Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 13, No. 2, 2025, 159-173.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

A Hybrid Feature Selection Technique Leveraging Principal Component Analysis And Support Vector Machines

Sayyed Mohammad Hoseini*, Majid Ebtia and Mohanna Dehgardi

Gahar Artificial Intelligence Research Group, Ayatollah Boroujerdi University, Boroujerd, Iran.

Article Info

Article History:
Received 31 December 2024
Revised 30 January 2025
Accepted 01 March 2025

DOI:10.22044/jadm.2025.15276.2632

Keywords:
K-nearest Neighbor, Feature Selection, Support Vector Machine, Principal Component Analysis, Naïve Bayes.

**Corresponding author: sm.hoseini@abru.ac.ir (S. M. Hoseini).*

Abstract

The abundance of high dimensional datasets and the computational limitations of data analysis processes in applying to high-dimensional data have made clear the importance of developing feature selection methods. The negative impact of irrelevant variables on prediction and increasing unnecessary calculations due to the redundant attributes lead to poor results or performance of the classifiers. Feature selection is, therefore, applied to facilitate a better understanding of the datasets, reduce computational time, and enhance prediction accuracy. In this research, we develop a composite method for feature selection that combines support vector machines and principal component analysis. Then the method is implemented to the K-nearest neighbor and the Naïve Bayes algorithms. The datasets utilized in this study consist of three from the UCI Machine Learning Repository, used to assess the performance of the proposed models. Additionally, a dataset gathered from the central library of Ayatollah Boroujerdi University was considered. This dataset encompasses 1,910 instances with 30 attributes, including gender, native status, entry term, faculty code, cumulative GPA, and the number of books borrowed. After applying the proposed feature selection method, an accuracy of 70% was obtained with only five features. Experimental results demonstrate that the proposed feature selection method chooses appropriate feature subset. The approach yields enhanced classification performance, as evaluated by metrics such as accuracy, F_1-score and Matthews correlation coefficient.

1. Introduction

Machine learning algorithms can be applied on datasets for different tasks such as classification, pattern recognition or clustering. Due to technological advances, many datasets are available that have a large number of features. In data analysis, features are also referred as columns, attributes, tuples or variables and they are the basic building blocks of datasets. The quality of the results obtained by a machine learning algorithm strongly dependents on the quality of the features in the dataset. Thus the identification and elimination of irrelevant and redundant variables are very important tasks in preprocessing stage. This procedure is referred as feature selection or variable elimination [1].

Feature selection plays a crucial role in data mining and machine learning for various reasons. By identifying and choosing the most relevant and significant features, the performance of the models can be enhanced. Irrelevant or redundant features have the potential to introduce noise and diminish the accuracy of the model [2]. The inclusion of these features can lead to overfitting, where the model achieves high accuracy on training data but

fails to generalize to new, unseen data. Feature selection is instrumental in mitigating overfitting by focusing on the most important features. Furthermore, using fewer features can result in faster training and prediction times, which is particularly vital for large datasets and real-time applications.

For high-dimensional datasets, feature selection methods have become an integral part of the learning process. Feature selection has many benefits such as model simplicity, lower computation requirement and higher predictor performance [1]. In feature selection we aim to select a subset of variables in the dataset such that the tasks are still well accomplished while reducing effects from noise or irrelevant variables [2]. To do this, a measure is needed which is used to determine the relevance of each feature with the target variable. Next a process must be introduced to find the more efficient features.

There are three categories for supervised feature selection methods: filter, wrapper and embedded methods [1,3]. Filter methods are usually pre-processing procedures that independently consider each feature in the dataset and rank the features. Then low ranked features are eliminated and the rest are applied to a predictor. The main advantages of these schemes are their low computational cost and good generalizability [2]. Filter methods rank features solely based on statistical properties such as correlation or mutual information, independent of any machine learning models. While these methods are simple and cost-effective, they may overlook complex interactions between features. This inability to identify nonlinear interactions can lead to poor performance in the final models, especially in problems where features simultaneously affect the output.

In wrapper methods, all subsets of the features are examined by the predictor performance [3]. The predictor will find a subset with highest predictor performance via a search algorithm. In other words, these algorithms use the learning method on a subset of features. However, this interaction with the classifier leads to better performance results than filter methods. Wrapper methods utilize a machine learning model to evaluate various subsets of features. Although these methods can find optimal feature sets that enhance model performance, they are prone to overfitting. This is because they optimize the model on the training data and may become sensitive to noisy patterns or specifics of the training dataset. Additionally, wrapper methods require intensive computations due to the large search space of feature combinations, which can be impractical for large datasets. Embedded schemes perform feature selection during the training process [4]. In these methods, the search for an optimal subset of features is embedded into the classifier. These methods are able to record dependencies at lower computational costs than wrapper schemes.

Support vector machines (SVM) method has become widely used in machine learning and data mining due to its ability to classify data effectively, even in high-dimensional spaces [5]. Its application to feature selection has garnered significant attention, particularly because of the capacity to handle large datasets, make accurate predictions, and select relevant features in the process. SVM is particularly suitable for feature selection because they inherently perform a form of feature weighting through their decision boundary. In the context of feature selection, the SVM algorithm can be used to identify the most discriminative features that contribute to the best classification performance [6]. The process often involves using a feature selection criterion, such as the weights of the support vectors or the recursive feature elimination technique, to assess the relevance of features.

While feature selection focuses on identifying and retaining the most relevant attributes from the original dataset, feature extraction transforms the existing features into a new set of representative variables, aiming to preserve essential information in a reduced form. Feature extraction technique is also used to reduce the number of features but with a completely different approach from the feature selection methods. Feature extractions such as linear discriminant analysis, wavelet transform, fast Fourier transform, and principle component analysis (PCA) create new features from the original ones in the dataset and then discarding the used features [7]. The process should be done in such a way that the new reduced set of features contains most of the information of the original dataset without losing important or relevant information. In this approach, new features are created using some transformations on the original feature space, that is the main difference between feature selection and feature extraction.

Despite advancements in feature selection and extraction techniques, many existing methods still face significant challenges. PCA is widely used for dimensionality reduction and eliminating irrelevant features; however, it may lead to the loss of critical information, negatively affecting classification accuracy. Conversely, SVM is effective for feature selection and improve class separability, but they are highly sensitive to parameter tuning and can struggle with complex

datasets. In this work, we present a composite feature selection method that prepare the dataset for the supervised learning algorithms. In the proposed method, we first use PCA method to extract new feature from the features in the dataset. Then support vector machine (SVM) is used to rank the new extracted features. Finally low ranked features are eliminated and the rest of the features are provided to the considered predictor algorithms.

The proposed hybrid approach leverages the strengths of both PCA and SVM to overcome these limitations and enhance classification performance. This study systematically compares the proposed method with conventional techniques, demonstrating its impact on improving predictive accuracy. While many hybrid methods combine filter and wrapper techniques, our approach leverages PCA to reduce redundancy and noise in the data while preserving essential information, followed by SVM to select the most discriminative features. This sequential combination ensures both efficiency (through PCA) and effectiveness (through SVM) in feature selection.

