Published by
NIGERIAN SOCIETY OF PHYSICAL SCIENCES
Available online @ https://journal.nsps.org.ng/index.php/jnsps

J. Nig. Soc. Phys. Sci. 7 (2025) 2273

Journal of the Nigerian Society of Physical Sciences

A feature selection and scoring scheme for dimensionality reduction in a machine learning task

Philemon Uten Emmoh[a,*], Christopher Ifeanyi Eke[b], Timothy Moses[b]

[a]*Department of Computer Science, Federal University Wukari, P.M.B 1020, Katsina-Ala Road, Wukari, Taraba State, Nigeria*
[b]*Department of Computer Science, Federal University of Lafia, P.M.B 146, Lafia, Nasarawa State, Nigeria*

Abstract

The selection of important features is very vital in machine learning tasks involving high-dimensional dataset with large features. It helps to reduce the dimensionality of a dataset and improve model performance. Most of the feature selection techniques have restrictions on the kind of dataset to be used. This study proposed a feature selection technique based on statistical lift measure to select important features from a dataset. The proposed technique is a generic approach that can be used in any binary classification dataset problem. The technique successfully determined the most important feature subset and outperformed the existing techniques. The proposed technique was tested on lungs cancer dataset and happiness classification dataset. The effectiveness of the proposed technique in selecting important features subset was evaluated and compared with other existing techniques, namely Chi-Square, Pearson Correlation and Information Gain. The proposed and the existing techniques were evaluated on five machine learning models using four standard evaluation metrics such as accuracy, precision, recall and F1-score. The experimental results of the proposed technique on lung cancer dataset shows that logistic regression, decision tree, adaboost, gradient boost and random forest produced a predictive accuracy of 0.919%, 0.935%, 0.919%, 0.935% and 0.935% respectively, and that of happiness classification dataset produced a predictive accuracy of 0.758%, 0.689%, 0.724%, 0.655% and 0.689% on random forest, k-nearest neighbor, decision tree, gradient boost and cat boost respectively, which outperformed the existing techniques.

DOI:10.46481/jnsps.2025.2273

Keywords: Algorithm, Dataset, Dimensionality reduction, Feature selection

Article History :
Received: 26 July 2024
Received in revised form: 04 November 2024
Accepted for publication: 05 November 2024
Published: 14 December 2024

Communicated by: O. Akande

1. Introduction

Machine learning researchers and engineers face significant challenges when processing high-dimensional data. In high-dimensional data, there are many features to be detected. There may be some unnecessary and unimportant features [1]. Several techniques have been developed to address the problem of reducing irrelevant variables when performing data mining, machine learning, and other modelling tasks. Feature selection is a method used to select subsets of original features by eliminating unimportant or redundant features while maintaining the original qualities of the features that aid in visualizing and comprehending [2]. According to Peng *et al.* [3], feature selection aids in data comprehension, minimizes the need for computation, mitigates the consequences of the dimensionality curse, and enhances the predictive capabilities of models.

The purpose of feature selection is to come up with sub-

*Corresponding author: Tel.: +234-803-520-2835.
Email address: philemon@fuwukari.edu.ng (Philemon Uten Emmoh)

sets of features from the input features that sufficiently represent the feature space [4]. According to Cherrington *et al.* [5], the choice of whether to keep important variables or remove unimportant ones can significantly affect how well a learning algorithm performs in the field of knowledge discovery. Before beginning any analysis, it is essential to carefully consider any potential consequences of these decisions. High-dimensional dataset can contain a lot of unimportant, noisy, and redundant information which may degrade the performance of learning models. So, not only the large quantities of high dimensional datasets are not profitable, but they have also brought great challenges [6, 7]. According to Meinshausen [8], it is not all the variables of a dataset that contain information of interest and relevance in the subject matter. Including irrelevant variables in a predictive model could negatively affect some evaluation metrics such as the Alkaike Information Criteria (AIC), variance, and degrees of freedom [9, 10]. Consequently, variable or feature selection is required to avoid including the less important variables in the models. Apart from the ability to produce a reduced number of features, a good feature selection algorithm should be easy to implement and should require minimal system resources such as space and memory [11].

Feature selection methods are normally divided into three main categories: filtering, wrapping, and embedded [5]. Filter methods often evaluate both the dataset's subset and its results. These approaches base their assessment on the relevance of the subset innate qualities and relationships with one another rather than machine learning algorithms. Some filtering techniques employ information gain, correlation or distance metrics to determine the relationship between each predictive feature and the target feature [12]. Other filter methods include the Pearson's correlation, Fisher score, t-statistics, information gain, ANOVA, variance threshold, chi-square, and many more. Furthermore, the wrapper feature selection techniques measure the effectiveness of the selected classifier algorithm as a metric to help choose the appropriate feature subset [13]. Mehmood *et al.* [14] explains that the wrapper methods incorporate part of the learning model to select feature subsets that iteratively train and evaluate models for accuracy. The subset that produces the best accuracy is returned as the best features for the dataset in question. The embedded methods select features based on the outcomes of the single machine learning model and feature importance [15].

They have in-built penalization procedures to control overfitting, such as the Ridge and Lasso regressions. After selecting the best features using any of the feature selection type mentioned above, a classifier can be employed to accurately classify the data after selecting the important features, or a regression model can be built to estimate the correlations between features. Most feature selection techniques is to perform feature selection with high accuracy within a short period of time so that it can be used as an input of predictive analysis in many fields such as clustering and classification [16]. Feature selection evaluation is essential to ascertain the efficacy of the feature selection technique. It basically involves evaluating the features that have been selected, and in turn, plays a vital role in the process of selecting critical features for machine learning tasks. During evaluation, it is often required to compare a newly proposed features selection approach with an existing one. The evaluation tasks would have been simple if the ground truth or the most important features had been known. However, this is not the case for data from the real-world. Real-world data lacks a ground truth. When many features are employed, classification or regression problems will have high time complexity and low performance. However, when reduced subsets and the most useful features are used, classification or regression problems will have low time complexity and good performance. Eke [17] investigated the impacts of feature selection techniques when dealing with high-dimensional data, and one of the impacts is by reducing the dimensionality space of the dataset for machine learning model to produce an optimal prediction accuracy in a lesser computational time.

This research introduces a filter selection technique named proposed algorithm (PA) to determine a predictive accuracy of features in a dataset using some classifiers such as logistic regression, decision tree, adaboost, gradient boosting, random forest, cat boost and k-nearest neighbours. The results of the proposed algorithm (PA) algorithm were compared with three existing filter selection algorithms namely chi-square, Pearson correlation and information gain to ascertain the efficacy of the proposed algorithm in accuracy prediction.

We present the summary of the contributions made by this study as follows: The novelty in the proposed algorithm compared to existing filter algorithms is attributable to the characteristics of lift which is the underlying measure implemented by the new algorithm. The proposed feature selection algorithm is a generic algorithm that can be implemented in any research area such as medical, finances, education and many other areas for binary classification problems. The proposed algorithm outperforms the existing techniques in evaluating important features subsets for binary classification modelling tasks.

