Analysis of alcohol abuse using improved artificial intelligence methods

D Kumari1 and A Swetapadma1
1School of Computer Engineering, KIIT Deemed to be University, Bhubaneswar, India
E-mail: aleena.swetapadma@gmail.com

Abstract. In this work, artificial intelligence (AI) based method has been proposed for analysis of alcohol abuse (AA) to evaluate drug risk and find an optimal method for analysis of AA. To detect alcohol abuse, twelve input features are selected. As the number of positive classes and negative classes are imbalanced, input features are processed using synthetic minority oversampling technique (SMOTE). The selected input features are given to the AI based method to predict whether an individual uses alcohol or not. Three AI methods (logistic regression, Naïve Bayes and gradient boost) have been used from which the most suitable method is chosen as optimal method for analysis of AA. Methods used for validation of the proposed AI based method are leave one out cross validation and test sample estimate. Among the three methods used, gradient boost method performs better than other method. Gradient Boost based AI method has 97.55% accuracy to predict alcohol user. Significant achievement of the proposed work is that the accuracy (nearly 97%) to detect alcohol abuse is much more than previously (nearly 70%) suggested methods. This is due to the efficient design of AI method implemented in this work. All the feature processing work and AI algorithm design work has been done using python software.

Keywords: Artificial intelligence, Alcohol abuse, Drug abuse detection, Gradient boosting.

1. Introduction
One of the most important issues considering the mental health nowadays is alcohol addiction. Drug addiction affects the youth physically and mentally, leading to decrease in performance considerably. Alcohol is considered to be one of the most consumed psychoactive substances. According to WHO (world health organization), every year alcohol consumption contributes to 3 million deaths across the globe and its harmful use leads to 5.1% of various global diseases. The prediction of alcohol usage by a person can be very helpful and important measures can be taken to prevent the individuals from getting addicted to alcohol. Various methods have been given by researchers for increasing the prediction rate of alcohol users.

Risky health behaviors are predicted by the personality factors of neuroticism, extraversion and conscientiousness [1]. These days, students are addicted towards drugs which hamper their health and career eventually. Machine learning and data mining algorithms are used to predict whether a teenager could be a drug addict as studied in [2]. This study focuses on detecting the individuals who are at high risk and those who are addicts are not focused. Most of the users of drug are in their adolescent stage. In [3], a machine learning approach is used to vectorize text message data based on contextual and semantic structure for use as input. It predicts alcohol consumption in adolescents and explains that how the contextual patterns of an individual’s text messaging communication reveal his or her wider behavioral patterns. In [4], the process of adapting well to substance use disorders has been discussed. In [5], adolescent substance use is predicted and it has been discussed that what is the connection between reward and cognitive control brain networks. In [6], the relationship between big
five personality traits and behavioral addictions has been suggested. In [7], drug consumption risk and importance of five factor model has been discussed. In [8], for the prediction of alcohol users, an artificial neural network-based method has been proposed. In [10], decision tree classifier has been used to predict if an individual is suffering from alcohol use disorder and seeks treatment or not. In the decision tree classifier, the ten measures were drinking behavior, depression and various psychological problems related to drinking and substance dependence. In [16], the different alcohols biomarkers with their pros and cons are discussed. Recently used biomarkers are ethanol derivative products which are very effective. The biomarkers which can be used in future for detection of pattern of use of alcohol are also analyzed. Trans-dermal biosensors are used in [17]. Extra-Trees machine learning algorithms are used with trans-dermal time series features to predict the drinking and non-drinking episodes of the individuals. In [18], treatment-seeking individuals with AUD on alcohol cue-reactivity (CR) are compared to non-treatment-seekers. The result showed that treatment seekers are not more cue-reactive than non-treatment seekers.

In this work, an AI based approach is proposed for alcohol abuse analysis. SMOTE has been used to pre-process the samples. A comparative study of three AI methods namely logistic regression, Naïve Bayes and gradient boost has been carried out to find the most suitable method. The paper is organized as following: Section 2 elaborates the oversampling method SMOTETomek. In [11], this method is explained and shows that SMOTETomek outperforms several existing methods and improves accuracy, section 3 describes various AI based methods like logistic regression, Naïve Bayes and one of the most efficient methods, gradient boost. The idea of this method is explained in [12], proposed method is described in section 4, section 5 describes the result and section 6 contains the conclusion.

