Hybrid filtering methods for feature selection in high-dimensional cancer data

Siti Sarah Md Noh[1], Nurain Ibrahim[1,3], Mahayaudin M. Mansor[1], Marina Yusoff[2,3]
[1]School of Mathematical Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Shah Alam, Malaysia
[2]School of Computing Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Shah Alam, Malaysia
[3]Institute for Big Data Analytics and Artificial Intelligence, Kompleks Al-Khawarizmi, Universiti Teknologi MARA, Shah Alam, Malaysia

Article Info

Article history:

Received Jan 31, 2023
Revised Apr 8, 2023
Accepted Apr 14, 2023

Keywords:

Breast cancer
Classification
Filtering method
High-dimensional data
Logistic regression
Prostate cancer

ABSTRACT

Statisticians in both academia and industry have encountered problems with high-dimensional data. The rapid feature increase has caused the feature count to outstrip the instance count. There are several established methods when selecting features from massive amounts of breast cancer data. Even so, overfitting continues to be a problem. The challenge of choosing important features with minimum loss in a different sample size is another area with room for development. As a result, the feature selection technique is crucial for dealing with high-dimensional data classification issues. This paper proposed a new architecture for high-dimensional breast cancer data using filtering techniques and a logistic regression model. Essential features are filtered out using a combination of hybrid chi–square and hybrid information gain (hybrid IG) with logistic regression as classifier. The results showed that hybrid IG performed the best for high-dimensional breast and prostate cancer data. The top 50 and 22 features outperformed the other configurations, with the highest classification accuracies of 86.96% and 82.61%, respectively, after integrating the hybrid information gain and logistic function (hybrid IG+LR) with a sample size of 75. In the future, multiclass classification of multidimensional medical data to be evaluated using data from a different domain.

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, Malaysia
Email: nurainibrahim@uitm.edu.my

1. INTRODUCTION

In the medical field of breast and prostate cancer, high-dimensional data is defined as when the number of variables or features exceeds the number of observations. Statistical scientists in academia and industry encounter this evidence daily. Most researchers have always worked with data that has many features. However, advances in data storage and computing capacity have resulted in the development of high-dimensional data in various fields, such as genetics, signal processing, and finance. The accuracy of classification algorithms tends to decline in high dimensions data due to a phenomenon known as the curse of dimensionality. When the dimensionality increases, the volume of space expands rapidly, and the accessible data becomes sparse.

Furthermore, recognizing areas where objects form groups with similar attributes is widely used to organize and search data. However, due to the high dimensionality of the data, it became sparse, resulting in increased errors. It is challenging to design an algorithm for dealing with high-dimensional data. The ability of an algorithm to give a precise result and converge to an accurate model decreases as dimensionality increases.

The next challenge in modeling high-dimensional data is to avoid overfitting. It is critical to develop a classification model that is capable of generalization. The classification model must perform admirably in both training and testing data. Nonetheless, the small number of samples on high-dimensional data can cause the classification model to be overfitted to the training data, resulting in poor model generalization ability. To avoid the abovementioned issues, feature selection must be applied to high-dimensional data beforehand to select only the significant features. The problem is determining the most efficient method to determine the relevant elements with less loss in different sample sizes. Several studies on high-dimensional classification reporting methods have been published in the literature. Liang *et al.* [1] proposed conditional mutual information-based feature selection with interaction to reduce performance error [2]. Tally *et al.* [3] discovered the genetic algorithm feature selection with a support vector machine classifier for intrusion detection, while Sagban *et al.* [2] investigated the performance of feature selection applied to cervical cancer data. Ibrahim and Kamarudin [4] applied filter feature selection method to improve heart failure data classification.

Guo *et al.* [5] proposed weighting and ranking-based hybrid feature selection to select essential risk factors for ischemic stroke detection, and Cekik *et al.* [6] developed a proportional rough feature selector to classify short text. Too many researchers focus on fusing feature selection, leaving traditional techniques like filter, wrapper, and embedded unstudied. A traditional feature selection method contains no hybrid or novel features. Way *et al.* [7] also investigated how small sample size affects feature selection and classification accuracy. More research is needed to understand how small sample sizes affect high-dimensional data. Wah *et al.* [8] compared information gain and correlation-based feature selection, wrapper sequential forward and sequential backward elimination to maximize classifier accuracy but still need to include the embedded technique. As a result, this study proposes the best integration method for evaluating high-dimensional data classification performance. As a result, breast cancer classification would improve with key features. As a result, much of this research topic requires further investigation.

The rest of the paper is organized as follows: section 2 explains material and method. Section 3 presents the results and the discussion of the experiment. Section 4 constitutes the conclusion.

2. MATERIAL AND METHOD

This section describes the structure, method, and procedure of the study. The methodology for this study includes data collection, preprocessing, feature extraction, and modeling. Patients with T1T2N0 breast carcinoma were studied at the Marie Curie Institute in Paris [9]. Data on prostate cancer is also being used, and the data were first analyzed by Singh *et al.* [10]. To remove unnecessary information, the data is preprocessed. The training set will be filtered using hybrid information gain (hybrid IG) and hybrid chi–square to select essential features. After that, the data is used to generate training and testing sets. Figure 1 depicts the conceptual research methodology for the study. To begin, two high-dimensional data sets, breast and prostate cancer, will be entered into the R software. The data will be preprocessed, which includes detecting, removing, or replacing missing values with appropriate ones and checking for redundant ones. Hence, 100 samples at random and 75 sample sizes before dividing the data into 70% and 70%, where 70% for training and 30% for testing [11], [12]. Training data samples were used as input for the classifier to learn the features and build a new learning model [13].

