DAIsy
Developing AI ecosystems improving diagnosis and care of mental diseases
DAIsy
24 partners
5 countries
ITEA Call 2021
Smart health

# DAIsy – Developing AI ecosystems improving diagnosis and care of mental diseases

ITEA 4 – 21016

# Work package 4 (WP4) : AI Technology Development

## Deliverable 4.2 : Develop tools for AI Uncertainty Quantification and Model Evaluation (doc & software)

| | |
|---|---|
| Document type | : Deliverable |
| Document version | : No. 1 |
| Document Preparation Date | : August 2025 |
| Classification | : Confidential |
| Due Date | : October 2025 |

# Contents

# I. Introduction

Artificial intelligence (AI)- based solutions are being increasingly integrated into the workflow of clinicians and enabling certain aspects of healthcare to be more accessible. Even though high-performance of models are one of the criteria for selecting and validating the candidate models for deployment, performance metrics are not enough by themselves. For an AI-model to be deemed trustworthy and dependable by clinicians they should be also be at least locally explainable, and their performance and working be reliable. Reliability of AI models not only depend on what features most influenced the AI models' decisions (AI Explainability), but also how the robust the models are across different problem complexities, be it with regards to learning from data of suboptimal quality (limited number of instances, significant amount of missingness, presence of bias) or  due to the algorithm's own instability (due to running into local minima, being operational across only a narrow range of hyperparameters, sensitive to variation of certain types of data, numerical error), or the due to the difference in the development and deployment environment and the difference in availability of resources in both. Using uncertainty quantification and model evaluation tools we can assess the robustness of AI models, and identify the sources of bias and potential vulnerabilities(Begoli, Bhattacharya, and Kusnezov 2019; Caldeira and Nord 2021; Varoquaux and Colliot 2023). In this deliverable document we discuss only those evaluation and uncertainty quantification metrices which have been applied by DAIsy work-package 4 (WP4) partners only, and is thus non-exhaustive.

# II. Model evaluation

Model evaluation metrics express how good or reliable a model is, in terms of their performance for a task. They are used for not only model selection, but even for model building stages, such as for hyperparameter optimization and feature selection (in statistical learning) (Raschka 2018). Based on the complexities of the input data and the interest of domain experts and end-users, selection of model performance metrics vary across use-cases, even if the technical task is the same. AI-model developers and users often compare multiple performance metrics to investigate the reliability of a model, as different performance metric capture different aspects of the model's decision(Rainio, Teuho, and Klén 2024; Raschka 2018; Varoquaux and Colliot 2023). However, further complications arise when two or more selected performance metrics are contradictory to each other, for example, high accuracy but low F-1 score, or low classwise accuracy for one class but high for another, in multi-class classification tasks.

Performance metrics for statistical models, 'traditional' machine learning, and deep learning models, for classification or regression tasks (including for object or event detection, prediction, event-to-date prediction) or clustering tasks differ considerably from that required for evaluating generative AI models, such quality of text generated by Large language models (LLMs) as summary of a session between patient and therapist, or a magnetic resonance or computed tomography (clinical) image reconstructed by generative adversarial networks (GANs)(Park et al. 2023; Park, Han, and Lee 2024).

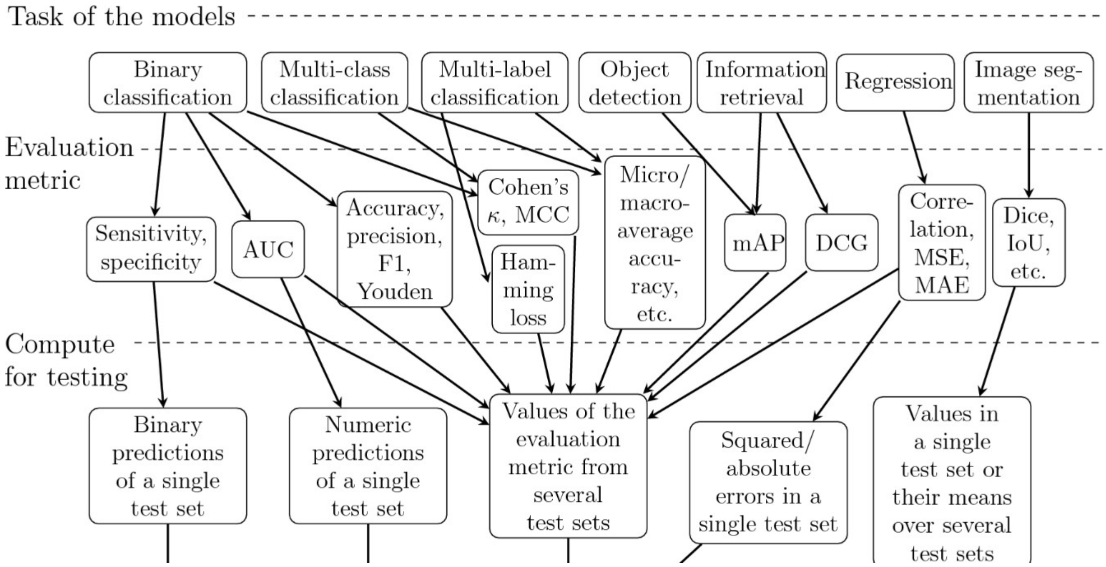

Task of the models
Binary classification
Multi-class classification
Multi-label classification
Object detection
Information retrieval
Regression
Image segmentation
Evaluation metric
Sensitivity, specificity
AUC
Accuracy, precision, F1, Youden
Cohen's κ, MCC
Hamming loss
Micro/ macro-average accuracy, etc.
mAP
DCG
Correlation, MSE, MAE
Dice, IoU, etc.
Compute for testing
Binary predictions of a single test set
Numeric predictions of a single test set
Values of the evaluation metric from several test sets
Squared/ absolute errors in a single test set
Values in a single test set or their means over several test sets

*Fig- 1: Evaluation metrics per task (excerpt from Fig.3 of Rainio O, Teuho J et al 2024)*

In this section we discuss the role of performance metrics in AI/ML model evaluation and thereby inferring results from the models.

## A. Evaluating statistical, machine learning, and deep learning models

Within DAIsy the following evaluation metrices have been used on machine learning (including deep learning) models.

### 1. Statistical metrics

a. **t-tests** is used to evaluate if the difference between two groups is statistically significant. It works under the assumption that variable being compared across the group is normally distributed and its variability is same across the groups(Nikolaou 2016).

b. **Analysis of Variance (ANOVA)** is used to evaluate if the difference between three or more groups are statistically significant, based on the group means of a single variable. Thus it is a more generalizable variant of *t-test(Nikolaou 2016)*.

c. **Multivariate ANOVA (MANOVA)** is similar to ANOVA, except the group means are based on multiple dependent variables.

d. **Correlation coefficients** gives the measure of strength of linear association between two variables that are independent, continuous and satisfy homogeneity of variance criteria. While Pearson coefficient require that the variable is normally distributed, Spearman is applicable for variables that are not normally distributed(Nikolaou 2016).