PCA is a linear method that selects features based on maximum variance in the data [8]. PCA can limit its effectiveness in feature ranking since it does not inherently capture nonlinear relationships among variables. However, integrating PCA with SVM, particularly using nonlinear kernels such as the RBF kernel, can mitigate this limitation. PCA eliminates correlated and irrelevant features, transforming data into a space where SVM can more effectively define decision boundaries. The proposed method is particularly suited for high-dimensional datasets, where PCA effectively extracts new features before SVM refines the feature set. This is a key distinction from other hybrid methods that may struggle with scalability or overfitting in such scenarios.

We consider the K-nearest neighbors (KNN) and Naïve Bayes (NB) methods to examine and compare the suggested feature selection method.

This paper is organized as follows: In Section 2, a brief review of feature selection techniques is presented. Section 7 is devoted to the proposed feature selection method. In Section 8, we apply the proposed feature selection scheme on several famous datasets to illustrate the efficiency and applicability of the method.

2. Literature Review

In this section, we briefly review related researches on feature reduction schemes. In machine learning, feature selection is the procedure of eliminating irrelevant and redundant features from a data set especially high dimensional one in order to improve the performance of the classifier [2,9]. Feature selection speeds up the training process and improves the predictive accuracy [9].

Similar to the machine learning viewpoint, Omuya et al. [10] classified the feature selection approaches into supervised, unsupervised, and semi-supervised methods. In supervised schemes which can either be filter, wrapper, or embedded models [11], relevant features based on labeled datasets are selected [4,11,12-14]. Filter approaches such as relief method, information gain and Fisher score method evaluate features independently from interrelationship between features and have poor performance [10].

The wrapper technique facilitates interaction between the features and model by considering feature dependencies. These models like genetic algorithms, greedy forward selection, simulated annealing and recursive feature elimination [13,14] assess features based on their interrelationships, thereby optimizing algorithmic performance. However, this approach often results in high computational costs and overfitting. Embedded techniques, such as Ridge regression and Lasso algorithms, choose features that increase significantly the accuracy of the classifier. These methods integrate the feature selection procedure into the classification algorithm [4].

For datasets containing a mixed of labeled and unlabeled data, semi-supervised methods [15,16] can be used to recognize the relevance of features. Unsupervised feature selection approaches such as PCA, feature similarity, and discriminant analysis [12,17-19] are implemented on unlabeled datasets to select relevant features. PCA involves assessing correlations between variables to identify the most important principal components. Using this technique [20], the size of datasets, including a large number of interrelated features, can be reduced. Thus one of the main applications of PCA is in dimensionality reduction. PCA identifies the principal components of the dataset and helps us to analyze some more new valuable features instead of examining all the features by extracting those features that provide more information. PCA generates new features. The main advantages of using it is reducing the execution time for the algorithm and preventing overfitting of the model . This technique works well in labeled datasets [16].

PCA is a linear transformation-based dimensionality reduction technique that derives new features by combining original ones. Consequently, this transformation may reduce feature interpretability, as principal components no longer directly correspond to the initial features

[21]. However, this trade-off is often justified by improved model accuracy and reduced computational complexity. To maintain interpretability, PCA loading scores can be analyzed to understand the contribution of each original feature to the principal components.

Although the PCA method involves extensive matrix computations and characteristic polynomial root-finding, several techniques have been developed to reduce its computational cost. One such technique is parallel computing [22-24], which allows computations to be distributed across multiple processors, significantly speeding up the process. Additionally, optimization methods like Incremental PCA [25] and Randomized PCA [26] have been introduced to further enhance efficiency. Incremental PCA updates the principal components incrementally as new data arrives, avoiding the need to recompute the entire dataset. This method is particularly useful for large-scale datasets where storing and processing all data at once is impractical. Randomized PCA, on the other hand, employs randomization techniques to approximate the principal components, leading to faster computations while maintaining high accuracy. These advancements in PCA methodologies have made it feasible to apply PCA to larger datasets and more complex problems, making it a valuable tool in various fields such as machine learning, data mining, and signal processing.

Cortes and Vapnik [27] developed SVM method that can be used in a variety of fields of study, including feature selection. The SVM algorithms are used for both classification and regression analysis. Support vector machines can be used for feature selection by leveraging the properties of the SVM algorithm to identify the most relevant features for classification or regression tasks. Feature selection using SVM involves identifying a subset of input variables that are most informative for the learning task, thereby improving model performance and reducing dimensionality. SVM can be used for feature selection as follows [28]:

- *Using Feature Weights*: After training an SVM model, the learned weights (coefficients) associated with each feature can be analyzed to identify the most influential features. In linear SVM, the magnitude of the weight vector w can indicate the importance of each feature. Features with higher absolute weights are considered more important for the classification task.
- *Recursive Feature Elimination (RFE)*: This technique involves iteratively training an SVM model on subsets of variables while systematically removing the least important features. During each iteration, feature weights are evaluated, and the least significant features are pruned until the desired number of features is achieved. RFE is effective for identifying a subset of variables that contribute the most to the model's performance.
- *Kernel Trick for Non-linear Feature Selection*: When using non-linear SVM with kernel functions (e.g., polynomial kernel, RBF kernel), the transformed feature space can highlight the most relevant features for the classification task. By examining the support vectors and their associated feature weights in the kernel-induced feature space, one can identify the most discriminative features.
- *Embedded Feature Selection*: Some SVM implementations, especially those with regularization (e.g., L^1 regularization), inherently perform feature selection during model training. Regularization terms penalize the inclusion of irrelevant features, effectively encouraging the model to focus on the most informative features.
- *SVM-Based Wrapper Methods*: These approach use SVM as a black-box model to evaluate subsets of features and select the best subset based on model performance. These methods employ SVM as a feature evaluator and optimize feature subsets based on the SVM model's performance.

By leveraging these techniques, SVM can effectively perform feature selection by identifying the most relevant features for the learning task, leading to improved model generalization, reduced overfitting, and enhanced interpretability.

Feature selection can also be categorized into univariate selection, feature importance and a combination of those [10]. In univariate selection, the relation between each two features and specially between the independent variables and the target variable are computed using correlation matrix. Information gain, Chi-squared, symmetrical uncertainty and minimum relevance maximum redundancy can be considered as univariate selection methods [1,29]. These approaches are commonly referred to as statistical-based feature selection methods as they utilize statistical tests to identify features with the powerful correlation to the output variable. Nevertheless, statistical methods could not be

proper for all datasets due to their reliance on the data type.

In feature importance techniques [30,31], the features are ranked based on their relevance or importance to the target variable. The ranking of the input features is done by the model. Xu et al. [30] explored the prediction of anticancer drug responses, which vary among patients due to genetic factors like mutations and RNA expression. They focused on the feature selection aspect for classification models by first employing an autoencoder network to reduce the dimensionality of genetic data and select significant input features. Subsequently, they utilized the Boruta algorithm to further refine the feature set for a random forest model used in predicting drug responses. However, their approach faces challenges in identifying key features and handling imbalanced datasets, which can affect generalizability. A feature importance ranking measure proposed by Zien et al. [31] for feature selection that demonstrated outstanding predictive accuracy, contingent on the distribution and size of input features. However, the computational time and stability of feature selection algorithms can be impacted by this measure.

