

Ensemble Filter-Embedded Feature Ranking Technique (FEFR) for 3D ATS Drug Molecular Structure

Yee Ching Saw¹, Zeratul Izzah Mohd Yusoh², Azah Kamilah Muda^{3*} and Ajith Abraham⁴

^{1,2,3} Computational Intelligence and Technologies Lab (CIT Lab)
Faculty of Information and Communication Technology,
Universiti Teknikal Malaysia Melaka,
Hang Tuah Jaya, Durian Tunggal,
76100 Melaka, Malaysia.
yycaw@hotmail.com; zeratul@utem.edu.my; azah@utem.edu.my

⁴ Machine Intelligence Research Labs (MIR Labs)
Scientific Network for Innovation and Research Excellence,
Auburn, WA, USA.
ajith@ieee.com

Abstract: The concern for illicit abused and trafficking of ATS drugs are continuously growing. This is due to the evolving of new and unfamiliar ATS drugs, present a significant challenge to the forensic staff and laboratory testing. This paper aims to explore the use of machine learning method in the 3D molecular structure of ATS drug identification. In order to perform the computational analysis, the 3D molecular structure of ATS drugs will be illustrated in the voxel format of data representation. This paper proposes a new ensemble feature selection technique of Filter-Embedded Feature Ranking Techniques (FEFR), which is the combination of the filter method (ReliefF) and embedded methods (Variable Importance based Random Forest). It is used to identify a subset of significant features with highly discriminative power in representing the molecular structure of ATS drugs. These selected significant features eventually improve the performance of identification task.

Keywords: Ensemble Feature selection, Filter- Embedded Feature Ranking Techniques (FEFR), ATS drug identification, Machine learning

I. Introduction

Abused of ATS drug is one of the most worrisome problems that causes a major impact to the societies and nations. This may be due to the easily available and widespread of illicit manufacture of ATS drug. The biggest amphetamine and methamphetamine manufacture in the world is in North America, South East Asia, and the Middle East. The study that conducted by UNODC organization shows that the illicit usage of ATS is widely expanding across the country and significantly between the years of 2009-2013 [1].

In addition, with the emergence of the new and wide range of unfamiliar ATS, an analysis of illicit ATS drugs is crucial. Illicit drug analyze is typically involves the process of identification and quantitation of the sample materials in order

to support the judicial process [2]. The existing new drug discovery process is a lengthy and costly process. Generally, it takes between ten and fifteen years of research for a single new drug discovery which costs about 1.8 billion dollars [3]. Aside from that, it requires a timely exchange of analytical data between laboratories and law enforcement authorities at the national, regional and international levels [4]. Moreover, the existing drug testing process only account for a specific number of existing drugs due to their complicated analytical process. As the patterns of drug abuse continually evolved, the extension of such process is essential so that a wider range of drugs can be taken into accounts.

The remainder of the paper is organized as follows. The next section briefly describes the machine learning concept the material and methods that used in this study, which includes the dataset description and overview concepts of the proposed method selection. Section 3 presents simulation experiment and results, which including performance measurements and result and discussion. Conclusions of this study are presented in the last section.

II. Machine Learning

Machine learning is a program that enables the computer to learn and analyze data by learning from the past experience [5]. It can act as a tool that helps in solving diagnostic and prognostic problems in drug discovery domain. Based on the literature, it has successfully proved to be benefitted in several domain areas such as text detection [6], image recognition [7], medical [8] and etc. In particular, within the drug discovery process, such as virtual screening, Quantitative Structure-Activity Relationship (QSAR) and Quantitative Structure-Property Relationship (QSPR) and prediction of protein structure, function and interaction [9]. Generally, the drug discovery involves screening large chemical libraries for

promising hits, and translating the hits into leads. The leads will be optimized into a drug candidate for further validation by clinical trials development [10].

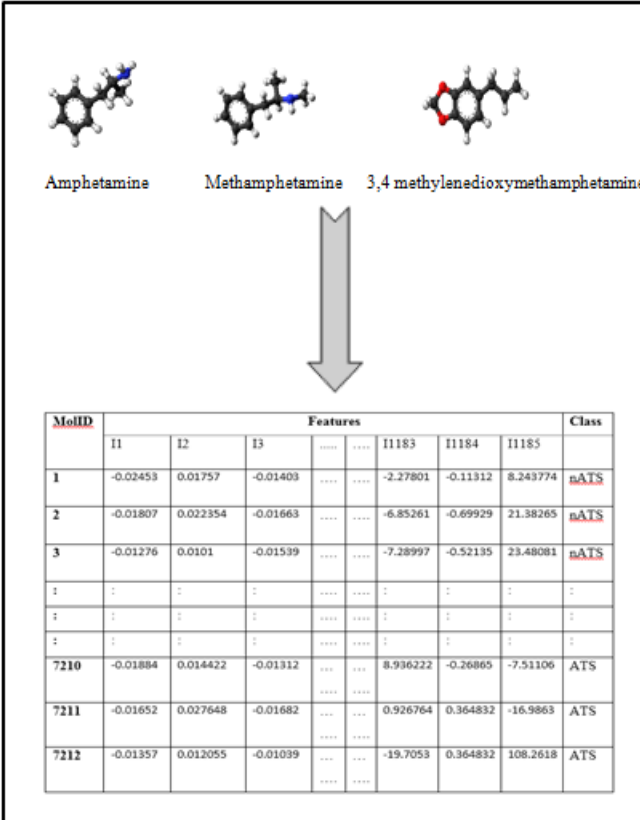
The rationale of adopting machine learning solutions in drug discovery domain is their capability to derive pattern from the input dataset, which can provide some basic knowledge for the development of approximation about the behavior of the samples [5]. Most of the machine learning methods are designed to process and learn from plenty of input features. It identifies the relation between the subset of features that represent the structure of a molecule compound. It will then learn to make an inference on the possible output for a new input set of given property. Therefore, these data-driven methods are considered as an appropriate alternative solution to identify the highly complex and non-linear patterns in ATS drugs dataset, which will then be used for prediction, detection, and identification of unknown or unfamiliar ATS drugs substances.

