

Hybrid embedded and filter feature selection methods in big-dimension mammary cancer and prostatic cancer data

Siti Sarah Md Noh¹, Nurain Ibrahim^{1,2}, Mahayaudin M. Mansor¹, Nor Azura Md Ghani¹,
Marina Yusoff^{1,2,3}

¹School of Mathematical Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Shah Alam, Malaysia

²Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Universiti Teknologi MARA, Shah Alam, Malaysia

³Faculty of Business, Sohar University, Oman, Malaysia

Article Info

Article history:

Received Nov 20, 2023

Revised Mar 6, 2024

Accepted Mar 15, 2024

Keywords:

Big-dimension data
Classification
Embedded method
Filter method
Logistic regression
Mammary cancer
Prostatic cancer

ABSTRACT

The feature selection method enhances machine learning performance by enhancing learning precision. Determining the optimal feature selection method for a given machine learning task involving big-dimension data is crucial. Therefore, the purpose of this study is to make a comparison of feature selection methods highlighting several filters (information gain, chi-square, ReliefF) and embedded (Lasso, Ridge) hybrid with logistic regression (LR). A sample size of $n=100, 75$ is chosen randomly, and the reduction features $d=50, 22$, and 10 are applied. The procedure for feature reduction makes use of the entire sample sizes. Each sample size's results are compared, including tests with no feature selection process. The results indicate that LR+ReliefF is the best method for mammary cancer data, whereas LR+IG is the best for prostatic cancer data, making the filter more suitable than embedded for big-dimension data. This study revealed that the sample's features and size influence the most effective method for selecting features from big-dimension data. Therefore, it provides insight into the most effective methods for particular features and sample sizes in high-dimensional data.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Nurain Ibrahim

School of Mathematical Sciences, College of Computing, Informatics and Mathematics

Universiti Teknologi MARA

Shah Alam, Selangor, Malaysia

Email: nurainibrahim@uitm.edu.my

1. INTRODUCTION

Feature selection was a source of inspiration for pattern recognition researchers, who typically employ it during the processing stage of machine learning. Feature selection is selecting an optimal subset of its original features to reduce the feature space based on a specific evaluation criterion. In addition to lowering feature dimensions, proper application of feature selection mitigates the effect of the curse of dimensionality to improve generalization performance [1], [2]. Utilizing the feature selection method will also enhance the interpretability of the model and reduce processing time [3]. There are three broad classifications for feature selection: filter-based, wrapper-based, and embedded-based. Filtering methods compute a score for each independent (features) model variable and then rank them according to their weights. It selects characteristics without using an algorithm. The wrapper employs a predefined algorithm and tests it on the model to identify significant features. In embedded methods, filter and wrapper methods are combined. In the training process, variable selection methods are utilized, and features are then selected analytically based on the objective of the learning model such as clustering on similar features [4].

Data has evolved in terms of characteristics and sample size in recent years. Numerous characteristics of medical data, including DNA microarray [5], brain tumours and Parkinson's disease [6], have resulted in an increased error. Due to rapid feature growth, a data set may become highly dimensional when the features number more than the sample sizes. A big-dimension data set with numerous irrelevant features and redundant information, for instance, may significantly degrade the performance of a learning algorithm. In addition, the lack of large samples to feed into the algorithm has become a limiting factor in identifying the best characteristics. Consequently, feature selection becomes essential when dealing with big-dimension data for machine learning tasks. On the other hand, the rapid increase in sample size and dimensionality presented the feature selection algorithm with significant challenges [7], [8].

Big-dimension data is familiar, and statistical scientists in academia and business work with it regularly. It is defined as data with more variables or features than the number of observations. The data is considered big-dimension when there are more variables or features than observations. When working with a dataset, researchers are accustomed to dealing with many samples relative to many features. However, due to recent improvements in data storage and processing capability, big-dimension data is now being produced in many industries. It is challenging to design a big-dimension algorithm, and the average execution time is proportional to the problem's dimensionality. As a result, it becomes increasingly more work for an algorithm to produce an accurate result and converge on the correct model as the number of dimensions increases. When utilizing big-dimension data, it is possible to overfit a model. Developing a classification model that can achieve a high level of generalization is essential. Despite this, a short sample size on big-dimension data may lead to overfitting the classification model to the training set, hindering its generalization capacity [9].

Numerous researchers in the past have investigated the topic of feature selection for big-dimension data [10], [11]. Despite this, more research needs to be conducted in the past to determine how the effect of different numbers of significant features and approaches with varying sizes of samples behave when applied to big-dimension data. The accuracy of statistical models is affected by these factors. Big-dimension data are frequently represented by microarray data [12], [13]. In light of the various sample sizes, the objective of this study was to determine the most efficient method for selecting the relevant microarray data features with minimal data loss via filters (information gain, chi-square, and ReliefF) and embedded selection methods (Lasso and Ridge) hybrid with logistic regression (LR) in big-dimension microarray data using a range of sample sizes and evaluating them concerning the size of the required features. Consequently, numerous aspects of this research issue need further investigation. In addition, this study investigates which of these approaches produces the highest quality outcomes. The remainder of the paper is structured as follows: material and procedure are explained in section 2. The experiment's findings and discussion are shown in section 3. The conclusion is presented in section 4.

2. METHOD

Figure 1 displays a new methodological approach for classifying big-dimension data in medical health. Firstly, two real-world big-dimension data will be input into the R software. Both data go through preprocessing: data cleaning (missing value and redundant), normalization, and recording. After that, we randomly selected full samples, 100 samples and 75 sample sizes. A 70:30 ratio was used for data splitting [10], [14].

