Using Machine Learning to Recommend Personalized Modular Treatments for Common Mental Health Disorders

Fabian Schmidt*, Karin Hammerfald†, Henrik Haaland Jahren‡, Ole André Solbakken†,
Amir H. Payberah*, and Vladimir Vlassov*
*School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden
† Department of Psychology, University of Oslo, Oslo, Norway
‡ Braive AS, Oslo, Norway
Email: *{fschm, payberah, vladv}@kth.se, †{karin.hammerfald, o.a.solbakken}@psykologi.uio.no,
‡henrik@braive.com

***Abstract*—So far, initial treatment recommendations for internet-based cognitive behavioral therapy (iCBT) decision support were mostly high-level or static. Personalized treatment recommendations could pave the way toward better treatment outcomes and adaptive treatments by leveraging information from past patients. We explore the disadvantages of multi-class recommendation and propose a modular approach using multi-label classification for treatment recommendations. Our machine learning-based treatment recommender composes treatment programs from a set of modules. It achieves a 79.02% F1-score on historically successful treatments, significantly outperforming the existing system by around 4% while offering other advantages such as interpretability and robustness. Using our recommendation as an initial starting point, clinicians can adjust the modular treatments to provide a more personalized treatment.**

***Index Terms*—Personalized treatment, Machine learning, Treatment recommendation, Internet-based cognitive behavioral therapy, Modular treatments, Common mental health disorders**

I. Introduction

Mental disorders are among the most prevalent illnesses, affecting nearly one billion people worldwide [1]. Although mental illness can be effectively treated with psychosocial interventions, only a minority of patients receive timely and effective treatment. To meet the increasing treatment need, more automated methods have been developed in the last two decades [2]. Among these, research has focused chiefly on internet-based psychotherapy [3], particularly internet-based cognitive behavioral therapy (iCBT). By now, significant evidence suggests that iCBT is efficient and effective in the treatment of mental disorders [4]–[7] and can reduce relapse rates dramatically [8]. However, more research needs to be done to understand which interventions work best for whom.

In psychotherapy, precision mental health means helping a clinician make treatment-related decisions to find the most promising path to change for a given patient using empirically based decision rules [9]. Machine learning (ML) techniques can play a crucial role in precision mental health by symptom and risk factor identification, prediction of symptom progression, and treatment personalization [10]. ML models to assist clinicians in navigating the psychotherapeutic process are called clinical decision support systems (CDSS) [11]. Personalized pre-treatment recommendations include suggestions based on pre-treatment patient features to select the optimal treatment. Literature on this topic suggests that ML can be applied to predict patients' expected response to different treatment options [12]–[14] or assign patients to matching therapists [15]. The transferability of these findings to real-life settings is questionable, and fine-grained treatment recommendations for specific interventions within one psychotherapeutic model may be more feasible in clinical practice [11].

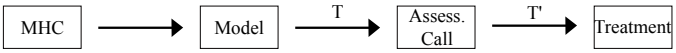

Fig. 1. Treatment recommendation process.

Braive[1] is one of the companies addressing this challenge. Braive offers a scalable, digital, on-demand solution with 11 psychotherapy treatment programs for adults suffering from common mental disorders (CMDs), including depression, anxiety, insomnia, and stress-related disorders. Braive's platform uses evidence-based iCBT techniques in a self-help or a blended treatment approach combining self-help with clinician feedback. The patients follow treatment *programs*, which comprise eight to 12 treatment *modules*. There are 32 modules in total, each representing distinct content. On average, a treatment program consists of nine modules (at least six and at most 12). Each module is designed to be completed in a week. Therapists give feedback on each completed module in written form or via a 30-minute video call.

Fig. 1 shows the process of Braive's current system for treatment program recommendation (*Braive's system)*. There are four steps: (i) mental health check (MHC), (ii) initial treat-

This work is funded by the Research Council of Norway (grant 321561) and partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP), funded by the Knut and Alice Wallenberg Foundation.

[1]http://www.braive.com

ment recommendation, (iii) assessment call, and (iv) treatment. Patients complete the MHC, which includes demographic variables and validated psychometric questionnaires to assess their mental health condition. The MHC provides *item* and *total* scores, with higher scores indicating more severe conditions. The MHC is described in Section II-A. Based on the MHC results, an initial treatment recommendation T is generated using a static decision tree, suggesting one of 17 programs for adults or adolescents. In the assessment call, the clinician presents the treatment recommendation T to the patient, and they discuss and agree on a suitable treatment T'. The patient may either follow the recommended treatment ($T = T'$) or opt for an alternative program ($T \neq T'$). Once the treatment T' is chosen, the patient begins the program, and their treatment trajectory is evaluated using general and symptom-specific questionnaires. Detailed information on monitoring treatment trajectories can be found in Section II-B.

Braive's system recommendations align with completed treatment programs in only 35.6% of cases. In the remaining cases, the treatment program is changed. Another notable limitation of Braive's system lies in its failure to consider valuable historical patient data, impeding its capacity to provide personalized recommendations due to its static nature.

As health technologies and applications make health and behavioral data available on a large scale, the ML approach seems particularly suitable to address this problem [10]. We hypothesize that training ML models can improve treatment program recommendations. To that end, we create datasets and train ML models on Braive patients' pre-treatment mental health evaluations and show that our ML models outperform Braive's system. Our findings suggest that individual symptom patterns before the beginning of treatment can be used for (i) personalizing treatment, (ii) supporting clinicians in choosing the most promising treatment option, and (iii) improving treatment outcomes.

In summary, this paper makes the following contributions:

1) We create a dataset comprising samples identified as successful treatments (Section II).
2) We train ML models for personalized treatment recommendations (Section III).
3) We evaluate and discuss the results of a multi-class vs. a more modular approach using multi-label classification with four ML architectures (Section IV and Section V).

