JOURNAL OF STATISTICAL MODELING & ANALYTICS (JOSMA)
(ISSN: 2180-3102)

UNIVERSITI MALAYA

Boosting Cancer Dataset Performance with Mutual Information-Based Feature Prioritization

Fung Yuen Chin[1*] and Yong Kheng Goh[2]

[1] *Department of Physical and Mathematical Science, Universiti Tunku Abdul Rahman, Kampar, Perak, Malaysia*

[2] *Department of Mathematical and Actuarial Sciences, Universiti Tunku Abdul Rahman, Kajang, Selangor, Malaysia*

**Corresponding author: chinfy@utar.edu.my*

RESEARCH ARTICLE

Received 25 October 2023
Accepted 19 January 2024
Published 30 May 2024

Abstract

In the field of statistical modelling, mutual information is a crucial and common concept, suitable for tasks such as selecting the most important features or classifying data into different categories. Feature selection addresses the challenge of high-dimensional data in building effective predictive models by identifying relevant attributes while mitigating the curse of dimensionality. Previous studies have benchmarked the effectiveness of statistical models against established results. To enhance this, a new benchmark method is proposed, exploiting ranking features via mutual information scores. Mutual information score is used to understand the relationship between underlying data and variables. The performance of the classification depends on its information content, which directly affects the performance of the statistical model. The technique simultaneously determines the optimal feature quantity to guide the feature selection process. The validation of these selected features is conducted through Z-score graphs. Experimental results show that this method can identify feature subsets better than using the full features by using the Support Vector Machine classifier. These advance promises to improve cancer analysis, enabling more sophisticated diagnostic and prognostic methods.

Keywords: Classification, Dimension reduction, Feature selection

1. Introduction

In the world of cancer data analysis, leveraging predictive models to generate important insights is critical. However, the complexity of high-dimensional data poses challenges due to the curse of dimensionality (Shi et al., 2022). Effective feature selection becomes an important strategy to deal with this complexity, aiming to identify relevant attributes while mitigating dimensionality-related issues (Adeboye et al., 2023).

Previous research has established practices for evaluating the effectiveness of statistical models against established benchmarks (Khaire & Dhanalakshmi, 2022; Khairuddin et al., 2023). However, there is still room for improvement, so a new reference benchmark method is proposed. This approach provides a novel dimension to this work by exploiting the prioritization of features via mutual information scores.

The amount of information in the data set directly affects the performance of the classification model (Jiang et al., 2023). Therefore, determining the reference feature quantity becomes crucial and becomes a guiding principle in the complex feature selection process. To verify the effectiveness of the selected features, Z-score plots were used to assess their importance. Empirical results highlight the effectiveness of this approach, revealing its ability to identify a subset of features that exceeds the performance achieved by including all features. This advancement will revolutionize the landscape of cancer data analysis, driving the development of more sophisticated diagnostic and prognostic methods.

The increase in the dimensionality of cancer datasets creates challenges in building effective statistical models (Chlioui et al., 2021). The proposed method solves this problem by facilitating the selection of relevant attributes, thus mitigating the curse of dimensionality. Although previous research has established benchmarking practices for statistical models, there is still room for improvement. The motivation for this study is to introduce a new “benchmark” approach that surpasses existing methods and sets a higher standard for model performance evaluation.

Recognizing that the quality of a dataset is closely related to its information content, this study attempts to exploit this connection. By strategically selecting features based on mutual information scores, this approach aims to improve information quality and subsequently enhance statistical model performance. Advances in this approach fill the need for more accurate diagnostic and prognostic methods in cancer analysis. By identifying subsets of features that outperform the full attributes, this research provides a practical tool for improving cancer analysis techniques and advancing medical decision-making.

The main goal of this study is to develop a new statistical model using the Support Vector Machine classifier for feature selection in high-dimensional cancer datasets. The proposed method ranks the features using mutual information scores, to identify the most relevant features to reduce the dimension of the data. By determining the number of reference features required for statistical modelling, the proposed method, including the identification and ranking of relevant features, enhances the effectiveness of cancer analysis models, surpassing traditional baselines and yielding overall improved performance. This feature selection method ensures more efficient and effective results in the statistical modelling process. (Ahuja & Sharma, 2021; Jimoh et al., 2021; Okwonu et al., 2023).

Mutual information has been widely used for feature selection in building statistical models. Battiti (1994) combined mutual information with a greedy selection method and proposed the mutual information-based feature selection (MIFS) algorithm. This method shows that mutual information can effectively measure relationships between features, including linear and nonlinear relationships. Peng et al. (2005) introduced a dimension reduction technique that enhances the relationship between features and labels while minimizing the redundancy between features. This proposed method called “minimum redundancy maximum correlation” (mRMR), focuses on selecting features that carry information about the class labels and do not reveal redundancy with each other.

MIFS does not take into account the interdependencies between features, which may lead to suboptimal feature subsets. mRMR often has difficulty handling high-dimensional data, and due to its greedy search approach, it may not effectively capture the most informative features. It doesn't scale well to large feature sets.

The concept of Joint Mutual Information (JMI) revolves around evaluating the collective mutual information shared by selected features and the categories they belong to (Yang and Moody, 1999). Bennassar et al. (2015) introduced the joint mutual information maximization (JMIM) and normalized joint mutual information (NJMIM) techniques. JMIM identifies features by gradually maximizing the mutual information between features and classes by considering previously selected features.

NJMIM enhances JMIM by normalizing mutual information scores and facilitating fair comparisons between different datasets. These methods strategically exploit the mutual information shared between features and class labels to help extract relevant and unique features, thereby improving data analysis and classification performance. JMI can be computationally expensive when processing large datasets, making it impractical for some real-world cancer datasets. It also doesn't fully account for functional redundancy.

