


# An Improved Supervised Classification Algorithm in Healthcare Diagnostics for Predicting Opioid Habit Disorder

Khushboo Jain, DIT University, India\*

 <https://orcid.org/0000-0002-4166-2591>

Akansha Singh, Noida Institute of Engineering and Technology, India

Poonam Singh, Hindustan College of Science and Technology, India

Sanjana Yadav, Hindustan College of Science and Technology, India

## ABSTRACT

Opioid habit disorder (OHD), which has become a mass health epidemic, is defined as the psychological or physical dependency on opioids. This study demonstrates how supervised machine learning procedures help us investigate and examine massive data to discover the hidden patterns in any disease to deliver adapted dealing and predict the disease in any patient. This work presents a generalized model for forecasting a disease in the healthcare sector. The proposed model was investigated and tested using a reduced feature set of the opioid habit disorder (OHD) dataset collected from the National Survey on Drug Use and Health (NSDUH) using an improved iterative dichotomiser 3 (pro-IDT) algorithm. The proposed healthcare model is also compared with further machine learning algorithms such as ID3, random forest, and bayesian classifier in Python programming. The performance of the proposed work and other machine-learning algorithms has estimated for accuracy, precision, misclassification rate, recall, specificity, and F1 score.

## KEYWORDS

Classification Algorithm, Decision Tree, Healthcare Diagnostic, Opioid, Opioid Habit Disorder, Random Forest, Supervised Learning, Support Vector Machine

## INTRODUCTION

Opioid habit disorder is characterized as dependence on opioids, a medication present in many legal prescription medications, and illicit drugs such as heroin. The World is in the middle of an outbreak of drug addiction, and it has now become a problem for public health. Growing the risk of relapse can be due to many causes, including mental disorders, intimate and social interaction difficulties, exposure to other opioid users, and previous stressful experiences (National Institute on Drug Abuse, 2020). Contemporary research has correlated those demographic, social, physical, and psychological factors with drug use diseases. However, several of these experiments aim to understand drug use disorders

DOI: 10.4018/IJRQEH.297088

\*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

by considering each function in isolation or one at a time. For example, the Department of Health and Human Services (DHHS) reports that the risk of drug misuse is raised by living with a disability (National Rehabilitation Information Center, 2011). Similarly, a correlation between mental illness and drug abuse (Saffer & Dave, 2002) is documented by the National Bureau of Economic Research. To decide who is likely to acquire OHD, it is essential to step from clarification to prediction, where intricate correlations can be considered between these characteristics, which will assist public health authorities in preparing and providing appropriate plans for action and assistance.

This study investigates supervised machine learning to construct a classifier based on an interconnected consideration of social, social, physical, and psychological characteristics to identify individuals at risk for OHD. It aims to improve algorithm performance using a general framework for Opioid Habit Disorder (OHD) Diagnostic Model in order:

- To investigate and examine massive data to discover the hidden patterns in any disease to deliver adapted dealing and predict the disease in any patient.
- To propose a generalized model (pro-IDT) for forecasting a disease in the healthcare sector.
- To compare the proposed model with machine learning algorithms such as ID3, Random Forest, and Support Vector Machine in Python programming.
- To access the performance analysis of proposed work and other machine-learning algorithms in accuracy, precision, misclassification rate, recall, specificity, and F1 score.

## BACKGROUND

To analyze factors impacting the use of opioids over a long period, multiple studies on long-term patterns in opioid use have been carried out. e.g., risk factors such as preoperative opioid use along with younger age, anxiety or depression, females, and other intrinsic factors are reported in (Bedard et al., 2017) with a higher number of refills or use at 12 months post-TKR. In (Martel et al., 2013), identical conclusions are drawn. It is seen in (Goesling et al., 2016) that the risk factors include more extreme pain, poor functioning, signs of depression, and a higher preoperative dosage of opioids. Good predictors for chronic opioid use are indicated for the form of surgery, prolonged hospital visits, discharge to the recovery unit, preoperative opioid use, higher comorbidity score, back pain, migraine, and smoking at baseline (Miotto et al., 2016). Also, previous use of opioids is often recognized as a significant risk factor for prolonged opioid use (Rozell et al., 2017). Stress or anxiety and discomfort are summarized in (Li et al., 2018) to be significantly linked to chronic pain following surgery or drug utilization postoperatively.

In addition to mathematical methods, machine learning techniques have also been used to study opioid use. The predominant logistic regression models are to find correlations between long-term opioid usage trends and risk factors (Acion et al., 2017; Ahn et al., 2016) or to forecast chronic opioid use or dependency on substances (Wadekar, 2019). The restricted study has begun to use more sophisticated machine learning approaches beyond logistic regression (McAnally, 2017). Hundreds of thousands of EHR reports of patients have been used in these experiments to train the models. However, the medical data are gathered unaware of the reasons for the use of opioids.

Also, the studies mentioned above have a common drawback that the EHR data, patient claims data, or history of opioid prescribing used in their research require the only study of long-term opioid usage trends, such as opioid use after six months of first prescription or opioid use on an ongoing basis. Thus, these trials, which rely on the long-term planned usage of opioids, are insufficient to identify JTR patients in terms of their short-term opioid needs and meet the need to help patients administer the correct number of opioids. Therefore, to describe the over-prescription nature of opioid usage in the short term, additional information sources are necessary other than EHR or prescription history.

