Optimized Analysis Using Feature Selection Techniques for Drug Discovery Detection

Abhay Dadhwal(✉) and Meenu Gupta

Department of Computer Science and Engineering, Chandigarh University, Punjab, India
dadhwal.abhay.abhay@gmail.com, meenu.e9406@cumail.in

Abstract. Machine learning is a tool with immense potential. One of the most important tasks of machine learning process is feature selection. To select best feature selection technique in Drug Discovery most studies could be noticed testing them with classifiers and selecting the one with highest score. But this study demonstrates that in such environment the features can be selected in a more effective manner to determine their quality. This study employs five feature selection techniques and utilizes their results collectively in a method called Priority feature selection. This method ranks the features and produces the most optimum set of features of this experiment. This standard is verified by testing this method by four classifiers where it produces results that surpass the performance of other feature selection techniques. This study also makes a big difference by producing and suggesting a feature selection method that attains maximum performance with all classifiers.

Keywords: Drug discovery · Feature selection · Classifier · Accuracy · Comparative analysis

1 Introduction

Small organic molecules that achieve their desired action by binding on a receptor at the target site can be called drugs. Their interactions are microscopic. It occurs in the molecular world. Drug Discovery as evident through the title is the process of finding possible new medicines. Subjects that encompass molecular theory like biology, chemistry, and pharmacology are covered in drug discovery. Ideally, experts and biologists would attempt to better understand the mechanisms of disease to build a molecule that can disrupt the disease agents. Yet, due to the presence of several complex interactions at the cellular level, it is difficult to proceed with the logical drug design process as good analytical skills are required. Some books provide great insight into the processes included in designing drugs [1]. Due to heavy resource consumption in drug discovery, success is noticed in high-yielding industries like cancer research lately. It is estimated that taking a drug from research to the market will cost an average of $2.6 billion and more than 10 years (can be seen in Fig. 1).

A. K. Luhach et al. (Eds.): ICAICR 2021, CCIS 1575, pp. 43–54, 2022.
https://doi.org/10.1007/978-3-031-09469-9_4

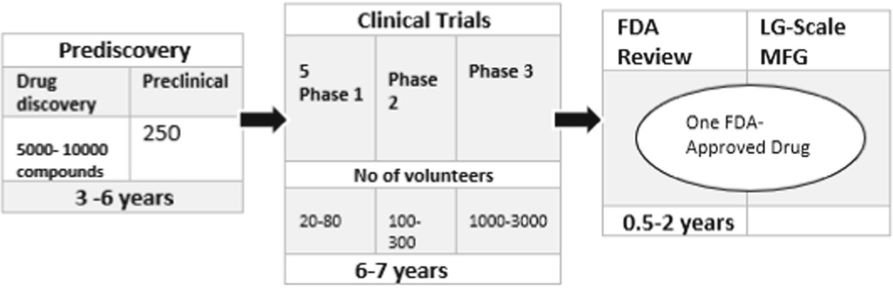

Fig. 1. Drug discovery pipeline

The need to stay relevant in the modern-day has pushed various industries [2–5] to practice machine learning techniques. Through the introduction of fresh concepts in AI like AlphaFold (AI algorithm of Google), Insilico Medicine, etc. the drug development job has seen a drastic change with work being reduced to days from years. A term called Synbiolic creates brand new molecules using Variational Autoencoder (VAE). Thus, significance of AI has become huge in Drug Discovery.

A wide variety of attributes or characteristics, such as topological indices, the characterization of three-dimensional molecular structures, quantum mechanical descriptors, and molecular field parameters, may be required to describe complex molecular compounds leading to tens or hundreds of thousands of characteristics. That is prohibitively high in some learning algorithms and in truth not all the features hold equal value. Therefore it is a good choice to choose the appropriate characteristics that are necessary to construct a quality model. Fundamental steps are still necessary, such as data cleaning and preprocessing [6, 7]. The larger the amount of unnecessary features, more difficult will be to find an acceptable decision function for the algorithm: the system may not converge to an optimal solution within a suitable period of time, or more data may be required to achieve a correct solution. To achieve best subset of features one can:

