Machine Learning-Based Prediction of Cannabis Addiction Using Cognitive Performance and Sleep Quality Evaluations

Abdelilah Elhachimi[1], Mohamed Eddabbah[2], Abdelhafid Benksim[3], Hamid Ibanni[4], Mohamed Cherkaoui[5]*
Department of Biology, University Cadi Ayyad Marrakech (UCAM), Marrakech, Morocco[1, 5]
The Higher School of Technology of Essaouira (ESTE) Cadi Ayyad University, Morocco[2]
Institute of Nursing Professions and Healthcare Techniques (ISPITS), Marrakech, Morocco[3]
National Association of Drug-Risk Reduction (RdR-Maroc), Marrakech, Morocco[4]

***Abstract*—Cannabis addiction remains a growing public health concern, particularly due to its impact on cognition and sleep quality. Conventional screening tools, such as structured interviews and self-assessments, often lack objectivity and sensitivity. This study aims to develop and compare machine learning (ML) models for the prediction of cannabis addiction using cognitive performance (Montreal Cognitive Assessment – MoCA) and sleep quality (Pittsburgh Sleep Quality Index – PSQI) features. A total of 200 participants aged 13 to 24 were assessed, including 103 diagnosed addicts and 97 controls. Principal Component Analysis (PCA) was used to reduce data dimensionality and enhance model robustness. The study evaluated six supervised machine learning algorithms, namely Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Multilayer Perceptron (MLP). Results showed that LR and MLP models achieved high sensitivity (85.71%) and specificity (100%) on the test set, outperforming the DSM-5-based CUD reference test (sensitivity = 71.43%). Although the RF and XGBoost models achieved perfect classification on the training set, their reduced performance on the test set indicates a potential overfitting issue. Integrating machine learning with validated psychometric assessments enables a more accurate and objective identification of cannabis addiction at early stages, thus supporting timely interventions and more effective prevention strategies.**

***Keywords*—Cannabis addiction; machine learning; cognitive assessment; sleep quality; predictive modeling**

I. Introduction

In recent decades, cannabis use has increased sharply, making it one of the most popular psychoactive substances worldwide. In this context, the United Nations Office on Drugs and Crime (UNODC) released a report in 2022 that outlined drug use trends from 2010 to 2020 [1]. According to this document, the number of individuals aged 15 to 64 who used a psychoactive substance in a given year remained relatively constant compared to previous years, with approximately 284 million people affected, or approximately 5.6% of the global population. Of these users, nearly 13.6%, or 38.6 million people, suffer from a drug use disorder. The report specifies that of these users, 13.6% of them, representing approximately 38.6 million individuals, have a drug use disorder. The report also indicates that cannabis was the most common psychoactive substance, with 209 million users, a significant increase of 23% compared to 2010. This consumption varies significantly by region. A systematic study revealed that the prevalence ranges from 0.42% to 43.90% in Europe, 1.40% to 38.12% in North and South America, 0.30% to 19.10% in Asia, and 1.30% to 48.70% in Oceania and Africa [2]. Numerous studies worldwide have extensively explored the cognitive, psychiatric, physical, and socioeconomic effects associated with cannabis use [3,4]. In this sense, this research has focused on cognitive deficits linked to cannabis use, including effects on executive functions, memory, attention, and decision-making abilities. The cognitive disorders are of greater concern due to their significant impact on the social, academic, and professional lives of those affected, particularly young adults whose cognitive abilities are still developing [5]. Furthermore, many studies highlight that sleep disturbances are generally associated with cannabis addiction. These disturbances concern the duration, quality, and efficiency of sleep [6], [7]. Sleep disturbances have been shown to be a predictive and aggravating factor in cannabis addiction [8], [9]. The sleep disturbance is measured using standardized instruments such as the Pittsburgh Sleep Quality Index (PSQI) [10].

Given these findings, cannabis addiction is often underdiagnosed due to the lack of objective, specific, and easily usable assessment tools for healthcare professionals [11]. Currently, screening for cannabis addiction relies primarily on structured or semi-structured clinical interviews and patient self-assessment [12]. These methods generate a significant amount of subjectivity and are vulnerable to various response biases [13]. Therefore, given these methodological constraints, it would be preferable to use innovative approaches that reliably and objectively identify predictive signs of addiction at their earliest stages. In this sense, artificial intelligence (AI), and more specifically predictive models-based ML, appear to be a promising solution to overcome these limitations. Machine learning offers major advantages, such as the ability to simultaneously analyze a large number of complex variables, identify subtle patterns in clinical and psychometric data, and produce accurate, reproducible predictions that can be generalized to new populations [14]. Indeed, these models are widely used in various fields of health, in diagnostic prediction, risk stratification, and therapeutic personalization [15].

Furthermore, the use of biomarkers such as cognitive status and sleep quality as features in an ML model could be a particularly interesting approach to predict cannabis addiction. Standardized cognitive tests such as the MoCA can offer a rapid, reliable and sensitive measure of global cognitive functions, allowing early detection of alterations associated with regular cannabis use [16]. Similarly, the objective assessment of sleep quality through the PSQI index provides a precise vision of the sleep disturbances often reported by cannabis users, in particular difficulties falling asleep, frequent night awakenings and poor subjective sleep quality [17]. These disorders could constitute interesting markers for early detection of cannabis.

The main objective of this article is to develop, validate, and compare several machine learning approaches to effectively predict cannabis dependence by combining in-depth cognitive assessments with objective measures of sleep quality. Furthermore, the development of a machine learning-based predictive model could provide a clinical decision-making tool for healthcare professionals. These predictive models could facilitate early screening, rapid intervention, and individualized therapeutic management.

This work constitutes an approach aimed at modernizing clinical practices in the field of addiction medicine. It also aims to strengthen primary and secondary prevention programs in the field of addiction medicine. Furthermore, this research can fill significant gaps in the current scientific literature and significantly contribute to improving knowledge regarding the complex interactions between cannabis use, cognitive functioning, and sleep quality. Furthermore, the use of ML in the field of addiction could contribute to the improvement and development of broader preventive strategies based on evidence.

To facilitate understanding and readability, the paper is structured as follows: the next section provides a detailed review of relevant scientific literature, highlighting existing screening methods and the growing interest in machine learning-based approaches for addiction prediction. Subsequently, we describe the methodology, including data collection procedures, evaluation instruments, as well as statistical and machine learning techniques employed. The results section presents a comparative analysis of the performances achieved by different machine learning models, while the discussion section provides an in-depth interpretation of these findings, placing them in context with previous research and emphasizing their clinical implications. Finally, the conclusion summarizes the primary contributions of this study, acknowledges its limitations, and outlines avenues for future research.

II. Literature Review

Screening cannabis addiction represents a significant public health challenge, particularly given its growing global prevalence. Traditional diagnostic methods have shown limitations, prompting the exploration of advanced approaches such as ML. Over recent years, ML has become transformative in addiction research, enabling detailed analyses and predictions of addictive behaviors and treatment outcomes. For example, Likhith et al. [18], successfully applied ML algorithms to predict smartphone addiction by analyzing behavioral indicators like app usage patterns, notification-checking frequency, and psychological factors such as stress and anxiety. Similarly, Kumara et al. [19], utilized ML models, including Random Forest (RF) and Convolutional Neural Networks (CNN), to accurately identify addiction risk factors, addiction types, and relapse probabilities, significantly enhancing prevention and treatment efficacy.