Meanwhile, the learning model was used during the testing phase to predict the test data. The training data will be filtered using two filter selection methods after the data has been split: information gain and chi-square. During this process, the important features were identified, and several features began to be reduced based on their ranking. The ranking was sorted based on the weight of the individual features, with a higher weight indicating a status of an importance feature. Following that, all of the identified essential features used in this study have d =50, 22, and 10, where d is the dimension of reduced features, and will be fed into the logistic regression data mining algorithm, from which a new model was built. The performance of each data was then predicted and evaluated using testing data. Finally, the testing data was used to make a prediction and the classifier's performance was evaluated using accuracy, sensitivity, specificity, and precision.

2.1. Data acquisition

The breast cancer data was first explored [9]. It was collected from a patient at Institute Curie for ten years from 1989 until 1999 or pT1T2N0 breast carcinoma. The data has clinicopathological characteristics of the tumor and the gene expression that had two classes where patients with no event after diagnosis were labeled good and patients with early metastatic will be labeled poor. The data consist of 2,905 genes with only 168 samples. Recent studies show that many researchers used breast cancer data to tackle the problem of high-dimensional data [14], [15]. The second data will be used in prostate cancer, which was initially analyzed by [10]. It was gathered from 1995 to 1997 from Brigham and Women's Hospital patients who were having radical prostatectomy surgery. The data contains 102 patterns of gene expression, of which 50 are from normal prostate specimens, and 52 are from tumors. The data, a collection of gene expression data based on oligonucleotide microarray, contains roughly 12,600 genes. Past studies show that there are a lot of past researchers that used prostate cancer data in investigating the classification of high-dimensional data [16]–[19]. The summary of these high-dimensional cancer data is shown in Table 1.

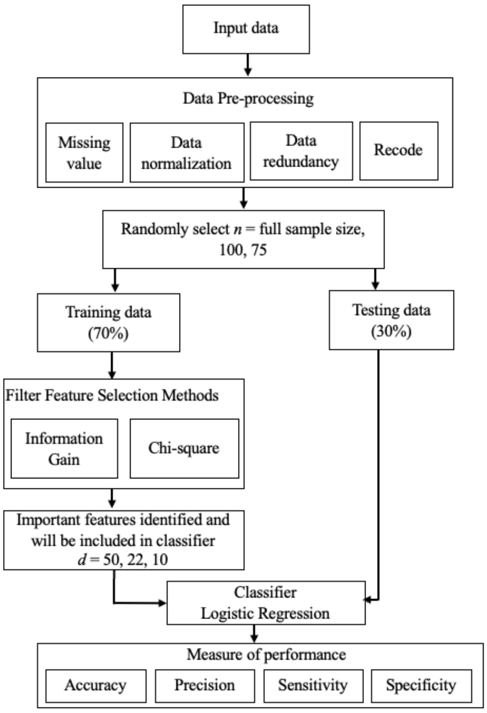

Figure 1. Conceptual research methodology

Table 1. Summary of high-dimensional cancer data

Data	No. of features	No. of samples	No. of classes	Reference
Breast cancer	2905	168	2	[9]
Prostate cancer	12600	102	2	[10]

2.2. Data preparation

Data preparation consists of preprocessing, identifying sample size, and selecting features. Pre-processing entails cleaning the data by detecting, removing, or replacing missing values with appropriate ones and examining redundant ones. The sample sizes are determined based on the previous study [20]. This study uses 75, 100, and the maximum sample size (full sample size of data set) as sample size configurations. The top 50 important, top 22 important, and top 10 important features are listed. The features identified are based on industry standards [21].

2.3. Filtering steps

For the feature selection process, filtering methods are chosen. Two filtering methods were employed to obtain three subsets of essential features. The filter selection method is a feature ranking technique that assesses the features independently, based on the data characteristic, without involving any learning algorithm or classifier. Each variable is scored using a suitable ranking criterion by assigning weight for each feature. The weight under the threshold value would be deemed unimportant and removed [22]. Then, all the reduced features would be input into the learning algorithm to assess the performance of the measurement.

2.4. Information gain

Based on the literature review, information gain was one of the most widely used univariate filters in evaluating the attributes [8], [23], [24]. This filter method analyzes a single feature at a time based on the information gained. It uses entropy measures to rank the variables. It is calculated by calculating the changes in entropy from a previous state to the known value state. The entropy for class features Y is presented as (1).

$$H(Y) = \sum_{i=1}^{n} p(y) log_2(p(y)) \quad (1)$$

The marginal probability density function for the random variable Y is denoted by $p(y)$. The marginal probability density function for the random variable Y is denoted by $p(y)$. There was a link between feature X and Y. It occurs in the training data, where the observed value for y was partitioned based on the second feature X. The result of partitioning makes the entropy of y produced by X regarding the partition less than that of Y before the partitioning. Hence, the entropy of Y after observing X is stated in (2). Once both entropy is computed, the differences are calculated to determine the gain value. The gain values from Y and X are the reduction in entropy values known as information gain. The calculation is as in (3).

$$H(Y|X) = \sum_{i=1}^{n} p(x) \sum_{i=1}^{n} p(y|x) log_2(p(y|x)) \quad (2)$$

$$IG(X) = H(Y) - H(Y|X) \quad (3)$$

The feature will each be ranked with its own information gain value. Higher information gain value will hold more information. After obtaining the information gain value, a threshold is needed to select the important features according to the order accepted. However, the weakness of using information gain is that it does not remove redundant data and needs to be more balanced with features that have more value, even though it only holds a little information.

