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

Public State of the Art Document

Document type	: Deliverable
Document version	: No. 1
Document Preparation Date	: Feb 2026
Classification	: Public
Due Date	: March 2026

25 years ITEA 4 Community 1998–2023

1. Table of Contents

Glossary 3
2. Introduction 4
3. Artificial Intelligence in Mental Healthcare 4
3.1 Types of AI Models 4
3.2 Clinical Applications 5
3.3. Ethics, Trust and Regulation 6
4. Tools and Techniques Used in DAISY 7
4.1 Software Tools and Frameworks 8
4.2 AI Models Applied in DAIsy 9
4.3 Explainability Methods and XAI Tools 11
5. Review of Scientific Literature 12
5.1 Key Publications (2020–2025) 12
5.2 Common Insights and Trends 14
5.3 Gaps and Challenges 15
6. DAIsy Approach and Innovations 16
6.1 Use Case Examples from Partners 16
6.2 Innovations and Contributions 17
6.3 Future Vision 18
7. Summary and Conclusion 18
8. References 19

Glossary

Artificial Intelligence:	A field of computer science focused on creating systems that can perform tasks typically requiring human intelligence.
Explainable AI (XAI):	AI techniques that allow humans to understand and trust the results and outputs of machine learning algorithms.
Machine Learning:	A subset of AI that enables systems to learn and improve from data without being explicitly programmed.
Clinical Decision Support System (CDSS):	A health information system that helps clinicians make data-driven decisions.
Neurofeedback:	A therapy technique that provides real-time feedback on brain activity to promote self-regulation.
Foundation Model:	A large pre-trained AI model that can be fine-tuned for a variety of specific tasks.
SHAP:	A method to explain the output of machine learning models based on Shapley values.
LIME:	A technique to explain the predictions of machine learning models locally around a given prediction.
Causal Inference:	The process of drawing conclusions about cause-and-effect relationships from data.
Wearable Sensor:	A device worn on the body that collects data related to health, behavior, or physical activity.
Multimodal Data:	Data that comes from multiple sources or types, such as text, audio, and images.
Psychometric Scale:	A standardized tool used to measure psychological variables like mood, anxiety, or cognitive function.
Personalized Treatment:	Medical care tailored to the individual characteristics of each patient.
Data Privacy:	The protection of personal data from unauthorized access or disclosure.
Interoperability:	The ability of different systems and organizations to work together and share information effectively.

2. Introduction

The DAIsy project, titled Developing AI ecosystems improving diagnosis and care of mental diseases, addresses urgent societal and clinical challenges in the field of mental healthcare. Conditions such as Major Depressive Disorder (MDD) and Eating Disorders (ED) affect millions of individuals across Europe and worldwide. These disorders are complex, often underdiagnosed, and difficult to treat effectively due to the interplay of biological, behavioral, and environmental factors.

DAIsy brings together a diverse consortium of hospitals, universities, SMEs, and industry partners from the Netherlands, Germanyand Türkiye. The project aims to develop AI-supported software solutions that improve diagnostic accuracy, assist in treatment selection, support behavioural change, and monitor treatment response in mental healthcare. With a focus on both youth and adult populations, DAIsy integrates clinical expertise with cutting-edge technological development.

This State-of-the-Art (SoTA) document provides an overview of the current landscape of AI applications in mental healthcare and situates the DAIsy project within it. It highlights the models, tools, and methods used across DAIsy's technical and clinical work packages and examines how these approaches align with recent academic literature and international developments. In doing so, the document clarifies how DAIsy contributes to the field and identifies the unique value it offers.

A general overview of AI in mental healthcare, including types of models, clinical use cases, and ethical considerations are given in this document. It also outlines the tools and techniques used in DAIsy, followed by a focused review of relevant scientific publications between 2020 and 2025by considering use-case examples, alignment with current scientific understanding, and future vision.

By presenting this public SoTA overview, the DAIsy consortium aims to ensure transparency, inform external stakeholders, and contribute to the broader dialogue around responsible and effective use of AI in mental health.

3. Artificial Intelligence in Mental Healthcare

3.1 Types of AI Models

The DAIsy project employs a diverse range of AI models, selected and developed to meet the clinical and technical demands of each use case in the domains of depression and eating disorders. These models vary in complexity, interpretability, and data requirements, and are applied across different stages of the patient pathway including diagnosis, treatment planning, and response monitoring.

Among the commonly used interpretable models are logistic regression, decision trees, k-nearest neighbours (KNN), and support vector machines (SVMs). These are primarily applied where model transparency and clinician interpretability are of high importance, especially in early diagnostic stages or in use cases with limited data volume.

In addition to these classical models, ensemble methods such as random forest and XGBoost are used to increase prediction robustness. These models are integrated in tasks such as patient classification, feature importance analysis, and early detection based on multimodal input data.

For more complex applications such as image, signal, or time-series analysis, DAIsy partners have employed deep learning approaches, particularly convolutional neural networks (CNNs) and transformer-based language models. These are used for tasks including emotion recognition, medical

image analysis, clinical text summarization, and patient monitoring through wearable data. CNNs are central to tools such as the food identification module in 5M Software's mobile application, while transformer models support speech-to-text and clinical narrative processing in Semlab and MEDrecord's platforms.

