A Predictive Machine Learning Model to Determine Alcohol Use Disorder

Ebrahimi, Ali; Wiil, Uffe Kock; Andersen, Kjeld; Mansourvar, Marjan; Nielsen, Anette Sogaard

Published in:
2020 IEEE Symposium on Computers and Communications (ISCC)

DOI:
10.1109/ISCC50000.2020.9219685

Publication date:
2020

Document version
Accepted manuscript

Citation for published version (APA):
https://doi.org/10.1109/ISCC50000.2020.9219685

Terms of use
This work is brought to you by the University of Southern Denmark through the SDU Research Portal.
Unless otherwise specified it has been shared according to the terms for self-archiving.
If no other license is stated, these terms apply:

• You may download this work for personal use only.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying this open access version

If you believe that this document breaches copyright please contact us providing details and we will investigate your claim.
Please direct all enquiries to puresupport@bib.sdu.dk

Download date: 30. Apr. 2021
A Predictive Machine Learning Model to Determine Alcohol Use Disorder

Ali Ebrahimi  
SDU Health Informatics and Technology, The Maersk Mc-Kinney Moller Institute,  
University of Southern Denmark  
Odense, Denmark  
aleb@mmmi.sdu.dk

Uffe Kock Wiil  
SDU Health Informatics and Technology, The Maersk Mc-Kinney Moller Institute,  
University of Southern Denmark  
Odense, Denmark  
ukwiil@mmmi.sdu.dk

Kjeld Andersen  
Research Unit of Clinical Alcohol Research, Department of Clinical Research, University of Southern Denmark  
Odense, Denmark  
kjeld.andersen@rsyd.dk

Marjan Mansourvar  
SDU Health Informatics and Technology, The Maersk Mc-Kinney Moller Institute,  
University of Southern Denmark  
Odense, Denmark  
marm@mmmi.sdu.dk

Anette Søgaard Nielsen  
Research Unit of Clinical Alcohol Research, Department of Clinical Research,  
University of Southern Denmark  
Odense, Denmark  
ansnielsen@health.sdu.dk

Abstract—Prediction of alcohol use disorder (AUD) may reduce the number of deaths caused by alcohol-related diseases. However, prediction of AUD based on patients’ historical clinical data is still an open research objective. This study proposes a method to predict AUD from electronic health record (EHR) data through supervised machine learning. The study creates a dataset based on the combination of EHR data with patient reported data from 2,571 patients in the Region of Southern Denmark. After that, the dataset is labeled into two categories, AUD positive (457) and AUD negative (2,114). This unique dataset is used to validate the proposed method for prediction of AUD using machine learning methods based on historical clinical data from EHRs.

Keywords—Predictive Model, Classification, Supervised Machine Learning, Alcohol Use Disorder.

I. INTRODUCTION

Alcohol use disorder (AUD) is associated with various diseases and it carries a high prevalence rate among Western populations particularly in Europe [1]. Excessive alcohol use leads to numerous diseases such as liver cirrhosis, chronic pancreatitis, upper gastrointestinal cancers, cardiomyopathy, polyneuropathy, and dementia. Alcohol is also a contributing cause for many road accidents [1]. It has been estimated that among the Danish population, 20 percent are heavy drinkers and 14 percent are harmful alcohol users [2]. From their study, Westman, et al. [3] note that the mortality rate due to high alcohol consumption in Denmark was higher than other European countries with similar populations such as Sweden and Finland. Consequently, it was analyzed that approximately five percent of all registered deaths in Denmark can be attributed to alcohol use, a common scenario in most Western countries [2, 3]. This indicates that AUD may cause shorter life expectancies among the Danes when compared to the general population.

The majority of patients suffering from AUD never receives specialized treatment for their addiction [4]. This might be because of the conventional methods which detect alcohol-related problems through self-test reports [5]. Distributing self-tests is time-consuming and the accuracy of self-test reports has been questioned especially for heavy drinkers and alcohol abusers [6]. In this regard, a prediction of AUD based on complex clinical reports via machine learning (ML) methods can help to overcome this problem. This information can offer insights into the patient's lifestyles which affect his/her medical conditions. Using this information, medical staff can become actively involved with their patients to understand their needs and to be able to discuss how treatments can best be arranged for the individual patient.

