MESG

MESTRADO EM ENGENHARIA DE SERVIÇOS E GESTÃO

Design and Evaluation of an Explainable Clinical Decision Support System for Depression-Related Triage Based on Digital Phenotyping

Francisco dos Santos Andrade

Master Thesis

Supervisor at FEUP: Prof. Fernando Cassola

Supervisor at INESC TEC: Dennis Paulino

U.PORTO

FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

2026-06-29

Abstract

Depression constitutes a significant public health problem and poses substantial challenges to mental health services, especially due to high demand, the episodic nature of traditional clinical assessment, and the limitations of retrospective self-reports. Digital phenotyping, combined with Machine Learning, has been proposed as a scalable way to complement these methods through the collection of passive behavioral signals. However, although smartphone and wearable data have been widely studied, human-computer interaction dynamics, such as mouse movement patterns, remain less explored in clinical contexts and are rarely integrated into explainable systems for clinical decision support.

This dissertation developed and evaluated an explainable clinical decision support system for the prioritization and review of cases related to depression, combining psychological self-reports and passive mouse dynamics features. The work followed a Design Science Research approach and produced three main artifacts: a predictive layer for prioritization (artifact A), an explainability layer based on TreeSHAP (artifact B), and a professional-oriented dashboard prototype (artifact C). The longitudinal analysis investigated intraindividual associations between mouse signals and future depressive severity through mixed linear models. The predictive pipeline combines subscales of the Depression Anxiety Stress Scales 21 items (DASS-21) with passive features extracted during the interaction with the Emotion Regulation Strategy Scale (RESS) questionnaire, evaluated through nested cross-validation to ensure generalization estimates without data leakage. The second artifact is an explainability layer based on TreeSHAP, which produces local and global explanations with high fidelity and completeness for five-factor representations. The third is a professional-oriented dashboard prototype, integrating a prioritization queue of artifact A's outputs, a case view, and an explanation layer from artifact B, formatively evaluated with psychologists through a think-aloud protocol, open-ended feedback, System Usability Scale (SUS), and adapted items from Explainable Artificial Intelligence (XAI) evaluation.

The results show that mouse dynamics did not exhibit statistically significant longitudinal associations with future depressive severity in the small available clinical sample, suggesting limitations in the temporal density of the data and in the ability to capture intra-individual variation. In contrast, the hybrid approach based on Random Forest achieved an Area Under the Receiver Operating Characteristic Curve (AUROC) of 0.856, higher than the baseline of self-report alone (0.772) and the exclusively passive configuration (0.790), suggesting a complementary value of passive signals when combined with self-reports. The Top-5 explanations preserved almost entirely the model's behavior, but formative evaluation showed that clinical utility mainly depends on the contextualization of passive signals and the integration of additional information. Thus, the dissertation contributes to the literature on digital phenotyping and clinical decision support by articulating hybrid modeling, explainability, and interface design in a system evaluated with mental health professionals, aimed at supporting review, screening, and referral, rather than automated diagnosis.

Sumário

A depressão constitui um problema relevante de saúde pública e coloca desafios significativos aos serviços de saúde mental, sobretudo devido à elevada procura, à natureza episódica da avaliação clínica tradicional e às limitações dos autorrelatos retrospetivos. A fenotipagem digital, aliada à aprendizagem automática, tem sido proposta como uma forma escalável de complementar estes métodos através da recolha de sinais comportamentais passivos. No entanto, embora os dados de smartphones e wearables tenham sido amplamente estudados, as dinâmicas de interação pessoa-computador, como padrões de movimento do rato, permanecem menos exploradas em contextos clínicos e raramente são integradas em sistemas explicáveis de apoio à decisão clínica.

Esta dissertação desenvolveu e avaliou um sistema explicável de apoio à decisão clínica para a priorização e revisão de casos relacionados com depressão, combinando autorrelatos psicológicos e características passivas de interações do cursor. O trabalho seguiu uma abordagem de *Design Science Research* e produziu três artefactos principais: uma camada preditiva para priorização (artefacto A), uma camada de explicabilidade baseada em *TreeSHAP* (artefacto B) e um protótipo de *dashboard* orientado para profissionais (artefacto C). A análise longitudinal investigou associações intraindividuais entre sinais de rato e severidade depressiva futura através de modelos lineares mistos. A camada preditiva combina subescalas da Escala de Depressão, Ansiedade e *Stress* com 21 itens (DASS-21) e características passivas extraídas durante a interação com o questionário da Escala de Estratégia de Regulação Emocional (RESS), avaliadas por meio de validação cruzada aninhada para garantir estimativas de generalização sem fuga de informação entre treino e teste. O segundo artefacto é uma camada de explicabilidade baseada em TreeSHAP, que produz explicações locais e globais com elevada fidelidade e completude para representações de cinco fatores. O terceiro é um protótipo de *dashboard* orientado para profissionais, integrando fila de priorização baseada nos resultados produzidos pelo artefacto A, vista de caso e camada de explicação do artefacto B, avaliado formativamente com psicólogos através de protocolo *think-aloud*, perguntas abertas, Escala de Usabilidade do Sistema (SUS) e itens adaptados de avaliação de Inteligência Artificial Explicável (XAI).

Os resultados mostram que as dinâmicas do cursor não apresentaram associações longitudinais estatisticamente significativas com severidade depressiva futura na pequena amostra clínica disponível, sugerindo limitações na densidade temporal dos dados e na capacidade de captar variação intraindividual. Em contraste, a abordagem híbrida baseada em Random Forest obteve um Área sob a Curva Característica de Operação do Recetor (AUROC) de 0,856, superior ao modelo de referência de autorrelato isolado (0,772) e à configuração exclusivamente passiva (0,790), sugerindo valor complementar dos sinais passivos quando combinados com autorrelatos. As explicações Top-5 preservaram quase integralmente o comportamento do modelo, mas a avaliação formativa mostrou que a utilidade clínica depende sobretudo da contextualização dos sinais passivos, e da integração de informação adicional. Assim, a dissertação contribui para a literatura em fenotipagem digital e apoio à decisão clínica ao articular modelação híbrida, explicabilidade e design de interface num sistema avaliado com profissionais de saúde mental, orientado para apoio à revisão, triagem e encaminhamento, e não para diagnóstico automatizado.

Acknowledgments

I would like to begin by thanking all those who made it possible to develop this dissertation in collaboration with INESC TEC and, in particular, with the HumanISE research center. Specifically, I would like to express my gratitude to Mr. Dennis Paulino, Mr. Thiago Netto, and Prof. Fernando Cassola for all the accessibility, understanding, availability, readiness, and assistance provided throughout the development of the dissertation. I would also like to thank all the people in the organization who were directly or indirectly involved in the work carried out, or who contributed to its success. Additionally, I would like to thank Dr. Paulo Pimentel, Dr. Liliana Mendes, the students Pedro Mendes, Ricardo Laranjinha and Leonor Pinto, who collaborated on the project, and all the other clinical professionals who dedicated themselves to responding to the questionnaire or participating in the semi-structured interviews.

To Professor Jorge Grenha, I am grateful for the support and the essential guidance during the initial phase of preparing the dissertation. I extend this gratitude to all the professors at FEUP with whom I had the privilege to collaborate and learn, as well as to the entire teaching and academic community that has accompanied my journey from preschool education to the present, and contributed to my academic, civic, and personal growth.

I would also like to thank all my friends and all the people who have crossed my path throughout my life and who, in one way or another, have influenced me, made me reflect, helped me learn, or conveyed something meaningful to me.

Finally, I would like to express my deep gratitude to my family, especially my parents and grandparents, for all their dedication and constant presence.

AI Use Declaration

In accordance with the guidelines issued by the University of Porto (June 2026), the following declaration documents the use of artificial intelligence tools during the preparation of this dissertation.

Table 1 AI Use Declarion

Field	Detail
Tools and versions	Claude Sonnet (Anthropic, claude.ai); OpenAI Codex o3 (ChatGPT Plus); Google Gemini Pro; Google NotebookLM; Figma Make
Period of use	December 2025-June 2026
Purpose	Academic writing support and linguistic revision; Conceptual explanations with examples; structural and argumentative feedback on drafted sections; literature organisation support; supervised source verification (NotebookLM); dashboard prototype visual design support (Figma Make); coding assistance for data analysis pipeline (Codex)
Parts of the work affected	All written chapters received linguistic and structural revision; dashboard mockups were developed with visual design support; coding assistance for data analysis pipeline (Codex)
What was generated	Brainstorming; Reformulation suggestions, thematic synthesis drafts from interview transcripts, subsequently verified and corrected by the author against original transcript, interview protocols, alternative phrasings, structural outlines, code snippets, and visual design elements
Human verification	All factual content, methodological decisions, analyses, references, reported metrics, and conclusions were independently verified by the author. All AI-generated suggestions were critically reviewed, modified where appropriate, and approved by the author before inclusion.
Evidence of process	Conversation logs available upon request

Artificial intelligence tools were used exclusively as support tools for writing, programming, design, and quality assurance. All scientific contributions, methodological choices, analyses, interpretations, and conclusions presented in this dissertation are solely the responsibility of the author.

Table of Contents

List of Tables

List of Figures

List of abbreviations

ACC — Accuracy

AI — Artificial Intelligence

AIC — Akaike Information Criterion

AUC — Area Under the Curve

AUROC — Area Under the Receiver Operating Characteristic Curve

BA — Balanced Accuracy

BETA — Black Box Explanations through Transparent Approximations

BIC — Bayesian Information Criterion

CDSS — Clinical Decision Support System

CFS — Correlation-Based Feature Selection

CI — Confidence Interval

CLARA — CLARA framework

COVID-19 — Coronavirus Disease 2019

CRISP-DM — Cross-Industry Standard Process for Data Mining

DALYs — Disability-Adjusted Life Years

DASS-21 — Depression Anxiety Stress Scales (21 items)

DMHTs — Digital Mental Health Tools

DSR — Design Science Research

DSS — Decision Support System

EHR — Electronic Health Records

EMA — Ecological Momentary Assessment

EU — European Union

F1 — F1 score

F2 — F2 score

FEUP — Faculty of Engineering of the University of Porto

GPS — Global Positioning System

GRU — Gated Recurrent Unit

HCI — Human–Computer Interaction

ICC — Intraclass Correlation Coefficient

INESC TEC — Institute for Systems and Computer Engineering, Technology and Science

JSON — JavaScript Object Notation

LIME — Local Interpretable Model-agnostic Explanations

LLM — Large Language Model

LMM — Linear Mixed Model

LSTM — Long Short-Term Memory

MAILA — Micro-Movement AI for Latent Assessment

ML — Machine Learning

NASSS — Non-adoption, Abandonment, Scale-up, Spread, Sustainability

NICE — National Institute for Health and Care Excellence

OCD — Obsessive–Compulsive Disorder

PHQ-9 — Patient Health Questionnaire-9

PRISM — Practical, Robust Implementation and Sustainability Model

Q-Q — Quantile–Quantile

RESS — Regulation of Emotion Systems Survey

RF — Random Forest

RQ — Research Question

SCS — System Causability Scale

SE — Standard Error

SHAP — SHapley Additive exPlanations

SUS — System Usability Scale

TMEA — Task–Modality–Explanation Alignment

UI — User Interface

ULSTMAD — Trás-os-Montes e Alto Douro Local Health Unit

USD — United States Dollar

WHO — World Health Organization

XAI — Explainable Artificial Intelligence

1 Introduction

1.1 Project Context and Motivation

Mental disorders impact one in eight people in the world and represent the largest contributor to the morbidity burden across the European Union (EU) region, impacting about one-third of all citizens annually (World Mental Health Report, 2022; Gutiérrez-Colosía et al., 2019). Depression is a mood disorder that can lead to suicide (Gopalakrishnan et al. 2024), and according to the World Health Organization (World Mental Health Report, 2022), approximately 280 million people in the world have this clinical condition, making it a leading cause of disability (Costantini et al. 2021). Additionally, the economic impact of this global health concern is significant, with projected annual costs in Europe for 2021 estimated at 92 billion euros due to decreased productivity levels (Costantini et al., 2021). This substantial burden occurs amid critical service capacity constraints, with estimates suggesting a significant shortage of trained professionals, suggesting that relying solely on in-person delivery models may not be sufficient to meet the global demand for evidence-based mental health care (Torous et al., 2021; WHO, 2022). This context motivates interest in scalable digital health technologies to augment traditional care.

The application of Machine Learning (ML) techniques in the psychiatry field has been recurrent and a subject of research over the last few decades. These types of applications have the potential to refine clinical management with early diagnosis and disease stratification, selection among drug treatments, treatment adjustment, and prognosis for personalized psychiatric care for each individual (Bzdok & Meyer-Lindenberg, 2018). In this context and leveraging growing smartphone and sensors usage patterns, a new approach coined digital phenotyping emerged in 2016 (Torous et al. 2016). It consists of measuring the human phenotype in real-time through the collection of personal digital data from individuals' personal devices, enabling correlating sensor data with self-reported information, and ultimately to predict individuals' mental states and enhance individual monitoring and intervention (H. Birk and Samuel 2020).

This introduces opportunities for passive, unobtrusive and potentially low-cost tracking and monitoring of individuals' mental health through sensing of behavioral dynamics during digital interactions (Hashia et al., n.d.; Huckvale et al. 2019; Insel 2017; Onnela and Rauch 2016; Torous et al. 2021), as a complement to episodic self-report, which could potentially be translated into clinical decision support systems and implemented in real-world settings (Bzdok and Meyer-Lindenberg 2018; Kamath et al. 2022).

However, despite this potential, translating these signals into concrete clinical interventions faces significant barriers (Bzdok and Meyer-Lindenberg 2018; Insel 2017).

While smartphone-based passive sensing has been widely explored (Bai et al. 2021; Braund et al. 2023; Fadul et al. 2024; Sahoo et al. 2025; Zhan et al. 2026), HCI-based signals such as mouse and keyboard dynamics remain underexplored in clinical contexts. Existing studies on these signals have primarily focused on experimental classification tasks or exploratory behavioral correlations rather than longitudinal monitoring or clinically integrated decision support systems. Recent work such as the MAILA framework (Weilnhammer et al., 2025)

suggests that cursor interaction dynamics can capture within-person behavioral patterns in general populations, but such approaches have not yet been operationalized within clinically validated decision support workflows.