A patient survey was performed in this study and used as a critical data source to analyze the actual consumption of opioids and estimate their degree. However, there is always a shortage of data

due to several factors, such as inability to participate or unacceptable answers, in the surveys (Ellis et al. 2019). Easy methods, such as discarding the participants with missed response values, are usually unsatisfactory in coping with missing data and frequently contribute to extreme biases in parameter estimation (Leurent et al., 2018). It has been generally agreed that all available evidence can be used to make a rational conclusion; the advantages of this are now well known.

Based on the missing mechanisms, the missing data in the surveys may belong to one of the three groups implemented in (Green et al. 2019): missing absolutely at random (MCAR), missing at random (MNAR), and missing at random (MNAR) (MAR). MCAR applies to the knowledge that the risk of being missed is entirely independent of all concern considerations. In this case, the data observed represents the complete data and can be used for interpretation without considering the missing portion of the data. The probability of a value being absent in the second class depends on the missing value itself. For analysis using MNAR data, additional untestable assumptions are needed as some relevant information is unnoticed. Finally, MAR means that the missing likelihood could be contingent on the observed data in these data but conditionally independent of the underlying significance provided in the observed data itself.

A classification model using data with missed target values and the labeled data is referred to as semi-supervised learning, where there are missing values in the target variable. Pseudo-labeling methods imputing the missing goal values are also leveraged by different semi-supervised learning paradigms (Cochran et al., 2014). Thus, a modern pseudo-labeling methodology using imputation can be developed and used tailored to the data obtained at SMH.

In summary, through current literature attempts, there is no available research that utilizes and estimates the actual extent of short-term opioid use in patients. The purposes of this paper are to gather and evaluate specific knowledge and, through a suggested supervised machine learning, to establish a classification model.

## **A GENERAL FRAMEWORK FOR OPIOID HABIT DISORDER (OHD) DIAGNOSTIC MODEL**

This section presents the general framework for OHD diagnostic model to forecast an opioid. This framework has three levels, and the workflow is described in Figure 1.

### **Databases and Data Collection**

The patient records were kept in the physical format by hospitals in previous days, and it is challenging to examine. In present times, the hospitals have started maintaining their patients' medical records in Automated Health Record (PHR) format due to computerization and future climate to more conveniently manage the patient data. The health care system can gather the data digitally using intelligent devices such as smartphones and even more advanced smartwatches with technology development. The data collected and diagnosed at the edge of these IoT-enabled devices must be processed further by sending this data directly to the cloud-based data repository (Chahal & Gulia 2016). Such technologies make multiple points of collection of data possible.

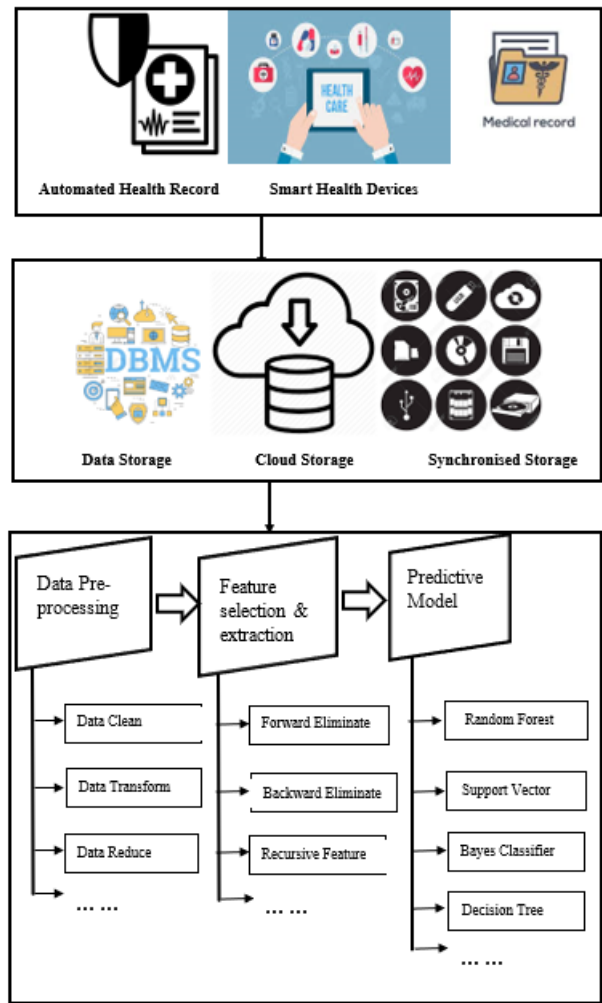
### **Data Backups**

Hospitals are used to maintain the archives for storing patient's health records. Hospitals retained patient data having minimal storage nature in their hospitals during the early days. Today, they have stored data in the cloud with improvements in storage systems. This role would permit a patient's data accessibility to the doctor from anywhere and at any time.

### **Opioid Habit Disorder (OHD) Diagnostic Model**

It has three subdivisions, which are as follows.

