Classifying Smoking Urges Via Machine Learning

Antoine Dumortier

University of Pittsburgh, Department of Electrical and Computer Engineering, Benedum Hall, Pittsburgh, PA 15260, USA

Ellen Beckjord

University of Pittsburgh, Department of Psychiatry, 5115 Centre Avenue, Suite 140, Pittsburgh, PA 15232, USA

Saul Shiffman

University of Pittsburgh, Department of Psychology, 510 BELPB, 130 N. Bellefield Avenue, Pittsburgh, PA 15260, USA

Ervin Sejdić*

University of Pittsburgh, Department of Electrical and Computer Engineering, Benedum Hall, Pittsburgh, PA 15260, USA

Abstract

Background and Objective: Smoking is the largest preventable cause of death and diseases in the developed world, and advances in modern electronics and machine learning can help us deliver real-time intervention to smokers in novel ways. In this paper, we examine different machine learning approaches to use situational features associated with having or not having urges to smoke during a quit attempt in order to accurately classify high-urge states. **Methods:** To test our machine learning approaches, specifically, Bayes, discriminant analysis and decision tree learning methods, we used a dataset collected from over 300 participants who had initiated a quit attempt. The three classification approaches are evaluated observing sensitivity, specificity, accuracy and precision. **Results:** The outcome of the analysis showed that algorithms based on feature selection make it possible to obtain high classification rates with only a few features selected from the entire dataset. The classification tree method outperformed the naive Bayes and discriminant analysis methods, with an accuracy of the classifications up to 86%. These numbers suggest that machine learning may be a suitable approach to deal with smoking cessation matters, and to predict smoking urges, outlining a potential use for mobile health applications. **Conclusions:** In conclusion, machine learning classifiers can help identify smoking situations, and the search for the best features and classifier parameters significantly improves the algorithms' performance. In addition, this study also supports the usefulness of new technologies in improving the effect of smoking cessation interventions, the management of time and patients by therapists, and thus the optimization of available health care resources. Future studies should focus on providing more adaptive and personalized support to people who really need it, in a minimum amount of

*Corresponding author

Email addresses: and148@pitt.edu (Antoine Dumortier), beckjorde@upmc.edu (Ellen Beckjord), shiffman@pinneyassociates.com (Saul Shiffman), esejdic@ieee.org (Ervin Sejdić)

time by developing novel expert systems capable of delivering real-time interventions.

Keywords: Smoking urges, smoking cessation, machine learning, supervised learning, feature selection.

1. Introduction

There are 1.1 billion smokers in the world (15.4% of the world population), and this number is expected to increase to 1.6 billion over the next two decades [1]. Worldwide, tobacco use causes more than 5 million deaths per year [1]; that is to say, one person dies every six seconds from a tobacco related disease. Furthermore, current trends show that it will lead to the death of more than 8 million people annually by 2030 [1]. In the developed world, tobacco is currently the single largest preventable cause of death and diseases, and the overall mortality among both male and female smokers is about three times higher than that among similar people who never smoked [2]. In the United States, cigarette smoking kills more than 480,000 Americans each year, with more than 41,000 of these deaths from exposure to secondhand smoke. As a result, the economic consequences for the country are critical. Smoking-related illness in the United States costs more than $289 billion a year, including at least $133 billion in direct medical care for adults and $156 billion in lost productivity [2]. In 2012, an estimated 18.1% (42.1 million) U.S. adults were current cigarette smokers and 78.4% (33 million) of these adults smoked every day [3]. However, in 2011, 68.8% of adult cigarette smokers wanted to stop smoking completely [4], and 42.7% had made a quit attempt in the past year [2]. Unfortunately, quit attempts are typically unsuccessful, ending in relapse (resumption of smoking), usually precipitated by moments of intense craving or urge to smoke [5].

Studies based on ecological momentary assessment (EMA) [6, 7] have been conducted. EMA involves repeated sampling of subjects' real world mood, thoughts and state of mind at specific and random times during the day, through completion of assessments in subject's daily routine using mobile technology [7, 8]. In [9, 10], a subset of features previously associated with lapses is used to analyze how craving, emotion, and social environment impact on smoking rate [9]. In [10], self-reports of contextual variables were analyzed to examine correlates of craving when cigarettes were smoked. Results showed, for example, that craving was higher when cigarettes were smoked while eating or drinking, during activity, and early in the day. On the other hand, craving does not appear to be related with the location, alcohol, or caffeine. However, there is variability in the evidence regarding the degree to which different contextual features are associated with smoking risk. For example, during a quit attempt [11], self-reported temptation episodes (i.e., intense craving to smoke) were associated with negative moods, exposure to others smoking, and consumption of food, coffee, or alcohol.

These previous studies, as well as the recent advances in computational algorithms, has led us to believe that machine learning approaches can be useful in the process of smoking cessation. To test our hypothesis, we developed an algorithm used to predict subjects smoking urges, with a focus on the development of a

general static algorithm intended for preliminary clinical testing. Our first goal was to carry out a comparative analysis of machine learning algorithms to determine if a classifier can be able to provide smoking urges' classifications. Then, implementing a feature selection algorithm, the second goal was to extract the most relevant features (in a given dataset) that can provide the best classifications of urges to smoke.

Section 2 describes the methodology for the different phases depicted. Before giving a presentation and technical information about the three classification methods used in this study, an explanation of how the input data selection is made and how data are organized is given. Validation techniques, which are implemented to split the dataset between a training dataset, to create the model, and a testing dataset, to test the model, are then presented. We also operate and compare feature selection algorithms in an attempt to exclude useless features and only select those which provide the best results. Section 3 presents the classification results and the final selected features, and section 4 overviews the implications of these results. A conclusion is provided in Section 5 followed by a list of references.

2. Methodology

2.1. Data collection

The data were collected at the University of Pittsburgh from 1990 to 1995. 349 smokers seeking to quit smoking were recruited through a media advertising and participants reported their smoking behavior, urges to smoke, and contextual information (e.g., mood, location) using a hand-held device at scheduled and random intervals five times a day for up to six weeks. The final dataset included 41 parameters (also called features) (Appendix A) and 29,959 environment reports from 248 unique subjects. The smoking urges are evaluated according to the value of one discrete attribute, which represents the urge rating at any point in time. This variable has its values on a 0 to 10 scale: 0 is for the lowest smoking urge, while 10 is for an intense smoking urge. In order to simplify the classification, the urge rating variable has been converted to a binary number. 0 (negative cases) has been attributed to values less than 5, and 1 (positive cases) has been attributed to values greater than or equal to 5. Out of these 29,959 reports, 70% (21,070 exactly) of them are 0, while 30% (8,889 exactly) of the reports are 1. Our objective was to identify features associated with high urges (i.e., urge rating greater than or equal to 5).