2.1. Data summarization and data preparation phase

In this research, two big-dimension data display the classifier's effectiveness on the chosen feature selection method. Gravier *et al.* [15] investigated the first big-dimension mammary cancer data dataset in 2010. The data consists of 2905 features with only 168 number of samples. The second dataset used is the big-dimension prostatic cancer dataset initially analyzed by Singh *et al.* [16]. The dataset contains 102 sample sizes and 12600 features.

Data preprocessing is a set of techniques that includes preparing and transforming data into a suitable form before the mining procedure [17]. Based on Sajesh and Srinivasan [18], it is known that Mahalanobis distance is not applicable to check for outliers in high dimensional data. Hence, this study would not proceed with detecting outliers for these big-dimension mammary cancer data. Data normalization was then applied to big-dimension prostatic cancer data so that the range of values is between 0 and 1 using the min-max normalization [19]. In contrast, no normalization was applied to big-dimension mammary cancer data following the previous study from Nurlaily *et al.* [20]. One of the advantages of min-max data normalization is that the relationship with the original attribute values is maintained [21].

2.2. Filter-based steps

Three filter-based methods were employed to obtain essential features, including information gain, chi-square and reliefF. A brief explanation of each filter method is as follows:

- Information gain is a univariate filter method for evaluating the attributes [22], [23]. This approach uses the information gathered to examine one characteristic at a time. Entropy measurements are used to rank the variables. Every feature will have a unique information gain value assigned to it. A higher information gain means that the feature contains more information.
- Chi-square assesses each feature's value using a discretization algorithm and a test of independence [24]. Using chi-square statistics for each class, this technique evaluates each feature separately [25], [26]. For any class, a significant characteristic will have a high chi-square value.
- ReliefF is an extension of the relief filtering step. The difference between relief and ReliefF is rather than a single hit and miss, ReliefF uses k nearest hits and misses and averages their impact to the feature weight [27].

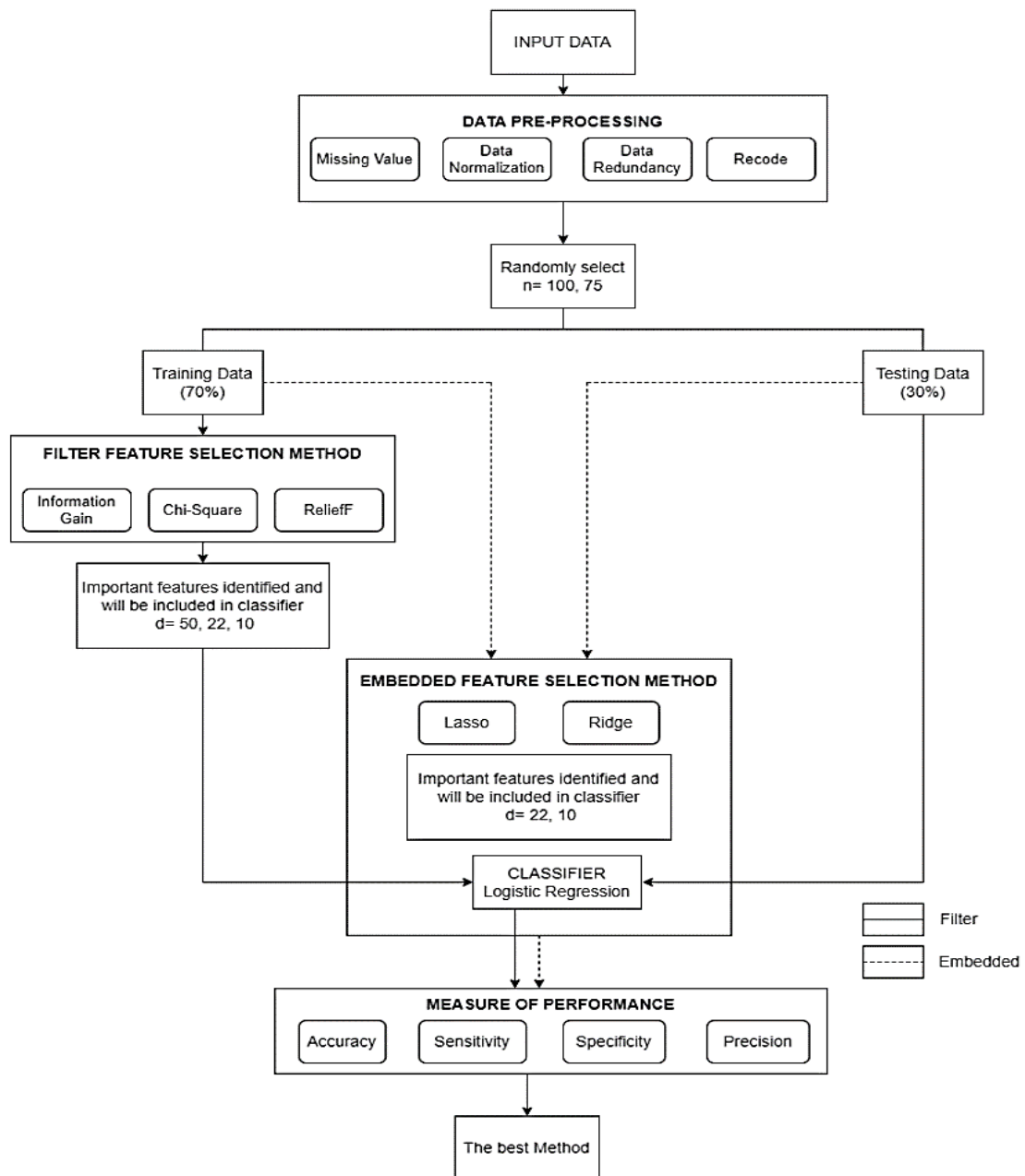


Figure 1. Conceptual research methodology

2.3. Embedded-based steps

This study involved two embed-based methods, such as lasso and ridge. The explanation for each method is being explored as follows:

