Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 8, S. 411-419, 2, 2024
© Telif hakkı IJANSER'e aittir
Araştırma Makalesi

IJANSER

https://as-proceeding.com/index.php/ijanser
ISSN: 2980-0811

International Journal of Advanced Natural Sciences and Engineering Researches
Volume 8, pp. 411-419, 2, 2024
Copyright © 2024 IJANSER
Research Article

Breast Cancer Prediction with Hybrid Filter-Wrapper Feature Selection

Tohid Yousefi*[1], Özlem Varlıklar[2]

[1]*Dokuz Eylul University, Computer Engineering Department, Turkey, tohid.yousefi@hotmail.com*
[2]*Dokuz Eylul University, Computer Engineering Department, Turkey, aktas.ozlem@deu.edu.tr*

(Received: 10 March 2024, Accepted: 12 March 2024)

(4th International Conference on Innovative Academic Studies ICIAS 2024, March 12-13, 2024)

ATIF/REFERENCE: Yousefi, T. & Varlıklar, Ö. (2024). Breast Cancer Prediction with Hybrid Filter-Wrapper Feature Selection. *International Journal of Advanced Natural Sciences and Engineering Researches, 8(2), 411-419.*

Abstract – Feature selection, the process of selecting a subset of relevant features for model construction, plays a pivotal role in machine learning tasks, particularly in enhancing model efficiency and performance. It aids in mitigating the curse of dimensionality, reducing computational costs, and improving the generalization of models. Among the various methods employed in feature selection, both filter and wrapper methods stand out for their effectiveness. However, the integration of hybrid versions of these methods holds promising prospects in further enhancing model performance. In a recent study utilizing a breast cancer dataset, encompassing 30 features, the utilization of traditional methods yielded an ROC AUC score of 0.943. Upon employing the hybrid feature selection technique proposed herein, the ROC AUC score surged to 0.954 after selecting a reduced set of 10 features. This significant improvement underscores the efficacy of the proposed method in enhancing model performance, thus affirming its superiority in optimizing predictive accuracy and robustness.

Keywords – Feature Selection, Filter Methods, Wrapper Methods, Hybrid Feature Selection, Hybrid Filter-Wrapper.

I. INTRODUCTION

Feature selection, which involves choosing the most relevant attributes for model building in machine learning, is crucial for optimizing model performance and efficiency. By selecting only the most important features, we can simplify the model, reduce computational complexity, and improve its ability to generalize to new data. However, as the number of features grows, especially in high-dimensional datasets, it can lead to challenges such as the curse of dimensionality [1, 2]. This phenomenon can cause models to become overly complex, making them more susceptible to overfitting and reducing their ability to generalize well to unseen data. Therefore, effective feature selection techniques are essential to mitigate these issues and ensure the robustness and reliability of machine learning models, particularly as the dimensionality of the data increases [3, 4].

Feature selection methods are foundational tools in the field of machine learning, playing a pivotal role in constructing precise models by identifying the most pertinent attributes. Their significance spans diverse domains, with particular emphasis placed on their utility in healthcare applications. In scenarios characterized by smaller dataset sizes, such as those often encountered in healthcare contexts, the judicious selection of features emerges as a critical factor in enhancing model performance [5]. For example, in this study, within the realm of breast cancer data analysis, the meticulous curation of relevant features holds the promise of enabling accurate predictions and facilitating a deeper understanding of the disease. Conversely, the oversight of proper feature selection procedures can result in models grappling with the identification of meaningful patterns within the data, thereby compromising performance and reliability. Therefore, the adoption of effective feature selection techniques stands as a cornerstone for optimizing model performance and fostering robust and dependable outcomes in machine learning endeavors, underscoring their paramount importance in contemporary research and application domains.

In this study, we focused on analyzing breast cancer data, employing feature selection methods to enhance predictive performance. Specifically, we utilized the filter method [6], which involves evaluating features independently of the model [7], and the wrapper method [8], which assesses subsets of features based on model performance [9]. By integrating these methods in a hybrid fashion, we achieved superior predictive accuracy with fewer features. This underscores the importance of feature selection techniques in machine learning applications. Initially, without feature selection (utilizing 30 features), the model achieved a ROC AUC score of 0.943. However, with our proposed method, utilizing only 10 features, the ROC AUC score improved to 0.954, demonstrating a 66% reduction in features while enhancing model performance. These findings underscore the effectiveness of our approach in improving model efficiency and predictive power.