2.5. Chi–square

Chi-square is a univariate filter algorithm that uses a test of independence evaluation to measure each feature's merit using a discretization algorithm [25]. This method assesses each feature individually by calculating chi–square statistics concerning each class [26], [27]. A relevant feature will have a high chi–square value for each class. The equation of measure is shown in (4). Given from (4), where V denotes the number of intervals, B is the number of classes, N denotes the total number of instances, R_i refers to the number of instances in the i^{th} range, B_j the number of instances in j^{th} class and finally, A_{ij} is the number of instances in the i^{th} range and j^{th} class.

$$X^2 = \sum_{i=1}^{V} \sum_{j=1}^{B} \frac{\left[A_{ij} - \frac{R_i * B_j}{N}\right]^2}{\frac{R_i * B_j}{N}}, \quad (4)$$

2.6. Logistic regression model

Logistic regression assigns each independent variable a coefficient that explains the contribution to variation in the independent variable. If the response is "Yes," the dependent variable will become 1. Otherwise, it will become 0. The predicted probabilities model is expressed as a natural logarithm (ln) of the odds ratio and the linear logistic model is shown as in (5) [28].

$$x_i b = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \ldots\ldots + \beta_k x_{ik} \quad (5)$$

$$ln\left[\frac{u_i}{1-u_i}\right] = x_i b = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \ldots\ldots + \beta_k x_{ik} \quad (6)$$

and

$$\frac{u_i}{1-u_i} = e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \ldots\ldots + \beta_k x_{ik}} \quad (7)$$

$$u_i = e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \ldots\ldots + \beta_k x_{ik}} - u_i e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \ldots\ldots + \beta_k x_{ik}} \quad (8)$$

$$u_i = \frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \ldots\ldots + \beta_k x_{ik}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \ldots\ldots + \beta_k x_{ik}}} \quad (9)$$

where $ln\left[\frac{u_i}{1-u_i}\right]$ is the log odds of the outcome, u_i is the binary outcome, $X_{i1}, X_{i2}, \ldots X_k$ is the independent variable, $\beta_{i1}, \beta_{i2} \ldots . \beta_{ik}$ is the regression coefficient and β_0 is the intercept.

2.7. Performance measures

The evaluation of performance on the algorithm used is defined from a matrix with the numbers of instances correctly and incorrectly classified for each class are called a confusion matrix. It is a table with two rows and two columns that displays the number of true positives (TP), true negatives (TN), false negatives (FN) and false positives (FP). Accuracy was shown as the most used metric in the past study [29], [30] calculates the percentage of correctly specified predictions which shows the effectiveness of the chosen algorithm. Equation (10) shows the calculation of accuracy. Furthermore, sensitivity is a test's ability to specify a positive class called a true positive rate. It is a ratio of true positives to the sum of true positives and false negatives. Mathematically it can be stated as (11). In addition, this study also used specificity as one of the performance metrics. Specificity is the test's ability to correctly identify the negative class called the true negative rate. It is calculated by dividing the true negative by the sum of the true negative and the false positive. It can be stated as (12). Nevertheless, precision is the last performance metric used in this study. Precision is the ability of a test to assign the positive event to the positive class. Equation (13) states the calculation for precision.

$$Accuracy = \frac{TP+TN}{(TP+TN+FP+FN)} \tag{10}$$

$$Sensitivity = \frac{TP}{(TP+FN)} \tag{11}$$

$$Specificity = \frac{TN}{(TN+FP)} \tag{12}$$

$$Precision = \frac{TP}{(TP+FP)} \tag{13}$$

3. RESULTS AND DISCUSSION

The study features several experiments to find a promising binary classification solution for breast cancer and prostate cancer binary classification solution. The analysis will be based on experiments integrating filtering methods with logistic regression. The methods are hybrid IG (d =22)+LR, hybrid IG (d =10)+LR, hybrid chi–square (d =22)+LR and hybrid chi–square (d =10)+LR. The measurement considers full sample size, 100 sample size, and 75 sample size of breast cancer and prostate cancer data.