- a. Lasso is a powerful method that performs two main tasks: regularization and feature selection. It was developed by Tibshirani (1996) [28] to perform parameter estimation and feature selection in regression analysis. Lasso regression aims to identify essential variables and the corresponding regression, resulting in a model with a minimum prediction error [29]. To do so, Lasso forces the total of the regression coefficient's absolute values to be smaller than a fixed value (λ) by putting a constraint on the model parameter, which will 'shrink' the regression coefficient towards zero and any variable that does not have a zero-coefficient value will be deemed significant and added to the model.
- b. Initially developed by Hoerl and Kennard in 1970, ridge was an ideal method for a dataset containing many features with non-zero coefficients and selected from the normal distribution [30]. Ridge regression is a method that was used when multicollinearity was identified. Multicollinearity will make the variance large and far away from the actual value.

2.4. Logistic regression model and model performances

Each independent variable in a LR is given a coefficient that indicates how much of the variance in the independent variable is explained. The dependent variable will become 1 if the answer is "Yes." If not, it will equal zero. The linear logistic model and the odds ratio's natural logarithm (\ln) are the two ways that the predicted probabilities model. A confusion matrix is used to establish the algorithm's performance assessment. According to previous research, accuracy was the most often used measure [31]–[33]. This study computes the proportion of accurately defined predictions, indicating the algorithm's efficacy, including accuracy, sensitivity, specificity, and precision as it is also being used in the previous research [34].

3. RESULTS AND DISCUSSION

To compare the feature selection approach between the filter and embedding in various sample sizes of big-dimension data for LR classification performance, we used sample sizes of 75, 100, and full samples. We also used the top 50 features, the top 22 features, and the top 10 features as benchmarks. It is noted that previous studies have discovered the impact of the wrapper feature selection method on prostate cancer. However, they did not explore the combination of the important features that affect the classification performances. The top 50 essential features were only used for the filter selection method because when applied to embedded Lasso, it had shrunk the coefficient to a minimum of 22 features, making comparison of the filter method with embedded become restricted. Hence, we would only use the top 50 essential features to compare between filter methods. Meanwhile, the top 22 essential features and top 10 essential features were utilized in the filter and embedded methods. Table 1, which explains the big-dimension mammary cancer data for the full sample, demonstrates the highest accuracy for hybrid ReliefF($d=50$)+LR with 74.51%. Hybrid chi-square ($d=50$)+LR acquired excellent sensitivity, obtaining the highest percentage of 84.62%. Overall, hybrid ReliefF($d=50$)+LR is the best method as each performance measure has a stable and consistent value. Meanwhile, the big-dimension prostatic cancer data for the full sample shows that hybrid ReliefF($d=50$)+LR had outperformed other techniques in all performance measures, obtaining the greatest accuracy, sensitivity, specificity, and precision with 80.65%, 80.00%, 81.25%, and 80.00% respectively.

Big-dimension mammary cancer data for $n=100$ shows that hybrid IG($d=50$)+LR has the most outstanding percentage with 66.67% accuracy, 72.22% sensitivity, 58.33% specificity, and 72.22% precision. In addition, the performance measures for big-dimension prostatic cancer show hybrid ReliefF($d=50$)+LR surpassing other methods by attaining the highest accuracy, specificity, and precision values with 83.33%, 90.00%, and 77.78%, respectively. According to Table 1, for big-dimension mammary cancer data for $n=75$, hybrid IG($d=50$)+LR gained the finest value for accuracy, sensitivity, and precision, with 86.96%, 94.44% and 89.47, respectively. Meanwhile, hybrid ReliefF ($d=50$)+LR shows up to be the best filter method in big-dimension prostatic cancer data for $n=75$ as it obtained the highest value in two out of four criteria, which was 65.22% accuracy and 55.56% precision.

Table 2 shows the embedded method hybrid Lasso($d=22$)+LR seems to work well in specificity and precision, seeing that it obtained the highest value of 91.67% each. In contrast, hybrid Ridge($d=22$)+LR, on the other hand, functions significantly in accuracy and sensitivity, attaining 72.55% and 92.31% for big-dimension mammary cancer data and $n=full$ sample. Moreover, hybrid chi-square($d=22$)+LR appeared to be the optimal approach for big-dimension prostatic cancer data as it obtained the finest values in three out of four measures, which were 90.32% in accuracy, 93.33% in sensitivity and 87.5% in precision. According to big-dimension mammary cancer data and $n=100$, hybrid Lasso($d=22$)+LR and hybrid Ridge($d=22$)+LR behaved the same way as in the full dataset, where hybrid Lasso($d=22$)+LR achieved the most outstanding values in specificity and precision with 100% each. In contrast, hybrid Ridge($d=22$)+LR obtained top values for accuracy and sensitivity with 63.33% and 100%, respectively, but did not work well in the opposite performance measures. Hybrid ReliefF($d=22$)+LR derived a consistent outcome for each performance measure, giving

63.33% accuracy, 77.78% sensitivity, 41.67% specificity, and 66.67% precision. Meanwhile, hybrid IG($d=22$)+LR, on the other hand, outshined others as it got the highest accuracy and sensitivity with fair specificity and precision values, which are %, 100%, 95%, and 90.91,% respectively when applied to big-dimension prostatic cancer data. When the specificity and precision values are considered, hybrid Lasso($d=22$)+LR and hybrid Ridge($d=22$)+LR outshine other methods, achieving 100% for both specificity and precision. The performance measures in big-dimension mammary cancer for $n=75$, confirm that hybrid chi-square($d=22$)+LR had an excellent and consistent performance in each measure, which are 73.91% in accuracy, 77.78% in sensitivity, 60.00% in specificity, and 87.50% in precision. However, hybrid Ridge($d=22$)+LR and hybrid IG($d=22$)+LR outperform others in terms of accuracy and sensitivity, with hybrid Ridge($d=22$)+LR having 82.61% accuracy and 100.00% sensitivity and hybrid IG($d=22$)+LR having 82.61% accuracy and 94.44% sensitivity, respectively.