The remaining parts of the paper are organized as follows: Section 2 provides the research review of related works. Section 3 describes the methods and material used in this study. Section 4 explains the results and discussion. Conclusion of the study is expressed in section 5

2. Review of filter variable selection methods and other related works

Filter methods select the features of a dataset without depending on any machine learning model [18]. They are executed as part of the preprocessing activity to rank features according to the order of importance. After the ranking, the modeler chooses the variables to be used in modelling based on their ranking score [4]. One of the merits of filter methods is their simplicity and independence from the type of classifier used. However, a major limitation of this method is lack of interaction with the classifier and the collinearity that might exist among features [19]. Some existing filter feature selection techniques are described in this section.

2.1. Chi square test

Chi square test checks independency between two events. The two events X,Y are defined to be independent if $P(XY) = P(X)P(Y)$ or $PP(X/Y) = P(X)$ and $P(Y/X) = P(Y)$ [20]. This is used in feature selection to test whether the occurrence of a specific input and the occurrence of a specific class are independent. When comparing two or more features, the feature with a higher Chi square score is considered the best for modelling. Chi square formula is given in equation (1). According to Pavya and Srinivasan [20] and Gajawada [21]:

$$X^2 = \sum_{i=1}^{r} \sum_{j=i}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}, \tag{1}$$

where X^2 is chi square, O is observed value, E is expected value. The last concern is an ensemble, as the authors describe, each feature selection method has its merits and demerits.

2.2. Pearson's correlation

Equation (2) is used to assess the significance of a feature and the equation is evaluated by Sahin and Chandrashekar [22]:

$$P_i = \frac{\mathrm{Cov}(X_i, Y)}{\sqrt{\mathrm{Var}(X_i) \times \mathrm{Var}(Y)}}, \tag{2}$$

where X_i and Y are, respectively, the i labeled variable and the outcome, $Cov()$and $Var()$are the correlation and variation respectively. A feature that produces a higher score is considered the best.

2.3. The t-statistic

The t-statistic is another means of feature importance scoring. It is computed using the formula in equation (3) Bodur and Atsa'am [4].

$$t_i = \frac{\left|\overline{X_1}(i) - \overline{X_0}(i)\right|}{\sqrt{\frac{d_1^2(i)}{n_1} + \frac{d_0^2(i)}{n_0}}}, \tag{3}$$

where $d_1, d_0, \bar{X}_1, \bar{X}_0$ is the standard deviation of the ith feature with class 1, the standard deviation of the i -th feature with class 0, the mean of the ith feature with class 1, and *the* mean of the feature with class 0 respectively; n_0 and n_1 are the overall records for classes 0 and 1, respectively.

2.4. Feature selection with risk ratios

A study by Bodur and Atsa'am [4] proposed an algorithm for variable ranking that makes use of risk ratios to score predictor variables. The formula for the algorithm is given in equation (4):

$$I_j = \left(\frac{\sum_{i=1,j=1}^{m,n} \mathrm{Wij}}{\sum_{i=1,j=1}^{m,n} \mathrm{Wij} + \sum_{i=1,j=1}^{m,n} \mathrm{Xij}} \right) \cdot \left(\frac{\sum_{i=1,j=1}^{m,n} \mathrm{Yij} + \sum_{i=1,j=1}^{m,n} \mathrm{Zij}}{\sum_{i=1,j=1}^{m,n} \mathrm{Yij}} \right), \tag{4}$$

where I_j is the importance score of the jth variable, W_{ij} are total observations with input = 1 and output = 1, X_{ij} are total observations with input = 1 and output = 0, Y_{ij} are total observations with input = 0 and output = 1, Z_{ij} are total observations with input = 0 and output = 0, m is the number of observations in the dataset while n is the number of predictor variables. For every predictor, the algorithm separately sums up the total number of observations in each of the four categories and evaluates the importance score of each predictor.

2.5. Other related works

A study by Pooja *et al.* [16] used another filter feature selection technique called Point-Biserial correlation feature selection which separates features into two subsets as relevant and irrelevant by computing the mean and deviation. The technique was used to select the relevant features for weather prediction from the dataset. The correlation coefficient separates the relevant features and irrelevant features to improve the feature selection accuracy and minimizes the time complexity. The performance of their technique was determined by comparing it with the existing method and their result shows that it is more accurate than the existing methods.

A study by Atsa'am [18] performed an experiment with the odds ratios variable selection technique to evaluate important features from the unimportant features in four datasets namely: Arrests For Marijuana Possession Data (Marijuana), Risk Factors Associated With Low Infant Birth Weight Data (Birth Weight), Spam e-Mail Data (Spam) and Beaver Body Temperature Data (Beaver). From the results obtained, marijuana and birth weight datasets, odds ratio = 73%, fisher score = 72%, pearson's 72%, varImp 71 and odds ratio = 61%, fisher's score=74%, pearson's = 61%, and varImp = 74% respectively. While spam dataset, having odds ratio = 61%, fisher's score = 60%, pearson's 60%, varImp = 60%, Beaver dataset is having odds ratio = 81%, fisher score = 80%, pearson's = 80%, varImp = 79%. The classification accuracies obtained from the experimental dataset was graphically represented. The algorithm was able to rank the features based on their importance but no algorithmically threshold separate the important features and the unimportant features.

Solorio-Fernández et al.[23] proposed a new method for selecting relevant and non-redundant features in supervised mixed datasets. The method combined Spectral Feature Selection and Information-theory based redundancy analysis, and it is called RnR-SSFSM. The Spectral Feature Selection was used to obtain a feature ranking of relevant features, then the RnR-SSFSM method was used to select a feature subset of relevant and non-redundant features through pairwise correlation analysis. The authors experimented with their technique on several real-world datasets using SVM, kNN and Random-Forest classifiers. Their results showed that RnR-SSFSM method generally obtained better results than other supervised feature selection methods; it was able to select feature subsets with low redundancy.

Also, Singer *et al.* [24] proposed a new model for measuring information-gain, called Weighted Information-Gain (WIGR). This model employed a weighted entropy function which considered various target class values. They tested this method on 12 datasets with less than 100 features, ranging from

7 to 32. It is an intriguing approach to measuring information-gain with promising applications in the field of data analysis.

Nurhayati *et al.* [25] carried out an experiment using chi square to evaluate the importance of using chi square in selecting important features from 700-training data and 30 test data were that were obtained from Corpus v1.0 Indonesian Movie Review. Sentiment analysis of documents was both tested with and without a chi-Square feature selection. An evaluation metrics of accuracy, precision, and recall were examined. With the chi-square test, the analysis of the sentiment yielded an accuracy, precision, and recall of 93.33%, 93.33%, and 93.33%, respectively. From these findings, the choice of using the chi-square test has an impact on assessing sentiment documents when applied with the Naïve Bayes model.

A study by Sun *et al.* [26] proposed an improved Fisher score model based on mutual information combined with second-order correlation between labels to preprocess multilabel data and optimize the performance of multilabel classification and its corresponding strategies. Then, a new classification margin based MNRS model was provided in multilabel neighborhood decision systems. In the study, a hybrid filter-wrapper feature selection techniques using an improved Fisher score model and a new MNRS model based on adaptive granularity was proposed. Their work first combined the mutual information and the second order correlation between labels to make a slight improvement to the conventional Fisher score method to consider not only the correlation between labels, but also multilabel datasets. Second, the self-classification margins of each sample were computed through a subset of the nearest homogeneous or heterogeneous samples, then a novel neighborhood radius developed based on the designed classification margin was presented. The study improved on the defects of the conventional MNRS model and proposed a new MNRS model to study the uncertainty measures of the dependency degree and significance of features. Their results show that the proposed algorithm was able to select feature subsets with small scale and strong classification ability for multilabel datasets, which has certain advantages over competitive multilabel feature selection technique. However, some problems were identified as follows: the improved Fisher score model ignored the correlation among features, the second-order correlation between labels was only considered when higher than second order on some multilabel datasets, and the computational complexity of their algorithm had no advantage over some state-of-the-art algorithms.