3.1. Result for top $d = 50$ important features

Performance measures on the full sample size, 100 sample size, and 75 sample size applied to the high-dimensional breast and prostate cancer data are shown in Table 2. The performance of feature selection methods was assessed using classification accuracy, sensitivity, specificity, and precision. As demonstrated in Table 2, the accuracy is highest for hybrid chi–square (d =50)+LR with 72.55% and hybrid IG (d =50)+LR reporting at 66.67% for full sample sizes. The worst accuracy for no feature selection applied with only 56.86% value can be found. Hybrid chi–square (d =50)+LR has good performance in sensitivity, where it obtained the highest value of 84.62%, whereas hybrid IG (d =50)+LR and a reasonably good sensitivity value of 76.92%. No feature selection method again has the worst sensitivity value with only 48.72%. However, it can be seen when comparing specificity and precision that no feature selection approach outperforms others with 83.33% and 90.48%, respectively but also obtained the worst accuracy and sensitivity value, which are 56.89% and 48.72%. For a sample size of 100, hybrid IG (d =50)+LR obtained the best value for all performance measures with 66.67% accuracy, 72.22% sensitivity, 58.33% specificity, and 72.22% precision. No feature selection approach is the worst feature selection method with the lowest accuracy, specificity, and precision of 60.00%, 41.67% and 65.00%, respectively. Furthermore, every feature selection method has the same sensitivity value of 72.22%.

Meanwhile, for a sample size of 75, it can be observed that hybrid IG (d =50)+LR holds the highest value for accuracy, sensitivity, and precision which are 86.96%, 94.44%, and 89.47% when applied to high-dimensional breast cancer data. However, when considering the specificity, it can be seen that no feature selection has the highest value among the rest, 80.00%. In summary, hybrid IG (d =50)+LR performs the best for breast cancer data since it achieves the best performance for all criteria. Even though other methods have been demonstrated to perform best in some performance metrics, such as no feature selection, which reaches the highest specificity value when applied to breast cancer data, when considering other performance values, hybrid IG (d =50)+LR still outperforms other methods in most performance metrics. As a result, the optimum filter selection strategy is hybrid IG (d =50)+LR for breast cancer for 75 sample sizes considering the top d =50 important features.

For prostate cancer data with full sample sizes, hybrid chi–square (d =50)+LR shows the highest accuracy, sensitivity, and precision percentages with 70.97%, 66.67%, and 71.43%. Contradict output can be seen when these feature selection methods were applied to prostate cancer with sample sizes of 100 and 75, where hybrid chi–square (d =50)+LR outperformed other methods with 80.00%, 90.00%, 75.00% and 64.29% of accuracy, sensitivity, specificity, and precision respectively. Hence, hybrid chi–square (d =50)+LR is the best method for full sample size, considering only 50 significant features for both high-dimensional breast and prostate cancer data.

Table 2. Performance measures of each filter method applied to top d =50 features and $n =$ full sample, 100 and 75 for breast cancer data

Data	Performance Measures	$n =$ full sample size			n =100			n =75		
		No Feature Selection	Hybrid IG (d=22) +LR (%)	Hybrid chi-square (d=22) + LR (%)	No Feature Selection	Hybrid IG (d=22) +LR (%)	Hybrid chi-square (d=22) + LR (%)	No Feature Selection	Hybrid IG (d=22) +LR (%)	Hybrid chi-square (d=22) + LR (%)
Breast Cancer	Accuracy	56.86	66.67	72.55	60.00	66.67	63.33	47.83	86.96	56.52
	Sensitivity	48.72	76.92	84.62	72.22	72.22	72.22	38.89	94.44	72.22
	Specificity	83.33	33.33	33.33	41.67	58.33	50.00	80.00	60.00	0.00
	Precision	90.48	78.95	80.49	65.00	72.22	68.42	87.50	89.47	72.22
Prostate Cancer	Accuracy	58.06	61.29	70.97	30.00	76.67	80.00	56.52	39.13	60.87
	Sensitivity	33.33	53.33	66.67	10.00	80.00	90.00	33.33	66.67	33.33
	Specificity	81.25	68.75	75.00	40.00	75.00	75.00	71.43	21.43	78.57
	Precision	62.50	61.54	71.43	7.69	61.54	64.29	42.86	35.29	50.00

3.2. Result for Top $d = 22$ important features

Performance measures of each filter method applied to top d =22 features, and $n = full$ sample, 100 and 75 for high-dimensional breast cancer data is demonstrated in Table 3. As can be seen in Table 3, the hybrid IG (d =22)+LR shows the highest accuracy and sensitivity of 68.63% and 79.49%, respectively. However, in terms of specificity and precision, no feature selection outperforms other methods with percentages of 83.33%. Thus, after considering only 22 features, filter hybrid IG (d =22)+LR is the optimal feature selection approach for the full sample size of high-dimensional breast cancer with top d =22 features based on accuracy and sensitivity. Each method's performance metrics were assessed using 100 sample sizes for high-dimensional breast cancer data. When comparing specificity, no feature selection, hybrid IG (d =22) +LR and hybrid chi-square shared the same performance of 41.67%. The more stable performance of no feature selection with a high value for each measure goes to achieving 60.00% accuracy, 72.22% sensitivity, and 65.00% precision. In addition, hybrid chi–square looks to perform the worst, scoring 46.67% in accuracy and 56.25% in precision.