Classification of clinical documents is one of the popular research areas in the domain of clinical data mining. It refers to the task of classifying clinical documents or reports into one or more than one predefined categories [7]. Classification techniques have been implemented on several types of clinical reports such as pathology reports, radiology reports, autopsy reports, and biomedical documents. These types of data used to detect cancer stages [8], to predict acute admission at hospitals [9], to identify pediatric traumatic brain injury [10], to predict the cause of death [7], and to predict AUD [11, 12].
To date, few studies have aimed to develop a predictive model for AUD based on ML approaches. Bi, et al. [13] proposed a support vector machine to construct a classifier to identify the drinking behavior of individual college students based on self-reports. Zhu, et al. [12] aimed to identify the features in a multivariate fashion that are predictive of alcohol dependence based on MRI data by employing ML algorithms.

Although the above studies claimed that they could successfully achieve their goals in investigating different cases for developing predictive models for predicting AUD problems, several factors distinguish their work from the current study. While several factors contribute to development of AUD [14], previous studies that had developed a predictive model for AUD [12, 13] did not consider patients’ historical clinical data. To our knowledge, prediction of AUD based on patients’ historical clinical data through ML techniques and algorithms has not yet been explored.

To address this gap, this study aims to develop an automated classification model by using supervised machine learning-based (SML-based) algorithms to classify historical patient data from Electronic Health Records (EHRs). Such a model can improve patient treatment by providing medical staff with a better knowledge of their patients by utilizing the information already stored in the EHRs, and subsequently improve the patients’ prognosis. In addition, the model can also assist staff by providing a better and faster prediction of AUD in admitted patients thereby saving lives, time and costs.

This paper is organized in following manner: Section 2 discusses the proposed method and available datasets, Section 3 presents the results and Section 4 presents limitations, conclusion, and future work.

II. METHOD

The research methodology proposed for this study encompasses five phases: data collection, data pre-processing, feature selection, data modeling, and evaluation metrics (see Fig. 1). In this study, clinical researchers have been involved throughout all the phases of the methodology. For example, clinical researchers were involved in the collection and pre-processing of data based on their domain knowledge. Moreover, medical reasoning about individual features was discussed in detail with them over several iterations in the feature selection phase.

A. Dataset collection

Our study population consist of patients aged 18-101 hospitalized at Odense University Hospital (OUH) in Denmark for at least 24 hours from January 2012 until June 2016. Data captured for this study is created by linking two sources, the Relay study [15-17] and EHRs from OUH.

The Relay study systematically collected patient reported data on diet, smoking, alcohol, and exercise habits from 3,534 patients admitted to the Gastrointestinal, Neurologic and Orthopedic Departments at OUH and to the Emergency Department at Hospital Sønderjylland in Aabenraa in the inclusion period, October 2013 to June 2016. In the Relay study, patients responded to a questionnaire with the Danish version of Alcohol Use Disorder Identification Test (AUDIT) embedded with the score between 0 and 40. According to the classification of Baror et al. [18], patients could be divided into two groups: AUDIT scores equal or less than 8 were categorized as AUD negative and the AUDIT scores greater than 8 were labeled as AUD positive. Of those 3,534 patients, 609 patients were labeled as AUD positive and 2,925 patients were labeled as AUD negative. These labels, in the current study, will be used as target values of the training dataset to build the classification model. Fig. 2 display the distribution of our sample data based on their gender, age group and the status of their AUD.

The EHR dataset is the medical records of patients who had a prior visit to OUH before participating in the Relay study. This dataset is repository historical account of the patients’ admissions and visits to the Gastrointestinal, Neurologic and Orthopedic Departments at OUH. The dataset contains clinical records including Personal Identification Number (called CPR number in Denmark) [19], age of patients, gender, the length of stay at the hospital, hospital department where the patient belonged to, hospital department where the patient received treatment, admission type, attending ICU, and being at the emergency department, ICD treatments, cares, operations, and more importantly diagnostic codes (based on ICD-10) and health-related conditions for the period of January 2012 until June 2016.

