



## OPEN ACCESS

EDITED BY  
Shabnam Sadeghi Esfahlani,  
Anglia Ruskin University, United Kingdom

REVIEWED BY  
JingTao Yao,  
University of Regina, Canada  
Bozena Kostek,  
Gdansk University of Technology, Poland

\*CORRESPONDENCE  
Sheela Ramanna  
✉ s.ramanna@uwinnipeg.ca

SPECIALTY SECTION  
This article was submitted to  
Pattern Recognition,  
a section of the journal  
Frontiers in Artificial Intelligence

RECEIVED 29 June 2022  
ACCEPTED 27 January 2023  
PUBLISHED 15 February 2023

CITATION  
Ramanna S, Ashrafi N, Loster E, Debroni K and  
Turner S (2023) Rough-set based learning:  
Assessing patterns and predictability of anxiety,  
depression, and sleep scores associated with  
the use of cannabinoid-based medicine during  
COVID-19. *Front. Artif. Intell.* 6:981953.  
doi: 10.3389/frai.2023.981953

COPYRIGHT  
© 2023 Ramanna, Ashrafi, Loster, Debroni and  
Turner. This is an open-access article  
distributed under the terms of the [Creative  
Commons Attribution License \(CC BY\)](#). The use,  
distribution or reproduction in other forums is  
permitted, provided the original author(s) and  
the copyright owner(s) are credited and that  
the original publication in this journal is cited, in  
accordance with accepted academic practice.  
No use, distribution or reproduction is  
permitted which does not comply with these  
terms.

# Rough-set based learning: Assessing patterns and predictability of anxiety, depression, and sleep scores associated with the use of cannabinoid-based medicine during COVID-19

Sheela Ramanna<sup>1\*</sup>, Negin Ashrafi<sup>1</sup>, Evan Loster<sup>2</sup>, Karen Debroni<sup>2</sup> and  
Shelley Turner<sup>2</sup>

<sup>1</sup>Department of Applied Computer Science, University of Winnipeg, Winnipeg, MB, Canada, <sup>2</sup>Ekosi Health  
Centre Corporation, Winnipeg, MB, Canada

Recently, research is emerging highlighting the potential of cannabinoids' beneficial effects related to anxiety, mood, and sleep disorders as well as pointing to an increased use of cannabinoid-based medicines since COVID-19 was declared a pandemic. The objective of this research is 3 fold: i) to evaluate the relationship of the clinical delivery of cannabinoid-based medicine for anxiety, depression and sleep scores by utilizing machine learning specifically rough set methods; ii) to discover patterns based on patient features such as specific cannabinoid recommendations, diagnosis information, decreasing/increasing levels of clinical assessment tools (CAT) scores over a period of time; and iii) to predict whether new patients could potentially experience either an increase or decrease in CAT scores. The dataset for this study was derived from patient visits to Ekosi Health Centres, Canada over a 2 year period including the COVID timeline. Extensive pre-processing and feature engineering was performed. A class feature indicative of their progress or lack thereof due to the treatment received was introduced. Six Rough/Fuzzy-Rough classifiers as well as Random Forest and RIPPER classifiers were trained on the patient dataset using a 10-fold stratified CV method. The highest overall accuracy, sensitivity and specificity measures of over 99% was obtained using the rule-based rough-set learning model. In this study, we have identified rough-set based machine learning model with high accuracy that could be utilized for future studies regarding cannabinoids and precision medicine.

## KEYWORDS

rough sets, machine learning, electronic health records, mental health, cannabinoid medicine, rough-fuzzy sets

## 1. Introduction

COVID-19 is an unprecedented health crisis causing a great deal of stress and sleep challenges for populations in Canada. Research is emerging highlighting the potential of cannabinoids' beneficial effects related to chronic pain (Lynch and Campbell, 2011), substance use (Hay et al., 2018), addiction (Prud'homme et al., 2015), and poor mental health (Lee et al., 2017; McGuire et al., 2018). Recent studies point to the clinically significant acute impacts the

pandemic is having on insomnia rates (Morin and Carrier, 2021). Where there is recent research which points to the potential positive impact cannabinoid may have regarding sleep (Ware and Ferguson, 2015; Sznitman et al., 2020), a 2017 review (Babson et al., 2017) of the literature on cannabinoids and sleep suggested mixed results and highlighted the need for further research.

With the availability of large amounts of patient data, machine learning (ML) techniques, specifically, supervised and deep learning classifiers, have made it possible to detect, diagnose and treat mental health disorders. Common dataset formats include: Electronic health records (EHR) (Ramesh et al., 2021), Social Media (e.g., Twitter, Reddit) (Tariq et al., 2019; Kim et al., 2021), Image (e.g., MRI) (Noor et al., 2020) and Audio (Xiao et al., 2016). Shatte et al. (2019), present an in-depth review of about 300 papers related to ML and its application in mental health. The most common ML models used include: support vector machines, decision trees, naive bayes, k-nearest neighbor, and neural networks (deep learning). Latent Dirichlet allocation (LDA) and sentiment analysis models were used for learning from textual and social media data. Predicting mental health from social media data is an interdisciplinary area also known as *human-centric machine learning* where human insights are combined with data driven predictions (Chancellor et al., 2019a). Ethical tensions in inferring mental health states of individuals from social media data are discussed in Chancellor et al. (2019b). In another study (Edo-Osagie et al., 2020), Twitter data was used in public health in surveillance, detection, and prevention of events. In Sharma and Verbeke (2020) XGboost classifier was used to assess the effectiveness of biomarkers to classify depression cases from healthy cases using a large dataset from Netherlands. In a recent study (Dobias et al., 2022), ML models were used to assess whether the adolescents with depressive symptoms had access to treatments and, if yes, where the treatments were received. Nemesure et al. (2021) proposed an ensemble of six classifiers to predict general anxiety disorder (GAD) and major depressive disorder (MDD) problems. Edgcomb and Zima (2019) discuss application of Natural Language Processing techniques to EHR phenotyping (unstructured text) which contain narrative text such as physician notes for improving mental health services. Rahman et al. (2020) present a survey of papers on mental health detection using ML techniques in Online Social Networks. A review of 2261 articles on the application of deep learning models in mental health outcomes was presented in Su et al. (2020). In this paper, various deep neural network architectures as well as different forms of clinical data (neuroimages, EMR, audio visual and social media) are discussed.