Table 3. Performance measures of each filter method applied to top $d = 22$ features and $n =$ full sample, 100 and 75 for high-dimensional breast and prostate cancer data

Data	Performance Measures	$n =$ full sample size			$n = 100$			$n = 75$		
		No Feature Selection	Hybrid IG (d=22) +LR (%)	Hybrid chi-square (d=22) + LR (%)	No Feature Selection	Hybrid IG (d=22) +LR (%)	Hybrid chi-square (d=22) + LR (%)	No Feature Selection	Hybrid IG (d=22) +LR (%)	Hybrid chi-square (d=22) + LR (%)
Breast Cancer	Accuracy	56.86	68.63	56.86	60.00	50.00	46.67	47.83	82.61	73.91
	Sensitivity	48.72	79.49	66.67	72.22	55.56	50.00	38.89	94.44	77.78
	Specificity	83.33	33.33	25.00	41.67	41.67	41.67	80.00	40.00	60.00
	Precision	90.48	79.49	74.29	65.00	58.82	56.25	87.50	85.00	87.50
Prostate Cancer	Accuracy	58.06	74.19	90.32	30.00	96.67	90.00	56.52	86.96	91.30
	Sensitivity	33.33	60.00	93.33	10.00	100.0	100.0	33.33	88.89	88.89
	Specificity	81.25	87.50	87.50	40.00	95.00	85.00	71.43	85.71	92.86
	Precision	62.50	81.82	87.50	7.69	90.91	76.92	42.86	80.00	88.89

For a sample size of 75, hybrid IG (d =22)+LR outperforms others in accuracy and sensitivity, with 82.61% accuracy and 94.44% sensitivity. Hence, it can be concluded that hybrid IG (d =22)+LR is the best feature selection method for high-dimensional breast cancer data. In addition, no feature selection yielded the lowest performance measure for accuracy, with 47.83% and 38.89% for sensitivity, making it the worst possible method to be applied. It can be said that filter hybrid IG (d =22)+LR is the best feature selection method for high-dimensional breast cancer. However, if the study wants to proceed with filtering techniques, it is better to use hybrid IG (d =22)+LR. It is the best feature selection method, with 22 features and 75 sample sizes for high-dimensional breast cancer respectively. These feature selection procedures were also applied to prostate cancer, and Table 3 displayed that hybrid Chi–Square (d =22)+LR is the best procedure for top d =22 and n = full sample and n =75 as it obtained the highest accuracies, sensitivity, specificity, and specificity precision.

3.3. Result for top $d = 10$ important features

The performance measures of each filter method applied to top d =10 features and n = full sample, 100 and 75 for high-dimensional breast cancer data are illustrated in Table 4. The performance metrics for high-dimensional breast cancer data using ten significant features by filtering technique (hybrid IG (d =10) + LR and hybrid chi–square (d =10)+LR), for full sample size and then compared to no feature selection. In high-dimensional breast data, no feature selection has the highest performance value for specificity and precision with 83.33% and 90.48%, respectively but performs poorly for the other two criteria having the lowest values for accuracy, 56.86%, and sensitivity, 48.72%. Hence, it can be concluded that no feature selection performs the worst as it gives out a very imbalanced output. Hybrid IG (d =10)+LR had the best performance as it gave out stable and high measurement values of 83.87%, 80.00%, 87.50%, and 85.71% for accuracy, sensitivity, specificity, and precision, respectively.

For a sample size of 100, as shown in Table 4, performance metrics were applied to two data, high-dimensional breast cancer data, where two different feature selection was used, which are filter (hybrid IG (d =10)+LR and hybrid chi–square (d =10)+LR) by 100 sample size and 10 important features. The results were compared with no feature selection. Hybrid IG (d =10)+LR and hybrid chi–square (d =10)+LR also gave a good result, with both having the same values for all metrics, which are 66.67% accuracy, 88.89% sensitivity, 33.33% specificity, and 66.67 precision.

For a sample size of 75, Table 4 illustrates the performance measure for high-dimensional breast cancer data when applying hybrid IG (d =10)+LR and hybrid chi–square (d =10) + LR, taking only 10 important features for 75. The high-dimensional breast cancer data shows that no feature selection performs poorly by achieving the lowest performance value for two out of four metrics, with 47.83% for accuracy and 38.89% for sensitivity, in terms of specificity and precision. Hence, it can be concluded that no feature selection performs the worst for high-dimensional breast cancer data. The method that can be seen giving out a high and consistent value in terms of accuracy, and sensitivity, is hybrid IG (d =10)+LR, with 69.57% and 77.78%. hybrid IG (d =10)+LR is the best feature selection method when applied to high-dimensional prostate cancer data with the configuration of n = full sample, n =100, and n =75 with d =10 features. Thus, it can be concluded that hybrid IG (d =10)+LR is the best feature selection method for high-dimensional breast cancer data. It can be said that the filter feature selection method works well for both high-dimensional breast cancer and prostate cancer data.