### 2. Classification performance metrics

a. **Accuracy**: measures the proportion of correct predictions out of all predictions made. It reflects model performance at a single classification threshold (which is by default 0.5 in Scikit-learn's prediction functions). However, focusing solely on accuracy can be misleading if the dataset is imbalanced.

b. **Sensitivity, specificity, precision, Area-under the ROC curve**:

- **True positive rate (TPR)**, also called sensitivity or recall, measures how effectively the model identifies positive instances $TPR = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP)+F\quad Negatives\ (FN)}$ .
- **False positive rate (FPR)** indicates how often the model incorrectly labels negative instances as positive and is given by FPR, where,

$FPR = \frac{False\ Positives\ (FP)}{False\ Positives\ (TP)+True\ Negatives\ (TN)}$.

- **Specificity or true negative rate (TNR)** is the measure of how well a model is able to identify true negatives (TN), i.e., $TNR = \frac{TN}{TN+F}$.
- **Area under the ROC curve (AUC)** captures the model's ability to distinguish between positive and negative classes at various threshold settings, providing a single value (ranging from 0 to 1) that summarizes this discriminative capability—higher AUC values indicate stronger performance. Consequently, the AUC balances the trade-off between TPR and FPR across different thresholds. It is also referred to as the C-statistic.
- **Precision or positive predictive value (PPV)** is given by $PPV = \frac{TP}{TP+FP}$.
- **F1-score** is the harmonic mean of precision and recall, where a score of 0 indicates worst and that of 1 indicates the best performance.

c. **Top k accuracy:** It is a metric similar to accuracy, except the classification is deemed correct as long as the prediction probability score of the correct class is among the *k* top probability scores.

d. **Classwise accuracy and balanced accuracy:** In a multi-class classification problem instead of overall accuracy, the classification accuracy per class is evaluated. Balanced accuracy (BA/ bAcc) is given by

$BA = \frac{Sensitivity+Specificity}{2}$.

The evaluation metrices sensitivity, specificity, precision, AUC, F1-score, BA, and Class-wise accuracy can account for inherent class imbalance. However while the first five of these are more suited for binary class classification, the last two and top k accuracy can handle multi-class classification.

TU/e used class-wise accuracy corresponding to the class with minimum number of samples (most difficult class), for hyperparameter selection. Thereafter, class-wise accuracies, macro-averaged accuracy (average of all classwise accuracies), class-wise AUC and macro-averaged classwise AUC were used to evaluate the discrimination performance of classifier models that trained on the baseline data of Eating Disorder patients from GGz OB. AMC used predictive performance metrices like bAcc, AUROC, sensitivity and specificity to report the performance of their models for antidepressant response prediction for depression (Poirot n.d.) and in prediction of response to methylphenidate for ADHD(Chen et al. 2025).
ARD Group used balanced accuracy during nested cross-validation to select model hyper-parameters on the imbalanced bipolar vs. unipolar dataset. Final performance was reported with class-wise recall (sensitivity), specificity, macro-averaged F1-score, and ROC-AUC, providing a balanced view of both class discrimination and overall model reliability.

## 3. Regression performance metrics

Mean square error  (MSE) and Mean average error (MAE) are indications of how far off the predictions are. While the MAE calculates an absolute error of the predictions, the MSE penalizes outliers harder(Carvalho, Pereira, and Cardoso 2019; Rainio et al. 2024).

a. MSE = $\frac{1}{N}\sum_{i=0}^{N}(x_i - \widehat{x_\iota})^2$

b. MAE = $\frac{1}{N}\sum_{i=0}^{N}|x_i - \widehat{x_\iota}|$

5M Software used these metrics when evaluating the volume predicted by the food volume models.

## 4. Metrics related to clustering

**Silhouette coefficient** is a measure of how well the generated clusters are defined. It is given by $S = \frac{1}{N}\left(\frac{(b_i - a_i)}{\max(b_i, a_i)}\right)$, where $i \in [1, N]$, $i$ refers to sample-i, $a_i$ is the average dissimilarity of sample-i to all other members of its assigned cluster, and $b_i$ is the dissimilarity between sample-i and members of all clusters except the cluster where sample-i belongs. The value ranges from –1 to 1, where a value close to 1 indicates better quality of clusters (well separated), and a negative value indicates overlapping and poorly defined clusters(Tan et al. n.d.).

TU/e used this metric to find hyperparameters (number of clusters) and evaluate cluster quality for clinical profiling of Eating disorder patients of GGz OB, based on their baseline data.

# B. Evaluating LLMs and other Generative AI

A broad spectrum of evaluation metrics has been employed to assess the performance of LLMs and GenAI across various healthcare and summarization tasks(Abbasian et al. 2024; Chang et al. 2024; Hu and Zhou 2024; Wang et al. n.d.). Within DAIsy the following evaluation metrics, either individually, or in combination, have been used.

## 1. Reference-based metrics

These metrics include Word Error Rate (WER), BLEU, ROUGE, Meteor, and BERT-score, which are applied to transcription quality and summarization tasks, offering quantitative comparisons to human-annotated ground truth.

a. **Word error rate (WER):** When word sequence hypothesised by automatic speech recognition(ASR) system is aligned with a reference transcription, the number of errors is computed as follows:

$$WER = \frac{I + D + S}{N} \times 100$$

Where S=sum of substitutions, I=insertions, D=deletions, and N=total words in the reference transcription(Ali and Renals n.d.).

b. **Bilingual Evaluation Understudy (BLEU)** is a widely used metric for evaluation of quality of text that is machine translated (thus generated) from a different language(Papineni et al. n.d.; Zhang et al. n.d.).

c. **Recall-Oriented Understudy for Gisting Evaluation (ROGUE)** is a measure of the quality of a machine generated summary by comparing to other (ideal, reference) human generated summaries(Lin n.d.).

d. **Metric for Evaluation of Translation with Explicit Ordering (METEOR)**: Based on harmonic mean of unigram precision and recall, this metric, like BLEU, evaluates the quality of machine translated outputs(Banerjee and Lavie 2005; Zhang et al. n.d.).

e. **BERTscore** introduced in (Zhang et al. n.d.) evaluates the quality of generated text by comparing cosine similarity between generated sentence and reference sentence.

### 2. Diarization Error Rate (DER)

The DER is used specifically for evaluating speaker diarization (which is the combination of speaker segmentation and identification). It is given by $DER = \frac{false\ alarm + missed\ detection + confusion}{total}$, where false alarm refers to the detection of non-speech duration as speech, missed detection refers to the reverse of it, confusion refers to assigning a speech segment to the incorrect speaker, and total is the total duration of speech segments across all speakers(Galibert 2013).

### 3. LLM-as-a-Judge and Ground Truth Evaluation

This metric serve as model-based and human-annotated approaches, respectively, to assess tasks like summarization and fact recall(Gu et al. n.d.; Zheng et al. n.d.).