A study by Alghamdi et al. (2018) reported using Gaussian Processes, Support Vector Machines, and Neural Networks algorithm to extract predictive patterns of cannabis use and the onset of first episode psychosis from clinical data. This study does not include any specific medical cannabis product (such as CBD or THC). The cannabis use feature consists of three values (never used, hash, or skunk). Another study by Choi et al. (2021) analyze behavior related to depression and suicide risk using machine learning algorithms (Logistic regression, Random Forest, and K-Nearest Neighbor) in adults that use marijuana (cannabis).

Since 1991, rough set theory has been applied extensively in medical informatics (Pawlak, 1991; hrn, 1999; Pattaraintakorn and Cercone, 2008; Hassanien et al., 2009; Gil Herrera et al., 2015; Pathan et al., 2020). More recently, rough set model was used to

analyze outpatient service quality in a hospital setting in China (Du et al., 2021). To the best of our knowledge, there are only a few papers related to the application of rough sets in mental health. In Shusaku and Kudo (2009), rough set theory was used to explore the relationship between human psychological state (scores of a psychological scale) and physiological state (level of the secretory biomarkers). In Nomura et al. (2010), the authors use rough sets instead of conventional linear correlation analysis for mining the relationship between a subjective stress scale and salivary cortisol stress biomarker. In Liu et al. (2011), a hybrid rough set and Taguchi-genetic algorithm (RS-HTGA) was proposed to determine the relationship between mental stress and biomedical signals. The efficacy of their model was tested on a clinical dataset comprising 362 cases (196 male, 166 female). In Liu et al. (2014), the RS-HTGA algorithm achieved sensitivity, specificity, and precision scores of 96%. In Mittal et al. (2014), the authors present an application of rough sets for attribute reduction to identify depressive episodes.

In this research, we seek i) to evaluate the relationship of the clinical delivery of cannabinoid-based medicine for anxiety, depression and sleep scores by utilizing machine learning specifically rough set methods; ii) to discover patterns based on patient features such as specific cannabinoid recommendations [includes medical cannabis products contain varying amounts of cannabidiol (CBD) and tetrahydrocannabinol (THC)], diagnosis information, decreasing/increasing levels of clinical assessment tools: GAD-7 (General Anxiety Disorder-7), PHQ-9 (Patient Health Questionnaire-9), and PSQI (Pittsburgh Sleep Quality Index) (Buysse et al., 1988) scores over a period of time including during the COVID timeline; and iii) to predict whether new patients could potentially experience either an increase or decrease in clinical assessment tool scores (pl. see [Supplementary Table 4, Appendix A](#) for scale values for each tool). PHQ-9 and GAD-7 scales are well-established instruments for screening for symptoms of depression and generalized anxiety respectively (Kroenke et al., 2001; Spitzer et al., 2006). The dataset for this study was derived from patient visits to Ekosi Health Centres in Manitoba and Ontario, Canada from January, 2019 to April, 2021. Extensive pre-processing and feature engineering was performed on the dataset. To determine the outcome of a patient's treatment, a class feature (Worse, Better, or No Change) indicative of their progress or lack thereof due to the treatment received was introduced. A two-class experiment (Worse or Better) was also explored. Well-known supervised machine learning classification algorithms: Random Forest (tree-based), RIPPER (rule-based) in addition to rough and fuzzy models were trained on the patient dataset. All experiments were conducted using a 10-fold CV stratified method. Also, prediction of new cases using the rough set-based classifier (LEM2 algorithm) is presented. Our results demonstrate that rough-set based classifier (with LEM2 algorithm) is superior to all other tested models in terms of overall classification accuracy (99.34% for the 3-class experiment), accuracy per class, sensitivity, and specificity values for both the 2-class and the 3-class experiments. A statistical *t*-test reveals that there is a difference between rough-set based classifier and other tested classifiers for the 3-class experiment.

Our results support the findings that the combination of THC and CBD appears to be most beneficial on GAD-7, PHQ-9, and PSQI scores for patients dealing with anxiety, depression, sleep disorders, chronic pain, and arthritis. In this study, we have identified

rough-set based machine learning model with high accuracy that could be utilized for future studies regarding cannabinoids and precision medicine. This research points to a novel application of rough and fuzzy rough classification learning to a case study involving cannabinoid medicine and anxiety, depression, and sleep pattern data.

## 2. Preliminaries

In this section, we present a brief review of rough and fuzzy rough set theory concepts that were used in this research. Specifically, we use different forms of fuzzy and rough nearest neighbor classification algorithms.

### 2.1. Rough sets

In classical set theory, we can classify whether elements either belong to a set or not. This is a precise or crisp set where the sets have sharp boundaries. However, when boundaries are *unsharp or vague*, it is difficult to classify elements uniquely to one set. In other words, this will result in a boundary region with elements that cannot be classified precisely. Rough set theory was proposed by Zdzislaw Pawlak in early 80's as a mathematical framework to analyze vague data and ill-defined objects based on an indiscernibility or equivalence relation (Pawlak, 1982; Pawlak and Skowron, 1994). Equivalence relations generate equivalence classes and the notion of indiscernibility is defined relative to a given set of attributes (Pawlak and Skowron, 2007). Due to the lack of knowledge (or uncertainty) that objects might belong to more than one set (or class), two approximation operators (*lower and upper*) are introduced in rough set theory to generate precise sets. In supervised classification, the advantage of rough set theory is that no prior or additional data is needed to categorize data into classes (Pawlak, 1991). [Supplementary Figure 1](#) shows the regions that emerge with rough set approximation. The lower approximation consists of the objects that certainly belong to the set (orange region) and upper approximation consists of objects that their membership is not certain (green region). The regions are depicted as squares only for the sake of illustration, but they can be of arbitrary shape. We should note that each granule can contain an arbitrary number of objects or may be empty. The oval denotes the target  $X$  which, in the case of supervised learning, is either a class or a pattern that needs to be learned.