The fundamental aspect to be considered in this study is how to virtually represent the ATS drugs substances in machine learning for computational analysis? Generally, drug substances can be represented based on their chemical structure which was also known as molecular structure shape. There are two types of structural representation to virtualise the ATS drug molecular structure which is two-dimensional (2D) molecular structure and three-dimensional (3D) molecular structure. Both 2D and 3D representation have their benefits and limitations respectively.

Traditionally, two-dimensional (2D) is the simplest molecular representation that has been widely applied to represent the chemical molecules. It presents in a compound and shows a complete structure on how each of the molecules bonded together. However, 2D representation is failed to predict the activity differences between the chemical compounds. This limitation can be overcome by adopting 3D representation. 3D representation is able to give a precise and clear information of the atom interaction between the binding affinity and the target proteins. However, a key challenge in 3D representation is it requires larger storage space and computational time compared to 2D representation [11]. Therefore, it is crucial to find a good representation (i.e. good feature subset) to perform machine learning task. Hence, thereby, the key concern of this study is to discover the unique features that exist in the ATS drug molecular structure by using feature selection methods. The discovery process for these unique features is indirect. This is because, in order to perform analysis on this sample, the 3D molecular structure of ATS drugs is represented in terms of voxel (volumetric pixel) data. Table 1 depicted the voxel representation of 3D molecular structures of ATS drugs.

With that, we are looking for voxel sites that containing high discriminative information that can best represent the ATS drugs. The selected feature subset will then validate based on the identification performance. The metric that is used widely to evaluate the quality of selected feature subset is based on the identification performance. A high identification performance that yields from the learning algorithm means the selected features are good. This is used to demonstrate the ability of the selected features to distinguish the class label associated with the sample data. This result will be used to

illustrate the solving capability of far-reaching problems, such as the deficiency in the traditional laboratory process.



MoID	Features						Class	
	I1	I2	I3	I1183	I1184	I1185		
1	-0.02453	0.01757	-0.01403	-2.27801	-0.11312	8.243774	ATS
2	-0.01807	0.022354	-0.01663	-6.85261	-0.69929	21.38265	ATS
3	-0.01276	0.0101	-0.01539	-7.28997	-0.52135	23.48081	ATS
1	:	:	:	:	:	:	:
1	:	:	:	:	:	:	:
1	:	:	:	:	:	:	:
7210	-0.01884	0.014422	-0.01312	8.936222	-0.26865	-7.51106	ATS
7211	-0.01652	0.027648	-0.01682	0.926764	0.364832	-16.9863	ATS
7212	-0.01357	0.012055	-0.01039	-19.7053	0.364832	108.2618	ATS

Table 1 Voxel representation of 3D ATS drug molecular structure

A. Gaps

Several kinds of research have been done such as [12]–[14] to identify the molecular structure of chemical substances in the literature. However, to the best of our knowledge, it is hard to find research work and references, specifically on ATS drug molecular structure's identification that adopting computational intelligence as an approach. Several computerized assessments have been introduced to address this particular problem of identifying chemical substances. To assess the similarity of these chemical substances, a numerical representation of chemical substance is required. A review of techniques that transform a chemical substance into numerical representations has been discussed by Nikolova and Jaworska [15]. Since there are thousands of compounds present in one drug element, the dataset which is the output from the feature extraction phase will be complex and large in size. Due to the high dimensionality of the dataset, feature selection is often conducted to select the most optimal features subset and obtain informative insights into the compounds. By performing feature selection, a better understanding of a dataset, a faster and more cost-effective predictor can be formed.

B. Feature Selection

Feature selection has become an active research area for decades and has been proven in both theory and practice application in the various domains [16]–[18]. The main

objective of feature selection is to select relevant feature subset with reduced size from the original dataset, at the same time does not decrease the classification accuracy significantly [16], [19], [20]. Feature selection can be broadly categorized into three groups, which are: filter, wrapper, and embedded methods [21]. Filter methods evaluate the feature subset based on the relevancy of the features correspond to the class label without involving induction algorithm. On the contrary, wrapper methods will select the feature subset based on the estimation of the learning method's accuracy using induction algorithm. Whereas, embedded methods evaluate the feature subset by embed the feature selection in the process of classifier construction. All of these three feature selection methods are designed to perform the same tasks with different design methodology. Each of them has their own advantages and weakness in different aspects such as model complexity, computational efficiency, and time efficiency.

In general, filter methods have the lowest model complexity, computational efficiency as well as time efficiency. This is because filter methods are independent and do not require any model learning. Nevertheless, filter methods have some limitations such as it is failed to encounter the dependencies between the features and do not have interaction with the classifier. On the other hand, wrapper method are generally has highest model complexity, computational efficiency and time efficiency among the three methods. This is due to the wrapper methods do take into account the interaction with classification algorithm, and iteratively perform cross-validation procedure on the learning model. Embedded method is introduced to complement the weakness of wrapper by incorporate the learning scheme in the feature selection model to reduce the cross-validation procedure and indirectly speed up the evaluation process. Hence, the goal of this paper is to propose an ensemble of filter and embedded approach in order to explore the knowledge and advantages of each approach, while mitigating their weakness.