A novel model type explored in DAIsy is the Kolmogorov–Arnold Network (KAN), developed and applied by the AMC clinical team. This approach combines mathematical transparency with the capacity to handle multi-source data including neuroimaging, environmental variables, and demographics. KAN has shown promise in differentiating between unipolar and bipolar depression, supporting treatment decision-making.

For real-time applications involving brain-computer interfaces and neurofeedback, DAIsy also integrates hybrid systems such as GLM-based modules and real-time SVM pipelines, particularly within the MultiPy framework developed by OFFIS.

In language-based tasks, large language models (LLMs) such as Whisper and VaultGemma-1B are applied for transcription, summarization, and privacy-preserving processing of clinical dialogues. These are used to reduce documentation burdens and enhance clinician-patient communication.

Each model within DAIsy is selected with a view to balance performance, interpretability, and data suitability. The project's focus on clinically useful AI means that model complexity is always considered in relation to its integration with real workflows, and tools are often combined to form pipelines that allow stepwise processing and explainability.

3.2 Clinical Applications

The DAIsy project delivers artificial intelligence solutions that respond to real clinical needs across different stages of mental healthcare. These applications are designed in collaboration with healthcare providers and are grounded in existing clinical workflows. The focus is on two major domains: depressive disorders and eating disorders. The work covers diagnosis, treatment support, behavioural monitoring and feedback, documentation, and patient engagement. All applications are developed with attention to clinical feasibility, privacy, and ethical standards.

One of the main areas of application is the early identification of depressive disorders. At Amsterdam UMC, a method based on Kolmogorov–Arnold Networks is used to support the differentiation between unipolar and bipolar depression. This model combines neuroimaging data with environmental and demographic variables to improve the accuracy and transparency of diagnosis.

In the context of eating disorders, GGZ Oost Brabant and the Dutch Eating Disorder Register make use of structured clinical data and validated questionnaires. Models such as logistic regression and support vector machines are applied to assist in the classification of anorexia nervosa and bulimia nervosa. ML-based approaches can provide valuable empirical insight into both diagnostic boundaries and within-diagnosis variability, informing future efforts toward more flexible and evidence-based ED classification frameworks.

Beyond classification, several DAIsy partners develop tools for behaviour tracking and self-monitoring, especially in youth-focused use cases. 5M Software has created a mobile application that includes a food recognition function based on image analysis. This feature helps young users receive guidance on food choices and supports therapeutic conversations. The explanations behind the model's decisions are presented in a user-friendly format to support transparency.

Vestel contributes to the development of emotion recognition tools using facial and voice analysis. These tools provide non-intrusive feedback to help adolescents recognize emotional states. This is intended to encourage reflection and strengthen engagement with therapeutic content delivered through a virtual interface.

In clinical settings, partners such as Semlab and MEDrecord work on reducing the administrative burden of mental health professionals. Using automatic speech recognition and summarization tools, clinician notes are transformed into structured text. These tools help save time and improve the quality of clinical documentation while protecting sensitive information. The systems are built to be compliant with privacy regulations and allow users to maintain control over data use.

In some use cases, sensor-based systems are applied. OFFIS explores the use of real-time signal processing in neurofeedback scenarios. These systems analyze data from EEG devices to support therapists during live sessions. Other partners work with wearable data to track physiological changes related to mental health, including heart rate and motion signals.

All applications developed in DAIsy aim to enhance mental healthcare by offering clinically meaningful insights, supporting user interaction, and facilitating decision-making. The solutions are designed to be practical, adaptable to various clinical contexts, and aligned with the expectations of both professionals and patients.

3.3. Ethics, Trust and Regulation

The DAIsy project places ethical responsibility and trust at the center of its development and deployment of artificial intelligence tools for mental healthcare. Given the sensitivity of psychiatric data and the potential impact of algorithmic decisions on patient care, ethical alignment and regulatory compliance are treated as essential project components from the outset.

A central goal of DAIsy is to ensure that the tools and models developed are not only technically valid but also acceptable, understandable, and trustworthy from a clinical and societal perspective. The project acknowledges that healthcare professionals and patients must be able to rely on the output of algorithmic systems in order to benefit from them. This trust cannot be established without a clear understanding of how decisions are made and on what basis. For this reason, explainability and interpretability are integrated into all technical work packages. Each model is supported by appropriate mechanisms for transparency, including model-agnostic explanation tools, interpretable architectures, and visualization components designed with end-users in mind.

DAIsy also takes into account the ethical concerns that arise when deploying data-driven systems in mental health. These include the potential for bias in training data, risks of over-reliance on automation, and the possibility of unintended effects in vulnerable populations. To address these, the consortium applies internal validation procedures and cross-disciplinary reviews involving clinicians, data scientists, and ethics experts.

Clinical partners are directly involved in co-designing the tools, ensuring that they are grounded in real practice needs and align with professional values. The project includes user feedback loops to assess trust and usability, particularly in settings involving youth or digitally mediated care.