Consequently, the integration of mouse dynamics with repeated self-reports within a hybrid modeling framework, examined through a small clinical dataset and translated into an interpretable clinician-facing decision support system prototype, remains an open research gap.

This gap motivates the present work. Building on recent evidence demonstrating that hybrid active-passive sensing improves predictive power over single-modality approaches (Kamath et al., 2022; Moshe et al., 2021) and that longitudinal within-person modeling captures symptom trajectories more effectively than cross-sectional snapshots (Bzdok & Meyer-Lindenberg, 2018; Weilnhammer et al., 2025; Liang et al., 2019) reflecting the inherently time-varying nature of mental health conditions, this work proposes the development and evaluation of a clinical decision support system (CDSS) driven by human-in-the-loop triage, which combines digital interaction dynamics with self-reports to support prioritization of depression-related cases. In addition to predictive performance, the approach considers key service-level and sociotechnical requirements for clinical decision support tools, including workflow alignment, interpretability, and trust (Abell et al., 2023; Holzinger et al., 2019).

This dissertation was developed within the scope of the DepressionSense project at INESC TEC, part of the broader DeepVybe research initiative, in close collaboration with multidisciplinary researchers (DeepVybe 2026).

1.2 Methodological Framework

Given the artifact-based nature of the project, the Design Science Research (DSR) framework was chosen as the overarching methodological reference. The development of a clinician-facing workflow for depression-related triage and prioritization based on digital phenotyping and machine learning requires the combination of technical development, clinical requirements, and formative evaluation, aligning with the DSR philosophy of creating and evaluating innovative artifacts that solve real-world problems and the processes that led to their construction (Hevner et al., 2004).

This study was informed by the interaction of the three cycles that compose Design Science Research: the relevance cycle, the design cycle, and the rigor cycle (Hevner, 2007; Hevner et al., 2004).

The relevance cycle is based on the real environment of a domain, composed of research requirements and criteria that allow for the evaluation of the results obtained in the surrounding environment. In this dissertation, that environment corresponds to the mental health screening and clinical review context, including the needs of clinicians, and the requirement that the system should support. Initial requirements were informed by the literature and by the project context, while interviews with clinical stakeholders were conducted in parallel with model development and helped refine the interpretation of the use case and the design of the clinician-facing interface.

The rigor cycle is composed of knowledge derived from the state of the art, which is important for ensuring the production of innovation in a rigorous way. In this study, the methodological and design choices were informed by prior work on digital phenotyping,

machine learning in mental health, human-computer interaction, explainable artificial intelligence, and clinical decision support systems.

The design cycle corresponds to the iterative development and evaluation of the dissertation artifacts, using the theories and methods acquired in the rigor cycles and addressing the requirements from the relevance cycles. Three main artifacts were produced iteratively: Artifact A, a predictive modeling layer for depression-related prioritization; Artifact B, an explainability engine based on feature attribution; and Artifact C, a clinician-facing dashboard prototype.

Accordingly, this chapter's structure is organized around the proposed research questions and artifacts. First, the construction of the analytical datasets and feature engineering pipeline is described. Then, the methodology used to develop and evaluate the predictive modeling layer is presented, covering both the longitudinal analysis of mouse dynamics and the cross-sectional machine learning pipeline. The chapter then describes the explainability methodology and, finally, the design and formative evaluation of the clinician-facing dashboard prototype.

1.3 Research Questions

This study was guided by four research questions (RQ1–RQ4), which were defined through an iterative process informed by the literature, the identified research gaps, the characteristics and constraints of the available DepressionSense dataset (DeepVybe 2026), and the evolving design objectives of the project.

The work was guided by the following core research questions:

1. Digital Phenotyping: To what extent are within-session mouse dynamics associated with intra-individual fluctuations in depression symptom severity in a small clinical sample?
2. Hybrid Modeling: To what extent can a hybrid machine learning pipeline combining self-reported and passive mouse-dynamics features support depression-related prioritization compared with questionnaire-only baselines in a small clinical sample?
3. Explainability & Trust: How can feature attribution from a hybrid model, including passive mouse-dynamics features, be integrated into a clinician-facing interface to support explanation and clinical review during depression-related prioritization?
4. Decision Support & Service Integration: How can a digital decision-support interface be designed to support clinicians in reviewing and prioritizing depression-related screening cases?

Methodologically, this study adopts the Design Science Research (DSR) framework to guide the iterative development of the proposed CDSS. The proposed DSS is operationalized through three artifacts: Artifact A (prioritization model), Artifact B (explainability layer), and Artifact C (clinical dashboard prototype).

1.4 Report Outline

The remainder of this dissertation is structured as follows.

Chapter 2, Literature Review, explores the current state of the art in digital phenotyping, machine learning applications in psychiatry, HCI-based behavioral signals, explainable AI, and the sociotechnical requirements of clinical decision support systems. This chapter establishes the theoretical and empirical foundations for the research gap addressed in the dissertation.

Chapter 3, Methodology, presents the Design Science Research approach adopted in the study and describes the data sources, preprocessing strategy, analytical datasets, and artifact development process. It details the methodological procedures used for the longitudinal analysis of mouse dynamics, the hybrid predictive modeling pipeline, the explainability layer based on feature attribution, and the design and formative evaluation of the clinician-facing dashboard prototype.

Chapter 4, Results, reports the outcomes of the empirical and design-oriented components of the study. It first presents the data preparation outcomes and analytical datasets, followed by the results for each research question. These include the longitudinal mixed-effects analysis for RQ1, the nested cross-validation and model selection results for RQ2, the global and local explanation results, objective Top-5 explanation metrics, and clinician feedback on the explanation layer for RQ3, and the formative evaluation of the dashboard prototype for RQ4.

Finally, Chapter 5, Conclusion, summarizes the main findings and contributions of the dissertation, discusses the limitations of the study, and presents directions for future research and development of explainable clinical decision support tools for depression-related prioritization.

2 Literature Review

The literature review is organized around RQ1-RQ4 to ensure that each research objective is grounded in prior evidence and directly informs the design and evaluation strategy of the proposed artifacts.

2.1 Mental Health Screening and the Measurement Gap

Mental disorders constitute a global public health concern affecting nearly a billion people around the world (Rehm and Shield 2019; *World Mental Health Report* 2022). Recent comprehensive estimates indicate that 418 million disability-adjusted life years (DALYs) could be attributable to mental disorders in 2019, representing 16% of the global disease burden (Arias et al. 2022). The associated economic value is estimated at approximately USD 5 trillion, equivalent to regional losses ranging from 4% of gross domestic product in Eastern sub-Saharan Africa to 8% in high-income North America (Arias et al., 2022). Across the EU region, mental disorders collectively impact approximately one-third of citizens annually, with mood and anxiety disorders ranking among the most prevalent conditions (Gutiérrez-Colosía et al. 2019; Louis 2022), accounting for 92 billion euros due to reduced productivity and healthcare expenditures (Costantini et al., 2021).

Within this broader landscape, major depressive disorder, which is characterized by emotional, cognitive, somatic, and behavioral symptoms (National Institute for Health and Care Excellence: Guidelines 2022), affects an estimated 280 million individuals worldwide, contributing substantially to disability-adjusted life years (DALYs) across all age groups (WHO, 2022). According to the same estimates, in 2019, 301 million people globally were living with anxiety disorders.

The COVID-19 pandemic has further underscored the urgency of effective mental health monitoring and intervention, with prevalence estimates indicating that one in four young people under 18 experienced clinically significant depressive symptoms during the outbreak, and 1 in 5 were experiencing clinically elevated anxiety symptoms (Racine et al. 2021).

This substantial burden is compounded by critical service capacity constraints. Mental health problems impact over one billion people worldwide annually (Torous et al., 2021), yet the World Health Organization's acknowledges that barriers to evidence-based care include lack of available services and funding (WHO, 2022). The magnitude of this service gap is illustrated by estimates suggesting that offering evidence-based mental health care in the United States alone would require an additional 4 million trained professionals (Torous et al., 2021). On a global scale, it is not feasible to propose that practices based entirely on in-person care will ever meet this demand, thereby motivating interest in scalable digital health technologies to augment traditional care delivery (Torous et al., 2021).

Despite this dual challenge of substantial burden and constrained service capacity, current screening and monitoring approaches remain further limited by episodic assessment paradigms.

Traditional depression screening is limited by its subjectivity, relying on clinical judgment and patient self-reporting (Huckvale et al. 2019). Factors such as recall bias, symptom heterogeneity, and the episodic nature of standard consultations (typically every 4-6 weeks)

may limit the temporal resolution required to track fluctuations in neurovegetative symptoms and functional impact (Bzdok and Meyer-Lindenberg 2018; Kamath et al. 2022; Shiffman et al. 2008).

Shiffman et al suggest that retrospective self-reports are highly susceptible to recall bias; patients consistently struggle to accurately aggregate and report their mood and behavioral fluctuations over extended periods, often disproportionately recalling only the most recent or intense emotional peaks (Shiffman et al., 2008). Furthermore, depression is a highly heterogeneous condition characterized by rapid fluctuations in emotional, cognitive, and neurovegetative symptoms (Fried & Nesse, 2015). Meta-analytic evidence indicates that commonly used risk factors have limited predictive accuracy for future suicide ideation and behavior (Ribeiro et al., 2016), which may partly reflect the highly dynamic nature of suicide risk, where clinically meaningful changes can occur over short time scales that are rarely captured by episodic clinical assessments.

Thomas Insel (2017) identified a critical flaw in current psychiatric practice: the over-reliance on subjective clinical judgment rather than objective measurement. He highlights that, in the absence of continuous tracking, clinicians may miss signs of deterioration in nearly 80% of patients. Crucially, Insel argues that the 'lack of measurement' in psychiatry is not due to missing biological markers, but rather the failure to objectively track fluctuations in mood, cognition, and behavior: a gap that motivates the shift toward continuous digital phenotyping explored in Section 2.2 (Insel, 2017).

2.2 Digital Phenotyping and Machine Learning in Psychiatry

The limitations of episodic assessments have motivated the exploration of digital technologies capable of capturing behavioral signals continuously and unobtrusively.

In this context and leveraging the greater accessibility of large volumes of data from digital devices, improved computing power, and less expensive data storage (Bzdok & Meyer-Lindenberg, 2018), a new approach coined digital phenotyping emerged in 2016. In the biological sciences, a phenotype is traditionally defined as the observable physical properties of an organism, produced by the interaction of the genotype with the environment (National Human Genome Research Institute 2026). First, Jain et al. (2015) introduced the term 'digital phenotype' to describe the quantification of disease through digital devices, effectively translating the biological concept of expressed traits into the realm of digital data (Jain et al. 2015). In 2016, Torous et al. (2016), formally defined digital phenotyping as the 'moment-by-moment quantification of the individual-level human phenotype in situ using data from personal digital devices'. In broader terms, it consists of measuring the human phenotype in real-time using data from computer usage, smartphones, and personal digital devices to correlate behavioral patterns with clinical states, aiming to measure and/or predict people’s mental health as well as their potential risk for mental ill health.

Torous et al. (2016) operationalized digital phenotyping in psychiatry through a scalable smartphone-based research platform capable of collecting two distinct streams of data simultaneously:

- Active Data: Information requiring user active participation for its generation, such as audio samples and self-report surveys.

- Passive Data: Information collected from sensors (e.g., GPS, accelerometer) and logs (e.g., call and text logs) without any direct involvement from the user.

Their work established an important methodological precedent for combining high-frequency behavioral data with traditional clinical scales while emphasizing transparency, protocol customization, privacy protection, and scientific reproducibility.

Digital phenotyping expanded to leverage diverse data streams, including smartphone sensors (GPS, accelerometer), wearable physiological signals, and computer interaction dynamics (mouse, keyboard), to capture behavioral patterns associated with mental health states attempting to enhance individual monitoring and intervention (Torous et al., 2021; Liang et al., 2019).

A recent systematic review by Choi et al. (2024), analyzing 40 studies of digital phenotyping published between 2010 and 2023, aimed to assess the effectiveness of this approach (through smartphone sensors) in identifying behavioral patterns associated with stress, anxiety, and mild depression in non-clinical adult populations. The results showed that passive sensors, including GPS, accelerometer, microphone, Bluetooth, and light sensors, have the potential to be effective in detecting these states, with 78% of the studies using machine learning models to make predictions.

Machine learning is generally defined as the ability of computational systems to learn patterns (i.e., parameters) from data and hence reach an optimal solution to a problem without being explicitly programmed by a human with rules for it. It is considered a subfield of artificial intelligence, and its application to the fields of psychology and psychiatry has been the subject of research for several decades. By modeling complex, high-dimensional data, ML has been facilitating the transition to tailored medicine at the individual level. It optimizes diagnoses, prognoses, and treatment outcome predictions, as well as enhancing the detection of biomarkers. Importantly, these methods offer an alternative to traditional statistical approaches based on group-level averages, which often struggle to capture heterogeneity and achieve consistent replication in psychiatric research (Dwyer et al., 2018). Bzdok also points to the ML applications' potential to refine clinical management with early diagnosis and disease stratification, selection between drug treatments, treatment adjustment, and prognosis for psychiatric care tailored to each individual, while providing a useful primer on these opportunities and challenges for clinical translation (Bzdok & Meyer-Lindenberg, 2018). Expanding on this potential, Bzdok and Meyer-Lindenberg (2018) argue that ML enables a fundamental shift towards 'Precision Psychiatry' by offering specific opportunities: single-subject prediction; redefining clinical subgroups; analysis of observational and contextual data; multi-class and multi-task predictive capability; and scalability and cost.

By 2018, the adoption of deep learning in psychiatric research remained limited, largely because such models require substantial volumes of data to mitigate the curse of dimensionality and reduce overfitting (Dwyer et al. 2018).