By performing ensemble method, different opinions and knowledge from different feature selection methods can take into consideration before making a decision. The key principle in forming an ensemble feature selection technique is composed of two main steps. Firstly, the components that used to ensemble must be determined (individual feature selection technique that will form the ensemble). Second, determine the method that will be employed to aggregate the result from each individual feature selection technique to one seamless whole result, which also referred as combination method. According to (Santana et al., 2007), there are three main strategies to aggregate the result, which is: fusion-based, selection-based, and hybrid methods. Fusion-based method will utilize the result of each individual components to produce a final outcome. On the other hand, the selection based methods will select only one of the most suitable technique to produce the final outcome. In the case of hybrid methods, both selection and fusion are adopting to complement each other to produce a more robust final result. In the case of our research, fusion-based ensemble method will be chosen to employ in our research due to our aim to retain the advantages of each component in the ensemble. The next section will review several previous studies that used

ensemble concepts in solving different problems in various fields.

C. Related Works

One of the earliest studies that done in the study by [22], multiple feature ranking techniques had been ensemble to resolve the problems in the domain of text classification. Three well-known filter feature ranking techniques have been employed in this study, such as document frequency thresholding, information gain, and the Chi-square method (χ^2_{max} and χ^2_{avg}). The experiment results shown that the proposed ensemble feature selection technique can achieve better performance in term of R-precision and microaveraged F1 compared to their non-combined feature selection counterparts.

In a later study by [23], explored a multi-criterion fusion-based recursive feature elimination (MCF-RFE) algorithm which composed of three scored-based filters feature ranking techniques, like Fisher's ratio, Relief, asymmetric dependency coefficient (ADC) and one embedded method which is absolute weight of Support Vector Machine (AW-SVM). The goal of this study is to enhance the stability and classification performance of the feature selection method. The performance of the proposed method (MCF-RFE) is evaluated by comparing to the benchmark of the SVM-RFE algorithm in term of classification error, the standard deviation of error estimation and feature stability. The results showed that MCF-RFE has good stability as compared to the benchmark SVM-RFE algorithm.

[24] introduced a general framework of ensemble feature ranking which composed of six filter ranking techniques with four different ranking aggregation procedures, which are Borda (BC), Condorcet (CD), Schulze (SSD) and Markov Chain (MC4). The effectiveness of the proposed technique is evaluated using 39 datasets that acquired from UCI. This Study employed three performance measurement to assess the classification performance of three chosen classifiers. The findings from the experiment showed that the SSD ranking aggregation method performs the best among the four aggregation methods.

[25] introduced a Global Optimisation Approach (GOA) to identify prominent features across several network traffic datasets in term of both spatial and temporal domains. GOA works by using six filter feature ranking techniques to rank each of the feature based on their frequency count. The optimal features will then select by using a cut-off that will discriminate from the unstable features. The goodness of the features will then assess by using a Random Forest framework. The findings of this experiment proved that GOA is a promising approach in terms of accuracy and stability in solving traffic classification problems.

After that, [26] also studies an iterative ensemble feature selection framework for solving the imbalance problem in the multiclass microarray dataset. They investigate the hybrid of two sampling methods (undersampling and oversampling), and three filter feature selection methods which include filter ranking, fast correlation-based filter selection (FCBF) and minimum redundancy maximum relevance (MRMR). They examined the performance of the proposed IRFS framework using six gene microarray data sets. The classification performance results show that the proposed framework

outperform other representatives state-of-the-art filter feature selection methods.

In this paper, we will address performing the feature selection by exploring the well-known ReliefF and Variable Importance based Random Forest (VI-RF). These techniques are based on the feature ranking criterion, which is a simple and efficient algorithm which have been widely applied in cheminformatics and used to analyze high-dimensional data and several improvements have been recently suggested [27]–[31]. Despite feature selection techniques have being used widely in cheminformatics, it has yet been implemented in ATS drugs identification. Therefore, in this research, both ReliefF and VIRF methods will be combined and apply to obtain the most optimal feature subset of unique characteristics features for ATS drugs identification.

III. The Material and Method

A. Data Collection

The dataset that used in this analysis is from ICGEB CRP Research Grant Programme Projects [32]. This data source contained 7212 sample records, which are 3602 of non-ATS drug molecular structure and 3610 of ATS drug molecular structure. Each instance is described by a fixed number of features, along with a class label. The features are recorded in voxel which aims to maintain the realistic properties of the 3D ATS molecular structure [33]. Voxel data are describe in the format of decimal value. This data source is used to train and test the proposed feature selection algorithm in this work. The characteristics of these datasets are presented in Table 2.

MolID	Features									Class
	I1	I2	I3	I1183	I1184	I1185		
1	-0.0245	0.01757	-0.0140	-2.2780	-0.1131	8.2437		n-ATS
2	-0.0180	0.02254	-0.0166	-6.8526	-0.6993	21.3826		n-ATS
3	-0.0127	0.0101	-0.0153	-7.2899	-0.5214	23.4808		n-ATS
:	:	:	:	:	:	:		:
:	:	:	:	:	:	:		:
:	:	:	:	:	:	:		:
7210	-0.0188	0.01442	-0.0131	8.9362	-0.2686	-7.51106		ATS
7211	-0.0165	0.02765	-0.0168	0.92676	0.36483	-16.9863		ATS
7212	-0.0135	0.01201	-0.0103	-19.7065	0.364832	108.2618		ATS

Table 2 Description of Dataset Used

B. Proposed Method

i. ReliefF

ReliefF is an approach which was extended by Kononenko in the year of 1994 to address the limitation of noisy and incomplete data and two-class classification problem [34]. The original Relief was introduced by Kira and Rendell in the year of 1992. The basic idea of ReliefF is to select a sample instance at random and search two nearest neighbors: one from the same class (nearest hit) and one from the different class (nearest miss). Then it will update the feature weighting vector according to the two nearest neighbors. The quality estimation of the selected features is based on the weight computation of the probability between the selected instances and their two nearest neighbors (nearest hit and nearest miss).