Improved mRMR is a method that was presented by Xie *et al.* [27] to enhance MRMR based on feature subsets, minimize the dimension of feature sets, and improve the performance of classification of samples by selecting important feature from the dataset. The authors used Pearson correlation coefficient and mutual information to measure the importance of a single feature by introducing an adjusted weights of two measurement criteria to rank the features of the candidate feature subsets. They calculated the features by using an incremental search method to determine the best features while eliminating the unimportant features. The conducted their experiment on seven datasets and their method effectively reduced the dimensionality of the dataset by eliminating the unimportant and the redundant features in a very minimal model training time and prediction. The results obtained from the improved method on the seven datasets were compared with four other existing techniques, including mRMR, information gain, symmetrical uncertainty, gain ratio, and relief. The Improved mRMR typically performs better than the other methods can effectively determine the best feature subset and enhances the performance of the classification modeling task.

Chen *et al.* [7] performed an experiment using Pearson correlation coefficient to reduce the dimensionality of some datasets which were obtained from the UNSW-NB15 datasets. The UNSW-NB15 dataset has 49 total dimensions. Pearson correlation coefficient as a statistical measure can only use a numerical data by which the dataset was converted to continuous numerical features while some of the features in the dataset are discrete features. They did some preprocessing by converting the discrete features into continuous numerical features by assigning numeric value to replace the noun. They used numerical normalization method to normalize the data for better evaluation as they used [0,1] as the range for the normalized value. They used calculated Pearson correlation coefficient between the basis features in the training set dataset to obtain the important features. They conducted Principal Component Analysis on the data set to get a selection of contrast features to compare with commonly used data dimensionality reduction techniques. Comparison of the results before and after feature selection revealed that while PCA feature selection's detection rate decreased and its false alarm rate increased, the method proposed in their paper's false alarm rate and detection rate were essentially unchanged, supporting the theoretical role of feature selection. In addition, the detection efficiency has increased as a result of effective dimensionality reduction of the data. This experiment significantly increased the detection efficiency while cutting the detection time by around 46%.

A filtering technique known as least loss technique was created by Thabtah *et al.* [28] to remove irrelevant characteristics from a dataset so that only the crucial features remain for high accuracy prediction performance. The approach selected the better rated features for classification modelling by ranking all the features according to feature relevance in the ascending order. To evaluate the effectiveness of the algorithms, they ran tests and found out that the algorithm can pick out important features from a dataset while removing the irrelevant features.

Eke [29] developed a multi-feature framework that employed five machine learning algorithms for sarcasm identification. Among the five learning algorithms, namely, support vector machine, decision tree, k-nearest neighbor, random forest and logistic regression, random forest attained the highest precision score of 94.7% as compared with the other models.

2.6. Summary of related works

A few related studies carried out by researchers are summarized in Table 1.

Table 1: Summary of related studies.

S/No.	Authors & reference	Method	Aim	Task	Drawback
1	Pooja *et al.* [16].	Point Biserial Correlated Feature Selection (PBCFS)	From the experiment and comparison carried out, PBCFS provides good accuracy better than the existing method, such as the hybrid neural model and SVR method by the following respective results 75%, 50% and 50%.	Classification and regression	Also, this method is more suitable to be applied on weather forecasting domain
2	Atsa'am, [18].	Relative odds or also called odds ratio (OR)	The algorithm operates independently of any machine learning model. This method improves the performance of Machine Learning by effectively eliminating the unimportant features.	Classification task	The algorithm is most suitable in heart care dataset
3	Bodur *et al.* [4].	Risk Ratios	The Relative Ratio (RR) Method was able to select important features effectively and the results obtained were better than other filters and wrapper methods.	Classification task	Also, the RR method can only be executed on any healthcare dataset with numeric data
4	Solorio-Fernández *et al.* [23].	Spectral feature selection and information-theory based redundancy analysis	It's a good filter selection method to be used in terms of mixed data.	Classification task	It's important to take into account the statistical test results, but ultimately their method appears to have a clear advantage in terms of accuracy when working with mixed data.
5	Singer *et al.* [24].	Weighted Information Gain	The classification features do not have to follow any natural ordering as compared with other ordinal classification. They conducted a numeric study that is based on well-known ordinal datasets with high level of non-monotonic noisy data.	Classification	This is not suitable for a high dimensional dataset. Also Neglect to take into account how the features interact and are relevant. An issue with overfitting could arise.
6	Nurhayati *et al.* [25].	Chi2	The method was able to select important feature for better classification prediction accuracy		This can only be applied on a low dimensional dataset
7	Sun *et al.* [26].	Fisher score based on Mutual Information and wrapper	The method is mostly used on a low dimensional dataset.	Classification	This approach doesn't show how the features relate to one another.
8	Xie *et al.* [27].	Improved maximal relevance and minimal redundancy method (IMRMR)	Their method was able to select the optimal features in a very minimal computational time during the classification process.	Classification	The method does not consider the independence between features. And that can also select unimportant features too.
9	Chen *et al.* [30].	Pearson correlation coefficient	Their work does not only reduce the dimensionality in the dataset but also revealed the relationship between the data.	classification	The experiment yields a positive result but cannot distinguish between independent and dependent variables
10	Thabtah *et al.* [31].	Least lost	Their work was aimed to reduce the dimensionality of a dataset by isolating the irrelevant features.	Classification	The method was evaluated and compared with other existing techniques using only one classifier.

3. Materials and methods

3.1. Design of the proposed algorithm (PA)

According to Vu *et al.* [32], lift is the ratio of the joint occurrence of an antecedent, X, and a consequent, Y, to the product of the marginal occurrences of X and Y. It assesses the relationship between X and Y: X and Y are independent when lift is said to be equal to 1. X and Y are positively associated when Lift > 1 while X and Y are negatively associated when lift is < 1. Greater relationship between X and Y is implied by lift values that are farther from 1.

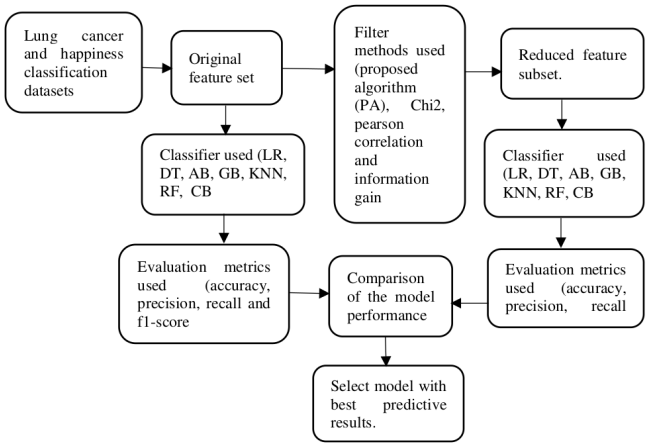


Figure 1: Framework of the proposed algorithm with three existing techniques.