Figure 1. General framework for Opioid Habit Disorder (OHD)



*Data Pre-Processing*

Data pre-processing performs an essential role in eliminating unnecessary and noisy data from data analytics. Data sources can often collect unnecessary and incomplete data. During the pre-processing data model, these issues were discussed. In machine learning, the following techniques deal with missing data:

- Ignore the missing fields by eliminating the missing values in the dataset: the most widely used data analytics method ignores missing fields. Only if you have appropriate samples in the dataset is this approach suitable.
- Use similar values from the previous dataset to complete the missing values: In this technique, the dataset’s historical values helped determine the missing values. It will be similar to the pre-existing value in this technique and fill relevant values in the missing fields. The final strategy was to complete the blank values using methods of computation.

- Utilize Mean to fill the blank value method: In this technique, features' mean calculated value is used to complete the values. When the dataset is limited in size, this technique would be useful.
- Use the possible value to complete the blank value: To achieve the empty values, we may use linear regression models.

However, we used the technique of "Eliminating the incomplete values by neglecting the blank fields in the dataset" as a pre-processing data methodology in this work.

### *Feature Selection and Extraction*

This methodology allows us to select the most relevant dataset features as the data source consists of the data generated with multiple attributes. In the dataset, some unwanted attributes are also generated, which do not affect the prediction performance so that such attributes can be removed during predictive model construction. For the data model, let us consider the dataset consists of " $N$ " attributes with " $M$ " samples, and it is denoted as in equation (1):

$$DS = (a_1, a_2, a_3, \dots, a_N) \quad (1)$$

Here " $a$ " is an attribute or feature in the dataset, and " $N$ " is the number of features. The classifier's accuracy may not depend on the complete set of attributes of the dataset " $DS$ " and it may rely only on the dataset's selected features. This selected feature subset is represented as in equation (2). Now the system is proposed with a Chi-square model to identify the best features from the given dataset:

$$DS' = \{a_1, a_2, a_3, \dots, a_n\}, \text{ where } n < N \quad (2)$$

A new subset will be created from the features that are selected with the help of Feature extraction. The original features of the data are transformed into a feature set that is in reduced form. This process is known as Feature extraction.

### *A Predictive Model for OHD*

In this framework, the prediction accuracy depends on the quality of data and the algorithms utilized during the prediction process (Chen et al., 2017).

## **ALGORITHMIC DESCRIPTIONS FOR OPIOID HABIT DISORDER (OHD) DIAGNOSTIC MODEL**

We address the setup on the experimental basis of the predictive analytics model (Jain & Kumar, 2020) for disease prediction in this section, as represented in Figure 1. In 1986, Ross Quinlan (Quinlan 1986) proposed the Iterative Dichotomiser3 algorithm, the conventional decision tree classification algorithm. ID3 is a supervised learning classification algorithm that utilizes information gain as a methodology for selecting attributes. The primary basis for the creation of an Iterative Dichotomiser 3 algorithm-based decision tree is as follows.

Following notation used in ID3 algorithm: Information entropy  $Entropy(DS')$  and information gain  $Info_{Gain}(DS', a)$  are stated as:

$$Entropy(DS') = -\sum_{i=1}^c p_i \log_2 p_i \quad (3)$$

$$Info_{Gain}(DS', a) = Entropy(DS') - \sum_{v=1}^V \left| \frac{DS'^V}{DS'} \right| \times Entropy(DS'^V) \quad (4)$$

where  $DS' = f\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$  denotes the sample learning dataset. The likelihood that a tuple in a sample learning data belongs to the class  $C_i$  is denoted by  $p_i$ .  $A = (a_1, a_2, \dots, a_d)$  signifies the attribute set of  $DS'$  and  $d = \{1, 2, \dots, i\}$ .

The ID3 approach can be used to obtain the decision tree model, which typically contains some internal nodes, few leaf nodes, and one root node. Here, the root node denotes the tests' set entirely, the attribute is denoted by the internal node, and the leaf node denotes a class name. A defined sequence corresponds to the path from the root node to each leaf node. There are many advantages to classifying the bulk data by implementing the ID3 algorithm, like useful instinctual functionality, simple decomposability, etc. Still, the following drawbacks of the ID3 algorithm cannot be ignored, such as:

- ID3 prefers choosing the attribute having more values (De Mántaras, 1991) by using the data gain as an attribute selection process since the knowledge gain equation value of such attribution type would be more than others.
- In the ID3 system, in the optimal attribution selection process, several logarithmic computations will waste much time calculating the knowledge gain. If the four arithmetic operations modify the logarithmic expression, the running time will increase rapidly in the entire building tree phase. In the course of creating a decision tree, it is difficult to regulate the tree size.
- Most improved approaches currently use pruning methods (Leung et al., 2001) to prevent over-fitted phenomena, which would result in the full two-step (i.e., modeling and pruning) phase of constructing decision tree models. That will save much time if a succinct decision tree is created in one step.

This work aims to solve the drawbacks mentioned above and propose an enhanced method for building a classification algorithm model. The enhanced form of the Iterative Dichotomiser- 3 algorithms is constructed to produce a decision tree model concisely and select an optimal attribute that is quite reasonable in a much shorter time for each internal node.