Table 1. Accuracy, sensitivity, specificity, and precision of filter methods for top $d=50$ features across various sample sizes for big-dimension dataset

Data		$n=$ full sample				$n=100$				$n=75$			
	Methods	Acc	Sen	Spe	Pre	Acc	Sen	Spe	Pre	Acc	Sen	Spe	Pre
Mammary cancer	Without feature selection	56.86	48.72	83.33	90.48	60.00	72.22	41.67	65.00	47.83	38.89	80.00	87.50
	Hybrid IG ($d=50$)+LR	66.67	76.92	33.33	78.95	66.67	72.22	58.33	72.22	86.96	94.44	60.00	89.47
	Hybrid chi-square ($d=50$)+LR	72.55	84.62	33.33	80.49	63.33	72.22	50.00	68.42	56.52	72.22	0.00	72.22
	Hybrid ReliefF ($d=50$)+LR	74.51	76.92	66.67	88.24	60.00	72.22	41.67	65.00	60.87	66.67	40.00	80.00
	Without feature selection	58.06	33.33	81.25	62.50	30.00	10.00	40.00	7.69	56.52	33.33	71.43	42.86
Prostatic cancer	Hybrid IG ($d=50$)+LR	61.29	53.33	68.75	61.54	76.67	80.00	75.00	61.54	39.13	66.67	21.43	35.29
	Hybrid chi-square ($d=50$)+LR	70.97	66.67	75.00	71.43	80.00	90.00	75.00	64.29	60.87	33.33	78.57	50.00
	Hybrid ReliefF ($d=50$)+LR	80.65	80.00	81.25	80.00	83.33	70.00	90.00	77.78	65.22	55.56	71.43	55.56

Table 2. Accuracy, sensitivity, specificity, and precision of filter and embedded methods for top $d=22$ features across various sample sizes for big-dimension dataset

Data		$n=$ full sample				$n=100$				$n=75$			
	Methods	Acc	Sen	Spe	Pre	Acc	Sen	Spe	Pre	Acc	Sen	Spe	Pre
Mammary cancer	Without feature selection	56.86	48.72	83.33	90.48	60.00	72.22	41.67	65.00	47.83	38.89	80.00	87.50
	Hybrid IG ($d=22$)+LR	68.63	79.49	33.33	79.49	50.00	55.56	41.67	58.82	82.61	94.44	40.00	85.00
	Hybrid chi-square ($d=22$)+LR	56.86	66.67	25.00	74.29	46.67	50.00	41.67	56.25	73.91	77.78	60.00	87.50
	Hybrid ReliefF ($d=22$)+LR	72.55	84.62	33.33	80.49	63.33	77.78	41.67	66.67	52.17	55.56	40.00	76.92
	Hybrid Lasso ($d=22$)+LR	43.14	28.21	91.67	91.67	53.33	22.22	100.0	100.0	73.91	88.89	20.00	80.00
	Hybrid Ridge ($d=22$)+LR	72.55	92.31	8.33	76.60	63.33	100.0	8.33	62.07	82.61	100.0	20.00	81.82
	Without feature selection	58.06	33.33	81.25	62.50	30.00	10.00	40.00	7.69	56.52	33.33	71.43	42.86
Prostatic cancer	Hybrid IG ($d=22$)+LR	74.19	60.00	87.5	81.82	96.67	100.0	95.00	90.91	86.96	88.89	85.71	80.00
	Hybrid chi-square ($d=22$)+LR	90.32	93.33	87.50	87.50	90.00	100.0	85.00	76.92	91.30	88.89	92.86	88.89
	Hybrid ReliefF ($d=22$)+LR	80.65	86.67	75.00	76.47	93.33	90.00	95.00	90.00	82.61	77.78	85.71	77.78
	Hybrid Lasso ($d=22$)+LR	74.19	86.67	62.50	68.42	73.33	20.00	100.0	100.0	100.0	100.0	100.0	100.0
	Hybrid Ridge ($d=22$)+LR	58.06	20.00	93.75	75.00	70.00	10.00	100.0	100.0	69.57	22.22	100.0	100.0

Demonstrated in Table 3 were the performance measures for big-dimension mammary cancer and big-dimension prostatic cancer data considering 10 significant features by using filtering technique and embedded feature selection method for full sample size, $n=100$, and $n=75$ sample sizes. The big-dimension mammary cancer data and $n=\text{full sample}$ indicated hybrid ReliefF($d=10$)+LR outshines others by giving out reliable percentages for all measures, which are 84.31% accuracy, 92.31% sensitivity, 58.33% specificity, and 87.80% precision. The big-dimension prostatic cancer and $n=\text{full sample}$ specified clearly that hybrid IG ($d=10$)+LR performed the best since it produced high and consistent measurement values, which are 83.87% for accuracy, 80.00% sensitivity, 87.50% specificity, and 85.71% precision.