The limitations of episodic assessments are common in some extent also to computational approaches. Bzdok and Meyer-Lindenberg (2018) warned that psychiatric research bears a “blind spot” regarding disease trajectories, as most ML applications focus on static, cross-sectional findings (Bzdok & Meyer-Lindenberg, 2018). This warning remains pertinent today. Recent literature confirms that despite the availability of sensors, the field has yet to fully transition to continuous monitoring. Dhanda et al. (2025), in a comprehensive review of public health ML, explicitly list 'Longitudinal Health Data Analysis' and 'Real-Time

Surveillance' as priority future directions rather than established standards. They note that while the shift from retrospective analysis to proactive, real-time surveillance through the adoption of explainable AI models is underway, it faces challenges in data granularity and consistency (Dhanda et al., 2025).

Consequently, the literature suggests that moving toward genuinely longitudinal modeling paradigms, potentially implemented through temporal deep learning architectures such as LSTMs or GRU (achieved high accuracies between 88-95%) operating on continuous sensor streams, is not merely a technical refinement, but a necessary evolution for capturing disease trajectories (Dhanda et al. 2025).

A structured review was carried out to identify primary empirical studies that implement and evaluate models based on passive digital signals for mental health assessment and monitoring, using relevant clinical outcomes (e.g., depression, anxiety). The evidence matrix was constructed to support the formulation and refinement of the research questions, particularly by examining the extent to which empirical studies had evaluated passive HCI-based interaction dynamics in longitudinal analyses or in relation to clinician-facing decision-support workflows for depression-related assessment. The selection was supplemented by citation tracking and iterative screening of results. As this is a structured, non-exhaustive review, it does not constitute definitive proof of the nonexistence of all possible implementations. Studies were grouped according to their primary sensing modality, although several digital phenotyping studies are inherently multimodal.

Additional details on the structured review strategy are provided in Appendix A.

Table 2 Structured overview of passive digital sensing studies for mental health assessment

Sensing modality	***Associated studies***	***Design***	***Longitudinal Analysis***	***Clinical Target & Ground Truth***	***Key metrics / results***	***Workflow Integration / CDSS***
Smartphone / wearable passive sensing	Bai et al., 2021; Moshe et al., 2021; Ware et al., 2020; Zhan et al., 2026; Zhan & Wang, 2026; Sahoo et al., 2025	Mostly longitudinal observational or prospective studies	Mostly yes (5/6), although the depth of temporal modeling varies	Depression, anxiety, stress, mood fluctuation, symptom severity, and treatment response, using repeated scales, EMA, clinical interviews, or treatment-response criteria	Moderate predictive or explanatory performance; examples include ACC ≈ 76.7%, AUC up to 0.84, F1 up to 0.77, and symptom-level F1 up to 0.86	Predictive Modeling Only
Smartphone keystroke dynamics	Fadul et al., 2024; Braund et al., 2023	Cross-sectional or short prospective studies using keyboard/typing metadata	No (0/2)	Depressive tendency and mental-health-related symptoms, using PHQ-9, and mental health-related self-report scales	Results range from very high performance in a very small clinical sample to weak or non-significant associations in a larger adolescent	Predictive Modeling Only

Mouse/cursor & touchscreen interaction dynamics	Zantvoort et al., 2025; Monaro et al., 2018; Weilnhammer et al., 2025 / MAILA	Heterogeneous designs, including clinical web-based assessment, laboratory tasks, and large-scale online studies	Mixed (1/3)	Treatment dropout, malingered vs. genuine depression, depression classification, and symptom change	sample Mixed results: modest clinical web performance, high laboratory classification accuracy, and moderate large-scale HCI prediction	Predictive Modeling Only

As illustrated in Table 2, a methodological divergence exists in the current landscape. Smartphone and wearable-based systems (e.g., Bai et al., 2021; Moshe et al., 2021; Ware et al., 2020; Zhan et al., 2026; Zhan & Wang, 2026; Sahoo et al., 2025; Braund et al., 2023 Fadul et al., 2024) more frequently adopt prospective or longitudinal designs and examine depression-related monitoring, symptom dynamics, or treatment response over time. Conversely, studies exploring Human-Computer Interaction (HCI) signals, specifically mouse and keyboard dynamics (e.g., Monaro et al., 2018; Zantvoort et al., 2025; Weilnhammer et al., 2025), remain underexplored and predominantly confined to cross-sectional designs, general population samples, or controlled experimental tasks.

This contrast reveals a critical translational gap. Despite the ubiquity of desktop computer interactions in occupational and domestic settings, their potential as continuous clinical biomarkers remains methodologically underexplored compared to mobile sensing. In addition, there seems to be limited consideration of workflow integration and support-oriented design in the realm of digital phenotyping. These disparities motivate a deeper examination of HCI-based behavioral biomarkers and the specific challenges hindering their longitudinal clinical application, as explored in the following section.

2.3 Mouse Dynamics as Behavioral Biomarkers

Motor behavior, including cursor-based interaction patterns, may provide an indirect behavioral proxy for mental status. This clinical plausibility is supported by evidence that psychomotor retardation, manifests as slowed fine motor movements, increased movement initiation time, and reduced kinematic efficiency (Bennabi et al. 2013).

The feasibility of using interaction dynamics to detect multiple emotional states in human-computer interactions was established early on by Epp et al. (2011), who demonstrated that keystroke dynamics could effectively classify emotions with accuracies ranging from 77.4 to 87.8% (Epp et al. 2011). Similarly, HCI research has shown that negative emotion can influence mouse cursor distance and speed during digital interaction, supporting the broader assumption that cursor dynamics may capture affective or psychomotor variation during task performance (Hibbeln et al. 2017). Yamauchi's 2013 study investigated the association between state anxiety and mouse activities using a large number of participants (N = 367 in two experiments) in a mundane perceptual task. This study demonstrated that the distributions of certain features, such as path lengths, velocity, and spatial features, provide important clues for anxiety detection, suggesting the feasibility of mouse trajectory analysis as a non-intrusive behavioral signal for affect-related states (Yamauchi 2013).

A machine learning approach relying on mouse and keystroke dynamics reported strong performance mainly for pleasant mood, while unpleasant mood was harder to detect, and the dataset was limited (10 participants), suggesting potential for mood-state detection (Aldrich et al., 2023). Yamauchi (2018) investigated cursor motion as an affective signal in choice-reaching tasks. The study extracted several trajectory features (for example, area under the curve and direction changes) and found that models trained on known participants could explain roughly 10–20% of the variance in positive affect and attentiveness for unknown participants.

Beyond affective states, studies on Mild Cognitive Impairment suggest that passive computer-use and mouse-dynamics measures can capture clinically relevant neurocognitive variation, including less efficient and more variable trajectories, longer pauses, and associations with executive functioning and attention, supporting their broader value as unobtrusive behavioral proxies (Kaye et al. 2014; Seelye et al. 2015). Similarly, stress-related studies suggest that mouse dynamics can function as unobtrusive behavioral proxies for stress (Banholzer et al. 2021; Sun et al. 2014).

Monaro et al. (2018) investigated whether mouse movement dynamics collected during a computerized depression questionnaire could distinguish genuine depressive responses from simulated ones. Their results showed that individuals with clinically consistent depression profiles exhibited slower and more variable mouse trajectories compared to simulators, reflecting patterns compatible with psychomotor slowing. Although conducted in a controlled task setting, this study provides clinically relevant evidence that mouse dynamics may encode depression-related motor signatures beyond self-report content (Monaro et al., 2018).

One of the most comparable conceptual works to this project is the CLARA framework. CLARA proposes an architecture that integrates HCI-derived signals (cursor, mouse, keystroke features) with repeated self-report instruments such as the DASS-21 in the workplace (Lee at al., 2019). CLARA is primarily conceptual: it lays out how HCI signals could be combined with repeated questionnaires within a screening workflow, but it has not yet been empirically validated as a working clinical decision support system (CDSS) (Lee et al. 2019).

More recently, Weilnhammer et al. (2025) demonstrated that it is possible to infer latent mental states (such as levels of distress, well-being, and self-reported diagnoses of depression or OCD) solely through the analysis of cursor and touchscreen movement patterns collected passively. Using a large dataset, their innovative framework (MAILA) consists of an unsupervised representation learning (using LSTM autoencoders and clustering) to segment raw cursor trajectories into a dictionary of recurring movement patterns, which are then used to predict psychological profiles via support vector regression. MAILA achieved cross-sectional correlations of $R = 0.20$ across mental health dimensions, an AUC of 0.64 for classifying individuals with a self-reported history of depression against the general population, comparable to neuroimaging-based markers, and an AUC of 0.73 for discriminating within-person deterioration from improvement across sessions. Predicting how much a person changed over time substantially outperformed predicting their state at a single point ($R = 0.48$ vs. $R = 0.20$), but only when a baseline cursor recording from the same participant was available; without it, performance dropped to $R = 0.15$, highlighting that repeated measurement from the same individual may carry meaningful additional information and that designing a model architecture that embeds temporal sequences may be beneficial.

This research provides evidence that passive cursor and touchscreen dynamics can encode information about self-reported mental states and their changes over time.

2.4 Hybrid and Multimodal Modeling Strategies

Kamath et al. (2022) demonstrated that the integration of passive sensing (EMA passive) and self-report depression scores (EMA active) showed better prediction power compared to passive or active EMA alone. Their LifeRhythm project demonstrated that integrating digital data enables the prediction of depressive symptoms through the combination of active and passive data. This study followed 182 university participants over eight months, using smartphone apps to collect passive data on location and social interaction, as well as Fitbit devices to monitor sleep and physical activity. By linking these automatic records to self-report questionnaires conducted every two weeks, the study concluded that the fusion of active and passive data has greater predictive power than each method analyzed individually. Similarly, Moshe et al. (2021) combined smartphone and wearable features with repeated self-reports (DASS-21) and found that multimodal models provided stronger prediction of depression symptoms than single-source feature sets (Moshe et al. 2021).

Given these studies and the emphasis on multimodal fusion in Liang et al. (2019), combining passive interaction signals (e.g. mouse dynamics) with active assessments (DASS-21/RESS) aligns with recommended hybrid designs for improving predictive power and clinical utility.

2.5 Technical Barriers in ML-based Digital Phenotyping

While digital phenotyping and machine learning offer promising pathways for continuous mental health monitoring, there are many common technical pitfalls that hinder clinical translation as noted by Bzdok and Meyer-Lindenberg (2018) and Ostojic (2024).

One of the most significant obstacles is the small size of current datasets; the authors note that psychiatry is still far from the threshold of 1 million examples needed for the full success of certain models (Bzdok & Meyer-Lindenberg, 2018; Dwyer et al., 2018). At the time of their review, many studies had relatively small (i.e., number of subjects) and insufficient phenotypic detail (e.g., medical history, comorbidities, progression in symptoms, treatment, and response), which limits the generalizability of complex models. While Bzdok and Meyer-Lindenberg identified this gap in 2018, Gopalakrishnan et al. (2024) confirm that this challenge remains unresolved, highlighting a persistent friction between small studies that prove the concept and large studies that show resilience and generalizability (Gopalakrishnan et al., 2024).

With small samples, models tend to "hallucinate" or learn the "noise" in the data rather than real patterns, which impairs the algorithm's ability to work with new patients. Ostojic et al. (2024) also argue that a large number of variables (features) for a small number of patients increases the data dispersion and the risk of overfitting exponentially (Ostojic et al., 2024).

Data leakage occurs when test data 'contaminates' the training process, leading to overly optimistic performance. This can arise not only from explicit leakage (e.g., performing feature selection prior to cross-validation), but also from the use of non-nested cross-validation, where the same data are used both for model selection or hyperparameter tuning and for performance estimation (Ostojic et al., 2024).

In psychiatric screening, datasets are often imbalanced (e.g., far more healthy controls than depressed patients). Ostojic et al. (2024) warn that standard accuracy metrics can be misleading in these scenarios, as models may simply predict the majority class. In addition, the same authors emphasize that development samples must proportionally reflect the target population to avoid sampling bias. Failure to do so introduces confounding factors (e.g., age, culture) and reinforces systemic discrimination. They warn that common handling methods, such as excluding participants with incomplete data, often result in biased estimates and a reduction in statistical power. Furthermore, there is great difficulty in ensuring that the models function consistently across different clinical centers, geographic populations and different types of hardware (e.g., different brands of brain scanners). Ostojic et al. (2024) also emphasize that models developed and tested on single local datasets do not guarantee functionality across different contexts. A model is only clinically useful if it can generalize predictions to new patients outside the original sample.

Finally, there is an inherent difficulty in explaining how complex algorithms (e.g., neural networks) reach a decision, which makes difficult the process of identifying the underlying biological mechanisms of the disease. Ostojic et al. (2024) warn that without interpretability, clinicians struggle to trust the system, as concluded by Tonekaboni et al. (2019). These authors emphasize that for clinical translation, accurate predictions are insufficient; the ability of an ML model to justify its outcomes and assist clinicians in rationalizing the model prediction is critical to establish trust (Tonekaboni et al. 2019). Addressing this interpretability gap through explainable AI methods is examined in Section 2.6, while the methodological safeguards addressing data quality and validation challenges adopted in the project are described in Chapter 3.

2.6 Explainability and Trust in Clinical Decision Support

While predictive accuracy is essential, the "Black Box" nature of advanced algorithms can be a primary barrier to clinical adoption as it compromises trust within clinical practitioners. Therefore, it is necessary to understand how a machine decision was made and to assess the quality of the explanation. Explainable AI (XAI) is the field dedicated to improving the clarity and comprehension of ML models, ensuring that algorithmic outputs are not just accurate, but actionable and trustworthy (Dhanda et al., 2025). In the health scope, it aims to allow healthcare providers to understand the forecasts generated by these models, and it is often a requirement for a Decision Support Systems (DSS).

This theoretical need is supported by empirical findings. Tonekaboni et al. (2019) surveyed clinicians in acute care settings and found that "trust" is not binary but contextual. Clinicians do not need to know every weight in a neural network; rather, they require justifiable outcomes: explanations that help them rationalize the model's prediction (e.g., risk scores, or other notifications concerning a patient’s status) (Tonekaboni et al., 2019).

Rosenbacke et al. (2024) provide a systematic review on how explainable AI influences clinician trust in clinical decision support systems. Across the included studies, the evidence is not uniformly positive: explainability can support trust and perceived usefulness in some settings, but effects are often context-dependent and sensitive to how explanations are presented and integrated into decisions. Although the majority of studies suggested that XAI has the potential to enhance clinicians’ trust in AI, this reinforces the view that “trust” is not a

fixed outcome of adding XAI, but an interaction between explanation design, clinical relevance, and workflow constraints (Rosenbacke et al. 2024).