The rationale of this idea is a feature is considered good when the probability of two nearest neighbors from the same class having the same value. Meanwhile, the features with their nearest neighbors from two different classes should have different values. Therefore, the larger the different between this probability, the better the features. The final output of ReliefF is a ranked list that sorted in descending order and the top ranked features is selected as the optimal features for the candidate solution.

ii. Variable Important based Random Forest (VI-RF)

VI-RF is an embedded feature selection technique, which selects the relevant features based on the variable importance yielded by random forest. In the context of random forest which made of an ensemble of decision trees, Breiman in the year of 2001 proposed a permutation test procedure in order to compute variable importance based on the classification error [35]. The difference in classification accuracy caused by the permutation is taking into account to define the variable importance. The prediction accuracy won't be affected by permuting the values of the variable that consists of purely random noise. Formally, the variable importance using random forest is computed based on two main principles: randomization and out-of-bag error (OOB) estimates. Let $B^{\wedge}(t)$ be the out-of-bag (OOB) sample for a tree, with $t \in \{1, \dots, ntree\}$. The importance measure for variable X_j in tree t is precisely defined as follows:

$$VI^{(t)}(x_j) = \frac{\sum_{i \in B^{\wedge}(t)} I(y_i = \hat{y}_i^{(t)})}{|B^{\wedge}(t)|} - \frac{\sum_{i \in B^{\wedge}(t)} I(y_i = \hat{y}_{i,\pi_j}^{(t)})}{|B^{\wedge}(t)|} \quad (4)$$

where $\hat{y}_i^{(t)} = f^{(t)}(x_i)$ is the predicted class for observation i before, and $\hat{y}_{i,\pi_j}^{(t)} = f^{(t)}(x_{i,\pi_j})$ the predicted class for observation i after permuting its value of variable X_j , i.e. $x_{i,\pi_j} = (x_{i,1}, \dots, x_{i,j-1}, x_{\pi_j(i)}, x_{i,j+1}, \dots, x_{i,p})$. Note that $VI^{(t)}(x_j) = 0$ by definition, if variable X_j is not in tree t .) The raw variable importance score for each variable is then computed as the mean importance over all trees: $\frac{\sum_{t=1}^{ntree} VI^{(t)}(x_j)}{ntree}$.

iii. Ensemble Method

Ensemble method is one of the active research within the machine learning area. An ensemble method is a combination of two or more learning algorithms through a

voting scheme to make a decision [36]. The proposed method of this study is to ensemble the two chosen feature selection techniques, by aggregate the knowledge of these two techniques. The proposed method can be summarized as follows:

1. Rank all the features using the two selected feature selection technique.
2. Identify the top ranked features subset from two selected feature selection technique.
3. Identify the common or overlap features between the feature subset of two top ranked lists

For better understanding, a flow chart that illustrates the idea behind our proposed method is presented in the Figure 1. Moreover, the algorithm of the proposed method is presented in Table 3.

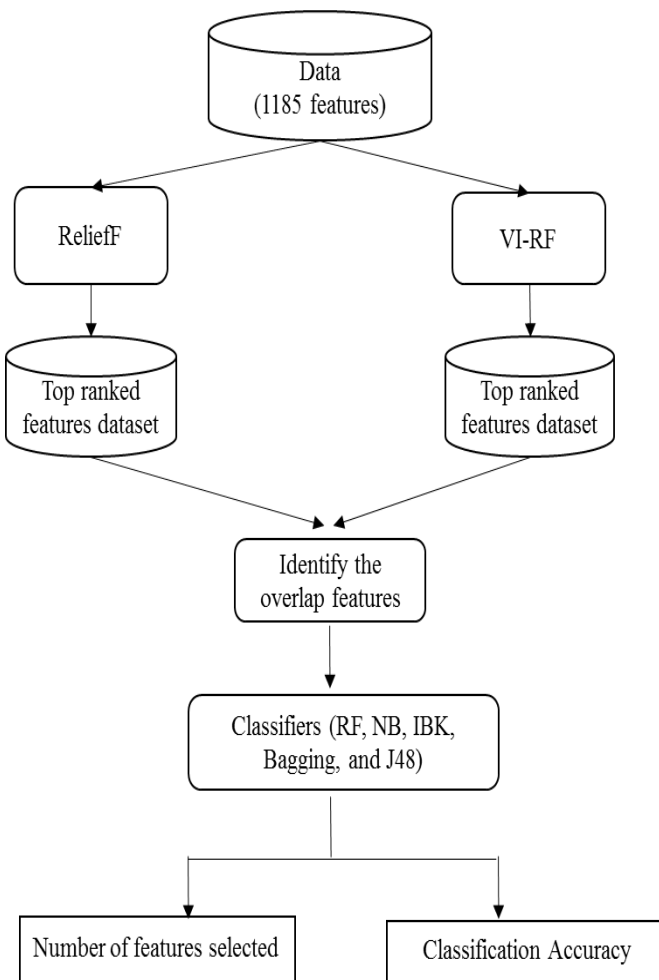


Figure 1. Flowchart for proposed method

Algorithm 1. Ensemble Component

Algorithm 1.1 (Filter and embedded feature ranking methods)

- Step 1: Let X_i be the feature set in the given dataset, where $X_i = \{X_1, X_2, X_3, \dots, X_{n-1}\}$ and C_i represents the class (i.e. ATS or non-ATS), where $C_i = \{C_1, C_2\}$.
- Step 2: For each Relief F filter method and VIRF embedded method, rank and sort the features X_i according to its importance in determining the output class C_i .
- Step 3: For both ranking list return by ReliefF and VIRF technique, select the top ranking features until further addition of the features degrade the classification performance
- Step 4: Output X' ; for each Relief F and VIRF feature ranking method.