In clinical addiction research, diverse data sources have been leveraged to build predictive models that enhance risk assessment and treatment outcomes. Feng et al. [20], utilized neuroimaging data derived from functional Magnetic Resonance Imaging (fMRI) to predict symptoms of internet addiction, identifying distinct connectivity patterns in brain networks associated with addiction-related behaviors. Likewise, Pyzowski et al. [21], demonstrated that electronic medical records and prescription drug monitoring data could effectively be integrated into machine learning frameworks to predict opioid addiction risk, highlighting factors such as psychiatric history and socioeconomic background. The same authors also demonstrated the utility of clinical laboratory data, including patient adherence and buprenorphine treatment outcomes, in predicting short-term relapse, thereby providing actionable insights for clinicians.

Additionally, recent studies explored innovative applications of ML models using alternative data sources, such as social media and self-reported surveys. Yang et al. [22], applied sentiment analysis using advanced Generative Adversarial Networks (GANs) on social media data, successfully predicting opioid relapse by identifying emotional triggers such as negativity or anxiety. Similarly, surveys capturing demographic data, phone usage behaviors, and psychological measures (e.g., stress or anxiety) have effectively predicted behavioral addictions, enabling early intervention strategies [18].

Furthermore, ML techniques have shown promise in distinguishing behavioral addictions like Internet gaming disorder from substance use disorders. Lee et al. [23], demonstrated the effectiveness of multimodal approaches combining neuropsychological assessments with Electroencephalography (EEG) features, achieving accuracy rates exceeding 70%. Such data-driven computational methods offer valuable insights into neural mechanisms underlying addictive behaviors, potentially informing targeted therapeutic interventions [24].

Despite these advancements, significant challenges remain. Bouhadja and Bouramoul [25], highlighted the critical role of data quality and diversity, emphasizing the need for robust, structured datasets to enhance predictive reliability. Unfortunately, the scarcity of comprehensive datasets remains a major hurdle in addiction research, often leading to overfitting and limited generalizability of ML models [26]. Moreover, methodological inconsistencies across studies, such as variability in study design and analysis methods, continue to pose barriers to reproducibility and validation of ML-derived predictions [27].

Another critical limitation identified in existing literature is the interpretability of ML models. Suva and Bhatia [28], noted that many ML algorithms function as "black boxes," making it difficult for clinicians to trust and implement their recommendations in real-world settings. In addition, ethical issues surrounding the use of sensitive personal data, including

risks of privacy breaches or misuse, remain prominent concerns in deploying these technologies clinically [29]. Finally, Bouhadja and Bouramoul [25], observed that models trained on specific populations may not generalize effectively to diverse clinical contexts, limiting their overall applicability.

Addressing these challenges requires concerted efforts to standardize methodologies, enhance data collection and preprocessing techniques, and improve model transparency. Integrating objective, clinically validated biomarkers such as cognitive performance and sleep quality indicators into predictive ML models may further improve accuracy and clinical utility. Studies by Ewert et al. [16], and Edwards and Filbey [10], have already emphasized cognitive impairment and sleep disturbances as reliable, objective markers associated with cannabis misuse, underscoring their potential value in predictive modeling for addiction.

In summary, advancing ML methodologies by addressing existing limitations can significantly enhance cannabis addiction screening, supporting early intervention and personalized treatment strategies. The current study aims precisely to contribute to these advancements, by proposing a robust ML-based predictive framework utilizing validated psychometric tools.

III. Methodology

A. Population Study

A population of 200 participants from both genders, aged between 13 to 24 years old, and who agreed to participate in this study. The sample of the study comprises two groups: 1) 103 participants clinically identified as cannabis addict; 2) 97 control patients without cannabis addiction.

The exclusion criteria included 1) patient refusing to participate in the study, 2) patient with serious behavioral issues who were unable to respond to the questions, 3) patients who, in addition to cannabis, were addicted to other psychoactive drugs.

This study was conducted at the primary health center for addiction in Marrakech. The study was authorized by the regional health directorate. The procedures were conducted in accordance with the guidelines of the Declaration of Helsinki. All participants were informed prior to data collection about the purpose of the study. Every participant and/or their legal representative has given their written informed consent.

B. Data Collection

Following a consultation with the psychiatrist-addictologist of Marrakech, an anonymous questionnaire is used to collect data aimed at obtaining information on the sociodemographic, clinical, cognitive and sleep quality characteristics of the study population.

1) The Montreal Cognitive Assessment (MoCA) test is a clinical neuropsychological test used to assess cognitive impairment. It consists of assessments of executive, visuo-spatial, denominative, memory, attention, language, abstraction, recall and orientation functions (Table I). The highest possible score is 30 points. When the score does not exceed the threshold of 26, the patient is identified as having a cognitive impairment [30].

TABLE I. Detailed Composition and Scores Associated with the MoCA Cognitive Assessment Test

Attribute Name	Component	Points Assigned
MoCA1	Visuospatial / Executive	5 points
MoCA2	Naming (Denomination)	3 points
MoCA3	Attention	6 points
MoCA4	Language	3 points
MoCA5	Abstraction	2 points
MoCA6	Memory	5 points
MoCA7	Orientation	6 points
MoCA	Total Score MoCA	30 points

2) The Pittsburgh Sleep Quality Index (PSQI) is a test used to assess sleep quality**.** Consisting of 11 questions with a maximum score of 21, aimed at quantifying sleep efficiency and quality. Each item is rated from 0 to 3, and the sum of the scores from the seven components constitutes the global PSQI score, which ranges from 0 to 21 (Table II). A global PSQI score above 5 reflects poor sleep quality [31].

TABLE II. Components and Scoring System of the Pittsburgh Sleep Quality Index (PSQI)

Attribute Name	Component	Description	Score (0-3)
PSQIC1	Subjective Sleep Quality	Overall perception of sleep quality	0 = Very good 3 = Very poor
PSQIC2	Sleep Latency	Time taken to fall asleep	0 = <15 min 3 = >60 min
PSQIC3	Sleep Duration	Total sleep hours per night	0 = >7h 3 = <5h
PSQIC4	Sleep Efficiency	Ratio of sleep time to time in bed	0 = >85% 3 = <65%
PSQIC5	Sleep Disturbances	Frequency of sleep interruptions and issues	0 = None 3 = Daily
PSQIC6	Use of Sleep Medication	Frequency of sleeping pill use	0 = Never 3 = Daily
PSQIC7	Daytime Dysfunction	Impact of sleep deprivation on daily activities	0 = None 3 = Severe

3) Cannabis Use Disorder (CUD) is a set of guidelines issued by the American Psychiatric Association to characterize problematic cannabis use in cognitive-behavioral, psychological and environmental terms. A score below 2 indicates no addiction, a score between 2 and 3 indicates mild addiction, a score between 4 and 5 indicates moderate addiction, and a score above 6 indicates severe addiction [32].

Participants with a cannabis addiction are identified and diagnosed using this DSM-5 questionnaire. Participants are divided into two classes: 0: Non-Addict and 1: Addict. All addicts, regardless of the intensity of their addiction, are included in the addict class. The gold standard for comparing the acquired models is the CUD questionnaire.