For big-dimension mammary cancer data and $n=100$, it was evident that hybrid Lasso($d=10$)+LR gave out consistent and high-performance values for each metric, which are 66.67% accuracy, 61.11% sensitivity, 75.00% specificity, and 78.57% precision. In contrast, hybrid Ridge($d=10$)+LR, under the same family as Lasso, had given weak results in specificity and precision, with 8.33% and 62.07%, respectively. The big-dimension mammary cancer data and $n=75$ shows that hybrid Ridge($d=10$)+LR obtained the highest values for accuracy and sensitivity, 82.61% and 77.27%, but performed poorly for specificity, gaining only 20%. Hence, hybrid Ridge($d=10$)+LR cannot become the best method since the results output is inconsistent. The method gives out a high and steady value for all measures: a hybrid chi-square($d=10$)+LR with 65.22% accuracy, 61.11% sensitivity, 80.00% specificity, and 91.67% precision value. Thus, hybrid chi-square($d=10$)+LR is the most effective method for big-dimension mammary cancer data. Meanwhile, the big-dimension prostatic cancer data and $n=75$ demonstrated that hybrid IG ($d=10$)+LR can be seen here to outshine others by obtaining the best accuracy value of 95.65% and 100% for both specificity and precision, making it the ideal method. Hybrid chi-square($d=10$)+LR and hybrid IG($d=10$)+LR were excellent methods for big-dimension mammary and prostatic cancer data, respectively.

Table 3. Accuracy, sensitivity, specificity, and precision of filter and embedded methods for top $d=10$ features across various sample sizes for big-dimension dataset

Data	Methods	$n=\text{full sample}$				$n=100$				$n=75$			
		Acc	Sen	Spe	Pre	Acc	Sen	Spe	Pre	Acc	Sen	Spe	Pre
Mammary cancer	Without feature selection	56.86	48.72	83.33	90.48	60	72.22	41.67	65	47.83	38.89	80	87.5
	Hybrid IG ($d=10$)+LR	82.35	89.74	58.33	87.5	66.67	88.89	33.33	66.67	69.57	77.78	40	82.35
	Hybrid chi-square ($d=10$)+LR	80.39	87.18	58.33	87.18	66.67	88.89	33.33	66.67	65.22	61.11	80	91.67
	Hybrid ReliefF ($d=10$)+LR	84.31	92.31	58.33	87.8	63.33	83.33	33.33	65.22	65.22	83.33	0	75
	Hybrid Lasso ($d=10$)+LR	76.47	94.87	16.67	78.72	66.67	61.11	75	78.57	73.91	94.44	0	77.27
	Hybrid Ridge ($d=10$)+LR	76.47	97.44	8.33	77.55	63.33	100	8.33	62.07	82.61	100	20	81.82
	Hybrid chi-square ($d=10$)+LR	77.42	80	75	75	90	100	85	76.92	91.3	100	85.71	81.82
Prostatic cancer	Without feature selection	58.06	33.33	81.25	62.5	30	10	40	7.69	56.52	33.33	71.43	42.86
	Hybrid IG ($d=10$)+LR	83.87	80	87.5	85.71	96.67	100	95	90.91	95.65	88.89	100	100
	Hybrid chi-square ($d=10$)+LR	77.42	80	75	75	90	100	85	76.92	91.3	100	85.71	81.82
	Hybrid ReliefF ($d=10$)+LR	80.65	86.67	75	76.47	90	90	90	81.82	95.65	100	92.86	90
	Hybrid Lasso ($d=10$)+LR	54.84	100	12.5	51.72	86.67	100	80	71.43	65.22	100	42.86	52.94
	Hybrid Ridge ($d=10$)+LR	48.39	93.33	6.25	48.28	36.67	90	10	33.33	39.13	88.89	7.14	38.1
	Hybrid chi-square ($d=10$)+LR	77.42	80	75	75	90	100	85	76.92	91.3	100	85.71	81.82

Table 4 shows the summarization of the best method in each feature size and sample in big-dimension mammary cancer data. The classification accuracy dropped from 74.51% to 66.67% as the sample size decreased from a full sample of 168 to 100. However, it showed improvement when the sample size was reduced to 75, achieving 86.96% accuracy in 50 significant features. The same scenario can also be seen in 20 important features where accuracy decreases when the full sample size decreases to 100 from 72.55% to 63.33% but then increases when the sample size is reduced to 75 obtaining 73.91%. In 20 significant features, the decline in sample size causes the classification accuracy to decrease from 84.31% in the full sample to 66.67% in 100 samples and drop to 65.22% in 75 samples.

Table 4. Summarization of the best method in each size of feature and sample in big-dimension mammary cancer data

Reduced features	Sample size		
	Full	100	75
50	Hybrid ReliefF+LR 74.51%	Hybrid IG+LR 66.67%	Hybrid IG+LR 86.96%
22	Hybrid ReliefF+LR 72.55%	Hybrid ReliefF+LR 63.33%	Hybrid chi-square+LR 73.91%
10	Hybrid ReliefF+LR 84.31%	Hybrid Lasso+LR 66.67%	Hybrid chi-square+LR 65.22%

Table 5 shows the summarization of the best method in each feature size and sample in big-dimension prostatic cancer data. The 50 significant features show an accuracy increase from 80.65% to 83.33% when the sample size was reduced from a full sample size of 102 to 100 but continued to fall to 65.22% as the sample size shrinks to 75 samples. In 22 essential features, sample size reduction shows an increase in classification accuracy by 9.68% when the full sample size is reduced to 75 samples, with an improvement in accuracy from 90.32% to 100%. The same findings can be detected in 10 significant features where accuracy increases by 11.78% when the full sample size is reduced to 75 samples. Even if there was a slight decrease as 100 samples were reduced to 75 with 96.67% to 95.65%, the results were still better than the accuracy in full sample size.