JMIM may be sensitive to the initial feature subset and may not always converge to the globally optimal feature set. It may miss important features or select redundant features. While NJMIM solves some of the problems of JMIM, it still has similar limitations in terms of initialization sensitivity and potential convergence issues. Its performance depends heavily on correct parameter tuning.

Liping (2015) introduced a method called conditional dynamic mutual information (CDMI) to address the limitations of traditional mutual information-based feature selection algorithms. CDMI overcomes the inaccuracy caused by fixed evaluation by dynamically evaluating mutual information throughout the selection process. It improves measurement accuracy by accurately assessing feature importance and information content by excluding features from further consideration after feature selection.

Besides mutual information-based techniques applied in building statistical models, in clinical medicine, multidimensional time series data are often used to analyze disease progression through data mining techniques such as classification and prediction. However, high data dimensionality may lead to inaccurate probability density distributions and increase computational complexity. This, combined with redundant and irrelevant features, hinders classification performance. Fang et al. (2015) proposed a method that combines Kozachenko-Leonenko entropy estimation for feature extraction with a selection algorithm based on class separability to address this issue.

Proteomic data analysis using mass spectrometry is an effective method for early disease diagnosis, especially in tumour detection. However, this method is challenged by limited samples, high dimensionality and noise interference. To address this problem, Qin et al. (2017) introduced a feature selection method combining support vector machine (SVM) and shape analysis. Unlike traditional techniques, their method considers both feature interactions and feature-class-label relationships, thereby improving classification accuracy.

Sluga and Uros (2017) introduced a feature selection technique based on quadratic mutual information, which depends on Cauchy-Schwarz divergence and Renyi entropy. This method uses a Gaussian kernel function to estimate direct quadratic mutual information and is good at capturing second-order nonlinear relationships. Unique advantages include seamless analysis of discrete and continuous data without the need for discretization and parameter-free design. Comparative evaluation with MIFS, MRMR and JMI highlights its efficiency and effectiveness in the fields of classification and regression.

Multi-label learning is common in fields such as information retrieval and bioinformatics, aiming to cope with noisy, redundant and high-dimensional data sets. This situation is exacerbated by the curse of dimensionality. Feature selection is an effective data preprocessing technique that has attracted attention for its role in improving computational efficiency, prediction accuracy, and data understanding. Many information theory-based feature selection methods are suitable for multi-label classification. However, many methods rely on heuristics or adaptations of single-label methods.

To fill up this gap, Sun et al. (2019) proposed a method based on mutual information to optimize fast solutions through constrained convex optimization. It contains label information, takes label correlations into account, and demonstrates the functionality and efficiency of different multi-label datasets. Due to its computational efficiency and result interpretability, feature selection plays a key role in dimensionality reduction in applications such as data mining and machine learning.

Wang et al. (2019) identified the limitation of the existing methods only considering individual relationships between candidate features and class vectors. They introduced the concept of equivalent partitioning and adopted the mutual information gain maximization (MIGM) criterion to evaluate candidate features.

Multi-label learning often involves high-dimensional data, leading to the "curse of dimensionality". To alleviate this situation, effective preprocessing through feature selection is crucial. Xiong et al. (2021) introduced a method to integrate label distribution learning into multi-label feature selection. This explains the difference in label importance. From a computing perspective, fuzzy similarity-based label enhancement algorithms convert logical labels into distributions. The corresponding feature selection algorithm uses fuzzy mutual information to measure feature importance.

In the field of multi-label data, the urgent focus on dimensionality has triggered great interest in feature selection. Existing information theory-based methods usually focus on feature correlations, which are determined solely by the information contribution of the features to the label set. However, they ignore two fundamental aspects: the rate of change of undetermined information and determined information. To address this issue, Hu et al. (2022) introduced a new feature correlation term, weight-based correlation (RW). The term covers both types of rates of change and takes into account their positive or negative impact on the assessment of relevance. Based on this, a multi-label feature selection method - correlation-based weighted feature selection (RWFS) is proposed.

CDMI can have difficulty handling noisy or incomplete data because it relies on accurate estimation of conditional dependencies, which can be challenging in such cases. MIGM does not consider feature redundancy and may select similar features, which may not significantly contribute to the statistical model performance. RW methods may not effectively capture the complex relationships between features in cancer data. It may not account for nonlinear interactions or hidden patterns.

Chronic kidney disease (CKD) is a life-threatening disease with a global impact. Early prediction and accurate classification are critical for effective management. Savitha and Rajiv (2023) introduced correlation-based weighted composite features (CWCF) and feature significance-based weighted composite features (FSWCF) algorithms to generate composite features from CKD indicators.

CWCF may not perform well when there are non-linear relationships or complex dependencies between features. It relies heavily on linear dependencies. FSWCF may be sensitive to the choice of significance measure and may perform poorly when the significance measure fails to capture the true importance of features in cancer data.

In recent decades, researchers have commonly used mutual information for feature selection to process high-dimensional data. However, previous research has mainly focused on identifying the most relevant features for building statistical models. Notably, there is a gap in the literature in determining the optimal number of features required to build statistical models and establish reference benchmarks.

In order to solve the limitations of existing feature selection methods and lay a solid foundation for the statistical model, this study has made certain improvements. The proposed method involves feature ranking to select the most relevant features to efficiently reduce the high-dimensional data. These relevant features are then used to establish the reference performance of the statistical model. The results show that the performance of the statistical model is better than the statistical model using the full features.

2. Methods

2.1 Mutual Information

Mutual information is used in information theory and is also important in statistics, especially in the fields of probability theory, machine learning, and data analysis (Zhu & Zeng, 2015). Its basic function is to evaluate the correlation between any two variables. It quantifies the extent to which understanding the value of one variable can reduce the inherent uncertainty associated with another variable. In cancer data analysis, especially in the field of feature selection, mutual information plays an important role.