Let  $U$  be a finite, non-empty universe of objects and let  $R \subseteq U \times U$  denote a binary relation on the universe  $U$ .  $R$  is called an *indiscernibility relation* and for rough sets, it has to be an *equivalence relation*. The pair  $(U, R) = \mathcal{A}$  is an *approximation space*  $\mathcal{A}$  (Stepaniuk, 1998). Let  $X \subseteq U$  be a target concept in this universe. Then the task is to create an approximated representation for  $X$  in  $U$  with the help of  $R$ . Let  $[x]_R$  denote the indiscernibility class of  $x$  i.e.  $y \in [x]_R \iff (x, y) \in R$ . Then, every equivalence class forms a *granule* or *partition* containing objects that are indiscernible for this approximation space  $\mathcal{A}$ . Therefore, every single item in a granule is considered identical and inseparable. These granules are approximated by the following means:

- **Lower approximation.** Intuitively, these are the objects which *certainly* belong to  $X$  with respect to  $\mathcal{A}$ .

$$\mathcal{L}_{\mathcal{A}}(X) = \{x \in U : [x]_R \subseteq X\}.$$

- **Upper approximation.** Intuitively, these are the objects which *may* belong to  $X$  with respect to  $\mathcal{A}$ .

$$\mathcal{U}_{\mathcal{A}}(X) = \{x \in U : [x]_R \cap X \neq \emptyset\}.$$

These two approximations will also form the following two regions:

- **Boundary region.** These are the objects occurring in the upper approximation but not in lower approximation of  $X$ .

$$\mathcal{B}_{\mathcal{A}}(X) = \mathcal{U}_{\mathcal{A}}(X) - \mathcal{L}_{\mathcal{A}}(X).$$

- **Negative region.** These are the objects that certainly don't belong to  $X$ .

$$U - \mathcal{U}_{\mathcal{A}}(X).$$

With this framework, we have two different types of sets: a set  $X$  is called a *crisp set* if and only if  $\mathcal{B}_{\mathcal{A}}(X) = \emptyset$ . Otherwise, it is called a *rough set*. The pair  $[\mathcal{U}_{\mathcal{A}}(X), \mathcal{L}_{\mathcal{A}}(X)]$  forms the *rough approximation* for  $X$  (see [Supplementary Section 1](#) for an illustration and list of symbols and their interpretation used in this paper).

### 2.2. Fuzzy rough sets

Fuzzy set theory was proposed by Zadeh (1997) as an extension of traditional set theory to deal with uncertainty and vagueness. In the context of fuzzy sets, let  $X$  denote the universe, a fuzzy set  $A \in X$  is characterized by a mapping  $X \rightarrow [0, 1]$  which is also called a membership function. A fuzzy relation  $R$  in  $X$  which is also a fuzzy set and is characterized by a mapping  $R: X \times X \rightarrow [0, 1]$  (Zadeh, 1997). Del Cerro and Prade (1986), Nakamura (1988), and Dubois and Prade (1990), introduced the idea of combining fuzzy and rough sets to develop soft similarity classes i.e., fuzzifying the approximations of rough set theory. Formally, a fuzzy rough set is a pair  $(A_1, A_2) \in (X, R)$  where  $A$  is a fuzzy set in  $X$  such that  $R \downarrow A = A_1$  and  $R \uparrow A = A_2$  and  $R$  is a fuzzy relation in  $X$  Cornelis et al. (2008). Fuzzy rough sets permit partial membership of an object to the lower and upper approximations and the approximate nature of information are modeled by means of fuzzy indiscernibility relations. In general,  $R$  can be considered as a fuzzy tolerance relation such  $R(x, x) = 1$  and  $R(x, y) = R(y, x)$  for all  $x, y$  in  $X$ . Let  $U$  be the universe and  $R$  the fuzzy tolerance relation in  $U$  which is a mapping  $U \rightarrow [0, 1]$  and  $A$  is a fuzzy set in  $U$ , the upper  $(R \uparrow A)$  and lower approximation of  $A$   $(R \downarrow A)$  is calculated by  $R$  using different methods. The general form for this calculation from Jensen and Cornelis (2011) is as follows:

$$(R \downarrow A)(x) = \inf_{y \in U} \mathcal{I}(R(x, y), A(y)) \tag{1}$$

$$(R \uparrow A)(x) = \sup_{y \in U} \mathcal{T}(R(x, y), A(y)) \tag{2}$$

where  $\mathcal{I}$  is an implicator and  $\mathcal{T}$  is a t-norm which are fuzzy logic connectives crucial for fuzzy rough hybridization. The Kleen-Diennes Implicator implemented in the WEKA platform<sup>1</sup> is defined as

$$\mathcal{T}_M = \min(x, y) \tag{3}$$

$$\mathcal{I}_M = \max(1 - x, y) \tag{4}$$

In the Fuzzy Rough Nearest Neighbor (FRNN) implementation, given a set of conditional attributes  $C$ ,  $R$  is defined as where  $R_a$  is the degree to which objects  $x$  and  $y$  are similar for attribute  $a$  Jensen and Cornelis (2011):

$$R(x, y) = \min_{a \in C} R_a(x, y) \tag{5}$$

The two options for  $R_a$  are:

$$R_a^1(x, y) = \exp\left(-\frac{(a(x) - a(y))^2}{2\sigma_a^2}\right) \tag{6}$$

$$R_a^2(x, y) = 1 - \frac{\|a(x) - a(y)\|}{|a_{max} - a_{min}|} \tag{7}$$

where  $\sigma_a^2$  is variance of attribute  $a$ ,  $a_{max}$ , and  $a_{min}$  are maximal and minimal values of attribute  $a$ . We have used option 2 given Equation (7). For the sake of completeness, we use the FRNN algorithm presented in Jensen and Cornelis (2011) implemented in WEKA in the Supplementary Section 1.

### 3. Materials

#### 3.1. Data preparation

The original dataset includes 541 unique patients and 32,514 records (for single and multiple visits). In this paper, patients with at least two different dates of a medical appointment with one of the Health Centers were considered (referred to as multiple visit dataset). The ages for youngest and oldest patients were 6 and 108 years respectively (with a mean value of 58.61). Additionally, this multiple dataset included 390 types of diagnoses with 75 unique cannabidiol formulations. After data cleaning, diagnoses types that were not of interest in this study removed, the multiple visit dataset was reduced to 354 patients from 375 patients. The final dataset after preprocessing for experimentation was: 8,281 records (2,911 male and 5,730 female).

- Patient Id : Since this feature uniquely identifies a patient, due to privacy reasons, this feature value was anonymized by removing each patient's name, date of birth, and any information that might reveal the patient's identity.
- Age: This feature gives the age of the patient where the minimum value for age is 6 and the maximum value is 108.