Table 4. Performance measures of each filter method applied to top d=10 features and n=full sample, 100 and 75 for breast and prostate cancer data

Data	Performance Measures	n = full sample size			n = 100			n = 75		
		No Feature Selection	Hybrid IG, (d=22) +LR (%)	Hybrid chi-square (d=22) + LR (%)	No Feature Selection	Hybrid IG (d=22) +LR (%)	Hybrid chi-square (d=22) + LR (%)	No Feature Selection	Hybrid IG (d=22) +LR (%)	Hybrid chi-square (d=22) + LR(%)
Breast Cancer	Accuracy	56.86	82.35	80.39	60.00	66.67	66.67	47.83	69.57	65.22
	Sensitivity	48.72	89.74	87.18	72.22	88.89	88.89	38.89	77.78	61.11
	Specificity	83.33	58.33	58.33	41.67	33.33	33.33	80.00	40.00	80.00
	Precision	90.48	87.50	87.18	65.00	66.67	66.67	87.50	82.35	91.67
Prostate Cancer	Accuracy	58.06	83.87	77.42	30.00	96.67	90.00	56.52	95.65	91.30
	Sensitivity	33.33	80.00	80.00	10.00	100.00	100.00	33.33	88.89	100.00
	Specificity	81.25	87.50	75.00	40.00	95.00	85.00	71.43	100.00	85.71
	Precision	62.50	85.71	75.00	7.69	90.91	76.92	42.86	100.00	81.82

3.4. Discussion

To classify the model's output, several performance evaluations were used. Metrics such as the gap between the two sets, the accuracy of both sets, and the precision, recall, and F1-score value reveal differences between the training and testing sets. The gap between the accuracy, training, and validation measures narrows to a reasonable level during model training. All models can classify patients. However, this study finds that the hybrid information gain with logistic regression (hybrid IG+LR) provides the best results for n =100 and n =75 with top d =50, n=full sample and n =75 with top d =22 after training and testing data were applied on the breast cancer data. Suprisingly, hybrid IG+LR is the best method for all sample sizes with top d =10. Furthermore, hybrid IG+LR is still outperformed other methods when it was applied to high-dimensional prostate cancer data. Specifically, the analysis involved n =100 with top d =22 and n =100 and n =75 with top d =10.

It is interesting to note that hybrid IG+LR performs the best for high-dimensional breast and prostate cancer data since it achieves the best performance for all criteria. Even though other methods have been demonstrated to perform at par within some performance metrics, no feature selection achieved the highest specificity value when applied to high-dimensional breast cancer data. As a result, the optimum filter selection strategy is hybrid IG+LR for high-dimensional breast cancer data for 75 sample sizes considering top d =50 and d =22 important features. In addition, filter hybrid chi–square+LR gives a feasible feature selection solution for high-dimensional breast and prostate cancer data. Table 5 shows the top d =22 and top =10 important features of each filter method applied n = full sample for breast and prostate cancer data. As shown in Table 5, several the same features appeared in the top d =22 and top d =10 for hybrid IG+LR and hybrid chi-square+LR when these methods were applied to high-dimensional breast cancer data. However, different features were selected by hybrid chi-square+LR when this method was used for high-dimensional prostate cancer data with top d =22 and top d =10 important features and full sample sizes.

The top 50 and 22 features outperformed the other configurations, with the highest classification accuracies of 86.96% and 82.61%, respectively, after integrating the hybrid information gain and logistic function (hybrid IG+LR) with a sample size of 75. In conclusion, this study shows that reducing in sample size resulted in increased classification accuracy. This finding was supported by Eckstein *et al.* [31] and Arbabshirani *et al.* [32] who also gave out the same result. Hence, it can be assumed that sample size does influence the classification accuracy. This study revealed that sample sizes influenced the hybrid IG+LR and performance and hybrid chi–square+LR. So, deciding the best feature selection methods to be applied to high-dimensional data is still challenging. However, this study showed that the recommended feature selection method is hybrid IG+LR for high-dimensional cancer data.

Table 5. Top d=22 and top d=10 essential features of each filter method applied n=full sample for breast and prostate cancer data

Data	Hybrid methods	n =full sample size
Breast Cancer (*n*=168)	Hybrid IG(*d*=22) +LR	*x.g1CNS507, x.g1int1354, x.g7E05, x.g1int429, x.g1int372, x.g1int1131, x.g1int1662, x.g1int1702, x.g1int382, x.g1CNS28, x.g1int1130, x.g1int154, x.g1int659, x.g1int373, x.g1CNS229, x.g1int361, x.g2B01, x.g1int663, x.g1int895, x.g1int1414, x.g1int1220, x.g1int380*
	Hybrid IG(*d*=10) +LR	*x.g1CNS507, x.g1int1354, x.g7E05, x.g1int429, x.g1int372, x.g1int1131, x.g1int1662, x.g1int1702, x.g1int382, x.g1CNS28*
	Hybrid chi-square (*d*=22)+LR	*x.g1CNS507, x.g7E05, x.g1int1354, x.g1int372, x.g1int1131, x.g1int382, x.g1int1702, x.g1int1662, x.g1CNS28, x.g1int1130, x.g1int659, x.g1int429, x.g1int1220, x.g1int373, x.g1CNS229, x.g1int380, x.g1int369, x.g1int361, x.g2B01, x.g1int663, x.g3F01, x.g1int375*
	Hybrid chi-square (*d*=10)+LR	*x.g1CNS507, x.g7E05, x.g1int1354, x.g1int372, x.g1int1131, x.g1int382, x.g1int1702, x.g1int1662, x.g1CNS28, x.g1int1130*
Prostate Cancer (*n*=102)	Hybrid IG(*d*=22) +LR	*x.V7247, x.V10494, x.V6866, x.V6462, x.V9850, x.V8850, x.V4365, x.V5757, x.V6185, x.V9172, x.V6620, x.V8566, x.V9034, x.V4241, x.V205, x.V5566, x.V5835, x.V12148, x.V8058, x.V8724, x.V7557, x.V8965*
	Hybrid IG(*d*=10) +LR	*x.V7247, x.V10494, x.V6866, x.V6462, x.V9850, x.V8850, x.V4365, x.V5757, x.V6185, x.V9172*
	Hybrid chi-square (*d*=22)+LR	*x.V10494, x.V9172, x.V10956, x.V6185, x.V4365, x.V11818, x.V9850, x.V6462, x.V8566, x.V7247, x.V8103, x.V8850, x.V10260, x.V10138, x.V6620, x.V942, x.V3794, x.V9034, x.V12153, x.V7820, x.V8965, x.V299*
	Hybrid chi-square (*d*=10)+LR	*x.V10494, x.V7247, x.V9850, x.V4365, x.V5757, x.V6185, x.V6866, x.V6462, x.V9172, x.V8566*