II. Treatment Recommendation Data

This section outlines creating datasets for training personalized treatment recommendation ML models. We follow a four-step approach: (i) introducing the MHC as model input, (ii) defining treatment trajectories, (iii) merging MHC and treatment trajectories, and (iv) filtering out unsuccessful treatments and providing justification for this strategy. The subsequent parts of this section elaborate on the study design and delve into each of the four steps in comprehensive detail.

In this study, we use de-identified clinical data from 1369 patients signing up for Braive treatment between 05/2021 and 10/2022. Data analysis was carried out between 06/2022 and 12/2022. All patients gave written consent for their de-identified data to be used in routine evaluations for service monitoring and improvement.

A. Step 1: Mental Health Check (MHC)

Before assigning a treatment program, a patient completes the MHC, which comprises demographic variables (gender, age, civil status, children, occupational situation) and validated psychometric questionnaires to assess the patient's mental health.[2] The questionnaires vary in the number of items, and each item has multiple answer options with corresponding scores. Patients choose the option that best reflects their situation, with higher scores indicating more severe symptoms.

B. Step 2: Monitoring Treatment Trajectories

Once a patient begins the treatment phase, the *treatment trajectory* is evaluated using two questionnaires for general assessment: K10 [29] and PSS [27] (*general* questionnaires). Depending on the treatment program, either K10 or PSS is used, but never both. A patient's treatment trajectory τ is the sequence of *total scores* s_t for the general questionnaires (1). Total scores s_t are the sum of *item scores* $s_{t,j}$ at each t. The trajectory begins at $t = 1$ and ends at $t = M$, where M is the maximum number of modules for that treatment program. If a patient quits a treatment program, no further questionnaire scores are available from that s_t. The number of questions N in a questionnaire remains the same at each t. For instance, a patient fills out the general assessment at any t by answering 10 questions ($N = 10$), producing the item scores $\{s_{t,1}, s_{t,2}, \ldots, s_{t,10}\}$.

$$\tau = \{s_1, s_2, \ldots, s_M\}, \quad s_t = \sum_{j=1}^{N} s_{t,j}, \quad t = 1, \ldots, M \quad (1)$$

In addition to the general questionnaires K10/PSS, the patients complete *symptom-specific* questionnaires. Six symptom-specific questionnaires are employed: GAD-7, PHQ-9, SPIN, ISI, PADIS, and KEDS. They are used more sparsely, usually at the beginning of every second module. Symptom-specific questionnaires are linked to specific modules, e.g., PADIS is used along with K10 for the panic disorder program producing only the scores $\{s_1, s_3, s_5, s_7\}$.

C. Step 3: Merging MHC and Treatment Trajectories

We merge MHC scores with treatment trajectories to create supervised ML training datasets. Matching is based on patient and treatment identifiers, as well as timestamps. Exactly one MHC is required per treatment trajectory. Correct timestamp order is crucial because MHC precedes treatment. Exceptions include cases where patients start treatment without an MHC, which are excluded. Multiple MHCs can be filled by a patient, such as post-treatment assessments or when changing treatment programs. The last treatment program is typically

[2]We use PHQ-9 [16], GAD-7 [17], KEDS [18], Mini-SPIN [19], PADIS [20], ISI [21], PC-PTSD-5 [22], GAF [23], IPDS [24], BGQ [25], SCOFF [26], Perceived Stress Scale (PSS) [27], and BDDQ [28].

the most successful. Fig. 2 exemplifies non-trivial merging to end up with the desired 1:1 mapping, where multiple MHCs and treatment trajectories need to be considered.

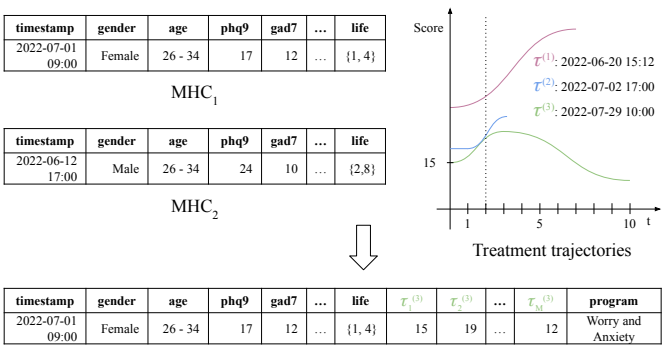

Fig. 2. Merging MHCs and corresponding treatment trajectories (τ).

D. Step 4: Successful Treatment

Only patients with positive treatment outcomes are included in the dataset to ensure the integrity of model training and prevent distortion. If the models were trained on a mixture of successful and unsuccessful MHC samples, the ground truth would become ambiguous. To address this ambiguity and maintain consistency, we employ a criterion to differentiate between successful and unsuccessful treatments. The reliable change criterion (RCC) [30], [31] serves as a classification rule [32], distinguishing whether patients have experienced significant change beyond measurement variability [31]. It has shown reliability in predicting treatment response in clinical populations [32]. Specifically, a *successful treatment* is determined by the difference in scores Δs between the initial score s_1 and the final available score s_M in the treatment trajectory surpassing the RCC threshold: $\Delta s = s_1 - s_M \geq RCC$.