To link the above-mentioned datasets, the CPR number was employed. On the first step, for the purpose of data security and to observe the GDPR rules, all CPR numbers were anonymized and then Relay and EHR datasets were linked using the anonymized CPR numbers. Since the EHR dataset is collected from OUH, those patients from Hospital Sønderjylland in Aabenraa were removed from our population. This means that the number of patients was reduced from 3,534 to 2,571. Moreover, patient’s records after the Relay interview date were excluded from the final dataset. This means that the dataset consists of the patient’s EHR data from 18 months prior to the admission to hospital and participation in the Relay study. The list, description and data type of variables from the final dataset are presented in a Table I.
ICU If the patient was transferred to the ICU ED If the patient visited the emergency department Length of stay The amount of time when patient spent at the hospital in each visit. Departments Main Departments Departments where offer main treatments to the patients. Secondary Departments Departments where offer additional treatments to the patient. Visit_type The type of visit to the hospital including outpatient visit and admission. Visit_start The time and date when the patient arrived at the hospital. Visit_end The time and date when the patient left the hospital. Action diagnosis Reason why patients visited the hospital. It is stored based on Danish version of ICD10 codes. Code Secondary diagnoseses, any kind of treatments, surgery, etc. Patients have between one and 100 codes.}

**TABLE I.** DESCRIPTION OF VARIABLES.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Data type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient_Id</td>
<td>Anonymized version of personal identification number</td>
<td>Numerical</td>
</tr>
<tr>
<td>AUD status</td>
<td>Either positive or negative. This is the class label as well.</td>
<td>Binary</td>
</tr>
<tr>
<td></td>
<td>AUD positive = 1</td>
<td>AUD Negative = 0</td>
</tr>
<tr>
<td>Gender</td>
<td>Male or female</td>
<td>Categorical</td>
</tr>
<tr>
<td>Age</td>
<td>Age of patients at time of Relay study.</td>
<td>Numerical</td>
</tr>
<tr>
<td>Interview_date</td>
<td>Date and time of interview during Relay study.</td>
<td>Categorical</td>
</tr>
<tr>
<td>Action diagnosis</td>
<td>Reason why patients visited the hospital.</td>
<td>Categorical</td>
</tr>
<tr>
<td>Code</td>
<td>Secondary diagnosises, any kind of treatments, surgery, etc. Patients have between one and 100 codes.</td>
<td>Categorical</td>
</tr>
<tr>
<td>Visit_start</td>
<td>The time and date when the patient arrived at the hospital.</td>
<td>Categorical</td>
</tr>
<tr>
<td>Visit_end</td>
<td>The time and date when the patient left the hospital.</td>
<td>Categorical</td>
</tr>
<tr>
<td>Main Departments</td>
<td>Departments where offer main treatments to the patients.</td>
<td>Categorical</td>
</tr>
<tr>
<td>Secondary Departments</td>
<td>Departments where offer additional treatments to the patient.</td>
<td>Categorical</td>
</tr>
<tr>
<td>Length of stay</td>
<td>The amount of time when patient spent at the hospital in each visit.</td>
<td>Numerical</td>
</tr>
<tr>
<td>ED</td>
<td>If the patient visited the emergency department prior to the admission</td>
<td>Binary</td>
</tr>
<tr>
<td></td>
<td>Yes = 1</td>
<td>No = 0</td>
</tr>
<tr>
<td>ICU</td>
<td>If the patient was transferred to the ICU</td>
<td>Binary</td>
</tr>
<tr>
<td></td>
<td>Yes = 1</td>
<td>No = 0</td>
</tr>
</tbody>
</table>

**Fig. 2.** Overview of study population.

**B. Data Preprocessing**

Following data collection, the process for data mining technique requires pre-processing of the data to properly analyze the data by ML algorithms. Depending on the type of the datasets, the pre-processing tasks might be different but, in most cases, it involves removing meaningless data from the collected dataset to improve the quality of the predictive models. In our data, each row represents a clinical report of a patient for a single visit or admissions to the Gastrointestinal, Neurologic or Orthopedic Departments at OUH. Each column entry represents a patient’s value to each variable.

First of all, we labeled all records in the dataset according to the AUDIT test score which was either AUD positive (AUDIT score greater than 8) or AUD negative (AUDIT score equal or less than 8). As presented in Table I, our dataset consists of several data types including categorical, numerical and binary value. Since ML algorithms only accepts numerical values as input, the categorical data must be encoded into numerical values in each category characterized with a number [20]. One of the widely used solution to overcome this problem is to convert the categorical variable to a set of binary features via one-hot-encoding technique. This method transforms a single variable with n observations and d distinct values, to d binary features with n observations each. Each observation indicating presence (1) or absence (0) of dichotomous binary variable. Therefore, one-hot-encoding is used to deal with the categorical data. Moreover, we had to deal with noisy data including missing value and weight differences between some values. Regarding the missing values (e.g., interview date), we decided to remove any patients from the dataset that had any missing value. The standard normalization technique was used to achieve a common scale for all variables. As it can be seen in the Fig.2, the EHR dataset is imbalanced. The EHR dataset consist of 2114 patients in majority class and 547 patients in minority class. To overcome imbalance in the dataset, the Synthetic Minority Oversampling (SMOTE) [21] technique is applied.