To achieve feature importance scoring, the proposed algorithm will depend on a statistical measure known as lift. Lift is a metric used in data mining to measure the relationship between the variables X and Y; whereby any value greater than 1 denotes a positive relationship. Based on Vu *et al.* [32], the formula for lift is given in equation (5):

$$\text{Lift} = \frac{aN}{(a+c)(a+b)}, \tag{5}$$

where a is observation total when input = 1 and output = 1, b is observations total when input = 1 and output = 0, c is Observations total when input = 0 and output = 1, N is observations total in the dataset.

The proposed algorithm (PA) was derived from the lift equation (5) which is mathematically represented as shown in equation (6):

$$\text{PA}\,[j] = \frac{\left(\sum_{i=1,j=1}^{m,n} \text{Aij}\right) \times N}{\left(\sum_{i=1,j=1}^{m,n} \text{Aij} + \sum_{i=1,j=1}^{m,n} \text{Cij}\right)\left(\sum_{i=1,j=1}^{m,n} \text{Aij} + \sum_{i=1,j=1}^{m,n} \text{Bij}\right)}, \tag{6}$$

where Proposed Algorithm (PA)[j] is the importance score for j^{th} feature (j = 1, . . . , n), the total number of observations when the input = 1 and the output = 1 is = *Aij*, the total number of observations when the input = 1 and the output = 0 is = *Bij*, the total number of observations when the input = 0 and the output = 1 is = *Cij*, and N is the sum of observations in the dataset. The design algorithm is shown in algorithm 1.

3.2. Proposed algorithm design framework

We employed a methodical approach to define the components of the experiments executed in this research and the datasets deployed to evaluate the effectiveness of the proposed feature selection technique on the three existing techniques. We provide more details in the proposed framework shown in Figure 1.

3.3. Evaluation metrics

Evaluation performance metrics were taken on each of the classifier to determine which of the algorithms performed better result than the other. Accuracy, precision, recall and F1-score measures were employed on the proposed algorithm and on the three existing algorithms which are chi-square, Pearson correlation, and information gain. According to Iwendi *et al.* [33], the definitions of these evaluation metrics are based on True Positive (TP), False Positive (FP), False Negative (FN) and True Negative (TN). The formulas are:

Accuracy is used to measure how many instances that are correctly classified. It is formulated in equation (7):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}. \tag{7}$$

Precision is used to compute the true-positive instances in relation to the false-positive instances. The metric measure can be seen in equation (8):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{8}$$

Recall is used to compute true-positive instances in relation to false-negative instances. This is represented in equation (9):

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{9}$$

F1-score is used to compute the average of recall and precision criterion; it is mathematically represented in equation (10):

$$F1 - \text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

3.4. Datasets

Most of the datasets contain noisy and irrelevant features that need to be removed. This cleaning process is important because, once done, the noisy, duplicates and irrelevant features will be removed leaving only important features to be used for modelling [34]. Several datasets were utilized to exhibit the efficacy of the suggested method in feature selection. All the datasets were obtained from Kaggle repository https://www.kaggle.com. All the datasets have a binary outcome. These datasets are enumerated below.

Lung cancer dataset:

Lung cancer dataset is a public dataset obtained from (https://www.kaggle.com/datasets/jillanisofttech/lung-cancer-detection) to helps people predict their cancer risk to take necessary decision based on status of the cancer risk. The features are all numeric except the gender and lung cancer (class). The gender attribute is labelled for male and female as male: 1, female: 0. Meanwhile the class attribute which is labeled as "YES" if it has lung cancer or "NO" appropriately 1, "NO": 0. The dataset contains 309 instances with 15 attributes with the class attribute in exception.

Table 2: Design algorithm for the proposed feature selection technique.

Algorithm 1:
First phase. Converting Dataset to binary
// Listing 1. Is the phase that transforms every input to binary
2 While j <= n //all columns in the dataset are counted
3 While i <= m //all rows in the dataset are counted
4 Compute average value for each column, store as $[Avg]_j$
5 IF D[i,j] < $[Avg]_j$ Then
6 D[i,j] = 0 //data points less than column average converted to 0
7 ELSE D[i,j] = 1 // Any data points that are higher than or equal to the column average are converted to 1.
8 END IF
9 i = i+1
10 j = j+1
11 End While
12 End While
Second phase counts the instances of a11, b10, c01
13 //Listing 2. This step counts a, b, c for each predictor
14 D = Array[1..m][1..n] As Integer //2-dim array of rows/columns
15 Class = Array[1..m] As Integer //1-dim array for class
16$Aj = 0 : Bj = 0 : Cj = 0$ As Integer //initialize sums of a, b, c
17 While j <= n // variable descriptions for the column index
18 While i<= m // variable descriptions for the row index
19 While y <= m // For the row index, apply the class variable
20 IF i = y THEN // index comparison among the input and the output
21 If D[i,j]=1 AND Class[i,j] = 1 THEN
22 $Aj = Aj + 1$ //counts a_{11}
23 END IF
24 IF D[i,j] = 1 AND Class[i,j] = 0 THEN
25 $Bj = Bj + 1$ //counts b_{10}
26 END IF
27 IF D[i,j] = 0 AND Class[i,j] = 1 THEN
28 $Cj = Cj + 1$ //counts c_{01}
29 END IF
30 END IF
31 j = j + 1
32 i = i + 1
33 y = y + 1
34 End While
35 End While
36 End While
Third phase compute the Proposed Algorithm (PA) for each of the columns
37 //Temporary variables
38 $upperProduct_j$, temp1, temp2, $lowerProduct_j$, using Integer: PA[j], as True
39 While j <= n
40 $upperProduct_j$= Aj × m
41 temp1 = A_j+C_j
42 temp2 = A_j+B_j
43 $lowerProduct_j$= $temp1 \times temp2$
44 PA[j] = $\frac{\text{upper Product}_j}{\text{lower Product}_j}$ //Computes proposed algorithm (PA) for each attr
45 j = j+1
46 End While
47 Print PA[j] for each variable

Happiness classification dataset:

This dataset is also a public dataset obtained from (https://www.kaggle.com/datasets/priyanshusethi/happiness-classification-dataset) based on a survey conducted to rate different metrics to predict if an individual is happy or unhappy in a city, he or she is residing in is extremely very important for someone's overall quality of life. This dataset contains 143 instances with 6 attributes with the class in exception. The binary response variable named "happy" is represented as 1 while "unhappy" is represented as 0. All the values in the dataset are numeric.

Table 3: The proposed algorithm on lung cancer dataset.

Classifier	ACC (%)	PREC (%)	REC (%)	F1 (%)
Logistic Regression	0.919	0.915	1.0	0.955
Decision Tree	0.935	0.962	0.962	0.962
ada boost	0.919	0.915	1.0	0.955
Gradient Boost	0.935	0.962	0.962	0.962
Random Forest	0.935	0.962	0.962	0.962

Table 4: Chi square on lung cancer dataset.

Classifier	ACC (%)	PREC (%)	REC (%)	F1 (%)
Logistic Regression	0.903	0.913	0.981	0.946
Decision Tree	0.887	0.927	0.944	0.935
ada boost	0.887	0.898	0.981	0.938
Gradient Boost	0.903	0.928	0.962	0.945
Random Forest	0.919	0.929	0.981	0.954

Table 5: Pearson Correlation on lung cancer dataset.