Analysis of the addictive profile according to DSM-5 criteria in our sample reveals a marked split between non-addicts and those with varying degrees of cannabis addiction. Among the participants, 46% (92 individuals) showed no signs of addiction, while 54% had some form of dependence: 18% suffered from mild addiction (36 individuals), 28% from moderate addiction (56 individuals) and 8% from severe addiction (16 individuals). These results indicate a significant prevalence of problematic cannabis use in our study population. Statistical analysis using the Chi-square test ($\chi^2 = 182.762$, $p < 0.001$) revealed a highly significant association between cannabis use and the development of a DSM-5 addictive disorder. The p-value of less than 0.001 confirms that this relationship is not due to chance, and suggests a direct link between consumption and dependence.

Looking at the distribution of addiction levels, it appears that all cases of moderate and severe addiction exclusively concern cannabis users. What's more, only 5.2% of individuals with mild addiction are classified as non-users, confirming that cannabis addiction is virtually non-existent among non-users; these results underline the high risk of dependence among users, with a high proportion of moderate to severe cases (36%), suggesting that regular use can lead to worsening addiction.

B. Feature Engineering

The input features consisted of 14 standardized sub-scores: seven from the MoCA (MoCAC1–MoCAC7) and seven from the PSQI (PSQIC1–PSQIC7), capturing key aspects of cognitive performance and sleep quality. All variables were normalized using z-score standardization to ensure comparability and facilitate model convergence.

PCA was applied exclusively to the training data to prevent information leakage. The first three components, which accounted for approximately 53% of the total variance, were retained for comparison purposes. Both original and PCA-transformed feature sets were fed into identical machine learning pipelines, allowing a controlled evaluation of their impact on classification performance.

C. Machine Learning Models

ML is a tool that allows a machine to acquire knowledge, build models, and analyze complex datasets without direct human intervention [33]. In this sense, ML has been widely used recently in many biomedical fields, including psychiatry and addiction [34], [35]. In general, the types of ML algorithms used in addiction research can be grouped as follows: supervised learning, unsupervised learning, deep learning (DL), and reinforcement learning (RL) [36].

In this study, six supervised machine learning algorithms were trained to predict cannabis addiction using cognitive and sleep quality features. All models were evaluated under two conditions: using the original standardized feature set and using the PCA-transformed data. This dual evaluation allowed us to assess the effect of dimensionality reduction on model performance.

1) Logistic regression: Logistic regression (LR) is a very popular supervised ML model. It is used to predict a categorical dependent variable from a set of explanatory variables [37]. LR calculates the probability that an observation belongs to a particular class [38]. In this paper, the LR model is configured without intercepts and with a linear solver "liblinear".

2) K-Nearest neighbors: The K-Nearest Neighbors (KNN) algorithm consists of calculating the distance between an unknown data point and its "k" nearest neighbors already classified. Subsequently, the label of the nearest class is assigned to the observation. The performance of the algorithm depends on the number of neighbors selected (k=1, 2, 3, ...), as well as the chosen distance (Euclidean, Manhattan, etc.)[39]. In this study, the optimization of the model focused on the choice of the number of neighbors (19, 21 and 23) to ensure robust decision-making, as well as on uniform weighting to simplify interpretation. The Manhattan distance was chosen as the metric because it is suitable for heterogeneous data and less sensitive to scale variations [40].

3) Support vector machine: The Support Vector Machine (SVM) is a robust linear classifier capable of distinguishing different classes from input data. Although there are infinitely many linear separators for classification problems, the SVM chooses an optimal one, ensuring maximum spacing between classes [41]. We used the Support Vector Machine (SVM) algorithm to classify addiction risk by optimizing its hyperparameters. The penalty C (0.1, 1, 10) was adapted to balance the separation margin and classification errors. Two kernels were tested: linear, suitable for separable data, and RBF, allowing to model complex relationships.

4) Random forest: Random forest (RF) is a series of decision trees built from a randomly selected subset of the training data. Each tree is built from a distinct portion of the data and participates in the final decision through a majority vote [42]. The model was parameterized with a number of trees varying between 50, 75 and 100, a maximum depth of 3, and feature selection based on the square root of the total number of explanatory variables (max_features = 'sqrt'). In addition, the minimum node split criterion was tested with 25 and 30 observations.

5) XGBoost: XGBoost is an improved model of the gradient boosting algorithm. In ML, extreme gradient boosting is a method that is used to reduce the number of errors in predictive data analysis [43], [44]. XGBoost is an assembly of decision trees that predict residuals and correct the errors of previous decision trees [45]. The particularity of this algorithm is the improvement in accuracy and execution speed. It uses advanced techniques like L1 and L2 regularization, subsampling, and missing value handling to improve its performance [44]. In this paper, the XGBoost model is configured with a number of estimators varying between 50, 75, and 100, a maximum depth of 2, a learning rate from 0.005 to 0.01, and subsampling of features and observations varying between 0.6 and 0.8 to reduce overfitting.

6) MLP Neural network: The MLP Neural Network model is an architecture based on the use of multilayer neural networks (Multilayer Perceptron) capable of modeling complex functions thanks to a hierarchy of interconnected hidden layers[46]. In this paper, the MLP network architecture was intentionally simplified to reduce its complexity and limit

overfitting. Three hidden layer sizes (10, 15, and 20 neurons) were tested. Two activation functions (relu and tanh) were evaluated, as well as the adam optimization algorithm for its speed and stability. Regularization was adjusted via the alpha parameter with three values (0.1, 0.5, 1.0) to control the model complexity. Finally, an initial learning rate of 0.001 was used to ensure stable convergence.

D. Experimental Design and Evaluation Strategy

The dataset was split into a training set (80%) and an independent test set (20%) using stratified sampling to preserve class distribution. All data preprocessing steps, including z-score standardization and dimensionality reduction via PCA, were strictly performed after the train-test split, and fitted exclusively on the training set to prevent data leakage. The PCA was applied on the correlation matrix of the training data, and the first three components were retained, accounting for approximately 53% of the total variance. This choice aimed to reduce dimensionality while preserving relevant variance and enabling visual analysis. PCA-transformed data were used for comparison with the full-feature models.

In addition to this train–test evaluation, we implemented a 5-fold stratified cross-validation on the training data to assess model robustness and generalization. Performance metrics accuracy, sensitivity, specificity, and AUC were computed on each fold and averaged. Standard deviations were also calculated to evaluate metric stability across folds.

Final validation was conducted on the unseen 20% test set. Although confusion matrices were generated for Random Forest and XGBoost for visualization purposes, the model evaluation relied exclusively on three key indicators: sensitivity, specificity, and precision. These metrics were selected for their clinical interpretability and direct relevance to addiction screening, and were compared to the reference diagnostic criteria of the CUD-DSM5, used as the gold standard. Additionally, ROC curves and AUC values were computed to assess the global discriminatory power of each classifier.

All experiments were conducted using Python 3.10, with scikit-learn (v1.2), XGBoost (v1.7), pandas (v1.5), and numpy (v1.23). A random seed (42) was fixed for reproducibility. Where applicable, hyperparameter tuning was performed using grid search, and class imbalance was handled using class_weight='balanced'.

1) ROC curve: The ROC curve is a graphical representation that shows the sensitivity and specificity for all possible classification threshold values. It is a visual device that helps to establish a balance between true positives and false negatives [47]. The closer the curve is to the upper left corner implies a better quality of the model.