- Clinical Assessment Tool (CAT): This feature indicates the type of the clinical measure assessment tool that was utilized to assess and score the patient. Three specific CAT types were observed in this study; the GAD-7 (General Anxiety Disorder-7), PHQ-9 (Patient Health Questionnaire-9), and PSQI (Pittsburgh Sleep Quality Index).
- CAT Value : The feature gives the values for each of the CAT types: GAD-7, PHQ-9, and PSQI.
- CAT Observation Date: This feature gives the date on which a CAT value was observed.
- Sex Id: This feature gives the gender and the distribution of the patients coded as 1: male (34.2%) and 2: female (65.8%).
- Cannabinoid recommendation: This feature indicates the specific cannabinoid recommendation. The medical cannabis products contain varying amounts of cannabidiol (CBD) and tetrahydrocannabinol (THC), two phytocannabinoids found in cannabis.
- Diagnosis: This feature indicates the diagnosis of the patient. There were 390 types of diagnoses and only 13 types were considered in this research.

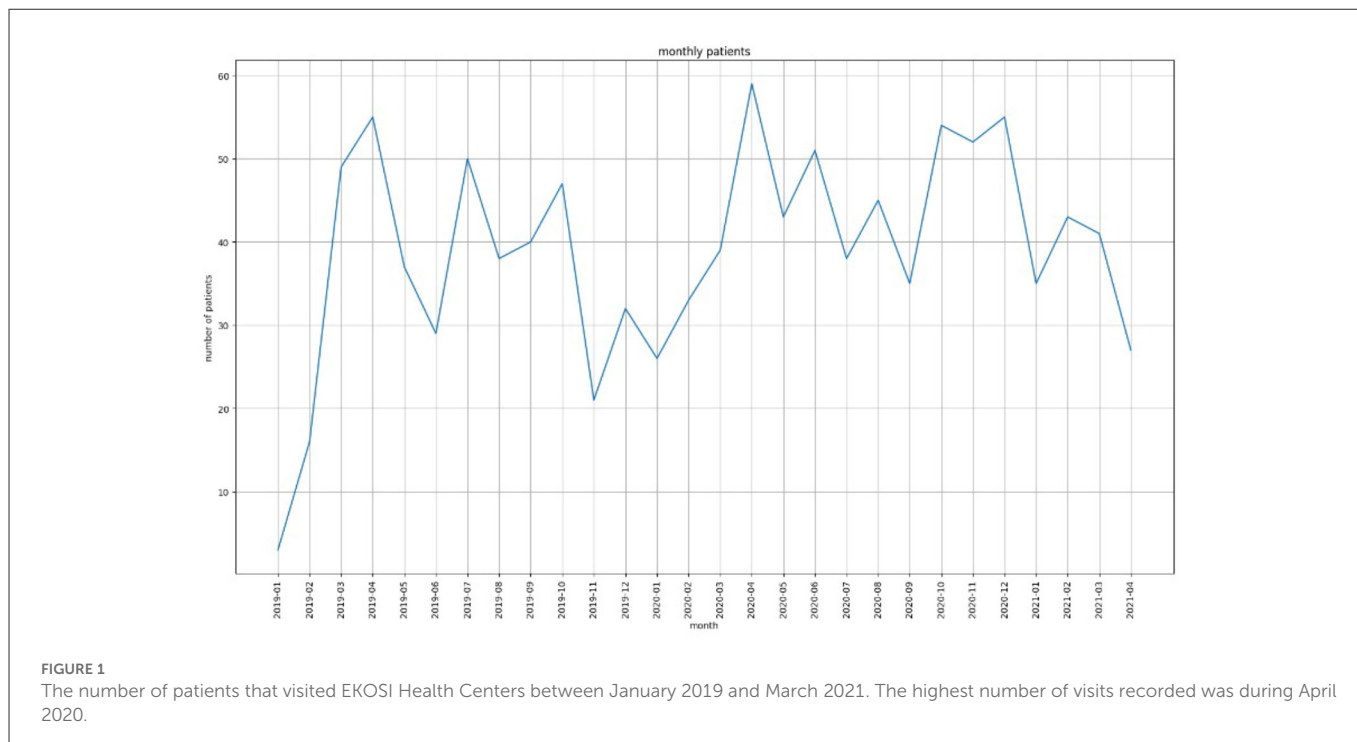
The raw data had several problems such as missing or invalid values, continuous values for dosage and similar diagnosis which required extensive preprocessing. In the following section, we discuss the preprocessing steps applied to the dataset.

#### 3.2. Preprocessing

The description of the steps are as follows:

- Invalid and missing values: Invalid and null values were found in gender and CAT value features and were removed. For example, there were 114 records that gender had a value other than 1 or 2. Also, in the original dataset, there were 18 records that CAT value greater than 27. There were very few records with missing values which were also removed.
- Diagnosis coding: The raw data consisted of 390 diagnoses categories. Some low occurring or categories not relevant to this study were removed (e.g.: ADHD, MS, Anemia, Vitiligo, Blood Clot, Schizophrenia, and Overweight). Other granular categories such as migraine, classical migraine, common migraine, and chronic migraine without aura were combined into the broader migraine category. In this study, we were primarily interested in chronic pain, so patients with migraine and headache were included in the chronic pain category.
- Cannabinoid recommendation coding: The values for this feature were continuous since they represent dosage values. Since we were only interested in a broad class of values, these values were converted into integers using regular expressions (using Python regular expression package).
- Multiple cannabinoid recommendations: Many patients (almost 40%) were recommended more than one cannabinoid product for one particular diagnosis in a single visit. This was primarily for cannabinoid product classes CBD and CBD AND THC: CBD. For such patients, the recommendation was changed to CBD AND THC: CBD (category 3). This resulted in duplicate records and these duplicate records were removed.

<sup>1</sup> <http://users.aber.ac.uk/rkj/book/wekafull.jar>



- Multiple CAT values: Some patients had a different value for GAD-7/PHQ-9/PSQI during a single visit. For this feature, records with largest CAT value (most severe) were recorded.
- Time of visit: All time values with a small difference during a single visit were standardized and 21 patients had a slight time difference in at least one record.
- CAT value coding: This generated feature was designed to merge CAT value and CAT types: A0-A3, D0-D4, and S0-S3 to represent anxiety (GAD-7), depression (PHQ-9), and sleep disorder (PSQI) severity level respectively.

One of the main objectives of this study was to detect patterns in the fluctuations of values for GAD-7, PHQ-9 and PSQI (clinical assessment tools) for a patient during a time period. Figure 1 shows the number of patients from 2019 to 2020. As the figure shows, the number of patients that visited the Center was the highest (59) during March 2020 which was also the start of the first wave of COVID. In particular, we were interested in the overall outcome of a patient's quality of life in terms of whether their GAD-7/PHQ-9/PSQI scores were increasing/decreasing/constant during the period of observation. In addition, this information had to be co-related with their cannabinoid product recommendation and diagnosis.