- Consider all features equally at first and calculate correlation value between target and every other feature.
- Select the subset that produces the most accurate result and also provides stronger generalizing power

1.1 Problem Statement and Contribution

In an environment of multiple feature selection techniques, recent studies could be noticed forming their results based on individual feature selection technique performance. They simply choose the technique that produces highest performance score out of all. But this study proves that it is possible to utilize multiple feature selection techniques collectively to determine the best set of features. It is carried out under the Priority feature selection method and the results produced prove that features selected through collective utilization deliver most optimum performance.

The individual and relative performance analysis of feature selection techniques is also demonstrated in this study as it always helps to understand more about the dataset in hand, for example, the relation between performance and feature-set size in general. The derived findings will lead to better future decisions especially in the field of pharmaceutical scientific research. In this paper most acknowledged feature selection techniques and classifiers are employed namely Pearson correlation coefficient, Chi-squared, Anova-F measure, Lasso and Tree-based (Random Forest).

Further sections are divided as follows Sect. 2 provides review of the previous studies and explains some areas where enhancements can be made. Section 3 acquaints with this study by describing the tools and techniques used. Section 4 presents the results obtained and the insights derived through their analysis. Final Sect. 5 summarizes the whole study by discussing important points and final outcome.

2 Literature Review

Since the recognition of the ability of feature selection there has been a lot of experimentation by researchers [8, 9]. The Drug Discovery process should be efficient; from cost to techniques applied everything should be efficiently sorted. Below mentioned are the studies that helped to familiarize with current techniques in Drug Discovery and to select the tools required for this experiment. Moreover, they helped to guide and define our research problem as described in the summary after these studies.

H. Shi et al. [10] and S. Redkar [11] laid the foundation for our study. They utilize feature selection methods for DTI prediction and also address the issue of class imbalance and high dimensionality. They employ WEKA tool and fewer number of feature selection methods are utilized and the ones producing higher score value are simply considered the final outcome. Choosing the available feature selection option from WEKA tool doesn't give much information about the suitability of features. More information on the selected features can make the process more robust in terms of performance. This forms the base of our study and is discussed more in the summary of this section.

T.Clifford [12] and K. Zhao [13] contributed in drug discovery domain by working on the feature space of dataset. They applied feature selection methods and classifiers. It proved the high efficiency that ML can provide in Drug Discovery. A larger feature-space can be experimented to produce results that relate more to the real-world as the real-world feature-space is usually large. A. Al. Marouf [14] and H. B. Chakrapani [15] considered domains of social media and SQL. They utilize multiple feature selection methods and help to evaluate their efficiency. PCC and Anova-based feature selection prevail most suitable due to their high robustness and accuracy suggesting their high consistency with present datasets of the world.

Now we summarize the above studies to pinpoint the areas in which our study is based:

The base studies such as H. Shi [10] and S. Redkar [11] of this experiment employ multiple feature selection techniques to determine the best among them. They formed their selection upon a simple criteria of high prediction accuracy and thus chose the one with best individual performance. In this study, collective utilization of multiple feature selection techniques is demonstrated to determine the best set of features, which is discussed more in the definition of Priority method. The results produced prove more robust performance than the base studies. It is noticed that the selected features through the method *WrapperSubetEval* in base studies are not capable of delivering peak performance with every classifier. But in this study a single technique of feature selection capable of delivering peak performance with every classifier is attained.

3 Methodology

This section provides the description about data source, the dataset, and techniques employed followed by the methodological framework.

3.1 Data Used

The dataset is drawn from online repository UCI https://archive.ics.uci.edu/ml/datasets, which is a highly trusted source. This is a drug dataset about chemical compounds and contains thousands of features. The instances are to be classified in one of the two classes as active or inactive. The dataset well resembles the real-world data in terms of its quality and size. Thus, the final outcomes perform equally good with any real-world data.