Composite feature selection approach uses a combination of two or more techniques. Various types of composite schemes have been considered in the literature [32-36]. Raghavendra et al. [34] implemented a combination of feature selection methods based on entropy value, mean value, and threshold value, along with forward selection and backward elimination techniques, on small-scale medical datasets. The method encounters challenges when dealing with noisy datasets and exhibits high computational complexity when applied to large-scale data. The method results in a stable model with improved performance, although the main issues include lengthy training times. Stability of feature selection schemes, which means that a small perturbation on the training data leads to a small different feature selection result, has been investigated in [32,33,35].

In this paper, we suggest a composite feature selection method based on a combination of PCA and SVM.

3. The proposed feature selection method

We now introduce the new composite method to reduce the dimensions of the data in order to improve the performance of the classification algorithms via selecting the best subset of features. It composes of two sequential stages: the feature extraction stage using PCA and the feature selection stage via SVM.

3.1. Principal Component Analysis

One of the famous statistical methods for dimensionality reduction is the PCA approach. This approach is useful and desirable when there are many variables with high correlation in the investigated data set. The main idea of the PCA approach is to find a number of uncorrelated linear combinations of the main correlated variables that contain the most variance structure of the original data. These uncorrelated linear combinations that give a better interpretation and understanding of the sources of changes in the data are called principal components. The directions of the principal components are chosen so that they are perpendicular to each other and successively maximize the variance of the projected data.

In practice, the first component is related to the direction in which the transformed observations have the highest variance. Next the second component is perpendicular to the first component and this time the variance of the points transformed on it is the highest, but its value is less than the variance of the first component. This process continues and the next main components are produced in order. One of the main important properties of the PCA approach is that the first few principal components account for a significant proportion of the changes in the main variables.

Let $X = \{x_1, x_2, \ldots, x_n\}$ denote the dataset of n observations, where each one has p numerical variables, i.e. $x_i = (x_i^1, x_i^2, \ldots, x_i^p)^T \in \mathbb{R}^p$ where T denotes transpose. In order to avoid biased results, we suppose that the dataset X is standardized. For convenience, we set these data values in an $n \times p$ data matrix X, whose i-th column is the i-th feature of the observations. PCA method calculate firstly the covariance matrix as follows:

$$Cov(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T, \qquad (1)$$

where $\bar{x}$ is the mean vector of the dataset X. The covariance matrix is utilized to assess the interdependence and correlation among variables. Subsequently, the spectral decomposition of the covariance matrix involves the use of eigenvectors $v_1, v_2, \ldots, v_p$ and eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. It should be pointed out that these eigenvectors are actually the principal components and the eigenvalues give the amount of variance carried in each principal component. Thus we rank the eigenvalues in descending order which means that the eigenvector corresponding to the first principal component is v_1 and the one corresponding to the

second principle component is v_2 and so on. Then we get the principal components in order of significance. Now we can keep all principle components or discard components with low eigenvalues. Then we form a matrix of the considered eigenvectors:

$$V = (v_1, v_2, v_3, \ldots, v_d),$$

where $d \leq p$ and V is a $p \times d$ matrix that has as columns the eigenvectors of the d first principle components. Finally, the feature matrix V is used to reorient the original dataset X to the axes represented by the principal components as follows:

$$Y = (X - \bar{x})V.$$

The transformed dataset Y is used to next feature selection or variable elimination stage. The advantages of this method can be mentioned as follows: reduction of noise in the data, prevention of overfitting and elimination of interdependent features.

PCA concerns with the converted space based on the eigenvectors, thus it may lose maximum classification sensitivity due to the removal of several sensitive features [37,38].The objective of the PCA model is to create a new set of features with reduced dimensions compared to the primary dataset documented in the literature. The approach transforms a p-dimensional dataset into a lower-dimensional dataset such as d, where $d \leqslant p$. But we use the PCA algorithm to convert the original p-dimensional dataset into a more useful space with the same dimension.

3.2. Support Vector Machines

Support vector machines for the dataset X is defined as the following optimization problem, aiming to maximize the margin between data classes while minimizing classification errors. This formulation ensures effective generalization on unseen data.

$$\textbf{Minimize } \frac{1}{2} \| w \|^2 .$$

subject to the following constraints for each training sample (x_i, y_i):

$$y_i (w^T x_i + b) \geq 1.$$

In addition to the constraints, the optimization problem may include soft-margin constraints to handle cases where the data is not linearly separable. This allows for some misclassifications by introducing a penalty for misclassified points, balancing the margin maximization and the classification error.

The solution to the maximization problem can be achieved through techniques like quadratic programming or convex optimization to determine the optimal values of w and b defining the separating hyperplane.

In the context of SVM, a kernel function transforms the input data into a higher-dimensional space, enabling the SVM to find a linear separation that may correspond to a nonlinear separation in the original input space.

The kernel function plays a crucial role in SVM because it allows the algorithm to efficiently find non-linear decision boundaries. Instead of explicitly mapping the input data to a higher-dimensional space, which can be computationally expensive, the kernel function calculates the inner product between the transformed data points without explicitly transforming them. This is known as the “kernel trick”. The most popular kernel functions include:

- Linear;
- Radial Basis Function (RBF);
- Polynomial;
- Sigmoid.

By using an appropriate kernel function, SVM can effectively model nonlinear relationships in the data and find optimal decision boundaries. The choice of the kernel function and its hyperparameters is an important consideration when using SVM for classification or regression tasks.

4. Classification algorithms

In this section we present two simple classifiers: K-Nearest Neighbors and Naïve Bayes. These classifiers are both popular classification algorithms in machine learning with their own unique benefits. Naïve Bayes is an algorithm with simple and easy to understand and efficient and fast for training and prediction, which performs well in multi-class prediction problems. This method works well with large datasets and handles missing data well. KNN is a non-parametric and flexible that can handle non-linear data with no assumptions about the data distribution. It is easy to interpret and explain which can handle multi-class classification problems.

4.1. Naïve Bayes

The Naïve Bayes classifier is an easy-to-understand and robust probabilistic approach that applies Bayes' theorem, operating under the assumption that features are independent [9]. It is widely used in machine learning for classification tasks, especially for text classification and spam filtering. This method is one of the machine

learning algorithms with conditional independence assumption. In other words, a group of simple classifiers, assuming the independence of random variables and based on Bayesian theorem, which is one of the most important and widely used concepts of probability, form the Bayesian classifier. From the Bayes rule, the probability of a sample x being class c can be calculate as follows:

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)},$$

where $P(c)$ the probability of the prior class, $P(x)$ the probability of the prior property, $P(c \mid x)$ the posterior probability, and $P(x \mid c)$ the likelihood estimation. The corresponding classifier is the function that assigns a class label $\hat{y}$ as follows:

$$\hat{y} = \arg\max_y P(y)\prod_{i=1}^{n} P(x_i \mid y).$$

Although it relies on a simplistic and "Naïve" assumption, Naïve Bayes often delivers surprisingly effective results, especially in text classification applications. It is robust to irrelevant features and can handle high-dimensional data efficiently. However, it may not capture complex relationships between features, and its performance can degrade if the independence assumption is severely violated [5].

4.2. K -Nearest Neighbors

As a popular machine learning approach, the KNN algorithm is used for regression and classification tasks. It identifies the K nearest data points to an input and predicts based on the average value for regression or the predominant class for classification [39].