2) AUC metric: AUC is a numerical indicator obtained based on the ROC curve. It illustrates the chance that the model produces the correct prediction based on a specific threshold and chosen attributes [48]. The closer it is to 1, the better the model will perform.

3) Test validity: Sensitivity and specificity are two statistical criteria used to assess a test's validity [49], [50]. The ratio of actual diseased patients to all diseased patients is known as the sensitivity [51]. The test's sensitivity shows how well it can identify patients who are ill [49]. Conversely, specificity is defined as the proportion of real healthy patients who are not recognised among all healthy patients[51]. The test's specificity shows how well it can rule out healthy people [49]. Another metric, such as accuracy, which is the proportion of all correct hits among all participants, can also be used to solidify the validity [49].

IV. Results

Exploratory Data Analysis (EDA) was organized in two parts: on the one hand, the evaluation of continuous variables such as age, Age of first Cannabis Use (AFCU), Cannabis Use Duration(CCD), MoCA, and PSQI scores, using the Kolmogorov-Smirnov (KS) test to test normality and the Mann-Whitney U test to compare distributions between addicted and non-addicted groups; On the other hand, the analysis focused on the cognitive components of the MoCA and the sleep quality parameters of the PSQI in order to highlight the statistical differences and their relevance in determining the addictive disorder. This methodology makes it possible to identify the most discriminating cognitive and behavioral markers, thus contributing to a better understanding and modeling of the risk of addiction. The Kolmogorov-Smirnov (KS) test was used to assess the normality of variable distributions in the addict and non-addict groups. A p-value < 0.05 indicates a significant deviation from a normal distribution, meaning that the variable does not follow a normal distribution (Table III).

The results in Table III shows that most variables are significantly different ($p < 0.01$) between the groups, indicating notable impacts of cannabis addiction on cognition and sleep quality. Indeed, scores on the various MoCA components were significantly lower in addicts, except for the abstraction component, which was not significant ($p = 0.054$). Furthermore, the total MoCA score is highly significant ($p < 0.01$), implying an overall impairment of cognitive functions in addicts. Regarding the PSQI, most PSQI components showed significant differences ($p < 0.01$), suggesting impaired sleep quality in addicts. Furthermore, the Sleep Duration component ($p = 0.182$) was not significant, indicating that sleep duration did not differ significantly between groups.

These results reinforce the idea that cannabis addiction negatively impacts cognitive function and sleep quality, although some aspects (abstraction and sleep duration) appear to be less affected. In this sense, PCA was therefore used to reduce the dimensionality of the dataset while retaining essential information from the MoCA and PSQI subcomponents. Applying PCA allows these subcomponents to be transformed into a reduced number of non-redundant variables, while maximizing the explained variance. The dimensionality reduction will allow the elimination of highly correlated variables to avoid information redundancy, the extraction of the main axes underlying cognitive deficits and sleep disorders in addicts, and the improvement of the efficiency of machine learning models by reducing the risk of overfitting.

TABLE III. COMPARATIVE ANALYSIS OF COGNITIVE AND SLEEP PARAMETERS BETWEEN ADDICTS AND NON-ADDICTS USING THE MOCA AND PSQI TEST

	Non-addict			Addict					
	Means ±SD	**KS**	**p-value**	**Means±SD**	**KS**	**p-value**	**Mann-Whitney U**	**Z**	**p-value**
Age	20.08±2.11	0.21	<0.01	21.23±2.84	0.238	<0.01	6323.5	3.272	0.001
AFCU	2.41±5.48	0.505	<0.01	15.75±2.46	0.223	<0.01	9352.5	11.068	<0.01
CCD	00	00	00	4±2.2	0.204	<0.01	9991.0	13.010	<0.01
TEST MoCA									
Visuospatial/ executive	4.55±0.48	0.383	<0.01	4,05±0.84	0.22	<0.01	3310.5	-4.490	<0.01
Naming	3±0.00			2.88±0.32	0,52	<0.01	4413.5	-3.459	<0.01
Attention	4,23±0,62	0,31	<0.01	3,79±0,97	0,24	<0.01	3684.0	-3.485	<0.01
Language	2,95±0.22	0,54	<0.01	2,5±0,56	0,34	<0.01	2820.5	-6.856	<0.01
Abstraction	1,33±0,56	0,43	<0.01	1,17±0,55	0,37	<0.01	4348.5	-1.925	0.054
Memory	4,40±0,55	0,34	<0.01	3,46±0,86	0.30	<0.01	2041.0	-7.895	<0.01
Orientation	5,95±0.2	0,54	<0.01	5,72±0,45	0,46	<0.01	3843.5	-4.435	<0.01
MoCA	26,40±1,06	0,22	<0.01	23,58±2,74	0,18	<0.01	1445.5	-8.805	<0.01
TEST PSQI									
Subjective sleep quality	1,10 ±0,74	0,29	<0.01	1,40 ±0,67	0,31	<0.01	6203.0	3.213	0.001
Sleep latency	1,41±0,79	0,31	<0.01	2,17±0,70	0,26	<0.01	7522.5	6.553	<0.01
Sleep duration	1,05±0,72	0,32	<0.01	0,80±0,60	0,36	<0.01	4537.0	-1.336	0.18
Habitual sleep efficiency	0,50±0.87	0,41	<0.01	0,96±1.1	0,24	<0.01	6736.5	4.651	<0.01
Sleep disturbances	1,25±0,61	0,38	<0.01	1.6 ±0,61	0,30	<0.01	6463.5	4.038	<0.01
Use of sleeping medication	0,14 ± 0.43	0.50	<0.01	1,32±1.1	0.20	<0.01	8105.5	8.527	<0.01
Daytime dysfunction and sleepiness	0,90±0.85	0.24	<0.01	1,46 ± 0.8	0.24	<0.01	6726.5	4.468	<0.01
PSQI	6,37±2,83	0,16	<0.01	10,02±3,57	0,12	0,001	8037.0	7.474	<0.01

M ± SD: Mean ± Standard Deviation., KS (D): Kolmogorov-Smirnov test statistic; (p < 0.05 indicates a significant deviation from normality).

The Mann-Whitney test comparing the Addict and Non-Addict groups. A p-value < 0.05 indicates a significant difference between groups; The Z value represents the standardized statistic of the Mann-Whitney U test

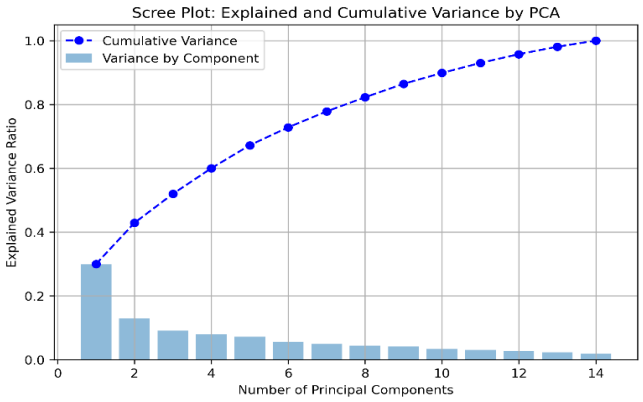

Fig. 1. Distribution of the variance explained and cumulative by the principal components from PCA.