In the fields of machine learning and statistics, mutual information is an important tool for feature selection in statistical models. Specifically, when there is a large amount of mutual information between a feature and the target variable, it indicates that the feature is important and informative enough to be included in the model. Mutual information plays an important role in dimensionality reduction, such as feature extraction and manifold learning. In this case, it helps to reduce the dimensions or select the features that are most relevant to the data.

The mutual information between two random variables X and Y is defined as:

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \quad (1)$$

where $I(X;\ Y)$ is the mutual information between X and Y, $p(x,y)$ is the joint probability distribution of X and Y and $p(x)$ and $p(y)$ are the marginal probability distributions of X and Y respectively.

2.2 The Reference Benchmark based on Prioritized Features

The proposed method defines the reference benchmark using feature ranking based on mutual information. Mutual information can evaluate the relationships between features, regardless of their linearity. When the mutual information is the highest, indicates there is a strong connection between the two features. To establish the reference benchmark, the mutual information of all features related to the class label is calculated, and then the features are ranked accordingly.

The most critical features are those that are most closely related to the class labels and represent them effectively. By incorporating more of these key features, a compact subset is formed that can robustly represent class labels. Through classifier performance evaluation, adding more features will improve the classification performance. This insight helps identify specific points in prioritized features where classification performance peaks, indicating the ideal feature quantity to achieve optimal classification results.

Given that microarray data often contain high-dimensional uncorrelated features and noise, it becomes critical to establish a more effective benchmark by focusing on relevant features, which are determined through mutual information and their correlation with class labels. The emphasis on relevant features is intended to go beyond classification methods that use full feature sets. This strategy can determine the best benchmark through feature ranking.

Additionally, it helps to build the necessary feature quantity (denoted as "k") for optimal performance. The feature quantity k is the key factor for classification. Previous research has mainly focused on selecting features but lacks guidance on determining the k feature benchmark to enable fair comparison of feature selection methods and enhance their evaluation.

2.3 Algorithm on Prioritized Features

Mutual information is an important measure to evaluate the similarity between attributes and labels. A higher score of mutual information indicates a stronger correlation between these elements. The scores of all features associated with class labels in the microarray dataset are calculated by using equation (1).

At first, the experimental data is normalised into the range [-1, 1]. Each feature and class label was then divided into three equal bins to classify the data set as low expression, normal expression, or high expression.

This binning method has the advantage of seamlessly handling missing data since the mutual information calculation relies only on the frequency counts of the remapped data, rather than the original values. It is important to note that mutual information does not take into account the specific relationship between attributes and labels, which makes it suitable for both linear and nonlinear scenarios. Features are then ranked based on their mutual information scores, and a graph depicting feature accuracy versus cumulative ranking is drawn.

The point on the graph with the highest accuracy determines the reference benchmark, indicating the feature quantity to use in the predictive model. Subsequently, evaluation tools such as confusion matrices and Receiver Operating Characteristic (ROC) curves are used to evaluate the performance of the selected subset of features. To prove that features are not randomly selected, a Z-score is applied to a subset of compact features. Figure 1 shows the flowchart of searching the reference benchmark and the feature quantity.

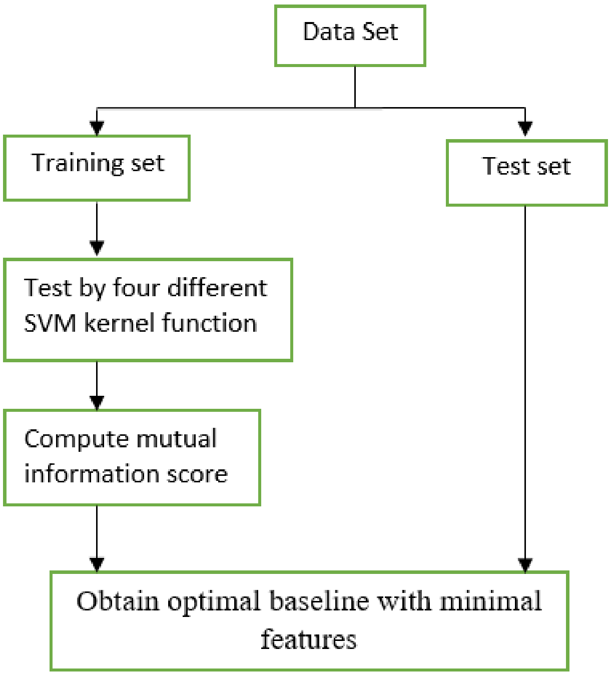
Data Set
Training set
Test set
Test by four different SVM kernel function
Compute mutual information score
Obtain optimal baseline with minimal features

Figure 1: Flowchart of searching the reference benchmark and the feature quantity.

Features selected by the proposed method will be used in a range of classifiers, including support vector machines (SVM), *K*-nearest neighbours (KNN) and decision trees (DT). This utilization will facilitate the creation of multiple statistical models, allowing for a comparative assessment of their predictive capabilities. In this study, the performance of statistical models was evaluated by applying different feature selection techniques, including regression and mRMR. The feature quantity selected by the proposed method, regression and mRMR will be the same across these evaluations. Subsequently, these results are compared with statistical models built using support vector machines (SVM), *K*-nearest neighbours (KNN) and decision trees (DT).

2.3.1 Support Vector Machine

Support vector machine (SVM) is a machine-learning method related to statistical modelling (Pisner & Schnyer, 2020). It operates as a supervised learning algorithm and is typically applied to tasks involving classification and regression. The core concept of SVM is to discover an optimal hyperplane that effectively partitions the data into different classes while maximizing the separation between these classes.