When employing the KNN algorithm, the distance between the input and each data point in the training set is computed. The K nearest data points are then selected based on their distances, and the majority class or average value of those neighbors is utilized to formulate the prediction. Therefore, it is necessary to specify a criterion to determine the distance between data points. We consider the Euclidean distance to get the distance between the two data points. In the K-nearest neighbor algorithm, a category is determined by selecting k samples from the training set that are most similar to each other. The decision for a new test sample is then based on the majority category or label within this selected neighborhood. In essence, the category assigned to the new sample should have the highest number of samples in the chosen neighborhood. Consequently, after calculating the Euclidean distance between the points, the elements are sorted based on this distance. Subsequently, the label that is most prevalent among the K neighbors is assigned to the unknown sample [39].

The value of K, i. e. the number of neighbors to consider, is a pivotal parameter in the KNN algorithm when making predictions. Choosing a small value for K can expose the algorithm to noise in the data, while a larger K may cause oversmoothing.

KNN is a straightforward and easy-to-understand algorithm, although it can be computationally demanding, particularly with extensive datasets. Furthermore, it can be influenced by the value of K and the choice of distance metric. Nonetheless, KNN remains widely utilized, particularly for small to medium-sized datasets and when interpretability is a priority.

5. Proposed Method

In this study, we propose a feature selection method based on PCA and SVM. Our proposed feature selection method leverages both PCA and SVM algorithms to more efficiently reduce the dimensionality of the datasets. We advocate for an initial feature selection using SVM, followed by presenting the remaining important features to the PCA algorithm for conversion into a new dataset, after which redundant features are once again removed. Finally, the data obtained from the principal component analysis algorithm are presented by KNN and NB methods.

The proposed method for this research includes the following steps:

1. The first step is to pre-process the data set, which involves data cleaning and normalization. Data cleaning includes detecting and then removing or modifying missing and noisy data from the data set. Then, the data set is normalized. There are several ways to normalize data, but we normalize any feature within the range of $[0, 1]$.
2. The first feature selection is done via SVM with a linear kernel, and important features are identified while irrelevant and insignificant features are removed. This is done based on the values of w, which represents the normal vector of the separation hyperplane. Each attribute is assigned to an axis, and the corresponding component in w indicates the importance of the associated attribute. This

value is a number in the range $[0, 1]$; the closer it is to one, the more important the attribute is.

3. In the third step, the features are extracted from the data obtained from SVM feature selection using the PCA method. This method extracts the features by converting the main feature space and creating new features based on the singular values of the data set. Then, the new features with low singular values are removed.
4. Finally, the data generated by the principal component analysis algorithm is presented to the K-nearest neighbor and simple Bayesian classifiers, and two models are constructed.

Existing feature selection and extraction methods often suffer from limitations such as reduced accuracy, inefficiency in handling complex patterns, and reliance on assumptions that may not hold in real-world scenarios. Traditional approaches, such as PCA for feature extraction and SVM for feature selection, each have their own drawbacks—PCA assumes linear relationships and is sensitive to data scaling, while SVM struggles with high-dimensional spaces and may not effectively eliminate redundant features. To address these issues, our proposed hybrid approach integrates PCA and SVM, leveraging the strengths of both methods to enhance feature relevance while mitigating their individual weaknesses. This integration improves classification accuracy and provides a more robust and adaptive feature selection mechanism.

To evaluate the efficiency improvement of our proposed method, we compare it with three different approaches. In the first approach, the dataset is utilized for KNN and NB classifiers without performing any feature selection. The second method involves the SVM algorithm to highlight essential features and remove redundancies prior to developing a classification model using the NB and KNN algorithms. The third approach employs the PCA method for feature selection, creating new features through transformation and subsequently eliminating the least important ones. Similar to the second approach, the new dataset is presented to the NB and KNN classifiers. Both these approaches involve a feature selection stage before classification.

6. Experimentation and Result Analysis

The experiments were conducted on three datasets, representing real-life classification problems obtained from the UCI Machine Learning Repository [40]. This repository is a collection of data generators, domain theories, and databases, used by the machine learning community for empirical analysis of machine learning algorithms. Established in 1987 by David Aha and other graduate students at UC Irvine, the repository has become a popular resource for students, teachers, and researchers globally for machine learning datasets. Each dataset consists of samples, each with a specific number of attributes. Table 1 offers a brief overview of all the datasets utilized. In Table 2, the accuracy results of different machine learning methods and feature selection approaches are reported.

Table 1. Description of datasets used in the experiment.

Database	d	n
Arrhythmia	279	452
German Credit	24	1000
Australian Credit	14	690
Colon	2000	62
Madelon	500	4400
SRBCT	2308	83
Leukemia	7129	72
Arcene	10000	900
Prostate Tumor	10509	102
Lung Cancer	12600	203

6.1. A practical case study on factors influencing student activity levels

We analyze data gathered from the central library of Ayatollah Boroujerdi University using our proposed methods. This dataset includes various student information such as gender, native status, entry term, faculty code, group code, educational group name, field code and study name, educational level, course details, overall grade point average, total credits earned, counts of conditional courses (both total and sequential), last term details, passed credits in the last term, grade point average for the last term, current student status, type of admission to higher education and the university, academically gifted student status, special student status, special enrollment status, years of attendance, number of semesters on leave, graduation level, and field of study at graduation.

Additionally, we have included an “Activity” column to indicate each student’s level of academic engagement. This activity level is determined by the number of books borrowed: a count of zero is denoted as ‘0’; if the average number of borrowed books is 20, counts below this are marked as ‘A’ and counts at or above this as ‘B’. After preprocessing to remove missing values and outliers, the dataset was refined to 1,910 samples and 30 features.

Table 2. Accuracy of machine learning methods obtained from various feature selection methods on the datasets.

Data set	Feature selection approach	Machine Learning method	ACC
Arrhythmia	Fuzzy Logic [41]	Bagging	0.4972
	Bagging [42]	NB	0.633
	Without FS [43]	Random forest	0.69
	Matched [44]	KNN	0.5865
German	Genetic algorithm-Particle Swarm Optimization [45]	KNN	0.7416
	LASSO [46]	SVM	0.9979
Australian	Genetic algorithm-Particle Swarm Optimization [45]	KNN	0.8609
	LASSO [46]	SVM	0.8797
Colon	Fast Correlation Based Filter Solution [47]	Hybrid Ensemble Method	0.814
	Without FS [48]	Artificial Neural Networks	0.9804
	PCA-Truncated Singular Value Decomposition [49]	KNN	0.875
	Clustering- Genetic Algorithm [50]	NB	0.745
Madelon	Fuzzy Logic [41]	Bagging	0.7599
	RNK-Genetic Algorithm [51]	KNN	0.8695
	Online PCA [52]	-	0.59
SRBCT	Gene Selection [53]	KNN	0.981
	PCA-Truncated Singular Value Decomposition [49]	KNN	0.8146
	PCA-Forward Selection[53]	KNN	0.9474
	Decision Tree [55]	Pearson PCA	0.9645
	Linear Discriminant Analysis [56]	Genetic algorithm	0.95
Leukemia	Information Gain [57]	KNN	100
	Fast Correlation Based Filter Solution [47]	Hybrid Ensemble Method	0.983
	Gene Selection [54]	KNN	0.979
	PCA-Truncated Singular Value Decomposition [49]	KNN	0.9305
	Decision Tree [55]	Pearson PCA	100
	Linear Discriminant Analysis [56]	Genetic algorithm	0.9421
	Neural Networks [58]	Genetic Algorithm-PCA	0.8823
	Clustering-Genetic Algorithm [50]	NB	0.9118
Arcene	RNK-Genetic Algorithm [51]	KNN	0.852
	Clustering-Genetic Algorithm [50]	NB	0.9032
Prostat Tumor	Fast Correlation Based Filter Solution [47]	Hybrid Ensemble Method	0.929
	Gene Selection [54]	KNN	0.935
	PCA-Truncated Singular Value Decomposition [49]	KNN	0.8614
	Decision Tree [55]	Pearson PCA	0.94
Lung cancer	Gene Selection [54]	KNN	0.937
	Decision Tree [55]	Pearson PCA	0.9674
	PCA [59]	KNN	0.96