Table 5. Summarization of the best method in each size of features and sample in big-dimension prostatic cancer data

Reduced features	Sample size		
	Full	100	75
50	Hybrid ReliefF+LR 80.65%	Hybrid ReliefF+LR 83.33%	Hybrid ReliefF+LR 65.22%
22	Hybrid chi-square+LR 90.32%	Hybrid IG+LR 96.67%	Hybrid Lasso+LR 100.0%
10	Hybrid IG+LR 83.87%	Hybrid IG+LR 96.67%	Hybrid IG+LR 95.65%

Presented in Table 6 was the list of selected features for the top 22 and 10 of the best methods applied to the full sample size of big-dimension mammary cancer data and big-dimension prostatic cancer data. The bold features indicated a similar feature in the top 22 and 10. The application of Hybrid ReliefF+LR in big-dimension mammary cancer data shows that features g1CNS507, g1int1354, g7E05, g1int429, g1int372, g1int1131, g1int1662, g1int1702, g1int382, g1CNS28, g1int1130, g1int154, g1int659, g1int373, g1CNS229, g1int361, g2B01, g1int663, g1int895, g1int1414, g1int1220, g1int380 were selected as the top 22 essential features whereas g1CNS507, g1int1354, g7E05, g1int429, g1int372, g1int1131, g1int1662, g1int1702, g1int382, g1CNS28 were chosen as the top 10 features. A few of the essential features identified in this study were consistent with another study by [35] in which they found that features g1CNS507, g1int1662, g1int382, and g1CNS229 were selected after using regularized estimation under structural hierarchy for classification (C-GRESH) for estimation purpose using quadratic logistics regression.

Table 6. Top $d=22$ and top $d=10$ essential features of best feature selection method applied to $n=\text{full}$ sample for breast and prostatic cancer data

Data	Hybrid Methods	$n=\text{full}$ sample size
Mammary Cancer ($n=168$)	Hybrid ReliefF($d=22$)	g1CNS507, g1int1354, g7E05, g1int429, g1int372, g1int1131, g1int1662, g1int1702, g1int382, g1CNS28, g1int1130, g1int154, g1int659, g1int373, g1CNS229, g1int361, g2B01, g1int663, g1int895, g1int1414, g1int1220, g1int380
	Hybrid ReliefF($d=10$)	g1CNS507, g1int1354, g7E05, g1int429, g1int372, g1int1131, g1int1662, g1int1702, g1int382, g1CNS28
	+LR	
Prostatic cancer ($n=102$)	Hybrid IG($d=22$)	V9850, V10234, V7584, V6185, V4365, V9937, V6390, V6866, V8306, V8878, V8729, V8058, V12148, V6220, V8850, V8566, V7247, V8330, V5890, V10956, V8527, V12414
	Hybrid IG($d=10$)	V9850, V6185, V4365, V7584, V8729, V6390, V10234, V8965, V12414, V8850
	+LR	

Furthermore, the Hybrid IG+LR utilized in big-dimension prostatic cancer data revealed that features V9850, V10234, V7584, V6185, V4365, V9937, V6390, V6866, V8306, V8878, V8729, V8058, V12148,

V6220, V8850, V8566, V7247, V8330, V5890, V10956, V8527 and V12414 as the top 22 essential features while features V9850, V6185, V4365, V7584, V8729, V6390, V10234, V8965, V12414 and V8850 as the top 10 features. These findings were supported by [35] who found that features V6185, V4365, and V8965 were essential in big-dimension prostatic cancer data. It can also be observed that all of the features in the top 10 are also presented in the top 22 except for one feature, which was feature V8965.

4. CONCLUSION

The analysis of the newly created feature set (50, 22, and 10) reveals that the optimal approaches for both sets are inconclusive, where Hybrid ReliefF+LR is the ideal method for big-dimension mammary cancer data while Hybrid IG+LR is the finest approach for big-dimension prostatic cancer data. Thus, this shows that different data require other feature selection methods. These results prove the filter selection method is more suitable for handling big-dimension data than the embedded selection method. Additionally, the findings indicate a smaller sample size in big-dimension data enhances classification performance. This study also identified the essential features of the most effective methods resulting from objective one to reflect objective two. The findings indicate that several features identified in this study are consistent with findings from earlier studies, including features for big-dimension mammary cancer data (g1CNS507, g1int1662, g1int382, g1CNS229, g1int1130, g1int1722) and big-dimension prostatic cancer data (V6185, V4365, and V8965). These discoveries help identify the genes that cause certain diseases and improve the precision of disease detection. In terms of the limitation, this study explored a comprehensive filter and embedded feature selection method applied to big-dimension data. However, further and in-depth studies may be necessary to confirm its effectiveness. Future studies may investigate other feature selection methods and other machine learning classification techniques to apply on these mammary cancer data and prostatic cancer data. In addition, these methods also can also be applied to other high-dimensional data.

ACKNOWLEDGEMENTS

Authors acknowledge the Universiti Teknologi MARA for funding under the Malaysian Research Assessment (MyRA) Grant Scheme 600-RMC/GPM LPHD 5/3 (055/2021). The authors also acknowledge Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Research Nexus UiTM (ReNeU), and College of Computing, Informatics and Mathematics UiTM.