From a regulatory standpoint, DAIsy adheres to European frameworks governing data protection, medical software, and AI systems. All clinical applications are developed with full compliance to the General Data Protection Regulation (GDPR), and partners handling patient data implement technical and organizational safeguards such as pseudonymization, secure storage, and audit mechanisms.

The tools developed by the project are classified and tested according to applicable medical device directives and software safety guidelines. The consortium actively monitors developments in regulatory frameworks related to AI in healthcare, including the European AI Act, and adapts accordingly to ensure future readiness.

By integrating ethical considerations into every stage of its design and implementation, DAIsy aims to develop AI-based solutions that are not only effective but also responsible, respectful of patient rights, and sustainable within the healthcare system. The project's approach reflects a shared commitment across partners to building digital technologies that serve people and support professionals, without replacing the human judgement at the heart of clinical care.

4. Tools and Techniques Used in DAISY

The DAIsy project applies a wide range of artificial intelligence methods and software tools tailored to the diagnosis, treatment planning, and monitoring of depressive and eating disorders. These techniques are selected according to the nature of each use case, the clinical objectives involved, and the type of data collected. The models are deployed within secure, privacy-compliant environments and are integrated with clinical workflows where appropriate.

DAIsy works with diverse data sources, including clinical forms, structured questionnaires, neuroimaging, speech recordings, sensor outputs, and visual content. Each partner contributes to the selection, development, and deployment of tools best suited for their data and clinical setting.

Model Architectures and Learning Approaches

A number of well-established machine learning models are used across the project. These include logistic regression, decision trees, k-nearest neighbors, and support vector machines. Such models offer clarity in how decisions are made, which is especially important in early diagnostic tasks and in interactions with clinicians.

In cases where larger feature sets or non-linear interactions are present, ensemble methods such as random forest and XGBoost are employed. These models are used in analyzing clinical intake data, patient self-assessments, and behavioral questionnaires.

Deep learning approaches are also used in selected applications. Convolutional neural networks are implemented, for example, in image-based food recognition tasks for youth with eating disorders. Transformer-based models are applied in natural language processing, such as in converting clinical conversations to structured summaries. Examples include the use of Whisper for transcription and VaultGemma-1B for summarization.

A unique model architecture in DAIsy is the Kolmogorov–Arnold Network, developed by the Amsterdam UMC team. This method is used to process combined neuroimaging, demographic, and environmental data to support the differentiation of unipolar and bipolar depression. The approach enables interpretable model outputs by analyzing the relative influence of input features.

In the Turkish use cases, additional pipelines were developed for processing magnetic resonance imaging data within hospital infrastructure. These workflows are integrated into clinical systems in a read-only manner, ensuring data is analyzed securely without direct extraction. This setup allows for automated support in distinguishing between depressive subtypes using imaging data.

Software Tools and Platforms

DAIsy uses a combination of open-source and custom-built tools. Common development frameworks include Scikit-learn for classical models, and PyTorch and TensorFlow for deep learning. Natural language models are managed through the Hugging Face ecosystem. Tools such as SHAP, LIME, and Captum are applied for explainability.

User-facing interfaces and dashboards are built using standard web technologies such as React or Vue, and data is stored in relational databases managed with appropriate security and access controls. Backend services are designed to separate data processing from user interaction, supporting modularity and minimizing system exposure.

Clinical data exchange is managed through secure connections to mental health care and hospital systems. Data access is limited by design, and most deployments operate in controlled, on-premise environments. In both Turkish use cases, AI tools are deployed within hospital networks, respecting local data protection laws and institutional policies. The systems function without requiring data to leave the facility, and each service is governed by access roles.

Explainability and Clinical Interpretation

Explainability is a core part of DAIsy's development philosophy. Tools such as SHAP and LIME are used to visualize feature importance and generate locally interpretable explanations for model predictions. These are integrated into clinical dashboards where applicable, allowing users to review and understand model behavior.

In some applications, attention mechanisms are used to enhance understanding of model focus, particularly in visual and text-based domains. Where available, graphical summaries are included to support user interaction.

All explanation methods are designed with the end user in mind, ensuring that results are meaningful to clinicians and, in some cases, patients. This supports better communication, trust, and ethical use of automated systems in mental healthcare.

4.1 Software Tools and Frameworks

The DAIsy project builds on a solid and diverse software foundation to support the development, training, deployment, and validation of artificial intelligence models in mental healthcare settings. Each tool and framework has been selected in accordance with the technical needs of specific use cases, data types, and integration environments.

For classical machine learning tasks, such as classification of structured clinical data or routine outcome monitoring (ROM) inputs, the project uses Scikit-learn, a widely adopted Python library for statistical and machine learning methods. This framework is particularly useful for deploying interpretable models including logistic regression, decision trees, and support vector machines, which are used in applications such as anomaly detection in anorexia nervosa recovery patterns and early screening for eating disorders.

In deep learning applications, PyTorch is primarily used across multiple work packages. Its flexibility allows partners to build and evaluate convolutional neural networks and transformer-based models, especially for image, audio, and natural language processing tasks. Examples include food recognition modules, emotion detection through visual or audio signals, and summarization of clinician-patient dialogues.