2. Feature processing

In this work, feature processing is done using a technique called SMOTE. SMOTE stands for synthetic minority oversampling technique. The concept of SMOTE was first introduced in 2002 [13]. It helps in dealing with the imbalanced datasets. When the severe class imbalance is present in the dataset then it is termed as an imbalanced dataset. Such datasets result in biased predictions and the inaccurate result is obtained in such cases. Re-sampling the datasets is then essential to prevent the classifier in predicting the minority classes inaccurately and thus leading to its poor performance [11]. The different ways to resample the datasets are that we either increase the minority classes or decrease the majority classes. The random under-sampling i.e. randomly removing the majority class observations seems like balances the dataset, but this approach may discard the observations that have important information and can lead to a bias. Random over-sampling of the minority observations by replicating them multiple times can prevent the loss of information which occurs in random under-sampling technique. However, it is prone to over-fitting as the information has been simply copied back to the dataset. The disadvantages of these approaches are removed in SMOTE. It creates new synthetic observations. The steps of this approach are mentioned as following:

i. The feature vector and its nearest neighbor are identified. Their difference is calculated.

ii. A random value between 0 and 1 is chosen and is multiplied with the difference obtained in the previous step.

iii. To the feature vector, a random number is added and a new point is thus identified on the line segment.

iv. For the identified feature vectors, the same process is repeated.

It is well understood since that after applying oversampling on the dataset, it should be followed by an under-sampling technique. A re-sampling technique which uses SMOTE with Tomek is used recently in a study [11]. Batista et al [14] proposed using Tomek links as a method which cleans the data and balances the imbalanced data. In [15], the journey of SMOTE has been discussed since 15 years of its inception and SMOTE with Tomek links is considered to be one of the most efficient
techniques for balancing the data. SMOTE increases samples in the dataset while Tomek is an undersampling technique for cleaning overlapping samples.

3. AI methods used

Various AI based methods are used for analysis of AA in this work such as logistic regression, Naïve Bayes and gradient boost. These methods are described in brief in the sections below for better understanding.

3.1. Logistic Regression Method

The Logistic regression is an improvement over the linear methods of classification. In this method, a transformation is done to probability. In case of binary classification where the classes have either 0 or 1 value, for given input x, the output is 1 and its probability is p(x)=p(g=1| X=x). The logit transformation is done as explained below. Here, \( \beta_0 + \beta x \) is the linear component where \( \beta \) is slope coefficient and \( \beta_0 \) is the intercept.

\[
\log \frac{p(x)}{1 - p(x)} = \beta_0 + \beta x
\]

(1)

The above expression can be written as following:

\[
p(x) = \frac{e^{\beta_0 + \beta x}}{1 + e^{\beta_0 + \beta x}}
\]

(2)

\[
p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta x)}}
\]

(3)

When \( p(x) > 0.5 \), output is 1. When \( p(x) < 0.5 \), output is 0. Linear regression is unbounded. When it is plug into above expression, it is made sure that probability lies between 0 and 1. Separation boundary lies between class 0 and class 1. When, \( p(x) = 0.5 \), then, \( \beta_0 + \beta x = 0 \) (obtained after putting \( p(x) = 0.5 \) in above equation). It is a straight-line equation. Logistic regression looks simple but it gives a very popular classifier. Apart from building the classification surfaces, the sensitivity analysis is also performed very efficient by it. It works very well in practice and is used by the people working in machine learning domain as well as by the statisticians. For the estimation of parameters of logistic regression, one needs to model the probability and maximize the likelihood of the data. In linear regression, error function is found and the optimization of the error function is performed. But here the optimization of the likelihood of the data is performed. \( P(D|\theta) \) is known as likelihood of \( \theta \). It is probability of D given parameters \( \theta \). The training data D is given and it is fixed, we are trying to find \( \theta \). \( P(D|\theta) \) is denoted by \( L(\theta) \) and \( \log L(\theta) = l(\theta) \). In above equation, \( \theta \) is \( \beta \). Input D consists of \( D = \{(x_1, g_1), (x_2, g_2)\ldots(x_N, g_N)\} \) where x is input and g is output. In binary classification, g is either 0 or 1. ie g \( \in \{0,1\} \). Then likelihood is

\[
L(\beta_0, \beta) = p(x)^g(1-p(x))^{(1-g)}
\]

(4)

The probability that x has the label 1, then \( L(\beta_0, \beta) = p(x) \). The probability that x has the label 0, then \( L(\beta_0, \beta) = 1-p(x) \).