Analysis of explained variance reveals that the first components express a large part of the variance (Fig. 1). Indeed, it is necessary to faithfully represent the data. In addition, the cumulative variance shows the progressive accumulation of explained variance. With 7 to 8 components, about 80% of the total variance is captured, indicating that a large part of the information is preserved. Beyond 10 components, the curve reaches a plateau within 100%, indicating that the last components bring little new information. The first component captures a significant part of the information (about 30%), followed by the second, which captures about 15%. The following components have a decreasing contribution, suggesting that only a few principal components are involved.

The Fig. 2 represents the projection of individuals into the space of the first three principal components resulting from the PCA. Each point corresponds to an individual, with a distinction between non-addicts (in blue) and addicts (in red). The observation shows that the two groups (addicts and non-addicts) occupy relatively distinct regions, although some areas present an overlap. Thus, the first three principal components express a significant part of the total variance, showing a separation between the groups in three dimensions. Furthermore, a more marked concentration of red and blue points in certain regions implies that the PCA has succeeded in capturing structural differences between the groups. The separation between the groups indicates that the principal components contain relevant discriminating information for the prediction of the diagnosis (addict vs. non-addict). However, the presence of an overlap states that some individuals are more difficult to classify, justifying the use of more complex models to improve accuracy. The obtained projection justifies the use of PCA in the preprocessing step for ML.

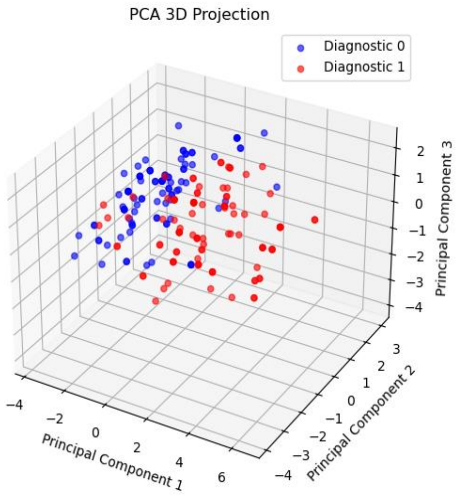

Fig. 2. Projection of individuals into the PCA component space according to addictive status.

In summary, the EDA highlighted redundancies and overlaps between some variables, justifying the application of ACP to better structure the information. In this sense, feature engineering focused on the selection of relevant indicators related to cognitive functions and sleep quality (MoCA and PSQI), with the aim of improving both the robustness and interpretability of the model. The performance of the models is assessed by exploiting the ROC curves shown in Fig. 3.

The results shows that all the tested models reveal an excellent classification capacity, with AUC values greater than 0.90. Indeed, on the test set, the best performances are obtained by the LR and the MLP neural network, both reaching an AUC of 0.94, closely followed by the RF (AUC = 0.93) and the SVM (AUC = 0.92).

These results show that the chosen variables, in particular the principal components resulting from the PCA, are congruent for the discrimination between addicted and non-addicted individuals. Regarding performance on the training set, the most complex models, such as RF (AUC = 0.97) and SVM (AUC = 0.96), have a very high learning capacity. Additionally, these two models show a more marked discrepancy with the results obtained on the test set, which may indicate overfitting. However, RL and the MLP neural network present stable and balanced results between training and testing, reflecting good generalization capacity.

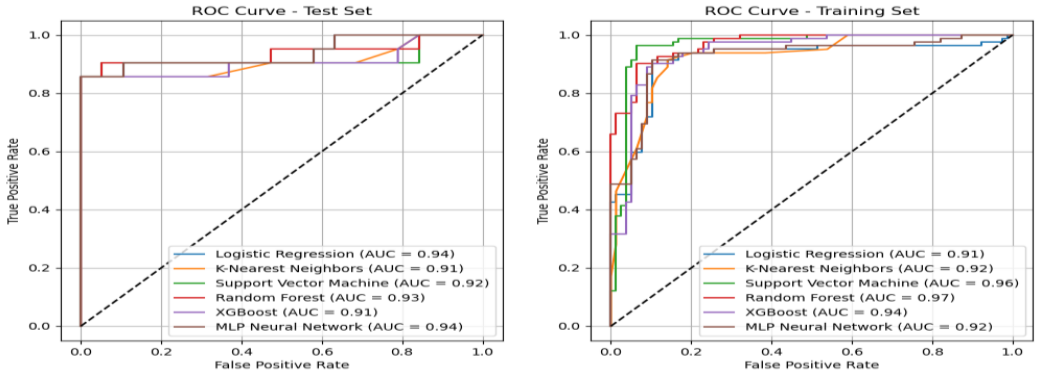

Fig. 3. ROC Curves for classification models on training and test sets for cannabis addiction prediction.

These observations indicate that all models perform well, but the most balanced approaches, such as an MLP can be favored for a robust practical implementation and reliable prediction of cannabis addiction risk.

Table IV summarizes the sensitivity, specificity, and accuracy of six machine learning models (LR, KNN, SVM, RF, XGBoost, and MLP Neural Network) compared to the clinical reference test (CUD-DSM5). Evaluations were conducted on both training and test datasets. On the test set, all models exhibited a specificity of 100%, indicating an excellent ability to correctly identify non-addicted individuals. However, sensitivity varied significantly across models, highlighting differences in their capacity to detect addicted subjects accurately.

The RF classifier demonstrated the highest performance on the test set, with a sensitivity of 90.48% and an accuracy of 95%, closely followed by LR and MLP Neural Network, each achieving sensitivities of 85.71% and accuracies of 92.5%. In contrast, the CUD-DSM5 clinical test, serving as the gold standard, had notably lower sensitivity (71.43%) and accuracy (85%), indicating a potential limitation in reliably identifying individuals with cannabis addiction.

Certain models, such as RF and XGBoost, achieved perfect performance on the training data but showed decreased performance on the test data, suggesting overfitting and thus limiting their practical applicability. Conversely, LR and the MLP Neural Network showed consistent performance between training and testing phases, highlighting their stability and suitability for real-world applications.

Overall, these results underscore that machine learning models significantly outperform the conventional CUD-DSM5 clinical screening tool in detecting cannabis addiction. LR and MLP models emerge as particularly reliable and generalizable options, combining high predictive accuracy with strong robustness.

To further examine the classification behavior of the best-performing models, confusion matrices were generated for Random Forest and XGBoost, both trained on PCA-transformed features (Fig. 4). These matrices display the number of true and false classifications for both training and test sets.