Figure 2 shows the trends in score values for a single patient at the peak of COVID. It can be seen that in Figures 2A, B there is no regular pattern for GAD-7/PHQ-9/PSQI scores.

### 3.3. Engineered feature—patient status

To determine the outcome of a patient's treatment, we introduced a new feature (status) indicative of their progress or lack thereof due to the treatment received over a period of time. Three values for status were decided: Worse, Better, or No Change. An additional reason for introducing these labels was to train classification models

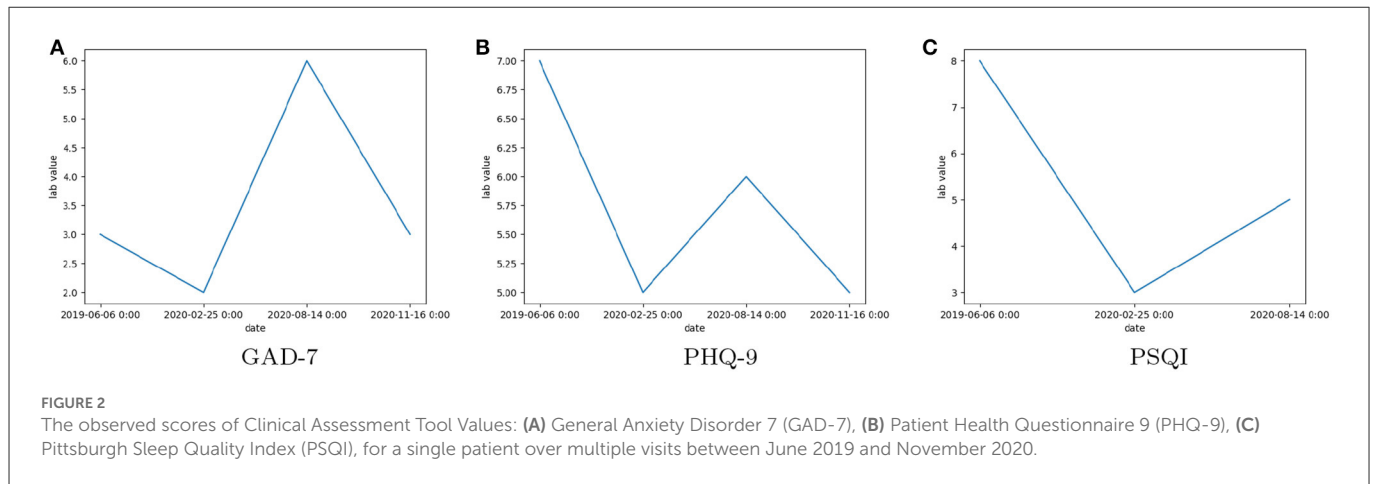
so that these models can be used to determine (or predict) the status of a new patient. The flowchart for computation of the value of this feature is given in Supplementary Section 2.2. The assumption behind this computation was that, since the score values for a disorder type does not follow any trend (as shown in Figure 2), a mean score value would be representative of a patient's score over the entire time period. In addition, there were unequal scores recorded for each patient during a time period. This problem was also observed for different disorder types as well. Hence, we separated the data into different CAT types first and then performed the labeling. This method also solved the problem of lack of observations of a CAT type with a time period for any given patient. The distribution of patient records based on i) labeled patient's status (Worse, Better, and No Change), ii) diagnosis (depression), and iii) CAT type for the four different types of cannabinoid formulations: CBD, THC:CBD, THC and CBD AND THC:CBD. The distribution of patients for other diagnoses (ex: Sleep Disorder, Chronic Pain, Arthritis, Anxiety, Depression) can be found in the Supplementary Section 3 and Supplementary Figures 5–9. However, chronic pain is the most frequent diagnosis and there were no patients with sleep disorder diagnosis who were recommended THC formulation.

## 4. Results

Results of the following algorithms are reported in Table 1: Random Forest, JRIP- Ripper algorithm in WEKA 3.7.2<sup>2</sup> as well the classical rough sets model Rough Sets implemented in the Rough Set Exploration System (RSES 2.2.2).<sup>3</sup> Many of the algorithmic

<sup>2</sup> <https://www.cs.waikato.ac.nz/ml/weka/>

<sup>3</sup> <https://www.roughsets.org/roughsets/software/>



methods used by RSES have their origins from rough set theory. The RSES software and underlying computational methods have been successfully applied in many studies and applications (Bazan and Szczuka, 2005). The most established classifiers in RSES are based on a set of decision rules. The LEM2 (Learning from Examples) rule-based algorithm is a covering technique that uses upper and lower approximation operators from rough sets to generate the classification results (Grzymala-Busse, 1992). The JRIP algorithm is an efficient implementation of the rule-based RIPPER (Repeated Incremental Pruning to Produce Error Reduction) algorithm (Cohen, 1995). Results from other fuzzy and rough sets algorithms implemented in WEKA are reported in the Supplementary Section 6. In Table 1, we provide the results (average values) in terms of classification accuracy (%), sensitivity(%), and specificity (%) for the four classifiers.

For the Random Forest classifier, the following parameters were used: maximum depth was set to 6 and number of trees was set to 10. For the JRIP classifier, one fold was used for pruning and two folds for growing the rules. For the FRNN classifier, 10 nearest neighbors were chosen, with Kleen-Diennes Implicator and Kleen-Diennes t-norm. For the results reported in Table 1, ten sets of training and testing pairs were used for experimentation across both platforms (10-Fold Cross Validation (CV) stratified method). We considered two forms of outcome of a patient's treatment: 2-class (Better or Worse) and 3-class (Worse, Better, or No Change) referred to as binary and ternary respectively. For the 2-class experiment, the Better class contains 5,157 records and the Worse class contains 3,124 records. For the 3-class, the Better class contains 5,157 records, Worse class contains 2,470 records and the No Change class contains 654 records.

- Average number of rules
  - Binary classification: JRIP: 157, LEM2: 2,843
  - Ternary classification: JRIP: 208, LEM2: 2,758.
- Execution time in secs per fold
  - Fuzzy rough NN: 0.01, Random Forest: 0.19–0.25
  - JRIP: 1–1.5, LEM2: 12.