**C. Feature Selection**

One of the most important steps in any data mining task is to select the most efficient features to build the best predictive ML model [22]. One of the characteristics of the real-life datasets is the high number of variables. Considering a dataset with n variables has 2^n possible features, therefore, our dataset with about 860 variables per patients has 2^{860} possible features. Feature selection basically focus on removal of redundant and irrelevant features of the original feature sets in order to select a small subset of features to build the predictive model.

Filter model, wrapper model and embedded models are the three main category of feature selection [23]. The main idea of filter model is to first rank features based on certain criteria and then use the features with the highest ranking to build the model [24]. The main advantage of this approach is to select features independent of any specific classifier. This approach is deployed as a measure to select novel features that can accurately and efficiently predict AUD based on our dataset. A vast number of feature selection algorithms are proposed in previous studies including, Chi Square statistics, Information gain, Ambiguity measure feature selection etc. [25]. Since in
this study most of the features are categorical features, we used Chi Square statistics ($P<0.05$) to select one set of potential features for further analysis (ModelI). Moreover, another set of potential features are selected based on the frequency of variable (specifically ICD codes) among patients (ModelII).

**D. Machine learning model development**

The selected features will then be used to build the AUD classification model using SML-based algorithms. Selecting an appropriate machine learning decision model is crucial. SML based algorithms employ a mapping function to predict dependent variable from independent variable and they are divided into two subtype including classification and regression. Here, the SML-based algorithms learn the classification rules from the features that were extracted from the labelled datasets or training set. Using SML-based algorithms to build an effective and efficient classification model may depend on various related factors and the most important factor among these is the features extracted from the clinical reports [7]. There are two types of SML-based algorithms namely, generative SML-based algorithms which learn the joint probability distribution and discriminative SML-based algorithms which learn the conditional probability distribution. In this study, the state of the art and the most widely used discriminative SML-based algorithms in clinical decision support [26], the Radical Basis Function (RBF) kernel support vector machines (SVM) was trained to examine both selected feature sets among two classes, AUD positive and AUD negative.

**E. Model validation**

The performance of the constructed predictive ML model will be measured using various performance metrics. These performance metrics include accuracy, precision, recall and F-measure. The values of these metrics can be computed by using the values of true positive (TP), false positive (FP), false negative (FN), and true negative (TN) from the confusion matrix. The outcome of the prediction of AUD’s classifiers are evaluated on the test data to report classification accuracy measured by various evaluation metrics including precision, recall, F-measure, and accuracy.

**III. RESULTS**

The study population consisted of 3,534 patients age 18 to 101 patients who had at least one visit to the OUH from January 2012 until June 2016. After excluding 963 patients due to missing value in one or more variable, 2,571 patients were available for analysis. Of those, 457 are labeled as AUD positive and 2,114 are labeled as AUD negative and about 54% of patients are female. Over the entire data collection period, about 14,079 clinical records were registered with at least one record for one patient and the most 135 records. The longest stay at the hospital was 5,574 minutes for an AUD negative patient (63-year-old male). 851 ICD codes were identified as action diagnosis (AD). The baseline characters of patients and final dataset are presented in Table II.

As shown in Table I, we have a variable called action diagnosis (AD). In total there are 850 different ICD codes as AD amongst all records, in which DM171 (frequency among records = 652), DM169 (frequency among records = 556), D1639 (frequency among records = 494), DT840 (frequency among records = 345) and DR298A (frequency among records = 343) are the top five AD in our dataset. Fig. 3 shows the top 20 ADs in our dataset. AD is one of the categorical variables that we have in our dataset. Using one-hot-encoding technique to convert this variable to the binary variable produced 851 features. Considering the top 28 ADs in which the frequency among records is greater than 100 was selected as features for development of ModelI. On the other hand, considering the result of Chi Square test ($P<0.05$), 178 ADs were selected as features for development of ModelII.