Algorithm 1.2 (Ensemble selection)

- Step 1: Combine selected output features X' ; of both ReliefF and VIRF method.
- Step 2: Identify and select the overlap features and remove the remaining features

Table 3 Algorithm for The Proposed Method

C. Classification algorithms

The performance of feature selection algorithm is evaluated by performing classification task. Classification is the process of predicting an unknown property based on the feature subset which corresponding to class labels. These tasks is performed in Weka environment using the default setting [37]. For this, five different learning algorithm are chosen. This is motivated by the “No free Lunch Theorem”, which means there is no one algorithm that can guarantee to works best for every problems [38]. Hence, several classification algorithm are employ in this work to get an overview of the results on different feature selection techniques. The five chosen learning algorithm are Random Forest, Naives Bayes, IBK, Bagging, and J48.

i. Random Forest (RF)

Random forests [35] are referred as an Hybrid method, which composed of multiple decision trees. In order to perform classification task on new data, each data points is evaluated with each of the tree in the forest. The result from the individual tree predictor is stored. The forest will choose the feature subset having the most votes.

The different between random forests with Bagging method [39] is that each tree in bagging method is creating from a random bootstrap sample of the original dataset. However, random forest will estimates the correlations between each tree and the attributes are used to implant the randomness into the resulting trees. Random forest have shown robust to the effect of noise. Furthermore, the unused example (out-of-bag) in each of tree can be used to estimate the error rate and internal correlation between the trees.

ii. Naive Bayes (NB)

Naives Bayes classifier is a simple yet efficient probabilistic classifier which utilize the Bayes’ theorem of conditional probability where it assume every feature to be class-conditionally independent [38]. In this learning algorithm, each instance is assume to be associate with a set of features and a class value is takes from a predefined set of values. Each feature will then be assumed to be class-conditionally independent, such as all the features x_1, x_2, \dots, x_n are conditionally independent to the class y .

iii. IBK

IBK is a k nearest neighbor classifier which utilized the normalized Euclidean distance under the lazy learners category in weka. The value k represent the number of neighbours. It works based on the hunch that the classification of an instance is almost similar to the classification of another nearest instance within the vector space. For example, this algorithm will start by ranking the neighbour X amongst a given set of N data (X_i, c_i) , $i = 1, 2, \dots, N$ associate with the class labels c_j ($j = 1, 2, \dots, K$) of the K most similar neighbours to predict the class for the unclassified vector X. In particular, the similarity of these instance is measure based on their Euclidean distance metric, and X will be assigned to the class label which gain the majority vote among the K nearest class labels [38].

iv. Bagging

Bagging is a meta-algorithm which Hybrid different classifier by using same data [39]. In other word, bagging also known as bootstrap aggregation. The idea behind this is to create a set of classifier with the bootstrap samples of the original data. The bagging algorithm works by assign a certain prediction to each bootstrap sample. The bagging will start by use a training set, A with size, n to generate a new training set, m. This process will continue repeatedly by selected uniformly random with replacement from A. Next, the selected classification algorithm is trained with training set m, and the result will be collected by averaging the voting from the overall prediction.

v. J48

Another term for J48 also known as C4.5, which is widely used in Weka [40]. This classifier is an extension of an ID3 decision tree algorithm which is based on the decision tree concept. The idea behind the decision tree is based on the hierarchical collection of rules that describe how break a huge collection of data into several groups based on their regularities. In C4.5 classifier, the decision tree works by learning from the training set repeatedly, and recursively partitioning the training examples according to the potential feature values in separating the classes.

IV. Result and Discussion

A. Performance Measurement

According to Common Scheme for the Evaluation of Forensic Software (COSEFOS) by Hildebrandt et al. [41] to evaluate the results of a forensic software, the results must be both reproducible and the potential error rate must be known. Therefore, the reliability of the feature selection will take into account to evaluate the efficiency of the proposed method. Reliability feature selection can be measured by the steadiness of a classifier's performance and the consistency in search for relevant features [42]. Both these aspect is essential to evaluate the result correspond to the evaluation scheme. In this case, steadiness of the classifier's performance is focused on the frequency and the significance of the possible errors that may occur. Meanwhile, the classifier's steadiness measure is fully dependent on the choice of feature selection methods. Hence, choosing right choice of feature selection methods is essential to ensure the steadiness of the classifier's performance.

Besides, the consistency of the feature selection methods is crucial as well. It focuses on the reproducibility and errors of the feature subset return by the feature selection algorithm. The ideas behind the consistency measure are to predict the class value of the instances, with the selected feature subset must be consistent. The requisite for the feature subset to be consistent is usually support with the criterion of finding a small feature set [43].

Therefore, the quality of the feature selection algorithms is measure by the number of selected features and classification accuracy of the selected features. In the early theoretical stages, classification accuracy was the most common performance metric that employed in evaluating classifier performance. Accuracy is usually expressed in percentage (%). It is the ratio between the total number of correctly classified instances and the total number of samples. It can be easily calculated using the formula as follows:

$$acc = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

Where True Positive (TP) is the number of correctly predicted positive examples. A True Negative (TN) is a number of correctly predicted negative example. A Type I error will occur when False Positive (FP) is a number of incorrect prediction that an example was positive when in fact it was negative. A Type II error False Negative (FN) will occur when the number of incorrect prediction that an example was negative when in fact it was positive. Table 4 illustrates the confusion matrix structure for a two-class problem, with positive and negative classes.

Total number of instances		Positive	Negative
		Actual class	
Predicted class	Positive	TP	FP
	Negative	FN	TN

Table 4 The Confusion Matrix Structure Returned By A Classifier

B. Experimental Results

In this section, we demonstrate the effectiveness of our proposed ensemble feature selection method. The experiments were conducted using two well-known feature selection methods, which are VIRF and ReliefF feature ranking algorithms. The results of classification accuracy were compared with the original dataset, and two baseline algorithms to show the efficiency and saliency of the selected feature subset by the proposed method.

i. Steadiness of classifier's performance

The selected features by each feature selection method are combined together in three different approaches to perform a qualitative comparison of different combining approaches to ensure the steadiness of our proposed FEFR feature ranking method based on the classification accuracy, which namely as Ensemble 1, Ensemble 2 and Ensemble 3 as depicted in Table 5. Meanwhile, the Ensemble 2 also represented our proposed FEFR feature ranking method. The number and area of selected features are marked and bolded.