3.2 Feature Selection Techniques

Pearson Correlation Coefficient: This easy-to-understand method comes under category of Filter Feature selection techniques. In this method, the correlation value between input and target features is calculated through below formula:

$$r = \frac{\sum(x - \overline{x})(y - \overline{y})}{\sqrt{\sum(x - \overline{x})^2 \sum(y - \overline{y})^2}} \tag{1}$$

where x and y denote input and target feature. Some recent studies describe utilization of PCC in different fields [16].

Tree-Based-Select From Model: An embedded method [17] that uses Random forest for feature selection by calculating node impurities. Information Gain is one of the attribute selection measures which can be elaborated as:

$$\text{Gain(S, A)} = \text{Entropy(S)}–\sum\nolimits_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{2}$$

where Values (A) is the all possible values for attribute A, and Sv is the subset of S for which attribute A has value v.

Lasso-Select From Model: It belongs to same category of Embedded techniques. It includes a regularizer which can drive parameters to zero which is the main reason behind its presence among various domains [18]. This is understood more by the following formula:

$$\propto \sum_{i=1}^{k} |w_i| \tag{3}$$

Higher the values of alpha, the fewer features have non-zero values. With high dimensionality, this technique becomes a promising choice due to presence of more irrelevant features.

Chi-square Method: It deals with degrees of freedom and observed and calculated values to detrmine scores between features. Its application is visible across different domains [19]. Formula for Chi-square calculation is described as:

$$\sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{4}$$

where O_i *is* number of observations in class i and E_i is no. of expected observations in class i if there was no relation.

Anova-f Method: Anova uses F-test to check if there is any significant difference between groups. It supposes hypothesis to be:-

h0 = All groups have same mean.
h1 = Between groups, there is at least one significant difference.

It's a relevant option with a dataset having numerical input and categorical output. Its usage and comparisons as a feature selection technique can be noticed in recent studies [20].

Priority Method: It is an interesting method which utilizes the results of all other feature selection techniques. Here, the importance of every feature is determined which depends upon the number of techniques that selected this feature.

$$Priority \propto count$$

Priority refers to the importance of a feature and count refers to number of feature selection techniques that selected the feature.

Lets say, if a feature is selected by all feature seletion techniques, then it is ranked higher. It is performed for every feature to finally obtain a set of ranked features. This ranked set is tested with different classifiers and with different feature-set sizes.

Methodological Approach:

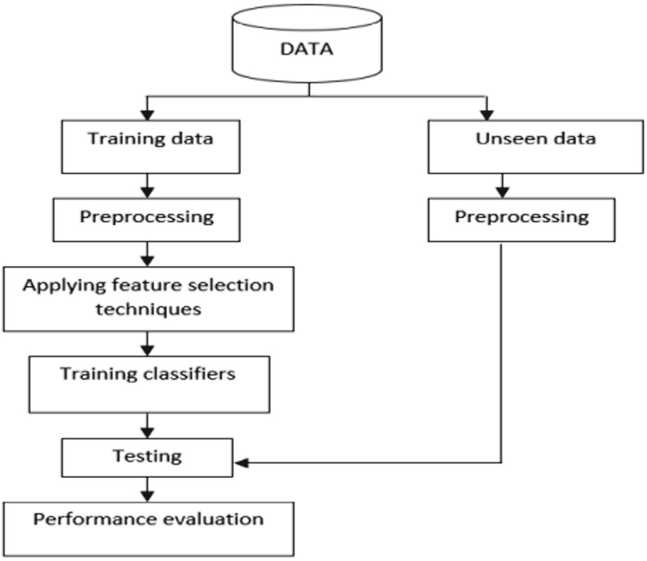

Fig. 2. Methodology

With training and unseen dataset provided, the common machine learning approach was followed as given in Fig. 2. Training and unseen dataset were scanned for erroneous values and preprocessing was performed to clean the dataset. Then, features were selected again and again in different numbers to find the combination with best performance by using feature selection techniques.

All steps after preprocessing were repeated again and again for different amount of features, classifiers and feature selection techniques as described more clearly in next section. This made for many number of combinations and for each combination, observations were noted to derive insights and ultimately figure out the one that gives the best performance.