2.2. Classification of smoking urges

Our unique dataset can be represented by the following form:

$$(\mathbf{X}, \mathbf{Y}) = (\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}, \ldots, \mathbf{X_k}, \ldots, \mathbf{X_n}, \mathbf{Y}) \tag{1}$$

where $\mathbf{Y}$ (binary column vector) is the target variable (the class) of the predictions (the urge rating). The matrix $(\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}, \ldots, \mathbf{X_n})$ represents the features used by the classifier (each $\mathbf{X_k}$ is a column of $\mathbf{X}$). $\mathbf{X}$ can

also be represented using the subject approach, $(\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \ldots, \mathbf{x_n})^T$ where each $\mathbf{x_k}$ is a line of $\mathbf{X}$. Therefore, we have:

$$\mathbf{X} = \begin{pmatrix} x_{A1} & x_{A2} & \ldots & x_{An} \\ x_{B1} & x_{B2} & \ldots & x_{Bn} \\ x_{C1} & x_{C2} & \ldots & x_{Cn} \\ \ldots & \ldots & \ldots & \ldots \end{pmatrix} = (\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n}) = \begin{pmatrix} \mathbf{x_1} \\ \mathbf{x_2} \\ \ldots \\ \mathbf{x_n} \end{pmatrix} \text{ and } \mathbf{Y} = \begin{pmatrix} y_A \\ y_B \\ y_C \\ \ldots \end{pmatrix}, y_k \in \{0, 1\}$$

2.2.1. The naive Bayes classifier

This method is based on the Bayes rule [12, 13], which provides that

$$P(\mathbf{Y} = y_i | \mathbf{X} = \mathbf{x_k}) = \frac{P(\mathbf{X} = \mathbf{x_k} | \mathbf{Y} = y_i) P(\mathbf{Y} = y_i)}{\sum_j P(\mathbf{X} = \mathbf{x_k} | \mathbf{Y} = y_j) P(\mathbf{Y} = y_j)} \tag{2}$$

where y_m is the mth row of $\mathbf{Y}$, that is to say the urge rating value for the mth subject, and x_k is the kth row in $\mathbf{X}$ (all the recorded values for the kth subject). The notation of (2) may be simplified as:

$$P(y_i | \mathbf{x_k}) = P(y_i) \times \frac{P(\mathbf{x_k} | y_i)}{\sum_j P(\mathbf{x_k} | y_j) P(y_j)} \tag{3}$$

In (3), given a value y_j of $\mathbf{Y}$, we assume that all the values of $\mathbf{x_k}$ are statistically independent of one another. With this assumption, we get

$$P(\mathbf{x_k} | y_i) = \prod_{w=1}^{n} P(x_w | y_i) \tag{4}$$

which leads to:

$$P(y_i | \mathbf{x_k}) = P(y_i) \times \frac{\prod_w P(x_w | y_i)}{\sum_j P(\mathbf{x_k} | y_j) P(y_j)} = P(y_i) \times \frac{\prod_w P(x_w | y_i)}{\sum_j P(y_j) \prod_w P(x_w | y_j)} \tag{5}$$

This is the crucial equation for the naive Bayes classifier [13] and even if the naive assumption is not entirely confirmed, good results can be expected from this method [14].

2.2.2. Discriminant analysis classifier

Similar to the naive Bayes classification method, the discriminant analysis assumes that the conditional probability density functions $P(\mathbf{x_k} | y_i = 0)$ and $P(\mathbf{x_k} | y_i = 1)$ are both normally distributed [15], with mean and covariance parameters $(\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0})$ and $(\boldsymbol{\mu_1}, \boldsymbol{\Sigma_1})$. In other words, the fitting function, which is used to generate the classifier, estimates the parameters of a Gaussian distribution for each class. Then a class ($y_i = 0$ or 1) is predicted by finding the smallest misclassification cost [16].

2.2.3. Decision tree learning

A decision tree is a structure used to predict the response (the class) to inputs. This method, regularly used in machine learning theory [17], is based on the construction of a binary tree where the nodes represent the tests made on the inputs. The results of the different tests give the direction to follow in the tree. Finally

the prediction can be read when a leaf node is reached [18]. In this paper we adopted classification trees, which provide binary (nominal) classification such as *true* (1) or *false* (0). These trees are very useful because they can provide easy to understand predictions in difficult conditions when many variables are present. Figure 1 shows an example of a classification tree with two classes ($Y = 1$ or 2) and two X variables (inputs). As can be seen, an advantage of the tree structure is its capability to deal with an important number of input variables, whereas a plot is limited to two (2 dimensions) or three (3 dimensions) input variables [19]. Further techniques, called *ensemble* methods [20], build more than one decision tree. We will deal with one of them, called *bagging* decision trees [21]. We decided to use bagging decision trees because it is amachine learning ensemble algorithm that is usually designed to improve the stability and accuracy of classification algorithms. Ensemble reduces variance and bagging algorithms are known to be an efficient way to minimize noise, bias, and variance, which are the main reasons for error in learning [22].

In the case where the classification is a three (or more)-class problem, the Support Vector Machine Decision Tree [23] proved to be an efficient alternative method. It allows to combine the computational benefits of the classification tree technique while keeping the precision of SVMs [24]. As described in [23], a high percentage of accuracy can be obtained with a multi-class classification problem, the computational time remaining very low.