We use the RCC from [31] with a 90% confidence interval (CI), derived in a four-step process. First, we determine Cronbach's alpha (α) [33], which measures the internal consistency of a test [33]. We calculate Cronbach's alpha for K10 and PSS and use normal population samples for symptom-specific questionnaires from [34]–[40]. For N number of questions in the questionnaire, average covariance between questions σ_j^2, and the overall variance of the total measured score σ_X^2, Cronbach's alpha is defined as follows.

$$\alpha = \frac{N}{N-1}\left(1 - \frac{\sum_j^N \sigma_j^2}{\sigma_X^2}\right). \tag{2}$$

In the second step, we calculate the standard deviation σ_1 on the item scores $\{s_{1,1}, s_{1,2}, \ldots, s_{1,N}\}$ at $t = 1$, with N number of questions, mean questionnaire score μ_1, and item scores $s_{1,j}$ and use it with α to compute the standard error SE in the third step as follows.

$$\sigma_1 = \frac{1}{N}\sum_{j=1}^{N}(s_{1,j} - \mu_1)^2, \tag{3}$$

$$SE = \sigma_1 \cdot \sqrt{2} \cdot \sqrt{1-\alpha}\,. \tag{4}$$

Finally, to provide a sufficient sample size to develop and test the prediction model, we multiply the z-score for the 90% CI with the SE to retrieve the $RCC = 1.645 \times SE$. A value greater than the RCC would only occur by the unreliability of measurement alone in less than 10% of times ($p < 0.1$) that two measurements are made on the same person [31]. We could and will improve the model in future iterations using a CI of 95% ($p < 0.05$) when more data is available.

III. Models for Treatment Recommendation

This section presents the models employed in our analysis, focusing on supervised learning models for tabular data tasks [41]–[43]. We categorize the training objectives into multi-class and multi-label classification, accompanied by corresponding performance metrics.

A. Training Objectives - Multi-class vs. Multi-label

When therapists followed ML-driven treatment recommendations based on pre-treatment patient characteristics in the first ten sessions, Lutz et al. [44] found an increase of 0.3 in effect size in an outpatient setting. Hence, we aim to recommend personalized treatments to patients using pre-treatment questionnaire scores from the MHC. We refer to treatment programs as T and labels at the time of ML training as y. We define the task in two ways that differ in label granularity: *multi-class* vs. *multi-label* classification.

1) Multi-class classification: The ground truth $y \in \{1, \ldots, C\}$ corresponds to a treatment program. In multi-class classification, labels are mutually exclusive: one treatment program y per MHC sample. The deviation from the ground truth is measured to determine the classifier's performance. During training, the model learns to minimize the deviation.

2) Multi-label classification: We define the task of predicting a treatment program on the modular level. Each treatment program is a composition of modules. Thus given a set of K distinct modules, a treatment program y can be represented as a binary vector $y = \{m_1, m_2, \ldots, m_K\}$, where

$$m_j = \begin{cases} 0 & \text{if module} \notin \text{treatment program} \\ 1 & \text{if module} \in \text{treatment program.} \end{cases} \tag{5}$$

This way, a treatment program is defined more granularly but irrespective of the order of the modules. For instance, suppose two treatments T_1 and T_2 with two modules m_2 and m_4 out of four total modules: $T_1 = \{m_2, m_4\}$ and $T_2 = \{m_4, m_2\}$. These two treatments will have the same binary vector representation $y = \{0, 1, 0, 1\}$.

In multi-class classification, we train a model that learns to map the MHC input to a class label $\{1, \ldots, C\}$. In multi-label classification, we train K classifiers (one classifier for each module) that each predicts whether a particular module m_j should be included, thereby transforming the approach from multi-class into multiple binary prediction tasks.

B. Machine Learning Models

We evaluate the following ML models in multi-class and multi-label classification: decision tree [45], random forest [46], and XGBoost [47]. These models have excelled on tabular data in the medical domain [41]–[43]. We also include logistic regression for multi-label classification. We employ a one-vs-the-rest (OvR) approach to produce binary modular predictions. In OvR, we fit one classifier for each class against all other classes. We use the scikit-learn library [48] to implement the models. We use Braive's system as a baseline.

C. Data Preprocessing

We transform the categorical features using one-hot encoding. We z-standardize the numeric MHC questionnaire scores using $(s_t - \mu)/\sigma$ where μ is the mean and σ the standard deviation of the questionnaire. We remove MHC and treatment trajectory pairings with fewer than two completed modules. Pairings below the successful treatment threshold are also excluded. We split the data into train, validation, and test sets.

IV. Experimental Results

In this section, we present the datasets and the metrics we use for our analysis and evaluate the results.

A. Datasets

We experiment on three datasets:

1) *D1*: We merge MHC total scores and general (K10/PSS) treatment trajectories. There are a total of 1528 patient samples. We then exclude those instances where the treatment is unsuccessful according to the success measure. Of the 1528 patient samples, 579 remain.
2) *D2*: We merge MHC item scores and general treatment trajectories. We explore whether item scores serve as better input for the model to make predictions. There are also 579 patient samples, though the number of features increases as there are more MHC item than total scores.
3) *D3*: We merge MHC total scores and symptom-specific treatment trajectories. We explore whether the symptom-specific questionnaires are a more sensitive measure of the symptoms and provide better results than general ones. There are 343 patient samples in this dataset.

We use the RCCs to distinguish between successful and unsuccessful treatments. Cronbach's alphas α, standard deviations σ and standard error of change (*SE*) to compute the RCCs are shown in Table I. General questionnaires are above, and symptom-specific questionnaires are below the dashed line. The RCC represents the minimum reduction needed from s_1 to s_M to fulfill the requirements to be a successful treatment defined in Section II-D. For instance, a patient with a K10 treatment trajectory must improve by at least 7.72 from the first score s_1 to the last score s_M to be included in the dataset.

TABLE I
Reliable change criteria for questionnaires.