3.2. Naïve Bayes Method

The probability of occurrence of class g given X is to be calculated

\[
p(g | X = x) = \frac{p(x | g)p(g)}{p(x)}
\]

(5)
where \( p(x) \) is given as following,

\[
p(x) = \sum_g p(x, g)
\]

(6)

It can be rewritten as,

\[
p(x) = \sum_g p(x \mid g)p(g)
\]

(7)

where \( p(x \mid g) \) is called Class condition distribution. Likelihood is a function of \( g \) where as \( p(x \mid g) \) is function of \( x \), conditioned on \( g \). The simplest assumption to get \( p(x \mid g) \) tractable is the Naive Bayes assumption. This assumption says that given the class label, the features are independent of each other. Following equation depicts the probability,

\[
p(x_1, x_2, \ldots, x_p \mid g) = \prod_{i=1}^{p} p(x_i = g)
\]

(8)

Some kind of parametric form for the conditional marginal distributions like mentioned in above equation is always recommended. The most usual such parametric form is Gaussian Naive Bayes. One more reason for go for Gaussian Naive Bayes is that, in case \( x_i \) is discrete value, then go for discrete distribution but if \( x_i \) is continuous value then we go for Gaussian distribution [12]. Gaussian Naive Bayes follows Gaussian or normal distribution for modeling the continuous valued features. The likelihood is

\[
P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma^2_y}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma^2_y}\right)
\]

(9)

Using Gaussian Naive Bayes algorithm, the model can be fit by using following steps:

i. It is assumed that Gaussian distribution is done for data description and there is no covariance between the dimensions.

ii. Within each level, mean and then standard deviation of the points is calculated.

iii. Find the z-score distance between every point and each class-mean. It is the ratio of distance from the class mean and standard deviation of that class.

![Gaussian Naive Bayes](image)

**Figure 1.** Gaussian Naive Bayes.
3.3. Gradient Boost Method

Boosting is considered among the most important ideas that have been introduced in the last two decades. The main concept on which boosting works is that it combines the output of weak classifiers and create a very accurate prediction method. Those classifiers are termed as the weak classifiers which have slightly better accuracy than random guessing method. In boosting, weak classification algorithm is repeatedly applied to the modified version of data sample obtained at each step. In this way, a sequence of weak classifiers is obtained. At the end, the predictions obtained at each step are combined using majority vote and final prediction is thus calculated. This method was called Adaptive Boosting which is also known as AdaBoost. However, five years after the introduction of AdaBoost algorithm, it was discovered that the algorithm is equivalent to forward stage wise additive modeling based on exponential loss. Gradient Boost is another popular boosting algorithm. This algorithm is also called Gradient Boosted decision trees as it is used in conjunction with the trees.

\[
T(x; \theta) = \sum_{j=1}^{J} \gamma_j I(x \in R_j)
\]  

Here, I is an identity function which is 1 if \(x\) belongs to region \(R_j\) and is 0 otherwise. Above equation sums over all regions, \(R_1\) to \(R_J\), \(Y_j\) is the output which will be produced if \(x\) lies in \(R_j\). Above equation is of regression tree where \(\theta\) is the specification of all \(R_j\) and \(Y_j\) for all the regions. If it is regression then we know that the loss function is going to be the squared loss. The loss is incurred when actual output is \(y_i\) and predicted output is \(Y_j\). The boosting with the trees gives the equation:

\[
f_m(x) = \sum_{m=1}^{M} T(x, \theta_m)
\]  

The output of all \(M\) trees is taken, that gives the output of the boosted trees. Above equation shows a collection of trees i.e. it is a forest.

\[
\theta_m = \arg\min_{\theta_m} \sum_{j=1}^{N} L(y_j, f_{m-1}(x_j) + T(x_j, \theta_m))
\]

To calculate the loss function, the output produced by \(m-1\) stage trees and the output produced by \(m^{th}\) tree are observed. Given the region \(R_j\), finding \(Y_j\) is easy. But finding region in general is difficult. In case of squared error, finding it is not difficult. In case of squared error loss, that tree is picked which predicts the best residual.