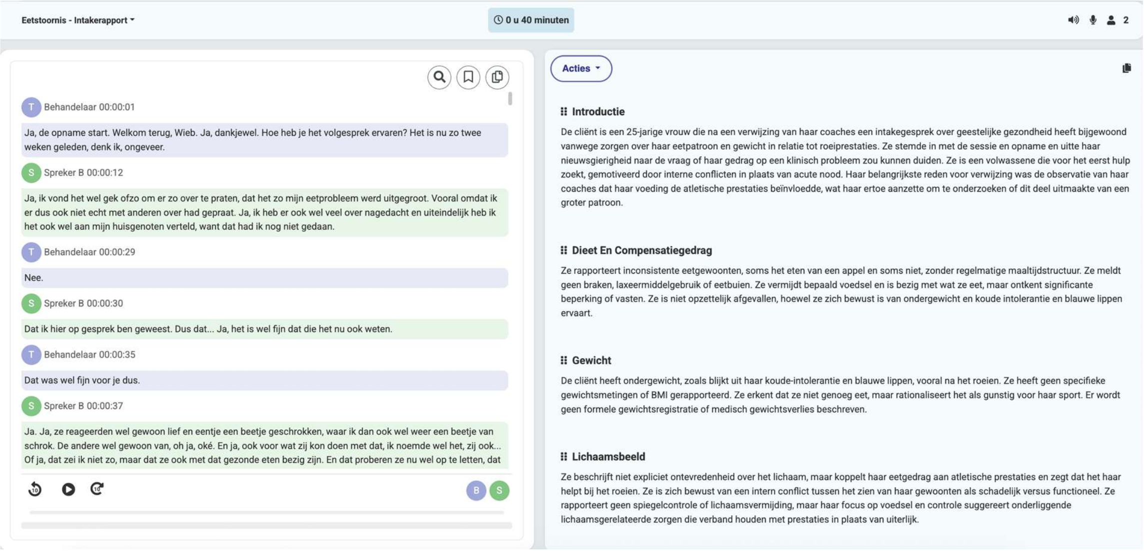
### 4. Retrieval-augmented generation (RAG)

In RAG contexts, metrics like Top-K hit rate, Precision@K, and answer quality dimensions such as relevance, fluency, coherence, and groundedness are used to evaluate both the retrieval and generation stages.

### 5. Guardrail evaluations

For broader safety and robustness, guardrail evaluations help detect harmful or risky content. Experiments also include meta-evaluation—assessing how well different metrics align with human judgment—and hardware benchmarking(e.g., evaluating performance on Macbook M4 vs. H100 GPU). Collectively, these metrics ensure a comprehensive and multi-layered assessment of generative AI capabilities in real-world applications.

MEDrecord used LLM-as-a-Judge Method, Ground Truth Evaluation, and ROUGE for quality control of transcription, diarization and report accuracy of Eating disorder interviews at GGz OB, benchmarked against the diarised transcription and language quality of the reports in multiple languages (See Fig- 2 and Fig- 3).



T Behandelaar 00:00:01

Ja, de opname start. Welkom terug, Wieb. Ja, dankjewel. Hoe heb je het volgesprek ervaren? Het is nu zo twee weken geleden, denk ik, ongeveer.

S Spreker B 00:00:12

Ja, ik vond het wel gek ofzo om er zo over te praten, dat het zo mijn eetprobleem werd uitgegroot. Vooral omdat ik er dus ook niet echt met anderen over had gepraat. Ja, ik heb er ook wel veel over nagedacht en uiteindelijk heb ik het ook wel aan mijn huisgenoten verteld, want dat had ik nog niet gedaan.

T Behandelaar 00:00:29

Nee.

S Spreker B 00:00:30

Dat ik hier op gesprek ben geweest. Dus dat... Ja, het is wel fijn dat die het nu ook weten.

T Behandelaar 00:00:35

Dat was wel fijn voor je dus.

S Spreker B 00:00:37

Ja. Ja, ze reageerden wel gewoon lief en eentje een beetje geschrokken, waar ik dan ook wel weer een beetje van schrok. De andere wel gewoon van, oh ja, oké. En ja, ook voor wat zij kon doen met dat, ik noemde wel het, zij ook... Of ja, dat zei ik niet zo, maar dat ze ook met dat gezonde eten bezig zijn. En dat proberen ze nu wel op te letten, dat

**Introductie**

De cliënt is een 25-jarige vrouw die na een verwijzing van haar coaches een intakegesprek over geestelijke gezondheid heeft bijgewoond vanwege zorgen over haar eetpatroon en gewicht in relatie tot roeiprestaties. Ze stemde in met de sessie en opname en uitte haar nieuwsgierigheid naar de vraag of haar gedrag op een klinisch probleem zou kunnen duiden. Ze is een volwassene die voor het eerst hulp zoekt, gemotiveerd door interne conflicten in plaats van acute nood. Haar belangrijkste reden voor verwijzing was de observatie van haar coaches dat haar voeding de atletische prestaties beïnvloedde, wat haar ertoe aanzette om te onderzoeken of dit deel uitmaakte van een groter patroon.

**Dieet En Compensatiegedrag**

Ze rapporteert inconsistente eetgewoonten, soms het eten van een appel en soms niet, zonder regelmatige maaltijdstructuur. Ze meldt geen braken, laxeermiddelgebruik of eetbuien. Ze vermijdt bepaald voedsel en is bezig met wat ze eet, maar ontkent significante beperking of vasten. Ze is niet opzettelijk afgevallen, hoewel ze zich bewust is van ondergewicht en koude intolerantie en blauwe lippen ervaart.

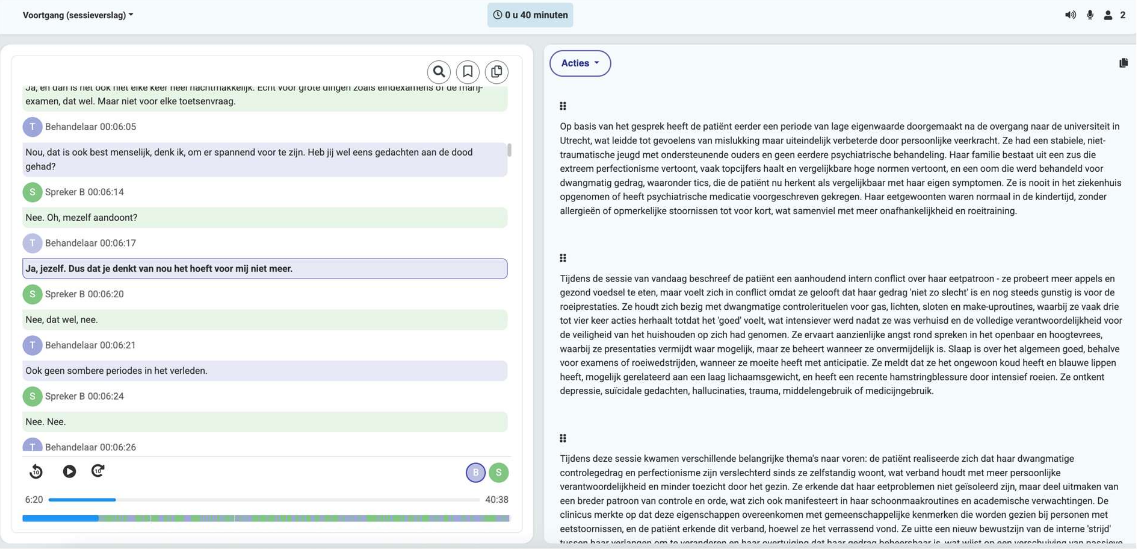
**Gewicht**

De cliënt heeft ondergewicht, zoals blijkt uit haar koude-intolerantie en blauwe lippen, vooral na het roeien. Ze heeft geen specifieke gewichtsmetingen of BMI gerapporteerd. Ze erkent dat ze niet genoeg eet, maar rationaliseert het als gunstig voor haar sport. Er wordt geen formele gewichtsregistratie of medisch gewichtsverlies beschreven.