## Databases and Data Collection

Based on the above challenges, this article describes the related solutions. A new technique is designed to bring the three solutions together and can be used in a shorter runtime than ID3 to create a more concise decision tree. In the ID3 algorithm, the optimal attribute selection is performed based on equation (4), but the logarithm algorithm increases the calculation difficulties. The decision tree construction would speed up faster if we could find a more straightforward computational formula. The simplification process is structured as follows.

As is known to all, the speed of running of the logarithmic expression is less as compared to the four arithmetical operations in two identical expressions. The entire decision tree building process is running speed would rapidly increase if the four-arithmetic operation replaced the logarithmic expression in equation (4). In more advanced mathematics, the sense of the Taylor formula would simplify functions that are complex according to differentiation theory. The Taylor method is an

extended form at any stage, and the Maclaurin formula is a function that can be expanded into Taylor's series at point zero.

The complexity of computing the entropy of information about the ID3 algorithm can be reduced based on the Maclaurin method's approximation formula, which allows creating a decision tree within a short time. The formula for Taylor is given as:

$$f(y) = f(y_0) + f'(y_0)(y - y_0) + o(y - y_0) \quad (5)$$

On substituting  $y_0 = 0$ , the above equation changes into the Maclaurin formula:

$$f(y) = f(0) + f'(0)\frac{y}{1!} + f''(0)\frac{y^2}{2!} + f'''(0)\frac{y^3}{3!} + \dots + f^n(0)\frac{y^n}{n!} + S(y) \quad (6)$$

where the value of  $S(y)$  will be:

$$S(y) = f^{n+1}(y)\frac{(y - y_0)^{n+1}}{(n+1)!} \quad (7)$$

To simplify the above equation, we are taking the value of  $f(x)$  as below:

$$f(x) = f(0) + f'(0)\frac{y}{1!} + f''(0)\frac{y^2}{2!} + f'''(0)\frac{y^3}{3!} + \dots + f^n(0)\frac{y^n}{n!} \quad (8)$$

Let us assume that  $n$  is the number of negative instances and  $p$  positive instances in the learning sample set  $D'$ , the entropy of  $D'$  in equation 10, will be written as:

$$Entropy(DS') = -\frac{p}{(n+p)}\log_2\frac{p}{(n+p)} - \frac{n}{(n+p)}\log_2\frac{n}{(n+p)} \quad (9)$$

Let us assume that there are  $V$  distinct values included in the attribute  $a_i$  of learning dataset  $D'$ , and each value contributes to either  $n_i$  negative instances or  $p_i$  positive instances, so the information gain of an attribute  $a_i$  can be represented as follow:

$$Info_{Gain}(DS', a) = Entropy(DS') - \sum_{v=1}^V \frac{n_i + p_i}{(n+p)} \times Entropy(DS'^v) \quad (10)$$

where:

$$Entropy(DS'^v) = -\frac{p_i}{(n_i + p_i)}\log_2\frac{p_i}{(n_i + p_i)} - \frac{n_i}{(n_i + p_i)}\log_2\frac{n_i}{(n_i + p_i)} \quad (11)$$

To simplify, if the principle  $\ln(x+1) \cong x$  is considered an assertion in the case of the minimal value of  $x$  and the constant can be ignored. Using this assumption, we have written Equation (8) and Equation (9) as follow:

$$Entropy(DS') = \left[ -\frac{1}{(n+p)\ln(2)} \left\{ -p \ln\left(\frac{p}{n+p}\right) - n \ln\left(\frac{n}{n+p}\right) \right\} \right] = \frac{2n \times p}{n+p} \quad (12)$$

Similarly, the term in equation (10) can be recomputed as:

$$\sum_{v=1}^V \frac{n_i + p_i}{(n+p)} \times Entropy(DS'^V) = \sum_{v=1}^V \frac{2n_i \times p_i}{(n_i + p_i)} \quad (13)$$

Thus, equation (10) can also be written as:

$$Info_{Gain}(DS', a) = \frac{2n \times p}{n+p} - \sum_{v=1}^V \frac{2n_i \times p_i}{(n_i + p_i)} \quad (14)$$

Equation (14) will then be used to measure each attribute's information gain, and the attribute having the highest gain of information will be chosen to be made the decision tree node. Equation (14), which only involves subtraction, addition, division, multiplication, and the expression of novel information gains, decreases the difficulties faced during computation and enhances the capacity to handle data to a great extent.

## ASSESSMENT OF (PRO-IDT) IN OPIOID HABIT DISORDER (OHD)

### Dataset

The National Survey on Drug Use and Well-being (NSDUH) datasets performed by the Administration of Substance Abuse and Mental Health Services using python programming language have experimented with the framework. Supervised machine learning requires a labeled data set where each entity is marked as having or not having OHD. It was created using the responses from the 2019 edition of the National Drug Use and Health Survey (NSDUH) conducted by the Administration of Substance Abuse and Mental Health Services (Dataset, 2019), as a specified data set was not readily available. The 2019 NSDUH study poses questions on drug use for the first time. Depending on whether opioid addiction or violence occurred or not, each finding was labeled as "OHD" or No-OHD." Table 1 lists the 11 individual features and their levels included in the labeled data collection.