Classifier	ACC (%)	PREC (%)	REC (%)	F1 (%)
Logistic Regression	0.887	0.898	0.981	0.938
Decision Tree	0.919	0.945	0.962	0.954
ada boost	0.903	0.913	0.981	0.946
Gradient Boost	0.887	0.927	0.944	0.935
Random Forest	0.903	0.928	0.962	0.945

Table 6: Information gain on lung cancer dataset.

Classifier	ACC (%)	PREC (%)	REC (%)	F1 (%)
Logistic Regression	0.887	0.898	0.981	0.938
Decision Tree	0.870	0.942	0.907	0.924
ada boost	0.903	0.913	0.981	0.946
Gradient Boost	0.887	0.912	0.962	0.936
Random Forest	0.919	0.945	0.962	0.954

4. Results and discussion

4.1. Experiments and results

The first option is to preprocess the features in the datasets when necessary. The second is to rank all the features in the dataset according to their importance using any feature selection techniques. The third option is to continue adding and dropping features with reasonably high-ranking values until the best threshold that evaluate the best improved model performance is chosen. A proposed feature selection algorithm known as proposed algorithm (PA) is designed to test the predictive accuracy of a model and to compare its predictive accuracy with three existing feature selection techniques. These existing techniques are Chi Square, Pearson Correlation and Information Gain. The best subsets generated from the dataset when ranked by the proposed algorithm and the other three existing techniques were selected for modelling. Python 3.9 programming language was employed to implement the proposed algorithm and performed the predictive accuracy comparison with the three existing techniques.

Results of the proposed algorithm and the three existing algorithms on lung cancer dataset

Table 3 presents the experimented results of the proposed algorithm on lung cancer dataset. As it can be seen in Table 3, five popularly known classifiers was used, and among the classifiers, in terms of accuracy, precision and F1-score, the decision tree, gradient boost and random forest attained the same results with an accuracies 0.935%, precisions of 0.962 and F1-score of 0.962%, which are considered to be higher as compared with that of the logistic regression and ada boost which are also having the same results with an accuracy of 0.919% and precision of 0.915%. Meanwhile, in terms of recalls, the logistic regression and ada boost attained same results with 1.0% each as considered to be higher than the decision tree, gradient boost and random forest.

Table 4 presents the experimented results of the chi square on lung cancer dataset. As it can be seen in Table 4, the five same popularly known classifiers used in experimenting the proposed algorithm was also used in experimenting the existing technique known as chi square on the same lung cancer dataset. The results show that, the random forest attained the highest accuracies, precisions and F1-scores of 0.919%, 0.929% and 0.954% respectively. But having the same performance with the logistic regression and ada boost with recalls of 0.981% respectively, which are higher than the decision tree and gradient boost in terms of recall.

Table 5 presents the experimented results of the Pearson correlation still on the lung cancer dataset. As it can be seen in Table 5, the same five popularly known classifiers used in experimenting the proposed algorithm was also used in experimenting the existing feature selection technique known as Pearson correlation on the same lung cancer dataset. The results show that, the decision tree attained the highest accuracies, precisions and F1-scores with 0.919%, 0.945% and 0.954% respectively. But scored lower in terms of recall than the logistic regression and the ada boost with recall of 0.981% respectively while having the same 0.962% recall with random forest

Table 6 presents the experimented results of information gain still on the lung cancer dataset. As it can be seen in Table 6, the same five popularly known classifiers used in experimenting the proposed algorithm was also used in experimenting the existing feature selection technique known as information gain on the same lung cancer dataset as to compare the result of the proposed algorithm with information gain in due cost. The results show that, the random forest attained the highest accuracies, precisions and F1-scores with 0.919%, 0.945% and 0.954% respectively. But scored lower in terms of recall than the logistic regression and the ada boost with recall of 0.981% each while having the same 0.962% recall with gradient boost and at the same time higher than decision tree that attained 0.907% recall.

Results of the proposed algorithm and the three existing algorithms on happiness classification dataset

Table 7 presents the experimented results of the proposed algorithm on happiness classification dataset. As it can be seen in Table 7, five popularly known classifiers was used, which are, random forest, K-nearest neighbour, decision tree, gradient

Table 7: Proposed algorithm (PA).

	ACC (%)	PREC (%)	REC (%)	F1 (%)
Random Forest	0.758	0.714	0.937	0.810
K-Nearest Neighbors	0.689	0.652	0.937	0.769
Decision Tree	0.724	0.7	0.875	0.777
Gradient Boost	0.655	0.636	0.875	0.736
Cat Boost	0.689	0.666	0.875	0.756

Table 8: Chi square on happiness classification dataset.

	ACC (%)	PREC (%)	REC (%)	F1 (%)
Random Forest	0.586	0.625	0.625	0.625
K-Nearest Neighbors	0.620	0.619	0.812	0.702
Decision Tree	0.586	0.625	0.625	0.625
Gradient Boost	0.586	0.611	0.687	0.647
Cat Boost	0.586	0.611	0.687	0.647

Table 9: Pearson correlation on happiness classification dataset.

	ACC (%)	PREC (%)	REC (%)	F1 (%)
Random Forest	0.689	0.684	0.812	0.742
K-Nearest Neighbors	0.620	0.619	0.812	0.702
Decision Tree	0.586	0.625	0.625	0.625
Gradient Boost	0.586	0.611	0.687	0.647
Cat Boost	0.586	0.611	0.687	0.647

boost and cat boost. Among these classifiers, random forest attained the best accuracy, precision and F1-score with 0.758%, 0.714% and 0.810% respectively. Meanwhile, having the same recall of 0.937% with K-nearest neighbour which is still greater than the recall of decision tree, gradient boost and cat boost.

Table 8 presents the experimented results of chi square on happiness classification dataset. As it can be seen in this Table 8, the same five popularly known classifiers that was used to experiment the proposed algorithm was also used here on the same dataset to compare the results of the chi square with the proposed algorithm in our later discussion. Among these classifiers, the K-nearest neighbor attained the best accuracy, recall and F1-score with 0.620%, 0.812% and 0.702% respectively. Meanwhile, having the best 0.619 precision as compared with gradient boost and cat boost which are having the same precision of 0.611% each but lower than the 0.625% precision of random forest and 0.625% precision of decision tree.

Table 9 presents the experimented results of Pearson correlation on happiness classification dataset. As it can be seen in this Table 9, the same five popularly known classifiers that was used to experiment the proposed algorithm was also used here on the same dataset to compare the results of the Pearson correlation with that of the proposed algorithm in our later discussion. Among these classifiers, the random forest attained the best accuracy, precision and F1-score of 0.689%, 0.684% and 0.742% respectively. Meanwhile, having the best recall of 0.812% as compared to the recall of k-nearest neighbour with the same 0.812%, but still higher than that of 0.625% recall of decision tree, 0.687% recall of gradient boost and 0.687% recall of cat boost.

Table 10 presents the experimented results of information gain on happiness classification dataset. As it can be seen in this Table 10, the same five popularly known classifiers that was used to experiment the proposed algorithm was also used here on the same dataset to compare the results of information gain with that of the proposed algorithm in our later discussion. Among these classifiers, the K-nearest neighbor attained the best accuracy, precision and F1-score of 0.620%, 0.692% and 0.620% respectively. Meanwhile, having the lower recall of 0.562% as compared to random forest with a recall of 0.625%.