II. MATERIAL AND METHOD

A. *Feature Selection*

Feature selection stands as a pivotal step within the domain of machine learning and data analysis, playing a crucial role in identifying and selecting the most pertinent attributes from a given dataset while filtering out irrelevant or redundant ones [10]. This process, illustrated in Figure 1, aims to enhance the performance, accuracy, and efficiency of machine learning models by reducing the dimensionality of the dataset and focusing solely on the most significant features. By doing so, feature selection assists in mitigating issues such as overfitting, thereby improving the model's ability to generalize to unseen data, reducing computational complexity, and enhancing interpretability [11]. Furthermore, through the meticulous curation of features, practitioners can streamline the modeling process, optimize predictive performance, and uncover deeper insights into the underlying data patterns. In essence, feature selection serves as a cornerstone in the data analysis pipeline, empowering researchers and analysts to make informed decisions and extract meaningful insights from complex datasets [12].

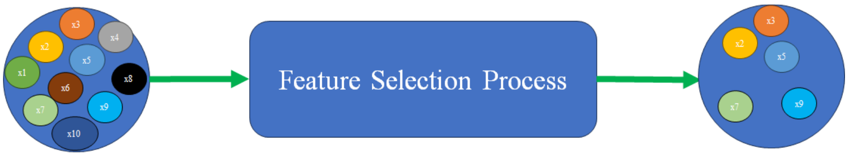

Figure 1. Feature selection process

As depicted in Figure 2, feature selection encompasses a variety of methods, each serving distinct purposes in optimizing model performance. Wrapper methods [8] assess feature subsets using predictive models to select the most informative subset [13], while filter methods [6] rely on statistical measures or correlation to evaluate feature relevance independently of model performance [14]. Embedded methods [15], on the other hand, integrate feature selection directly into the model training process, refining feature importance during training [16]. Ensemble methods [17], such as bagging and boosting, combine predictions from multiple weak learners to create a stronger model [18]. Hybrid methods [19] merge the strengths of different approaches to achieve more effective feature selection [20]. As the field of feature selection continues to evolve, new techniques emerge and existing ones are refined, advancing our understanding and application of feature selection in machine learning and data analysis.

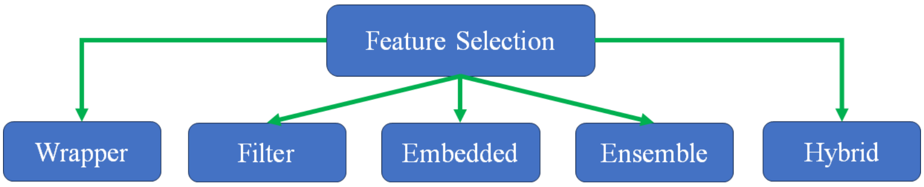

Figure 2. Feature selection methods

B. *Wrapper Method*

Wrapper method [8] is another vital technique in feature selection, which differs from filter methods by incorporating the actual performance of a predictive model in the selection process [21]. As depicted in Figure 3, Unlike filter methods that evaluate features independently, wrapper methods employ a specific model to evaluate subsets of features iteratively, determining the subset that yields the best performance. This iterative process typically involves training and evaluating the model with different combinations of features, making it computationally expensive but potentially more accurate than filter methods. Wrapper methods aim to select the optimal subset of features that maximizes model performance, thereby enhancing predictive accuracy [7, 21].

Wrapper methods are essential in feature selection, employing a predictive model to evaluate subsets of features iteratively and select the best-performing subset. These methods, such as Sequential Forward Feature Selection (SFS) [22] start with an empty set of features and gradually add features one by one based on their individual contribution to model performance [23]. Sequential Backward Feature Selection (SBS) [24], on the other hand, begins with the full set of features and sequentially removes the least relevant features until optimal performance is achieved [25]. Sequential Floating Forward Feature Selection (SFFS) [26] and Sequential Floating Backward Feature Selection (SBFS) [26] are extensions of SFS and SBS, respectively, allowing for the possibility of adding or removing multiple features at each iteration [27]. Exhaustive Feature Selection (EFS) [28] is another wrapper method that evaluates all possible feature combinations to identify the optimal subset, ensuring the best possible performance at the cost of increased computational complexity [29].