Given the *n* data points in the training dataset as:

$$\{(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4) \dots \dots (x_n, y_n), \} \tag{2}$$

where y_n is in the range of -1 to 1, is used to indicate which classes should x_n belong to. Each x_n is a *p*-dimensional real vector. Next, *n* represents the number of samples. The equation of the hyperplane is defined by:

$$w^T x + b = 0 \tag{3}$$

where *b* is scalar and *w* is a *p*-dimensional vector. The separating hyperplane is perpendicular to the vector *w*. The parameter *b* is added to increase the margin.

The hyperplane tends to pass through the origin when b is absent. This causes the restriction of the solution. The parallel hyperplanes, H_1, and H_2 can be defined as:

$$H_1 = w^T x + b = 1 \tag{4}$$

$$H_2 = w^T x + b = -1. \tag{5}$$

SVM problem can be formulated as:

$$w.x_i - b \geq 1 \quad or \quad w.x_i - b \leq -1. \tag{6}$$

By combining these two formulae, it can be written as:

$$y_i(w.x_i - b) \geq 1, 1 \leq i \leq n. \tag{7}$$

The optimal hyperplane separating the data for which $|w|$ should be minimized to maximize the separability:

$$Min\ \varphi(w) = \frac{1}{2}\|w\|^2 \tag{8}$$

subject to the constraints:

$$y_i(w^T.x_i + b) \geq 1, i = 1,2, \ldots..n \tag{9}$$

where $\varphi(w)$ is essentially half of the squared Euclidean norm of the weight vector.

This optimization problem can be solved by using Lagrange’s function:

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i(y_i(w^T.x_i + b) - 1) \tag{10}$$

where α_i represents the Lagrange multiplier. The saddle points must be found to minimize the Lagrange equation in (10) for *w* and b and have to be maximized for non-negative α_i. The saddle point can be obtained by partial differentiation:

$$\frac{\partial L}{\partial w_0} = 0, w_0 = \sum_{i=1}^{n} \alpha_i\, y_i x_i \tag{11}$$

$$\frac{\partial L}{\partial b_0} = 0, \sum_{i=1}^{n} \alpha_i\, y_i = 0. \tag{12}$$

Substitute the equation (11) and equation (12) into equation (10). Now changing the primal form to dual form which only consists $\boldsymbol{\alpha}$ as the parameters.

$$L(\alpha) = \sum \alpha_i - \sum_{i=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \tag{13}$$

subject to the constraints:

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \qquad \alpha_i \geq 0, \qquad i = 1,2,3 \ldots n. \tag{14}$$

With this, an optimal hyperplane can be found and used for separating different classes in a dataset.

2.3.2 *K*-nearest neighbours

K-Nearest Neighbors (KNN) is an effective machine-learning algorithm for classification and regression tasks (Zhang, 2016). It follows the principle, where the prediction of a new data point relies on the characteristics of its nearest neighbors in the training data set. KNN determines the class of a new data point by evaluating the class labels of its *k* nearest neighbours. The class with the most neighbours becomes the predicted class for the new data point. In a regression task, KNN computes the average of the target variable values of the *k* nearest neighbours and assigns that average as the prediction for a new data point.

The value of *k* determines the number of neighbours considered when making predictions. Higher *k* improves model stability but may smooth out fine-grained patterns in the data. In contrast, lower *k* can capture local variations but may be sensitive to noise. In order to identify the closest data point to a specific query point, the distance between the query point and all other data points will be calculated. These distance measures play a crucial role in establishing decision boundaries, which in turn divide query points into different regions. The distance functions in KNN are given as:

$$Euclidean\ distance = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2} \tag{15}$$

where $x_i$and y_i are the data points.

2.3.3 Decision trees

A decision tree is a tree-like model used for classification and regression tasks in machine learning and statistics (Somvanshi et al., 2016). Decisions in a decision tree are selected by a series of conditions applied to the input data. Each node in the tree represents a decision or evaluation related to a specific characteristic or attribute.

The decision-making process can be evaluated using a series of "if-else" statements. At each internal node, the condition is evaluated, and depending on whether it is true or not, the tree continues to the next relevant node, finally reaching the leaf node where the final decision or prediction is made. The configuration of a decision tree and the criteria for specifying each node depends on the specific algorithm used, such as Classification and Regression Trees (CART). These algorithms employ various statistical measures such as Gini impurity or information gain to determine the most informative conditions for splitting the data at each node and constructing a valid decision tree.

2.4 Z-Score

Given the relatively small sample sizes in microarray data analysis, feature selection may be affected by chance. In critical applications such as cancer diagnosis and treatment, algorithms that reliably select relevant features are needed. This study uses *Z*-score analysis to evaluate the robustness of the algorithm in feature selection (Li et al., 2001; Jirapech-Umpai & Aitken, 2005). *Z*-scores quantify the importance of selected features. High *Z*-score values indicate non-random feature selection.

Furthermore, an algorithm is considered more robust if it selects features with higher *Z*-score values among the selected features. The *Z*-score of the feature is defined as

$$Z = \frac{f_i - E(f_i)}{\sigma} \tag{16}$$

where f_i was the frequency of the feature selected, and σ was the standard deviation of f_i. Let *N* be the total feature quantity and $E(f_i)$ was the expected number of times feature *i* was selected and $f_{\bar{C}}$ be the average number of selected features, then the probability of feature *i* was selected as $\frac{f_{\bar{C}}}{N}$ and

$$E(f_i) = P(f_i)K \tag{17}$$

$$\sigma = \sqrt{P(f_i)\big(1 - P(f_i)\big)K} \tag{18}$$

where *K* is the number of replicates.