6.2. Evaluation measures of model performance

Evaluating the performance of classifiers, algorithms can be measured by various criteria [5]. To do this, the dataset are divided into train and test subsets. The train dataset is used to obtain model of classifier and then by applying the model on the test dataset, the model performance can be evaluated. The following parameters can be calculated for two class target problems:

- TP denotes the number of correctly classified instances of a specific class;
- TN denotes the number of correctly classified instances that did not belong to the specific class;
- FN is the number of instances that incorrectly assigned to another class.
- FP is the number of instances that incorrectly assigned to the specific class;

Here, we use three measures: accurac, Matthews correlation coefficient and F_1 score that can be calculated as follows [60]:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN},$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}},$$

$$F_1 = 2\left(\frac{Precision \times Sensitivity}{Precision + Sensitivity}\right),$$

where

$$Sensitivity = \frac{TP}{TP + FN}, \quad Precision = \frac{TP}{TP + FP}.$$

7. Result Analysis

In this section, we present the results and evaluate the performance of the models based on specific criteria across the datasets. The evaluation criteria for the three datasets are reported in Tables 2-11.

For Arrhythmia dataset we get the following results. In the training dataset, the KNN model outperforms all three evaluation criteria in the second view, while the NB model excels in the third view. However, in the fourth view, the NB model surpasses the KNN model, indicating that even with reduced dimensions and fewer features, the models perform well.

In the experimental dataset, the KNN model demonstrates better results in MCC and ACC criteria in the fourth view, while the NB model excels in all evaluation criteria in the same perspective. Overall, reducing dimensions has led to effective predictions by the models. The NB model outperforms the KNN model in all evaluation criteria in the fourth approach. For the Arrhythmia dataset, the combination of SVM and PCA with the NB algorithm provides the best

performance. Applying the approaches on the German bank dataset leads to the following results. In the training database, the KNN model performs better in the fourth view, while the NB model excels in all three criteria in the same perspective. The composite model performs well in the reduced dimension scenario, with the NB model reporting superior results.

In the testing dataset, the KNN model outperforms the other views in all evaluation criteria in the fourth view, while the NB model performs well in the third view. Both models demonstrate effective performance despite the reduced dimensions. the combination of SVM and PCA significantly enhances the performance of KNN, demonstrating the effectiveness of feature reduction.

The Australian bank dataset gives the following results. In the KNN training dataset, the third view yields better results for all evaluation criteria. The NB model also shows promising results in the fourth view, indicating the effectiveness of the proposed composite model. In the testing dataset, the KNN model performs better in the fourth view for the MCC criterion and excels in the F_1 and ACC criteria in the second view. The NB model demonstrates superior performance in all evaluation criteria in the fourth approach, outperforming the KNN model. For the Australian credit dataset, the combination of SVM and PCA with the NB algorithm provides the best performance. The Australian credit dataset shows significant improvement with feature reduction, particularly using SVM and PCA combined. NB benefits most from this combination.

PCA often leads to better performance compared to no feature reduction. It reduces dimensionality while preserving important information. SVM feature selection, generally improves performance further, especially when combined with PCA. Consistently, the combination of SVM and PCA provides the best results across different datasets, indicating the power of combining these techniques. In addition, KNN shows significant improvement with feature reduction, particularly with the SVM and PCA combination. Also, NB benefits from feature reduction, but the impact is more pronounced with SVM and PCA combined.

Overall, the use of combined methods such as SVM and PCA significantly improves the performance of various machine learning algorithms. The NB algorithm also shows the best performance in many cases, especially when combined with feature selection methods.

7.1. Result analysis on factors influencing student activity levels

The proposed analysis provides important insights into the factors influencing student activity levels, highlighting the relative significance of each feature in predicting engagement. The “Last term” feature is the strongest predictor with a weight of 0.65, indicating that recent academic performance greatly impacts student engagement. “Overall grade point average” follows closely with a weight of 0.49, signifying a strong correlation between cumulative performance and activity levels. Significant predictors also include “Educational group name” (0.47) and “Entry term” (0.44), suggesting that both the timing of enrollment and the educational grouping affect student engagement. Additionally, “Number of semesters on leave” (0.42) implies that academic interruptions can influence activity levels.

While these key features play a major role, it is important to recognize that other factors with lower weights might still be relevant in specific contexts. Further investigation, combined with qualitative insights and field expertise, can deepen our understanding of the dynamics of student engagement.

The analysis also highlights the effectiveness of advanced feature selection techniques. The evaluation results obtained using the proposed methods are reported in Table 5. In the Faculty of Basic Sciences, applying SVM and PCA for dimensionality reduction improved KNN labeling accuracy to 70.7%, while in the Faculty of Engineering, the NB algorithm using SVM and PCA feature selection achieved a highest MCC of 0.407 and an accuracy of 70.2%. These results demonstrate how such techniques can enhance classification performance across various academic fields.

8. Acknowledgment

We would like to extend our heartfelt gratitude to the anonymous referees for their valuable time, insightful comments, and constructive suggestions. Their feedback has significantly improved the quality and clarity of this work. We deeply appreciate their expertise and dedication, which have been instrumental in shaping the final version of this manuscript.

This paper was supported in part by Ayatollah Boroujerdi Universtiy, project grant IR01101401000683 on data driven analysis in academic library of Ayatollah Boroujerdi University using machine learning methods.

Table 3. Model performance based on the datasets.