Some modules, such as the real-time neurofeedback classifier developed in the MultiPy toolbox, rely on additional libraries including NumPy, Pandas, SciPy, and Numba, which support efficient numerical computations and streaming signal analysis. For real-time signal communication, PyLSL is used to handle data exchange between EEG/fNIRS pipelines and the model inference layer. Visualization components of these systems are built using PySide6, PyQtGraph, and Matplotlib, offering interactive plots and accuracy metrics for clinician review.

Natural language processing and speech recognition tasks utilize open-source models, including Whisper and VaultGemma, integrated into clinical applications to automate documentation and support structured reporting. These models are managed and adapted using tools from the Hugging Face ecosystem, which also provides pre-processing and deployment utilities.

Model explanation tools such as SHAP, LIME, and ELI5 are widely used across different pipelines to ensure interpretability of predictions. These tools are particularly important in models applied to clinical decisions, where visualizing feature contributions supports transparency and trust.

Frameworks and services are deployed within isolated environments using Docker containers, often managed through Kubernetes clusters. These deployments follow a microservice architecture, allowing models to run independently while integrating with upstream and downstream components. Interfaces are described using Swagger (OpenAPI) to ensure standardization and developer accessibility.

Data pipelines, particularly in imaging-related applications such as Turkish Use Case 2, utilize neuroimaging preparation tools like FSL. These components generate feature representations from MRI data, which are then processed by machine learning models through interfaces connected to clinical visualization tools such as the VTK-based Viewer and Doctor Dashboard.

Package management and reproducibility are ensured through Conda environments with fixed dependency versions. This setup minimizes cross-component conflicts and supports consistent execution across partners.

All software tools and frameworks in DAIsy are selected with clinical integration, performance efficiency, and long-term sustainability in mind. The modular and reproducible design of each system enables secure deployment within clinical infrastructures and facilitates future scaling or adaptation as technologies evolve.

4.2 AI Models Applied in DAIsy

The DAIsy project incorporates a variety of artificial intelligence models, each selected based on the specific needs of the clinical use case, the structure of the available data, and the deployment requirements in real-world healthcare environments. Rather than focusing on a single modeling approach, DAIsy integrates interpretable, statistical, deep learning, and hybrid models across its platform and services.

Interpretable and Classical Models

In several DAIsy use cases, especially those involving structured clinical questionnaires or digital diaries, interpretable models such as logistic regression, decision trees, support vector machines (SVMs), and k-nearest neighbors (KNN) are utilized. These models are preferred in contexts where clinical transparency and explainability are essential. For example, in the Therapy Assistant, patient-reported measures such as PHQ-9 scores and digital behavior markers are processed to support

therapy planning. The prediction of patient progress in this system is supported by statistical models embedded in the backend infrastructure.

Ensemble Models and Feature-Based Techniques

For more complex classification tasks involving high-dimensional features or class imbalance, DAIsy applies random forest and XGBoost models. These are commonly used in dashboards and backend services where multiple patient data streams are aggregated. For example, clinicians can view AI-supported predictions for therapy response or detect early signs of disengagement based on aggregated digital behavior, as implemented in components like the Therapist Dashboard and FitSprite Nutrition Portal.

These models offer the advantage of providing feature importance measures, which can be interpreted through post-hoc explainability tools such as SHAP or LIME.

Deep Learning Models

In use cases where data includes images, audio, or natural language, deep learning methods are implemented. The FitSprite Nutrition platform uses a set of three deep learning models to estimate food intake from photos:

- A food recognition model for identifying food types
- A segmentation model for isolating food portions
- A depth estimation model for calculating portion sizes based on image depth maps
- These models are built using the TensorFlow framework and run within Docker containers.

In the domain of speech-based clinical documentation, DAIsy integrates Whisper and VaultGemma models for speech-to-text transcription and medical summarization, especially in MEDrecord and Semlab use cases. These models help reduce clinician burden and ensure consistent documentation in psychiatry settings.

Hybrid and Domain-Specific Models

A key innovation within the project is the implementation of Kolmogorov–Arnold Networks (KAN) by the clinical partner Amsterdam UMC. This model architecture is used to classify patient subtypes by combining structured clinical data, neuroimaging features, and environmental exposures. The KAN approach balances interpretability and performance by highlighting which variables (e.g., environmental or anatomical) contribute most to a given classification result.

Real-Time Models for Neurofeedback

For real-time therapeutic applications, such as MultiPy, machine learning models are integrated into neurofeedback systems that process EEG and fNIRS data. These pipelines, developed by OFFIS and partners, include components for signal quality assessment, feature extraction, and adaptive feedback generation. Models in this context are based on SVMs, linear classifiers, and rule-based logic that respond in real time during clinical sessions.

Clinical Imaging Models

In Turkish Use Case 2, structural MRI data is processed using a dedicated AI pipeline for the differential diagnosis of depressive subtypes. The pipeline includes:

- Region-based feature extraction from MRI
- Classification using trained models

- Integration with clinical findings for context-aware results

The outcome is returned to the clinician via structured reports stored in the PACS system. The entire pipeline is managed through a UPS-based orchestration model, ensuring traceability and compliance with hospital workflows.

DAIsy's model diversity ensures that technical robustness is matched with clinical relevance. By using interpretable models where required and more complex methods where appropriate, the project reflects a balanced and responsible approach to AI development in mental healthcare.