Table 10: Information gain on happiness classification dataset.

	ACC (%)	PREC (%)	REC (%)	F1 (%)
Random Forest	0.517	0.571	0.5	0.533
K-Nearest Neighbors	0.620	0.692	0.562	0.620
Decision Tree	0.551	0.636	0.437	0.518
Gradient Boost	0.517	0.571	0.5	0.533
Cat Boost	0.551	0.588	0.625	0.606

Table 11: Classification accuracy comparison on lung cancer dataset.

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Chi square	0.919	0.929	0.981	0.954
Pearson correlation	0.919	0.945	0.981	0.954
Information gain	0.919	0.945	0.981	0.954
Proposed	0.935	0.962	1.0	0.962

Table 12: Classification accuracy comparison on happiness classification dataset.

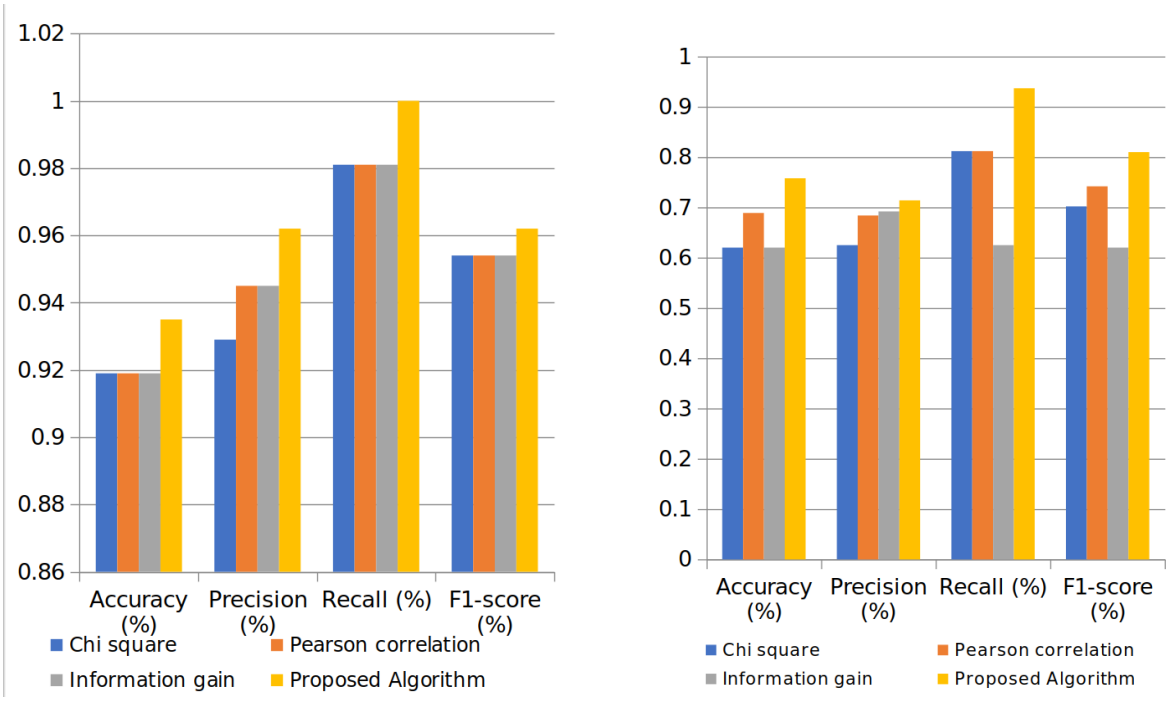
Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Chi square	0.620	0.625	0.812	0.702
Pearson correlation	0.689	0.684	0.812	0.742
Information gain	0.620	0.692	0.625	0.620
Proposed	0.758	0.714	0.937	0.810

Table 11 presents the compared accuracy. Precision, recall and F1-score results of the proposed algorithm with the three existing techniques on the lung cancer dataset. As can be seen in this Table 11, the proposed algorithm outperformed the three existing techniques in terms of accuracy, precision, recall and F1-score which is also illustrated in Figure 2a.

Table 12 presents the compared accuracy. Precision, recall and F1-score results of the proposed algorithm with the three existing techniques on happiness classification dataset. As can be seen in Table 12, the proposed algorithm outperformed the three existing techniques in terms of accuracy, precision, recall and F1-score of 0.785%, 0.714%, 0.937% and 0.810% respectively, this is also illustrated in Figure 2b.

4.2. Discussion

To determine the predictive accuracy power of the proposed algorithm and that of the three existing algorithms. A test of predictive accuracy was carried out. We evaluated five machine learning models on the proposed algorithm and the three existing techniques by deploying the lung cancer dataset and the happiness classification dataset with two classes each. The results of the experiment show that the proposed algorithm attained the best result on lung cancer dataset and the happiness classification dataset with accuracy of 0.935% and 0.758% respectively. We provide a theoretical discussion analysis of the



(a) (b)

Figure 2: (a) Comparison of the proposed algorithm with existing techniques on lung cancer and (b) Comparison of the proposed algorithm with existing techniques on happiness classification dataset.

results based on machine learning models in the sub-section below.

4.2.1. *Machine learning model*

There are several experiments that deployed machine learning models but known of them have justified that a single machine learning classifier can outperformed the others in diverse types of dataset [35]. As such, there is need to deploy different machine learning classifiers to determine which one that produced the best and optimal result in a particular dataset. To build a predictive model, a machine learning model has to be applied to the feature selection subset selected by the feature selection technique. In this work, we deployed five machine learning models to the lung cancer and happiness classification datasets. On the lung cancer dataset, we applied logistic regression, decision tree, ada boost, gradient boost and random forest. Out of the five classifiers, the decision tree, gradient boost and random forest performed better in the proposed feature selection algorithm while random forest performed the best when applied chi square and information gain. Decision tree on the other hand performed better when applied to Pearson correlation. On the Happiness classification dataset, we used random forest, k-nearest neighbor, decision tree, gradient boost and cat boost classifiers. The random forest performed the best accuracy result when applied to the proposed feature selection algorithm and Pearson correlation, whereas the K-Nearest Neighbors outperformed the others on chi square and information gain feature selection techniques.

4.2.2. *Results of the proposed algorithm, chi square, Pearson correlation and information gain*

This section present the results of the proposed algorithm, chi square, Pearson correlation and information gain on lung cancer dataset as can be presented and explained in Tables 3-6, respectively. The results of the proposed algorithm, chi square, Pearson correlation and information gain on happiness classification dataset are also presented and explained in Tables 7-10, respectively. From the two datasets, the proposed algorithms performed better results than the three existing techniques in terms of accuracy, precision, recall and F1-score which can are presented in Table 11 and Table 12.

4.2.3. *Comparison of the proposed algorithm with the existing techniques on lung cancer dataset*

We conducted a comprehensive experiment on the dataset to assess the importance of the proposed algorithm for feature selection in model prediction accuracy using the five machine learning models as earlier discussed. The comparison experiment is illustrated in Table 11, and the result shows that the proposed algorithm attained accuracy, precision, recall and F1-score of 0.935%, 0.962%, 1.0% and 0.962% respectively, which outperformed the three counterparts (the existing feature selection techniques). The visual representation of this comparison is depicted in Figure 2a.