Dataset	Feature selection approach	No. features	Classifier	Train			Test		
				MCC	F_1	ACC	MCC	F_1	ACC
Arrhythmia	--	279	KNN	0.383	0.277	0.628	0.375	0.278	0.637
			NB	0.395	0.607	0.371	0.103	0.168	0.164
	PCA	143	KNN	0.432	0.322	0.65	0.372	0.261	0.637
			NB	0.364	0.581	0.335	0.195	0.186	0.208
	SVM	8	KNN	0.383	0.291	0.631	0.425	0.291	0.659
			NB	0.515	0.652	0.695	0.421	0.316	0.648
	PCA and SVM	8	KNN	0.374	0.29	0.628	0.449	0.272	0.67
			NB	0.502	0.641	0.686	0.515	0.408	0.703
German credit	--	24	KNN	0.366	0.672	0.755	0.288	0.63	0.725
			NB	0.418	0.704	0.737	0.406	0.699	0.73
	PCA	9	KNN	0.52	0.752	0.81	0.386	0.687	0.755
			NB	0.375	0.671	0.702	0.413	0.701	0.73
	SVM	12	KNN	0.341	0.655	0.748	0.251	0.598	0.72
			NB	0.315	0.648	0.735	0.416	0.703	0.765
	PCA and SVM	4	KNN	0.443	0.711	0.782	0.495	0.744	0.795
			NB	0.288	0.644	0.703	0.398	0.698	0.74
Australian credit	--	14	KNN	0.733	0.865	0.867	0.671	0.832	0.84
			NB	0.616	0.8	0.807	0.606	0.785	0.804
	PCA	5	KNN	0.729	0.863	0.865	0.701	0.85	0.855
			NB	0.556	0.751	0.769	0.603	0.771	0.797
	SVM	4	KNN	0.78	0.889	0.891	0.595	0.795	0.804
			NB	0.703	0.851	0.853	0.626	0.812	0.818
	PCA and SVM	1	KNN	0.711	0.854	0.855	0.709	0.847	0.847
			NB	0.714	0.853	0.853	0.744	0.861	0.862
Colon	--	2000	KNN	0.752	0.859	0.884	0.309	0.525	0.684
			NB	0.563	0.721	0.721	0.069	0.521	0.526
	PCA	40	KNN	0.703	0.827	0.86	0.093	0.49	0.631
			NB	0.795	0.898	0.907	0.287	0.627	0.632
	SVM	789	KNN	1	1	1	0.408	0.68	0.737
			NB	0.621	0.766	0.767	0.233	0.614	0.631
	PCA and SVM	30	KNN	0.55	0.716	0.791	0.449	0.636	0.737
			NB	0.898	0.949	0.953	0.268	0.593	0.684
Madelon	--	500	KNN	0.357	0.668	0.673	0.214	0.598	0.604
			NB	0.466	0.733	0.733	0.244	0.622	0.628
	PCA	60	KNN	0.488	0.741	0.742	0.383	0.688	0.69
			NB	0.331	0.666	0.666	0.205	0.602	0.602
	SVM	190	KNN	0.472	0.723	0.728	0.375	0.675	0.681
			NB	0.344	0.672	0.672	0.259	0.629	0.63
	PCA and SVM	20	KNN	0.686	0.843	0.843	0.613	0.806	0.806
			NB	0.271	0.636	0.636	0.262	631	0.631

Table 4. Model performance based on the datasets.

Dataset	Feature selection approach	No. features	Classifier	Train			Test		
				MCC	F_1	ACC	MCC	F_1	ACC
SRBCT	--	2308	KNN	1.000	1.000	1.000	0.696	0.785	0.760
			NB	1.000	1.000	1.000	1.000	1.000	1.000
	PCA	50	KNN	1.000	1.000	1.000	0.742	0.825	0.800
			NB	1.000	1.000	1.000	0.779	0.828	0.840
	SVM	986	KNN	1.000	1.000	1.000	0.896	0.939	0.920
			NB	1.000	1.000	1.000	1.000	1.000	1.000
	PCA and SVM	10	KNN	0.977	0.983	0.983	1.000	1.000	1.000
			NB	1.000	1.000	1.000	1.000	1.000	1.000
Leukemia	--	7129	KNN	0.917	0.956	0.960	0.690	0.817	0.818
			NB	1.000	1.000	1.000	1.000	1.000	1.000
	PCA	30	KNN	0.810	0.895	0.900	0.909	0.952	0.954
			NB	0.775	0.887	0.900	0.598	0.790	0.818
	SVM	2851	KNN	1.000	1.000	1.000	0.908	0.952	0.954
			NB	1.000	1.000	1.000	1.000	1.000	1.000
	PCA and SVM	10	KNN	1.000	1.000	1.000	1.000	1.000	1.000
			NB	0.956	0.977	0.980	0.904	0.949	0.955
Arcene	--	10000	KNN	1.000	1.000	1.000	0.471	0.732	0.733
			NB	0.899	0.949	0.95	0.047	0.511	0.55
	PCA	60	KNN	1.000	1.000	1.000	0.509	0.749	0.75
			NB	0.633	0.814	0.814	0.222	0.600	0.600
	SVM	3832	KNN	1.000	1.000	1.000	0.489	0.744	0.750
			NB	0.885	0.942	0.943	0.312	0.653	0.667
	PCA and SVM	40	KNN	1.000	1.000	1.000	0.661	0.826	0.833
			NB	0.710	0.855	0.857	0.471	0.732	0.733
Prostate Tumor	--	10509	KNN	1.000	1.000	1.000	0.575	0.765	0.774
			NB	0.295	0.612	0.634	0.506	0.676	0.710
	PCA	40	KNN	1.000	1.000	1.000	0.556	0.77	0.774
			NB	0.549	0.731	0.746	0.609	0.758	0.774
	SVM	3772	KNN	1.000	1.000	1.000	0.575	0.765	0.774
			NB	0.556	0.772	0.775	0.609	0.758	0.774
	PCA and SVM	10	KNN	0.836	0.915	0.915	0.747	0.870	0.871
			NB	0.696	0.844	0.845	0.631	0.801	0.807
Lung Cancer	--	12600	KNN	0.828	0.716	0.915	0.722	0.642	0.869
			NB	0.933	0.963	0.965	0.775	0.813	0.885
	PCA	40	KNN	0.857	0.734	0.929	0.803	0.716	0.902
			NB	0.93	0.958	0.965	0.651	0.638	0.836
	SVM	5150	KNN	0.886	0.826	0.944	0.830	0.842	0.918
			NB	1.000	1.000	1.000	0.900	0.886	0.951
	PCA and SVM	10	KNN	0.944	0.908	0.972	0.865	0.866	0.934
			NB	0.944	0.945	0.972	0.831	0.814	0.918

Table 5. The results obtained from the proposed method on the data from the central library of Ayatollah Boroujerdi University, categorized by faculties.

Faculty	Feature selection approach	No. of features	Classifier	Test		
				MCC	F_1	ACC
Basic Sciences	-	30	KNN	0.279	0.639	0.639
			NB	0.380	0.682	0.686
	PCA	8	KNN	0.303	0.651	0.651
			NB	0.326	0.662	0.662
	SVM	11	KNN	0.349	0.674	0.674
			NB	0.406	0.693	0.697
	PCA and SVM	5	KNN	0.424	0.707	0.707
			NB	0.380	0.682	0.686
Humanities	-	30	KNN	0.356	0.675	0.678
			NB	0.237	0.556	0.597
	PCA	8	KNN	0.289	0.643	0.644
			NB	0.315	0.657	0.657
	SVM	11	KNN	0.321	0.655	0.658
			NB	0.190	0.577	0.590
	PCA and SVM	5	KNN	0.360	0.677	0.678
			NB	0.232	0.599	0.610
Engineering	-	30	KNN	0.322	0.661	0.661
			NB	0.329	0.619	0.642
	PCA	8	KNN	0.300	0.647	0.649
			NB	0.314	0.639	0.648
	SVM	11	KNN	0.333	0.667	0.667
			NB	0.239	0.570	0.601
	PCA and SVM	5	KNN	0.417	0.708	0.708
			NB	0.211	0.557	0.589

9. Conclusion

In this study, a composite feature selection method based on PCA and SVM techniques was proposed and compared to Naïve Bayes and K-nearest neighbor algorithms using various datasets. The results indicate that the proposed feature selection scheme enhances overall performance, outperforming other mentioned approaches in computational cost, accuracy, F_1-score and MCC.