\[
y_i = f_{m-1}(x_i), \gamma_m
\]

For two class problems and exponential loss function, it becomes exactly the same as Adaboost applied on decision trees. In case of differentiable loss function, the equation will be 

\[
L(f) = \sum_{i=1}^{N} L(y_i, f(x_i))
\]

\[
f = (f(x_1), ..., f(x_n))
\]

where \(f(x_i)\) is output obtained at a point \(x_i\). The same solution be \(h_0\). Then \(f_0 = h_0\) and it can be written as
Steepest descent will be used for optimization and $h_m = -\rho_m g_m$.

$$g_m = \left[ \frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]$$

The direction in which it should be proceed is $-g_m$ and it gives the direction of the steepest descent. $\rho_m$ is the step size taken in that direction.

$$\rho_m = \arg \min_{\rho} L(f_{m-1} - \rho g_m)$$

Once $\rho_m$ is obtained then $f_m = f_{m-1} - \rho_m g_m$. In the given equation $g_m$ does not have any constraint. At this point to fit a tree, approximation of $g_m$ will be done.

$$\theta_m = \arg \min_{\theta} \sum_{i=1}^{N} \left(-g_m - T(x_i, \theta)\right)^2$$

In above equation, squared error is used. While building the $m$th stage decision tree, regression is performed. The particular gradient descent direction is predicted for a particular input value $x_i$. The squared error is as following,

$$\frac{1}{2}(y_i - f(x_i))^2$$

The value of $-g_m$ can be calculated as equal to $y_i - f(x_i)$. At the end thus residual $y_i - f(x_i)$ is predicted. Using the basic concept of boosting same residual will be obtained. In gradient boosting, always regardless of loss function, regression with trees is performed. In case of classification, deviance is the loss function.

$$I(G_k) - p_k(x_i)$$

If $i$th class is $G_k$ then the probability of $x_i$ will be to lie in class $k$.

$$\gamma_m = \arg \min_{\gamma} \sum_{X_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$$

For all the part where the trees grow i.e. while performing stage wise boosting process, regression is used. When the final output is obtained, at that point, the true loss function can be applied. At every point, gradient is used to boost the performance. Gradient boost is an efficient AI based methods and it is widely used in many application these days.
4. Proposed AI based method
In this work, AI based method for alcohol abuse analysis has been carried out which is divided into various steps. The steps involved in AI based method are shown in Figure 2. In this work, python software has been used to carry out all the algorithm design and implementation work.

Figure 2. Flow diagram of proposed method.

4.1. Input Features Used
To validate proposed method, the data has been collected from UCI machine learning repository [9]. Table 1 shows the details of all the features used for the validation of the proposed method. The data consist of various types of data such as integer, categorical etc. The data sets are first normalized and the inputs are generated. The targets of the inputs are set to user and non-user.

4.2. Feature processing and selection
The original data contains 1817 alcohol users and 68 non users. An imbalanced data always affects the accuracy of the proposed method. Hence the data should be made balanced before it is used as input to classification algorithms. In this work, SMOTE Tomek is used as the pre-processing method. The features distribution before and after pre-processing has been shown in Figure 3. Figure 3(a) shows the data before processing and Figure 3(b) shows the data after processing with SMOTE.

4.3. Design of AI based method
In this work, three different AI methods have been used for AA analysis i.e. logistic regression algorithm, Naïve Bayes algorithm and gradient boosting algorithm. Design of each of the method has been described below in detail.
Table 1. Features used and its values