Holzinger et al. (2019) distinguish between explicability (technically explaining how the model works) and causability (explaining the result in a way that aligns with human causal reasoning, and to what extent that explanation was effective for understanding). They argue that for a doctor to trust an AI, the system must map its mathematical findings to clinical concepts, allowing the expert to validate the prediction against their medical knowledge, acting as a transparent filter, instead of replacing the clinician (Holzinger et al. 2019).

The article distinguishes two types of systems for achieving explainability (Holzinger et al., 2019):

- Post-hoc Systems: Aim to provide local explanations for specific decisions of existing models (e.g., LIME, BETA, SHAP), allowing the visualization of contributions of pixels or features to a prediction (S. Lundberg and Lee 2017; M. T. Ribeiro et al. 2016).
- Ante-hoc Systems: Are interpretable models by design (e.g., linear regression, decision trees, fuzzy inference systems), which are inherently more transparent.

Using metrics like the System Causability Scale (SCS) proposed by Holzinger et al. (2019), researchers can quantify the perceived quality of explanations in a human-AI interface beyond standard performance metrics and from a human-centered perspective (Holzinger et al. 2020).

Beyond explanation methods themselves, successful clinical adoption requires that XAI should be embedded within existing clinical workflows. Shulha et al. (2024) demonstrated, through a modified design thinking approach involving clinicians, that explainability features must align with time constraints, decision points, and clinical narratives (Shulha et al. 2024). This highlights that explainability is not solely a model-level property, but also a service/systems' design challenge within DSS.

Recently, Bergomi et al. (2025) investigated the reasons behind clinicians' attitudes toward explainable AI (XAI) methods applied with predictive techniques, commonly found in clinical settings. Comparing three XAI techniques, namely SHAP, Bayesian networks, and AraucanaXAI, using real-world clinical cases, they discovered that understandability and actionability dimensions, which they defined as how easily a clinician could intuitively grasp the explanation and the extent to which an explanation enabled an informed decision, respectively, have a statistically significant positive association (i.e., positive correlation). In addition, they found that clinicians often prefer simpler feature-attribution explanations such as SHAP but also highlighted the risk of increased compliance and automation bias (86% agreement, even with errors) when explanations are perceived as overly authoritative. The findings reinforce that explainability must be evaluated not only in terms of transparency, but also in terms of its impact on clinical judgment and decision behavior (Bergomi et al. 2025).

2.7 Clinical Translation and Service Integration

The landscape of Digital Mental Health Tools (DMHTs) has expanded rapidly over the last decade (Torous et al., 2021). Despite this rapid growth, evidence and real-world adoption have not progressed at the same pace. The Banbury Forum consensus highlights a persistent deployment gap in digital mental health in the US, where tools may show promise in controlled settings but struggle to achieve sustained uptake and impact in routine care due to

limited workflow integration. In practice, effectiveness depends not only on predictive performance or app features, but also on how the technology is operationalized as part of a service with defined roles, escalation pathways, and ongoing monitoring (Mohr et al. 2021).

Abell et al. (2023) reviewed barriers and facilitators to CDSS implementation in hospitals using the NASSS framework. Their findings show that successful implementation depends less on model performance alone and more on sociotechnical fit, particularly workflow alignment, usefulness of outputs, technical integration, users' trust in the evidence base and input data, and the fit between the CDSS and clinicians' roles. The review also highlights recurring barriers such as poor interface design, alert fatigue, interoperability problems, disruption of clinical workflow routines, and concerns about professional autonomy, reinforcing that CDSS adoption requires implementation planning, user training, and being supportive of clinical judgment rather than prescriptive (Abell et al. 2023).

To better specify these operational requirements, Hopkin et al. (2025) propose a conceptual framework that characterizes DMHTs across domains relevant for implementation. In a triage-focused CDSS, explicitly defining the system's Purpose (triage vs. diagnosis) and expected Human Responsiveness (timeliness and responsibility for follow-up) can support clearer workflow integration and safer escalation policies. Similarly, characterizing the requirements for Functionality and Human Interaction helps inform system design decisions, including dashboard/workflow integration points and realistic service-level expectations (Hopkin et al. 2025).

Building on this service-oriented framing, Trinkley et al. (2020) argue that successful clinical decision support requires combining established CDSS design best practices with an implementation science lens. By integrating the PRISM model with CDSS design principles, the authors outline an iterative process that spans stakeholder engagement, interface design, usability testing, deployment, and ongoing performance evaluation and maintenance. Importantly, this perspective reinforces that operational success depends on minimizing workflow disruption and cognitive burden, preventing alert fatigue through careful targeting of notifications, and preserving clinical autonomy by keeping recommendations actionable rather than prescriptive. Complementing these implementation-oriented recommendations, Bayor et al. (2025) employed a systematic review of CDSS design research approaches through a user-centered lens and highlight a recurring gap between usability, focused prototypes and real-world integration. Across the reviewed studies, user-centered design practices are widely adopted; however, many CDSS remain stand-alone tools, with insufficient attention to interoperability and embedding into existing clinical systems and workflows. The authors identify key design challenges spanning usability and user experience, perceived validity and reliability, data quality assurance, and the complexity of integration into practice (Bayor et al. 2025).

Empirical evidence further illustrates how integration challenges materialize at the interface level in clinician-facing triage scenarios. Shulha et al. (2024) proposed a methodology intended to support adoption-oriented design to overcome the barriers to the adoption of Artificial Intelligence (AI) in clinical practice, using a COVID-19 prognosis case study. The authors adapted the traditional Design Thinking process (Empathize, Define, Ideate, Prototype, Test) by adding two theoretical layers-NASSS and a trust-oriented evaluation framework-to explicitly account for feasibility and adoption-related requirements early in the design cycle. At the same time, their iterative prototyping illustrates the interface-level

integration issues highlighted in CDSS design research, where workflow alignment depends on how risk information is structured and accessed. Their findings also emphasize that the alignment of AI with human and clinical values depends on the integration of end users throughout the entire design process, and not just in final validation. In addition, they highlight operational risks such as cognitive overload from overly dense explanations and the potential misinterpretation of contextual patient information as model inputs, reinforcing the need for clear UI separation between “clinical context” and “features used by the model” in CDSS dashboards (Shulha et al., 2024). Sunjaya et al. (2022) applied a design thinking approach to develop and test a primary care CDSS, combining low-fidelity wireframing with repeated think-aloud evaluations and incremental interface refinement. This work provides a pragmatic blueprint for early-stage DSS development: test interface concepts rapidly with end users, identify points of workflow disruption, and iteratively adjust the presentation and timing of decision support outputs to improve actionability and acceptance (Sunjaya et al. 2022).

Taken together, this literature suggests that clinician-facing CDSS should be evaluated not only by predictive performance, but also by their ability to fit clinical workflows, present actionable and non-prescriptive outputs, preserve clinician autonomy, and minimize cognitive burden. These considerations directly inform the design and formative evaluation of the DepressionSense CDSS prototype.

2.8 Ethical and Sociological Considerations

Birk, R. H., & Samuel, G (2020) sought to close the gap in research on the fact that, despite increasing research and efforts to digitally track and predict poor mental health, there had been little sociological and critical engagement with this field before 2020. Thus, they developed their work aiming to introduce and explore digital phenotyping advancements and formulate three main critiques of the technology behind it (Birk & Samuel, 2020): these approaches may embed assumptions about “normal” behavior, reduce complex social life to technical proxies, and use biological language that risks reifying mental health problems while underemphasizing social, cultural, and economic context.

Beyond these foundational sociological critiques, as digital phenotyping approaches clinical implementation, recent literature highlights some additional ethical challenges such as the bioethical principle of informed consent, data leakage, the novel ethical dilemma of "bystander privacy", as sensors might inadvertently capture data from individuals around the patietn (Kamath et al., 2022), secondary data uses abuse (e.g., employers accessing mental health inferences) (Choi et al. 2024; Insel 2017). Moreover, deploying machine learning models trained on non-representative datasets risks perpetuating algorithmic bias, which could exacerbate existing healthcare disparities for marginalized groups (Dhanda et al., 2025; Huckvale et al., 2019), and induce or exacerbate symptoms of certain vulnerable patients from knowing that are being continuously monitored (Onnela & Rauch, 2016).

Overall, these critiques motivate a cautious, human-in-the-loop use of digital phenotyping signals in clinical decision support.

2.9 Summary and Research Gaps

The literature review highlights that while digital phenotyping holds promise for mental health monitoring, human-computer interaction dynamics (such as mouse movements) remain underexplored in clinical settings compared to smartphone and wearable data. Furthermore, the integration of passive behavioral signals into explainable Clinical Decision Support Systems (CDSS) designed to support human-in-the-loop clinical review and decision-making remains limited. Existing studies have predominantly focused on the predictive performance of passive sensing approaches, with comparatively less attention devoted to the translation of such models into clinically meaningful decision-support tools.

To address these gaps, this dissertation investigates the use of mouse dynamics and self-report measures for depression-related case prioritization, develops an explainability layer to support interpretation of model outputs, and evaluates a clinician-facing CDSS prototype through formative assessment with mental health professionals.

Compared to the reviewed literature, it stands out from works focused solely on the performance of passive sensors by integrating modeling, explainability, and formative evaluation with professionals.

To address these gaps, this dissertation positions itself beyond the current state of the art, as represented by recent comprehensive frameworks such as CLARA and MAILA. CLARA is primarily conceptual: it lays out how HCI signals could be combined with repeated questionnaires within a screening workflow, but it has not yet been empirically validated as a working clinical decision support system (CDSS) (Lee et al. 2019).

Compared with MAILA, and the broader literature review, this work introduces three key different aspects. First, it evaluates a depression-related binary prioritization model combining mouse-dynamics features and DASS-based self-reports, whereas MAILA primarily evaluates a binary classifier from touchscreen trajectories alone (RQ1). Second, it examines longitudinal intra-individual associations in a small clinical sample, while MAILA's longitudinal follow-up analyses were mainly conducted in a large online/general-population setting (RQ1). Third, it translates the predictive model into an explainable clinician-facing CDSS prototype, evaluating model outputs in relation to triage and clinical review utility, and clinician interpretability, rather than only as scalable markers of latent mental states (RQ3/RQ4).

3 Methodology

3.1 Data Sources and Study Context

The dataset for this study was sourced from patients at the psychology department of Trás-os-Montes and Alto Douro Local Health Unit (clinical group) and from a local factory (control group). The study protocol was reviewed and approved by the ethics committee.

The original study protocol was developed under the project "Screening Depression Using Artificial Intelligence Algorithms in Human-Computer Interaction Analysis", whose objective was to explore whether machine learning models trained on human-computer interaction patterns collected during digital psychological questionnaires could support depression-related screening. The protocol framed the collection of mouse and keyboard interaction data during validated questionnaire completion as a digital phenotyping approach for identifying behavioral indicators associated with depressive states.

For the clinical group, recruitment was planned within the Clinical Psychology Service of the Trás-os-Montes and Alto Douro Local Health Unit. Eligible participants were adults, portuguese-speaking, capable of understanding the study procedure and providing informed consent and undergoing psychological assessment in the service. Prior to the digital screening, participants underwent a traditional assessment conducted by a psychologist. During this evaluation, individuals identified as presenting a relevant clinical symptomatological profile related to depression were allocated to the clinical group and invited to complete the structured digital interaction. The study protocol specified that invited participants received written informed consent, which was also explained orally, with emphasis on voluntary participation, the possibility of withdrawal at any time, and the absence of clinical or institutional consequences in case of refusal.

The inclusion criteria comprised being aged 18 years or older, undergoing psychological assessment at ULSTMAD, being able to understand the procedure and sign informed consent, and voluntarily agreeing to participate. Exclusion criteria included severe cognitive impairment preventing comprehension or task execution, lack of portuguese fluency, refusal or discomfort after explanation, and other clinically defined exclusion conditions specified in the protocol. At the protocol level, the available demographic characterization was therefore limited to adult age, portuguese-language fluency, and clinical-service recruitment status; further demographic characterization was constrained by the variables available in the received dataset.

The data collection procedure was designed to take place on a dedicated laptop in the Clinical Psychology Service. Participants were welcomed, seated in front of the computer, presented with the informed consent form, and then asked to complete the digital instruments autonomously. Additional details on the original study protocol, including recruitment procedures, informed consent, inclusion and exclusion criteria, data collection procedures, privacy safeguards, and ethical considerations, are provided in Appendix B.

During each session interaction, participants completed a structured digital screening interaction with two distinct psychological questionnaires: DASS-21 (measuring depression, anxiety, and stress during the past week) (Moshe et al. 2021) and RESS (evaluating emotion

regulation) (De France and Hollenstein 2017). Simultaneously, behavioral interaction signals were recorded unobtrusively in the background from a specialized JavaScript library.

The control group followed the same general digital assessment and passive interaction data collection procedure, with adaptations to the recruitment setting.

The initial set of behavioral data included 285 JSON files with raw interaction records, of which 211 belonged to the clinical group and 74 to the control group. After the inventory process, selection of canonical files, and linking between passive records and questionnaire scores, 276 DASS-21 sessions with valid matches were identified, associated with 103 clinical participants and 68 control group participants. Whereas some participants from the clinical group completed multiple sessions over time, the control group sessions occurred only once.

Figure 1 summarizes the dissertation development pipeline.

