, A.K. XyGen: Synthetic data generator for feature selection. *Softw. Impacts* **2023**, *15*, 100485. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.