Table 13: Proposed algorithm compared with other studies on lung cancer dataset.

Author	Classifier	Accuracy (%)
Dewi *et al.*[36]	Decision Tree	0.89
Li *et al.* [37]	Random forest	0.893
Patil *et al.* [38]	Logistic Regression	0.84
Proposed Algorithm	Random Forest	0.935

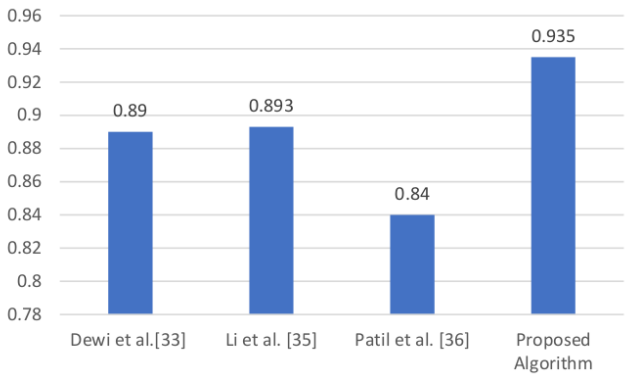


Figure 3: Accuracy prediction of the proposed algorithm with other studies.

4.2.4. *Comparison of the proposed algorithm with the existing techniques on happiness classification dataset*

We also conducted a comprehensive experiment on the dataset to assess the importance of the proposed algorithm for feature selection in model prediction accuracy using the five machine learning models as earlier discussed. The comparison experiment is illustrated in Table 12, and the outcome of the comparison shows that the proposed algorithm with accuracy, precision, recall and F1-score of 0.758%, 0.714%, 0.937% and 0.810% respectively achieved the highest results than the existing techniques. The visual representation of this comparison is depicted in Figure 2b.

Generally, the proposed algorithm scored better than the chi square, Pearson correlation and information gain in all the classifiers items of accuracy, precision, recall and F1-score on the two datasets deployed as can be spotted in Figure 2a and Figure 2b. This implies that, the proposed algorithm significantly improves the prediction accuracy over the three counterparts (findings) as can be summarized in Table 11 and Table 12 for lung cancer and happiness classification dataset respectively. The proposed algorithm achieves feature selection process by first of all, converting the dataset into binary digit. Thereafter, the four variables needed to evaluate the lift measure as can be seen in equation (5), which is the fundamental approached used to develop the proposed algorithm are evaluated from the binary dataset. The lift score for each features values and the outcome are computed. The features with the highest scores are used for the model classification. The limitation of the proposed algorithm is that, it cannot be implemented on a dataset with negative values and can only be implemented on a binary classification problem.

4.2.5. *Comparison of the proposed algorithm with other studies on lung cancer dataset*

Table 13 represents the prediction accuracy performance of the proposed algorithm compared with other research studies on lung cancer classification, whereas Figure 3 depicts the visualization. The proposed algorithm attained higher accuracies than other studies as can be seen in Table 13. This justified that the proposed algorithm is a good feature selection technique that can be adopted by other research to carry out high-dimensional data feature selections for optimal accuracy prediction.

5. Conclusion

A key consideration in supervised machine learning algorithms is feature selection, which focuses on removing irrelevant characteristics from a training dataset so that only the crucial features remain for modelling prediction. Most high-dimensional datasets come with features that are redundant, duplicative and unimportant which are not really needed in modeling. As such, those irrelevant features have to be removed in order to produce good predictive accuracies. Feature selection technique is needed to select the important features from the dataset. There exist several feature selection approaches, in this research; we focused on filter approach which is a statistical approach that operates independently of any machine learning models. This means, the feature selection preprocessing selection activity has to be done before applying any machine learning models for final accuracy prediction. A proposed algorithm was developed, tested and compared with three popular existing feature selection techniques, namely chi square; Pearson correlation and information gain on lung cancer dataset and happiness classification dataset. The results of the proposed algorithm produced higher accuracies than the counterpart as can be observed in Figure 2a for the lung cancer dataset and Figure 2b for the happiness classification dataset. One of the key features of the proposed algorithm is its ability to search the entire feature space, ranging from 0 to infinity. This implies that, the proposed algorithm can take values from 0 to infinity. The limitation of the proposed algorithm is its inability to evaluate dataset with negative values and is only restricted to binary classification datasets. This proposed algorithm is recommended to be used when filtering features of non-negative binary classification dataset of any domain application.

Data availability

All the datasets used in this research were obtained from Kaggle repository at https://www.kaggle.com.

Acknowledgement

We appreciate everyone who contributed to the success of this research work. May God bless you abundantly.

References

[1] D. T. Patel, N. Honest, P. Vyas & A. Patel, "Univariate and multivariate filtering techniques for feature selection and their applications in field of machine learning", in *Applying data science and learning analytics throughout a learner's lifespan*, G. Trajkovski, M. Demeter & H. Hayes (Eds.), IGI Global, 2022, pp. 73-93. https://doi.org/10.4018/978-1-7998-9644-9.ch004.

[2] X. Zhang & J. Gao, "Measuring feature importance of convolutional neural networks", IEEE Access **8** (2020) 196062. https://doi.org/10.1109/ACCESS.2020.3034625.

[3] H. Peng, F. Long & C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance & min-redundancy", IEEE Trans. Pattern Anal. Mach. Intell. **27** (2005) 8. https://doi.org/10.1109/TPAMI.2005.159.

[4] E. K. Bodur and D. D. Atsa'am, "Filter variable selection algorithm using risk ratios for dimensionality reduction of healthcare data for classification", Processes **7** (2019) 4. https://doi.org/10.3390/pr7040222.

[5] M. Cherrington, F. Thabtah, J. Lu & Q. Xu, "Feature selection: filter methods performance challenges", in *2019 International Conference on Computer and Information Sciences (ICCIS)*, IEEE, Apr. 2019, pp. 1–4. https://doi.org/10.1109/ICCISci.2019.8716478.

[6] A. F. R. Araújo, V. O. Antonino & K. L. Ponce-Guevara, "Self-organizing subspace clustering for high-dimensional and multi-view data", Neural Networks **130** (2020) 6. https://doi.org/10.1016/j.neunet.2020.06.022.

[7] Y. Li, Y. Chai, H. Yin & B. Chen, "A novel feature learning framework for high-dimensional data classification", Int. J. Mach. Learn. Cybern. **12** (2021) 2. https://doi.org/10.1007/s13042-020-01188-2.

[8] N. Meinshausen, "Hierarchical testing of variable importance", Biometrika **95** (2008) 2. https://doi.org/10.1093/biomet/asn007.

[9] A. Biancolillo, K. H. Liland, I. Måge, T. Næs & R. Bro, "Variable selection in multi-block regression", Chemom. Intell. Lab. Syst. **156** (2016) 8. https://doi.org/10.1016/j.chemolab.2016.05.016.

[10] G. Heinze & D. Dunkler, "Five myths about variable selection", Transpl. Int. **30** (2017) 1. https://doi.org/10.1111/tri.12895.

[11] R. B. O'Hara and M. J. Sillanpää, "A review of Bayesian variable selection methods: what, how and which", Bayesian Anal. **4** (2009) 1. https://doi.org/10.1214/09-BA403.