The key observation that can be drawn from the experiment result is the feature subset of Ensemble 2 yields the highest identification accuracy, followed by Ensemble 1 and Ensemble 3. Ensemble 1 combining both resulted feature subset, including the overlap features which yields 416 features. Ensemble 2 selecting the overlap/ common features from the optimal feature lists which yield just 176 features. Whereas Ensemble 3 which combines both of the techniques in a way that removed the entire overlap features, this result from a feature set with 240 features. The average classification performance is measured by the average results of five chosen classifiers, which is RF, NB, IBK, Bagging, and J48.

Ensemble 1, which containing top features that selected by both ReliefF and VIRF techniques including overlap and non-overlap features achieved a moderate result of 77.745%, Whereas feature subset that resulted from Ensemble 3, without the overlap features, achieved the lowest classification accuracy of 75.380%. Ensemble 2 (our proposed method) which containing overlap features are the most significant features with the classification accuracy of 78.131%. This result proved that the overlap features that selected by both different techniques are capable of providing satisfactory discrimination results in classifying ATS drugs.

ii. Consistency of the feature selection methods

In this section, the feature set selected by proposed ensemble FEFR feature ranking methods were compared with

the original dataset and each individual conventional feature selection method before ensemble. The comparison results of classification accuracy are shown in Table 6 and associated with two histogram graphs that used to better visualize the classification results, Figure 2 and Figure 3.

The classification performance obtained by using the original dataset (76.414%) has improved to 78.131%, with only 176 numbers of features. Similarly, the classification performance obtained by using ReliefF (77.401%) and VI-RF (77.608%) has improved to 78.131%, with only 176 numbers of features when the ensemble feature selection is employed in the ATS drugs dataset. The identification of such features implies that they are sufficient to reflect the incidence of a particular ATS drug from non-ATS drugs. Hence, it can be seen that the overlap features selected by the ensemble of both ReliefF and VI-RF are more effective than the feature selected by the single conventional feature selection method.

Based on the findings, it is apparent that the FEFR feature ranking methods use fewer features to offers improved classification performance with respect to original dataset, and each individual feature selection method before ensemble. This show that, FEFR technique is more effective than the features selected by the conventional method. This is because both techniques have different ability to mitigate the potential effect of irrelevant features and selecting the most informative features by using the different ranking criteria. Thereby, the overlap features were considered as the most significant features that agreed upon the both techniques.

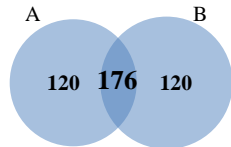
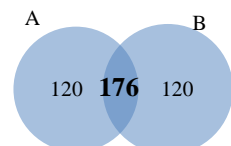
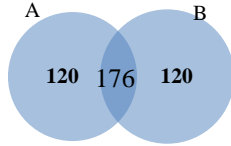
Ensemble Method	No. of selected features	Random Forest (%)	Naive Bayes (%)	IBK (%)	Bagging (%)	J48(%)	Avg (%)
Ensemble 1 	416	82.917	72.684	76.082	81.877	75.166	77.745
Ensemble 2 	176	82.862	73.253	76.844	81.226	76.47	78.131
Ensemble 3 	240	80.838	69.856	73.697	79.396	73.114	75.380

Table 5 Overall Classification Performance On ATS Drug Dataset Based On Three Different Ensemble Approaches

- *Notes: Ensemble 1: Aggregate both top ranked features
 Ensemble 2: Select overlap features
 Ensemble 3: Remove overlap features

	No. of selected features	Random Forest (%)	Naïve Bayes (%)	IBK (%)	Bagging (%)	J48 (%)	Avg (%)
Original dataset	1185	82.169	68.968	74.265	81.683	74.986	76.414
ReliefF	296	81.986	72.312	75.048	80.647	77.013	77.401
VI-RF	296	81.596	73.653	75.066	80.616	77.11	77.608
FEFR (Proposed method)	176	82.862	73.253	76.844	81.226	76.47	78.131

Table 6 Average Overall Classification Accuracy On ATS Drug Dataset Based On Different Feature Rankers

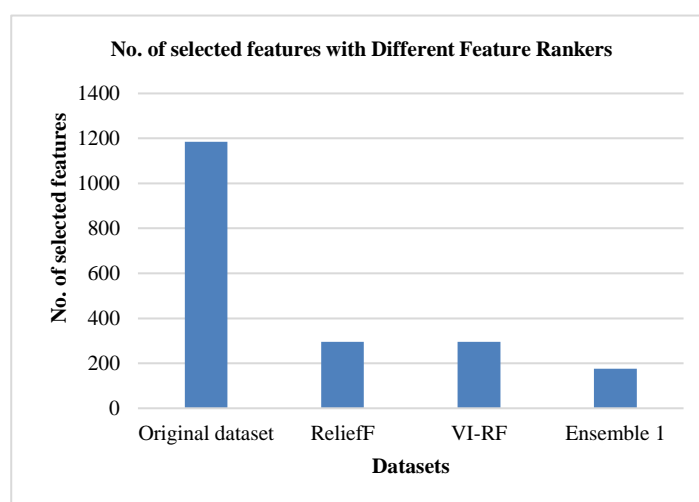


Figure 2 Comparison of classification model performance using different subset size selected by different feature ranking techniques