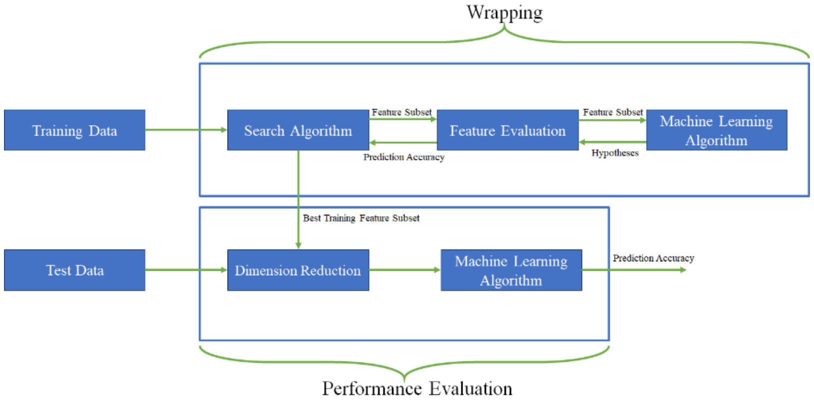

Figure 3. The general framework of wrapper method

C. *Filter Method*

The filter method [6] is one of the fundamental techniques in feature selection, aimed at identifying the most relevant features based on their intrinsic characteristics [30]. As depicted in Figure 3, Unlike wrapper methods that incorporate a predictive model, filter methods evaluate features independently of the model. They typically rely on statistical measures or predefined criteria to assess the importance of features [6].

For instance, one common approach is ANOVA (Analysis of Variance) [31], which measures the variance between different groups of data to determine feature relevance [32]. Another popular method is mutual information [33], which quantifies the mutual dependence between two variables and is particularly useful for identifying nonlinear relationships [34]. Additionally, the chi-square (χ^2) [35] test assesses the independence between categorical variables, making it suitable for feature selection in classification tasks [36]. These methods serve as invaluable tools in preprocessing data and selecting informative features prior to model training, contributing to improved model performance and interpretability.

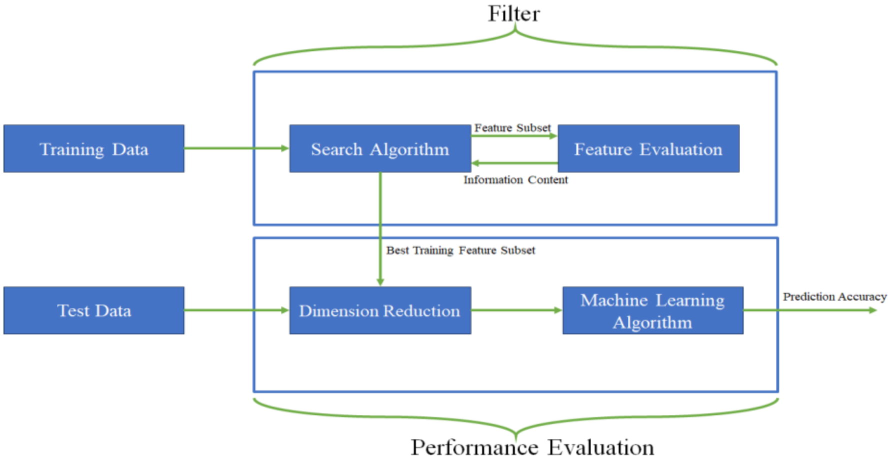

Figure 4. The general framework of filter method

D. *Random Forest*

Random Forest [37, 38], as a machine learning algorithm, operates by combining multiple decision trees to make a prediction. Each decision tree is trained on a randomly sampled subset of the data and selects the best split using a random subset of features. Subsequently, the predictions from each tree are aggregated to make a final prediction. This method is resilient to overfitting and can be utilized for both classification and regression problems. Random Forest performs well on large and complex datasets and enhances model interpretability by providing feature importance rankings [39].

E. *Hybrid Filter-Wrapper Method*

Feature selection plays a crucial role in machine learning by enhancing model performance and reducing computational costs. Without feature selection, models may suffer from high complexity and inefficiency due to the inclusion of irrelevant or redundant features, leading to increased training times and computational expenses.