4 Experimental Results and Analysis

Following observations were noted by applying combinations of type of techniques and amount of features. Performance with training set helps us to know whether the model has learned the training set or not and in what manner did that happen but what helps to relate with real world is unseen dataset performance, hence it is noticed more in following analysis.

4.1 Analysis

The feature set size varied from hundred features as least to all features at most. By increasing the feature set by 100, observations were recorded again and again, starting from 100 features. This continued till 1000 features, after which they were increased by 500. This process promised deeper analysis. The term higher or larger feature sets denotes all feature sets that exceed the length of 700 and lower sets are the ones below. Relative performance evaluation after the analysis of every two techniques is provided based on the performance within the more significant order set. It serves as a checkpoint of performance of feature selection techniques describing local optimum performance. It plots overall accuracy with classifier. As minimum outliers are observed in this study therefore overall accuracy can be trusted to reflect true overview of the performance of techniques with the feature sets. The margins are excluded from the Figs. 3, 4 and 5 as the focus is on the pattern of relative performance. Tables 1, 2 and 3 contain those margins or performance scores. When not specified, assume the documented analysis with unseen set.

The Tables 1, 2 and 3 provide an overview of performance for each feature selection technique by displaying results of different classifiers. Overall accuracy attained among low and high order feature sets on unseen data is shown in the table. They help to understand the performance and contribute in the analysis of the respective classifier and feature selection method. Along every table, detailed analysis that includes both training and unseen dataset is given with respect to every classifier to help acquire useful insights.

Performances of feature selection techniques can be analyzed as:

Pearson Correlation Coefficient (PCC) and Tree-based (TB): Table 2 helps to describe the general performance of PCC and TB method with classifiers on both feature sets which also serves to document its detailed analysis as below:

Low & High-order Sets: The majority of combinations with both PCC and TB display a tendency to yield best performance with low-order sets like BNB, Ad and LR. LR could attain the highest score and outperform all the other combinations.

Table 1. Pearson correlation and tree-based method

Overall Accuracy (Unseen set)				
Classifiers	PCC		TB	
	Low-order	High-order	Low-order	High-order
LR	90.17%	89.46%	89.01%	86.77%
RF	89.27%	89.66%	89.74%	89.85%
BNB	90.3%	90.3%	90.30%	90.3%
Ad	88.81%	87.87%	88.17%	87.42%

In almost every scenario low-order sets outperform high-order ones suggesting more efficiency associated with them than high-orders sets.

Relative Performance: PCC vs TB: As low order sets perform consistently better with all classifiers than high order sets so their behaviour with feature selection techniques is studied more through the following figure to realize any underlying patterns and enhance the performance.

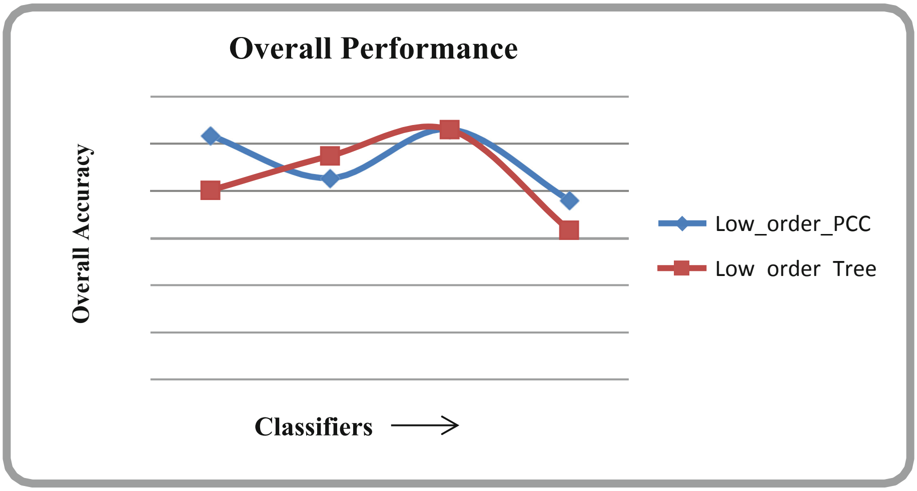

Fig. 3. Relative Performance of PCC and TB

PCC produces more amount of high scores within low order sets as compared to Tree based method with Adaboost and Logistic Regression. Thus, PCC stands as a better choice than TB method.