TABLE IV. COMPARATIVE PERFORMANCE METRICS (SENSITIVITY, SPECIFICITY, AND ACCURACY) OF MACHINE LEARNING MODELS VERSUS THE CUD-DSM5 REFERENCE TEST ON TRAINING AND TEST DATASETS

	Sensitivity		Specificity		Accuracy	
	Train (%)	Test (%)	Train (%)	Test (%)	Train (%)	Test (%)
LR	87.8	85.71	87.18	100.0	87.5	92.5
KNN	96.34	76.19	89.74	100.0	93.13	87.5
SVM	93.9	80.95	88.46	100.0	91.25	90.0
RF	100.0	90.48	100.0	100.0	100.0	95.0
XGBOOST	100.0	80.95	100.0	100.0	100.0	90.0
MLP	90.24	85.71	89.74	100.0	90.0	92.5
CUD	69.51	71.43	100.0	100.0	84.38	85.0

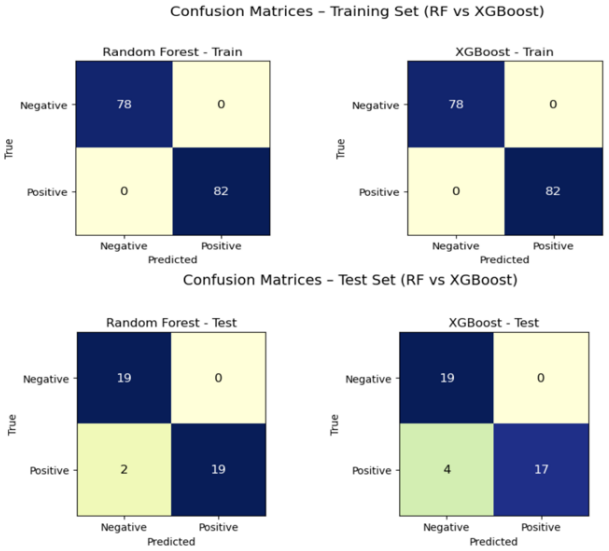

Fig. 4. Confusion matrices for training and test sets: comparison between random forest and XGBoost models.

On the training data, both models achieved perfect classification: all positive and negative cases were correctly identified, yielding 100% sensitivity and 100% specificity.

On the test set, both classifiers maintained perfect specificity (no false positives). However, some false negatives were observed: RF misclassified two positive cases (sensitivity = 90.48%), and XGBoost misclassified four (sensitivity = 80.95%). Accuracy on the test set was 95.0% for Random Forest (95% CI: 83.50% to 98.62%) and 90.0% for XGBoost (95% CI: 76.95% to 96.04%).

These results suggest that while both models effectively avoided false alarms, RF provided a slightly better recall and generalization on unseen data compared to XGBoost.

To evaluate the generalization performance of the classification models, a 5-fold cross-validation was conducted using both the original standardized features and the PCA-transformed components. Table V summarizes the average performance across folds, including accuracy, sensitivity, specificity, and AUC, each reported with their corresponding standard deviation.

Among the models trained on the raw feature set, RF achieved the best overall performance with an average accuracy of 0.938 ± 0.034, sensitivity of 0.915 ± 0.049, specificity of 0.962 ± 0.031, and an AUC of 0.985 ± 0.022. XGBoost followed closely, showing the highest sensitivity (0.927 ± 0.045) and a strong AUC (0.970 ± 0.033), suggesting its effectiveness in detecting positive cases. The MLP classifier also demonstrated competitive performance, with an AUC of 0.968 ± 0.039.

When PCA was applied prior to training, a slight decrease in performance was observed across most models. For instance, the AUC for RF (PCA) dropped to 0.967 ± 0.035, and for XGBoost (PCA) to 0.941 ± 0.030. The decline was more pronounced for models such as LR (PCA) and MLP (PCA), where sensitivity and AUC values were considerably lower compared to their raw counterparts.

These results indicate that while PCA offers a reduction in dimensionality, it does not necessarily improve classification performance. Models trained on the full standardized feature set consistently outperformed those using the reduced component space.

TABLE V. 5-FOLD CROSS-VALIDATED CLASSIFICATION PERFORMANCE OF MACHINE LEARNING MODELS WITH AND WITHOUT PCA. MEAN ± STANDARD DEVIATION (SD) ARE REPORTED FOR EACH METRIC

Models	Accuracy (±SD)	Sensitivity (±SD)	Specificity (±SD)	AUC (±SD)
RF(Raw)	0.938 ± 0.034	0.915 ± 0.049	0.962 ± 0.031	0.985 ± 0.022
XGBoost (Raw)	0.931 ± 0.050	0.927 ± 0.045	0.937 ± 0.068	0.970 ± 0.033
LR (Raw)	0.856 ± 0.058	0.842 ± 0.097	0.873 ± 0.079	0.936 ± 0.054
SVM (Raw)	0.919 ± 0.054	0.879 ± 0.098	0.962 ± 0.050	0.973 ± 0.030
KNN (Raw)	0.856 ± 0.058	0.793 ± 0.122	0.924 ± 0.047	0.949 ± 0.039
MLP (Raw)	0.925 ± 0.061	0.926 ± 0.060	0.923 ± 0.101	0.968 ± 0.039
RF(PCA)	0.912 ± 0.070	0.915 ± 0.063	0.911 ± 0.084	0.967 ± 0.035
XGBoost (PCA)	0.875 ± 0.059	0.866 ± 0.089	0.884 ± 0.064	0.941 ± 0.030
LR (PCA)	0.850 ± 0.054	0.843 ± 0.081	0.859 ± 0.061	0.916 ± 0.053
SVM (PCA)	0.887 ± 0.058	0.926 ± 0.062	0.846 ± 0.065	0.901 ± 0.043
KNN (PCA)	0.856 ± 0.070	0.852 ± 0.117	0.859 ± 0.024	0.923 ± 0.022
MLP (PCA)	0.869 ± 0.054	0.877 ± 0.089	0.859 ± 0.061	0.889 ± 0.074

V. DISCUSSION

This study demonstrates the potential of machine learning (ML) to enhance early screening for cannabis addiction by leveraging objective measures from standardized cognitive and sleep quality assessments. The integration of features from the MoCA and PSQI allowed the ML models to identify addiction-related patterns with high accuracy.

Our results reveal that LR and the MLP achieved the most balanced performance on the independent test set, with sensitivity and specificity both reaching 85.71% and 100%, respectively. These models outperformed the CUD-DSM5 gold standard, which reached only 71.43% sensitivity. This suggests that ML models can detect subtle psychometric variations associated with cannabis addiction that conventional tools may overlook.

Despite the high training accuracy observed in RF and XGBoost models (100%), their performance dropped on the test set (sensitivity = 90.48% and 80.95%, respectively), revealing overfitting. This highlights the necessity of cross-validation and model regularization, especially when working with limited sample sizes.

To ensure robustness, a 5-fold stratified cross-validation was conducted. Cross-validated performance metrics (Accuracy, Sensitivity, Specificity, and AUC) were computed for both raw and PCA-transformed datasets. Across all models, cross-validation confirmed high stability. Notably, RF trained on raw data achieved the highest AUC (0.985 ± 0.022), followed closely by XGBoost (0.970 ± 0.033) and MLP (0.968 ± 0.039).

The study also explored PCA as a dimensionality reduction strategy. Although PCA was applied strictly on the training data to prevent information leakage, its benefit on model performance was mixed. While it helped reduce collinearity and improve interpretability, models trained on raw features often outperformed those using PCA-transformed features. This can be attributed to the fact that the first three components retained only 53% of the variance, potentially omitting important information.

Our findings are consistent with prior literature that employed ML in addiction detection. For example, Lee et al. [23], used EEG and neuropsychological data for behavioral addiction classification, and Coelho et al.[12] showed moderate sensitivity for clinical tests like CUDIT-R. In contrast, our ML pipeline, using readily available psychometric tools, achieved higher classification performance and better generalization.