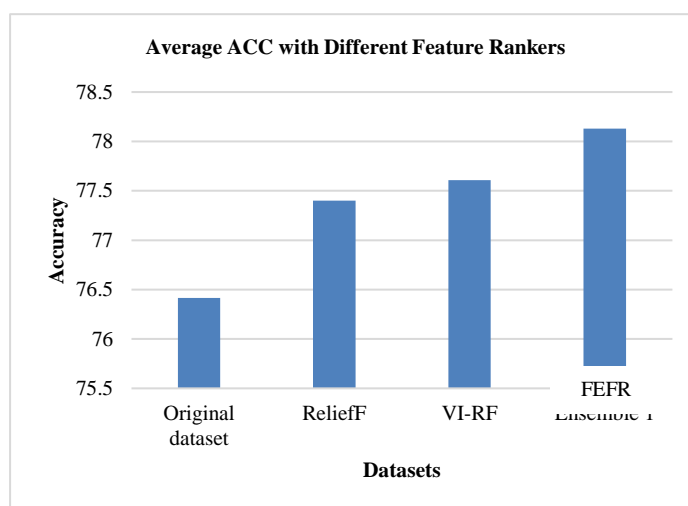


Figure 3 Comparison of classification model performance using different feature ranking techniques

V. Conclusions

In this study, an ensemble filter-embedded feature ranking (FEFR) scheme, which composed of the filter and embedded feature ranking methods, is proposed for the ATS drug identification problem. The overlap features that selected by both ReliefF and VI-RF feature selection algorithm is identified as features with high discriminative power. The ability of the selected overlap features in improved classification performance with a smaller number of features is assessed. The presented feature selection technique can be used for the automatic identification of ATS drugs and would help facilitate in early laboratory testing in the detection of specific ATS drug.

Acknowledgment

The authors would like to thank the Universiti Teknikal Malaysia Melaka and Faculty of Information Technology and Communication for providing excellent research facilities.

References

- [1] United Nations Office on Drugs and Crime, "The Challenge of Synthetic Drugs in East and South-East Asia and Oceania Trends and Patterns of Amphetamine-type Stimulants and New Psychoactive Substances," 2015.
- [2] Waters, *Drugs of Abuse Analysis Application Handbook*. 2007.
- [3] S. M. Paul *et al.*, "How to improve R&D productivity: the pharmaceutical industry's grand challenge," *Nat. Rev. Drug Discov.*, vol. 9, no. 3, pp. 203–14, 2010.
- [4] U. N. O. on D. and C. UNODC, "Recommended Methods For The Identification And Analysis Of Amphetamine, Methamphetamine And Their Ring-Substituted Analogues In Seized Materials," 2006.
- [5] E. Alpaydm, *Introduction to Machine Learning*, Second Edi. 2010.
- [6] W. Liao and Y. Wu, "An Integrated Approach for Multilingual Scene Text Detection," vol. 8, pp. 33–41, 2016.
- [7] T. Saitoh, T. Shibata, and T. Miyazono, "Feature Points based Fish Image Recognition," vol. 8, pp. 12–22, 2016.
- [8] M. David, M. Hirsch, J. Karin, E. Toledo, and S. Akselrod, "An estimate of fetal autonomic state by time-frequency analysis of fetal heart rate variability," *J. Appl. Physiol.*, vol. 102, no. 3, pp. 1057–1064, 2006.
- [9] N. Wale, "Machine learning in drug discovery and developmen," *Drug Dev. Res.*, vol. 72, no. 1, pp. 112–119, 2011.
- [10] J. P. Hughes, S. S. Rees, S. B. Kalindjian, and K. L. Philpott, "Principles of early drug discovery," *Br. J. Pharmacol.*, vol. 162, no. 6, pp. 1239–1249, 2011.
- [11] W. Shin, X. Zhu, M. G. Bures, and D. Kihara, "Three-Dimensional Compound Comparison Methods and Their Application in Drug Discovery," pp. 12841–12862, 2015.
- [12] C. Y. Tseng and J. Tuszynski, "A unified approach to computational drug discovery," *Drug Discov. Today*, vol. 20, no. 11, pp. 1328–1336, 2015.
- [13] R. Dutt and A. K. Madan, "Predicting biological activity: Computational approach using novel distance based molecular descriptors," *Comput. Biol. Med.*, vol. 42, no. 10, pp. 1026–1041, 2012.
- [14] P. Csermely, T. Korcsmáros, H. J. M. Kiss, G. London, and R. Nussinov, "Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review," *Pharmacol. Ther.*, vol. 138, no. 3, pp. 333–408, 2013.
- [15] N. Nikolova and J. Jaworska, "Approaches to Measure Chemical Similarity— a Review," *QSAR Comb. Sci.*, vol. 22, no. 910, pp. 1006–1026, 2003.
- [16] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.
- [17] A. Jovic, K. Brkic, and N. Bogunovic, "A review of feature selection methods with applications,"