| Input Features | Values and Number of samples | Input Features | Values and Number of samples |
|----------------|------------------------------|----------------|------------------------------|
| Age (18-24) years – 643 | Sensation seeing -2.07848 - 71 | Male – 943 | Impulsive -2.55524 - 20 |
| (25-34) years – 481 | -1.54858 - 87 | | -1.37983 - 276 |
| (35-44) years – 356 | -1.18084 - 132 | | -0.71126 - 307 |
| (45-54) years – 294 | -0.84637 - 169 | | -0.21712 - 355 |
| (55-64) years – 93 | -0.52593 - 211 | | 0.19268 - 257 |
| over 65 years – 18 | -0.21575 - 223 | | 0.52975 - 216 |
| Gender Female – 942 | | | 0.88113 - 195 |
| Male – 943 | 0.07987 - 219 | | 1.29221 - 148 |
| | 0.40148 - 249 | | 1.86203 - 104 |
| | 0.76540 - 211 | | 2.90161 - 7 |
| Education Left school before 16 years - 28 | Agreeableness (Ascore) (12 – 20) score - 3 | Country Australia - 54 | Conscientiousness (Cscore) (17 – 20) score - 5 |
| Left school at 16 years - 99 | (21 – 30) score - 57 | Canada - 87 | (21 – 30) score - 131 |
| Left school at 17 years - 30 | (31 – 40) score - 602 | New Zealand - 5 | (31 – 40) score - 650 |
| Left school at 18 years - 100 | (41 – 50) score - 750 | Other - 118 | (41 – 50) score - 942 |
| Some college or university but no | (51 – 60) score - 216 | Republic of Ireland - 20 | (51 – 60) score - 157 |
| certificate or degree - 506 | | UK - 1044 | |
| Professional certificate/diploma | | USA - 557 | |
| - 270 | | | |
| University degree - 480 | | | |
| Masters degree - 283 | | | |
| Doctorate degree - 89 | | | |
| Ethnicity Asian - 26 | Extraversion (Escore) (16 – 20) score - 12 | Neuroticism (Nscore) (12 – 15) score - 13 | Openness (Oscore) (24 – 30) score - 30 |
| Black - 33 | (16 – 20) score - 56 | (12 – 15) score - 13 | (16 – 20) score - 56 |
| Mixed-Black/Asian - 3 | (21 – 30) score - 143 | (16 – 20) score - 56 | (21 – 30) score - 143 |
| Mixed-White/Asian – 20 | (31 – 40) score - 832 | (16 – 20) score - 56 | (31 – 40) score - 832 |
| Mixed-White/Black - 20 | (41 – 50) score - 781 | (16 – 20) score - 56 | (41 – 50) score - 781 |
| Other - 63 | (51 – 59) score - 94 | (16 – 20) score - 56 | (51 – 59) score - 94 |
| White - 1720 | | (16 – 20) score - 56 | |
4.3.1. Logistic regression algorithm-based method
In this work, logistic regression (LR) algorithm is used to predict AA user and non user. LR-AA module is designed using various input features. In this work, the algorithms are designed in two ways. One is test sample method in which samples has been divided into 85% and 15% for training and testing respectively. Second method is leaving one out cross validation (LOOCV). The target of the method is drug user as ‘1’ and nonuser as ‘0’. The LR-AA module is designed with input and target based on different parameters of the algorithm. Performance of the method is analyzed in terms sensitivity, specificity, accuracy etc.

Figure 3. (a) Original features (b) Features after processing with SMOTE.

4.3.2. Naïve Bayes algorithm-based method
In this work, Naïve Bayes (NB) algorithm also has been used for prediction of AA user and non user. Gausian NB algorithm is used in this work and NB-AA module is designed using input features The target of the method is drug user as ‘1’ and nonuser as ‘0’. The NB-AA module is designed with input and target based on different parameters of the algorithm.

4.3.3. Gradient Boosting algorithm based method
In this work, gradient boosting (GB) algorithm has been used for prediction of AA user and non user. The GB-AA module is designed with input and target based on different parameters of the algorithm. Gradient boost classifier with out-of-bag estimates is used which is a heuristic to estimate optimal number of iterations for boosting. In this work, 1200 subsamples are taken where learning rate=0.01 and random state= 3. The target of the method is drug user as ‘1’ and nonuser as ‘0’.

5. Results
In this work, performance of AI based method has been validated with all 12 input features as given in Table 1 and selected 5 input features. The selected 5 features are age, education, neuroticism, gender and sensation seeking. AI based method has been validated using both test sample estimate and LOOCV. All these three measures used for performance analysis depend on negative and positive values. Here TP=true positive, TN=true negative, FP= false positive and FN= false negative. Accuracy of the AI based method is

\[
\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP} \times 100
\] (23)
Sensitivity is calculated as

\[ \text{Sensitivity} = \frac{TP}{TP + FN} \times 100 \]  

(24)

Specificity is calculated as

\[ \text{Specificity} = \frac{TN}{TN + FP} \times 100 \]  

(25)