## 5. Discussion

From the results in Table 1, the Rough Sets classifier (with LEM2 algorithm) gives the best overall result in terms of overall classification accuracy, accuracy per class, sensitivity, and specificity values for both the 2-class and the 3-class experiments. The best accuracy (99.34%) was obtained in the ternary classification. It is important to note in both cases (binary and ternary), the classification accuracy is over 99%. The class distribution in both experiments are highly imbalanced with the Better class having almost 2.5 times more records than the Worse class and 7.9 times the No Change class. The per class accuracy results are also consistently better in the ternary classification across all three classes. FRNN and JRIP classifiers are second best in terms of overall accuracy. The parameter settings for FRNN (number of K neighbors) and for JRIP were tuned to get the best results. Overall, FRNN gives the next best results. In terms of sensitivity (or the true positive rate), the best result for the ternary classification is with the better class (Rough Sets classifier- 99.6%). This is not surprising since there are more training examples for this class. In terms of specificity, the best result for the ternary classification is with the No Change class (or the true negative rate) is 99.9% which is consistent with the accuracy results for this class. Overall, rule-based models (RSES and JRIP) seem to do better than tree-based ensemble (Random Forest) and Fuzzy Rough Nearest Neighbor (FRNN) models for this dataset. However, the number of rules using the LEM2 algorithm is almost 13 times more than the JRIP classifier. This is also reflected in the time it takes for classification. In terms of execution time, FRNN is the best performing classifier. We have presented classification results of other nearest neighbor implementations using different forms of fuzzy and rough sets (Supplementary Table 4).

## 6. Conclusion

One of the objectives of this research was to predict whether new patients could potentially experience either an increase or decrease in clinical assessment tool scores. We demonstrate this by presenting results for 15 cases (ternary classification) with the LEM2 classifier in Supplementary Figure 8. Based on the results of a paired *t*-test

TABLE 1 Results—binary and ternary class experiments.

Metric	Fuzzy rough NN	Random Forest	JRIP	Rough sets (RSES)
Mean accuracy overall (binary)	97.11	96.22	97.16	99.20
Mean accuracy overall (ternary)	96.79	95.82	96.52	99.34
Accuracy (binary–better)	97.9	97.6	99.1	99.4
Accuracy (binary–worse)	95.7	93.9	93.8	98.78
Accuracy (ternary–better)	97.0	96.3	96.6	99.3
Accuracy (ternary–worse)	97.3	96.7	97.6	99.4
Accuracy (ternary–no change)	99.2	98.5	98.7	99.8
Sensitivity (binary–better)	97.9	97.6	99.1	99.4
Sensitivity (binary–worse)	95.7	93.9	93.8	98.78
Specificity (binary–better)	95.7	93.9	93.8	98.78
Specificity (binary–worse)	97.9	97.6	99.1	99.4
Sensitivity (ternary–better)	97.9	98.1	99.1	99.6
Sensitivity (ternary–worse)	95.0	93.6	93.5	98.7
Sensitivity (ternary–no change)	94.2	87.7	86.5	99.3
Specificity (ternary–better)	95.5	93.6	92.4	98.9
Specificity (ternary–worse)	98.2	98.0	99.4	99.7
Specificity (ternary–no change)	99.6	99.5	99.7	99.9

(Supplementary Section 6), there is no statistical difference between FRNN and Rough Set classifier (highlighted in blue) in the binary case. For the ternary classification, there is a difference between Rough Set classifier and the other classifiers (JRIP, RF, and FRNN).

The combination of THC and CBD appears to be most beneficial on GAD-7, PHQ-9, and PSQI scores for patients dealing with anxiety, depression, sleep disorders, chronic pain, and arthritis. CBD alone overall had a positive effect on GAD-7, PHQ-9, and PSQI scores across all conditions but not as pronounced of a positive effect as a THC:CBD combination. THC alone worsened GAD-7 scores for all conditions except for arthritis patients which suggests THC may increase anxiety in patients. Other research has suggested that the combination of cannabinoids and terpenes or the “Entourage Effect” enhances the therapeutic potential of cannabinoids (Russo, 2019; Ferber et al., 2020). Our results appear to support this finding. Practically, this study highlights the need for additional research to further identify predictability, patterns and understand the efficacy and real-world evidence regarding cannabinoids, especially the combination of cannabinoids, for anxiety, depression, sleep disorders, chronic pain, and arthritis. We have identified rough-set based machine learning model with high accuracy that could be utilized for future studies regarding cannabinoids and precision medicine. Further research on the interaction and synergy of cannabinoids and terpenes, using this model, may lead to new and valuable insights, for the benefit of patients and health care practitioners alike.

The introduction of precise milligram cannabinoid dosing feature (numeric values), would lead to the application of discretization and normalization methods. Currently all patient features are nominal. The methodology to determine the

engineered feature (status) for each patient would need further examination as the number of multiple visits for a patient would increase and fluctuate over time. Even though FRNN algorithm seems to be the best performing in terms of runtime performance, we believe that the rule-based algorithm (e.g., LEM2) facilitates easier interpretability of decisions in a clinical setting.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Ethics statement

Approval for the secondary use of data was granted by the University of Winnipeg Human Ethics Research Board (UHREB) Protocol#WR001. Written informed consent from the participants’ legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

## Author contributions

SR designed the materials and methods and writing of the paper. NA performed data processing and experiments. EL and KD provided and analyzed the data as well as assisted in editing the draft. ST contributed to the research. All authors contributed to the article and approved the submitted version.

## Funding

This research was supported by the MITACS Accelerate grant IT-18982.

## Acknowledgments

SR acknowledges the NSERC Discovery Grants program.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## References