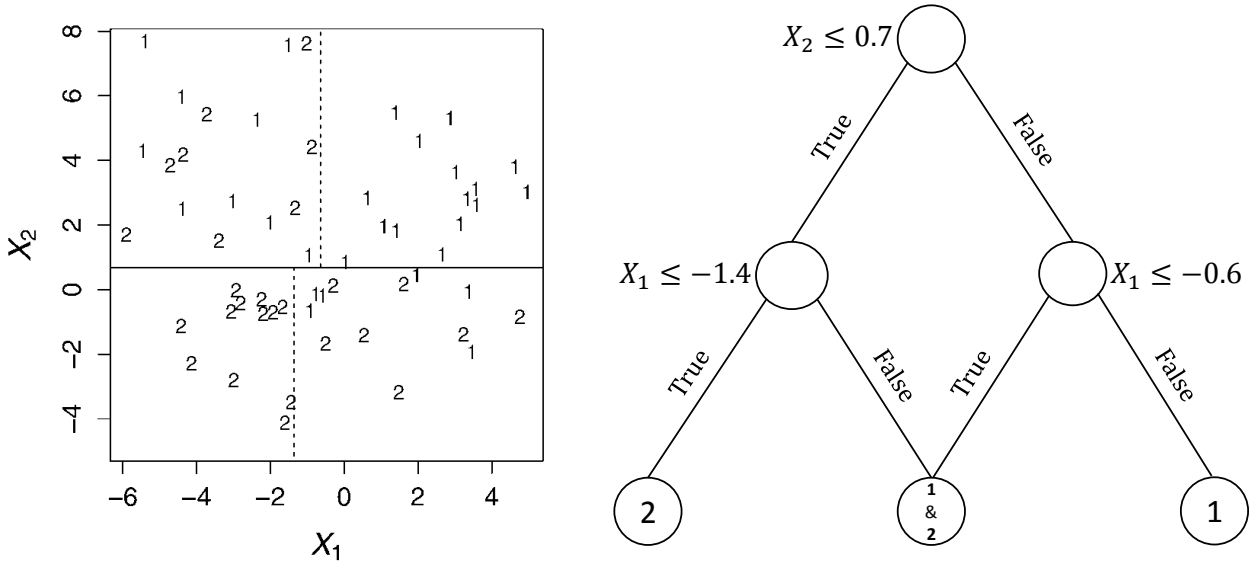


Figure 1: Partitions (left) and decision tree structure (right) for a classification tree model with two classes (1 and 2).

2.3. Dimensionality reduction and features selection

Reducing the number of features is an important challenge in machine learning. The dimensionality of the datasets used to perform data mining algorithms has increased a lot since data can be acquired automatically,

but not useful or correlated features can lead to overfitting when we deal with a large amount of data [25]. That is why it is necessary to use dimensionality reduction algorithms.

2.3.1. Principal component analysis

Principal Component Analysis (PCA) turns out to be an adequate technique to decrease the dimensionality of a dataset constituted of numerous interdependent variables [26, 27]. It can be characterized as an unsupervised algorithm because it disregards class labels. In essence, PCA attempts to reduce the dataset's dimension by finding a few orthogonal linear combinations, the principal components, of the original variables with the largest variance [28]. Those orthogonal components are then precisely organized in order to first extract those with the largest variation, and then to remove those having the smallest contribution to the dataset's variation [26]. In other words, the algorithm has to retain as much as possible of the variation already existing in the dataset.

2.3.2. Linear discriminant analysis

In contrast to PCA, Linear Discriminant Analysis [29, 30] is a supervised algorithm that computes the directions (the linear discriminants) that will represent the axes maximizing the separation between multiple classes. Although one would find obvious that LDA could provide better results than PCA for a multi-class classification task with known class labels (i.e., a supervised learning task), it is not always the case. Classical LDA makes a projection of the data onto a lower-dimensional vector space in order to maximize the ratio of the between-class distance to the within-class distance and achieve maximum differentiation [31]. Over the past years, LDA has received a lot of extensions.

2.3.3. Feature selection

Dimensionality reduction is not only useful to accelerate algorithm execution, but it actually helps with the final classification accuracy as well. Removing un-informative or dis-informative data can help the algorithm to achieve better performances on new data, and this can be done by searching for the best subset of attributes in a given dataset. The purpose of feature selection is firstly to preserve the relevant [25] features and to get rid of irrelevant and redundant features [32], [33]. A definition of a *relevant* feature can be found in [32]. Variable or feature selection can be potentially beneficial in many points. It facilitates the visualization and the understanding of the data [25], and can also reduce computation times and storage needs. However, its main purpose often remains to reduce a potentially high dimension probability distribution to a simpler one containing a smaller amount of features, whilst maintaining high performance by discarding those least useful [34]. This technique has been shown to increase the classification accuracy in numerous cases [33], [35].

As presented in [36], there are two principal ways to reduce the dataset's dimensionality: feature extraction and feature selection. The latter, chosen for this study and designed to select a subset of the existing features

without transforming them, has two variants (presented in figure 2) for selecting the best set of input variables that would maximize the model's accuracy and minimize variance [36, 37]:

- Filter Method: The subset selection procedure is independent of the learning algorithm. It uses a relevance criterion and it is generally a pre-processing step. It is mainly known as a fast and general method, but with a tendency to select larger subsets [36, 37, 38];
- Wrapper Method: The subset selection is based on the learning algorithm used to train the model. Unlike the filter methods, every subset is evaluated in the specific context of the learning algorithm. Its accuracy is an advantage, but the execution is slow and each classification problem requires a new subset selection [36, 39].

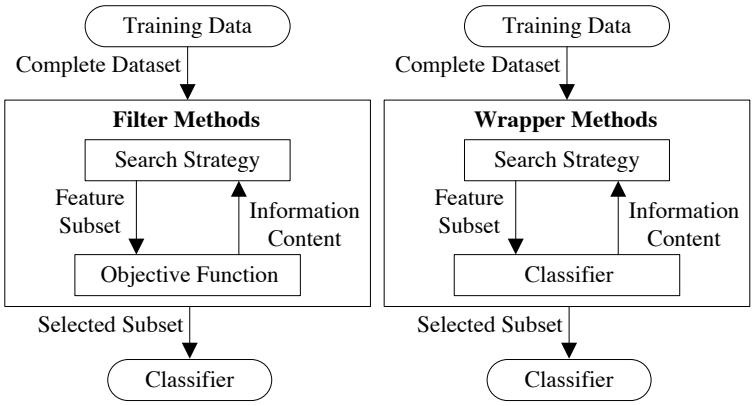


Figure 2: Summary of the two variants of a feature selection method: filters (left) and wrappers (right).

The interpretation of the principal components (in the PCA algorithm) or the linear discriminants (for the LDA) can sometimes be very complex. Although the new variables are uncorrelated and conceived as linear combinations of the original variables they do not necessarily coincide with meaningful physical quantities [28]. In this study, such loss of interpretability is not acceptable, that is why a sequential feature selection technique was chosen. It is a filter method that presents the key benefits of performing feature selection on our initial dataset, with easily interpretable results, allowing to identify which features is relevant to our specific smoking cessation problem. It selects a subset of features from the input data X that best predict the data in Y by sequentially selecting features until there is no improvement in the predictions. The improvement criterion is based on the misclassification rate, but a different criterion (e.g., classification accuracy, misclassification rate, specificity,...) can be adopted depending on the classification problem. The sequential feature selection uses three distinct sub-algorithms:

1. the search algorithm, that looks for the feature subset optimizing the criterion.
2. the evaluation algorithm, which evaluates the chosen criterion.