Questionnaire	α	σ_1	*SE*	*RCC*
K10	0.85	7.17	3.94	7.72
PSS	0.83	6.50	3.78	7.41
GAD-7 (D)	0.89	3.41	1.60	2.63
PHQ-9 (UK)	0.83	6.47	3.77	6.20
SPIN (UK)	0.94	9.30	3.22	5.30
ISI (N)	0.90	6.20	2.77	4.56
PSS (D)	0.85	6.41	3.51	5.77
PADIS (AUS)	0.86	2.74	1.45	2.38
KEDS (SWE)	0.75	6.18	4.37	7.19

TABLE II
Test set results in percent (%) for multi-class prediction on the total scores dataset (D1).

Model	*Prec.*	*Recall*	*F1*
Braive's system	**42.21**	35.60	**34.84**
Decision Tree	30.75	34.78	31.46
Random Forest	31.99	**39.13**	**34.86**
XGBoost	30.08	36.52	32.41

B. Evaluation Metrics

We perform 5-fold cross-validation for multi-class and multi-label classification and evaluate test set results using precision, recall, and F1 to compare multi-class and multi-label. Additionally, we evaluate multi-label results with Hamming loss (HL) [49] and the Jaccard index [50], also known as the Intersection over Union (IoU), to provide a more comprehensive evaluation. We plot the receiver operating characteristic (ROC) curves for multi-label results and provide the area under the curve (AUC). The metrics for supervised classification are computed in a OvR manner for each class (see Section III-B). We present the micro and weighted averages for the metrics.

C. Multi-class Treatment Program Prediction

We train the models mentioned in Section III first as multi-class and second as multi-label classification. We measure the results using the metrics from Section IV-B. Table II shows the test set results for the multi-class classification on D1. The model predicts the treatment program, e.g., Depression and Anxiety or Depression and Sadness. The F1-score of Braive's system (34.84) is only matched by random forest (34.86). The decision tree (31.46) and XGBoost (32.41) fail to match the performance of Braive's system. Overall, we observe that the multi-class performance is not convincing.

To investigate the cause, suppose three treatment programs: T_a, T_b, and T_c. The predictive performance of a supervised ML model is evaluated by comparing the prediction $\hat{y}$ to the ground truth y. If we evaluate a multi-class classification, $\hat{y}$ and y take the form of one of the treatments T_a, T_b and T_c, usually represented as a dummy vector of shape $1 \times |C|$ where $|C|$ is the number of classes. A classifier is trained and predicts one of the three treatments. However, treating each misclassification $\hat{y} \neq y$ equally wrong is problematic because there is an inherent similarity between some treatment programs that share modules.

TABLE III
TEST SET RESULTS IN PERCENT (%) FOR MULTI-LABEL CLASSIFICATION ON D1-D3.

Model	Total scores (D1)					Item scores (D2)					Symptom-specific (D3)				
	HL ↓	*IoU* ↑	*Prec.* ↑	*Recall* ↑	*F1* ↑	*HL* ↓	*IoU* ↑	*Prec.* ↑	*Recall* ↑	*F1* ↑	*HL* ↓	*IoU* ↑	*Prec.* ↑	*Recall* ↑	*F1* ↑
Braive's system	14.68	64.91	79.15	74.39	75.36	14.68	64.91	79.15	74.39	75.36	13.82	66.76	**81.28**	76.27	77.02
Log. Regression	11.74	70.30	79.49	78.47	77.68	13.46	67.87	77.75	76.91	76.61	12.42	69.97	79.30	76.35	77.17
Decision Tree	12.93	69.02	75.39	77.97	75.87	14.65	65.11	74.67	74.43	73.66	11.44	69.56	74.21	76.02	74.90
Random Forest	**10.90**	**71.54**	81.02	**80.13**	78.66	12.27	68.20	78.56	77.44	75.89	11.22	**71.18**	77.96	**78.10**	**77.22**
XGBoost	11.25	71.37	**81.24**	79.63	**79.02**	**12.18**	**69.02**	**80.87**	**77.96**	**77.69**	**11.01**	70.61	74.42	77.77	75.83

The similarity is better captured if we consider the modular representation of a treatment program. Now the treatment programs are represented as $T_a = [0, 1, 1, 1, 1]$, $T_b = [0, 1, 1, 0, 1]$ and $T_c = [1, 0, 0, 1, 0]$. Treatment programs can be of a different label but very similar on a modular level. For instance, T_a and T_b share three of the five modules. T_b and T_c share no modules. This could lead to distortions in a multi-class setting when training on similar MHC input with two classes T_a and T_b. The model is forced to distinguish between the two classes, but only marginal differences exist in the input.

This is reflected in a quantitative evaluation. When evaluating the classifier, either T_a, T_b, or T_c is predicted. Suppose the ground truth is $y = T_a$. In multi-class classification, we get the F1-scores 1, 0, and 0, respectively. In multi-label classification, we get the F1-scores 1, 0.75, and 0.25, respectively.

D. Multi-label Modular Treatment Prediction

Instead of using programs as labels, we redefine labels as modules, shifting the training objective to multi-label classification while utilizing the same dataset (D1). Table III presents the test set results for modular treatment recommendation, with Braive's system achieving a 75.36% F1-score. ML models surpass the baseline, with logistic regression (77.68%) outperforming the decision tree (75.87%). Random forest and XGBoost outperform Braive's system in all metrics, with XGBoost achieving the highest F1-score at 79.02%, followed by random forest at 78.66%.

Fig. 3 displays the ROC curves for multi-label treatment recommendations based on total scores. The micro-averaged AUC is reported for each ML model, demonstrating superior performance compared to a random classifier. Random forest and logistic regression slightly outperform XGBoost with an AUC of 0.94, while the decision tree performs the worst at 0.92. Similar results are obtained for item scores, where random forest achieves the highest AUC of 0.94, followed by logistic regression (0.93) and the decision tree (0.90).