Lasso and Chi-Square Method: Table 2 describes general performance of these methods and helps to document their analysis below:

Low & High-Order Sets: The prediction model of all classifiers attained better classification accuracy with low-orders sets and the highest was observed from LR. RF and BNB could attain high scores with high-orders sets but not as consistently.

Table 2. Lasso and Chi-square Method

Overall Accuracy (Unseen set)				
Classifiers	Lasso		Chi-sq	
	Low-order	High-order	Low-order	High-order
LR	90.45%	84.44%	89.95%	89.75%
RF	89.30%	89.40%	89.61%	89.82%
BNB	90.3%	90.3%	90.30%	90.30%
Ad	87.63%	88.26%	88.14%	88.48%

Thus, behaviour with the feature sets suggests the presence of necessary features in higher amount with low-order sets.

Relative Performance: Lasso & Chi-square: Following figure is plotted to understand the relative performance of these two methods under low-order sets (Fig. 4):

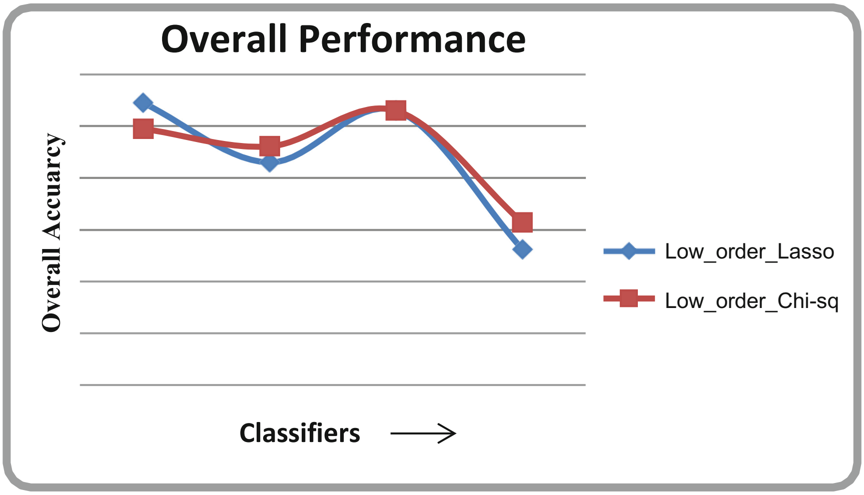

Fig. 4. Relative Performance of Lasso and Chi-sq

Lasso and Chi-square both run closely and display high performance. LR tops the score value with Lasso but Chi-sq excels with two classifiers suggesting equal capability for both the techniques.

Anova-f and Priority Method: Below Table 3 describes general behaviour of these methods:

Table 3. Anova-f and Priority method

Overall Accuracy (Unseen set)				
Classifiers	Lasso		Chi-sq	
	Low-order	High-order	Low-order	High-order
LR	90.45%	84.44%	89.95%	89.75%
RF	89.30%	89.40%	89.61%	89.82%
BNB	90.3%	90.3%	90.30%	90.30%
Ad	87.63%	88.26%	88.14%	88.48%

Low & High-order Sets: The number of samples correctly classified under low-order sets were higher than other section with most classifiers. Though, RF attained best with high-order but it was outperformed by LR. This reflects higher redundancy or impure features among larger feature sets.