3. the performance function algorithm, which is in our case one of the three classifiers used in this study (i.e., naive Bayes, discriminant analysis, or classification tree).

2.4. Evaluation measures

Based on the outcome of a classification test, we calculated the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) [40], which were then used to calculate the following metrics:

- the **True Positive Rate** (TPR) (or **Sensitivity**, or **Recall**):

$$TPR = \frac{TP}{TP + FN} \tag{6}$$

  It describes the test's capability to identify a class correctly, and measures the proportion of positive subjects correctly identified as such.

- the **True Negative Rate** (TNR) (or **Specificity**):

$$TNR = \frac{TN}{FP + TN} \tag{7}$$

  It describes the test's capability to eliminate a class correctly, and measures the proportion of negative subjects correctly identified as such.

- the **Accuracy** (ACC):

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

  It represents the proportion of true results, both TP and TN, in the whole population.

- the **Positive Prediction Value** (PPV) (or **Precision**):

$$PPV = \frac{TP}{TP + FP} \tag{9}$$

  It represents the proportion of true positives among all the positive results (both TP and FP).

These four metrics will be used to compare the outcomes of three classification analyses.

The first analysis is based on the whole dataset in order to understand the accuracies of the algorithms when the entire dataset is utilized.

The second classification will be based on the selected features that result from the feature selection algorithm.

Finally, eight clinically relevant and potentially actionable features are used for the last classification analysis, as they have been shown in previous studies to be related to the probability of lapse and/or relapse in a smoking cessation attempt [9], [10]. These features and their evaluation scales are presented in Table 1.

Table 1: Previously-identified features

Previously-identified features	**Evaluation scale**
Day of week	1 to 7, with 1 = Monday
Availability of cigarettes	Yes or No
Alcohol consumption	Yes or No
Confidence in ability to resist smoking	1 to 4, 1 = No, 4 = Yes
Location of the subject	Home, Work-place, Other's home, Bar/Restaurant, Vehicle, Outside, Other
Subject's mood	Very bad, Bad, Neutral, Good, Very good
Presence of people smoking near the subject	Yes or No
Is the weekday a weekend day?	Yes or No

2.5. Experimental settings

For the design of a practical classifier, we very often have to deal with only one dataset to perform the study. Thus, we use cross-validation to partition the initial dataset into two sub-dataset [41]. The first part, the training sample, is used to train the algorithm and build the classifier. On the other hand, the validation sample, which represents the remaining part of the data, is used to test and validate the classifier. The two main techniques used to perform cross-validation are the *leave-one-out* cross validation [42] and the *k-fold* cross-validation [43].

In this study, we chose to use 10-fold cross-validation [43] to test our algorithms, since a leave-one-out cross-validation algorithm, which is in general more efficient, would take more than a year to compute according to the number of observations.

3. Results

Table 2 presents the different selected features that resulted from the feature selection algorithm. Respectively four, eleven, and four features were selected for the NB, DA, and CT algorithm. One feature, the day of week, appears in both the previously-identified and CT selected features. Two features, the day of the study and the tense level (Does the subject feel tense?), have been selected with the three classification methods. Two others, the energy or arousal level (Does the subject feel energetic?) and the restlessness level (Does the subject feel restless?), have been selected with two classification methods out of the three studied in this paper.

When the entire dataset is used, the classification tree model (Figure 3c) has an average accuracy of 69.3% for the four metrics, but the sensitivity is less than 80%. For the naive Bayes (Figure 3a) and discriminant analysis classification methods (Figure 3b), the average accuracy is respectively 67% and 68.3%.

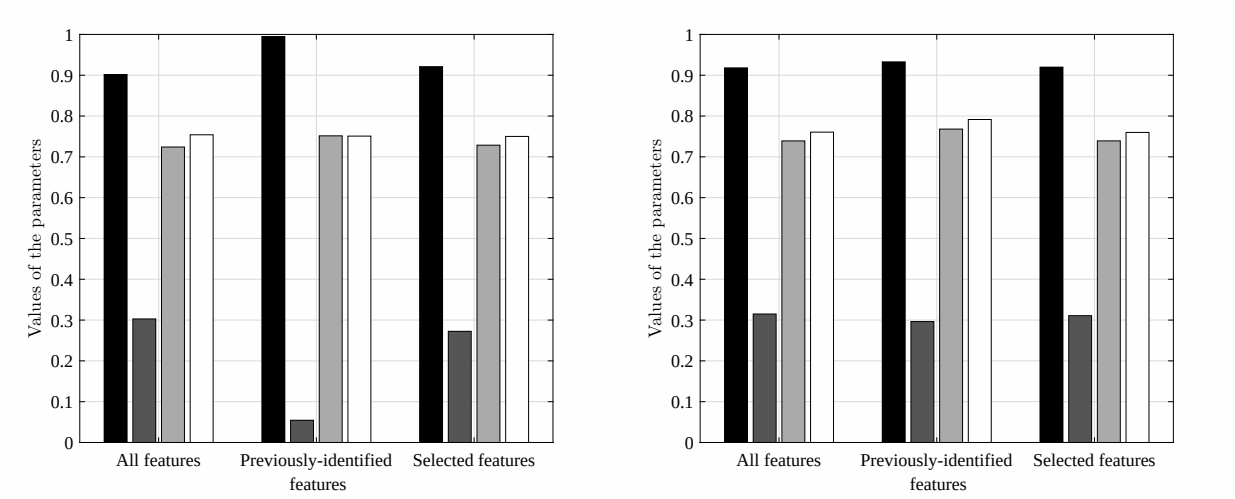
Table 2: Selected features

Selected Features		
NB	**DA**	**CT**
Day of the study	Day of the study	Day of the study
Feeling tense?	Feeling tense?	Feeling tense?
Feeling energetic?	Feeling energetic?	Day of week
Interacting with others?	Feeling restless?	Feeling restless?
	Is the subject alone?	
	What is the subject's arousal level?	
	Drinking coffee?	
	Feeling contented?	
	Inactive?	
	Feeling miserable?	
	Feeling sad?	