**Lichaamsbeeld**

Ze beschrijft niet expliciet ontevredenheid over het lichaam, maar koppelt haar eetgedrag aan atletische prestaties en zegt dat het haar helpt bij het roeien. Ze is zich bewust van een intern conflict tussen het zien van haar gewoonten als schadelijk versus functioneel. Ze rapporteert geen spiegelcontrole of lichaamsvermijding, maar haar focus op voedsel en controle suggereert onderliggende lichaamsgerelateerde zorgen die verband houden met prestaties in plaats van uiterlijk.

*Fig- 2: Screenshot illustrating speech-to-text diarization with LLM-as-a-Judge method in MEDrecord’s Healthtalk for DAIsy clinical partner GGZ Oost-Brabant.*



Ja, en dan is het ook niet elke keer heel nachtmakkelijk. Echt voor grote dingen zoals eindexamens of de marij-examen, dat wel. Maar niet voor elke toetsenvraag.

Behandelaar 00:06:05

Nou, dat is ook best menselijk, denk ik, om er spannend voor te zijn. Heb jij wel eens gedachten aan de dood gehad?

Spreker B 00:06:14

Nee. Oh, mezelf aandoont?

Behandelaar 00:06:17

**Ja, jezelf. Dus dat je denkt van nou het hoeft voor mij niet meer.**

Spreker B 00:06:20

Nee, dat wel, nee.

Behandelaar 00:06:21

Ook geen sombere periodes in het verleden.

Spreker B 00:06:24

Nee. Nee.

Behandelaar 00:06:26

Op basis van het gesprek heeft de patiënt eerder een periode van lage eigenwaarde doorgemaakt na de overgang naar de universiteit in Utrecht, wat leidde tot gevoelens van mislukking maar uiteindelijk verbeterde door persoonlijke veerkracht. Ze had een stabiele, niet-traumatische jeugd met ondersteunende ouders en geen eerdere psychiatrische behandeling. Haar familie bestaat uit een zus die extreem perfectionisme vertoont, vaak topcijfers haalt en vergelijkbare hoge normen vertoont, en een oom die werd behandeld voor dwangmatig gedrag, waaronder tics, die de patiënt nu herkent als vergelijkbaar met haar eigen symptomen. Ze is nooit in het ziekenhuis opgenomen of heeft psychiatrische medicatie voorgeschreven gekregen. Haar eetgewoonten waren normaal in de kindertijd, zonder allergieën of opmerkelijke stoornissen tot voor kort, wat samenviel met meer onafhankelijkheid en roeitraining.

Tijdens de sessie van vandaag beschreef de patiënt een aanhoudend intern conflict over haar eetpatroon - ze probeert meer appels en gezond voedsel te eten, maar voelt zich in conflict omdat ze gelooft dat haar gedrag 'niet zo slecht' is en nog steeds gunstig is voor de roeiprestaties. Ze houdt zich bezig met dwangmatige controlerituelen voor gas, lichten, sloten en make-uproutines, waarbij ze vaak drie tot vier keer acties herhaalt totdat het 'goed' voelt, wat intensiever werd nadat ze was verhuisd en de volledige verantwoordelijkheid voor de veiligheid van het huishouden op zich had genomen. Ze ervaart aanzienlijke angst rond spreken in het openbaar en hoogtevrees, waarbij ze presentaties vermijdt waar mogelijk, maar ze beheert wanneer ze onvermijdelijk is. Slaap is over het algemeen goed, behalve voor examens of roeiwedstrijden, wanneer ze moeite heeft met anticipatie. Ze meldt dat ze het ongewoon koud heeft en blauwe lippen heeft, mogelijk gerelateerd aan een laag lichaamsgewicht, en heeft een recente hamstringblessure door intensief roeien. Ze ontkent depressie, suïcidale gedachten, hallucinaties, trauma, middelengebruik of medicijngebruik.

Tijdens deze sessie kwamen verschillende belangrijke thema's naar voren: de patiënt realiseerde zich dat haar dwangmatige controlegedrag en perfectionisme zijn verslechterd sinds ze zelfstandig woont, wat verband houdt met meer persoonlijke verantwoordelijkheid en minder toezicht door het gezin. Ze erkende dat haar eetproblemen niet geïsoleerd zijn, maar deel uitmaken van een breder patroon van controle en orde, wat zich ook manifesteert in haar schoonmaakroutines en academische verwachtingen. De clinicus merkte op dat deze eigenschappen overeenkomen met gemeenschappelijke kenmerken die worden gezien bij personen met eetstoornissen, en de patiënt erkende dit verband, hoewel ze het verrassend vond. Ze uitte een nieuw bewustzijn van de interne 'strijd' tussen haar verlangen om te veranderen en haar overtuiging dat haar gedrag beheersbaar is, wat wijst op een verschuiving van passieve

*Fig- 3: Screenshot illustrating speech-to-text diarization and summarization with LLM-as-a-Judge method in MEDrecord's Healthtalk for DAIsy clinical partner GGZ Oost-Brabant.*

MATERNA used Ollama with RAG, to evaluate with the help of annotated data, the  dataset of medical guidelines for mental healthcare. This served to identify trends and determine whether Ollama could be a useful tool in conjunction with RAG document integration to support therapists in their work. To this end, ollama was tested on a local system (M4 Pro chip, Apple and an H100 GPU).

Semlab used (Custom) Word Error Rate, Meteor, BertScore, Bleu, to assess the quality of transcription on ED intake interviews at GGZ OB, from the Python package Jiwer v3.1.0, and the Evaluate library v0.4.2 from Huggingface. They used DER from Pyannote (version 3.2.1) to evaluate diarization, i.e., speaker segmentation and identification on aforementioned ED intake interviews. They used Rouge metric from the Rogue Python library v1.0.1, and BertScore from Python-based toolbox of the same name (v0.3.13), to evaluate LLM models and prompts on summarization of (parts of) these intake interviews.

Philips used a broad overview of both reference-free and reference-based metrics to obtain an overview of LLM model evaluation tooling and methods (unsupervised, model-based, supervised and GenAI based) for a variety of tasks (eg QA, summarization, data-to-text, dialogue, etc). They also used Information retrieval (IR) and Answer generation (AG) for RAG based chatbots in healthcare domain, and meta-evaluation to assess evaluation metrics (LLM aaJ).

# III. Uncertainty quantification

The performance metric value of a model might be affected by the initialization condition of the model, the variation between the training, test and validation sets, or noise in the data that the model trained on and was applied to. This is why an isolated performance metric value is not enough to judge a model, and uncertainty quantification is necessary.