Given the classification approach, we have two different classes, AUD positive (1) and AUD negative (0). RBF SVM was employed to build two predictive models in which ModelI is build using 33 features and ModelII is built using 188 features which is presented in Table III. ModelI is based on 10,122 clinical records and ModelII is based on 22,398 clinical records (after handling imbalance issue in the dataset). Both SVM classifiers were implemented based on 80% of the final dataset as training sets and evaluated on the other 20%.

Fig. 4 display confusion matrix of both SVM classifiers in which among 2,025 clinical records as test set of ModelI, 1,041 of them are AUD positive in which the SVM classifier could correctly predict 997 of them as AUD positive. On the other hand, out of 4,480 clinical records as test set of ModelII, 2,223 of them are AUD positive in which the SVM classifier could correctly predict 1,926 of them as AUD positive. Moreover, out of 984 clinical record as testing set of the ModelII for prediction of AUD negative, the SVM classifier could correctly predict 880 of them. In contrast, of 4,480 clinical records as test set of ModelI, 2,257 of them were AUD negative in which the SVM classifier could correctly predict 1,669 of them.
Table IV represents the performance metrics of the SVM classifiers. In ModelI, about 91% of samples are supported for AUD negative class and 92% of samples are supported in the AUD positive class. Besides, in ModelI about 79% of samples are supported the AUD negative class, while only about 81% of the samples are supported the event class (AUD positive). Although both models perform well, ModelII with fewer features could outperform ModelI (37 features in ModelII as compared to 188 features in ModelI).

**TABLE II. CHARACTERISTICS OF STUDY POPULATION**

<table>
<thead>
<tr>
<th>Variable</th>
<th>AUD Positive</th>
<th>AUD Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>%</td>
</tr>
<tr>
<td>Patients</td>
<td>457</td>
<td>17.7</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>122</td>
<td>4.7</td>
</tr>
<tr>
<td>Male</td>
<td>335</td>
<td>13.2</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-20</td>
<td>24</td>
<td>0.9</td>
</tr>
<tr>
<td>21-40</td>
<td>85</td>
<td>3.3</td>
</tr>
<tr>
<td>41-60</td>
<td>174</td>
<td>6.7</td>
</tr>
<tr>
<td>61-80</td>
<td>165</td>
<td>6.4</td>
</tr>
<tr>
<td>81-110</td>
<td>9</td>
<td>0.3</td>
</tr>
<tr>
<td>Clinical Records</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Action Diagnosis</td>
<td>2,366</td>
<td>16.8</td>
</tr>
<tr>
<td>ED</td>
<td>214</td>
<td>1.51</td>
</tr>
<tr>
<td>ICU</td>
<td>26</td>
<td>0.2</td>
</tr>
</tbody>
</table>

**TABLE III. SELECTED FEATURES.**

<table>
<thead>
<tr>
<th>Feature ID ModelI</th>
<th>Feature ID ModelII</th>
<th>Feature Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-178</td>
<td>1-27</td>
<td>Action diagnosis</td>
</tr>
<tr>
<td>179</td>
<td>28</td>
<td>Female</td>
</tr>
<tr>
<td>180</td>
<td>29</td>
<td>Male</td>
</tr>
<tr>
<td>181</td>
<td>30</td>
<td>Age</td>
</tr>
<tr>
<td>182</td>
<td>31</td>
<td>Length of stay</td>
</tr>
<tr>
<td>183</td>
<td>32</td>
<td>ICU</td>
</tr>
<tr>
<td>184</td>
<td>33</td>
<td>ED</td>
</tr>
<tr>
<td>185-188</td>
<td>34-37</td>
<td>Main Departments</td>
</tr>
</tbody>
</table>

**TABLE IV. PERFORMANCE METRICS OF THE SVM CLASSIFIER.**

<table>
<thead>
<tr>
<th>Model</th>
<th>AUD Positive</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>ModelI</td>
<td></td>
<td>0.77</td>
<td>0.87</td>
<td>0.81</td>
<td>2223</td>
</tr>
<tr>
<td></td>
<td>AUD Negative</td>
<td>0.85</td>
<td>0.74</td>
<td>0.79</td>
<td>2257</td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td>0.80</td>
<td></td>
<td></td>
<td>4480</td>
</tr>
<tr>
<td>ModelII</td>
<td>AUD Positive</td>
<td>0.90</td>
<td>0.94</td>
<td>0.92</td>
<td>1041</td>
</tr>
<tr>
<td></td>
<td>AUD Negative</td>
<td>0.93</td>
<td>0.89</td>
<td>0.91</td>
<td>984</td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td>0.92</td>
<td></td>
<td></td>
<td>2025</td>
</tr>
</tbody>
</table>