Relative Performance: Anova & Priority: Figure 5 reflects the relative performance of these methods under low-order sets:

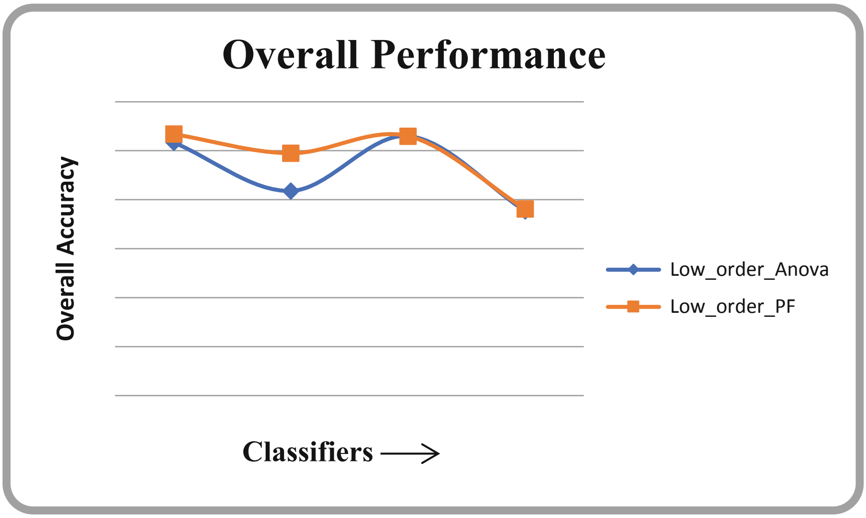

Fig. 5. Relative Performance of Anova nd Priority

Evident from figure, Anova-selected features form decent model but the best is attained with Priority-selected features.

Overall Analysis: The findings of the experiment are summarized below:

Tree-Based method of feature selection indicates high uncertainity in its performance. It fails to deliver high scores with LR on low order sets. Between Lasso and Chi-square the latter responds better to diversity suggesting the higher quality selection of features by this method. It was able to deliver scores within a narrower range than Lasso. But Anova surpasses this narrow range of Chi-square suggesting a more precise feature selection. Thus, Anova emerges more robust than Chi-square.

The sixth method namely Priority technique of feature selection surpasses the performance of all other feature selection techniques. It was able to produce the best results out of all with minimum amount of features. Such performance was expected due to its simple and highly effective mechanism. It utilized the results of all techniques to perform ranking and thus provided best set of features. This work provides useful insights and suggests a robust and secure feature selection method after the analysis study.

5 Conclusion

This research aimed to study the behaviour of different techniques to acquire useful insights and thus form a model with best tools that promises desired performance. It achieves this task through utilization of some smart techniques and by insightful analysis throughout the experiment. Several scenarios of different feature set size were studied. The experiment with feature sets led to the conclusion that a model needs to be tested with different feature sets to arrive at the most efficient solution. Otherwise, big differences having high impact on efficiency can be overlooked. In this study, general behaviour displaying high quality of smaller sets was largely noticed. Best feature set was agreed to be that of 200 features attained through Priority method which is very efficient when compared with original feature set size. This method attained maximum performance with every classifier. Logistic Regression and Priority method of feature selection were concluded to be the most appropriate combination in terms of high accuracy, high reliability and modern-day relevance. Both of these together make up for a good quality prediction model. Quality performance of Priority method also suggests that combining the results of all features selection techniques is very effective method and delivers more reliable and accurate prediction. Evident from the analysis, this study yielded many insights and produced quality performance.