To compare the multi-class and multi-label approaches, we map the multi-class labels to modular representations of treatment programs and compare $\hat{y}$ to y. The modular representation is limited to binary vectors representing predefined modular treatments, with maximum $|C|$ distinct binary vectors.

Table IV presents the results for modular evaluation, showing significant improvements in all metrics compared to multi-class performance in Table II. The best-performing model achieves an F1-score of 76.00%, more than doubling the previous result of 34.86%. It is important to note that these improvements stem from the evaluation approach change rather than an actual performance increase. Nonetheless, even in the modular evaluation, multi-label models outperform multi-class models. The modular approach promotes personalized treatment, offering robust and interpretable models trained for each module. The models focus on learning signals directly related to specific modules, enhancing customization during retraining.

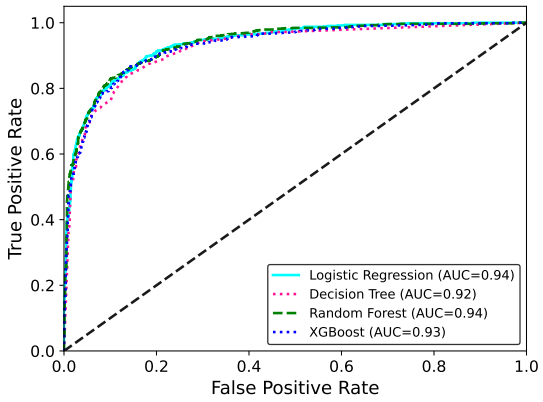

Fig. 3. Micro-averaged ROC for modular predictions on MHC total scores.

TABLE IV
TEST SET RESULTS IN PERCENT (%) FOR MULTI-CLASS WITH MODULAR EVALUATION (D1).

Model	*HL* ↓	*IoU* ↑	*Prec.* ↑	*Recall* ↑	*F1* ↑
Braive's system	**14.40**	65.63	**80.17**	74.98	**76.00**
Decision Tree	15.30	65.66	73.67	76.12	73.95
Random Forest	14.51	**67.09**	73.66	78.36	75.66
XGBoost	14.78	66.81	73.25	**78.11**	75.17

The recommendations do not reflect the order of modules, except for naturally ordered modules like introduction modules. Future work will explore the ordering of modules using approaches similar to [51]. Ordering remains the responsibility of clinicians for the models presented in this study.

V. DISCUSSION

A. Personalized Treatment Recommendations

We present models for personalized treatment recommendations. We show that multi-class predictions of a treatment program are sensitive to the evaluation scheme. Furthermore,

multi-class recommendations lack the granularity to provide personalized recommendations. Multi-label recommendations, on the other hand, allow for the composition of treatments tailored to the individual needs of each patient. Our study can be placed within the broader context of using ML approaches for mental health treatment predictions. Other researchers used prognostic indices (see [52] for an overview) or compared treatment outcome estimates for different treatment alternatives (e.g., [53], [54]) or the expected change for different treatment strategies (e.g., [55]–[57]) to predict optimal treatment. Different methods were used for treatment selection. We, in contrast, predict more fine-grained treatment recommendations as modular components focused on interventions rather than strategies. On a technical level, this means higher complexity in terms of the number of classes: 11 treatment programs in multi-class and 32 in multi-label classification.

B. Item vs. Total Scores

We train and evaluate multi-label models for the item (D2) and total scores (D1). The results are shown in Table III. We hypothesized that training on item scores might improve performance because item scores contain more information than total scores. There are myriad possibilities to arrive at the same total score. Hence, knowing the degree to which each question contributed could be valuable. We observe no improvements on MHC item scores compared to total scores. Models trained on the total scores (D1) outperform their counterparts trained on item scores (D2) on all metrics. Total scores seem to contain enough information for the models already to learn meaningful features.

C. General vs. Symptom-specific

As the RCCs are the basis for subsetting our datasets, the quality of the datasets is directly affected by the quality of the measure for treatment trajectories. Symptom-specific questionnaires are more specific for symptom development, thus functioning as better indicators for the treatment trajectory of a patient. Hence, symptom-specific scores could distinguish better which treatments were successful and which were not. We expect models on the symptom-specific dataset D3 to perform at least as well as those trained on the general questionnaire datasets D1 and D2, possibly even better. We see improvements in the F1-score for three models when comparing D2 to D3 (Table III). The XGBoost model trained on D2 performs better. Compared to D1, the D3 models perform worse. The baseline improves from 75.36% to 77.02%, indicating that Braive's system performance is better reflected in symptom-specific scores.

D. Inclusion Criteria of Patient Samples

We exclude patient samples classified as unsuccessful based on the RCC. The dataset decreases from 1528 to 579 samples. Already operating with a small dataset, we inhibit the ability to learn from more samples. We argue that unsuccessful treatments should not be considered recommendations as the treatment did not lead to the desired decrease in psychometric scores. Similar strategies were conducted by [12]. They only included 50% of the patients with the strongest PAIs (Personalized Advantage Index), a measure to indicate the difference between the outcome predictions of two treatment approaches.

We include all patients with at least two completed modules and a degree of improvement between the first and last completed module greater than or equal to the RCC. We calculate RCCs for each primary outcome measure and the K10 by applying a change equal to $z = 1.645$ on the basis of a standard deviation unit. However, alternative strategies for defining the training dataset could be considered if data quantity was not the primary concern. Studies examining the dose-response relationship in psychotherapy in routine care settings recommend dosages between four and 26 sessions for less severe symptoms and short-term treatments [58]. As even higher dosages are needed for severe psychopathology and open-ended treatments [59], it is debatable whether a successful treatment outcome after only two sessions in our sample is attributable to treatment. Furthermore, a common definition of psychotherapy treatment response follows [30] definition of clinically significant change ($RCC < 1.96 \times \sigma$ and pass halfway cut-off to normal population compared to patient population), which is a stricter change criterion.