Predictive uncertainty (PU) is the cumulative of *aleatoric uncertainty* (AU) and *epistemetic uncertainty* (EU). AU, also known as data uncertainty, is due to noise or uncertainty present in the data on which the model is trained, validated or applied. Aleatoric uncertainty is further divided into (i) homoscedastic uncertainty, corresponding to noise that is constant across the input space, and (ii) heteroscedastic uncertainty, referring to noise which varies with the input.

Epistemetic uncertainty is the one arising out of lack of knowledge (include lack of enough training data). Epistemetic uncertainty can be expressed mathematically as a probability distribution over the model parameters, often requiring Bayesian approach to define the model likelihood. For example, the soft-max likelihood in models performing classification task, and Gaussian likelihood for regressor models (Abdar et al. 2021). Uncertainty Quantification can even be in terms of simple measures such as confidence intervals via normal approximation, or bootstrapping, or using standard deviation of performance metrics across different folds of data, which are among the more commonly used metrics in clinical research(Raschka 2018).

In this deliverable the uncertainty quantification and model evaluation tools developed and used by the DAIsy partners are discussed in the following sections:

## 1. Variance and Standard deviation

Evaluation of variance and standard deviation of performance metrices (e.g., balanced accuracy) across cross-validation folds were used to estimate model uncertainty. These measures reflected how stable the model's predictions were across different data splits, helping assess the reliability of results given the limited dataset size(Rainio et al. 2024).

## 2. Variance inflation Factor (V-i-F)

V-i-F is an uncertainty quantification metric used to quantify multicollinearity and the stability of coefficient estimates in regression problems.

## 3. Confidence interval (CI)

CI (Carvalho et al. 2019; Rainio et al. 2024) provides a range of values within which a statistical measure (such as accuracy or C-statistic) has α probability of falling, and α is generally 95 or 99%. Thus, narrower the interval, more robust is the model's performance.

## 4. Probabilistic outcome

Probabilistic outcome of a model provides a model's confidence in its decision. Instead of providing a crisp outcome the model provides a probability distribution or likelihood that an instance belongs to the different classes in the training.

## 5. Token level probabilities

Many LLMs provide probability distributions over the next token in their output (Gupta et al. 2024). Analysing these probabilities (e.g., average token probabilities, minimum token probability, entropy of the distribution) can offer a local measure of confidence for generated text. When running multiple LLMs (or the same LLM with different initializations/parameters/fine-tuning) the observed inconsistencies (or divergence) of the outputs among ensemble members often indicates higher uncertainty.

## 6. Agreement between constituents of an ensemble

Disagreement among ensemble members often indicates higher uncertainty.

ARD Group used standard deviation, variance of model performance across cross-validation folds, to estimate predictive uncertainty and feature relevance when classifying bipolar vs. unipolar depression based on structural MRI and clinical data. Similarly, TU/e used these measures on the performance metrics like class-wise accuracy and AUC of ML classifiers RF, SVM, LDA, LVQ, and the baseline log-LASSO, that were trained for the detection of ED type among patients who had intake at GGz Oost-Brabant. This was also evaluated over 5 cross-validation folds. OFFIS used these measures of uncertainty on the performance metrics like accuracy, F1, and AUC of models like SVM, LDA and logistic regression, in their toolbox for neurofeedback modulation.

5M Software used the class probabilities of predictions by their food recognition model to inform users of the certainty of the predictions, enabling them to override these if necessary.

MATERNA used probabilistic score to predict the focus level and calmness of a user for neurofeedback modulation.

MEDrecord used the measure token level probabilities to evaluate the final report generated by their finetuned LLMs. Furthermore, while running multiple LLMs (or the same LLM with different initializations/parameters/fine-tuning) for all text-2-text generation tasks, they extracted uncertainty of models' decisions by observing the consistency (or divergence) of their outputs, i.e., higher the disagreement between the models of an ensemble of models, the higher the uncertainty in the decisions of these models.

# REFERENCES

Abbasian, Mahyar, Elahe Khatibi, Iman Azimi, David Oniani, Zahra Shakeri Hossein Abad, Alexander Thieme, Ram Sriram, Zhongqi Yang, Yanshan Wang, Bryant Lin, Olivier Gevaert, Li Jia Li, Ramesh Jain, and Amir M. Rahmani. 2024. ‘Foundation Metrics for Evaluating Effectiveness of Healthcare Conversations Powered by Generative AI’. *Npj Digital Medicine 2024 7:1* 7(1):1–14. doi:10.1038/s41746-024-01074-z.

Abdar, Moloud, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. 2021. ‘NC-ND License A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges’. *Information Fusion* 76:243–97. doi:10.1016/j.inffus.2021.05.008.

Ali, Ahmed, and Steve Renals. n.d. ‘Word Error Rate Estimation for Speech Recognition: E-WER’. 20–24. https://github.com/qcri/e-wer.

Banerjee, Satanjeev, and Alon Lavie. 2005. ‘METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments’. 65–72.

Begoli, Edmon, Tanmoy Bhattacharya, and Dimitri Kusnezov. 2019. ‘The Need for Uncertainty Quantification in Machine-Assisted Medical Decision Making’. *Nature Machine Intelligence 2019 1:1* 1(1):20–23. doi:10.1038/s42256-018-0004-1.

Caldeira, João, and Brian Nord. 2021. ‘Deeply Uncertain: Comparing Methods of Uncertainty Quantification in Deep Learning Algorithms’. *Mach. Learn.: Sci. Technol* 2:15002. doi:10.1088/2632-2153/aba6f3.

Carvalho, Diogo V., Eduardo M. Pereira, and Jaime S. Cardoso. 2019. ‘Machine Learning Interpretability: A Survey on Methods and Metrics’. *Electronics 2019, Vol. 8, Page 832* 8(8):832. doi:10.3390/ELECTRONICS8080832.

Chang, Yupeng, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. ‘A Survey on Evaluation of Large Language Models’. *ACM Transactions on Intelligent Systems and Technology* 15(3). doi:10.1145/3641289.

Chen, Mingshi, Zarah van der Pal, Maarten G. Poirot, Anouk Schrantee, Marco Bottelier, Sandra J. J. Kooij, Henk A. Marquering, Liesbeth Reneman, and Matthan W. A. Caan. 2025. ‘Prediction of Methylphenidate Treatment Response for ADHD Using Conventional and Radiomics T1 and DTI Features: Secondary Analysis of a Randomized Clinical Trial’. *NeuroImage: Clinical* 45:103707. doi:10.1016/J.NICL.2024.103707.

Galibert, Olivier. 2013. ‘Methodologies for the Evaluation of Speaker Diarization and Automatic Speech Recognition in the Presence of Overlapping Speech’. doi:10.21437/Interspeech.2013-303.

Gu, Jiawei, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Zhouchi Lin, Yuanzhuo Wang, Lionel Ni, Wen Gao, and Jian Guo. n.d. *A Survey on LLM-as-a-Judge*. https://awesome-llm-as-a-judge.github.io/.

Gupta, Neha, Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh, Rawat Aditya, Krishna Menon, Sanjiv Kumar, Google Research, and New York. 2024. *Language Model Cascades: Token-Level Uncertainty and Beyond*.