Importantly, our study underscores that simple, interpretable models such as LR can perform on par with, or better than, complex models, particularly when paired with appropriate feature engineering and validation strategies. This is especially relevant in clinical practice, where transparency and reproducibility are crucial for model adoption.

However, limitations must be acknowledged. The relatively small and localized sample limits external validity. Additionally, addiction status was treated as a binary label, which oversimplifies the continuum of substance use behavior. Future work should explore multi-class or severity-level prediction, and test the models in broader populations and longitudinal frameworks.

In summary, ML models trained on cognitive and sleep features offer a promising and cost-effective approach to support early detection of cannabis addiction. Their integration into clinical workflows could enhance existing screening strategies, promoting timely intervention and better patient outcomes.

VI. CONCLUSION

This study demonstrated the potential of machine learning techniques in improving the early detection of cannabis addiction using objective and validated assessment tools such as the Montreal Cognitive Assessment and the Pittsburgh Sleep Quality Index. The predictive models developed, particularly LR and MLP achieved high sensitivity and specificity, outperforming traditional clinical tools such as the DSM-5-based screening. Models trained on raw standardized features performed better than those using PCA-transformed data, indicating that dimensionality reduction was unnecessary in this context. These findings support the potential of ML enhanced screening tools to assist clinicians in the early identification of at-risk individuals based on routine assessments.

Future research should aim to validate these findings on larger and more heterogeneous populations. Incorporating complementary data such as neuroimaging, genetic markers, or behavioral tracking could enhance prediction accuracy. Longitudinal studies are also needed to evaluate the ability of these models to monitor addiction trajectories over time. Finally, embedding explainable AI mechanisms would improve clinical interpretability and foster greater trust in real-world applications.

ACKNOWLEDGMENT

The authors would like to thank the participants in this study as well as the RdR-Maroc center in Marrakech for hosting this research.

REFERENCES

[1] UNODC. Global Overview : Drug Demand. 2022.

[2] Wang Q, Qin Z, Xing X, Zhu H, Jia Z. Prevalence of Cannabis Use around the World: A Systematic Review and Meta-Analysis, 2000-2024. China CDC Wkly 2024;6:597–604. https://doi.org/10.46234/ccdcw2024.116.

[3] Volkow ND, Swanson JM, Evins AE, DeLisi LE, Meier MH, Gonzalez R, et al. Effects of cannabis use on human behavior, including cognition,motivation, and psychosis: A review. JAMA Psychiatry 2016;73:292–7. https://doi.org/10.1001/jamapsychiatry.2015.3278.

[4] Hall W, Leung J, Lynskey M. The Effects of Cannabis Use on the Development of Adolescents and Young Adults 2020. https://doi.org/10.1146/annurev-devpsych-040320.

[5] Hinckley Jesse D. and Dillon J. Developmental Impact. In: Riggs Paula and Thant T, editor. Cannabis in Psychiatric Practice: A Practical Guide, Cham: Springer International Publishing; 2022, p. 45–59. https://doi.org/10.1007/978-3-031-04874-6_4.

[6] Baumer AM, Nestor BA, Potter K, Knoll S, Evins AE, Gilman J, et al. Assessing changes in sleep across four weeks among adolescents randomized to incentivized cannabis abstinence. Drug Alcohol Depend 2023;252. https://doi.org/10.1016/j.drugalcdep.2023.110989.

[7] Gaston SA, Alhasan DM, Jones RD, Braxton Jackson W, Kesner AJ, Buxton OM, et al. Cannabis use and sleep disturbances among White, Black, and Latino adults in the United States: A cross-sectional study of National Comorbidity Survey-Replication (2001-2003) data. Sleep Health 2023;9:587–95. https://doi.org/10.1016/j.sleh.2023.06.003.

[8] A Khurshid K. Relationship between sleep disturbances and addiction. Ment Health Addict Res 2018;3. https://doi.org/10.15761/mhar.1000162.

[9] Ouellet J, Spinney S, Assaf R, Boers E, Livet A, Potvin S, et al. Sleep as a Mediator Between Cannabis Use and Psychosis Vulnerability: A Longitudinal Cohort Study. Schizophr Bull Open 2023;4. https://doi.org/10.1093/schizbullopen/sgac072.

[10] Edwards D, Filbey FM. Are Sweet Dreams Made of These? Understanding the Relationship between Sleep and Cannabis Use. Cannabis Cannabinoid Res 2021;6:462–73. https://doi.org/10.1089/can.2020.0174.

[11] López-Pelayo H, Batalla A, Balcells MM, Colom J, Gual A. Assessment of cannabis use disorders: A systematic review of screening and diagnostic instruments. Psychol Med 2015;45:1121–33. https://doi.org/10.1017/S0033291714002463.

[12] Coelho SG, Hendershot CS, Quilty LC, Wardell JD. Screening for cannabis use disorder among young adults: Sensitivity, specificity, and item-level performance of the Cannabis Use Disorders Identification Test – Revised. Addictive Behaviors 2024;148:107859. https://doi.org/10.1016/j.addbeh.2023.107859.

[13] Striley CW, Hoeflich CC. Intricacies of Researching Cannabis Use and Use Disorders Among Veterans in the United States. American Journal of Psychiatry 2021;179:5–7. https://doi.org/10.1176/appi.ajp.2021.21111125.

[14] Cesarelli G, Ponsiglione AM, Sansone M, Amato F, Donisi L, Ricciardi C. Machine Learning for Biomedical Applications. Bioengineering 2024;11. https://doi.org/10.3390/bioengineering11080790.

[15] Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. New England Journal of Medicine 2019;380:1347–58. https://doi.org/10.1056/nejmra1814259.

[16] Ewert V, Pelletier S, Alarcon R, Nalpas B, Donnadieu-Rigole H, Trouillet R, et al. Determination of MoCA Cutoff Score in Patients with Alcohol Use Disorders. Alcohol Clin Exp Res 2018;42:403–12. https://doi.org/10.1111/acer.13547.

[17] Bensalah Y, Sabir M, Elomari F. Sleep disorders and addiction A study of 100 patients. European Psychiatry 2024;67:S304–S304. https://doi.org/10.1192/j.eurpsy.2024.633.

[18] Likhith S, Chitteti C, Dharani M, Nivedhitha V, Geethika NG, Godwin V. Machine Learning Model for Prediction of Smartphone Addiction. 2024 International Conference on Expert Clouds and Applications (ICOECA), IEEE; 2024, p. 924–9. https://doi.org/10.1109/ICOECA62351.2024.00163.

[19] Kumara UGHT, Siriwardana SSA, Weerasinghe L, Shavindi RAKI, Chiranjeewa HPRC, Siriwardana S. A Machine Learning Approach to Analyze the Drug Addiction. 2023 5th International Conference on Advancements in Computing (ICAC), 2023, p. 113–8. https://doi.org/10.1109/ICAC60630.2023.10417256.

[20] Feng Q, Ren Z, Wei D, Liu C, Wang X, Li X, et al. Connectome-based predictive modeling of Internet addiction symptomatology. Soc Cogn Affect Neurosci 2024;19. https://doi.org/10.1093/scan/nsae007.

[21] Pyzowski P, Herbert B, Malik WQ. Machine Learning Applied to Clinical Laboratory Data Predicts Patient-Specific, Near-Term Relapse in Patients in Medication for Opioid Use Disorder Treatment 2020. https://doi.org/10.1101/2020.08.10.20163881.