Assuming that treatment outcome in more complex psychiatric problems (e.g., major depressive disorder plus comorbid OCD) is disproportionately often unsuccessful in our data, such patients would not adequately be represented in the training data. Using routine care data instead of high-quality research data to train our model may result in clinical decisions that do not align with scientific evidence, perpetuating suboptimal choices [60]. Yet, undue reliance on research data in algorithm development may neglect crucial clinical knowledge and potentially compromise the quality of care [61].

VI. Conclusion and Future Work

We presented ML models for treatment recommendations based on pre-therapy patient assessment. We evaluated two classification objectives, i.e., multi-class vs. multi-label, and four architectures, i.e., logistic regression, decision tree, random forest, and XGBoost. Using multi-label classification, we improved the baseline, i.e., Braive's system, on F1-scores in all experiments. We explored the role of the evaluation scheme in multi-class and multi-label classification. Multi-label outperformed multi-class models in modular evaluations. Additionally, the modular approach is preferable because it offers more granular models. This paper provides the groundwork for implementing a new treatment recommendation model in production at Braive. This treatment selection model can support decision-making by providing data-informed predictions, which can be adapted by continuously monitoring the patient's condition throughout treatment.

References

[1] S. Dattani, H. Ritchie, and M. Roser, "Mental health," *Our World in Data*, 2021, https://ourworldindata.org/mental-health.

[2] G. Andersson, "Internet interventions: Past, present and future," *Internet interventions : the application of information technology in mental and behavioural health*, vol. 12, pp. 181–188, 2018.

[3] D. C. Mohr, K. R. Weingardt, M. Reddy, and S. M. Schueller, "Three problems with current digital mental health research ... and three things we can do about them," *Psychiatric services*, vol. 68, no. 5, pp. 427–429, 2017.

[4] S. Gainsbury and A. Blaszczynski, "A systematic review of internet-based therapy for the treatment of addictions," *Clinical psychology review*, vol. 31, no. 3, pp. 490–498, 2011.

[5] V. Spek, P. Cuijpers, I. Nyklíček, H. Riper, J. Keyzer, and V. Pop, "Internet-based cognitive behaviour therapy for symptoms of depression and anxiety: a meta-analysis," *Psychological medicine*, vol. 37, no. 3, pp. 319–328, 2007.

[6] G. Andrews, A. Basu, P. Cuijpers, M. Craske, P. McEvoy, C. English, and J. Newby, "Computer therapy for the anxiety and depression disorders is effective, acceptable and practical health care: an updated meta-analysis," *Journal of anxiety disorders*, vol. 55, pp. 70–78, 2018.

[7] P. Carlbring, G. Andersson, P. Cuijpers, H. Riper, and E. Hedman-Lagerlöf, "Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis," *Cognitive behaviour therapy*, vol. 47, no. 1, pp. 1–18, 2018.

[8] F. Holländare, S. Johnsson, M. Randestad, M. Tillfors, P. Carlbring, G. Andersson, and I. Engström, "Randomized trial of internet-based relapse prevention for partially remitted depression," *Acta Psychiatrica Scandinavica*, vol. 124, no. 4, pp. 285–294, 2011.

[9] W. Lutz, J. Rubel, B. Schwartz, V. Schilling, and A.-K. Deisenhofer, "Towards integrating personalized feedback research into clinical practice: Development of the trier treatment navigator (ttn)," *Behaviour Research and Therapy*, vol. 120, p. 103438, 07 2019.

[10] A. Thieme, D. Belgrave, and G. Doherty, "Machine learning in mental health: A systematic review of the hci literature to support the development of effective and implementable ml systems," *ACM transactions on computer-human interaction*, vol. 27, no. 5, pp. 1–53, 2020.

[11] W. Lutz, D. Zimmermann, V. N. L. S. Müller, A.-K. Deisenhofer, and J. A. Rubel, "Randomized controlled trial to evaluate the effects of personalized prediction and adaptation tools on treatment outcome in outpatient psychotherapy: study protocol," *BMC psychiatry*, vol. 17, no. 1, pp. 306–306, 2017.

[12] B. Schwartz, Z. Cohen, J. Rubel, D. Zimmermann, W. Wittmann, and W. Lutz, "Personalized treatment selection in routine care: Integrating machine learning and statistical algorithms to recommend cognitive behavioral or psychodynamic therapy," *Psychotherapy Research*, vol. 31, pp. 1–19, 05 2020.

[13] S. C. van Bronswijk, R. J. DeRubeis, L. H. J. M. Lemmens, F. P. M. L. Peeters, J. R. Keefe, Z. D. Cohen, and M. J. H. Huibers, "Precision medicine for long-term depression outcomes using the personalized advantage index approach: cognitive therapy or interpersonal psychotherapy?" *Psychological medicine*, vol. 51, no. 2, pp. 279–289, 2021.

[14] L. Lorenzo-Luaces, R. J. DeRubeis, A. van Straten, and B. Tiemens, "A prognostic index (pi) as a moderator of outcomes in the treatment of depression: A proof of concept combining multiple variables to inform risk-stratified stepped care models," *Journal of affective disorders*, vol. 213, pp. 78–85, 2017.

[15] J. Delgadillo, J. Rubel, and M. Barkham, "Towards personalized allocation of patients to therapists," *Journal of consulting and clinical psychology*, vol. 88, no. 9, pp. 799–808, 2020.