4. CONCLUSION

This paper attempts to provide more detailed investigations regarding high-dimensional breast cancer and prostate data. This research compares filter feature selection methods in different sample sizes. Logistic regression with hybrid IG+LR demonstrates improvement in binary classification accuracy, especially for small sample sizes. It can be said that the filter feature selection method works well on high-dimensional breast cancer and prostate cancer data. The result is significant for many features and a small sample size. In addition, the sample size configuration affects the feature selection and classification performance. It resulted in integrating hybrid IG+LR with a sample size of 75, with the top 50 and 22 important features outperforming other configurations. Thus, this integration is expected to be used in other types of high-dimensional data. In the future, evaluating the multiclass classification from a different domain is recommended.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Research Nexus UiTM (ReNeU), and specifically thanks to research management centre (RMC), Universiti Teknologi MARA for financial support under the Malaysian Research Assessment (MyRA) Grant Scheme 600-RMC/GPM LPHD 5/3 (055/2021).

REFERENCES

[1] J. Liang, L. Hou, Z. Luan, and W. Huang, "Feature selection with conditional mutual information considering feature interaction," *Symmetry*, vol. 11, no. 7, Jul. 2019, doi: 10.3390/sym11070858.

[2] R. Sagban, H. A. Marhoon, and R. Alubady, "Hybrid bat-ant colony optimization algorithm for rule-based feature selection in health care," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 6, pp. 6655–6663, Dec. 2020, doi: 10.11591/ijece.v10i6.pp6655-6663.

[3] M. T. Tally and H. Amintoosi, "A hybrid method of genetic algorithm and support vector machine for intrusion detection," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 1, pp. 900–908, Feb. 2021, doi: 10.11591/ijece.v11i1.pp900-908.

[4] 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*, 2023. doi: 10.1063/5.0109918.

[5] Y. Guo, F.-L. Chung, G. Li, and L. Zhang, "Multi-label bioinformatics data classification with ensemble embedded feature selection," *IEEE Access*, vol. 7, pp. 103863–103875, 2019, doi: 10.1109/ACCESS.2019.2931035.

[6] R. Cekik and A. K. Uysal, "A novel filter feature selection method using rough set for short text data," *Expert Systems with Applications*, vol. 160, Dec. 2020, doi: 10.1016/j.eswa.2020.113691.

[7] T. W. Way, B. Sahiner, L. M. Hadjiiski, and H.-P. Chan, "Effect of finite sample size on feature selection and classification: a simulation study," *Medical Physics*, vol. 37, no. 2, pp. 907–920, Jan. 2010, doi: 10.1118/1.3284974.

[8] Y. B. Wah, N. Ibrahim, H. A. Hamid, S. A. Rahman, and S. Fong, "Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy," *Pertanika Journal of Science and Technology*, vol. 26, no. 1, pp. 329–340, 2018.

[9] E. Gravier *et al.*, "A prognostic DNA signature for T1T2 node-negative breast cancer patients," *Genes, Chromosomes and Cancer*, vol. 49, no. 12, pp. 1125–1134, Dec. 2010, doi: 10.1002/gcc.20820.

[10] D. Singh *et al.*, "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, no. 2, pp. 203–209, Mar. 2002, doi: 10.1016/S1535-6108(02)00030-2.

[11] 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, pp. 1–15, Feb. 2021, doi: 10.1155/2021/4832864.

[12] B. Vrigazova, "The proportion for splitting data into training and test set for the bootstrap in classification problems," *Business Systems Research Journal*, vol. 12, no. 1, pp. 228–242, May 2021, doi: 10.2478/bsrj-2021-0015.

[13] V. Nasteski, "An overview of the supervised machine learning methods," *Horizons.B*, vol. 4, pp. 51–62, Dec. 2017, doi: 10.20544/HORIZONS.B.04.1.17.P05.

[14] M. Abd-elnaby, M. Alfonse, and M. Roushdy, "A hybrid mutual information-LASSO-genetic algorithm selection approach for classifying breast cancer," in *Digital Transformation Technology*, 2022, pp. 547–560. doi: 10.1007/978-981-16-2275-5_36.