2.5 Assessment method

The accuracy of the prioritized features measured using the SVM was defined as follows:

$$accuracy = 1 - \frac{false\ negative + false\ positive}{false\ negative + false\ positive + true\ positive + true\ negative}. \tag{19}$$

The value of the accuracy ranges from 0-100%. The receiver operating characteristic curve (ROC) illustrates the trade-off between a true positive rate and a false positive rate at different thresholds. It provides insights into the model's ability to differentiate between positive and negative cases at different decision boundaries. The area under curve (AUC) measures the overall ability of a model to distinguish between different classes or categories. Higher AUC values indicate better discrimination and model performance.

3. Results and Discussion

3.1 Dataset

This study used publicly available microarray data from two reputable sources: the UCI Machine Learning Repository and the National Center for Biotechnology Information (NCBI). These datasets are shown in Table 1 and consist of two binary classification datasets and one multi-class classification dataset. These datasets were chosen because of their high dimensionality, small sample size, and widespread use in published studies, which makes them well-suited for applying our proposed algorithm to establish a reference benchmark.

Table 1. Summary of the downloaded dataset.

Dataset	No. of attribute	No. of sample	Type of classification
Prostate cancer	2135	102	Binary
Lung cancer	1626	181	Binary
Skin cancer	22215	15	3 classes

The process begins by randomly dividing the full data set into two subsets: training and test sets, with a ratio of 7:3. This ratio is a common choice and widely used in machine learning. The training set includes the full features, represented by $X = \{x_1, x_2, \dots, x_N\}$ and the N-dimensional class C. These training samples serve as input to the algorithm. The next step involves normalizing the training set so that its values fall in the range [-1, 1]. To achieve this, each feature is partitioned into three equal bins.

Additionally, labels are ordered based on the number of labels they represent. The algorithm computes the joint probability mass function for each feature and label. To determine the best benchmark for all datasets, the feature ranking method outlined in Section 3 is adopted. The whole process is repeated 50 times and the average score of mutual information of each feature is calculated. Repeating this process 50 times can be used as a strategic approach to enhance the robustness and reliability of performance or feature importance estimates. This repetition helps mitigate the effects of random variation, thereby improving the stability of machine learning.

3.2 Reference Benchmark

To establish a reference benchmark for the data, the algorithm outlined in Section 3 will be employed. Figures 2-4 show the performance of prioritized features in prostate cancer, lung cancer, and skin cancer datasets. From the outcome of the algorithm, the highest accuracy will serve as the reference benchmark for the dataset. This reference benchmark aligns with the feature quantity needed to attain it. Importantly, the optimality is often not achieved using the full feature set as a benchmark.

Full-featured benchmarks will always produce lower accuracy levels compared to the proposed reference benchmarks. Including all features simultaneously introduces redundancy and noise into the model, causing unpredictable fluctuations in the accuracy of the statistical model. As a result, the overall accuracy of the statistical model decreases. These findings provide clear insights into how to determine the reference benchmarks and the feature quantity required to reach peak performance levels.

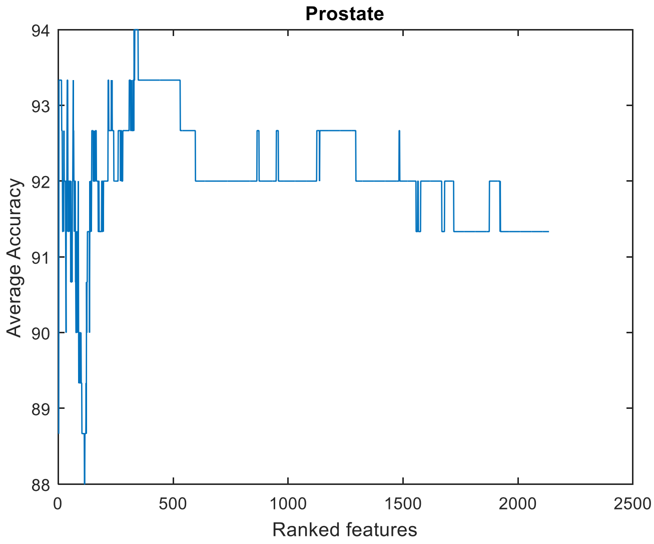

Figure. 2. The performance of the prioritized features for prostate cancer data set.

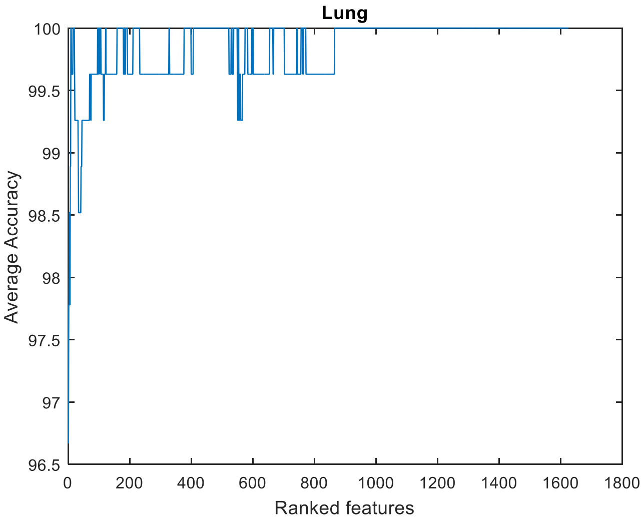

Figure 3. The performance of the prioritized features for lung cancer data set.