5.1. Performance analysis of LR-AA AI module

Logistic regression algorithm-based AI method is tested against the trained LR-AA module. The AI based method has been designed using all 12 features and selected 5 features. Performance of the LR-AA AI method for test sample estimate is given in Table 2. It can be observed that if imbalance data is taken then sensitivity is 0 while the specificity is 100. After balancing the samples with SMOTE it was observed that sensitivity is 69.76 while the specificity is 67.71 using 12 features. It was observed that sensitivity is 68.75 while the specificity is 68.77 using 5 features. The accuracy in prediction is nearly same for both features.

| Number of Features | Features Processing | Sensitivity | Specificity | Total | Accuracy |
|--------------------|---------------------|-------------|-------------|-------|----------|
| 12                 | Without SMOTE       | 0.00        | 100.0       | 100.0 | 96.46    |
|                    | With SMOTE          | 69.76       | 67.71       | 137.48| 68.68    |
| 5                  | Without SMOTE       | 0.00        | 100.0       | 100.0 | 96.46    |
|                    | With SMOTE          | 68.75       | 68.77       | 137.52| 68.76    |

Performance of the LR-AA AI method for LOOCV is given in Table 3. It can be observed that if imbalanced data is taken then sensitivity is 0 while the specificity is 100 when imbalanced data is taken which should not be considered as correct method. After balancing the samples with SMOTE it was observed that sensitivity is 71.71 while the specificity is 71.88 using 12 features. It was observed that sensitivity is 70.64 while the specificity is 64.98 using 5 features. The highest accuracy of the LR based AI method is 71.79% with 12 features. Figure 4(a) shows the ROC curve for highest accuracy obtained and Figure 4(b) shows the performance of the LR method with LOOCV.

| Number of Features | Features Processing | Sensitivity | Specificity | Total | Accuracy |
|--------------------|---------------------|-------------|-------------|-------|----------|
| 12                 | Without SMOTE       | 0.00        | 100.0       | 100.0 | 96.39    |
|                    | With SMOTE          | 71.71       | 71.88       | 143.59| 71.79    |
| 5                  | Without SMOTE       | 0.00        | 100.0       | 100.0 | 95.59    |
|                    | With SMOTE          | 70.64       | 64.98       | 135.63| 67.81    |
5.2. Performance analysis of NB-AA AI module

Naïve Bayes algorithm based AI method is tested against the trained NB-AA module. Performance of the NB-AA AI method has been shown in Table 4 for test sample estimate. It can be observed that if imbalance data is taken then sensitivity is 14.29 for 12 features and 0 for 5 features which should not be considered. After balancing the samples with SMOTE it was observed that sensitivity is 65.63 while the specificity is 67.44 using 12 features. It was observed that sensitivity is 67.38 while the specificity is 68.32 using 5 features. The accuracy in prediction is nearly same for both features.

Table 4. Performance of NB-AA AI module using test sample method

| Number of Features | Features Processing | Sensitivity | Specificity | Total | Accuracy |
|--------------------|---------------------|-------------|-------------|-------|----------|
| 12                 | Without SMOTE       | 14.29       | 96.74       | 111.025 | 94.69    |
|                    | With SMOTE          | 65.63       | 67.44       | 133.06 | 66.48    |
| 5                  | Without SMOTE       | 0.0         | 96.45       | 96.45  | 96.11    |
|                    | With SMOTE          | 67.38       | 68.32       | 135.70 | 67.84    |

Figure 4. (a) ROC for LR method with LOOCV (b) Performance for LR method with LOOCV.
Performance of the NB-AA AI method with LOOCV is given in Table 5. It can be observed that if imbalance data is taken then sensitivity is 7.35 for 12 features and 0 for 5 features which should not be considered as correct method. After balancing the samples with SMOTE it was observed that sensitivity is 73.47 while the specificity is 62.08 using 12 features. It was observed that sensitivity is 72.48 while the specificity is 61.59 using 5 features. The highest accuracy of the NB based AI method is nearly 67%.