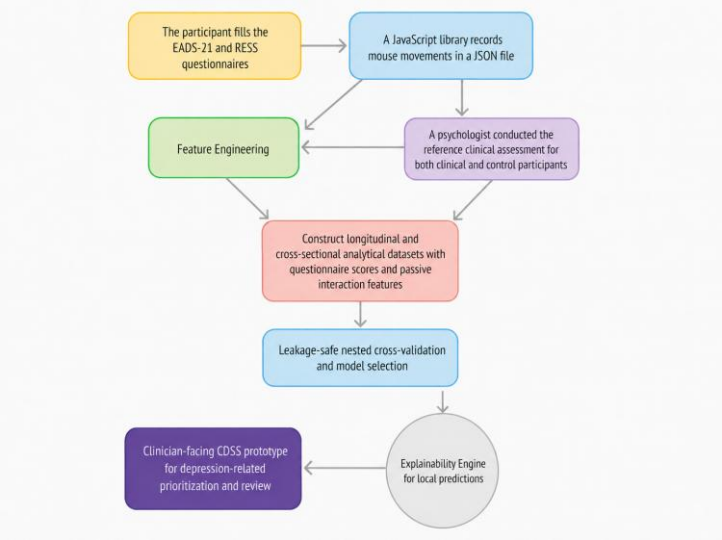

Figure 1 Dissertation development pipeline (Adapted from (Netto et al. 2025))

3.2 Data Preparation and Analytical Dataset Construction

Data Inventory, Canonical File Selection, and Quality Checks

The first main practical tasks of this project included organizing the available data, describing their structure and content, exploring relevant patterns, assessing data quality, and preparing the available data, closely following the first stages of the CRISP-DM methodology (Schröer et al. 2021).

As the project data were distributed across apparently repeated multiple folders and file types, including anonymization files, behavioral interaction logs, and clinical scoring files, it was necessary to understand the structure, completeness, and redundancy of the available raw data to ultimately establish a consistent basis for subsequent dataset construction.

The data inventory procedure consisted of checking logging consistency across JSON representations from different folders, quantifying the number of questionnaire-level records available in the repository, selecting a canonical source for each session-questionnaire, and using anonymization table files to map questionnaire spreadsheets to the canonical behavioral

session chosen, among other steps. The latter was especially important because the subsequent analyses required combining active self-report information, such as DASS-21 scores, with passive interaction features extracted from mouse and keyboard dynamics.

The inventory process, therefore, acted as a quality-control step, ensuring that only sessions with interpretable and linkable data were carried forward into the feature engineering and modeling stages.

Additionally, linked sessions DASS-21 distributions and missingness were analyzed per subscale and cohort through boxplots and histograms. Repeated sessions from the clinical cohort were also inspected to assess whether a longitudinal modeling design would be feasible.

This stage established a cleaned data foundation from which the longitudinal and cross-sectional analytical datasets were later derived.

Passive Signal Source Selection and Feature Engineering

Initially, the chosen source for the passive signal collection was constrained to the DASS-21 due to its intrinsic connection to the depression-related disorders.

During the iterative development of the modeling pipeline, an additional methodological concern was identified regarding the comparability of passive behavioral signals extracted during DASS-21 completion since the clinical and control groups completed the screening in different environments, with different screen resolutions. This could be relevant for distance and time-related mouse features, which may be influenced by the size of the questionnaire.

This issue was investigated by comparing interface probes obtained through browser developer tools with the empirical distribution of click coordinates and trajectory spans across groups. The actual trajectories showed greater vertical extent in the control and greater distance travelled, consistent with a vertically longer task. Thus, the passive features of DASS-21 could reflect interface differences, especially in metrics such as distance, speed, acceleration, and other trajectory measures, rather than participant-level behavioral interaction patterns.

To mitigate this risk, the passive signal used in the cross-sectional predictive pipeline (RQ2) was shifted from DASS-21 to the RESS questionnaire. This decision was based on additional layout and coordinate checks, suggesting that the RESS interaction environment was substantially more comparable across the clinical and control groups. In contrast to the DASS-21 module, the RESS interface showed substantially higher comparability across groups. The final button was consistently identified as the last click across sessions, the vertical position of questionnaire items was stable across the clinical and control environments, and the global vertical trajectory span was nearly identical between groups. Interface probes further confirmed that the representative clinical and control RESS layouts shared the same vertical structure, with only a small horizontal offset attributable to viewport width differences. Since this offset did not affect features based on relative movement between consecutive points, geometric normalization was not applied. The DASS-21 scores were retained as active self-report predictors, while the passive behavioral features were extracted from RESS interactions.

This mitigation should not be interpreted as fully eliminating measurement bias, since residual hardware and context-related differences may still affect interaction dynamics.

The initial extraction of the feature set was guided by a broad review of the literature at the intersection of Human-Computer Interaction (HCI) and affective computing. The goal was to capture a broad spectrum of mouse dynamics frequently associated with the detection of stress, anxiety, and generic emotional states in normative or non-clinical populations (e.g., Banholzer et al., 2021; Sun et al., 2014; Yamauchi, 2013; Aldrich et al., 2023; Khan et al., 2024; Monaro et al., 2018; Zantvoort et al., 2025). Although these metrics have not been exclusively validated for the diagnosis of depression or using standardized clinical scales in their original studies, their inclusion ensured the representation of the main kinematic, spatial, and temporal dimensions documented in the digital phenotyping literature. The considered feature set is presented in Appendix C.

3.3 Clinical Requirements Elicitation

To ground the proposed system in clinical needs, a requirements elicitation process was conducted with mental health professionals in parallel with model development, consistent with the relevance cycle of the DSR framework and user-centered CDSS design principles. The semi-structured interview aimed essentially to validate clinical needs, integration points in the workflow, preferences regarding the type of output, interpretability requirements, and possible adoption limitations.

The interview protocol was mostly based on the conceptual framework proposed by Hopkin et al intended to provide a common language and consistent definitions to describe the characteristics of DMHTs and to facilitate communication between healthcare professionals and others across multidisciplinary boundaries. Additional guidance was drawn from prior work on user-centered CDSS design, implementation considerations, and explainable clinical decision support systems (Bayor et al., 2025; Shulha et al., 2024; Trinkley et al., 2020). The interview script is structured around six domains: current clinical workflow and triage practices; system purpose and intended functionality; human responsiveness and integration; trust and explainability requirements; risk trade-off preferences; and adoption barriers.

To maximize participation given scheduling constraints, the protocol was administered in two modalities: synchronously via online interview (n=2) and asynchronously via a structured Google Forms questionnaire (n=4), of which two responses were consolidated in time to inform the initial interface design. Participants included clinical psychologists working in hospital-based and private clinical settings. This interview protocol and forms link are provided in the Appendix D.

3.4 Artifact A: Predictive Modeling Layer

RQ1: Longitudinal Mixed-Effects Modeling

RQ1 was addressed through an exploratory longitudinal analysis designed to assess if a deviation in the mouse dynamics behavior for an individual could be associated with depression symptom severity (DASS-21) in the next session. This analysis was intended to investigate the longitudinal relevance of passive behavioral signals, and the result could be relevant to inform the predictive modeling layer that constitutes Artifact A.

The dataset for this research question was constrained to participants from the clinical group who underwent at least two sessions on separate days. Each observation corresponded to a

temporally ordered pair of sessions, where predictors were extracted from the current session (t), and the dependent variable corresponded to DASS-21 depression severity at the following session (t+1). The final RQ1 analytical dataset included 37 clinical participants and 91 consecutive session pairs. This structure allowed the analysis to preserve temporal ordering while assessing whether current-session deviations in mouse dynamics were associated with subsequent depression symptom severity.

A Linear Mixed Model was the selected model for the task for the following reasons: (i) natively can handle unbalanced data as different individuals completed a different number of sessions and provide the flexibility to incorporate time-varying covariates, such as the session-specific mouse dynamics (Chi et al. 2019); (ii) repeated sessions would break the dependence assumption for standard regression models, which LMMs account for through random effects (Snijders and Bosker 1999); (iii) the random-intercept specification allows stable between-participant heterogeneity to be modeled while estimating associations between session-level predictors and depression severity (Snijders and Bosker 1999).

A random-intercept specification was adopted to account for stable between-participant differences in depression severity. This allowed the model to represent the fact that participants may differ substantially in their baseline symptom levels, while still estimating the average association between session-level deviations in mouse dynamics and depression severity at the following session. Random slopes were not included, given the limited number of repeated observations per participant and the exploratory nature of the analysis (Bell et al. 2019).

The longitudinal dataset included 37 participants, which was considered acceptable for the exploratory estimation of fixed effects in a Linear Mixed Model. Simulation studies suggest that fixed-effect estimates and their standard errors can remain accurate with approximately 30 to 50 Level-2 units. However, while the point estimates of random-intercept variances may remain unbiased in such samples, their standard errors are frequently underestimated when the number of groups is below 100, leading to unreliable confidence intervals (Maas and Hox 2005). Therefore, this sample size was considered more appropriate for exploratory inference on fixed-effect associations than for drawing strong conclusions about between-participant variance components.

This research question requires an intra-individual interpretation, which means it is intended to analyze how each individual's own variation from the mouse dynamics features is associated with subsequent symptom severity. Thus, the main predictors, including current depression severity and mouse-dynamics features, were centered around each participant’s own mean before model estimation. This allowed the analysis to focus on within-person fluctuations rather than on stable between-person differences in mouse behavior or current depression severity (Snijders and Bosker 1999).

The main fixed-effect predictors included current-session DASS-21 depression severity, time since baseline, and a compact set of pre-specified mouse-dynamics features. For the main model (LMM), whose objective is hypothesis testing on intrapersonal changes over time (focusing purely on Level 1 variables), rather than the accurate prediction of future responses, the full and pre-specified theoretical model should be maintained (Harrell, 2015). Nevertheless, the multilevel modelling literature indicates that more complex designs require more observations, so the specification was deliberately kept conservative (Brysbaert 2018).

The complete feature space considered during the broader feature engineering phase is reported in Appendix C, while the longitudinal model retained only a reduced subset of mouse-dynamics features.

Thus, the final subset was selected to preserve parsimony while retaining mouse-dynamics dimensions that were available in the engineered feature set and conceptually aligned with prior empirical work on depression-related mouse or touchscreen dynamics, including temporal, speed-related, acceleration/kinematic, and trajectory-irregularity measures examined in or motivated by Monaro et al. (2018) and Zantvoort et al. (2025).

Features with highly skewed distributions were truncated at the upper tail to reduce the leverage of extreme values while preserving all participants in the analysis, and were standardized through the z-score scaling technique to improve coefficient comparability.

Accordingly, two model specifications were considered and pre-specified. The main specification included current depression severity, days since baseline, and four mouse-dynamics predictors: jitter, average cursor speed, acceleration mean, and mean time between mouse movement events. A more parsimonious specification retained current depression severity, days since baseline, jitter, and acceleration mean.

Model estimates were evaluated using fixed-effect coefficients, confidence intervals, and Wald-type z-tests, as reported by the mixed-effects model summaries. These tests assessed whether each fixed-effect coefficient differed from zero, corresponding to the null hypothesis of no conditional association between each predictor and depression severity at the following session, while holding the remaining predictors constant. The pre-specified p-value threshold was 0.05. A null random-intercept model was first estimated to quantify the baseline proportion of variance attributable to stable between-participant differences using the intraclass correlation coefficient (ICC) (Snijders & Bosker, 1999). Marginal and conditional R^2 values were computed to summarize the variance explained by the fixed effects alone and by the full mixed-effects model, respectively (Nakagawa and Schielzeth 2013). Model diagnostics included visual inspection of residual patterns and distributional assumptions, with attention to linearity, homoscedasticity, normality of level-1 residuals, and normality of participant-level random intercepts (Harrell , 2015; Snijders and Bosker 1999). Model fit was additionally assessed by comparing the null, main, and parsimonious specifications using AIC and BIC (Bell et al., 2019). For these comparisons, models were refitted using maximum likelihood to ensure comparability across different fixed-effect structures. All analyses were conducted in Python, using appropriate libraries.

RQ2: Cross-Sectional / Hybrid Machine Learning Pipeline

Unlike RQ1, the objective of RQ2 was to assess the predictive ability of machine learning models in a single moment in time, using mouse-based features as a complement to self-reports to distinguish symptom profiles from control-group profiles. This cross-sectional binary classification approach aims to support a depression-related prioritization task, constituting the predictive layer (artifact A) of the clinical decision support system.

This modeling strategy was motivated by the available data structure, the limited density of the longitudinal data observed during RQ1, and the research gap identified in the literature regarding clinically oriented hybrid decision-support pipelines. This approach reduces

reliance on the smaller repeated-measures clinical subset used in RQ1 and focuses on the intended initial prioritization use case.

The number of total questionnaire sessions available in the RQ2 cleaned dataset was 267, including 199 from the clinical and 68 from the control group. The session-level dataset was constrained to account only for the first sessions from the clinical group. The motivation to exclude these observations relies on the attempt to mitigate the familiarity bias that the clinical group might have, introduced by repeated exposure to the paper version of the questionnaires, before the first digital assessment. This resulted in a final RQ2 dataset of 152 participants/sessions: 84 clinical and 68 control.

The classification target was defined as group membership, distinguishing participants from the clinical group from those in the control group. This target was treated as an operational proxy for depression-related prioritization rather than as a direct diagnostic label. Accordingly, model outputs were interpreted as prioritization scores intended to support clinical review, not as definitive diagnostic predictions. This interpretation was aligned with the decision-support orientation of the dissertation, in which predictive outputs were intended to inform, rather than replace, clinical judgment.

The RQ2 modeling strategy compared multiple pipelines defined by both input and model type. Three feature configurations were constructed:

1. DASS-21-only baseline, using only active self-report symptom scores, serving as a reference baseline;
2. HCI_RESS-only pipeline, using passive mouse-interaction features extracted during the RESS questionnaire without any self-report information;
3. hybrid DASS-21 + HCI_RESS pipeline, combining active clinical self-report information with passive behavioral features.

These feature configurations were evaluated using both Random Forest and XGBoost classifiers, as both are suitable for tabular datasets and can model non-linear relationships. (Fadul et al. 2024; Lamba et al. 2026; Zantvoort et al. 2025).

For each algorithm, an independent 5x5 Nested Cross-Validation was applied to reduce optimistic bias, preventing overfitting in small sample sizes and avoiding data leakage during preprocessing (Lamba et al., 2026; Ostojic et al., 2024; Zhan et al., 2026). The inner loop was solely focused on optimizing the specific hyperparameters of that model, feature selection, and threshold selection, while the 5 outer loops provided an unbiased assessment of its generalization ability.