IV. DISCUSSION AND CONCLUSION

Currently, a vast amount of data has been (and is being) compiled in EHRs. It has long been recognized that clinical reports are beneficial for secondary use. A number of researchers has deployed clinical data mining to mine useful information (such as medical concepts or medical entity) from clinical reports [27]. Various applications of clinical data mining include clinical information extraction [28], clinical relation extraction [29], clinical document clustering [30] and classification of clinical documents [7].

In this study we have presented a methodology for prediction of AUD using ML techniques based EHR. The proposed method consists of five stages including data collection, pre-processing, feature selection, predictive model development, and evaluation. We explained the process of data collection along with pre-processing and two set of master features selected for development of a predictive model based on SVM algorithm. The main contribution of this study is the development of a novel predictive model for prediction of AUD as well as comparing the results of two SVM models with different size of feature set. To our knowledge, this is the first study of its kind based on EHR data. Although ModelII could outperform the performance of ModelI, we believe there are a variety of aspects and features in ModelI, which has high impact on the final accuracy. There are several reasons why ModelI could not perform well including high dimensionality, missing data, sparseness, etc.

In applications of clinical data mining, it is important to prepare quality data. In general, data cleaning and preparation take approximately 80% of the total workflow of a data mining task. It is expected that the input data of machine learning algorithms is nicely distributed, contains no missing or incorrect values, and has well-engineered features to prevent disguising useful hidden patterns in the data, low accuracy and consequently poor prediction quality. Real-world datasets may be noisy, inconsistent, and incomplete, which are the cause of disguised useful patterns.

![Fig. 4. Confusion Matrix of the SVM Classifiers.](image-url)
algorithms is nicely distributed, contains no missing or incorrect values, and has well-engineered features to prevent disguising useful hidden patterns in the data, low accuracy and consequently poor prediction quality. Real-world datasets may be noisy, inconsistent, and incomplete, which are the cause of disguised useful patterns.

The feature selection step is another crucial task in the process of building a prediction model based on SML algorithms. Feature selection is the task to design the discriminative features, which have the highest probability of being one of the finalized features. The preliminary analysis of our data shows that in the 14,079 clinical records, there are 851 recorded diagnosis among all patients which each of them has the possibility to be among the most important features to predict AUD. This number can be high for many classification models and sometimes including all features may produce less accurate models. Moreover, the high dimensional feature vector increases computational complexity and requires more time to develop a classification model.

Dimensionality reduction is one of the most popular tasks in many data mining studies to remove redundant and noisy variables from datasets. Feature extraction and feature selection methods are the most commonly used to overcome high dimensionality. Feature extraction methods produce a newly constructed feature set based on the projection of the original feature. Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Canonical Correlation Analysis (CCA) techniques are examples of feature extraction methods.

Feature selection methods try to select a small portion of features based on their importance to the target value or class labels. Basically, feature selection methods are divided into three categories including filter model, wrapper model, and embedded models. The main advantage of the filter model is that the effects of the selected features on the final performance of the algorithm is not considered. Wrapper mode covers this limitation in which it builds based on the dependency of the feature on the performance of the model. Basically, wrapper models iterate over two steps, first select a subset of feature, and then evaluate the performance of the predictive model which was built based on the subset features. The main and major disadvantage of this process is the high computational cost [31]. Embedded model, on the other hand, has the main advantage of both wrapper and filter model in which feature selection are embedded with classifier construction which is considering the integration of wrapper model as well as filter model [32].

In our future work a comparative study will be conducted to compare the performance of different ML techniques including SVM, KNN, DR, DNN, and BN based on master features which are the results of dimensionally reduction techniques. Also, we will compare the results of dimensionality reduction techniques including PCA, LDA, and CCA and feature extraction techniques including filter model, wrapper model, and embedded models based on the performance of the developed predictive models to identify the AUD risk factors. Consequently, we aim to identify the most effective factors and techniques to identify AUD patients.

REFERENCE