As depicted in Figure 5, our proposed hybrid method involves several steps to efficiently select the most relevant features from the dataset. Initially, the method applies basic filter-based feature selection techniques to identify and remove constant, quasi-constant, and duplicate features. Then, it performs correlation analysis to eliminate features with high correlation above a certain threshold. Next, the method utilizes filter feature selection methods such as ANOVA, chi-square, and mutual information to select the most important features. Subsequently, it employs wrapper methods including Sequential Forward Feature Selection (SFS), Sequential Backward Feature Selection (SBS), Sequential Floating Forward Feature Selection (SFFS), Sequential Floating Backward Feature Selection (SBFS), and Exhaustive Feature Selection (EFS) to further refine the feature subset. Finally, the method divides the dataset into training and testing sets using the selected features, trains machine learning algorithms with fewer features, performs model tuning, and evaluates the results.

In conclusion, our hybrid method offers a superior approach to feature selection, resulting in improved model performance with fewer features and reduced computational costs. By systematically selecting the most informative features and utilizing both filter and wrapper methods, our approach ensures more efficient and effective machine learning model training and deployment.

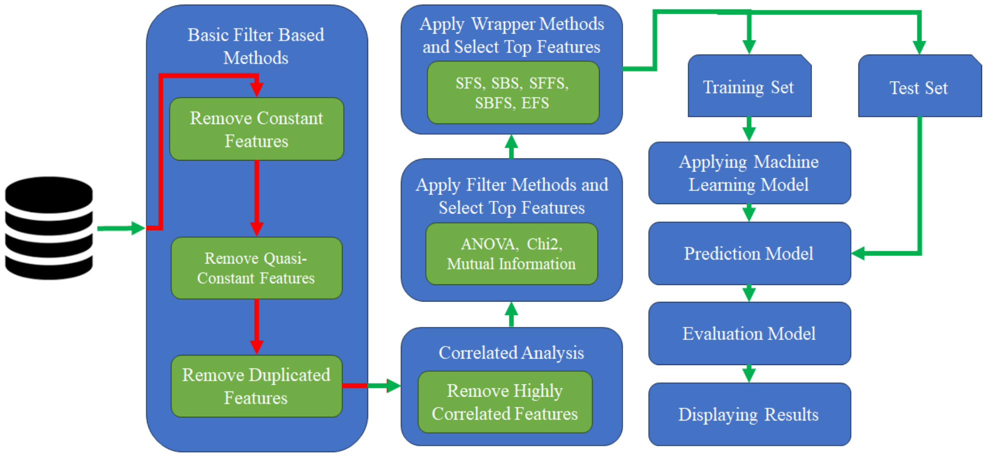
Basic Filter Based Methods
Remove Constant Features
Remove Quasi-Constant Features
Remove Duplicated Features
Apply Wrapper Methods and Select Top Features
SFS, SBS, SFFS, SBFS, EFS
Apply Filter Methods and Select Top Features
ANOVA, Chi2, Mutual Information
Correlated Analysis
Remove Highly Correlated Features
Training Set
Test Set
Applying Machine Learning Model
Prediction Model
Evaluation Model
Displaying Results

Figure 5. Flowchart of the proposed hybrid filter-wrapper method

F. *Dataset*

The dataset contains 30 features associated with breast cancer, excluding id and diagnosis. These features are utilized to predict the presence of breast cancer in individuals. You can access the dataset for free on Kaggle through the following link: (https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset).

III. RESULTS

In this study, initially, a modeling process was conducted without applying any feature selection, where predictions for breast cancer were made using all features available in the dataset. Subsequently, employing the proposed method resulted in achieving better predictive performance with 66% fewer features. The rationale behind this improvement lies in the utilization of both filter and wrapper methods in the proposed approach, which enabled the selection of the best features, thereby facilitating the creation of a robust model. By leveraging the combined strength of filter and wrapper techniques, our method effectively identified the most informative features, leading to enhanced modeling outcomes.

The code used in this research is openly accessible on both Kaggle and GitHub platforms. Readers who are interested can find the code repository on GitHub via the following link: (https://github.com/tohid-yousefi/Breast-Cancer-Prediction-with-Hybrid-Filter-Wrapper-Feature-Selection). Furthermore, the code is also accessible on Kaggle using the following link: (https://www.kaggle.com/code/tohidyousefi/breast-cancer-prediction-with-hybrid-method).