Table 5. Performance of NB-AA AI module using LOOCV

| Number of Features | Features Processing | Sensitivity | Specificity | Total | Accuracy |
|--------------------|---------------------|-------------|-------------|-------|----------|
| 12                 | Without SMOTE       | 7.35        | 98.18       | 105.53| 94.90    |
|                    | With SMOTE          | 73.47       | 62.08       | 135.55| 67.78    |
| 5                  | Without SMOTE       | 0.00        | 99.83       | 99.83 | 95.44    |
|                    | With SMOTE          | 72.48       | 61.59       | 134.07| 67.03    |

5.3. Performance analysis of GB-AA AI module

Gradient boosting algorithm based AI method is tested and performance of the GB-AA AI method is given in Table 6. It can be observed that if imbalance data is taken then sensitivity is 0 and specificity is 96.45. After balancing the samples with SMOTE it was observed that sensitivity is 98.87 while the specificity is 96.43 using 12 features. It was observed that sensitivity is 98.43 while the specificity is 93.00 using 5 features. The accuracy in prediction is nearly 96% for both features.

Table 6. Performance of GB-AA AI module using test sample method

| Number of Features | Features Processing | Sensitivity | Specificity | Total | Accuracy |
|--------------------|---------------------|-------------|-------------|-------|----------|
| 12                 | Without SMOTE       | 0.0         | 96.45       | 96.45 | 96.11    |
|                    | With SMOTE          | 98.87       | 96.43       | 195.30| 97.61    |
| 5                  | Without SMOTE       | 0.0         | 96.47       | 96.47 | 96.47    |
|                    | With SMOTE          | 98.43       | 93.00       | 191.43| 95.56    |

Performance of the GB-AA AI method with LOOCV is given in Table 7. It can be observed that if imbalance data is taken then sensitivity is 0 and 7.69. After balancing the samples with SMOTE it was observed that sensitivity is 99.29 while the specificity is 94.21 using 5 features. The highest accuracy of the GB based AI method is 97.55% with 12 features. Figure 5 (a) shows the ROC curve for highest accuracy obtained and Figure 5 (b) performance of GB method with LOOCV. Hence GB based AI method can be used efficiently for user and non user of alcohol correctly.

Table 7. Performance of GB-AA AI module with LOOCV

| Number of Features | Features Processing | Sensitivity | Specificity | Total | Accuracy |
|--------------------|---------------------|-------------|-------------|-------|----------|
| 12                 | Without SMOTE       | 7.69        | 96.42       | 104.11| 95.81    |
|                    | With SMOTE          | 98.76       | 96.40       | 195.16| 97.55    |
| 5                  | Without SMOTE       | 0.00        | 100.00      | 100.00| 96.39    |
|                    | With SMOTE          | 99.29       | 94.21       | 193.50| 96.61    |

5.4. Evaluation of all AI methods

Based on the results obtained as discussed in the previous section, AI based method has been analyzed. Figure 6 (a) shows the measures used to evaluate the correctness of AI methods for test
sample estimate method. Sensitivity, specificity, combined both and accuracy of AI method with 12 features has been shown. All the measures used are higher for gradient boosting method than all other method.

Figure 5. (a) ROC for GB method with LOOCV (b) Performance for GB method with LOOCV.

Figure 6(b) shows the measures used to evaluate the correctness of AI methods with LOOCV method. Sensitivity, specificity, combined both and accuracy of AI method with 12 features has been shown. All the measures used are higher for gradient boosting method than all other method. Hence it can be concluded that GB method is the optimal method for analysis of alcohol abuse.

Figure 6. (a) Performance of the AI method for test sample estimate (b) Performance of the AI method for LOOCV.
5.5. Comparison

For prediction of alcohol user and non-user various methods have been suggested in literature. The proposed AI-based method has been compared in terms of data set used, methods used, attribute used, specificity, sensitivity and combined as given in Table 8. In [6], author has used linear discriminant analysis in which combined sensitivity and specificity is 138.58. In [8], author has used artificial neural network in which accuracy is 98.70%. But drawback of the method is that it is designed with imbalanced data and it may not get that accuracy if data is balanced. Proposed GB-based AI method gas combined sensitivity and specificity of 195.16. The proposed gradient boosting-based AI method has highest specificity among all. Hence AI method can be used for alcohol abuse analysis effectively.

| Parameters | Fehrman et al [6] | Kumari et al [8] | Proposed Method |
|------------|-------------------|------------------|------------------|
| Features used | 5 | 12 | 12 |
| Pre-processing technique used | Sparse pca | No | SMOTE |
| Methods used | Linear discriminant analysis | Artificial Neural Network | Logistic regression | Gradient boosting |
| Sensitivity | 75.34 | - | 71.71 | 98.76 |
| Specificity | 63.24 | - | 71.88 | 96.40 |
| Combined | 138.58 | 98.70 | 143.59 | 195.16 |