Using respectively the previously-identified features, and then a selected subset (feature selection), the metrics' mean is 63.8% and 66.8% for the NB classifier, 69.7% and 68.2% for the DA classifier, and 69.3% and 68.9% for the CT classifier. Besides, as can be seen on Figure 3, none of the feature selection algorithms succeeded in providing both higher sensitivity and higher specificity, compared to the case where the whole dataset is used. Moreover, it is important to note that the results obtained when a feature selection algorithm is used are dependent on the selected features themselves, but are independent on the number of selected features. The purpose of the feature selection algorithm is to preserve the relevant features [25] and to get rid of irrelevant and redundant features [33]. That is why a variable useless or irrelevant by itself can still improve the algorithm performance when grouped with other variables [44].

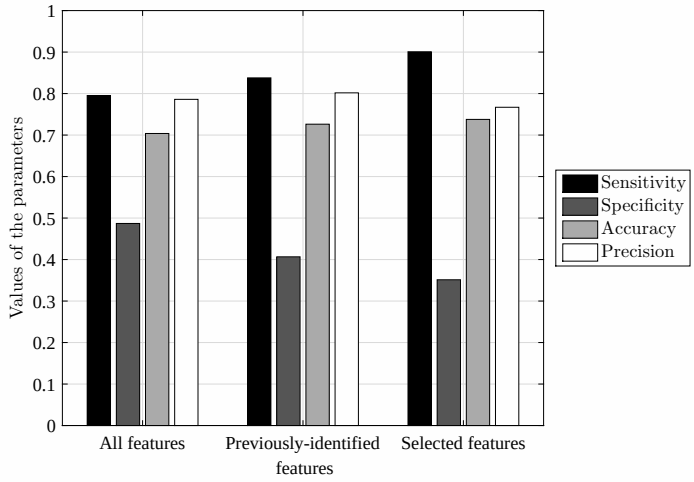
The two first techniques (Figures 3a and 3b) present almost no significant differences when a feature selection algorithm is used. With both methods, sensitivity is slightly higher but specificity is lower than in cases without feature selection. For the classification tree model (Figures 3c), sensitivity is 10.6% higher but specificity is 13.6% lower than in the classification tree without feature selection case, and they are respectively 6.3% higher but 5.5% lower than in the classification tree with the previously-identified features case. The previously-selected and the selected features seem to increase the sensitivity, at the cost of a decreased specificity. In other words, the test's capability to identify a class correctly is high, but this involves a lower capability to eliminate a class correctly. For example, the sensitivity in the NB with feature selection case is 2% higher than in the without feature selection case, but the specificity is 3% lower.

To evaluate the importance of the feature selection algorithm and to emphasize its efficiency, we display



(a) Naive Bayes classiffication method.

(b) Discriminant Analysis method.

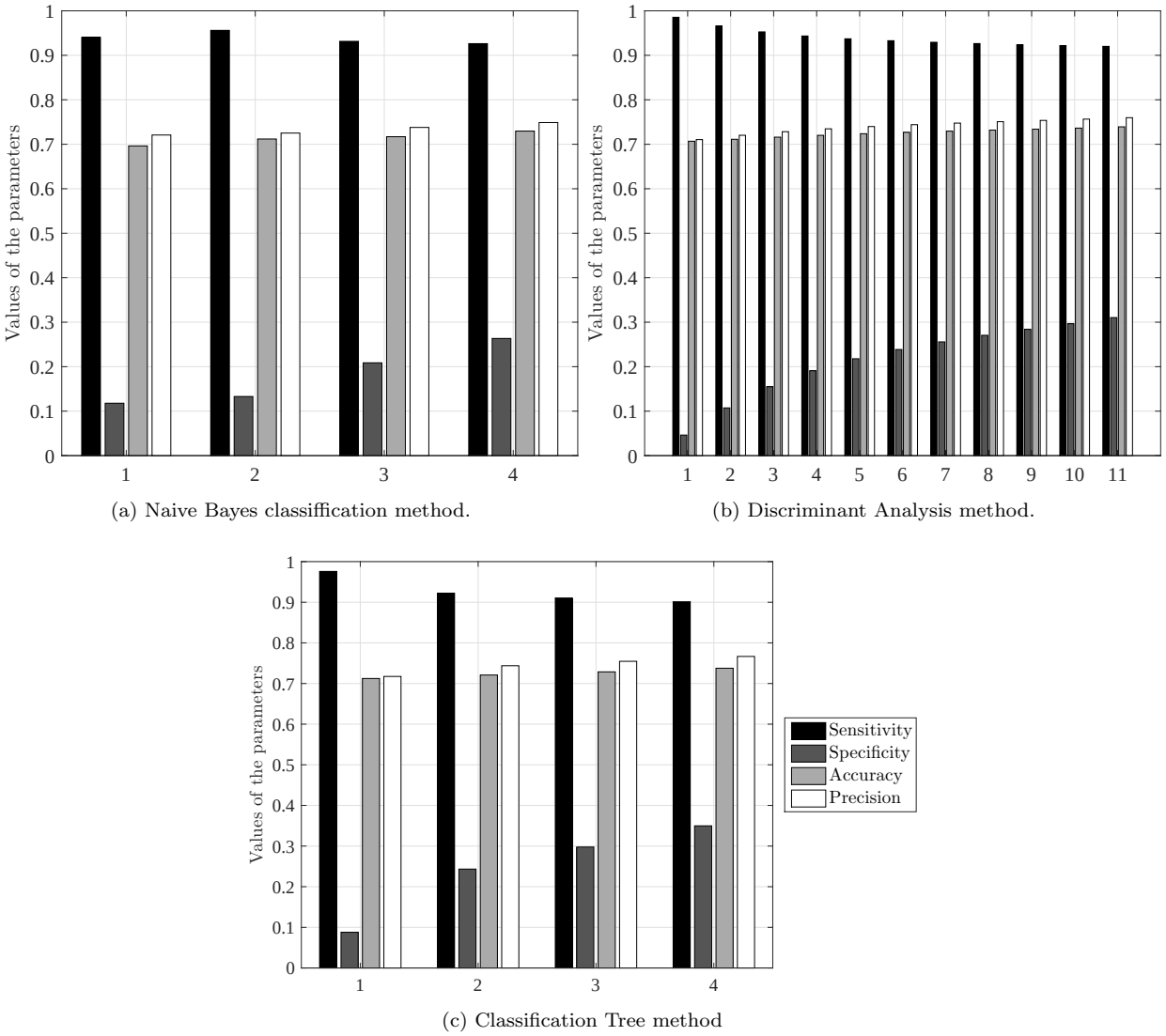


(c) Classification Tree method

Figure 3: Comparison of three classification methods with different datasets.

the results when a different number of features is used for the classification (Figure 4). This number appears on the X axis. The features are selected among the selected features presented in Table 2 and all the possible combinations of features are covered. There are 15 combinations (number of possible combinations with four features) for NB and CT, and 2047 (number of possible combinations with 11 features)for the DA method. In other words, each kth group of four bar plots presented in the following figures provides the results when k features are used for the classification. The displayed results represent the mean value for each possible combination of features.