Hu, Taojun, and Xiao-Hua Zhou. 2024. ‘Unveiling LLM Evaluation Focused on Metrics: Challenges and Solutions *’.

Lin, Chin-Yew. n.d. ‘ROUGE: A Package for Automatic Evaluation of Summaries’.

Nikolaou, Vasilis. 2016. ‘Statistical Analysis: A Practical Guide for Psychiatrists’. *BJPsych Advances* 22(4):251–59. doi:10.1192/APT.BP.115.014696.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. n.d. ‘BLEU: A Method for Automatic Evaluation of Machine Translation’.

Park, Seong Ho, Kyunghwa Han, Hye Young Jang, Ji Eun Park, June Goo Lee, Dong Wook Kim, and Jaesoon Choi. 2023. ‘Methods for Clinical Evaluation of Artificial Intelligence Algorithms for Medical Diagnosis’. *Radiology* 306(1):20–31. doi:10.1148/radiol.220182.

Park, Seong Ho, Kyunghwa Han, and June Goo Lee. 2024. ‘Conceptual Review of Outcome Metrics and Measures Used in Clinical Evaluation of Artificial Intelligence in Radiology’. *Radiologia Medica* 129(11):1644–55. doi:10.1007/S11547-024-01886-9/FIGURES/3.

Poirot, M. G. n.d. ‘Artificial Intelligence and MRI for the Prediction of Treatment Outcome in Depression’.

Rainio, Oona, Jarmo Teuho, and Riku Klén. 2024. ‘Evaluation Metrics and Statistical Tests for Machine Learning’. *Scientific Reports 2024 14:1* 14(1):1–14. doi:10.1038/s41598-024-56706-x.

Raschka, Sebastian. 2018. ‘Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning’.

Tan, Pang-Ning, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. n.d. *Introduction to Data Mining*. Pearson Education.

Varoquaux, Gael, and Olivier Colliot. 2023. ‘Evaluating Machine Learning Models and Their Diagnostic Value’. *Neuromethods* 197:601–30. doi:10.1007/978-1-0716-3195-9_20/TABLES/2.

Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. n.d. ‘SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems’.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. n.d. ‘BERTSCORE: EVALUATING TEXT GENERATION WITH BERT’. https://github.com/Tiiiger/bert_score.

Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, and Joseph E. Gonzalez. n.d. ‘Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena’.

# APPENDIX

In the appendix we share excerpts from various public and internal dissemination of DAIsy partners, which formed the building blocks of this deliverable document.

## A. Relevant excerpts from TU/e contributions

*Detecting Anomalies in Clinical Data Using Unsupervised Learning: A Study on Anorexia (BSc. End Project report, P. Kwaspen, TU/e)*

***From Chapter 2, page 23***

*The primary metric used for evaluation is the Fβ score with β = 2, which places greater emphasis on recall relative to precision. This is especially appropriate in a clinical setting, where failing to detect a true RFS case (a false negative) may have more severe consequences than incorrectly flagging a non-RFS case (a false positive). Furthermore, since RFS cases are relatively rare, metrics like accuracy can be misleadingly high due to the large number of true negatives.*

*The F2-score is defined as:*

$$F_2 = (1 + \beta^2 ).\frac{Precision\ .Recall}{(\beta^2.Precision)+Recall}, \text{ with } \beta = 2 \qquad (2.9).$$

*This formulation biases the metric toward recall, making it well-suited for evaluating anomaly detection performance in imbalanced, high-risk clinical scenarios such as RFS detection. The F2-score serves as the primary evaluation metric, and the following standard classification metrics are also reported to provide a more complete picture of model performance. To interpret these metrics, it is helpful to understand the basic categories used in binary classification:*

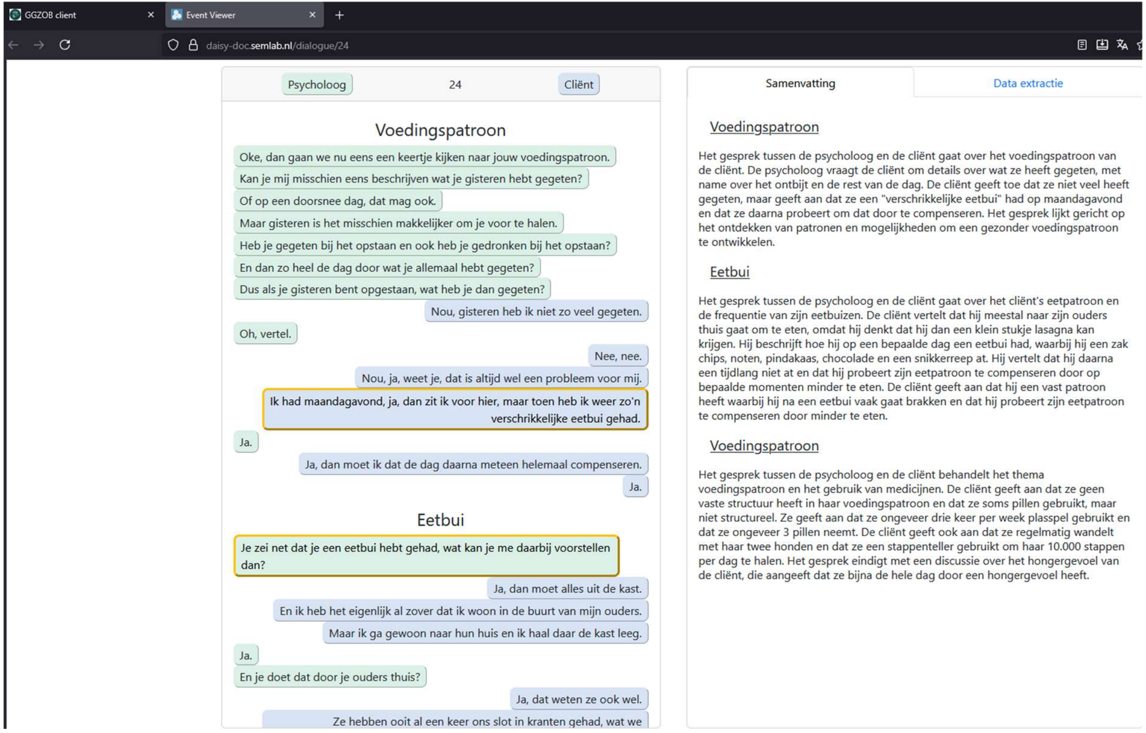
- *True Positive (TP): An RFS case correctly identified as an anomaly.*
- *False Positive (FP): A non-RFS case incorrectly flagged as an anomaly.*
- *True Negative (TN): A non-RFS case correctly identified as normal.*
- *False Negative (FN): An RFS case incorrectly classified as normal.*

*Using these definitions, the following performance metrics are computed:*

- *Precision: the proportion of true positives among all instances predicted as positive:*
  $$Precision = \frac{TP}{TP+F} \qquad (2.10).$$
- *Recall (Sensitivity): the proportion of true positives among all actual positive instances:*
  $$Recall = \frac{TP}{TP+FN} \qquad (2.11).$$
- *Accuracy: the proportion of all correctly classified instances (both positives and negatives) over the total number of instances:*
  $$Accuracy = \frac{TP+T}{TP+TN+FP+FN} \qquad (2.12).$$

*These metrics will be computed using the predictions generated by each model on the test set and compared across different anomaly detection methods and preprocessing configurations. This evaluation directly addresses sub-question 1 and 3, which seek to determine which model performs best in detecting anomalies and what the effect of rate of change features is. By grounding model evaluation in clinically validated labels, the assessment provides a more robust and interpretable measure of real-world performance.*

## B. Screenshot of application developed by Semlab for GGZ OB.



Voedingspatroon

Oke, dan gaan we nu eens een keertje kijken naar jouw voedingspatroon.