[15] S. L. Verghese, I. Y. Liao, T. H. Maul, and S. Y. Chong, "An empirical study of several information theoretic based feature extraction methods for classifying high dimensional low sample size data," *IEEE Access*, vol. 9, pp. 69157–69172, 2021, doi: 10.1109/ACCESS.2021.3077958.

[16] M. Rostami, S. Forouzandeh, K. Berahmand, M. Soltani, M. Shahsavari, and M. Oussalah, "Gene selection for microarray data classification via multi-objective graph theoretic-based method," *Artificial Intelligence in Medicine*, vol. 123, Jan. 2022, doi: 10.1016/j.artmed.2021.102228.

[17] B. Ghaddar and J. Naoum-Sawaya, "High dimensional data classification and feature selection using support vector machines," *European Journal of Operational Research*, vol. 265, no. 3, pp. 993–1004, Mar. 2018, doi: 10.1016/j.ejor.2017.08.040.

[18] H. Lu, J. Chen, K. Yan, Q. Jin, Y. Xue, and Z. Gao, "A hybrid feature selection algorithm for gene expression data classification," *Neurocomputing*, vol. 256, pp. 56–62, Sep. 2017, doi: 10.1016/j.neucom.2016.07.080.

[19] N. Ibrahim, "Variable selection methods for classification : application to metabolomics data," *University of Liverpool*, no. March, 2020.

[20] J. Cho, K. Lee, E. Shin, G. Choy, and S. Do, "How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?," *Prepr. arXiv.1511.06348*, Nov. 2015.

[21] G. Heinze, C. Wallisch, and D. Dunkler, "Variable selection - a review and recommendations for the practicing statistician," *Biometrical Journal*, vol. 60, no. 3, pp. 431–449, May 2018, doi: 10.1002/bimj.201700067.

[22] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, Jan. 2014, doi: 10.1016/j.compeleceng.2013.11.024.

[23] M. Al Khaldy, "Resampling imbalanced class and the effectiveness of feature selection methods for heart failure dataset," *International Robotics & Automation Journal*, vol. 4, no. 1, Feb. 2018, doi: 10.15406/iratj.2018.04.00090.

[24] R. Porkodi, "Comparison of filter based feature selection algorithms : An overview," *International Journal of Innovative Research in Technology & Science(IJIRTS)*, vol. 2, no. 2, pp. 108–113, 2014.

[25] C. De Stefano, F. Fontanella, and A. Scotto di Freca, "Feature selection in high dimensional data by a filter-based genetic algorithm," in *EvoApplications 2017: Applications of Evolutionary Computation*, 2017, pp. 506–521. doi: 10.1007/978-3-319-55849-3_33.

[26] S. DeepaLakshmi and T. Velmurugan, "Empirical study of feature selection methods for high dimensional data," *Indian Journal of Science and Technology*, vol. 9, no. 39, Oct. 2016, doi: 10.17485/ijst/2016/v9i39/90599.

[27] A. Pedersen *et al.*, "Missing data and multiple imputation in clinical epidemiological research," *Clinical Epidemiology*, vol. 9, pp. 157–166, Mar. 2017, doi: 10.2147/CLEP.S129785.

[28] U. Grömping, "Practical guide to logistic regression," *Journal of Statistical Software*, vol. 71, no. 3, 2016, doi: 10.18637/jss.v071.b03.

[29] M. Mafarja and S. Mirjalili, "Whale optimization approaches for wrapper feature selection," *Applied Soft Computing*, vol. 62, pp. 441–453, Jan. 2018, doi: 10.1016/j.asoc.2017.11.006.

[30] 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, Jun. 2017, doi: 10.1016/j.neucom.2017.02.057.

[31] F. Eckstein *et al.*, "Effect of training set sample size on the agreement, accuracy, and sensitivity to change of automated U-net-based cartilage thickness analysis," *Osteoarthritis and Cartilage*, vol. 29, pp. S326--S327, Apr. 2021, doi: 10.1016/j.joca.2021.02.427.

[32] M. R. Arbabshirani, S. Plis, J. Sui, and V. D. Calhoun, "Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls," *NeuroImage*, vol. 145, pp. 137–165, Jan. 2017, doi: 10.1016/j.neuroimage.2016.02.079.

BIOGRAPHIES OF AUTHORS

Siti Sarah Md Noh received the Diploma in Statistics in 2018 and Bachelor of Science (Hons) (Statistics) in 2021 at Universiti Teknologi MARA. She is now currently a student, undergoing Master of Science 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 the Bachelor of Science (Hons) (Statistics), Master of Science Applied Statistics from Universiti Teknologi MARA in 2013 and 2014, and PhD 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 Media, Universiti Teknologi MARA, Malaysia and she also is currently 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 experienced in teaching A-level students for UK, US and Germany bounds at INTEC Education College in 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 PhD 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 currently a senior lecturer in Statistics at the School of Mathematical Sciences, Universiti Teknologi MARA and an external supervisor register at the School of Mathematical Sciences, The University of Adelaide. His research interest focuses on data analytics, time series, R computing, quantitative finance and vulnerable populations. He can be contacted at email: maha@uitm.edu.my.

Marina Yusoff is currently a senior fellow researcher 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 in 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 recent 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 marina998@uitm.edu.my.