4.3 Explainability Methods and XAI Tools

Explainability plays a central role in the design and implementation of AI components within the DAIsy project. Given the clinical setting in which these tools are deployed, transparency and trust are considered essential. Both intrinsically interpretable models and post-hoc explanation tools are used to make AI-supported decisions understandable to healthcare professionals and, where relevant, to patients and caregivers.

Post-Hoc Explainability Tools

DAIsy makes use of several widely recognized model-agnostic explanation libraries:

SHAP (SHapley Additive exPlanations) is used to quantify the contribution of each input feature to a given model prediction. It is applied in multiple work packages to support visual explanations, particularly in use cases involving structured clinical data, such as risk prediction or therapy adherence assessment.

LIME (Local Interpretable Model-agnostic Explanations) is used to generate locally interpretable surrogate models, especially in image-based and emotion detection use cases. For example, LIME is applied in the FitSprite Nutrition app to explain food classification results from CNN-based models.

ELI5 is applied during model debugging and feature sensitivity analysis, helping technical teams understand how model outputs change in response to variations in the input data.

These tools support developers and clinicians in reviewing model behavior and verifying whether the system's decisions align with clinical logic and ethical expectations.

Visualization and Model Attention

In tasks involving vision or text-based models, additional explanation strategies are employed:

Attention maps are used in models analyzing visual data such as facial expressions and food images. These maps help visualize which parts of the input contribute most to a classification decision. For example, in emotion recognition modules developed by Vestel, attention mechanisms highlight regions of interest on the user's face to support transparent interaction.

In speech summarization and transcription tasks, attention weights within transformer-based language models provide an indication of which parts of a sentence or dialogue influence the summary output. This is important in MEDrecord's documentation tools, where clinicians must ensure the generated content reflects the clinical reality.

Interpretability by Design

Beyond post-hoc techniques, many DAIsy models are interpretable by nature. Logistic regression, decision trees, and support vector machines are applied in tasks where full traceability of predictions is needed. These models are especially useful in use cases related to early screening, patient motivation classification, and treatment planning.

In addition, Kolmogorov–Arnold Networks (KAN) applied by Amsterdam UMC are designed to provide interpretable insights from heterogeneous input sources such as MRI-derived features, environmental exposure variables, and clinical scales. These models output variable importance rankings that clinicians can review to understand which inputs are most associated with a given classification.

Integration with Clinical Interfaces

DAIsy ensures that explanation outputs are not only computed but also presented in ways that are accessible to clinicians. Visual feedback such as SHAP plots, LIME masks, or attention heatmaps are integrated into user-facing dashboards or mobile applications where appropriate.

For instance, in the Therapist Dashboard, feature contributions are visualized to assist therapists in interpreting AI-predicted patient trajectories. In Turkish Use Case 2, model results from MRI classification pipelines are returned with variable-specific indicators to guide diagnosis refinement without overwhelming the clinician with technical details.

Explainability methods in DAIsy are not limited to algorithmic transparency but are designed to support real-world clinical interpretation, enhance trust, and promote responsible AI use. By combining technical rigor with human-centered design, the project strengthens the connection between predictive analytics and clinical decision-making.

5. Review of Scientific Literature

5.1 General Review

This section presents a curated review of nine peer-reviewed scientific publications identified in DAIsy Deliverable D4.1 as central to the intersection of explainable artificial intelligence (XAI) and mental healthcare. These studies, published between 2020 and 2025, include review papers, original research articles, exploratory studies, and expert perspectives. All selected publications are clinically grounded and directly relevant to the project's focus areas, such as depression and eating disorders. Studies based solely on social media data or non-clinical questionnaires were deliberately excluded to ensure clinical validity.

Overview of Reviewed Publications:

Ghosh et al. (2024)

A narrative review analyzing machine learning (ML) solutions applied in the clinical management of eating disorders (EDs). It highlights strengths and limitations of XAI in ED diagnosis and treatment, and the need for clinically interpretable models. This article directly informs DAIsy's approach to model selection for ED-focused use cases.

Byeon (2023)

A review of AI-based depression prediction methods. It underscores the trade-off between model complexity and explainability and highlights performance metrics and regulatory implications. DAIsy's clinical partners addressing depressive disorders have adopted a similar interpretability-performance balance.

De Franceschi et al. (2025)

An original study introducing an ensemble of Kolmogorov–Arnold Networks (KANs) for differentiating psychiatric patients from healthy controls. The paper emphasizes combining neuroimaging, environmental, and demographic features. DAIsy's Amsterdam UMC use case employs KANs in a comparable clinical context.

Joyce et al. (2023)

This review introduces the TIFU framework (Transparency and Interpretability for Understandability) and discusses tools like SHAP and LIME. It recommends using black-box models as preprocessing tools, followed by simpler interpretable models for final decision layers a methodology closely reflected in DAIsy's explainability pipeline.

de Oliveira et al. (2025)

An exploratory study assessing the effect of XAI on the trust of mental health professionals. Using LIME and SHAP, the study evaluates AI-assisted suicide prevention tools. The findings support DAIsy's emphasis on clinician education and trust in AI systems.