We have identified essential features from the reduced features discussed above from each of the following supervised classification algorithms: decision tree, Bayesian Classification, Random forest trees, and improved decision tree.

### Assessment Parameters

The primary system's goal is to assist doctors in shortening the time for diagnosis so that they can start treatment as early as possible. Using the confusion matrix, the accuracy of the prediction model is determined. With the actual count and expected count, the uncertainty matrix is formulated. The uncertainty matrix is shown in Table 2, with the quantity of correct and incorrect predictions.



**Table 1. Attributes of data collected from the National Survey on Drug Use and Health (NSDUH)**

S. No	Attributes	Level
1	Gender	Male and Female
2	Age	Married, Widowed, Divorced or Separated and Never Been Married
3	Age Group	12-17, 18-25, 26-34, 35-49, 50-64 and 65 or Older
4	Income	Less than \$20,000, \$20,000-\$49,999, \$50,000-\$74,999, \$75,000 greater than
5	Employment	Full Time, Part Time, Unemployed and others (including students, persons keeping the house or caring for children full time, retired or disabled persons, or other persons not in the labor force)
6	Education	High School Graduate, Some College and College Graduate
7	One or more cognitive, visual, dressing, walking, hearing, and running errands disability.	-
8	Any mental illness, including psychological distress and suicidal thoughts	-
9	The first use of alcohol before 18 years	-
10	The first use of marijuana before 18 years	-
11	Overall health	Excellent, Very Good, Good, Fair/Poor

**Table 2. Confusion Matrix for classification algorithms**

	Prediction-No	Prediction-Yes
Actual-Yes	False Negative (FN)	Real Positive (TP)
Actual-No	Real Negative (TN)	False Positive (FP)

We have used the following metric to evaluate the performance of various supervised learning algorithms:

1. **Accuracy:** Accuracy determines how much the prediction model classifies the proper output. The number of accurately corrected predictions can be defined as the number of all predictions in total, and it is marked in equation (15) as:

$$Accuracy = \frac{TP + TN}{Total\ predictions} \quad (15)$$

2. **Precision:** Precision attempts to explain how accurate it is when the prediction model categorizes it. It can be described in the form of a classifier that predicts the predictions that are favorable to the overall optimistic predictions, i.e. (both true positive and false positive) and is expressed in equation (16) as:

$$Precision = \frac{TP}{TP + FP} \quad (16)$$

3. **Specificity:** Specificity is intended to determine how much no is expected by the prediction model when it is no. It can be described as a negative classifier for the prediction of real negatives. It is called the Real Negative Rate, too. It is expressed in equation (17) as:

$$Specificity = \frac{TN}{TN + FP} \quad (17)$$

4. **Misclassification rate:** The misclassification rate determines how much the prediction model classifies the wrong performance. It can be described in the form of a classifier, which makes improper forecasts. It has also known as the rate of error and is expressed in equation (18) as:

$$Error\ rate = \frac{FP + FN}{Total\ predictions} \quad (18)$$

5. **Recall:** The purpose of the recall is to define how much yes is predicted by the prediction model when it is actually yes. It can be represented in the form of a classifier capable of predicting positive values for values that are positive. It is often referred to as 'Sensitivity' or 'Recall.' and is expressed in equation (19) as:

$$Recall = \frac{TP}{TP + FN} \quad (19)$$

6. **F1 score:** It is defined as the weighted average of the actual positive levels (recall) and consistency. F1 is a useful general measure of a model's precision that combines accuracy with recall. It has considered perfect when it is 1 or 100 percent, while the model is a complete disaster when it is 0. It is expressed in equation (17) as:

$$F1\ score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (20)$$

## RESULTS AND DISCUSSION

There were 42,316 observations in the labeled dataset, of which only 427, or around 1 percent, were labeled as OHD. The OHD class is, therefore, very unusual. This imbalance between the classes of OHD and No-OHD renders the issue of classification very difficult. The labeled data set was divided 80-20 using stratified sampling to counter this disproportion. To construct a balanced learning collection, the partition containing 80 percent of the data was downsampled. The partition was used as the test set with 20 percent data. This downsampling process generates a balanced learning set, but the test set stays imbalanced. On the balanced learning set, a random forest classifier was refined and tested on the test set. This method of train-testing was repeated ten times. The classifier's precision, accuracy, sensitivity, specificity, misclassification, recall, and F1-score were computed from the confusion matrix for each execution. Across all runs, the average metrics were computed. The python

**Table 3. Results of the NSDUH dataset using various machine-learning algorithms**

SNO	Assessment Parameter	Decision Trees	Bayesian Classifier	Random forest	Improved Decision Tree
1	Accuracy	98.74	97.95	99.04	99.21
2	Precision	63.75	49.35	68.14	75.58
3	Specificity	99.43	98.97	99.53	99.67
4	Misclassification rate	1.26	2.05	0.96	0.79
5	Recall	58.20	48.65	66.19	68.01
6	F1 Score	60.85	49.00	67.15	71.60

implementation was performed using the following libraries and packages: pandas, NumPy, sklearn, tree, RandomForestRegressor, DecisionTreeClassifier.