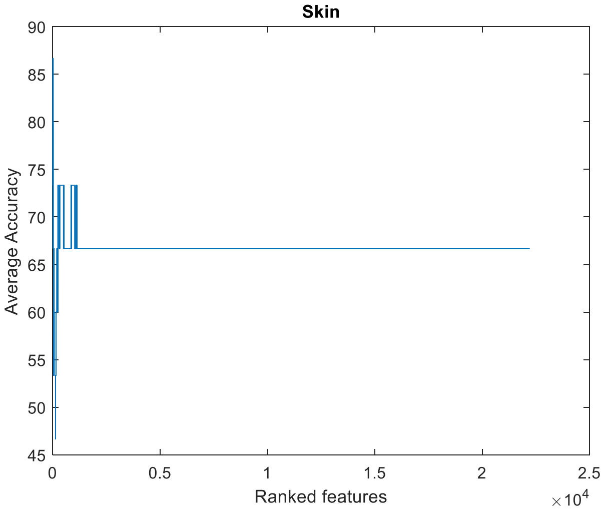

Figure 4. The performance of the prioritized features for the skin cancer dataset.

Table 2 shows the comparison between the benchmark accuracy achieved when using the full feature set and the reference benchmark obtained by the proposed method explained in Section 2.3. It also indicates the feature quantity required to reach this reference benchmark. After removing redundant and noisy features, the prediction model performance is improved. This allows the model to focus on relevant information, reduce overfitting, improve generalization capabilities, and enhance overall stability. Therefore, this process helps in making consistent and reliable predictions across different data sets. The results consistently show that the proposed algorithm consistently outperforms the benchmark accuracy achieved using all features.

Table 2. The benchmark uses full features and the reference benchmark with the feature quantity.

Dataset	Benchmark	Full features	Reference benchmark	No. of features	Dimension to reduce
Prostate cancer	91.33%	2135	94%	330	84.54%
Lung cancer	100%	1626	100%	9	99.45%
Skin cancer	66.67%	22215	86.67%	2	99.99%

Furthermore, it significantly reduces the feature quantity required to achieve a reference benchmark compared to full features. These selected features are obtained by the proposed algorithm and ranked according to their relevance, ensuring that they contain the most relevant information. This highlights a significant reduction in data dimensionality, equivalent to a 90% reduction in the original size. Therefore, the reduction in the feature quantity obtained by the proposed algorithm is expected to enhance the predictive power of the model as it reduces the complexity associated with a large number of dependent variables.

3.3 Performance of the Statistical Model

The feature quantity obtained by the algorithm is crucial for feature selection. There is a need to ensure that any feature selection method does not exceed a predetermined feature quantity while maintaining the same level of accuracy. Therefore, a new recommendation is introduced specifying the maximum feature quantity allowed during feature selection. This approach differs from previous research practices, which used all features to establish a benchmark, resulting in uncertainty about the ideal feature quantity for a predictive model. This new approach has an advantage in solving this problem.

Table 3 provides the results of 10-fold cross-validation with the average accuracy (Acc) when using different classifiers such as support vector machine (SVM), K-nearest neighbour (KNN), and decision tree (DT). These results show that the proposed method performs better than mRMR and regression methods when selecting the same feature quantity. In this study, mRMR is only applied to binary datasets.

Table 3: The 10-fold cross-validation and average accuracy using SVM, KNN and DT.

Dataset	Method	SVM-CV	SVM-Acc	KNN-CV	KNN-Acc	DT-CV	DT-Acc
Prostate cancer	Proposed	93.24	94	90.99	89.03	80.85	83.87
	Regression	91.67	92.67	88.61	86.67	80.28	82
	mRMR	91.67	86.45	89.58	84.52	84.23	78.06
Lung cancer	Proposed	99.21	100	98.43	100	95.59	97.78
	Regression	97.64	98.15	98.43	98.52	94.80	96.67
	mRMR	98.74	97.78	98.11	98.15	95.59	94.07
Skin cancer	Proposed	83.33	100	86.67	85	75.00	66.67
	Regression	60	86.67	70	75	56.67	6.67

3.4 Evaluation of the Statistical Model

Table 4 provides the confusion matrices and AUC for the three datasets. The AUC for the three datasets are above 0.9.

Table 4. The output of the confusion matrix and the ROC curve.

Dataset	TP	TN	FP	FN	Prediction speed	Training time	AUC
Prostate cancer	33	31	6	2	34 obs/sec	13.087 sec	0.94
Lung cancer	105	21	1	0	130 obs/sec	9.5591 sec	1
Skin	5	2	3	2	0.04 obs/sec	962.58 sec	0.97

3.5 Z Score of the Selected Features

Various feature selection methods face the same challenge: determining whether the selected features are consistent or contribute to the predictive model. To address this issue, the Z-score is used to evaluate the features selected by the proposed algorithm. Due to the relatively small sample size of our microarray data, Z-scores can be used as a reliable means of evaluating selected characteristics, ensuring that the selected features are not the result of chance.

The Z-score is a metric that assigns higher values to features that are truly important to the predictive model (as opposed to randomly selected features). When analysing the Z-scores associated with features in the dataset, it helps to identify and confirm the truly important features. Features that are truly representative and necessary for the predictive model will produce high Z-score values. Therefore, the Z-score indicates that the selected features are relevant and not chosen randomly. Figure 5-7 shows Z-scores plotted against the features of each data set.