Figure 4 shows that the sensitivity seems to be slightly decreasing while the number of features is increasing



(a) Naive Bayes classiffication method.

(b) Discriminant Analysis method.

(c) Classification Tree method

Figure 4: Comparison of the number of features (selected among features in Table 2) versus the performance of each classifier.

from one to the total number of selected features. Negatively, the three other metrics increase with the number of features. The evolution of accuracy and precision is hardly notable, but the increase of the specificity is very significant: 15% for NB and 26% for DA and CT. These results highlight the behavior of the feature selection algorithm. It is actually designed to determine the exact number of features that provide the best classification results. That is why the classification results decrease if some features are removed from the selected dataset. Besides, among the different selected features (4 for the NB algorithm, 11 for DA, and 4 for CT), it is not possible to identify one particular feature that is better than the remaining one, as these are fairly nonlinear feature selection process.

In addition to the improvement of the global prediction's accuracy that can be observed with the subset of

features, a faster computation time has also been noticed. Algorithms with feature selection are approximately ten times faster than algorithms based on the entire dataset, and this aspect can be relevant in the case where some analysis has to be run on a less-powerful platform.

4. Discussion

In this study, we considered three machine learning methods to classify situations with strong smoking urges. First of all, without using a feature selection algorithm, the presented results showed that the urge rating can be accurately classified into two states: a high smoking urge versus little or no urge. The naive Bayes method and discriminant analysis have the highest sensitivity ($\approx 90\%$), but the lowest specificity ($\approx 30\%$). The discriminant analysis is the method that provides, on average, the most accurate classifications when the entire dataset is used.

However, evaluating dozens of parameters several times a day can be very time consuming, and lead to unnecessary rejection of potential therapies to be delivered via modern electronic devices such as smartphones. Therefore, it was beneficial to find a reduced set of features that can be used to classify the urge rating. The selected features appear to be dominated by measures of emotional state, especially in the discriminant analysis. It is known that negative emotional states are related to craving and lapsing [45, 5], but it is interesting to see that algorithms have extracted multiple (up to six for he discriminant analysis) indicators of emotional state. Furthermore, the meaning of this selection is that each feature makes an incremental contribution to the final classification since a feature is selected by the algorithm when its selection involves a decrease of the misclassification rate. Coffee consumption has also been associated with smoking and was selected in the discriminant analysis approach. It is interesting to notice that the feature related to drinking alcohol is not taken into consideration, as studies have shown that the effect of alcohol drinking on urges to smoke is stronger than coffee [5]. Finally, the feature that specifies the day of the study has probably been selected because craving decreases as abstinence progresses [46]. Regarding the one that specifies the week day ("Day of week"), it could differentiate a smoking activity happening during week days from a smoking activity happening during weekends.

Feature selection algorithms have respectively selected four and eleven features for the classification tree method and the discriminant analysis. These two classifiers resulted in almost the same mean ($\approx 68\%$) for the four metrics evaluated in this study (sensitivity, specificity, accuracy and precision), and the first provides the highest specificity (35.1% against 31.1%) while the other has the highest sensitivity (92% against 90.1%). These results also highlight that the algorithms are better at correctly classifying the true presence of an urge rather than the true absence of one. This is valuable for the classifications because it is preferable to inaccurately assume an urge is present (and risk unnecessary urge query) than miss a real urge (and risk not intervening to prevent smoking). In addition, sensitivity and specificity are inversely proportional [47, 48].

That can also explain why the specificity is low: as the sensitivity increases, the specificity decreases and vice versa. Finally, according to these results, one of these two classifiers could be finally chosen to make the classifications.

In a future project, the selected features could be implemented in a personal mobile application in order to assist a subject in his/her smoking cessation process. The algorithm would be able to estimate his/her urge rating, which could allow the application to provide more adaptive and personalized support. According to the chosen method, as few as four have to be reported by a subject at each data collection to facilitate a highly sensitive classification of high-urge state. Combined with the popularity of mobile devices and text messaging, users could have the ability to send and receive short and instant messages to deliver a smoking-cessation intervention. This could lead to further studies designed to target specific socio-professional group, like college students for example since programs meant to help smokers quit are expected at both the high school and the college levels to decrease the percentage of young adults who are addicted to nicotine [49]. Future studies could also use the classification of smoking urges via machine learning to explore factors affecting smoking initiation, heavy smoking, and quit behaviors among workers and determine the association between smoking among young adults and social factors in the work environment.

However, there is an important clinical difference between knowing a high urge is likely to occur now (classification) and knowing a high urge is likely to occur next (prediction). While these analyses showed that urges can be correctly classified, the advanced machine learning algorithms adopted in this study could be reused, with new suitable time variables, in order to predict both smoking urges and time occurrences of these events.

4.1. Limitations of the study

In this research work, machine learning algorithms have been used to classify smoking urges. The variety of the methods allowed us to find the most suitable classifier. On the other hand, the specificity of the different algorithms can also be seen as a relative weakness of our approach. As our results have shown, these algorithms can be relatively less predictive, if suitable features are not easily identifiable. However, the question of suitable smoking related features in machine learning remains open, as contextually relevant features do not necessarily produce the most accurate results.

The accuracy of machine learning algorithms is closely related to the dataset used to train the classifiers. Questionnaires used to collect data were good for measuring attitudes from research participants, and they were administrated to a large population, considering that all these data points were collected in the early 1990s. The response rate was high since the patients' goal was to quit smoking.

Finally, this smoking cessation approach can be improved in the future. The first limitation was the number of features, but we highlighted that it can significantly be reduced. Also, this study is limited to classifications, but future studies could use similar methods to address the prediction problem. The tobacco

dependence is an incessant situation that often requires repeated intervention for success [50]. Implementing these algorithms in electronic devices would allow to collect more data, in a wide variety of real-time situations. This would certainly improve the quality of the classifiers and increase the classification accuracy, leading to more efficient counseling and therapies for smoking cessation.