- Ieee*, pp. 1200–1205, 2015.
- [18] J. B. O. Mitchell, “Machine learning methods in chemoinformatics,” *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, vol. 4, no. October, p. n/a-n/a, 2014.
- [19] P. Refaailzadeh, L. Tang, and H. Liu, “On comparison of feature selection algorithms,” *Proc. AAAI Work. Eval. Methods Mach. Learn. II*, pp. 34–39, 2007.
- [20] M. Sewell, “Feature Selection,” 2007. [Online]. Available: <http://machine-learning.martinsewell.com/feature-selection/feature-selection.pdf>.
- [21] I. Guyon, “An Introduction to Variable and Feature Selection Introduction,” vol. 3, pp. 1157–1182, 2003.
- [22] J. O. S. Olsson and D. W. Oard, “Combining feature selectors for text classification,” pp. 798–799, 2006.
- [23] F. Yang and K. Z. Mao, “Robust feature selection for microarray data based on multicriterion fusion,” *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 8, no. 4, pp. 1080–1092, 2011.
- [24] R. C. Prati, “Combining feature ranking algorithms through rank aggregation,” *Proc. Int. Jt. Conf. Neural Networks*, no. Cmcc, 2012.
- [25] A. Fahad and A. Harthi, “Designing an Accurate and Efficient Classification Approach for Network Traffic Monitoring,” 2015.
- [26] J. Yang, J. Zhou, Z. Zhu, X. Ma, and Z. Ji, “Iterative ensemble feature selection for multiclass classification of imbalanced microarray data,” *J. Biol. Res.*, vol. 23, no. 1, p. 13, 2016.
- [27] A. Palczewska, J. Palczewski, R. M. Robinson, and D. Neagu, “Interpreting random forest classification models using a feature contribution method (extended),” *2013 IEEE 14th Int. Conf. Inf. Reuse Integr.*, pp. 1–30, 2013.
- [28] A. L. Teixeira, J. P. Leal, and A. O. Falcao, “Random forests for feature selection in QSPR models - An application for predicting standard enthalpy of formation of hydrocarbons,” *J. Cheminform.*, vol. 5, no. 2, pp. 1–15, 2013.
- [29] A. S. Chen, N. J. Westwood, P. Brear, G. W. Rogers, L. Mavridis, and J. B. O. Mitchell, “A Random Forest Model for Predicting Allosteric and Functional Sites on Proteins,” *Mol. Inform.*, pp. 125–135, 2016.
- [30] R.P.L.DURGABAI, “Feature Selection using ReliefF Algorithm,” *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 3, no. 10, pp. 8215–8218, 2014.
- [31] N. Aniceto, A. A. Freitas, A. Bender, and T. Ghafourian, “A novel applicability domain technique for mapping predictive reliability across the chemical space of a QSAR: reliability-density neighbourhood,” *J. Cheminform.*, vol. 8, no. 1, p. 69, 2016.
- [32] A. K. Muda, “CRP - ICGEB Research Grants Completed in 2016 (A New 3D Descriptor of Synthetic Drug Molecular Structure for Drug Analysis - CRP/13/010),” 2012.
- [33] A. E. Kaufman, “Voxels as a computational representation of geometry,” *Comput. Represent. Geom.*, vol. d, 1994.
- [34] I. Kononenko, “ReliefF for estimation and discretization of attributes in classification, regression, and ILP problems,” *Artif. Intell. Methodol. Syst. Appl.*, pp. 1–15, 1996.
- [35] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [36] T. G. Dietterich, “Ensemble Methods in Machine Learning,” *Mult. Classif. Syst.*, vol. 1857, pp. 1–15, 2000.
- [37] I. H. Witten, E. Frank, and M. a. Hall, *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition*, vol. 54, no. 2. 2011.
- [38] R. O. Duda, P. E. Hart, and D. G. Stork, “Pattern Classification,” *New York: John Wiley, Section*. p. 680, 2001.
- [39] L. Breiman, “Bagging Predictors,” *Mach. Learn.*, vol. 24, no. 421, pp. 123–140, 1996.
- [40] J. R. Quinlan, “C 4.5: Programs for machine learning,” *Morgan Kaufmann Ser. Mach. Learn.*, 1993.
- [41] M. Hildebrandt, S. Kiltz, and J. Dittmann, “A common scheme for evaluation of forensic software,” *Proc. - 6th Int. Conf. IT Secur. Incid. Manag. IT Forensics, IMF 2011*, pp. 92–106, 2011.
- [42] K. F. and S. P. Nguyen, H.T., “Reliability in A Feature-Selection Process for Intrusion Detection,” *Reliab. Knowl. Discov.*, no. January, pp. 3–27, 2012.
- [43] A. Arauzo-azofra, “Consistency measures for feature selection.pdf,” vol. 22, pp. 1–22, 1997.

Author Biographies



Saw Yee Ching was born in Perak, Malaysia on November 26, 1991. She received her Bachelor of Computer Science in Interactive Media from the Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka. She is currently pursuing his Master of Science in Information and Communication Technology, also in Universiti Teknikal Malaysia Melaka. Her research interests are including intelligent system, pattern recognition and software engineering.



Zeratul Izzah Mohd. Yusoh is a lecturer in the Department of Intelligent Computing and Analytics, Faculty of Information & Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM) where she has been a faculty member since 2003. Zeratul completed her PhD in Queensland University of Technology, Australia. Her PhD work is focusing on SaaS Resource Management System in Cloud Computing using Evolutionary Computation. Zeratul's did her Master at University of Edinburgh, Scotland and her undergrad study is at Universiti Teknologi Malaysia (UTM), Johor. She is also a Certified IT Professional (Malaysia). Her research interests lie in the area of intelligent systems, with a focus on developing soft computing solutions to improve the system's quality. Through her past research in Dyslexia area, she has won two gold medals, one silver medal, and one bronze medal at both international and national levels. The product of the research has also been commercialized and applied for patent. In recent years, she has focused on evolutionary computation techniques, Software as a Service (SaaS) and Cloud computing area.



Azah Kamilah Muda is a Associate Professor at Faculty of ICT, UTeM. She has appointed as Deputy Dean of Post Graduate and Research since 2015. She received her PhD in 2010 from Universiti Teknologi Malaysia, specializing in image processing. Her research interest includes fundamental studies on data analytics using soft computing techniques, pattern analysis and recognition, image processing, machine learning, computational intelligence and hybrid systems. Her current research work is on pattern analysis of molecular computing for drug analysis, data analytic for various application and root cause analysis in manufacturing process.



Ajith Abraham received Ph.D. degree from Monash University, Melbourne Australia and a Master of Science Degree from Nanyang technological University, Singapore. His research and development experience includes over 17 years in the Industry and Academia spanning different continents in Australia, America, Asia and Europe. He works in a multi-disciplinary environment involving computational intelligence, network security, sensor networks, e-commerce, Web intelligence, Web services, computational grids, data mining and applied to various real world problems. He has authored/co-authored over 350 refereed journal/conference papers and book chapters and some of the papers have also won best paper awards at international conferences and also received several citations.