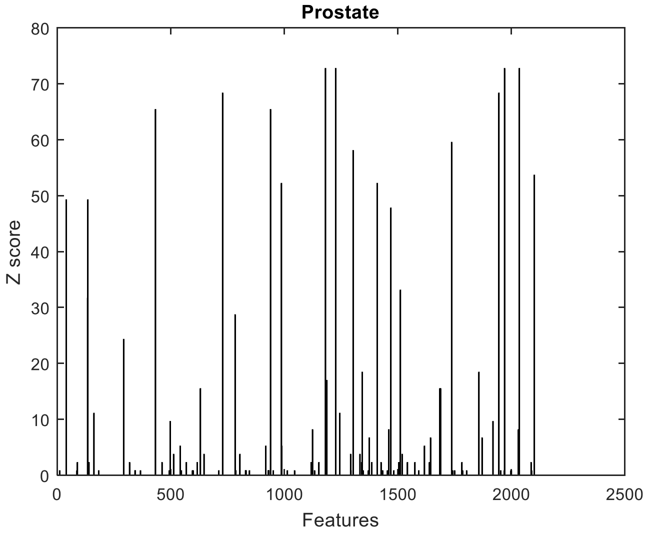

Figure 5. *Z*-score versus the features for the prostate data set

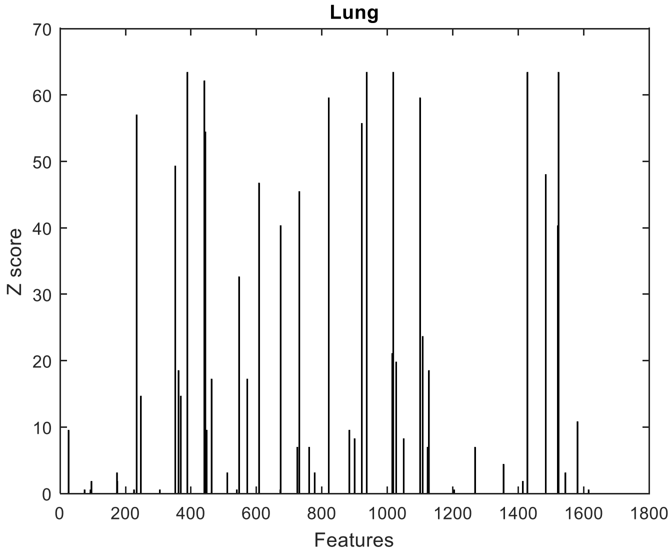

Figure 6. *Z*-score versus the features for lung cancer data set

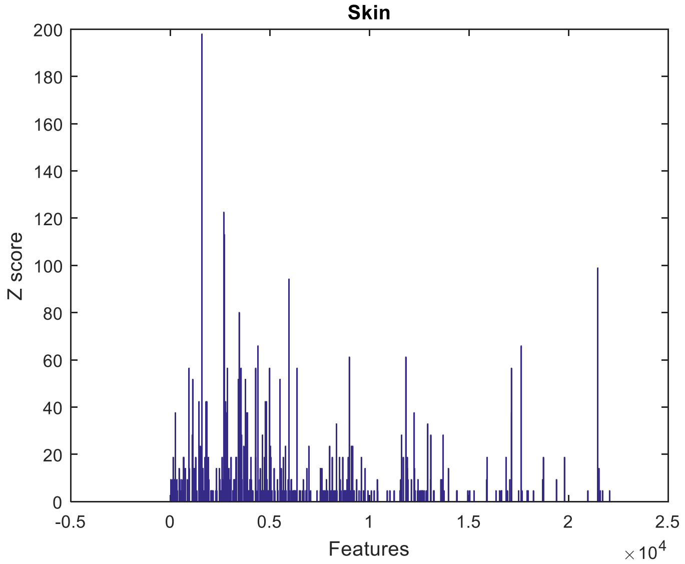

Figure 7. *Z*-score versus the features for skin cancer data set

4. Conclusion

In this study, mutual information has been use for feature selection. Mutual information offers several advantages in this regard. It can evaluate the relationship between features and class labels, whether they are linear or non-linear, making it versatile. Furthermore, mutual information can be calculated without problems even when dealing with missing data. Unlike some statistical methods that rely on the assumption of normality, mutual information does not require such an assumption.

Features are evaluated based on their mutual information scores, with higher scores indicating more information content. The results show that the reference benchmark that excludes irrelevant, redundant, and noisy features outperforms the benchmark that includes all features. This reference benchmark also determines the feature quantity required to build the predictive model. This study uses feature selection techniques and compares benchmarks using all features, emphasizing the evaluation of current methods and comprehensive benchmarks.

The proposed method defines the reference benchmark of a dataset before selecting features to build a predictive model. The robustness of selected features is evaluated using Z-scores to ensure they were not randomly chosen. Subsequently, the sensitivity tests on the predictive model using ROC curves and the area under the curve to measure model effectiveness. The proposed algorithm, based on mutual information scores, assists researchers in establishing a more informative benchmark, which in turn impacts the predictive model's performance. The proposed algorithm applies to high-dimensional data before constructing a statistical model.

Future research can explore the relationships between selected features by building network models using mutual information. Furthermore, addressing the feature selection problem under tie conditions, where the contribution of features to label categories may vary, deserves further study. Merging two datasets to increase sample size and information content is an upcoming challenge. Feature selection in small sample scenarios is also a pressing issue. Finally, addressing the imbalanced class problem in feature selection is crucial to prevent prediction models from being biased toward specific classes.

5. Acknowledgements

The authors thank the anonymous reviewers for their insightful comments and feedback.

6. References

Adeboye, N. O., Bashiru, K. A., Afolabi, H. A., & Ojurongbe, T. (2023). Diagnosing Sexually Transmitted Disease from Some Symptoms Using Machine Learning Models: Diagnosis of Sexually Transmitted Diseases. *Journal of Statistical Modeling & Analytics*, 5(1). https://doi.org/10.22452/josma.vol5no1.5

Ahuja, R., & Sharma, S. C. (2021). Exploiting Machine Learning and Feature Selection Algorithms to Predict Instructor Performance in Higher Education. *Journal of Information Science and Engineering*, 37(5), 993-1009. https://doi.org/10.6688/JISE.202109_37(5).0001

Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5(4), 537-550. https://doi.org/10.1109/72.298224

Bennasar, M., Hicks, Y., & Setchi, R. (2015). Feature selection using Joint Mutual Information Maximisation. *Expert Systems with Applications*, 42(22), 8520-8532. https://doi.org/10.1016/j.eswa.2015.07.007