A. *Creating a Random Forest Model Without Using the Hybrid Filter-Wrapper Method*

In this section, we initially constructed a model for breast cancer diagnosis using all features available in the dataset (excluding id and diagnosis), without performing any feature selection. For prediction purposes, we employed the random forest algorithm. At the conclusion of the study, we obtained a ROC AUC value of 0.943 for the test data, as depicted in Table 1 and Figure 6. Notably, this result was achieved without conducting any feature selection, highlighting the potential drawbacks in terms of time and costs associated with utilizing all features indiscriminately.

Table 1. Metrics of random forest model without using the hybrid filter-wrapper method (for test data)

	Precision	**Recall**	**F1-score**	**ROC AUC**
0	0.95	0.98	0.96	
1	0.97	0.90	0.93	
				0.943

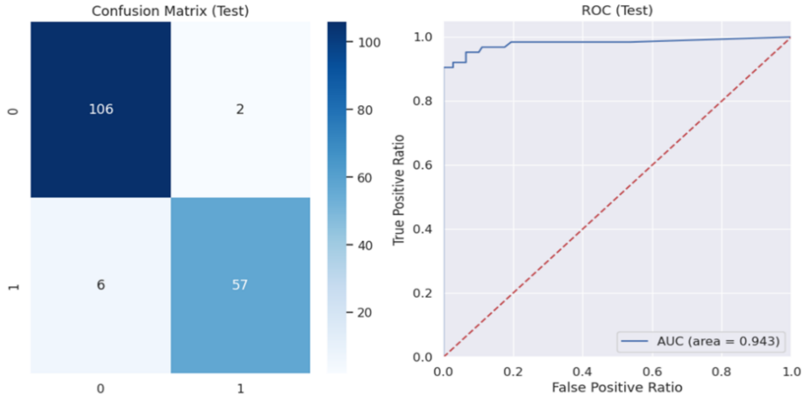

Figure 6. Confusion matrix and roc auc graphic of random forest model without using the hybrid filter-wrapper method

B. *Creating a Random Forest Model Using the Hybrid Filter-Wrapper Method*

Our proposed hybrid method for breast cancer diagnosis combines both filter and wrapper techniques to optimize feature selection. By integrating various filter methods such as ANOVA, CHi2, and MUTUAL_INFORMATION with wrapper methods including SFS, SBS, SFFS, SBFS, and EFS, we achieved notable results. Specifically, as observed in Table 2 and Figure 6, the MUTUAL_INFORMATION-SBFS method demonstrated superior performance, attaining a ROC AUC value of 0.954 while utilizing only 10 out of the initial 30 features. This outcome underscores the effectiveness of our approach in enhancing predictive accuracy while reducing computational costs and processing time. Additionally, upon examining Table 3, it becomes evident that our method outperforms other techniques, further reinforcing the efficacy of our approach.

Table 2. Metrics of random forest model using the hybrid filter-wrapper method (for test data)

	Precision	**Recall**	**F1-score**	**ROC AUC**
0	0.96	0.97	0.97	
1	0.95	0.94	0.94	
				0.954

Table 3. All Metrics of the random forest model using the hybrid filter-wrapper method (for test data)

	ROC AUC
ANOVA-SFS	0.943
ANOVA-SBS	0.938
ANOVA-SFFS	0.943
ANOVA-SBFS	0.951
ANOVA-EFS	0.907
CHI2-SFS	0.943
CHI2-SBS	0.934
CHI2-SFFS	0.943
CHI2-SBFS	0.946
CHI2-EFS	0.915
MUTUAL_INFORMATION-SFS	0.942
MUTUAL_INFORMATION-SBS	0.934
MUTUAL_INFORMATION-SFFS	0.938
MUTUAL_INFORMATION-SBFS	**0.954**
MUTUAL_INFORMATION-EFS	0.913