All three feature selection approaches in this study improve the performance of NB and KNN classifiers. However, the proposed composite PCA and SVM method achieves significantly improved performance. Therefore, it can be concluded that compared to using all features, classifier performance is enhanced through feature selection. Additionally, the experimental results indicate that the proposed composite model of PCA and SVM successfully decreases the dimensionality of the data.

References

[1] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[2] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[3] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1, pp. 273–324, 1997.

[4] S. A. Ali Shah, H. M. Shabbir, S. U. Rehman, and M. Waqas, "A comparative study of feature selection approaches: 2016-2020," *International Journal of Scientific & Engineering Research*, vol. 11, no. 2, pp. 469–478, 2020.

[5] J. Han, M. Kamber, and J. Pei, Data Mining, 3rd ed., Morgan Kaufmann/Elsevier, Waltham, MA, 2012.

[6] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines." Machine Learning, vol. 46, no. 1-3, pp. 389-422, 2002. https://doi.org/10.1023/A:1012487302797.

[7] F. Song, Z. Guo, and D. Mei, "Feature selection using principal component analysis," in *2010 International Conference on System Science, Engineering Design and Manufacturing Informatization*, vol. 1, pp. 27–30, 2010.

[8] G. R. Naik, Advances in Principal Component Analysis: Research and Development. Springer, 2019.

[9] H. Zhang, "The optimality of Naïve Bayes," in V. Barr and Z. Markov, Eds., *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*, Miami Beach, FL, USA: AAAI Press, pp. 562–567, 2004.

[10] E. O. Omuya, G. O. Okeyo, and M. W. Kimwele, "Feature selection for classification using principal component analysis and information gain," *Expert Systems with Applications*, vol. 174, p. 114765, 2021.

[11] S. Kashef, H. Nezamabadi-Pour, and B. Nikpour, "Multilabel feature selection: A comprehensive review and guiding experiments," *WIREs Data Mining and Knowledge Discovery*, vol. 8, no. 2, 2018.

[12] S. Solorio-Fernández, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, "A review of unsupervised feature selection methods," *Artificial Intelligence Review*, vol. 53, no. 2, pp. 907–948, 2019.

[13] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," *Chapman and Hall/CRC*, pp. 37–64, 2014 (Copyright © 2015 by Taylor & Francis Group, LLC).

[14] W. Zheng et al., "Multifeature based network revealing the structural abnormalities in autism spectrum disorder," *IEEE Transactions on Affective Computing*, vol. 12, no. 3, pp. 732–742, 2021.

[15] R. Sheikhpour, M. A. Sarram, S. Gharaghani, and M. A. Z. Chahooki, "A survey on semi-supervised

feature selection methods," *Pattern Recognition*, vol. 64, pp. 141–158, 2017.

[16] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed, "Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 5, pp. 971–989, 2016.

[17] T. A. Alhaj et al., "Feature selection using information gain for improved structural-based alert correlation," *PLOS ONE*, vol. 11, no. 11, p. e0166017, 2016.

[18] Z. Zhao, L. Wang, and H. Liu, "Efficient spectral feature selection with minimum redundancy," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 24, no. 1, pp. 673–678, 2010.

[19] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, ACM, 2010.

[20] J. Nobre and R. F. Neves, "Combining principal component analysis, discrete wavelet transform and XGBoost to trade in the financial markets," *Expert Systems with Applications*, vol. 125, pp. 181–194, 2019.

[21] P. Sanguansat, Ed., *Principal Component Analysis*. IntechOpen: Rijeka, 2012.

[22] M. A. Jabri, "High Performance Principal Component Analysis with ParAL," *Neuromorphic LLC*, Oct. 1998.

[23] K. Song, B. Zhang, W. Li, L. Yan, and X. Wang, "Research on parallel principal component analysis based on ternary optical computer," *Optik*, vol. 241, p. 167176, 2021.

[24] N. Funatsu and Y. Kuroki, "Fast parallel processing using GPU in computing L1-PCA bases," in TENCON 2010-2010 IEEE Region 10 Conference, Fukuoka, Japan, Nov. 2010, pp. 2087-2090.

[25] D. A. Ross, J. Lim, R. S. Lin, and M. H. Yang, "Incremental learning for robust visual tracking," *International Journal of Computer Vision*, vol. 77, pp. 125-141, 2008.

[26] N. Halko, P. G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," SIAM Review, vol. 53, no. 2, pp. 217-288, 2011.

[27] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[28] T. Hastie, The Elements of Statistical Learning, Springer Series in Statistics, 2nd ed., Springer, New York, NY, 2017.

[29] A. I. Pratiwi and Adiwijaya, "On the feature selection and classification based on information gain for document sentiment analysis," *Applied Computational Intelligence and Soft Computing*, vol. 2018, pp. 1–5, 2018.

[30] X. Xu, H. Gu, Y. Wang, J. Wang, and P. Qin, "Autoencoder based feature selection method for classification of anticancer drug response," *Frontiers in Genetics*, vol. 10, 2019.

[31] A. Zien, N. Krämer, S. Sonnenburg, and G. Rätsch, "The feature importance ranking measure," in *Advances in Neural Information Processing Systems*, vol. 21, pp. 694–709, Springer Berlin Heidelberg, 2009.

[32] I. Kamkar, S. K. Gupta, D. Phung, and S. Venkatesh, "Exploiting feature relationships towards stable feature selection," in 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), vol. 37, IEEE, pp. 1–10, 2015.

[33] W. W. B. Goh and L. Wong, "Evaluating feature-selection stability in next-generation proteomics," *Journal of Bioinformatics and Computational Biology*, vol. 14, no. 05, p. 1650029, 2016.

[34] S. Raghavendra and M. Indiramma, "Hybrid data mining model for the classification and prediction of medical datasets," International Journal of Knowledge Engineering and Soft Data Paradigms, vol. 5, no. 3/4, p. 262, 2016.

[35] B. Xin, L. Hu, Y. Wang, and W. Gao, "Stable feature selection from brain sMRI," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 29, no. 1, 2015.

[36] A. Mehrabinezhad, M. Teshnelab, and A. Sharifi, "Autoencoder-PCA-based Online Supervised Feature Extraction-Selection Approach," *Journal of AI and Data Mining*, vol. 11, no. 4, pp. 525-534, 2023. doi: 10.22044/jadm.2023.12436.2390.

[37] R. Adhao and V. Pachghare, "Feature selection using principal component analysis and genetic algorithm," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 23, no. 2, pp. 595–602, 2020.

[38] I. T. Jolliffe, Principal Component Analysis for Special Types of Data, Springer New York, New York, NY, pp. 338–372, 2002.

[39] L. Peterson, "K-nearest neighbor," Scholarpedia, vol. 4, no. 2, p. 1883, 2009.

[40] M. Lichman. (n.d.). UCI Machine Learning Repository. [Online]. Available: http://archive.ics.uci.edu/ml.

[41] H. Nosrati Nahook, and M. Eftekhari, "A new method for feature selection based on fuzzy logic." *Computational Intelligence in Electrical Engineering*, vol. 4, no. 1, pp. 71-84, 2013.