Chlioui, I., Idri, A., Abnane, I., & Ezzat, M. (2021). Ensemble Case based Reasoning Imputation in Breast Cancer Classification. *Journal of Information Science and Engineering*, 37(5), 1039-1051. https://doi.org/10.6688/JISE.202109_37(5).0004

Fang, L., Zhao, H., Wang, P., Yu, M., Yan, J., Cheng, W., & Chen, P. (2015). Feature selection method based on mutual information and class separability for dimension reduction in multidimensional time series for clinical data. Biomedical Signal Processing and Control, 21, 82-89. https://doi.org/10.1016/j.bspc.2015.05.011

Jiang, L., He, J., Pan, H., Wu, D., Jiang, T., & Liu, J. (2023). Seizure detection algorithm based on improved functional brain network structure feature extraction. *Biomedical Signal Processing and Control*, 79(Part 1), 104053. https://doi.org/10.1016/j.bspc.2022.104053

Jimoh, R. G., Abisoye, O. A., & Uthman, M. M. B. (2021). Ensemble Feed-Forward Neural Network and Support Vector Machine for Prediction of Multiclass Malaria Infection. *Journal of Information and Communication Technology*, 21(1), 117-148. https://doi.org/10.32890/jict2022.21.1.6

Jirapech-Umpai, T., & Aitken, S. (2005). Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. Bioinformatics, 21(6), 168-174.

Khaire, U. M., & Dhanalakshmi, R. (2022). Stability of feature selection algorithm: A review. *Journal of King Saud University - Computer and Information Sciences*, 34(4), 1060-1073. https://doi.org/10.1016/j.jksuci.2019.06.012

Khairuddin, A. R., Alwee, R., & Haron, H. (2023). Hybrid Neighbourhood Component Analysis with Gradient Tree Boosting for Feature Selection in Forecasting Crime Rate. *Journal of Information and Communication Technology*, 22(2), 207-229. https://doi.org/10.32890/jict2023.22.2.3

Li, L., Weinberg, C.R., Darden, T.A., & Pedersen, L.G. (2001). Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics, 17(2), 1131–1142.

Liping, W. (2015). Feature Selection Algorithm Based on Conditional Dynamic Mutual Information. International Journal on Smart Sensing and Intelligent Systems, 8(1), 316-337. https://doi.org/10.21307/ijssis-2017-761

Liu, S., Li, X., Hu, C., et al. (2022). Spammer detection using multi-classifier information fusion based on evidential reasoning rule. Scientific Reports, 12, 12458. https://doi.org/10.1038/s41598-022-16576-7

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8), 1226-1238. https://doi.org/10.1109/TPAMI.2005.159

Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In A. Mechelli & S. Vieira (Eds.), Machine Learning (pp. 101-121). Academic Press. ISBN 9780128157398. DOI: https://doi.org/10.1016/B978-0-12-815739-8.00006-7

Okwonu, F. Z., Ahad, N. A., Hamid, H., Muda, N., & Sharipov, O. S. (2023). Enhanced Robust Univariate Classification Methods for Solving Outliers and Overfitting Problems. *Journal of Information and Communication Technology*, 22(1), 1-30. https://doi.org/10.32890/jict2023.22.1.1

Qin, Q., Li, J., Zhang, L., Yue, Y., & Liu, C. (2017). Combining Low-dimensional Wavelet Features and Support Vector Machine for Arrhythmia Beat Classification. Scientific Reports, 7(1), 6067. https://doi.org/10.1038/s41598-017-06596-z

Savitha, S., & Rajiv Kannan, A. (2023). A Novel Technique Based on Mutual Information Weighted Feature Selection to Predict Chronic Kidney Disease. Date of Publication Not Specified, 1(491–504).

Shi, C., Xin, X., & Zhang, J. (2022). A novel multigranularity feature-selection method based on neighborhood mutual information and its application in autistic patient identification. *Biomedical Signal Processing and Control*, 78, 103887. https://doi.org/10.1016/j.bspc.2022.103887

Sluga, D., & Lotrič, U. (2017). Quadratic Mutual Information Feature Selection. Entropy, 19, 157. https://doi.org/10.3390/e19040157

Sun, Z., Zhang, J., Luo, Z., Cao, D., & Li, S. (2018). A Fast Feature Selection Method Based on Mutual Information in Multi-label Learning. In Communications in Computer and Information Science (Vol. 917).

Somvanshi, M., Chavan, P., Tambade, S., & Shinde, S. V. (2016). A review of machine learning techniques using decision tree and support vector machine. In 2016 International Conference on Computing Communication Control and Automation (ICCUBEA) (pp. 1-7). Pune, India. DOI: 10.1109/ICCUBEA.2016.7860040.

Wang, X., Guo, B., Shen, Y., Zhou, C., & Duan, X. (2019). Input Feature Selection Method Based on Feature Set Equivalence and Mutual Information Gain Maximization. IEEE Access, 7, 151525-151538. https://doi.org/10.1109/ACCESS.2019.2948095

Xiong, C., Qian, W., Wang, Y., & Huang, J. (2021). Feature selection based on label distribution and fuzzy mutual information. Information Sciences, 574, 297-319. https://doi.org/10.1016/j.ins.2021.06.005

Yang, H. H., & Moody, J. (1999). Data visualization and feature selection: new algorithms for nongaussian data. *In Proceedings of the 12th International Conference on Neural Information Processing Systems (NIPS'99)* (pp. 687-693). MIT Press.

Zhang, Z. (2016). Introduction to machine learning: K-nearest neighbors. Annals of Translational Medicine, 4(11). https://doi.org/10.21037/atm.2016.03.37

Zhu, Z., & Zeng, G. (2015). A Unified Definition of Mutual Information with Applications in Machine Learning. Mathematical Problems in Engineering, 2015, 201874. https://doi.org/10.1155/2015/201874