- Alghamdi, W., Stamate, D., Stahl, D., Zamyatin, A., Murray, R., and Di Forti, M. (2018). "A new machine learning framework for understanding the link between cannabis use and first-episode psychosis," in *Studies in Health Technology and Informatics*, 9–16.
- Babson, K. A., Sottile, J., and Morabito, D. (2017). Cannabis, cannabinoids, and sleep: a review of the literature. *Curr. Psychiat. Rep.* 19, 1–12. doi: 10.1007/s11920-017-0775-9
- Bazan, J. G., and Szczuka, M. (2005). "The rough set exploration system," in *Transactions on Rough Sets III, LNCS* (Berlin Heidelberg: Springer) 3400, 37–56. doi: 10.1007/11427834\_2
- Buysse, D., Reynolds, I. I. I., C., Monk, T., Berman, S. R., and Kupfer, D. (1988). The pittsburgh sleep quality index: A new instrument for psychiatric practice and research. *Psychiat. Res.* 28, 193–213. doi: 10.1016/0165-1781(89)90047-4
- Chancellor, S., Baumer, E. P., and De Choudhury, M. (2019a). Who is the "human" in human-centered machine learning: The case of predicting mental health from social media. *Proc. ACM Hum. Comput. Inter.* 3, 1–32. doi: 10.1145/3359249
- Chancellor, S., Birnbaum, M. L., Caine, E. D., Silenzio, V. M., and De Choudhury, M. (2019b). "A taxonomy of ethical tensions in inferring mental health states from social media," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 79–88. doi: 10.1145/3287560.3287587
- Choi, J., Chung, J., and Choi, J. (2021). Exploring impact of marijuana (cannabis) abuse on adults using machine learning. *Int. J. Environ. Res. Public Health* 18, 10357. doi: 10.3390/ijerph181910357
- Cohen, W. W. (1995). "Fast effective rule induction," in *Twelfth International Conference on Machine Learning* (Morgan Kaufmann) 115–123. doi: 10.1016/B978-1-55860-377-6.50023-2
- Cornelis, C., De Cock, M., and Radzikowska, A. M. (2008). "Fuzzy rough sets: from theory into practice," in *Handbook of Granular Computing* (John Wiley and Sons, Ltd.) 533–52. doi: 10.1002/9780470724163.ch24
- Del Cerro, F., and Prade, H. (1986). Rough sets, twofold fuzzy sets and modal logic fuzziness in indiscernibility and partial information. *Mathem. Fuzzy Syst.* 88, 103–120.
- Dobias, M. L., Sugarman, M. B., Mullarkey, M. C., and Schleider, J. L. (2022). Predicting mental health treatment access among adolescents with elevated depressive symptoms: Machine learning approaches. *Admin. Policy Mental Health Serv. Res.* 49, 88–103. doi: 10.1007/s10488-021-01146-2
- Du, M.-L., Tung, T.-H., Tao, P., Chien, C.-W., and Chuang, Y.-C. (2021). Application of rough set theory to improve outpatient medical service quality in public hospitals based on the patient perspective. *Front. Public Health* 9, 739119. doi: 10.3389/fpubh.2021.739119
- Dubois, D., and Prade, H. (1990). Rough fuzzy sets and fuzzy rough sets\*. *Int. J. General Syst.* 17, 191–209. doi: 10.1080/03081079008935107
- Edgcomb, J. B., and Zima, B. (2019). Machine learning, natural language processing, and the electronic health record: innovations in mental health services research. *Psychiatr. Serv.* 70, 346–349. doi: 10.1176/appi.ps.201800401
- Edo-Osagie, O., De La Iglesia, B., Lake, I., and Edeghere, O. (2020). A scoping review of the use of twitter for public health research. *Comput. Biol. Med.* 122, 103770. doi: 10.1016/j.compbiomed.2020.103770
- Ferber, S. G., Namdar, D., Hen-Shoval, D., Eger, G., Koltai, H., Shoval, G., et al. (2020). The case for the entourage effect and conventional breeding of clinical cannabis: No strain, no gain. *Curr. Neuropharmacol.* 18, 1969. doi: 10.2174/1570159X17666190903103923
- Gil Herrera, E., Aden Buie, G., Tsalatsanis, A., Barnes, L. E., and Djulbegovic, B. (2015). Rough set theory based prognostic classification models for hospice referral. *BMC Med. Inform. Decis. Mak.* 15, 580–585. doi: 10.1186/s12911-015-0216-9
- Grzymała-Busse, J. (1992). Lers-a system for learning from examples based on rough sets. *Intell. Decis. Support* 11, 3–18. doi: 10.1007/978-94-015-7975-9\_1
- Hassanien, A. E., Abraham, A., Peters, J. F., and Schaefer, G. (2009). "Rough sets in medical informatics applications," in *Applications of Soft Computing*, eds. J., Mehnen, M., Köppen, A., Saad, and A., Tiwari (Berlin, Heidelberg: Springer) 23–30. doi: 10.1007/978-3-540-89619-7\_3
- Hay, G. L., Baracz, S. J., Everett, N. A., Roberts, J., and Costa, P. A., Arnold, J. C., et al. (2018). Cannabidiol treatment reduces the motivation to self-administer methamphetamine and methamphetamine-primed relapse in rats. *J. Psychopharmacol.* 32, 415–420. doi: 10.1177/0269881118799954
- hrn, A. (1999). *Discernibility and rough sets in medicine: tools and applications*. PhD thesis, Norwegian University of Science and Technology, Norway.
- Jensen, R., and Cornelis, C. (2011). "Fuzzy-rough nearest neighbour classification," in *Transactions on rough sets XIII* (Berlin Heidelberg: Springer) 56–72. doi: 10.1007/978-3-642-18302-7\_4
- Kim, J., Lee, D., and Park, E. (2021). Machine learning for mental health in social media: Bibliometric study. *J. Med. Internet Res.* 23, e24870. doi: 10.2196/24870
- Kroenke, K., Spitzer, R., and Williams, J. B. (2001). The phq-9: validity of a brief depression severity measure. *J. Gen. Intern. Med.* 16, 606–613. doi: 10.1046/j.1525-1497.2001.016009606.x
- Lee, J. L., Bertoglio, L. J., Guimares, F. S., and Stevenson, C. W. (2017). Cannabidiol regulation of emotion and emotional memory processing: relevance for treating anxiety-related and substance abuse disorders. *Br. J. Pharmacol.* 174, 3242–3256. doi: 10.1111/bph.13724
- Liu, T., Chen, Y., Zheng, Z., Wang, C., Hou, Z., Chen, C., et al. (2011). "Using evolutionary rough sets on stress prediction model by biomedical signal," in *International Conference on Machine Learning and Cybernetics, ICMLC 2011* (Guilin, China: IEEE) 319–323. doi: 10.1109/ICMLC.2011.6016702
- Liu, T.-K., Chen, Y.-P., Hou, Z.-Y., Wang, C.-C., and Chou, J.-H. (2014). Noninvasive evaluation of mental stress using by a refined rough set technique based on biomedical signals. *Artif. Intell. Med.* 61, 97–103. doi: 10.1016/j.artmed.2014.05.001
- Lynch, M. E., and Campbell, F. (2011). Cannabinoids for treatment of chronic non-cancer pain; a systematic review of randomized trials. *Br. J. Clin. Pharmacol.* 72, 735–744. doi: 10.1111/j.1365-2125.2011.03970.x
- McGuire, P., and Robson, P., Cubala, W. J., and Vasile, D., Morrison, P. D., Barron, R., et al. (2018). Cannabidiol (CBD) as an adjunctive therapy in schizophrenia: a multicenter randomized controlled trial. *Am. J. Psychiat.* 175, 225–231. doi: 10.1176/appi.ajp.2017.17030325
- Mittal, T., Gupta, P., and Chakraverty, S. (2014). "Application of rough sets in diagnosis of the depressive state of mind," in *2014 Recent Advances in Engineering and Computational Sciences (RAECS)* 1–6. doi: 10.1109/RAECS.2014.6799520
- Morin, C. M., and Carrier, J. (2021). The acute effects of the COVID-19 pandemic on insomnia and psychological symptoms. *Sleep Med.* 77, 346–347. doi: 10.1016/j.sleep.2020.06.005
- Nakamura, A. (1988). Fuzzy rough sets. *Note Multiple-Valued Logic Japan* 9, 1–8.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/frai.2023.981953/full#supplementary-material>