Kan je mij misschien eens beschrijven wat je gisteren hebt gegeten?

Of op een doorsnee dag, dat mag ook.

Maar gisteren is het misschien makkelijker om je voor te halen.

Heb je gegeten bij het opstaan en ook heb je gedronken bij het opstaan?

En dan zo heel de dag door wat je allemaal hebt gegeten?

Dus als je gisteren bent opgestaan, wat heb je dan gegeten?

Nou, gisteren heb ik niet zo veel gegeten.

Oh, vertel.

Nee, nee.

Nou, ja, weet je, dat is altijd wel een probleem voor mij.

Ik had maandagavond, ja, dan zit ik voor hier, maar toen heb ik weer zo'n verschrikkelijke eetbui gehad.

Ja.

Ja, dan moet ik dat de dag daarna meteen helemaal compenseren.

Ja.

Eetbui

Je zei net dat je een eetbui hebt gehad, wat kan je me daarbij voorstellen dan?

Ja, dan moet alles uit de kast.

En ik heb het eigenlijk al zover dat ik woon in de buurt van mijn ouders.

Maar ik ga gewoon naar hun huis en ik haal daar de kast leeg.

Ja.

En je doet dat door je ouders thuis?

Ja, dat weten ze ook wel.

Ze hebben ooit al een keer ons slot in kranten gehad, wat we

Voedingspatroon

Het gesprek tussen de psycholoog en de cliënt gaat over het voedingspatroon van de cliënt. De psycholoog vraagt de cliënt om details over wat ze heeft gegeten, met name over het ontbijt en de rest van de dag. De cliënt geeft toe dat ze niet veel heeft gegeten, maar geeft aan dat ze een "verschrikkelijke eetbui" had op maandagavond en dat ze daarna probeert om dat door te compenseren. Het gesprek lijkt gericht op het ontdekken van patronen en mogelijkheden om een gezonder voedingspatroon te ontwikkelen.

Eetbui

Het gesprek tussen de psycholoog en de cliënt gaat over het cliënt's eetpatroon en de frequentie van zijn eetbuizen. De cliënt vertelt dat hij meestal naar zijn ouders thuis gaat om te eten, omdat hij denkt dat hij dan een klein stukje lasagna kan krijgen. Hij beschrijft hoe hij op een bepaalde dag een eetbui had, waarbij hij een zak chips, noten, pindakaas, chocolade en een snikkerreep at. Hij vertelt dat hij daarna een tijdlang niet at en dat hij probeert zijn eetpatroon te compenseren door op bepaalde momenten minder te eten. De cliënt geeft aan dat hij een vast patroon heeft waarbij hij na een eetbui vaak gaat brakken en dat hij probeert zijn eetpatroon te compenseren door minder te eten.

Voedingspatroon

Het gesprek tussen de psycholoog en de cliënt behandelt het thema voedingspatroon en het gebruik van medicijnen. De cliënt geeft aan dat ze geen vaste structuur heeft in haar voedingspatroon en dat ze soms pillen gebruikt, maar niet structureel. Ze geeft aan dat ze ongeveer drie keer per week plasspel gebruikt en dat ze ongeveer 3 pillen neemt. De cliënt geeft ook aan dat ze regelmatig wandelt met haar twee honden en dat ze een stappenteller gebruikt om haar 10.000 stappen per dag te halen. Het gesprek eindigt met een discussie over het hongergevoel van de cliënt, die aangeeft dat ze bijna de hele dag door een hongergevoel heeft.

Fig- A 1: *The AI application from Semlab was developed and finetuned using the evaluation metrices described in this deliverable documents. More of the demo of this application can be found at :* https://daisy-demo.semlab.nl/demo.

## C. Relevant excerpts from AMC contributions

---

*Prediction of methylphenidate treatment response for ADHD using conventional and radiomics T1 and DTI features: Secondary analysis of a randomized clinical trial*(Chen et al. 2025)

---

***From Section 2.4***

*A conventional analysis of T1 and DTI feature distributions was performed using Bayesian analysis using Cohen's d as effect size and 95 % Highest Density Interval (HDI) as confidence interval. Here, the implementation of Bayesian Estimation Supersedes the t-test (BEST) was used (Kruschke, 2013, Kruschke and Liddell, 2018). Conventional and radiomics machine learning model performance was evaluated using balanced accuracy (bAcc), precision, recall, F1 score, Receiver Operating Characteristic (ROC) curve, Area Under the ROC Curve (AUC-ROC) value, and AUC of Precision-Recall Curve (AUC-PRC). To understand how the features contribute to the model during training, SHapley Additive exPlanations (SHAP) values were generated to analyze feature contributions and importance (Lundberg et al., 2020). Exact binomial tests were applied to compare the model performances across participant and age-based subgroups and chance (Sundjaja et al., 2023). Statistical significance was set at $p < 0.05$, with analyses conducted using SPSS (v28.0) and Python 3.10. Further details are in I.4 Supplement Methods.*

***From Section 3.3***

*Performance of the conventional model is shown in Table 2. For the total cohort, the conventional model significantly outperformed chance during treatment (bAcc 63 %, AUC-ROC 0.69, $p = 0.04$), but not post-treatment (bAcc 42 %, AUC-ROC 0.34, $p = 0.77$). During treatment, models performed better for the total cohort compared to the children (bAcc 32 %, AUC-ROC 0.33) and adults (bAcc 36 %, AUC-*

*ROC 0.41) separately. The model's performance was significantly worse than chance for children during treatment (p = 0.01) but not for adults (p = 0.15). Post-treatment, predictive performance significantly declined and did not surpass chance for whole-group and subgroup analyses. As the children subgroup had limited good vs. poor responders (n = 3 vs. n = 20), the conventional model failed to classify.*

***From Section 3.4***

*During treatment, the radiomics model performed better than chance (bAcc of 68 %, AUC-ROC of 0.73, p = 0.003) for the total cohort. Performance slightly surpassed the conventional model, but this was not statistically significant (p = 0.61). For the children subgroup, the radiomics model performed significantly better (bAcc of 64 %, AUC-ROC 0.62, p = 0.02) than the conventional model (bAcc 32 %, AUC-ROC 0.33). ROC curves of Radiomics models during treatment are displayed in Fig. 4. Similar to the conventional models, the predictive performance of the radiomics models significantly diminished post-treatment, and it failed to classify the children.*