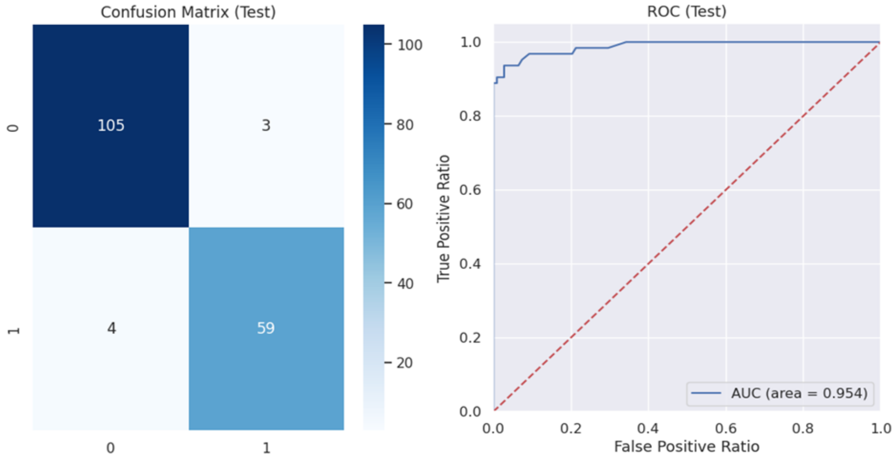

Figure 7. Confusion matrix and roc auc graphic of random forest model using the hybrid filter-wrapper method

IV. CONCLUSION

Feature selection is a crucial step in machine learning, allowing for the identification of the most relevant features while discarding irrelevant or redundant ones. This process offers several advantages, including improved model performance, enhanced interpretability, reduced overfitting, and decreased computational costs. In our study, we initially conducted breast cancer diagnosis using all features available in the dataset (30 features) without performing any feature selection, resulting in a ROC AUC value of 0.943. Subsequently, we adopted a hybrid approach combining filter and wrapper methods. Among the hybrid models developed, the MUTUAL_INFORMATION-SBFS method stood out, achieving a ROC AUC value of 0.954 using only 10 features, or 33% of the original feature set. This result serves as evidence of the effectiveness of our model and emphasizes the importance of feature selection methods in developing efficient models for breast cancer diagnosis. By leveraging these methods, we were able to develop a model with reduced time, complexity, and cost while maintaining high diagnostic accuracy for breast cancer.

REFERENCES

1. Miao J. and Niu L., "A survey on feature selection," Procedia computer science, vol. 91, no. pp. 919-926, 2016.
2. Shardlow M., "An analysis of feature selection techniques," The University of Manchester, vol. 1, no. 2016, pp. 1-7, 2016.
3. Wang S., Tang J., and Liu H., "Feature Selection," vol. no. pp. 2017.
4. Kalousis A., Prados J., and Hilario M., "Stability of feature selection algorithms: a study on high-dimensional spaces," Knowledge and information systems, vol. 12, no. pp. 95-116, 2007.
5. Remeseiro B. and Bolon-Canedo V., "A review of feature selection methods in medical applications," Computers in biology and medicine, vol. 112, no. pp. 103375, 2019.