[12] U. F. Njoku, A. Abelló, B. Bilalli & G. Bontempi, "Impact of filter feature selection on classification: an empirical study", 2022. [Online]. Available: https://upcommons.upc.edu/bitstream/handle/2117/369043/paper8.pdf;jsessionid=CA1B96892A128489FF39036944684EE7?sequence=1.

[13] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr & J. M. O'Sullivan, "A review of feature selection methods for machine learning-based disease risk prediction", Front. Bioinforma. **2** (2022) 6. https://doi.org/10.3389/fbinf.2022.927312.

[14] T. Mehmood, K. H. Liland, L. Snipen, and S. Sæbø, "A review of variable selection methods in partial least squares regression", Chemom. Intell. Lab. Syst. **118** (2012) 8. https://doi.org/10.1016/j.chemolab.2012.07.010.

[15] G. Soledad, 'Alternative feature selection methods in machine learning", 2021. [Online]. https://www.kdnuggets.com/2021/12/alternative-feature-selection-methods-machine-learning.html.

[16] S. B. Pooja & R. S. Balan, "Point biserial correlated feature selection of weather data", Int. J. Eng. Adv. Technol. **8** (2019) 6. https://doi.org/10.35940/ijeat.F7891.088619.

[17] C. I. Eke, A. A. Norman & L. Shuib, "Multi-feature fusion framework for sarcasm identification on twitter data: A machine learning based approach", PLoS One, **16** (2021) 6. https://doi.org/10.1371/journal.pone.0252918.

[18] D. D. Atsa'am, "Feature selection algorithm using relative odds for data mining classification", in *Big data analytics for sustainable computing*, A. Haldorai & A. Ramu (Eds.), IGI Global, 2020, pp. 81–106. https://doi.org/10.4018/978-1-5225-9750-6.ch005.

[19] S. DeySarakar & S. Goswami, "Empirical study on filter based feature selection methods for text classification", Int. J. Comput. Appl. **81** (2013) 6. https://doi.org/10.5120/14018-2173.

[20] K. Pavya & D. B. Srinivasan, "Feature selection techniques in data mining: a study", Int. J. Sci. Dev. Res. **2** (2017) 6. www.ijsdr.org.

[21] S. K. Gajawada, "Chi-square test for feature selection in machine learning", 2019. [Online]. https://towardsdatascience.com/chi-square-test-for-feature-selection-in-machine-learning-206b1f0b8223.

[22] G. Chandrashekar & F. Sahin, "A survey on feature selection methods", Comput. Electr. Eng. **40** (2014) 16. https://doi.org/10.1016/j.compeleceng.2013.11.024.

[23] S. Solorio-Fernández, J. F. Martínez-Trinidad & J. A. Carrasco-Ochoa, "A supervised filter feature selection method for mixed data based on spectral feature selection and Information-theory redundancy analysis", Pattern Recognit. Lett. **138** (2020) 321. https://doi.org/10.1016/j.patrec.2020.07.039.

[24] G. Singer, R. Anuar & I. Ben-Gal, "A weighted information-gain measure for ordinal classification trees", Expert Syst. Appl. **152** (2020) 11337. https://doi.org/10.1016/j.eswa.2020.113375.

[25] Nurhayati, A. E. Putra, L. K. Wardhani & Busman, "Chi-square feature selection effect on naive bayes classifier algorithm performance for sentiment analysis document", in 2019 7th International Conference on Cyber and IT Service Management (CITSM), IEEE, Nov. 2019, pp. 1–7. https://doi.org/10.1109/CITSM47753.2019.8965332.

[26] L. Sun, T. Wang, W. Ding, J. Xu & Y. Lin, "Feature selection using Fisher score and multilabel neighborhood rough sets for multilabel classification", Inf. Sci. (Ny). **578** (2021) 887. https://doi.org/10.1016/j.ins.2021.08.032.

[27] S. Xie, Y. Zhang, D. Lv, X. Chen, J. Lu & J. Liu, "A new improved maximal relevance and minimal redundancy method based on feature subset", J. Supercomput. **79** (2023) 3. https://doi.org/10.1007/s11227-022-04763-2.

[28] F. Thabtah, F. Kamalov, S. Hammoud & S. R. Shahamiri, "Least Loss: A simplified filter method for feature selection", Inf. Sci. (Ny). **534** (2020) 9. https://doi.org/10.1016/j.ins.2020.05.017.

[29] C. I. Eke, A. A. Norman & L. Shuib, "Multi-feature fusion framework for sarcasm identification on twitter data: A machine learning based approach", PLoS ONE **16** (2021) e0252918. https://doi.org/10.1371/journal.pone.0252918.

[30] P. Chen, F. Li & C. Wu, "Research on intrusion detection method based on Pearson correlation coefficient feature selection algorithm", J. Phys. Conf. Ser. **1757** (2021) 012054. https://doi.org/10.1088/1742-6596/1757/1/012054.

[31] F. Thabtah, F. Kamalov, S. Hammoud & S. R. Shahamiri, "Least loss: A simplified filter method for feature selection", Inf. Sci. (Ny). **534** (2020) 9. https://doi.org/10.1016/j.ins.2020.05.017.

[32] K. Vu, R. A. Clark, C. Bellinger, "The index lift in data mining has a close relationship with the association measure relative risk in epidemiological studies", BMC Med. Inform. Decis. Mak. **19** (2019) 112. https://doi.org/10.1186/s12911-019-0838-4.

[33] C. Iwendi, S. Khan, J. H. Anajemba, M. Mittal, M. Alenezi & M. Alazab, "The use of ensemble models for multiple class and binary class classification for improving intrusion detection systems", Sensors **20** (2020) 9. https://doi.org/10.3390/s20092559.

[34] C. I. Eke, A. A. Norman, Liyana Shuib & H. F. Nweke, "Sarcasm identification in textual data: systematic review, research challenges and open directions", Artif. Intell. Rev., **53** (2020) 6. https://doi.org/10.1007/s10462-019-09791-8.

[35] A. Alhazmi, R. Mahmud, N. Idris, M. E. Mohamed Abo & C. I. Eke, "Code-mixing unveiled: Enhancing the hate speech detection in Arabic dialect tweets using machine learning models", PLoS One **19** (2024) 7. https://doi.org/10.1371/journal.pone.0305657.

[36] Dewi Widyawati and Amaliah Faradibah, "Comparison Analysis of Classification Model Performance in Lung Cancer Prediction Using Decision Tree, Naive Bayes & Support Vector Machine", Indones. J. Data Sci. **4** (2023) 2. https://doi.org/10.56705/ijodas.v4i2.76.

[37] D. Li, G. Li, S. Li & A. Bang, "Classification prediction of lung cancer based on machine learning method", Int. J. Healthc. Inf. Syst. Informatics **19** (2023) 1. https://doi.org/10.4018/IJHISI.333631.

[38] R. Patil, C. G. Sinchana, P. Tejashwini, K. N. Tejaswini & V. V. Ganiga, "Lung cancer prediction system using logistic regression approach", International Research Journal of Modernization in Engineering Technology and Science **2** (2020) 656. https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://www.irjmets.com/uploadedfiles/paper/volume2/issue_12._december_2020/5379/1628083215.pdf.