6. Conclusion

Drug risk analysis has been done by various researchers by using various methods where accuracy in prediction is a matter of concern. In this work, AI-based method has been used to predict the alcohol user and non-user. A comparative study of three AI methods namely LR, NB and GB has been carried out. It has been observed that gradient boosting method is more accurate to predict alcohol user than other two methods. The sensitivity and specificity obtained for the proposed method is 75.39 and 78.59 respectively which is much improved than the method proposed previously. The result of the proposed method shows that it can be implemented effectively for analysis of alcohol abuse. The future possibility of the proposed work is to predict the time of alcohol abuse correctly.

References

[1] Vollrath M, Torgersen S 2002 Who takes health risks? a probe into eight personality types Personality and Individual Differences 32 1185–1197.
[2] Bulut F, Bucak IO 2014 An urgent precaution system to detect students at risk of substance abuse through classification algorithms Turk J Elec Eng & Comp Sci 22 690–707.
[3] Bergh A 2019 A Machine Learning Approach to Predicting Alcohol Consumption in Adolescents From Historical Text Messaging Data Computational and Data Sciences (MS) Theses 5-28.
[4] Belcher AM, Volkow ND, Moeller FG, Ferre S 2016 Personality traits and vulnerability or resilience to substance use disorders Trends in cognitive sciences 18 211-217.
[5] Weissman DG, Schriber RA, Fassbender C, Atherton O, Krafft C, Robins RW, Hastings PD, Guyerb AE 2015 Earlier adolescent substance use onset predicts stronger connectivity between reward and cognitive control brain networks Developmental Cognitive Neuroscience 16 121–129.
[6] Andreasen CS, Griffiths MD, Gjertsen SR, Krossbakken E, Kvaam S, Pallesen S 2013 The relationships between behavioral addictions and the five-factor model of personality Journal of Behavioral Addictions 2 90–99.
[7] Fehrman E, Muhammad A, Mirkes E, Egan V, Gorban A 2017 The Five Factor Model of personality and evaluation of drug consumption risk Data Science Studies in Classification, Data Analysis, and Knowledge Organization231-242.
[8] Kumari D, Kilam S, Nath P, Swetapadma A 2018 Prediction of alcohol abused individuals using artificial neural network Int. j. inf. tecnol. 10 233–237.
[9] https://archive.ics.uci.edu/ml/datasets/Drug consumption.
[10] Lee M, Sankar V, Hammer A, Kennedy W, Barb J, McQueen P, Leggio L 2019 Using machine learning to classify individuals with alcohol use disorder based on treatment seeking statusEClinical Medicine 12 70-78.
[11] Wang Z, Wu C, Zheng K, Niu X, Wang, X 2019 SMOTEtomek-Based Resampling for Personality Recognition IEEE Access 7 129678-129689.
[12] Friedman J 2001 Greedy Function Approximation: A Gradient Boosting Machine The Annals of Statistics 29 1189-1232.
[13] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP 2002 SMOTE: Synthetic Minority Over-sampling Technique Journal of Artificial Intelligence Research 16 321–357.
[14] Batista GEAPA, Prati RC, Monard MC 2004 A study of the behaviour of several methods for balancing machine learning training data. SIGKDD Explor 6 20–29.
[15] Fernandez A, Garcia A, Herrera F, Chawla NV 2018 SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary Journal of Artificial Intelligence Research 61 863-905.
[16] Ghosh S, Jain R, Jhanjee S, Rao R, Mishra AK 2019 Alcohol Biomarkers and their Relevance in Detection of Alcohol Consumption in Clinical Settings Int Arch Subst Abuse Rehabil 11-8.
[17] Fairbairn CE, Kang D, Bosch N 2020 Using machine learning for real-time BAC estimation from a new-generation transdermal biosensor in the laboratory Drug and Alcohol Dependence 216 108205.
[18] Venegas A, Ray LA 2019 Comparing alcohol cue-reactivity in treatment seekers versus non-treatment-seekers with alcohol use disorder The American Journal of Drug and Alcohol Abuse 46 1-8.