All preprocessing steps were embedded within the inner loop, transformations were fitted only on the corresponding training folds, and applied to the validation fold to reduce the risk of data leakage (Ostojic et al. 2024), including a Correlation-Based Feature Selection (CFS) technique, applied only to the passive feature subset, preserving the DASS-21 variables as clinically meaningful self-report predictors.

In each of the five outer folds, the training split was further divided into five inner folds, on which the hyperparameters were optimized using grid search (Appendix E), based on the AUROC criteria, which provides a threshold-independent measure of discriminative/ranking performance according to predicted risk or priority. Brier score was used to summarize the quality of probabilistic predictions. Threshold-dependent metrics included accuracy, balanced

accuracy, recall, specificity, precision, and F2-score. Balanced accuracy was included to account for performance across both classes, while the F2-score was considered because it gives greater weight to recall, reflecting the triage-oriented preference for reducing false negatives. After the best inner-loop configuration was selected, the model was refitted on the full outer-training split and evaluated on the untouched outer-test fold. This was conducted across all the outer folds, after which the final average metrics are reported as mean ± standard deviation across the five outer evaluations.

Following performance estimation via nested cross-validation, a final refitted model for Artifact B and C was trained using the full RQ2 dataset. Model selection was based exclusively on inner-loop performance, to avoid using outer-loop estimates for algorithm selection, which would constitute data leakage (Wilimitis and Walsh 2023). The final model was retrained on the full RQ2 dataset using consensus features, those selected consistently across all outer folds, and modal hyperparameters, following the stability-based refitting strategy proposed by Parvandeh et al. (2020).

All machine learning analyses were conducted in Python.

3.5 Artifact B: Explainability Layer

In line with the project goals and requirements identified in the literature review, this artifact addresses the need to create an explanation layer to make the results interpretable in a clinical context, exploring how to integrate feature-attribution explanations from models based on cursor-derived features into an interface aimed at clinicians. It consists of an explanation layer built on top of the final hybrid Random Forest, designed to translate model outputs into faithful, compact, case-level explanations suitable for clinician-facing review.

The chosen final model requires post-hoc explanations as it can be considered a black-box model (Abbas et al. 2025; Roychowdhury et al. 2025). The selection of XAI techniques was guided by the Task–Modality–Explanation Alignment (TMEA) framework, which states that the utility of an explanation is maximized when it aligns with the clinical task, the data modality, and context constraints (Abbas et al., 2025). Given that the data used are structured and tabular (EHR and passive behavioral dynamics), the TMEA framework primarily recommends the use of model-agnostic feature attribution methods, as they can ensure greater trust among users (Abbas et al., 2025). In this context, SHAP and LIME are the most widely adopted techniques (Abbas et al, 2025; Mienye et al., 2024; Roychowdhury et al., 2025). SHAP was selected a priori as the primary explanation method because it is well-suited to tree-based models and tabular clinical data, and it supports both local and global interpretations through the aggregation of local attributions (Roychowdhury et al., 2025; Abbas et al., 2025). Furthermore, for tree-based models like Random Forests, TreeSHAP provides a fast and exact computation of SHAP values, making it a practical choice for this study (S. M. Lundberg et al. 2019).

Additionally, given the risk-stratification nature of the clinical task at hand, the TMEA framework recommends a model-agnostic attribution (Abbas et al., 2025). SHAP values represent the average marginal contribution of a feature across all possible combinations of features, explaining exactly how you go from the model's average prediction (the base value) to the local prediction for that specific patient. Each Shapley value corresponds to a single feature, and it can take a negative or positive value depending on whether it is rising or

diminishing the score risk given by the model of being associated with the clinical group class (S. Lundberg and Lee 2017).

Several charts were generated to visualize and analyse both the global and local feature importance.

Mean absolute TreeSHAP values and a beeswarm plot were used to summarize the overall influence and direction of features across the full refitted dataset.

In addition, eight representative cases were selected based on quantile criteria over the predicted probability distribution, covering multiple and varied profiles (from low to high-risk cases). For each case, a waterfall plot showed how the model moved from its average expected probability to the final predicted probability for that participant, displaying the five largest local SHAP contributions and grouping the remaining variables as "other features". The explanations were limited to the top five features to preserve interpretability and avoid overloading the clinician-facing interface, aligning with evidence that a moderate explanation granularity (few factors) maximizes clinical performance (Abbas et al., 2025). These waterfall explanations directly informed Artifact C, where the same information was translated into simplified case-level mockups with a priority score, top drivers, contribution and features values, and a short narrative for clinical review to complement the visual explanation. The latter followed a deterministic template-based approach, using predefined sentence blocks conditioned on the model's threshold status and on the sign of the top local SHAP contributions (see below). Also, mouse-derived features were translated into a comprehensible and non-causal format.

"O modelo sinalizou este caso como {acima/abaixo do limiar operacional}, com score de prioridade de {probabilidade} face ao limiar operacional de 35%. {frase sobre fatores que aumentam prioridade} {frase sobre fatores que reduzem prioridade} {cautela obrigatória}"

SHAP method was subjected to internal validation through objective metrics such as sparsity, completeness, and fidelity (Mienye et al., 2024), and to subjective evaluation during the usability tests with clinicians.

Fidelity measured the proportion of instances for which the classification obtained from the Top-5 SHAP approximation matched the classification of the original model. Completeness was computed as the proportion of variance in the original model predictions captured by the Top-5 SHAP approximation. Sparsity was examined as a descriptive check on the a priori choice of presenting five detailed features, verifying the average number of non-zero SHAP feature attributions used by the final top-5 explanation across the dataset.

The perceived interpretability and usefulness of the explanations were evaluated subjectively during the clinician-facing prototype evaluation, whose protocol is described in more detail in the Artifact C methodology. These usability tests assessed both the dashboard interface and the explanation layer embedded within it.

The explanation evaluation included open-ended feedback and two adapted Likert scale items based on Bergomi et al., focusing on understandability and actionability dimensions. Participants reviewed representative case mockups and were asked to comment on the clarity, usefulness, and clinical relevance of the displayed priority score, top contributing factors, and narrative explanation.

3.6 Artifact C: Clinical Dashboard Prototype

Built upon the artifacts A and B, initial clinical feedback, and literature guidelines, artifact C consists of a static clinician-facing mockup, whose goal is to support initial screening/prioritization and clinical review after the first digital assessment. This artifact addresses RQ4 by exploring how to design a clinical decision support system dashboard and combine questionnaire information, predictive model outputs, and explanation components to support clinicians in reviewing and prioritizing depression-related screening cases. The static dashboard mockup was developed using Figma.

Mockups' prototyping was informed essentially by two complementary sources, consistent with user-centered design principles (Abras et al. 2004): requirements gathering in the initial interview cycle with clinicians; and ii) literature recommendations on CDSS/XAI design that focus on workflow integration, AI explanations clarity, and cognitive overload reduction (Abbas et al., 2025; Mienye et al., 2024; Shulha et al., 2024; Tonekaboni et al., 2019; Trinkley et al., 2020).

From these sources, five design principles were operationalized: (1) present clinical context from the individual as a complement to the AI system prediction or explanation, and distinguishing them clearly to prevent clinicians from assuming the model used all visible data (Shulha et al. 2024); (2) display the model's estimated probability along risk category to help clinicians calibrate their own trust in the system (Roychowdhury et al. 2025; Tonekaboni et al. 2019); (3) restrict the information presented to the essential (e.g., only the Top-5 variables) to keep the screen clean, avoiding overwhelming the clinician and maximizing human speed and accuracy under time pressure (Abbas et al. 2025); (4) recur to autonomy-preserving, non-causal language throughout the interface, and use the AI system as a complement to clinical tasks rather than fully automating them (Roychowdhury et al. 2025); (5) interface actions should be directly actionable, with minimal navigation overhead or complexity (Sunjaya et al., 2022; Trinkley et al., 2020).

Additionally, clinical requirements were synthesized from the four interview responses collected during the elicitation process (Section 3.3) through a domain-by-domain thematic analysis organized around the interview protocol structure. This synthesis is reported in Appendix F. Subsequently, a separate requirements matrix was constructed by combining the interview synthesis with CDSS/XAI design recommendations from the literature and the scope constraints of the current prototype. This matrix was used to guide the design of Artifact C and to document which requirements were implemented, partially represented, or reserved for future iterations. The complete matrix can be found in Appendix F, while Table 3 below represents a compact one.

Table 3 Design requirements and corresponding mockup implementation status

Desing Requirement	*Mockup implementation*	*Status*
Adaptability to different clinical contexts	Generic clinical review/triage support language.	Partial
Consider longitudinal evolution	Not implemented; scope limited to the first digital assessment.	Future
Present prioritization/signaling rather than diagnosis	Priority score and "Above threshold" / "Below threshold" label.	Implemented
Priority-ordered case queue	Initial queue ordered by priority score.	Implemented
Support human review and potential	Actions such as "View details", "View	Partial

follow-up	responses", and "Mark for review".	
Provide concise visual explanations	Top-5 contributing factors and aggregation of "other factors".	Implemented
Provide complementary explanatory text	Brief narrative interpreting the signaling and final caution.	Implemented
Separate self-report, HCI, and additional context	Separate blocks for self-report, interaction, explanation, and additional context.	Implemented
Avoid overinterpretation of HCI signals	Warnings supporting clinical review without automatic diagnosis or causal inference.	Implemented
Prioritize sensitivity in signaling	Recall/F2-oriented operational threshold and human review.	Implemented
Present summary first, details later	Overview with metrics/queue; detail view with explanation and context.	Implemented
Address oversight, privacy, and operational workload	Human oversight made explicit; privacy/workload discussed in the evaluation.	Partial
Represent inconsistency between self-report and interaction patterns	Contextual indicator: "Possible discrepancy to explore".	Contextual/Future

To assess the proposed interface, a semi-structured evaluation protocol was prepared in line with recent clinical XAI evaluation practices (Roychowdhury et al., 2025), combining a think-aloud usability task, open-ended questions, two adapted XAI evaluation items, and the System Usability Scale (SUS). An application-grounded evaluation approach was adopted, aiming to align the evaluation with the actual context of tool use (Shulha et al. 2024). The evaluation was conducted with two mental health experts with a background in psychology, who interacted with simulated case mockups populated with real participant data, including DASS-21 depression scores, RESS information, and model-derived prioritization outputs.

Each session began with a short preamble explaining the intended context of use, the static nature of the mockup, the triage/prioritization purpose of the system, the meaning of the priority score and operational threshold, and the main limitations of the available data. Participants were asked to think aloud while interacting with the interface to facilitate a discussion about the various features of the tool.

After this introduction, participants first reviewed the prioritization dashboard and were asked to describe what they interpreted from the page, including which information stood out, whether the ranking was clear, whether the meaning of the priority score and threshold label was understandable, and other details. They were then shown simulated case mockups populated with real participant data. Three cases with different case roles were selected: a high-priority convergent case, where the model's high score was matched by strong self-reports; a high-priority case mainly influenced by passive interaction patterns, allowing the evaluation of the interpretation of signals less directly aligned with self-reports; and a case near the operational threshold (borderline), useful for exploring ambiguity and follow-up decisions. This selection made it possible to see if the interface was understandable in both situations with clear signals and in more ambiguous or potentially conflicting ones.

During case review, participants were encouraged to comment on the clarity of the relationship between self-reports, model score, and explanatory factors; the added value of the textual narrative relative to the visual SHAP output; and whether the separation between model inputs and additional contextual information was clearly communicated, among other open-ended questions resulting from the ongoing discussion.

At the end of the session, participants completed a short questionnaire. The questionnaire first collected basic participant profile information, including session identifier, professional profile, years of professional experience, prior experience with AI systems, and perceived potential usefulness of AI in clinical or professional work. It then included two adapted XAI items, followed by the System Usability Scale, and an open-ended question to collect suggestions on how the interface could be enhanced or made more useful in a real clinical context.

To specifically assess the explainability engine and the quality of information generated by Artifact B, two items adapted from clinical XAI evaluations were used, focusing on two relevant dimensions: the understandability of the explanation and its usefulness in supporting clinical review (actionability) (Bergomi et al., 2025). These items were used to evaluate whether the visual and narrative explanation helped explain the model's signaling and whether it could support screening or follow-up decisions. At the same time, the overall usability of the dashboard and the integration of its elements into the workflow, corresponding to Artifact C, were assessed using the System Usability Scale (SUS), a widely used scale for the subjective evaluation of interactive system usability (Brooke 1996; Holzinger et al. 2019).

Qualitative feedback from the formative evaluation was analyzed through a structured thematic synthesis using a hybrid deductive-inductive approach. The interview and think-aloud material were first organized and deductively coded according to the evaluation protocol sections: the initial dashboard view, case-level reviews, which included the explanation components, and final open-ended feedback. Each coded excerpt was then mapped to the relevant research question and artifact: Artifact B/RQ3 for comments related to model explanations, SHAP factors, narrative explanations, passive-feature interpretation, and understandability and actionability regarding the explanation section; and Artifact C/RQ4 for comments related to dashboard organization, score interpretation, workflow fit, clinical usefulness, and usability. Additional recurrent observations were incorporated inductively when they emerged from the participants' comments. The analysis was conducted through a supervised coding matrix and was descriptive and illustrative, given the formative purpose and small number of participants. The analysis are reported in the Appendix G.

Given the formative and exploratory nature of the evaluation and the small sample size, scale results were interpreted descriptively rather than inferentially. Aligning with recent XAI evaluation guidelines, qualitative data from think-aloud protocols and open-ended responses provided the primary source of interpretive evidence to gain deeper insights into user experiences (Roychowdhury et al., 2025). The sessions protocol and forms link are provided in Appendix H.

4 Results & Interpretation

4.1 Dataset Characterization and Preparation

The process of inventory and data preparation (Section 3.3) resulted in 276 sessions with matching between passive records and questionnaire scores, of which 208 belonged to the clinical group and 68 to the control group, associated with 103 participants in the clinical group and 68 in the control group. No missing responses were identified in the questionnaires within this subset.

The figure below illustrates boxplots that represent the total sum of the response scores of each of the three subscales and allows a comparison between the two study groups. As expected, the clinical group showed generally higher values in the depression, anxiety, and stress subscales. However, there was considerable overlap between the groups, especially in the depression subscale, which is consistent with the mixed and episodic nature of depressive symptoms (Fried and Nesse 2015).