Kerz et al. (2023)

This research presents "MentalRoBERTa" a transformer-based NLP model for detecting emotion, personality, and mental health conditions. It uses LIME-based explanations. DAIsy partners MEDrecord and Vestel explore similar XAI strategies for patient interaction modules.

Ostojic et al. (2024)

A review of practical challenges in applying ML to psychiatry, including integration issues, model opacity, and clinician skepticism. The article reinforces DAIsy's alignment with responsible AI development and multi-stakeholder usability.

Sheu (2020)

A review focused on interpretable deep learning (DNN) models in psychiatric research. It presents multiple XAI libraries (Captum, tf-explain) and highlights the limitations of black-box architectures. DAIsy's CNN-based components benefit from similar techniques in imaging use cases.

Starke et al. (2022)

A qualitative expert review examining the ethical expectations of clinicians toward AI in psychiatry. It argues that explainability is essential for real-world adoption and discusses tensions between clinical practice and existing AI guidelines. DAIsy integrates these ethical insights across use-case development.

Summary and Alignment with DAIsy

These publications reveal that XAI is not only a technological tool but also a central element of clinical trust and regulatory compliance. The models and tools reviewed, such as SHAP, LIME, KAN, SVM, LogReg, and attention mechanisms, directly influence the methodological choices made in DAIsy. In multiple cases, partner-specific implementations echo strategies discussed in these works, demonstrating strong alignment with current academic and clinical standards. Furthermore, issues of transparency, model trust, and human-in-the-loop design, as emphasized in the literature, are reflected across DAIsy's explainability workflows and stakeholder engagement activities.

Table 1:Overview of state-of-the-art explainable AI techniques in mental healthcare

	Paper title	Paper type and use-case(s)
1	**Review of machine learning solutions for eating disorders** (Ghosh et al. 2024)	Narrative review on the clinical management of eating disorders using state of the art ML/AI techniques
2	**Advances in machine learning and explainable artificial intelligence for depression prediction** (Byeon 2023)	Review paper involving Depression prediction with ML/AI
3.	**Ensemble KAN: Leveraging Kolmogorov Arnold Networks to discriminate individuals with psychiatric disorders from Controls** (De Franceschi et al. 2025)	Original research paper involving ensemble of KANs for differentiation of healthy controls and two conditions in Psychiatry, namely psychosis and depression.
4.	**Explainable artificial intelligence for mental health through transparency and interpretability for Understandability** (Joyce et al. 2023)	Review paper involving XAI for Psychosis and Depression.
5.	**Effect of explainable AI on trust of mental health professionals in an AI-based system for suicide prevention** (de Oliveira et al. 2025)	Exploratory study investigating how XAI tools impact the trust of mental health professionals on tools for Depression, such as the Boamente systems for suicide risk prediction.
6.	**Toward explainable AI (XAI) for mental health detection based on language behaviour** (Kerz et al. 2023)	Original research on XAI for text-based data, focussing on social media, for detection of five Mental health conditions, including depression and bipolar disorder.
7.	**The challenges of using machine learning models in psychiatric research** (Ostojic et al. 2024)	Review paper that discusses various challenges of applications of ML/AI in Mental healthcare, including but not limited to explainability of models.
8.	**Illuminating the Black Box: Interpreting Deep Neural Network Models for Psychiatric Research** (Sheu 2020)	Review paper on interpretable DNN for Psychiatry in general
9.	**Explainability as fig leaf? An exploration of experts' ethical expectations towards machine learning in psychiatry** (Starke et al. 2022)	Qualitative expert review of challenges of application of opaque ML/AI in Psychiatry use-cases.

5.2 Common Insights and Trends

The reviewed literature reveals several recurring insights that shape the development of explainable artificial intelligence in mental healthcare. These trends directly inform the methodological and ethical decisions made in the DAIsy project. Despite variations in clinical focus, model complexity, and data types, a number of shared themes are evident across the nine publications considered in DAIsy's state-of-the-art analysis.

Interpretability as a Foundation for Trust

Across all reviewed sources, explainability is consistently linked to trust. Papers such as those by Joyce et al. (2023) and de Oliveira et al. (2025) highlight how mental health professionals are more likely to accept AI-supported systems when they understand how decisions are made. Tools like SHAP and LIME are often presented not only as technical aids, but as mechanisms that allow professionals to verify and contextualize AI outputs within their clinical reasoning.

This emphasis on transparency is echoed in the DAIsy project, where explainability tools are embedded in clinical interfaces and model choices prioritize interpretability when needed. The shared view is that black-box models may be useful in processing complex data, but the final decision-making layer must be accessible and auditable.

The Trade-off Between Performance and Interpretability

Many of the reviewed articles, including those by Byeon (2023), Sheu (2020), and Ostojic et al. (2024), discuss the balance between high-performing but opaque models and simpler, more interpretable alternatives. This trade-off is particularly important in psychiatric applications, where small sample sizes and multi-source data are common.

DAIsy addresses this challenge by combining model types. For example, deep learning models may be used to extract features from images or text, while final classifications are made using linear or tree-based models. This layered approach reflects the broader trend of combining flexibility with explainability.