6. Chandrashekar G. and Sahin F., "A survey on feature selection methods," Computers & Electrical Engineering, vol. 40, no. 1, pp. 16-28, 2014.
7. Li J., Cheng K., Wang S., Morstatter F., Trevino R.P., Tang J., and Liu H., "Feature selection: A data perspective," ACM computing surveys (CSUR), vol. 50, no. 6, pp. 1-45, 2017.
8. Kohavi R. and John G.H., "Wrappers for feature subset selection," Artificial intelligence, vol. 97, no. 1-2, pp. 273-324, 1997.
9. Liu H., Motoda H., Setiono R., and Zhao Z., "Feature selection: An ever evolving frontier in data mining," Feature selection in data mining, pp. 4-13, 2010. pp. 4-13, 2010.
10. Ramchandran A. and Sangaiah A.K., Unsupervised anomaly detection for high dimensional data—An exploratory analysis, Elsevier, 2018.
11. Jimenez-del-Toro O., Otálora S., Andersson M., Eurén K., Hedlund M., Rousson M., Müller H., and Atzori M., Analysis of histopathology images: From traditional machine learning to deep learning, Elsevier, 2017.
12. Dey N., Borra S., Ashour A.S., and Shi F., Machine learning in bio-signal analysis and diagnostic imaging, Academic Press, 2018.
13. Talavera L., "An evaluation of filter and wrapper methods for feature selection in categorical clustering," International Symposium on Intelligent Data Analysis, pp. 440-451, 2005. pp. 440-451, 2005.
14. Dash M. and Liu H., "Feature selection for classification," Intelligent data analysis, vol. 1, no. 1-4, pp. 131-156, 1997.
15. Zheng A. and Casari A., Feature engineering for machine learning: principles and techniques for data scientists, " O'Reilly Media, Inc.", 2018.
16. Ang J.C., Mirzal A., Haron H., and Hamed H.N.A., "Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection," IEEE/ACM transactions on computational biology and bioinformatics, vol. 13, no. 5, pp. 971-989, 2015.
17. Opitz D. and Maclin R., "Popular ensemble methods: An empirical study," Journal of artificial intelligence research, vol. 11, no. pp. 169-198, 1999.
18. Saeys Y., Abeel T., and Van de Peer Y., "Robust feature selection using ensemble feature selection techniques," Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2008, Antwerp, Belgium, September 15-19, 2008, Proceedings, Part II 19, pp. 313-325, 2008. pp. 313-325, 2008.
19. Kabir M.M., Islam M.M., and Murase K., "A new wrapper feature selection approach using neural network," Neurocomputing, vol. 73, no. 16-18, pp. 3273-3283, 2010.
20. Peng Y., Wu Z., and Jiang J., "A novel feature selection approach for biomedical data classification," Journal of Biomedical Informatics, vol. 43, no. 1, pp. 15-23, 2010.
21. Jović A., Brkić K., and Bogunović N., "A review of feature selection methods with applications," 2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO), pp. 1200-1205, 2015. pp. 1200-1205, 2015.
22. Whitney A.W., "A direct method of nonparametric measurement selection," IEEE transactions on computers, vol. 100, no. 9, pp. 1100-1103, 1971.
23. Inza I., Larranaga P., Blanco R., and Cerrolaza A.J., "Filter versus wrapper gene selection approaches in DNA microarray domains," Artificial intelligence in medicine, vol. 31, no. 2, pp. 91-103, 2004.
24. Marill T. and Green D., "On the effectiveness of receptors in recognition systems," IEEE transactions on Information Theory, vol. 9, no. 1, pp. 11-17, 1963.
25. Ladha L. and Deepa T., "Feature selection methods and algorithms," International journal on computer science and engineering, vol. 3, no. 5, pp. 1787-1797, 2011.
26. Pudil P., Novovičová J., and Kittler J., "Floating search methods in feature selection," Pattern Recognition Letters, vol. 15, no. 11, pp. 1119-1125, 1994.
27. Mlambo W., Cheruiyot W.K., and Kimwele M.W., "A survey and comparative study of filter and wrapper feature selection techniques," Int. J. Eng. Sci, vol. 5, no. 8, pp. 57-67, 2016.
28. Galatenko V., Shkurnikov M.Y., Samatov T., Galatenko A., Mityakina I., Kaprin A., Schumacher U., and Tonevitsky A., "Highly informative marker sets consisting of genes with low individual degree of differential expression," Scientific Reports, vol. 5, no. 1, pp. 14967, 2015.
29. Galatenko V.V., Maltseva D.V., Galatenko A.V., Rodin S., and Tonevitsky A.G., "Cumulative prognostic power of laminin genes in colorectal cancer," BMC Medical Genomics, vol. 11, no. 1, pp. 77-81, 2018.
30. Kullback S. and Leibler R.A., "On information and sufficiency," The annals of mathematical statistics, vol. 22, no. 1, pp. 79-86, 1951.
31. St L. and Wold S., "Analysis of variance (ANOVA)," Chemometrics and intelligent laboratory systems, vol. 6, no. 4, pp. 259-272, 1989.
32. AJPAS A., "A Feature Selection Based on One-Way-Anova for Microarray Data Classification," AJPAS JOURNAL, vol. 3, no. pp. 1-6, 2016.
33. Shannon C.E., "A mathematical theory of communication," The Bell system technical journal, vol. 27, no. 3, pp. 379-423, 1948.
34. Peng H., Long F., and Ding C., "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," IEEE transactions on pattern analysis and machine intelligence, vol. 27, no. 8, pp. 1226-1238, 2005.

35. Kass G.V., "An exploratory technique for investigating large quantities of categorical data," Journal of the Royal Statistical Society: Series C (Applied Statistics), vol. 29, no. 2, pp. 119-127, 1980.
36. Esmael B., Arnaout A., Fruhwirth R., and Thonhauser G., "A statistical feature-based approach for operations recognition in drilling time series," International Journal of Computer Information Systems and Industrial Management Applications, vol. 4, no. 6, pp. 100-108, 2012.
37. Breiman L., "Random forests," Machine learning, vol. 45, no. pp. 5-32, 2001.
38. Breiman L., "Bagging predictors," Machine learning, vol. 24, no. 2, pp. 123-140, 1996.
39. Biau G. and Scornet E., "A random forest guided tour," Test, vol. 25, no. pp. 197-227, 2016.