[42] E. Namsrai, T. Munkhdalai, M. Li, J. H. Shin, O. E. Namsrai, and K. H. Ryu, "A feature selection-based ensemble method for arrhythmia classification," *Journal of Information Processing Systems*, vol. 9, no. 1, pp. 31-40, 2013.

[43] R. Jain, P. R. Betrabet, B. A. Rao, and N. S. Reddy, "Classification of cardiac arrhythmia using improved feature selection methods and ensemble classifiers," in *Journal of Physics: Conference Series*, vol. 2161, no. 1, p. 012003, 2022, IOP Publishing.

[44] M. Tunç and G. B. Cangöz, "Classification of the cardiac arrhythmia using combined feature selection algorithms," *Turkish Journal of Science and Technology*, vol. 19, no. 1, pp. 147-159, 2024.

[45] Liang, C. F. Tsai, and H. T. Wu, "The effect of feature selection on financial distress prediction," Knowledge-Based Systems, vol. 73, pp. 289-297, 2015.

[46] Y. Zhou, M. Shamsu Uddin, T. Habib, G. Chi, and K. Yuan, "Feature selection in credit risk modeling: an international evidence," Economic Research-Ekonomska Istraživanja, vol. 34, no. 1, pp. 3064-3091, 2021.

[47] A. Rouhi and H. Nezamabadi-Pour, "A hybrid-based feature selection method for high-dimensional data using ensemble methods," *Iranian Journal of Electrical and Computer Engineering*, vol. 60, no. 4, p. 283, 2018.

[48] M. A. Rahman and R. C. Muniyandi, "Feature selection from colon cancer dataset for cancer classification using artificial neural network," *International Journal of Advanced Science, Engineering and Information Technology*, vol. 8, no. 4-2, pp. 1387-1393, 2018.

[49] O. O. Petinrin, F. Saeed, N. Salim, M. Toseef, Z. Liu, and I. O. Muyide, "Dimension reduction and classifier-based feature selection for oversampled gene expression data and cancer classification," Processes, vol. 11, no. 7, p. 1940, 2023.

[50] S. DeepaLakshmi and T. Velmurugan, "Benchmarking attribute selection techniques for microarray data," *ARPN Journal of Engineering and Applied Sciences*, vol. 13, no. 11, pp. 3740-3748, 2018.

[51] C. De Stefano, F. Fontanella, and A. Scotto di Freca, "Feature selection in high dimensional data by a filter-based genetic algorithm," in Applications of Evolutionary Computation: 20th European Conference, EvoApplications 2017, Amsterdam, The Netherlands, Apr. 19-21, 2017, Proceedings, Part I 20, pp. 506-521, Springer International Publishing.

[52] A. Mehrabinezhad, M. Teshnelab, and A. Sharifi, "Autoencoder-PCA-based online supervised feature extraction-selection approach," *Journal of AI and Data Mining*, vol. 11, no. 4, pp. 525-534, 2023.

[53] K. Yang, Z. Cai, J. Li, and G. Lin, "A stable gene selection in microarray data analysis," *BMC Bioinformatics*, vol. 7, pp. 1-16, 2006.

[54] F. H. Yağın, Z. Küçükakçalı, İ. B. Çiçek, and H. G. Bağ, "The effects of variable selection and dimension reduction methods on the classification model in the small round blue cell tumor dataset," Middle Black Sea Journal of Health Science, vol. 7, no. 3, pp. 390-396, 2021.

[55] M. Hamim, I. El Mouden, M. Ouzir, H. Moutachaouik, and M. Hain, "A novel dimensionality reduction approach to improve microarray data classification," *IIUM Engineering Journal*, vol. 22, no. 1, pp. 1-22, 2021.

[56] S. Karimi and M. Farrokhnia, "Leukemia and small round blue-cell tumor cancer detection using microarray gene expression data set: Combining data dimension reduction and variable selection technique," Chemometrics and Intelligent Laboratory Systems, vol. 139, pp. 6-14, 2014.

[57] L. Y. Chuang, C. H. Ke, and C. H. Yang, "A hybrid both filter and wrapper feature selection method for microarray classification," arXiv preprint arXiv:1612.08669, 2016.

[58] S. J. Susmi, H. K. Nehemiah, and A. Kannan, "Hybrid dimension reduction techniques with genetic algorithm and neural network for classifying leukemia gene expression data," *Indian Journal of Science and Technology*, vol. 9, no. 1, pp. 1-8, 2016.

[59] T. K. Abuya, "Lung cancer prediction from Elvira biomedical dataset using ensemble classifier with principal component analysis," *Journal of Data Analysis and Information Processing*, vol. 11, no. 2, pp. 175-199, 2023.

[60] D. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness markedness & correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

یک روش ترکیبی برای انتخاب ویژگی‌ها با بهره‌گیری از تحلیل مؤلفه‌های اصلی و ماشین‌های بردار پشتیبان

سیدمحمد حسینی*، مجید ابتیاع و مهنا دهگردی

گروه پژوهشی هوش مصنوعی گهر، دانشگاه آیت الله بروجردی(ره)، بروجرد، ایران.

ارسال ۲۰۲۴/۱۲/۳۱ بازنگری ۲۰۲۵/۰۱/۳۰؛ پذیرش ۲۰۲۵/۰۳/۰۱

چکیده:

با توجه به افزایش داده‌های چندبعدی و محدودیت‌های پردازشی، اهمیت یافتن ویژگی‌های کلیدی داده بیش از پیش مشهود شده است. ویژگی‌های غیرضروری می‌توانند هم باعث کاهش دقت پیش‌بینی شوند و هم محاسبات اضافی ایجاد کنند که کارایی مدل‌های طبقه‌بندی را پایین می‌آورد. از این رو، انتخاب مناسب ویژگی‌ها به درک بهتر داده‌ها، کاهش زمان پردازش و افزایش دقت پیش‌بینی کمک می‌کند. در این پژوهش، روشی ترکیبی ارائه شده که از قدرت ماشین بردار پشتیبان و تحلیل مؤلفه اصلی بهره می‌برد تا ویژگی‌های مهم را استخراج کند. این روش سپس بر روی الگوریتم‌های نزدیک‌ترین همسایه و بیز ساده پیاده‌سازی شده است. برای ارزیابی عملکرد، از سه مجموعه داده از مخزن UCI و یک مجموعه داده جمع‌آوری شده از کتابخانه مرکزی دانشگاه آیت‌الله بروجردی استفاده شده است؛ این مجموعه شامل ۱۹۱۰ نمونه با ۳۰ ویژگی نظیر جنسیت، وضعیت بومی، ترم ورود، کد دانشکده، معدل کل و تعداد کتاب‌های امانت‌شده می‌باشد. نتایج نشان می‌دهد که با انتخاب تنها ۵ ویژگی اصلی، به دقتی معادل ۷۰٪ دست یافته‌ایم. این روش انتخاب ویژگی نه تنها مجموعه مناسبی از ویژگی‌ها را شناسایی می‌کند، بلکه عملکرد طبقه‌بندی را از نظر معیارهایی مانند دقت، امتیاز F و ضریب همبستگی متیوز به طور چشمگیری بهبود می‌بخشد.

کلمات کلیدی: کا نزدیکترین همسایه، انتخاب ویژگی، ماشین بردار پشتیبان، تحلیل مولفه اصلی، بیز ساده.