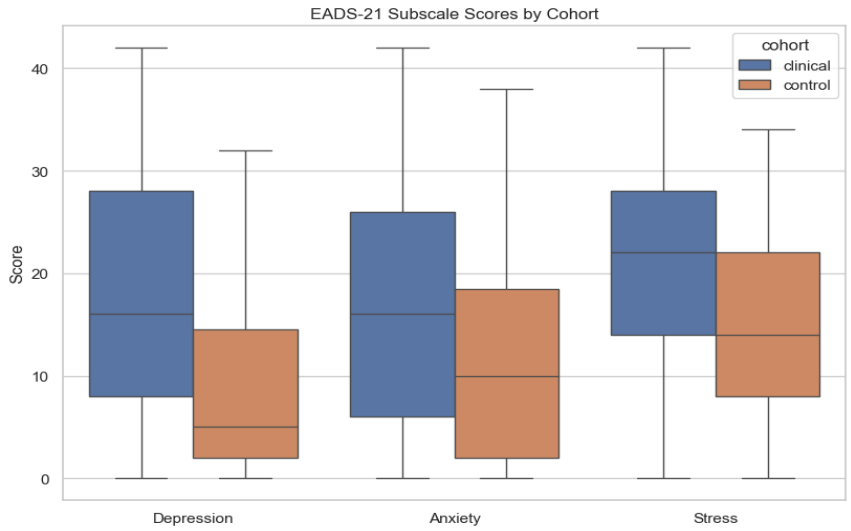

Figure 2 DASS-21 subscale score distributions by study group

From this base dataset, two analytical datasets were constructed, whose characteristics are summarized in Appendix I.

4.2 RQ1: Longitudinal Associations Between Mouse Dynamics and Depression Severity

RQ1 aimed to evaluate, in a longitudinal and exploratory manner, whether intra-individual variations in mouse dynamics during a session were associated with the severity of depressive symptoms in the following session. This analysis used the longitudinal dataset described in Section 4.1, where each observation corresponded to a pair of consecutive sessions, in which the behavioral and clinical features of the current session (t) were used to explain the total depression score on the DASS-21 scale in the following session (t+1). To this end, mixed linear models with random intercepts per participant were estimated. Table 4 compares the null, main, and parsimonious models.

Table 4 Comparison of longitudinal mixed-effects model specifications for RQ1

	Fixed effects	*AIC*	*BIC*	*Marginal R^2*	*Conditional R^2*	*ICC*	*Converged*
Null model	Random intercept only	664.14	671.67	0.000	0.755	0.755	Yes
Main model	Current depression + time + 4 mouse features	671.82	694.42	0.011	0.777	0.774	Yes
Parsimonious model	Current depression + time + 2 mouse features	669.46	687.04	0.007	0.769	0.767	Yes

Regarding the null model, about 76% of the variation in subsequent depression scores could be attributed to stable differences between participants (ICC ≈ 0.76), suggesting that individual baseline explains substantially more variance than session-to-session fluctuations, which limits the explanatory power available to behavioral signals. Models with predictors have higher AIC/BIC than the null model, indicating that the features do not improve the fit enough to offset the added complexity. Additionally, the marginal R^2 values are very low (0.011 and 0.007), confirming that fixed effects explain less than 2% of the variance, while the high conditional R^2 (≈0.77) is completely dominated by the random intercept. These values show that these predictors added little explanatory power compared to the model with compared with the null model, which included only a participant-level random intercept.

The Wald test on fixed effects asks the following question, allowing for an answer to the research question under analysis:

"For each fixed effect coefficient, does the Wald test assess whether intra-individual variation in a specific feature of the cursor's dynamics in the current session is associated with variation in the severity of depressive symptoms in the next session, while keeping the other predictors in the model constant?"

Among the models including fixed effects, the parsimonious model was retained for interpretation because it provided the most appropriate balance between fit and interpretability (AIC/BIC results), given the limited number of repeated observations. Fixed-effect estimates for this model are reported in Table 5.

Table 5 Fixed-effect estimates for the parsimonious longitudinal mixed-effects model

Predictor	*Estimate*	*SE*	*95% CI*	*z*	*p-value*
Intercept	18.034	1.922	[14.267, 21.801]	9.383	<.001
Depression severity (t)	−0.981	0.689	[−2.331, 0.370]	−1.423	.155
Days since baseline	−0.177	0.750	[−1.648, 1.293]	−0.237	.813
Jitter	−0.084	0.688	[−1.433, 1.264]	−0.123	.902
Acceleration mean	−0.268	0.636	[−1.515, 0.978]	−0.422	.673

The parsimonious linear mixed model yielded no statistically significant fixed-effect associations between within-person mouse-dynamics features and subsequent depression severity. All predictor coefficients had wide confidence intervals crossing zero and p-values

above the .05 threshold, indicating no reliable within-person associations between current-session mouse dynamics and depression severity at the following session.

Among the fixed-effect predictors, current depression severity showed the strongest tendency ($\beta = -0.981$, $z = -1.423$, $p = .155$). The negative coefficient indicates that, within the same participant, sessions with higher current depression severity tended to be followed by lower depression severity at the next session, after controlling for time since baseline and mouse-dynamics features. However, this effect was not statistically significant and should therefore be interpreted cautiously.

The mouse-dynamics features included in the parsimonious model, jitter and acceleration mean, showed small and non-significant coefficients. These results provide no evidence that within-person deviations in these session-level kinematic features were associated with depression severity at the following session in this longitudinal subset.

Figure 3 presents the diagnostic plots used to assess the linear mixed model assumptions, including residual dispersion across fitted values, normality of Level-1 residuals, and normality of participant-level random intercepts.

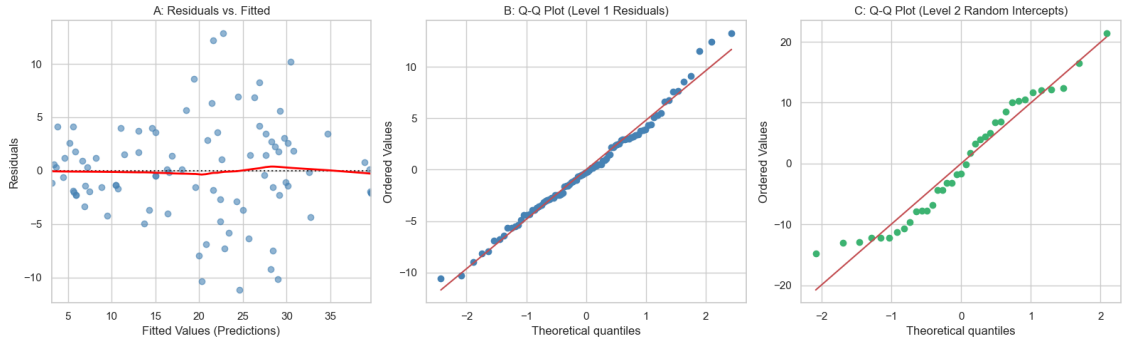

Figure 3 Diagnostic plots for the parsimonious longitudinal mixed-effects model

Diagnostic inspection of the linear mixed model suggested that the main model assumptions were broadly acceptable, although some deviations were observed. The residuals-versus-fitted plot did not indicate a clear violation of linearity, as the smoothed trend remained close to zero across the range of fitted values. However, the dispersion of Level-1 residuals appeared larger around intermediate fitted values, suggesting possible heteroscedasticity in the standard LMM specification.

The Q-Q plot of Level-1 residuals indicated an approximately normal distribution, with only minor deviations in the tails. The Q-Q plot of the participant-level random intercepts showed a slight S-shaped curvature and some deviation in the lower tail, suggesting a modest departure from perfect normality. However, this deviation was not considered to invalidate the fixed-effect interpretation, since simulation evidence suggests that even substantial violations of higher-level random-effect normality have limited impact on fixed-effect estimates and their standard errors in linear mixed models (Bell et al., 2019). Overall, the diagnostics did not indicate severe assumption violations, although the possible heteroscedasticity and limited number of repeated observations support a cautious interpretation of the model estimates.

Overall, the results of RQ1 do not provide strong statistical evidence that intra-individual variations in mouse dynamics during a session are linked to the severity of depressive symptoms in the following session. The high ICC value (0.755) suggests that much of the observed variation was related to baseline individual differences. Thus, the passive signals analyzed didn't show clear longitudinal value in this small clinical sample, though the results should be seen as exploratory rather than definitive evidence of no effect.

4.3 RQ2: Predictive Modeling for Depression-Related Prioritization

RQ2 assessed the cross-sectional discrimination capacity between participants in the clinical group and the control group from machine learning models, using the dataset described in Section 4.1. Three input configurations were compared: a baseline based only on the results of the three subscales of the DASS-21 questionnaire, a pipeline based only on the HCI features extracted during the completion of the RESS questionnaire, and a hybrid pipeline that combined DASS-21 with passive HCI_RESS features. Each configuration was evaluated with Random Forest and XGBoost models.

Table 5 presents the metrics obtained in the outer loop of the nested cross-validation, aggregated as the mean across all their respective values from the 5 outer folds along with the corresponding standard deviations. AUROC was used as the main discrimination metric independent of the threshold, since its logic is associated with the discriminative or ranking/prioritization power of the model, which aligns with the objective of both the research question and artifact A. In other words, this metric allows answering the following question, and, consequently, the research question:

"What is the probability that an individual from the clinical group receives a higher probability than an individual from the control group?"

The Brier score was also used to evaluate the probabilistic quality of the predictions. Additionally, threshold-dependent metrics referring to the outer folds were reported, considering both the threshold optimized by balanced accuracy and the threshold guided by the F2-score, given the clinical interest in prioritization scenarios with higher sensitivity.

Table 6 Outer-loop performance metrics for the RQ2 model configurations

Configuration	*Classifier*	*AUC outer*	*AUC inner*	*Brier score*	*Accuracy, BA threshold*	*Accuracy, F2 threshold*
DASS-21-only baseline	RF	0.772 ± 0.046	0.800 ± 0.024	0.202 ± 0.018	0.658 ± 0.049	0.612 ± 0.079
HCI_RESS-only	RF	0.790 ± 0.038	0.805 ± 0.025	0.184 ± 0.016	0.703 ± 0.055	0.618 ± 0.038
HCI_RESS-only	XGBoost	0.763 ± 0.057	0.796 ± 0.013	0.205 ± 0.024	0.644 ± 0.062	0.599 ± 0.046
DASS-21 + HCI_RESS	RF	0.856 ± 0.055	0.877 ± 0.017	0.156 ± 0.023	0.789 ± 0.067	0.743 ± 0.059
DASS-21 + HCI_RESS	XGBoost	0.863 ± 0.034	0.876 ± 0.013	0.151 ± 0.020	0.730 ± 0.052	0.731 ± 0.072

Through the analysis of the table, it is possible to verify that the relatively small differences between the inner AUC and outer AUC averages may suggest the absence of overfitting. Configurations with passive features as the only inputs achieved performance similar to the metrics reported by the baseline. In the case of the passive model, whose algorithm was the Random Forest ensemble model, the performance was even superior for all metrics, although with modest differences. The hybrid models outperformed the others across all reported metrics.

These results suggest that the complement between the different sources of signals, namely cursor interaction and self-report symptom questionnaires, can help discriminate individuals with clinical depressive symptom profiles from a control group in the present dataset.

Furthermore, the analysis of the reported results also constitutes evidence that models combining information from cursor interaction and symptom information from self-reports show an apparent superiority compared to models that rely solely on the latter type of information, thus addressing the research question under analysis. Nevertheless, these findings should be interpreted as evidence for prioritization support rather than diagnostic classification, given the operational nature of the clinical/control target and the limited sample size. In addition, the counter-intuitive direction of temporal mouse features, where shorter event intervals between mouse movements contributed positively to the clinical group score, suggests that the model may have partially learned group-level differences in digital familiarity or task pacing rather than purely depression-related behavioral patterns. Clinical participants, recruited from a hospital psychology service with prior exposure to the paper versions of the questionnaires, may have completed the digital screening more quickly and with less hesitation than control participants, independently of their current symptom severity.

In line with the methodology presented in 3.4, for the selection of the model to be used in Artifact B/XAI and Artifact C, only the information from the inner loop was considered, where the optimization of the models and thresholds took place.

As the two hybrid models showed very similar internal AUROC values, an additional inspection of the threshold-dependent internal metrics was carried out. This analysis allowed the evaluation of not only the average discrimination but also the operational profile of the models under two distinct thresholds: a balanced threshold, optimized for balanced accuracy, and an F2-score-oriented threshold, more sensitive to detecting positive cases. The latter was adopted, given the preference for minimizing false negatives over false positives expressed by participants in the initial requirement-gathering interviews. Table 7 summarizes this comparison for the hybrid models.

Table 7 Inner-loop threshold comparison for the hybrid model configurations

Model/configuration	*Inner AUROC*	*Threshold strategy*	*Threshold*	*Recall*	*Specificity*
Hybrid RF	0.877±0.017	Balanced accuracy	0.504 ± 0.060	0.819 ± 0.065	0.787 ± 0.067
Hybrid RF	0.877±0.017	F2-oriented	0.336 ± 0.060	0.964 ± 0.013	0.489 ± 0.175
Hybrid XGBoost	0.876±0.013	Balanced accuracy	0.558 ± 0.207	0.804 ± 0.118	0.808 ± 0.116
Hybrid XGBoost	0.876±0.013	F2-oriented	0.199 ± 0.086	0.961 ± 0.023	0.467 ± 0.119

The hybrid Random Forest was selected as the final model because it achieved the highest inner-loop AUROC, although only marginally above the hybrid XGBoost model, and showed a slightly more stable operational profile under the F2-oriented threshold. In particular, both models achieved similarly high recall under the F2-oriented strategy, but the Random Forest retained slightly higher specificity and lower threshold variability. This made it the preferred model for the prioritization-oriented use case and for downstream explanation and dashboard generation.

The final model was then refitted on the full cross-sectional dataset to generate the prediction scores and explanations used in Artifacts B and C. The final refitted model used the consensus features selected across the nested cross-validation procedure and the modal hyperparameters, following a stability-based refitting strategy. The final consensus features were the following: depression, anxiety, stress, total time, click count, mouse move count, average cursor speed, acceleration mean, jitter, mean time between mouse movement events, and maximum time between mouse movement events.