The decision tree classification algorithm can predict the probability of developing OHD with an accuracy of 98.74 percent, an accuracy of 63.75 percent, a precision of 99.43 percent, a recall of 58.20, and an F1 score of 6085, and an error rate of 1.26 percent. The Bayesian classifier predicted OHD with an accuracy of 97.95 percent, an accuracy of 49.35 percent, a precision of 98.97 percent, a recall of 48.65, and an F1 score of 49, and an error rate of 2.05 percent. The random forest classifier will predict that adults are likely to grow OHD with a 99.04 percent precision, a 68.14 percent accuracy, a 99.53 percent specificity, a 66.19 recall, a 67.15 F1 score, and a 2.05 percent error rate. The enhanced decision tree classifier can predict the highest accuracy of 99.21 percent, accuracy of 75.58 percent, precision of 99.67 percent, recall of 68.01, F1 score of 71.60, and error rate 0.79 percent for adults likely to develop OHD.

The average specificity was higher than the average sensitivity in this case. The classifier would have skipped a higher number of people with OHD, which is not a desirable scenario. Whereas if the classifier's average sensitivity is greater than the specificity average. That means that it correctly recognizes many OHD individuals but at the expense of incorrectly flagging some individuals. The F1 score of more than 60 percent indicates good results for clinical psychology applications like this one. Thus, while the OHD class's prevalence is only about 1% in the imbalanced test collection, the enhanced decision reports promising results.

The improved decision tree classifier also ranks the 11 independent characteristics in its significance plot shown in Table 3 in order of their significance for predicting OHD. OHD's best

**Figure 2. Comparison of Percentage Accuracy, Precision and Specificity for Decision Tree, Bayesian Classifier, Random Forest and Proposed Improved Decision Tree**

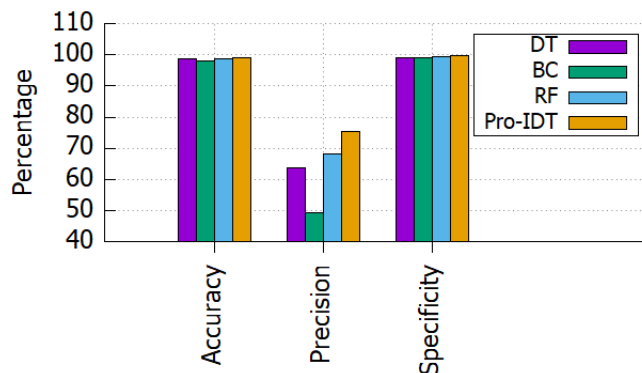


Figure 3. Comparison of Misclassification rate for Decision Tree, Bayesian Classifier, Random Forest, and Proposed Improved Decision Tree

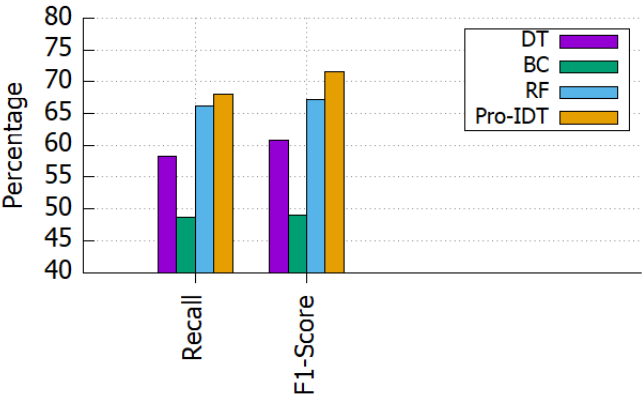
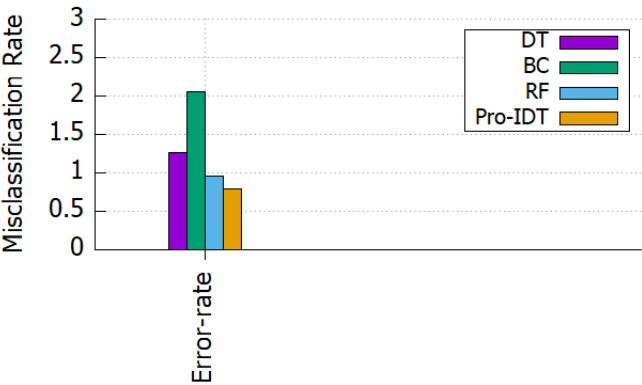


Figure 4. Comparison of Percentage Recall and F1-score for Decision Tree, Bayesian Classifier, Random Forest, and Proposed Improved Decision Tree



predictor is early initiation or' first use of marijuana before 18 years of age.' That is uncertain and unexpected since two disorders, often comorbid with OHD, rate higher than mental disorder and disability. The best preventive measure could, therefore, be to curb early marijuana initiation. That illustrates the critical role that parents, teachers, and clinicians can play in alleviating the opioid epidemic.