[16] K. Kroenke and R. L. Spitzer, "The phq-9: a new depression diagnostic and severity measure," *Psychiatric annals*, vol. 32, no. 9, pp. 509–515, 2002.

[17] R. L. Spitzer, K. Kroenke, J. B. Williams, and B. Löwe, "A brief measure for assessing generalized anxiety disorder: the gad-7," *Archives of internal medicine*, vol. 166, no. 10, pp. 1092–1097, 2006.

[18] A. Besèr, K. Sorjonen, K. Wahlberg, U. Peterson, Å. Nygren, and M. Åsberg, "Construction and evaluation of a self rating scale for stress-induced exhaustion disorder, the karolinska exhaustion disorder scale," *Scandinavian journal of psychology*, vol. 55, no. 1, pp. 72–82, 2014.

[19] K. M. Connor, K. A. Kobak, L. E. Churchill, D. Katzelnick, and J. R. Davidson, "Mini-spin: A brief screening assessment for generalized social anxiety disorder," *Depression and anxiety*, vol. 14, no. 2, pp. 137–140, 2001.

[20] P. J. Batterham, A. J. Mackinnon, and H. Christensen, "The panic disorder screener (padis): development of an accurate and brief population screening tool," *Psychiatry research*, vol. 228, no. 1, pp. 72–76, 2015.

[21] C. M. Morin, *Insomnia: Psychological assessment and management.* Guilford press, 1993.

[22] A. Prins, M. J. Bovin, D. J. Smolenski, B. P. Marx, R. Kimerling, M. A. Jenkins-Guarnieri, D. G. Kaloupek, P. P. Schnurr, A. P. Kaiser, Y. E. Leyva, and Q. Q. Tiet, "The primary care ptsd screen for dsm-5 (pc-ptsd-5): development and evaluation within a veteran primary care sample," *Journal of general internal medicine*, vol. 31, no. 10, pp. 1206–1211, 2016.

[23] R. C. Hall, "Global assessment of functioning: a modified scale," *Psychosomatics*, vol. 36, no. 3, pp. 267–275, 1995.

[24] D. Langbehn, B. Pfohl, S. Reynolds, L. Clark, M. Battaglia, L. Bellodi, R. Cadoret, W. Grove, P. Pilkonis, and P. Links, "The iowa personality disorder screen: Development and preliminary validation of a brief screening interview," *Journal of personality disorders*, vol. 13, no. 1, pp. 75–89, 1999.

[25] K. M. Shear, C. T. Jackson, S. M. Essock, S. A. Donahue, and C. J. Felton, "Screening for complicated grief among project liberty service recipients 18 months after september 11, 2001," *Psychiatric Services*, vol. 57, no. 9, pp. 1291–1297, 2006.

[26] J. Morgan, F. Reid, and H. Lacey, "The scoff questionnaire: Assessment of a new screening tool for eating disorders," *British Medical Journal (Clinical research ed.)*, vol. 319, no. 7223, pp. 1467–8, 2000.

[27] S. Cohen, T. Kamarck, and R. Mermelstein, "A global measure of perceived stress," *Journal of health and social behavior*, vol. 24, no. 4, pp. 385–396, 1983.

[28] S. G. Mancuso, N. P. Knoesen, and D. J. Castle, "The dysmorphic concern questionnaire: A screening measure for body dysmorphic disorder," *Australian & New Zealand Journal of Psychiatry*, vol. 44, no. 6, pp. 535–542, 2010.

[29] R. C. Kessler, G. Andrews, L. J. Colpe, E. Hiripi, D. K. Mroczek, S.-L. Normand, E. E. Walters, and A. M. Zaslavsky, "Short screening scales to monitor population prevalences and trends in non-specific psychological distress," *Psychological medicine*, vol. 32, no. 6, pp. 959–976, 2002.

[30] N. S. Jacobson and P. Truax, "Clinical significance: A statistical approach to defining meaningful change in psychotherapy research," *Journal of consulting and clinical psychology*, vol. 59, no. 1, pp. 12–19, 1991.

[31] C. Evans, F. Margison, and M. Barkham, "The contribution of reliable and clinically significant change methods to evidence-based mental health," *BMJ Mental Health*, vol. 1, no. 3, pp. 70–72, 1998.

[32] W. Lutz, B. Schwartz, and J. Delgadillo, "Measurement-based and data-informed psychological therapy," *Annual Review of Clinical Psychology*, vol. 18, pp. 71–98, 2022.

[33] L. J. Cronbach, "Coefficient alpha and the internal structure of tests," *psychometrika*, vol. 16, no. 3, pp. 297–334, 1951.

[34] B. Löwe, O. Decker, S. Müller, E. Brähler, D. Schellberg, W. Herzog, and P. Herzberg, "Validation and standardization of the generalized anxiety disorder screener (gad-7) in the general population," *Medical Care*, vol. 46, no. 3, pp. 266–274, 2008.

[35] I. Cameron, J. Crawford, K. Lawton, and I. Reid, "Psychometric comparison of phq-9 an hads for measuring depression severity in primary care," *The British journal of general practice: the journal of the Royal College of General Practitioners*, vol. 58, no. 546, pp. 32–6, 2008.

[36] K. Connor, J. Davidson, L. Churchill, A. Sherwood, E. Foa, and R. Weisler, "Psychometric properties of the social phobia inventory (spin)," *The British journal of psychiatry: the journal of mental science*, vol. 176, pp. 379–86, 2000.

[37] J. Filosa, P. Omland, K. Langsrud, K. Hagen, M. Engstrøm, O. K. Drange, A. Knutsen, E. Brenner, H. Kallestad, and T. Sand, "Validation of insomnia questionnaires in the general population: The nord-trøndelag health study (hunt)," *Journal of Sleep Research*, vol. 30, 10 2020.