- Nemesure, M. D., Heinz, M. V., Huang, R., and Jacobson, N. C. (2021). Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. *Sci. Rep.* 11, 1–9. doi: 10.1038/s41598-021-81368-4
- Nomura, S., Santoso, H., and Kudo, Y. (2010). “Data mining the relationship between psychological and physiological stress measurement using rough set analysis,” in 2010 *IEEE International Conference on Systems, Man and Cybernetics*, 3622–3627. doi: 10.1109/ICSMC.2010.5641756
- Noor, M. B. T., Zenia, N. Z., Kaiser, M. S., Mamun, S. A., and Mahmud, M. (2020). Application of deep learning in detecting neurological disorders from magnetic resonance images: a survey on the detection of alzheimer’s disease, parkinson’s disease and schizophrenia. *Brain Inform.* 7, 1–21. doi: 10.1186/s40708-020-00112-2
- Pathan, M. S., Jianbiao, Z., John, D., Nag, A., and Dev, S. (2020). Identifying stroke indicators using rough sets. *IEEE Access* 8, 210318–210327. doi: 10.1109/ACCESS.2020.3039439
- Pattaraintakorn, P., and Cercone, N. (2008). Integrating rough set theory and medical applications. *Appl. Math. Lett.* 21, 400–403. doi: 10.1016/j.aml.2007.05.010
- Pawlak, Z. (1982). Rough sets. *Int. J. Comput. Infor. Sci.* 11, 341–356. doi: 10.1007/BF01001956
- Pawlak, Z. (1991). *Rough Sets: Theoretical Aspects of Reasoning About Data*. Dordrecht and Boston: Kluwer Academic Publishers. doi: 10.1007/978-94-011-3534-4
- Pawlak, Z., and Skowron, A. (1994). “Rough membership functions,” in *Advances in the Dempster-Shafer Theory of Evidence*, 251–271.
- Pawlak, Z., and Skowron, A. (2007). Rudiments of rough sets. *Inf. Sci.* 177, 3–27. doi: 10.1016/j.ins.2006.06.003
- Prud’homme, M., Cata, R., and Jutras-Aswad, D. (2015). Cannabidiol as an intervention for addictive behaviors: a systematic review of the evidence. *Substance Abuse*, 9, SART-S25081. doi: 10.4137/SART.S25081
- Rahman, R. A., Omar, K., Mohd Noah, S. A., Danuri, M. S. N. M., and Al-Garadi, M. A. (2020). Application of machine learning methods in mental health detection: A systematic review. *IEEE Access* 8, 183952–183964. doi: 10.1109/ACCESS.2020.3029154
- Ramesh, J., Keeran, N., Sagahyroon, A., and Aloul, F. (2021). “Towards validating the effectiveness of obstructive sleep apnea classification from electronic health records using machine learning,” in *Healthcare* (Multidisciplinary Digital Publishing Institute) 9, 1450. doi: 10.3390/healthcare9111450
- Russo, E. B. (2019). The case for the entourage effect and conventional breeding of clinical cannabis: No “strain,” no gain. *Front. Plant Sci.* 9, 1969. doi: 10.3389/fpls.2018.01969
- Sharma, A., and Verbeke, W. J. (2020). Improving diagnosis of depression with xgboost machine learning model and a large biomarkers dutch dataset (n = 11,081). *Front. Big Data* 3, 15. doi: 10.3389/fdata.2020.00015
- Shatte, A. B., Hutchinson, D. M., and Teague, S. J. (2019). Machine learning in mental health: a scoping review of methods and applications. *Psychol. Med.* 49, 1426–1448. doi: 10.1017/S0033291719000151
- Shusaku, N., and Kudo, Y. (2009). An application of rough set analysis to a psycho-physiological study - assessing the relation between psychological scale and immunological biomarker. *JACIII* 13, 352–359. doi: 10.20965/jaciii.2009.p0352
- Spitzer, R., Kroenke, K., Williams, J. B., and Lowe, B. (2006). A brief measure for assessing generalized anxiety disorder: The gad-7. *Arch. Intern. Med.* 166, 1092–1097. doi: 10.1001/archinte.166.10.1092
- Stepaniuk, J. (1998). *Approximation Spaces, Reducts and Representatives*. Physica-Verlag, HD, Heidelberg 109–126. doi: 10.1007/978-3-7908-1883-3\_6
- Su, C., Xu, Z., Pathak, J., and Wang, F. (2020). Deep learning in mental health outcome research: a scoping review. *Translat. Psychiatry* 10, 1–26. doi: 10.1038/s41398-020-0780-3
- Sznitman, S. R., Vulfsons, S., Meiri, D., and Weinstein, G. (2020). Medical cannabis and insomnia in older adults with chronic pain: a cross-sectional study. *BMJ Support. Palliat. Care* 10, 415–420. doi: 10.1136/bmjspcare-2019-001938
- Tariq, S., Akhtar, N., Afzal, H., Khalid, S., Mufti, M. R., Hussain, S., et al. (2019). A novel co-training-based approach for the classification of mental illnesses using social media posts. *IEEE Access* 7, 166165–166172. doi: 10.1109/ACCESS.2019.2953087
- Ware, M. A., and Ferguson, G. (2015). Review article: Sleep, pain and cannabis. *J. Sleep Diso. Ther.* 4, 1–5. doi: 10.4172/2167-0277.1000191
- Xiao, B., Huang, C., Imel, Z. E., Atkins, D. C., Georgiou, P., and Narayanan, S. S. (2016). A technology prototype system for rating therapist empathy from audio recordings in addiction counseling. *PeerJ. Comput. Sci.* 2, e59. doi: 10.7717/peerj-cs.59
- Zadeh, L. (1997). Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy Sets Syst.* 177, 111–127. doi: 10.1016/S0165-0114(97)00077-8