5. Conclusion

In this paper, we proposed a machine learning approach to detect smoking urges. Specifically, we considered approaches based on the naive Bayes, the linear discriminant analysis and classification trees algorithms. Firstly, it has been shown that the three classifiers can each classify urge ratings with reasonable sensitivity and specificity. Then, combined with a feature selection algorithm, the classifier based on a discriminant analysis method extracted eleven features, while the ones based on naive Bayes classifier or classification tree selected four features. These feature selection algorithms enabled us to obtain a sensitivity and specificity of respectively 90% and 35%, showing that smoking urges can be accurately predicted with a reduced dataset. This also suggests that algorithms developed through machine learning approaches may be useful in guiding predictive mobile interventions, as well as providing a more accurate support to existing smoking cessation processes. Besides, taking advantage of the final number of features that have to be reported, patients would be able to update their situation more often than they initially did. More observations would conjointly contribute to increase the amount of available data and thus improving the accuracy of the machine learning algorithms. Eventually, the results would help doctors and healthcare clinicians to discuss the role that smoking plays so that they can provide more accurate behavioral counseling using intelligent systems technology.

The current research limitation is the available dataset. With additional and more recent data, new investigations could be conducted and the new features could increase the classification's accuracy.

Since there is a significant clinical difference between classifications and predictions, these machine learning approaches could be reused, with appropriate time variables, in order to predict both smoking urges and time occurrences of these events. These kind of algorithms could be implanted into mobile systems for smoking cessation purposes. They would use the contextual data reported by users as well as data passively sensed by the mobile device to acquire information, and continually predict high smoking urges. The main positive aspect we could get from it would be the ability to provide real-time intervention before a patient relapse, and without requiring any medical intervention. Mobile devices are nowadays powerful enough to work with powerful algorithms and communicate with remote servers. Thus mobile health applications can adopt more sophisticated approaches and continuously update their predictive algorithms. Finally, since behavioral counseling coupled with pharmacotherapy is considered to be one of the most efficient smoking cessation intervention [51], this research could help with the optimization of dosage, regularity, and duration of both

behavioral counseling and pharmacotherapies.

Acknowledgments

Research reported in this publication was supported by the National Center For Advancing Translational Sciences of the National Institutes of Health under Award Number KL2TR000146. This project was partly also supported by the National Institutes of Health Grant Number 5UL1TR000005-09. Collection of the data used in these analysis was funded by National Institutes of Health (National Institute on Drug Abuse) grant DA06084 to Saul Shiffman. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Appendix A. List of the 41 parameters initially collected

Name	Question	Possible responses
ACONFIDE	Confident in ability to abstain?	NO !! no ?? yes ?? YES !!
ALCOHOL	Drinking alcohol?	No or Yes
ALLOWED	Smoking regulations/allowed?	Forbidden, Discouraged, Allowed
ALONE	Were you alone?	No or Yes
AVAIL	Availability of cigarettes?	No or Yes
BAFFECT	Overall feeling?	Very bad, Bad, Neutral, Good, Very good
BAROUSE	Arousal/energy level?	Very low, Low, Moderate, High, Very high
BURGE	Rate urge to smoke?	0 to 10
COFFEE	Drinking coffee or tea?	No or Yes
CONTENTE	Feeling contented?	NO !! no ?? yes ?? YES !!
DAY	Day of the study?	Number
DOW	Day of week?	1 to 7
EATING	Food or drink 15 min pre-episode?	No or Yes
ENERGETI	Feeling energetic?	NO !! no ?? yes ?? YES !!
FRUSTRAN	Feeling frustrated?	NO !! no ?? yes ?? YES !!
HAPPY	Feeling happy?	NO !! no ?? yes ?? YES !!
HARDCONC	Hard to concentrate?	NO !! no ?? yes ?? YES !!
HUNGRY	Feeling hungry?	NO !! no ?? yes ?? YES !!
INACTIVE	Inactive?	No or Yes
INTERACT	Interacting with others?	No or Yes
IRRITABL	Feeling irritable?	NO !! no ?? yes ?? YES !!
LEISURE	Activities Leisure?	No or Yes
LOCATION	Where were you?	Home, Work-place, Other's home, Bar/Restaurant, Vehicle, Outside, Other
MEALSNAC	Food Meal or Snack?	No, Yes-Meal, Yes-Snack
MISERABL	Feeling miserable?	NO !! no ?? yes ?? YES !!
NEGAFF	Subject's mood?	Very bad, Bad, Neutral, Good, Very good
OTHERACT	Activities Other-not listed?	No or Yes
OTHERSMO	Were people smoking?	No or Yes
RESTLESS	Feeling restless?	NO !! no ?? yes ?? YES !!
SAD	Feeling sad?	NO !! no ?? yes ?? YES !!

SLEEPY	Feeling sleepy?	NO !! no ?? yes ?? YES !!
SPACEY	Feeiling spacey?	NO !! no ?? yes ?? YES !!
TELEPHON	Activities Telephone?	No or Yes
TENSE	Feeling tense?	NO !! no ?? yes ?? YES !!
TIRED	Feeling tired?	NO !! no ?? yes ?? YES !!
TYPEINAC	Type of inactivity?	Waiting, Between activities, Doing nothing
TYPEINTE	Type of interaction with others?	Socializing, For Business, Household Issues, Arguing, Other Interaction
TYPEWORK	Type of work?	Job, House/Personal, Other
WANTSHOU	Your activities were?	Wants, Shoulds, Both
WHERESMOK	Where were others smoking?	In my group, In view
WORK	Activities Working/Chores?	No or Yes

References

[1] World Health Organization: WHO report on the global tobacco epidemic, 2011: warning about the dangers of tobacco: executive summary (2011)

[2] U.S. Department of Health and Human Services: The health consequences of smoking-50 years of progress: A report of the surgeon general. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health **17** (2014)

[3] Agaku, I.T., King, B.A., Dube, S.R.: Current cigarette smoking among adults-United States 2005-2012. Morbidity and Mortality Weekly Report **63**(2), 29–34 (2014)

[4] Centers for Disease Control and Prevention: Quitting smoking among adults-United States 2001-2010. Morbidity and Mortality Weekly Report **60**(44), 1513 (2011)

[5] Shiffman, S.: Relapse following smoking cessation: a situational analysis. Journal of consulting and clinical psychology **50**(1), 71 (1982)