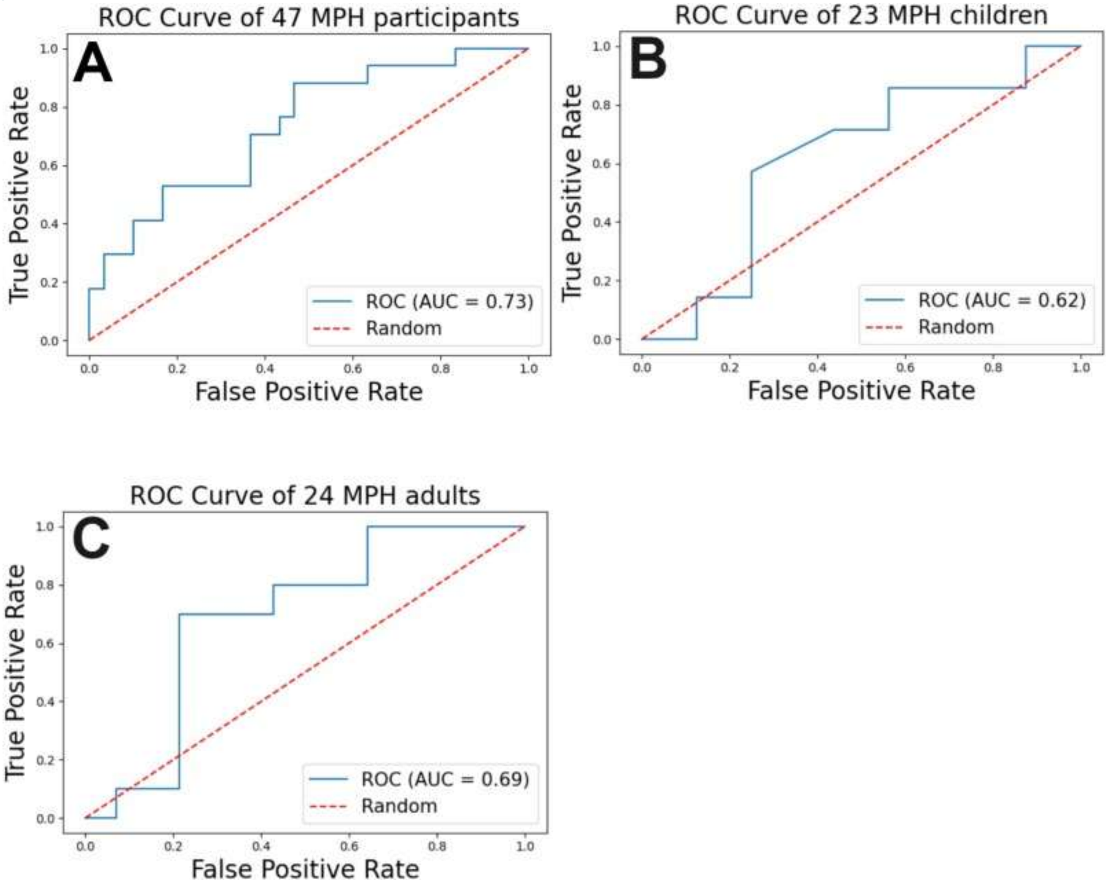


*Fig. 4. Radiomics models' ROC curves with MPH group during treatment for CGI-I (A: total cohort; B: children subgroup; C: adult subgroup).*

***From Section 3.5***

*In children, three outliers with high CJV and DTI motion scores were identified (Fig. A7). The Spearman Rank Correlation Test revealed that 380 features from the Radiomics model were highly correlated with age and less so with CJV and DTI motion (Fig. A8). After employing ComBat and GAMs for harmonization, the age correlation was effectively neutralized, though it inadvertently introduced additional correlation with CJV (Fig. A9). The subsequent removal of three outliers successfully eliminated this CJV correlation (Fig. A10).*

*Sensitivity analyses excluding these outliers revealed a significant impact on both the conventional and radiomics model performance during treatment for the total cohort (conventional model: bAcc 59 %, AUC-ROC 0.59; radiomics model: bAcc 58 %, AUC-ROC 0.57). The performance metrics after excluding these outliers, for both the total cohort and the children subgroup are shown in Supplement Table A4.*

*Moreover, we compared our models including and excluding an additional feature selection step (see 2.4, Supplement Table A4). Feature selection improved model performance only for the radiomics model in the adult subgroup during treatment, achieving a bAcc of 70 % and an AUC-ROC of 0.89. Finally, we assessed to what extent our results differ when we use symptom severity (CGI-S) as outcome compared to treatment response (CGI-I) (Supplement Table A5). Contrary to CGI-I, conventional and radiomics models with CGI-S performed better post-treatment compared to during treatment, with the conventional model achieving a bAcc of 66 % and AUC-ROC of 0.63, and the radiomics model reaching a bAcc of 70 % and AUC-ROC of 0.73. At post-treatment, for the adult subgroup, the conventional model with CGI-S demonstrated robust performance, achieving a balanced accuracy (bAcc) of 84 % and an AUC-ROC of 0.77.*

*Artificial intelligence and MRI for the prediction of treatment outcome in depression(Poirot n.d.)*

***From Chapter 3, pages 75-76***

*Primary predictive performance metrics are mean balanced accuracy (bAcc) and AUROC across repeats and folds. In addition, we report sensitivity and specificity. Balanced accuracy simplifies model evaluation independently of a priori outcome rates. We also computed the natural logarithm of the diagnostic odds ratio ln(DOR) to allow a comparison to a previous meta-analysis.10 For our primary hypothesis that our predictive models significantly outperform chance, we calculated the 95% confidence intervals for bAcc by chance, using one-tailed binomial tests using the number of unique test samples.32 Chance is defined as the a priori response rate, the best alternative method. As a supplementary (post-hoc) analysis, we predict continuous clinical outcome measures directly; see the Supplementary Methods for more details.*

*To test our second hypothesis on treatment specificity, we tested if external validation performance in subgroups B (placebo-treated) and C (sertraline-treated placebo-non-responders) was significantly lower than on internal validation in Subgroup A (sertraline-treated) by performing a one-sided dependent t-test between Subgroup B and the independent test set A, and between Subgroup C and test set A. T-test samples consisted of the mean performance estimates for all modeling configurations, i.e., pretreatment and early-treatment prediction of remission and response. We excluded configurations that did not perform significantly better than chance from this analysis to avoid regression toward the no-information rate.*

***From Chapter 5, page 179***

*The primary performance metrics for classifiers are the area under the receiver operating characteristic (AUROC), the balanced accuracy (bAcc), and the F1-score. For regressors, these were the root mean squared error (RMSE), the mean absolute error (MAE), and the explained variance (R2). Here, the error is defined as the difference between the predicted and true week 8 HAM-D score.*

*We define three analysis samples fitting our research analyses. In our primary analysis, we performed internal validation within analysis sample A using leave-one-out (LOO) cross-validation (CV). For the remaining analyses, we trained on analysis sample A and tested on B, tested on and C. All validations were nested in 5-times repeated 5-fold CV to allow for Bayesian hyperparameter optimization implemented in Scikit-Optimize (v.0.10.1).*[62]