REFERENCES

- [1] W. Qian, Y. Xiong, W. Ding, J. Huang, and C. M. Vong, "Label correlations-based multi-label feature selection with label enhancement," *Engineering Applications of Artificial Intelligence*, vol. 127, 2024, doi: 10.1016/j.engappai.2023.107310.
- [2] S. K. Priya and K. P. Karthika, "An embedded feature selection approach for depression classification using short text sequences," *Applied Soft Computing*, vol. 147, doi: 10.1016/j.asoc.2023.110828.
- [3] H. Askr, M. A. Salam, and A. E. Hassanien, "Copula entropy-based golden jackal optimization algorithm for high-dimensional feature selection problems," *Expert Systems with Applications*, vol. 238, 2023, doi: 10.1016/j.eswa.2023.121582.
- [4] N. M. Mahfuz, M. Yusoff, and Z. Ahmad, "Review of single clustering methods," *IAES International Journal of Artificial Intelligence*, vol. 8, no. 3, pp. 221–227, 2019, doi: 10.11591/ijai.v8.i3.pp221-227.
- [5] S. Swetha, G. N. Srinivasan, and P. Dayananda, "A hybrid multiple indefinite kernel learning framework for disease classification from gene expression data," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, pp. 844–855, 2023, doi: 10.14569/IJACSA.2023.0140690.
- [6] M. Redhya and K. S. Kumar, "Refining PD classification through ensemble bionic machine learning architecture with adaptive threshold based image denoising," *Biomedical Signal Processing and Control*, vol. 85, 2023, doi: 10.1016/j.bspc.2023.104870.
- [7] X. Sun and J. Chai, "Random forest feature selection for partial label learning," *Neurocomputing*, vol. 561, 2023, doi: 10.1016/j.neucom.2023.126870.
- [8] Y. Xue, C. Zhang, F. Neri, M. Gabbouj, and Y. Zhang, "An external attention-based feature ranker for large-scale feature selection," *Knowledge-Based Systems*, vol. 281, 2023, doi: 10.1016/j.knosys.2023.111084.
- [9] Y. Liu, X. Lu, W. Peng, C. Li, and H. Wang, "Compression and regularized optimization of module stacked residual deep fuzzy system with application to time series prediction," *Information Sciences*, vol. 608, pp. 551–577, 2022, doi: 10.1016/j.ins.2022.06.088.
- [10] S. S. Noh, N. Ibrahim, M. M. Mansor, and M. Yusoff, "Hybrid filtering methods for feature selection in high-dimensional cancer data," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 6, pp. 6862–6871, 2023, doi: 10.11591/ijece.v13i6.pp6862-6871.
- [11] G. Ge and J. Zhang, "Feature selection methods and predictive models in CT lung cancer radiomics," *Journal of Applied Clinical Medical Physics*, vol. 24, no. 1, pp. 1–16, 2023, doi: 10.1002/acm2.13869.
- [12] W. Gardner, D. A. Winkler, P. J. Pigram, D. L. J. Alexander, D. Ballabio, and B. W. Muir, "Effect of data preprocessing and machine learning hyperparameters on mass spectrometry imaging models," *Journal of Vacuum Science & Technology A*, vol. 41, no. 6, 2023, doi: 10.1116/6.0002788.
- [13] L. Y. Jia, T. Wang, A. G. Gad, and A. Salem, "A weighted-sum chaotic sparrow search algorithm for interdisciplinary feature selection and data classification," *Science Report*, vol. 13, no. 1, pp. 1–28, 2023, doi: 10.1038/s41598-023-38252-0.
- [14] Q. H. Nguyen *et al.*, "Influence of data splitting on performance of machine learning models in prediction of shear strength of soil," *Mathematical Problems in Engineering*, vol. 2021, 2021, doi: 10.1155/2021/4832864.




- [15] E. Gravier *et al.*, "A prognostic DNA signature for T1T2 node-negative breast cancer patients," *Genes Chromosomes Cancer*, vol. 49, no. 12, pp. 1125–1134, Dec. 2010, doi: 10.1002/gcc.20820.
- [16] D. Singh *et al.*, "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, no. 2, pp. 203–209, 2002, doi: 10.1001/archsurg.1980.01380080083019.
- [17] S. A. Alasadi and W. S. Bhaya, "Review of data preprocessing techniques in data mining," *Journal of Engineering and Applied Sciences*, vol. 12, no. 16, pp. 4102–4107, 2017.
- [18] T. A. Sajesh and M. R. Srinivasan, "Outlier detection for high dimensional data using the comedian approach," *Journal of Statistical Computation and Simulation*, vol. 82, no. 5, pp. 745–757, May 2012, doi: 10.1080/00949655.2011.552504.
- [19] T. N. Nuklianggraita, A. Adiwijaya, and A. Aditsania, "On the feature selection of microarray data for cancer detection based on random forest classifier," *Journal Infotel*, vol. 12, no. 3, pp. 89–96, 2020, doi: 10.20895/infotel.v12i3.485.
- [20] D. Nurlaili, S. W. Purnami, and H. Kuswanto, "Support vector machine for imbalanced microarray," *AIP Conference Proceedings*, vol. 2194, no. 1, 2019.
- [21] G. Manikandan, N. Sairam, S. Sharmili, and S. Venkatakrishnan, "Achieving privacy in data mining using normalization," *Indian Journal of Science and Technology*, vol. 6, no. 4, pp. 4268–4272, 2013, doi: 10.17485/ijst/2013/v6i4.16.
- [22] M. A. Khaldy, "Resampling imbalanced class and the effectiveness of feature selection methods for heart failure dataset," *International Robotics & Automation Journal*, vol. 4, no. 1, 2018, doi: 10.15406/iratj.2018.04.00090.
- [23] R. Porkodi, "Comparison of filter based feature selection algorithms: an overview," *International Journal of Innovative Research in Technology & Science*, vol. 2, no. 2, pp. 108–113, 2014.
- [24] C. D. Stefano, F. Fontanella, and A. S. D. Freca, "Feature selection in high dimensional data by a filter-based genetic algorithm," *Applications of Evolutionary Computation: 20th European Conference*, vol. 10199, pp. 506–521, 2017, doi: 10.1007/978-3-319-55849-3_33.
- [25] S. Deepa Lakshmi and T. Velmurugan, "Empirical study of feature selection methods for high dimensional data," *Indian Journal of Science and Technology*, vol. 9, no. 39, pp. 1–6, 2016, doi: 10.17485/ijst/2016/v9i39/90599.
- [26] A. B. Pedersen *et al.*, "Missing data and multiple imputation in clinical epidemiological research," *Clinical Epidemiology*, vol. 9, pp. 157–166, 2017, doi: 10.2147/CLEP.S129785.
- [27] R. J. Urbanowicz, M. Meeker, W. L. Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: Introduction and review," *Journal of Biomedical Informatics*, vol. 85, pp. 189–203, 2018, doi: 10.1016/j.jbi.2018.07.014.
- [28] R. Shrinkage, "Regression Shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B*, vol. 58, no. 1, pp. 267–288, 1996, [Online]. Available: jstor.org/stable/2346178.
- [29] N. Ibrahim, "Variable selection methods for classification: application to metabolomics data," Ph.D Thesis, Department of Biostatistics, University of Liverpool, Liverpool, United Kingdom, 2020.
- [30] J. O. Ogutu, T. S. -Streeck, and H. P. Piepho, "Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions," *BMC proceedings*, vol. 6, no. s10, 2012, doi: 10.1186/1753-6561-6-S2-S10.
- [31] N. Ibrahim and A. N. Kamarudin, "Assessing time-dependent performance of a feature selection method using correlation sharing T-statistics (corT) for heart failure data classification," *AIP Conference Proceedings*, vol. 2500, 2023, doi: 10.1063/5.0109918.
- [32] M. Mafarja and S. Mirjalili, "Whale optimization approaches for wrapper feature selection," *Applied Soft Computing*, vol. 62, pp. 441–453, 2018, doi: 10.1016/j.asoc.2017.11.006.
- [33] A. Mirzaei, Y. Mohsenzadeh, and H. Sheikhzadeh, "Variational relevant sample-feature machine: a fully bayesian approach for embedded feature selection," *Neurocomputing*, vol. 241, pp. 181–190, 2017, doi: 10.1016/j.neucom.2017.02.057.
- [34] N. Ibrahim, H. A. A. Rahman, A. A. Azran, M. A. M. Faddillah, and M. A. Q. M. Qamarudin, "Prediction of water quality for the selangor rivers using data mining approach," *Journal of Sustainability Science and Management*, vol. 18, no. 9, pp. 171–183, 2023.
- [35] H. Jiang and Y. Dong, "Structural regularization in quadratic logistic regression model," *Knowledge-Based Sys.*, vol. 163, pp. 842–857, 2019, doi: 10.1016/j.knosys.2018.10.012.