[6] Stone, A.A., Shiffman, S.: Ecological momentary assessment (EMA) in behavorial medicine. Annals of Behavioral Medicine (1994)

[7] Shiffman, S., Stone, A.A., Hufford, M.R.: Ecological momentary assessment. Annu. Rev. Clin. Psychol. **4**, 1–32 (2008)

[8] Shiffman, S.: Ecological momentary assessment (EMA) in studies of substance use. Psychological assessment **21**(4), 486 (2009)

[9] Shiffman, S., Rathbun, S.L.: Point process analyses of variations in smoking rate by setting, mood, gender, and dependence. Psychology of Addictive Behaviors **25**(3), 501 (2011)

[10] Dunbar, M.S., Scharf, D., Kirchner, T., Shiffman, S.: Do smokers crave cigarettes in some smoking situations more than others? situational correlates of craving when smoking. Nicotine & tobacco research **12**(3), 226–234 (2010)

[11] Shiffman, S., Gnys, M., Richards, T.J., Paty, J.A., Hickcox, M., Kassel, J.D.: Temptations to smoke after quitting: a comparison of lapsers and maintainers. Health Psychology **15**(6), 455 (1996)

[12] Lewis, D.D.: Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. Springer (1998)

[13] Mitchell, T.M.: Machine Learning, First edition edn. McGraw Hill (1997)

[14] Manning, C.D., Raghavan, P., Schtze, H.: Introduction to Information Retrieval vol. 1. Cambridge University Press (2008)

[15] Venables, W.N., Ripley, B.D.: Modern Applied Statistics with S. Springer (2002)

[16] Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annals of eugenics **7**(2), 179–188 (1936)

[17] Rokach, L.: Data Mining with Decision Trees: Theory and Applications. Series in machine perception and artificial intelligence. World Scientific (2007)

[18] Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: Classification and Regression Trees. CRC press (1984)

[19] Loh, W.-Y.: Classification and regression trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **1**(1), 14–23 (2011)

[20] Maclin, R., Opitz, D.: Popular ensemble methods: An empirical study. Journal Of Artificial Intelligence Research **11**, 169–198 (2011)

[21] Breiman, L.: Bagging predictors. Machine learning **24**(2), 123–140 (1996)

[22] Breiman, L.: Heuristics of instability and stabilization in model selection. The annals of statistics **24**(6), 2350–2383 (1994)

[23] Zhang, Y., Wang, S., Dong, Z.: Classification of alzheimer disease based on structural magnetic resonance imaging by kernel support vector machine decision tree. Progress In Electromagnetics Research **144**, 171–184 (2014)

[24] Kumar, M.A., Gopal, M.: A hybrid svm based decision tree. Pattern Recognition **43**(12), 3977–3987 (2010)

[25] Tang, J., Alelyani, S., Liu, H.: Feature selection for classification: A review (1997)

[26] Zhang, Y., Wu, L.: An MR brain images classifier via principal component analysis and kernel support vector machine. Progress In Electromagnetics Research **130**, 369–388 (2012)

[27] Jolliffe, I.: Principal Component Analysis. Wiley Online Library (2002)

[28] Fodor, I.K.: A survey of dimension reduction techniques (2002)

[29] Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley (2000)

[30] Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic press (2013)

[31] Ye, J., Janardan, R., Li, Q.: Two-dimensional linear discriminant analysis. In: Advances in Neural Information Processing Systems, pp. 1569–1576 (2004)

[32] Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence **97**(12), 273–324 (1997)

[33] Dash, M., Liu, H.: Feature selection for classification. Intelligent data analysis **1**(3), 131–156 (1997)

[34] Kira, K., Rendell, L.A.: The feature selection problem: Traditional methods and a new algorithm. In: AAAI, vol. 2, pp. 129–134 (1992)

[35] Koller, D., Sahami, M.: Toward optimal feature selection (1996)

[36] Zhang, Y., Wang, S., Phillips, P., Ji, G.: Binary pso with mutation operator for feature selection using decision tree applied to spam detection. Knowledge-Based Systems **64**, 22–31 (2014)

[37] Kojadinovic, I., Wottka, T.: Comparison between a filter and a wrapper approach to variable subset selection in regression problems. Citeseer

[38] Yu, L., Liu, H.: Feature selection for high-dimensional data: A fast correlation-based filter solution. In: ICML, vol. 3, pp. 856–863 (2003)

[39] Karegowda, A.G., Jayaram, M., Manjunath, A.: Feature subset selection problem using wrapper approach in supervised learning. International journal of Computer applications **1**(7), 13–17 (2010)

[40] Stehman, S.V.: Selecting and interpreting measures of thematic classification accuracy. Remote Sensing of Environment **62**(1), 77–89 (1997)

[41] Arlot, S., Celisse, A.: A survey of cross-validation procedures for model selection. Statistics surveys **4**, 40–79 (2010)

[42] Shao, J.: Linear model selection by cross-validation. Journal of the American statistical Association **88**(422), 486–494 (1993)

[43] Geisser, S.: The predictive sample reuse method with applications. Journal of the American Statistical Association **70**(350), 320–328 (1975)

[44] Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. The Journal of Machine Learning Research **3**, 1157–1182 (2003)

[45] Marlatt, G.A., Gordon, J.R.: Relapse Prevention: Maintenance Strategies in the Treatment of Addictive Behaviors. Guilford Press (1985)

[46] Shiffman, S., Engberg, J.B., Paty, J.A., Perz, W.G., Gnys, M., Kassel, J.D., Hickcox, M.: A day at a time: predicting smoking lapse from daily urge. Journal of abnormal psychology **106**(1), 104 (1997)

[47] Stojanović, M.M.: Understanding sensitivity, specificity and predictive values. Vojnosanitetski pregled **71**(11) (2014)

[48] Parikh, R., Mathai, A., Parikh, S., Sekhar, G.C., Thomas, R., *et al.*: Understanding and using sensitivity, specificity and predictive values. Indian journal of ophthalmology **56**(1), 45 (2008)

[49] Everett, S.A., Husten, C.G., Kann, L., Warren, C.W., Sharp, D., Crossett, L.: Smoking initiation and smoking patterns among US college students. Journal of American College Health **48**(2), 55–60 (1999)

[50] Laniado-Laborín, R.: Smoking cessation intervention: an evidence-based approach. Postgraduate medicine **122**(2), 74–82 (2010)

[51] Stead, L.F., Perera, R., Lancaster, T., et al.: Telephone counselling for smoking cessation. Cochrane Database Syst Rev **3**(3) (2006)