## CONCLUSION

This study shows how it is possible to use machine learning to predict adults at risk for OHD. In identifying the interactions between characteristics that increase this risk, machine learning is also promising. It also demonstrates that collections of public domain data such as NSDUH carry extensive data that can be exploited to enhance our knowledge of drug use disorders and mental illnesses. Such data sets contain many observations, a variety of features are gathered for each observation, data is cleaned and recoded, and surveys are carried out in a statistically sound manner. The proposed architecture is based on a decision support system to predict disease in health care diagnostic. The

essential features were identified using the ChiSquare method. We have also tested the health care diagnostic model with ID3, pro-IDT, Random forest, and SVM methods in the HCV dataset. The experimental results show that pro-IDT proves to be a promising solution for all the parameters used for comparison. The accuracy, precision, and specificity are improved a lot, whereas the misclassification rate is reduced. By adopting the pro-IDT approach, the F1 score is improved while maintaining an acceptable recall value.

We will be extending this work to envision and visualize the other possibilities of disease difficulties using a few advanced algorithms in classifications for the healthcare industry. Also, we will work on the cloud-based health care monitoring system. Furthermore, the classification model can be used to screen the expected opioid heavy users and identify the critical features leading to intensive opioid usage and overdose, so that specific intervention plans through continuous usage monitoring and consulting can be developed to reduce their dependence addiction. Finally, a rigorous exploration and proof of the soundness and superiority of the pseudo-labeling method will be investigated regarding the proposed methodology.

## **ACKNOWLEDGMENT**

The Open Access Publication fee was waived by the publisher for this article.

## REFERENCES

- Acion, L., Kelmansky, D., Laan, M. D. V., Sahker, E., Jones, D., & Arndt, S. (2017). Use of a machine learning framework to predict substance use disorder treatment success. *PLoS One*, 12(4), e0175383. doi:10.1371/journal.pone.0175383 PMID:28394905
- Ahn, W. Y., Ramesh, D., Moeller, F. G., & Vassileva, J. (2016). Utility of machine-learning approaches to identify behavioral markers for substance use disorders: Impulsivity dimensions as predictors of current cocaine dependence. *Frontiers in Psychiatry*, 7(34). Advance online publication. doi:10.3389/fpsyt.2016.00034 PMID:27014100
- Bedard, N. A., Pugely, A. J., Westermann, R. W., Duchman, K. R., Glass, N. A., & Callaghan, J. J. (2017). Opioid use after total knee arthroplasty: Trends and risk factors for prolonged use. *The Journal of Arthroplasty*, 32(8), 2390–2394. doi:10.1016/j.arth.2017.03.014 PMID:28413136
- Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease prediction by machine learning over big data from healthcare communities. *IEEE Access: Practical Innovations, Open Solutions*, 5, 8869–8879. doi:10.1109/ACCESS.2017.2694446
- Cochran, B. N., Flentje, A., Heck, N. C., Van Den Bos, J., Perlman, D., Torres, J., Valuck, R., & Carter, J. (2014). Factors predicting development of opioid use disorders among individuals who receive an initial opioid prescription: Mathematical modeling using a database of commercially-insured individuals. *Drug and Alcohol Dependence*, 138, 202–208. doi:10.1016/j.drugalcdep.2014.02.701 PMID:24679839
- Dataset. (2019). *National Survey on Drug Use and Health (NSDUH)*. Retrieved October 10, 2020, from <https://www.datafiles.samhsa.gov/study-series/national-survey-drug-use-and-health-nsduh-nid13517>
- De Mántaras, R. L. (1991). A distance-based attribute selection measure for decision tree induction. *Machine Learning*, 6(1), 81–92. doi:10.1023/A:1022694001379
- Ellis, R. J., Wang, Z., Genes, N., & Ma'ayan, A. (2019). Predicting opioid dependence from electronic health records with machine learning. *BioData Mining*, 12(1), 1–19. doi:10.1186/s13040-019-0193-0 PMID:30728857
- Goesling, J., Moser, S. E., Zaidi, B., Hassett, A. L., Hilliard, P., Hallstrom, B., Clauw, D. J., & Brummett, C. M. (2016). Trends and predictors of opioid use following total knee and total hip arthroplasty. *Pain*, 157(6), 1259–1265. doi:10.1097/j.pain.0000000000000516 PMID:26871536
- Green, C. A., Perrin, N. A., Hazlehurst, B., Janoff, S. L., DeVeugh-Geiss, A., Carrell, D. S., Grijalva, C. G., Liang, C., Enger, C. L., & Coplan, P. M. (2019). Identifying and classifying opioid-related overdoses: A validation study. *Pharmacoepidemiology and Drug Safety*, 28(8), 1127–1137. doi:10.1002/pds.4772 PMID:31020755
- Hemlata, C., & Gulia, P. (2016). Big Data Analytics. *Research Journal of Computer and Information Technology Sciences*, 4(2), 1–4.
- Jain, K., & Kumar, A. (2020). An energy-efficient prediction model for data aggregation in sensor network. *Journal of Ambient Intelligence and Humanized Computing*, 11(11), 5205–5216. doi:10.1007/s12652-020-01833-2
- Leung, C. S., Wong, K. W., Sum, P. F., & Chan, L. W. (2001). A pruning method for the recursive least squared algorithm. *Neural Networks*, 14(2), 147–174. doi:10.1016/S0893-6080(00)00093-9 PMID:11316231
- Leurent, B., Gomes, M., Faria, R., Morris, S., Grieve, R., & Carpenter, J. R. (2018). Sensitivity analysis for not-at-random missing data in trial-based cost-effectiveness analysis: A tutorial. *PharmacoEconomics*, 36(8), 889–901. doi:10.1007/s40273-018-0650-5 PMID:29679317
- Li, X., Chaovalitwongse, W. A., Curran, G., Tilford, J. M., Felix, H., & Martin, B. C. (2018). Using machine learning to predict opioid overdoses among prescription opioid users. *Value in Health*, 21, S24. doi:10.1016/j.jval.2018.04.151
- Martel, M. O., Wasan, A. D., Jamison, R. N., & Edwards, R. R. (2013). Catastrophic thinking and increased risk for prescription opioid misuse in patients with chronic pain. *Drug and Alcohol Dependence*, 132(1-2), 335–341. doi:10.1016/j.drugalcdep.2013.02.034 PMID:23618767
- McAnally, H. B. (2018). Opioid dependence risk factors and risk assessment. In *Opioid Dependence* (pp. 233–264). Springer. doi:10.1007/978-3-319-47497-7\_10

- Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, 6(1), 1–10. doi:10.1038/srep26094 PMID:27185194
- National Institute on Drug Abuse. (2020, May). *The Science of Drug Use: A Resource for the Justice Sector*. Retrieved October 15, 2020, from <https://www.drugabuse.gov/drug-topics/criminal-justice/science-drug-use-resource-justice-sector>
- National Rehabilitation Information Center. (2011, January). *Substance Abuse and Individuals with Disabilities*. Retrieved October 15, 2020, from <https://naric.com/?q=en/publications/volume-6-number-1-january-2011-substance-abuse-individuals-disabilities>
- Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81–106. doi:10.1007/BF00116251
- Rozell, J. C., Courtney, P. M., Dattilo, J. R., Wu, C. H., & Lee, G. C. (2017). Preoperative opiate use independently predicts narcotic consumption and complications after total joint arthroplasty. *The Journal of Arthroplasty*, 32(9), 2658–2662. doi:10.1016/j.arth.2017.04.002 PMID:28478186
- Saffer, H., & Dave, D. (2005). Mental illness and the demand for alcohol, cocaine, and cigarettes. *Economic Inquiry*, 43(2), 229–246. doi:10.1093/ei/cbi016
- Wadekar, A. (2019, July). Predicting Opioid Use Disorder (OUD) Using A Random Forest. In *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC) (Vol. 1, pp. 960-961)*. IEEE.

## APPENDIX

National Survey on Drug Use and Health (NSDUH) The National Survey on Drug Use and Health (NSDUH) series, formerly known as the National Household Survey on Drug Abuse, is a major source of longitudinal data on the use of illegal drugs, alcohol, and tobacco, as well as mental health problems, among US civilians aged 12 and up. The survey examines psychiatric and/or drug use disorders, as well as treatment for these disorders, to monitor patterns in particular substance use and mental illness interventions and determine the effects of these conditions. Identification of groups at high risk for drug use initiation and concerns among people with co-occurring substance use disorders and mental illness are two examples of how the NSDUH data can be used. Downloadable public-use data files from the NSDUH are available in SAS, SPSS, STATA, and ASCII formats, as well as online analysis with SDA. The R-DAS can be used to analyze NSDUH restricted-use data files online. The National Survey on Drug Use and Health is funded by the Substance Abuse and Mental Health Services Administration's Center for Behavioral Health Statistics and Quality (formerly Office of Applied Studies). For more information, visit the NSDUH website. (Data Source: <https://www.datafiles.samhsa.gov/>).

*Khushboo Jain is currently working as an Assistant Professor in DIT, University, Dehradun. She is engaged in the teaching profession for the last 7 years. She is also pursuing her Ph.D. in Computer Science Engineering in the field of "Wireless Sensor Networks" from Banasthali Vidyapith. She received her M.Tech degree from Banasthali Vidyapith and B.Tech degree from UPTU Lucknow, India. Her research interests include Sensor Networks, Machine Learning, Data Mining, Data Prediction & Data Aggregation. She has published has more than 15 research papers in International journals and conferences indexed in SCI and Scopus.*

*Akansha Singh is working as an Assistant Professor in the Department of Mathematics, NIET, Greater Noida. She has published many research papers in international journal/conference. She is Editorial Board Member and Reviewer of many International Journals. Her research interests include Cryptography, Wireless sensor networks and Network security.*

*Poonam Singh is currently working as an Assistant Professor in Hindustan College of Science & Technology,*

*Sanjana Yadav is currently working as an Assistant Professor in Hindustan College of Science & Tech, Mathura. She is engaged in teaching profession for more than 8 years. She received her M.Tech degree from MODY university, Lakshmangarh, Rajasthan and B.Tech degree from UPTU Lucknow, India. Her research interests includes Image Processing, Data Analytics, Sensor networks and Machine Learning. She has published more than 5 research papers in reputed journals and conferences.*