BIOGRAPHIES OF AUTHORS






Siti Sarah Md Noh    received a Diploma in Statistics in 2018 and a Bachelor of Science (Hons) (Statistics) in 2021 at Universiti Teknologi MARA. She is a student undergoing a Master of Science in Applied Statistics at Universiti Teknologi MARA Shah Alam. Her research interest includes data mining, applied statistics, and medical statistics. She can be contacted at email: sarah97noh@gmail.com.






Nurain Ibrahim    received a Bachelor of Science (Hons) (Statistics), Master of Science Applied Statistics from Universiti Teknologi MARA in 2013 and 2014, and a Ph.D. in Biostatistics from University of Liverpool, United Kingdom, in 2020. She is currently a senior lecturer with the School of Mathematical Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Malaysia. She is also an associate fellow researcher at the Institute for Big Data Analytics and Artificial Intelligence (IBDAAI). Her research interests include biostatistics, data mining with multivariate data analysis, and applied statistics. She has experience teaching A-level students for the UK, US, and Germany at INTEC Education College from 2013 to 2014. She can be contacted at email: nurainibrahim@uitm.edu.my.






Mahayaudin M. Mansor    obtained a B.Sc.App. (Hons) (Mathematical Modelling) from Universiti Sains Malaysia in 2005, a M.Sc. (Quantitative Science) from Universiti Teknologi MARA in 2011, and a Ph.D. in Statistics from The University of Adelaide, Australia, in 2018. He worked in banking and insurance before teaching Statistics and Business Mathematics at the Universiti Tun Abdul Razak Malaysia. After completing his Ph.D. studies, he worked as a researcher specializing in data management and quantitative research at the Australian Centre for Precision Health, University of South Australia. He is a senior lecturer in Statistics at the School of Mathematical Sciences, Universiti Teknologi MARA and an external supervisor registered at the School of Mathematical Sciences, The University of Adelaide. His research interests include data analytics, time series, *R* computing, quantitative finance, and vulnerable populations. He can be contacted at email: maha@uitm.edu.my.



Nor Azura Md Ghani    is a Professor at the School of Mathematical Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Malaysia, and the Chair of the IEEE Computer Society Malaysia Chapter. She currently serves as the Director at the Research Management Center, Universiti Teknologi MARA, Malaysia. Her expertise is big data, image processing, artificial neural networks, statistical pattern recognition, and forensic statistics. She is the author or coauthor of many high-impact journals and conference proceedings at national and international levels. She can be contacted at email: azura158@uitm.edu.my.



Marina Yusoff    is a Deputy Director at the Institute for Big Data Analytics and Artificial Intelligence (IBDAAI) and Associate Professor of Computer and Mathematical Sciences, Universiti Teknologi MARA Shah Alam, Malaysia. She has a Ph.D. in Information Technology and Quantitative Sciences (Intelligent Systems). She previously worked as a Senior Executive of Information Technology at SIRIM Berhad, Malaysia. She is most interested in multidisciplinary research, artificial intelligence, nature-inspired computing optimization, and data analytics. She applied and modified AI methods in many research and projects, including recent hybrid deep learning, particle swarm optimization, genetic algorithm, ant colony, and cuckoo search for many real-world problems, medical and industrial projects. Her current projects are data analytic optimizer, audio, and image pattern recognition. She has many impact journal publications and contributes as an examiner and reviewer to many conferences, journals, and universities' academic activities. She can be contacted at email: marina998@uitm.edu.my.