[38] E. Klein, E. Brähler, M. Dreier, L. Reinecke, K. Müller, G. Schmutzer, K. Wölfling, and M. Beutel, "The german version of the perceived stress scale – psychometric characteristics in a representative german community sample," *BMC Psychiatry*, vol. 16, 05 2016.

[39] P. Batterham, A. Mackinnon, and H. Christensen, "The panic disorder screener (padis): Development of an accurate and brief population screening tool," *Psychiatry Research*, vol. 228, 04 2015.

[40] R. Persson, K. Österberg, N. Viborg, P. Jönsson, and A. Tenenbaum, "Two swedish screening instruments for exhaustion disorder: Cross-sectional associations with burnout, work stress, private life stress, and

personality traits," *Scandinavian journal of public health*, vol. 45, p. 381–388, 2017.

[41] Z. Xu and Z. Wang, "A risk prediction model for type 2 diabetes based on weighted feature selection of random forest and xgboost ensemble classifier," in *2019 eleventh international conference on advanced computational intelligence (ICACI)*. IEEE, 2019, pp. 278–283.

[42] H. Xu, H. Wang, C. Yuan, Q. Zhai, X. Tian, L. Wu, and Y. Mi, "Identifying diseases that cause psychological trauma and social avoidance by gcn-xgboost," *BMC bioinformatics*, vol. 21, pp. 1–16, 2020.

[43] P. Zhang, F. Li, R. Zhao, R. Zhou, L. Du, Z. Zhao, X. Chen, and Z. Fang, "Real-time psychological stress detection according to ecg using deep learning," *Applied Sciences*, vol. 11, no. 9, p. 3838, 2021.

[44] W. Lutz, A.-K. Deisenhofer, J. Rubel, B. Bennemann, J. Giesemann, K. Poster, and B. Schwartz, "Prospective evaluation of a clinical decision support system in psychological therapy," *Journal of consulting and clinical psychology*, vol. 90, no. 1, pp. 90–106, 2022.

[45] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, pp. 81–106, 1986.

[46] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[47] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[49] R. W. Hamming, "Error detecting and error correcting codes," *The Bell system technical journal*, vol. 29, no. 2, pp. 147–160, 1950.

[50] A. H. Murphy, "The finley affair: A signal event in the history of forecast verification," *Weather and forecasting*, vol. 11, no. 1, pp. 3–20, 1996.

[51] K. C. Fernandez, A. J. Fisher, and C. Chi, "Development and initial implementation of the dynamic assessment treatment algorithm (data)," *PLOS ONE*, vol. 12, no. 6, pp. 1–16, 2017.

[52] A. M. Chekroud, J. Bondar, J. Delgadillo, G. Doherty, A. Wasil, M. Fokkema, Z. Cohen, D. Belgrave, R. DeRubeis, R. Iniesta, D. Dwyer, and K. Choi, "The promise of machine learning in predicting treatment outcomes in psychiatry," *World Psychiatry*, vol. 20, no. 2, pp. 154–170, 2021.

[53] R. DeRubeis, Z. Cohen, N. Forand, J. Fournier, L. Gelfand, and L. Lorenzo-Luaces, "The personalized advantage index: Translating research on prediction into individualized treatment recommendations. a demonstration," *PloS one*, vol. 9, no. 1, pp. 1–8, 2014.

[54] W. Lutz, C. Leach, M. Barkham, M. Lucock, W. Stiles, C. Evans, R. Noble, and S. Iveson, "Predicting change for individual psychotherapy clients based on their nearest neighbors," *Journal of consulting and clinical psychology*, vol. 73, no. 5, pp. 904–13, 2005.

[55] J. A. Rubel, A. J. Fisher, K. Husen, and W. Lutz, "Translating person-specific network models into personalized treatments," *Psychotherapy and psychosomatics*, vol. 87, no. 4, pp. 249–251, 2018.

[56] M. Y. Ng, J. L. Schleider, R. L. Horn, and J. R. Weisz, *Psychotherapy for children and adolescents: From efficacy to effectiveness, scaling, and personalizing.*, 7th ed. Hoboken, NJ: Wiley, 2021.

[57] J. M. G. Penedo, B. Schwartz, J. Giesemann, J. A. Rubel, A.-K. Deisenhofer, and W. Lutz, "For whom should psychotherapy focus on problem coping? a machine learning algorithm for treatment personalization," *Psychotherapy Research*, vol. 32, no. 2, pp. 151–164, 2022, pMID: 34034627.

[58] L. Robinson, J. Delgadillo, and S. Kellett, "The dose-response effect in routinely delivered psychological therapies: A systematic review," *Psychotherapy research*, vol. 30, no. 1, pp. 79–96, 2020.

[59] M. Nordmo, J. T. Monsen, P. A. Høglend, and O. A. Solbakken, "Investigating the dose-response effect in open-ended psychotherapy," *Psychotherapy research*, vol. 31, no. 7, pp. 859–869, 2021.

[60] D. S. Char, M. D. Abràmoff, and C. Feudtner, "Identifying ethical considerations for machine learning healthcare applications," *American journal of bioethics*, vol. 20, no. 11, pp. 7–17, 2020.

[61] J. J. Fenton, S. H. Taplin, P. A. Carney, L. Abraham, E. A. Sickles, C. D'Orsi, E. A. Berns, G. Cutter, R. E. Hendrick, W. E. Barlow, and J. G. Elmore, "Influence of computer-aided detection on performance of screening mammography," *The New England journal of medicine*, vol. 356, no. 14, pp. 1399–1409, 2007.