Ethical Alignment and Clinical Relevance

Ethical concerns are a central theme in expert-oriented publications such as Starke et al. (2022) and Ostojic et al. (2024). These works underline the importance of aligning AI tools with clinical guidelines, professional values, and patient rights. The literature highlights that explainability is not only a technical feature but also a safeguard for ethical deployment.

In DAIsy, this principle is reflected in the co-design of tools with clinicians, in the documentation of model behavior, and in the integration of feedback loops. AI outputs are intended to support, not replace, human decision-making, reinforcing ethical expectations found in the literature.

Multi-modality and Real-World Constraints

Several studies, including De Franceschi et al. (2025) and Kerz et al. (2023), emphasize the increasing use of multi-modal data, such as combining neuroimaging, environmental variables, and behavioral patterns. These studies report that such combinations improve prediction accuracy, but also increase the complexity of interpretation.

DAIsy follows a similar path. Use cases include data from structured forms, wearable sensors, speech, and medical imaging. The trend toward multi-modality is matched by a commitment to modular system design, enabling clinicians to see how each data type contributes to the overall decision.

5.3 Gaps and Challenges

Despite the growing interest in explainable AI for mental healthcare, the literature reveals several ongoing challenges that are also relevant to the DAIsy project:

- One of the most common gaps is the limited availability of high-quality, clinically validated datasets, especially for conditions such as eating disorders or early-stage depression. Many studies rely on small or highly specific samples, which can affect model generalizability.

- Another challenge is the lack of consensus on how to evaluate explainability in clinical settings. While tools like SHAP and LIME are widely used, there is no standard metric for measuring whether explanations are meaningful to clinicians or patients.
- Balancing model performance with interpretability also remains a known limitation. More complex models may achieve higher accuracy but are often difficult to explain. Conversely, simpler models can be easier to interpret but may underperform in some scenarios.
- Finally, integration into clinical workflows is an area where many systems fall short. Studies often demonstrate model performance in isolation, but fewer provide evidence of successful use in real-world mental health settings.

DAIsy addresses these challenges through a modular approach, stakeholder feedback loops, and a focus on ethical alignment. However, the issues raised in the literature continue to shape ongoing design and validation efforts.

6. DAIsy Approach and Innovations

6.1 Use Case Examples from Partners

The DAIsy project includes six distinct clinical use cases across Turkey, the Netherlands, and Germany. Each use case targets specific diagnostic or treatment challenges in mental health, particularly focusing on various forms of depressive disorders and eating disorders. These real-world scenarios guide the design, development, and validation of the DAIsy platform, ensuring that technical innovations are grounded in clinical needs and ethical standards.

Turkish Use Case 1 – Major Depressive Disorders

Partners: NP Istanbul Brain Hospital, Vestel Health, Bewell Teknoloji

This use case focuses on enhancing the diagnosis, treatment, and monitoring of patients with severe forms of depression, including melancholic, catatonic, seasonal, and psychotic subtypes. It combines traditional psychiatric evaluations with psychometric scales (such as Beck and SCL-90), and data from wearable devices.

Partners contribute to various components: NP Brain validates AI-supported clinical decision tools; Vestel and Bewell develop monitoring systems using embedded and wearable technologies; Vestel provides secure data management infrastructure.

The integrated system is tested in real hospital workflows, with attention to safety, privacy compliance (KVKK and GDPR), and clinical usability.

Turkish Use Case 2 – Differentiating Bipolar and Unipolar Depression

Partners: ARD Grup (lead), Cerrahpaşa Medical Faculty

This use case addresses the clinically important challenge of distinguishing between bipolar and unipolar depressive disorders. By combining structured psychiatric evaluations with MRI data, the use case supports the development of AI models that assist clinicians in making more accurate and timely diagnostic decisions.

ARD leads model development and deployment, while Cerrahpaşa provides clinical data and validation support. Explainability techniques are integrated to support trust and adoption in real clinical settings.

Netherlands Use Case 1 – Major Depressive Disorder

Partners: AMC (Academic Medical Center), GGZ Oost-Brabant, TU/e, Philips, 5M Software, Semlab, KnowL Solutions

This use case focuses on the use of neuroimaging biomarkers and behavioral data to support personalized treatment decisions in MDD. AMC builds on its DEPREDICT platform, which streamlines the processing of MRI data to predict antidepressant treatment response.

Other partners contribute tools for activity monitoring, clinical data integration, and user-facing dashboards to support treatment planning. The aim is to improve precision in diagnosis and provide interpretable AI insights for clinicians.

Netherlands Use Case 2 – Eating Disorders

Partners: GGZ Oost-Brabant, TU/e, 5M Software, MEDrecord, Semlab, Philips

This use case targets multiple challenges in the treatment of eating disorders, from intake to long-term follow-up.

Key innovations include:

- Automatic summarization of intake interviews (Semlab, MEDrecord),
- Interpretable classification and treatment-response prediction (TU/e),
- Psychoeducation chatbot for patient engagement (Philips),
- Mobile monitoring and personalized feedback app (5M Software).

The overall system aims to reduce clinical burden, improve adherence, and support personalized care with explainable and privacy-aware AI tools.