References

1. Poduri, R. (ed.): Drug Discovery and Development. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-5534-3
2. Hooda, N., Bawa, S., Rana, P.S.: B2FSE framework for high dimensional imbalanced data: a case study for drug toxicity prediction. Neurocomputing **276**, 31–41 (2018)
3. Hooda, N., Bawa, S., Rana, P.S.: Fraudulent firm classification: a case study of an external audit. Appl. Artif. Intell. **32**(1), 48–64 (2018)
4. Hooda, N., Bawa, S., Rana, P.S.: Optimizing fraudulent firm prediction using ensemble machine learning: a case study of an external audit. Appl. Artif. Intell. **34**(1), 20–30 (2020)
5. Bhardwaj, R., Hooda, N.: Prediction of pathological complete response after neoadjuvant chemotherapy for breast cancer using ensemble machine learning. Inform. Med. Unlocked **16**, 100219 (2019)
6. Zelaya, C.: Towards explaining the effects of data preprocessing on machine learning. In: IEEE 35th International Conference on Data Engineering (ICDE), pp. 2086–2090 (2019). https://doi.org/10.1109/ICDE.2019.00245
7. Celik, O., Hasanbasoglu, M., Aktas, M., Kalipsiz, O., Kanli, A.: Implementation of data preprocessing techniques on distributed big data platforms. In: 4th International Conference on Computer Science and Engineering (UBMK) (2019). https://doi.org/10.1109/ubmk.2019.8907230
8. Suto, J., Oniga, S., Sitar, P.P.: Comparison of wrapper and filter feature selection algorithms on human activity recognition'. In: 6th International Conference on Computers Communications and Control (ICCCC), pp. 124–129. https://doi.org/10.1109/ICCCC.2016.7496749, (2016)
9. Dhote, Y., Agarwal, S., Deen, A.J.: A survey on feature selection techniques for ınternet traffic classification. In: International Conference on Computational Intelligence and Communication, pp. 1375–1380 (2015). https://doi.org/10.1109/CICN.2015.267

10. Shi, H., Liu, S., Chen, J., Li, X., Ma, Q., Yu, B.: Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure. Genomics **111**(6), 1839–1852 (2019)
11. Redkar, S., Mondal, S., Joseph, A., Hareesha, K.S.: A machine learning approach for drug-target ınteraction prediction using wrapper feature selection and class balancing. Molecul. Inform. **39** (2020). https://doi.org/10.1002/minf.201900062
12. Clifford, T., Bruce, J., Ajayi, T.O., Matta, J.: Comparative analysis of feature selection methods to ıdentify biomarkers in a stroke-related dataset. In: IEEE Conference on Computational Intelligence in a Stroke-Related Dataset, pp. 1–8 (2019). https://doi.org/10.1109/CIBCB.2019.8791457
13. Zhao, K., So, H.C.: Drug repositioning for schizophrenia and depression/anxiety disorders: a machine learning approach leveraging expression data. IEEE J. Biomed. Health Inform. **23**(3), 1304–1315 (2019)
14. Marouf, A., Hasan, K.Md., Mahmud, H.: Comparative analysis of feature selection algorithms for computational personality prediction from social media. IEEE Trans. Comput. Soc. Syst. (2020). https://doi.org/10.1109/TCSS.2020.2966910
15. Chakrapani, H.B., Chourasia, S., Saha, A., Swathi, J.N.: Predicting performance analysis of system configurations to contrast feature selection methods. In: International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), pp. 1–7 (2020). https://doi.org/10.1109/ic-ETITE47903.2020.106
16. Zhi, X., Yuexin, S., Jin, M., Lujie, Z., Zijian, D.: Research on the Pearson correlation coefficient evaluation method of analog signal in the process of unit peak load regulation. In: 13th IEEE International Conference on Electronic Measurement & Instruments (ICEMI), 522–527 (2017). https://doi.org/10.1109/ICEMI.2017.8265997
17. Bachu, V., Anuradha, J.: A review of feature selection and its methods. Cybern. Inf. Technol. **19**(1), 3 (2019)
18. Chen, J., Li, T., Zou, Y., Wang, G., Ye, H., Lv, F.: An ensemble feature selection method for short-term electrical load forecasting. In: IEEE 3rd Conference on Energy Internet and Energy System Integration (EI2) **170**, 22–29 (2019). https://doi.org/10.1016/j.apenergy.2016.02.114
19. Ikram, S.T., Cherukuri, A.K.: Intrusion detection model using fusion of chi-square feature selection and multi class SVM. J. King Saud Univ. Comput. **29** (2017). https://doi.org/10.1016/j.jksuci.2015.12.004
20. Doan, D.M., Jeong, D.H., Ji, S.: Designing a feature selection technique for analyzing mixed data. In: 10th Annual Computing and Communication Workshop and Conference (CCWC) (2020). https://doi.org/10.1109/CCWC47524.2020.9031193