[22] Yang Z, Nguyen L, Jin F. Predicting Opioid Relapse Using Social Media Data 2018.

[23] Lee JY, Song MS, Yoo SY, Jang JH, Lee D, Jung YC, et al. Multimodal-based machine learning approach to classify features of internet gaming disorder and alcohol use disorder: A sensor-level and source-level resting-state electroencephalography activity and neuropsychological study. Compr Psychiatry 2024;130. https://doi.org/10.1016/j.comppsych.2024.152460.

[24] Wilkinson CS, Luján M, Hales C, Costa KM, Fiore VG, Knackstedt LA, et al. Listening to the Data: Computational Approaches to Addiction and Learning. Journal of Neuroscience, vol. 43, Society for Neuroscience; 2023, p. 7547–53. https://doi.org/10.1523/JNEUROSCI.1415-23.2023.

[25] BOUHADJA A, BOURAMOUL A. A Review on Recent Machine Learning Applications for Addiction Disorders. 2022 4th International Conference on Pattern Analysis and Intelligent Systems (PAIS), IEEE; 2022, p. 1–8. https://doi.org/10.1109/PAIS56586.2022.9946888.

[26] De Mattos BP, Mattjie C, Ravazio R, Barros RC, Grassi-Oliveira R. Craving for a Robust Methodology: A Systematic Review of Machine Learning Algorithms on Substance-Use Disorders Treatment Outcomes. Int J Ment Health Addict 2024. https://doi.org/10.1007/s11469-024-01403-z.

[27] Barenholtz E, Fitzgerald ND, Hahn WE. Machine-learning approaches to substance-abuse research: emerging trends and their implications. Curr Opin Psychiatry 2020;33:334–42.

[28] Suva M, Bhatia G. Artificial Intelligence in Addiction: Challenges and Opportunities. Indian J Psychol Med 2024. https://doi.org/10.1177/02537176241274148.

[29] Tahir GA. Ethical Challenges in Computer Vision: Ensuring Privacy and Mitigating Bias in Publicly Available Datasets 2024.

[30] Marceau EM, Lunn J, Berry J, Clin Neuro M, Kelly PJ, Solowij N. The Montreal Cognitive Assessment (MoCA) is sensitive to head injury and cognitive impairment in a residential alcohol and other drug therapeutic community. J Subst Abuse Treat 2016;66:30–6. https://doi.org/10.1016/j.jsat.2016.03.002.

[31] Park BK. The Pittsburg Sleep Quality Index (PSQI) and associated factors in middle-school students: A cross-sectional study. Child Health Nursing Research 2020;26:55–63. https://doi.org/10.4094/chnr.2020.26.1.55.

[32] Crocq M-Antoine, Guelfi J-Daniel, Boyer P, Pull C-Bernard, Pull-Erpelding M-Claire, American psychiatric association. DSM-5 : manuel diagnostique et statistique des troubles mentaux. 2015.

[33] Cresta Morgado P, Carusso M, Alonso Alemany L, Acion L. Practical foundations of machine learning for addiction research. Part I. Methods and techniques. American Journal of Drug and Alcohol Abuse 2022;48:260–71. https://doi.org/10.1080/00952990.2021.1995739.

[34] Bi Q, Goodman KE, Kaminsky J, Lessler J. What is machine learning? A primer for the epidemiologist. Am J Epidemiol 2019;188:2222–39. https://doi.org/10.1093/aje/kwz189.

[35] Fusar-Poli P, Werbeloff N, Rutigliano G, Oliver D, Davies C, Stahl D, et al. Transdiagnostic risk calculator for the automatic detection of individuals at risk and the prediction of psychosis: Second replication in an independent national health service trust. Schizophr Bull 2019;45:562–70. https://doi.org/10.1093/schbul/sby070.

[36] Mak KK, Lee K, Park C. Applications of machine learning in addiction studies : A systematic review. Psychiatry Res 2019;275:53–60. https://doi.org/10.1016/j.psychres.2019.03.001.

[37] Peng C-YJ, So T-SH. Logistic Regression Analysis and Reporting: A Primer. Understanding Statistics 2002;1:31–70. https://doi.org/10.1207/S15328031US0101_04.

[38] Baby Saral G, Priya R. Digital screen addiction with KNN and -Logistic regression classification. Mater Today Proc 2021. https://doi.org/10.1016/j.matpr.2020.11.360.

[39] Giustolisi O, Laucelli D. Improving generalization of artificial neural networks in rainfall-runoff modelling. Hydrological Sciences Journal 2005;50:439–57. https://doi.org/10.1623/hysj.50.3.439.65025.

[40] Verma J, Nath M, Tripathi P, Saini KK. Analysis and identification of kidney stone using Kth nearest neighbour (KNN) and support vector machine (SVM) classification techniques. Pattern Recognition and Image Analysis 2017;27:574–80. https://doi.org/10.1134/S1054661817030294.

[41] Noble WS. What is a support vector machine? Nat Biotechnol 2006;24:1565–7. https://doi.org/10.1038/nbt1206-1565.

[42] Choi J, Jung HT, Choi J. Marijuana addiction prediction models by gender in young adults using random forest. Online Journal of Nursing Informatics (OJNI) 2021;25.

[43] Nurma Yulita I, Ardiansyah F, Prabuwono AS, Ramdhani MR, Wicaksono MA, Trisanto A, et al. Recyclable Waste Classification using SquezeeNet and XGBoost. vol. 14. n.d.

[44] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol. 13-17- August-2016, Association for Computing Machinery; 2016, p. 785–94. https://doi.org/10.1145/2939672.2939785.

[45] Boldini D, Grisoni F, Kuhn D, Friedrich L, Sieber SA. Practical guidelines for the use of gradient boosting for molecular property

prediction. J Cheminform 2023;15. https://doi.org/10.1186/s13321-023-00743-7.

[46] Elansari T, Ouanan M, Bourray H. A novel Mathematical Modeling for Deep Multilayer Perceptron Optimization: Architecture Optimization and Activation Functions Selection. Statistics, Optimization and Information Computing 2024;12:1409–24. https://doi.org/10.19139/soic-2310-5070-1990.

[47] Cook JA. ROC curves and nonrandom data. Pattern Recognit Lett 2017;85:35–41. https://doi.org/10.1016/j.patrec.2016.11.015.

[48] Tian Y, Shi Y, Chen X, Chen W. AUC maximizing support vector machines with feature selection. Procedia Comput Sci 2011;4:1691–8. https://doi.org/10.1016/j.procs.2011.04.183.

[49] Parikh R, Mathai A, Parikh S, Chandra Sekhar G, Thomas R. Understanding and using sensitivity, specificity and predictive values. Indian J Ophthalmol 2008;56:45. https://doi.org/10.4103/0301-4738.37595.

[50] Monaghan TF, Rahman SN, Agudelo CW, Wein AJ, Lazar JM, Everaert K, et al. Foundational Statistical Principles in Medical Research: Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value. Medicina (B Aires) 2021;57:503. https://doi.org/10.3390/medicina57050503.

[51] Boyce D. Evaluation of Medical Laboratory Tests. Orthopaedic Physical Therapy Secrets. 3rd ed., Elsevier; 2017, p. 125–34. https://doi.org/10.1016/B978-0-323-28683-1.00017-5.