German Use Case 1 – Multimodal Neurofeedback

Partners: University Hospital Bonn (UKB), OFFIS, BEE Medic, Materna

This use case develops an open-source toolbox for neurofeedback therapy by integrating EEG and fNIRS signals in real time. The system aims to improve self-regulation in patients undergoing therapy by providing visual and acoustic feedback based on brain activity.

The tool fills a significant technological gap by enabling real-time, multimodal neurofeedback and is currently undergoing clinical trials for usability and therapeutic effectiveness.

German Use Case 2 – Virtual Therapy Assistant

Partners: OFFIS (lead), UKB, Ascora, Materna

This use case introduces a virtual assistant designed to support patients with depression outside of clinical settings. The system synthesizes behavioral and sensor data to offer personalized, low-burden support.

It also explores the integration of Large Language Models (LLMs) to provide conversational assistance grounded in clinical guidelines. The goal is to enhance self-reflection and help patients maintain engagement with therapy in their daily routines.

6.2 Innovations and Contributions

DAIsy adopts a modular and clinically grounded approach to the development of artificial intelligence tools for mental healthcare. The project addresses real-world challenges related to depression and eating disorders through a combination of structured clinical data, multi-modal input, and feedback from end users.

A key aspect of the DAIsy methodology is its emphasis on explainability. Rather than focusing solely on performance metrics, the project prioritizes transparency, interpretability, and alignment with clinical practice. This is reflected in the selection of models that balance complexity and clarity, the use of post-hoc explanation tools, and the inclusion of clinician input in interface design and model evaluation.

DAIsy also supports the integration of AI components into existing clinical systems. Technical solutions are developed with attention to security, data protection, and interoperability. In several use cases, models are deployed within hospital environments or on secure backend infrastructures, ensuring that sensitive data remains under institutional control.

The project fosters collaboration across technical and clinical partners. Different use cases contribute unique insights and tools, ranging from speech-based documentation systems and nutrition monitoring applications to imaging-based classification and real-time neurofeedback tools. While each solution is designed for its own context, shared architectural principles enable reuse and adaptation across partners.

In terms of innovation, DAIsy explores model architectures such as Kolmogorov–Arnold Networks for combined neuroimaging and clinical data, the use of transformer models for summarizing clinical dialogue, and mobile applications that deliver AI-assisted support directly to users. These technologies are complemented by user dashboards and visualization tools designed to enhance trust and usability.

6.3 Future Vision

The DAIsy project aims to contribute to the responsible and effective integration of AI in mental healthcare by advancing explainability, trust, and clinical usability. Looking ahead, DAIsy's vision includes expanding its modular platform to accommodate a wider range of psychiatric conditions and care settings.

Future work will explore the scalability of DAIsy components, the integration of emerging data sources, and the refinement of human-AI interaction in clinical environments. Continued collaboration between technical and clinical partners will support the safe adoption of AI-driven tools that are transparent, ethical, and aligned with patient care priorities.

7. Summary and Conclusion

This document has presented an overview of the current state of artificial intelligence in mental healthcare, with a focus on how the DAIsy project fits within this evolving landscape. It outlined the types of AI models and tools being developed and used, the clinical challenges being addressed, and the importance of ethical, interpretable, and trustworthy solutions.

The DAIsy project contributes to this field by applying AI in ways that are clinically relevant, explainable, and privacy-conscious. Its six use cases demonstrate how different types of data, including medical imaging, clinical assessments, behavioral indicators, and wearable sensor information, can be brought together to support care in real settings.

Across all activities, DAIsy prioritizes the integration of AI into existing clinical workflows. The project promotes the use of tools that are understandable by healthcare professionals, support informed decision-making, and respect patient safety and data protection standards.

Looking ahead, DAIsy provides a strong foundation for further development in this area. The insights gained through its multidisciplinary and international collaborations can support future efforts to make AI more usable, effective, and responsible in mental health services.

8. References

Byeon, H. (2023). *Advances in machine learning and explainable artificial intelligence for depression prediction*. International Journal of Environmental Research and Public Health.

de Oliveira, T. A., et al. (2025). *Effect of explainable AI on trust of mental health professionals in an AI-based system for suicide prevention*. Journal of Affective Disorders.

De Franceschi, L., et al. (2025). *Ensemble KAN: Leveraging Kolmogorov-Arnold Networks to discriminate individuals with psychiatric disorders from controls*. NeuroImage: Clinical.

Ghosh, S., et al. (2024). *Review of machine learning solutions for eating disorders*. Frontiers in Psychiatry.

Joyce, D. W., et al. (2023). *Explainable artificial intelligence for mental health through transparency and interpretability for understandability (TIFU)*. Journal of Medical Internet Research.

Kerz, E., et al. (2023). *Toward explainable AI for mental health detection based on language behaviour*. Computers in Human Behavior.

Ostojic, L., et al. (2024). *The challenges of using machine learning models in psychiatric research*. Translational Psychiatry.

Sheu, Y.-H. (2020). *Illuminating the black box: Interpreting deep neural network models for psychiatric research*. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging.

Starke, G., et al. (2022). *Explainability as fig leaf? An exploration of experts' ethical expectations towards machine learning in psychiatry*. AI & Society.