4.4 Explainability of the Selected Model

RQ3 aimed to analyze how the results of the final predictive model could be transformed into explainable components suitable for integration into a clinical interface to assist in the task of signaling clinical priority related to depression. For this stage, the hybrid Random Forest model selected in RQ2 was used, utilizing the entire dataset retrained with the consensus features and the hyperparameters defined in the previous process.

The explanations were generated with TreeSHAP, an approach suitable for tree-based models and tabular data. SHAP contributions were calculated for the positive class, that is, for the probability of the case belonging to the clinical group. Thus, positive SHAP values indicate that a variable contributed to increasing the probability assigned by the model, while negative values indicate a contribution in the opposite direction.

Figures 4 presents distinct representations of the model's overall explanation: a global mean plot and a beeswarm. The first (on the left) ranks the features according to their mean absolute TreeSHAP value, indicating the average magnitude of each feature’s contribution to the model output, while the beeswarm allows simultaneously observing the relative importance of the variables, and both distribution and direction of individual SHAP contributions across participants, with color indicating the relative value of each feature.

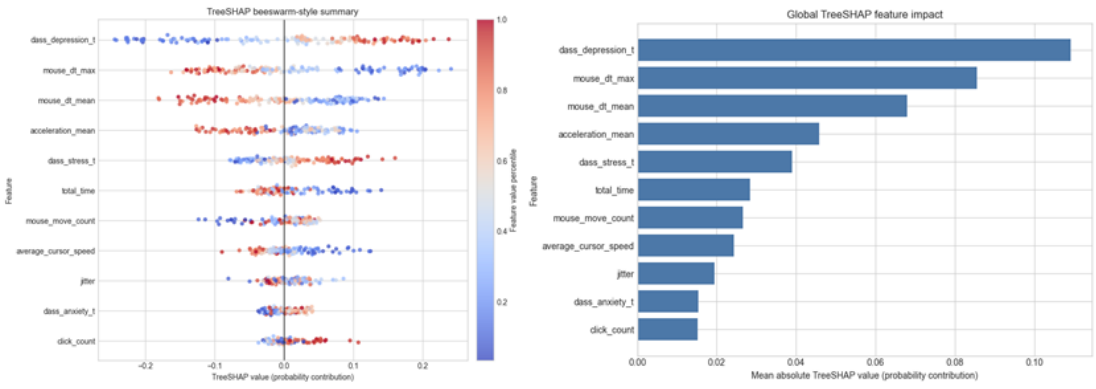

Figure 4 Global TreeSHAP explanations for the selected hybrid model: beeswarm summary and mean absolute feature impact

The global TreeSHAP plots show that the final hybrid model relied on both self-reported symptoms and passive RESS interaction features. The DASS-21 depression score had the highest mean contribution, followed by several mouse-interaction features, particularly maximum and mean time between mouse movement events and acceleration mean. The beeswarm plot suggests that higher depression and stress scores generally increased the model's priority score. For several HCI_RESS features, the direction was less clinically straightforward. Lower values of the mean and maximum time between mouse movement events, lower acceleration, lower average cursor speed, and lower total response time tended to contribute positively to the priority score in the global distribution, while higher values more often contributed negatively. This should be interpreted with caution, because it may reflect interaction style, familiarity, haste, or layout characteristics, not necessarily a direct psychological marker.

Figure 5 illustrates how a local waterfall SHAP explanation was translated into the clinician-facing explanation component of the prototype. The selected case represents a high-priority convergent profile, with a 100% priority score and elevated DASS-21 depression and stress scores. The interface presents the main positive contributors to the model output, grouped into self-report features and RESS interaction patterns, while aggregating the remaining variables as "other factors". A short narrative summary complements the visual explanation and reinforces that the output should be interpreted as support for triage and clinical review, not as an automatic diagnosis or causal inference.

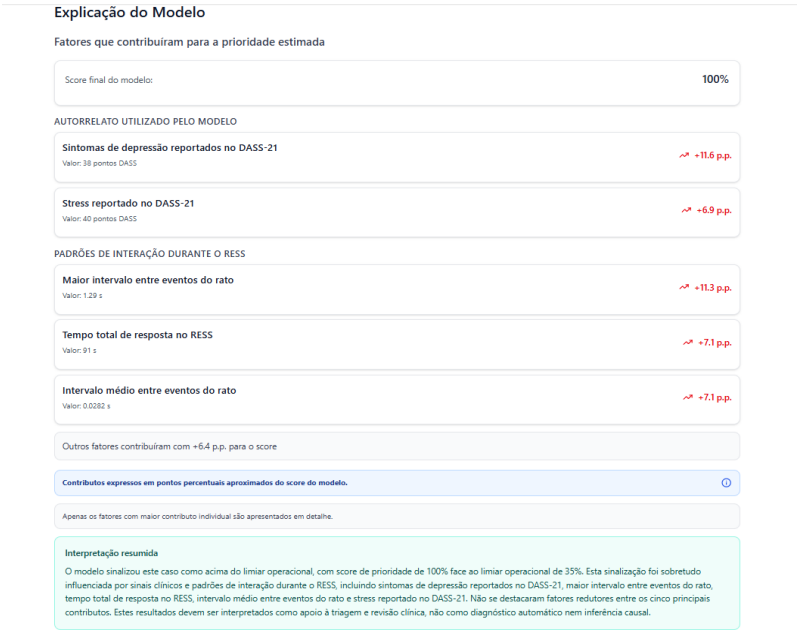

Figure 5 Clinician-facing explanation component for a high-priority convergent case

Table 7 summarizes the metrics used to objectively assess whether the compact version of the explanations, limited to the Top-5 features per case, sufficiently preserves the behavior of the full model and its simplicity.

Table 8 Objective metrics for compact Top-5 explanations

Metric	*Value*
Completeness	0.980
Fidelity	0.993
Sparsity	5.0

Fidelity indicates that, in almost all cases, the Top-5 approximation would lead to the same classification as the original model. Completeness shows that this approximation still captures almost all the variation in the probabilities predicted by the model. Sparsity confirms that the explanation format consistently presented five non-zero feature contributions per case.

The subjective evaluation of Artifact B focused on the perceived understandability and actionability of the explanation components embedded in the case-level mockups. These components included the displayed contributing factors, the short narrative explanation, and the interpretation of passive mouse-interaction features when they contributed to the model output. Since the priority score is both a model output and a dashboard element, comments about the score were considered in this section only when they concerned the interpretation of the model explanation. Broader comments about dashboard organization, additional clinical context, visual layout, and workflow are reported under RQ4.

The protocol included convergent, passive-driven, and borderline cases. As not all variants were shown to both participants, cross-participant findings focus on the convergent and passive-driven cases.

Two psychology professionals participated in the formative evaluation, with 19 and 6 years of professional experience, respectively. Both reported having good knowledge of, or prior experience with, AI systems in their work, and both rated their belief that AI could improve the effectiveness of their clinical or professional work as 4 out of 5. The two adapted explanation items yielded positive descriptive results. Both participants strongly agreed that the explanation helped them understand why the model had signaled the case with that priority score (P1 = 5/5; P2 = 5/5). The perceived usefulness of the explanation for deciding whether a case should be reviewed or clinically followed was also positive, with P1 rating this item as 4/5 and P2 as 5/5.

Table 9 Adapted XAI evaluation item scores for the explanation layer

Item	*P1*	*P2*
Explanation helped understand why the model assigned this priority	5/5	5/5
Explanation would be useful for deciding whether the case should be reviewed or followed clinically	4/5	5/5

The qualitative feedback was broadly consistent with these ratings, although some important concerns were raised, and it helped clarify the conditions under which the explanation was perceived as more or less actionable.

In the convergent high-priority case, the explanation layer was not rejected by either participant and was generally perceived as useful for supporting case interpretation, although this perception followed a brief clarification of the mechanical form of the explanation, suggesting that some initial guidance may be needed for users unfamiliar with this type of tool:

"That means percentage points. [...] What we are seeing is the explanation of the model, therefore, they are the factors that most contributed to increasing the score. And we have several weights." (Francisco / facilitator)

"I like this, [but] I do not know if it is very intuitive for everyone, for all professionals." (P1, translated)

The same interviewee made a similar observation about the narrative explanation.

"For example, 'no factors stood out among the five main contributions'. I know what it means, but it is not very intuitive." (P1, translated)

P2 also accepted the explanation section, although the evidence was less spontaneous and mainly expressed through agreement during the guided review.

"The explanation seems okay." (P2, translated)

P2 similarly agreed that elements related to machine learning may be less intuitive for professionals without technical background.

The interpretation of passive mouse-interaction features emerged as the most sensitive aspect of Artifact B. In the passive-driven/discordant case, both participants were cautious about the clinical meaning of mouse-derived variables. P1 interpreted mouse-related patterns as potentially reflecting technological familiarity rather than depression-related behavior.

"If it is related to the mouse, it only indicates one thing to me: that the person has no experience using a mouse." (P1, translated)

The same interviewee further emphasized that future use of the tool should include information about mouse-use familiarity.

"Whoever uses this tool in the future has to indicate their experience with the mouse." (P1, translated)

This suggests that mouse-interaction features were not clinically self-explanatory and could be confounded by digital or motor familiarity.

P2 recognized the potential value of passive signals in discordant cases, noting that people who do not identify depressive symptoms may still be flagged as requiring clinical evaluation.

"Even people who do not identify depressive symptoms can be identified as having depressive symptoms and [...] need clinical evaluation." (P2, translated)

Although P2 understood that the system used mouse movements as an additional signal and recognized the usefulness of the explanation, P2 had low interpretative confidence in the mouse factors.

P2 expressed a related concern, focused on the actionability of a high score driven by passive features. She stated that she did not feel able to interpret the mouse-related variables directly.

"I do not understand anything about the mouse issues." (P2, translated)

In such a case, she explained that she would return to the self-reported symptom scores.

"I would automatically look at the symptoms [...] anxiety, stress." (P2, translated)

When faced with a case showing a high priority score but no expressive self-reported symptoms, she indicated that this would make her question whether to contact the person for an evaluation.

"So, it has 94%, but no symptoms. Then maybe I would question whether I would actually call the person to schedule the evaluation or not." (P2, translated)

This indicates that passive HCI explanations may generate a useful alert, but may not be sufficient on their own to support confident clinical action.

P2 then clarified that additional contextual information would help interpret such a case.

"If I saw the medication, if I saw the suicidal ideation, the two questions on suicidal ideation, attempts and thoughts, sleep, physical activity, it could help..." (P2, translated)

Although some of these elements belong primarily to the broader dashboard and clinical context layer discussed under RQ4, they are relevant to RQ3 because they condition how clinicians interpret model explanations when passive features dominate the priority score.

Taken together, these findings suggest that compact feature-attribution explanations may support clinical review, especially when the model output is aligned with expressive self-report scores, and that they should be accompanied by simpler wording and clearer interpretive support in what regards the clinical narrative. In passive-driven cases, actionability appeared to take a more cautious form: the explanation could act as a trigger for further clinical verification, such as contacting the person, scheduling a follow-up assessment, or collecting additional contextual information, rather than directly supporting a definitive clinical decision on its own. This interpretation is consistent with the positive questionnaire ratings, since the evaluated action was whether a case should be reviewed or clinically followed, not whether the system could independently determine diagnosis or treatment.

4.5 RQ4: Formative evaluation of the dashboard prototype for clinicians

RQ4 focused on understanding how the dashboard prototype intended for healthcare professionals could support the clinical review and prioritization of cases related to depression screening. The evaluation of Artifact C, therefore, focused on the perceived usability, usefulness, and integration of the dashboard's elements, as well as areas for improvement, with a focus on its clinical utility. The results combine descriptive SUS ratings with comments collected through the think-aloud protocol and open feedback obtained during the formative evaluation. Illustrative images from the mockup can be found in the Appendix J.

Two psychology professionals participated in the evaluation, with 19 and 6 years of professional experience, respectively. The quantitative usability results were positive, although they should be interpreted descriptively due to the small number of participants and the formative nature of the assessment. Both participants scored 82.5/100 on the SUS. They also gave positive ratings to items related to ease of use, feature integration, trust, and ease of learning. In particular, both disagreed that the system was difficult or uncomfortable to use, and both indicated that they would feel confident using it. However, the qualitative feedback clarified various conditions under which the dashboard would need improvements before use in a real context.

Table 10 SUS scores and item-level responses for the dashboard prototype

SUS Result	***P1***	***P2***
SUS Score	**82.5/100**	**82.5/100**
I would like to use this system frequently	5/5	4/5
I found the system unnecessarily complex	3/5	1/5
I thought the system was easy to use	4/5	5/5
I think I would need technical support to be able to use this system	1/5	2/5
I found the various functions in this system were well integrated	4/5	4/5
I thought there was too much inconsistency in this system	1/5	2/5
I imagine that most people would learn to use this system very quickly	4/5	4/5
I found the system very cumbersome to use	1/5	1/5
I felt very confident using the system	4/5	4/5
I needed to learn a lot of things before I could get going with this system	2/5	2/5

The first view evaluated was the prioritization dashboard, including the case queue, summary graphs, priority score, and threshold status. The initial visual impression was positive. P1 described the dashboard as intuitive and highlighted the visual presentation of the information.

"It seems very intuitive [...].The graphs, perhaps the sober colors." (P1, translated)

P1 also considered the ranking easy to understand and reacted positively to the typography, although she suggested that some text could be slightly larger.

"I think it is good. I think it is easy to understand." (P1, translated)

P2's feedback on the initial view was also globally positive, as she did not identify relevant problems with the page, case queue, or graphs during the dashboard review.

"Yes, it seems fine to me." (P2, translated)

Taken together, these comments suggest that the first view presented a positive initial usability profile, especially in terms of visual organization and perceived readability. However, the evidence regarding autonomous prioritization was more limited. In both sessions, the cases were explored mainly through the guided protocol and not through a free choice of which case to open first. Therefore, the results support the clarity of the case queue and navigation at a basic level, but do not robustly demonstrate that the dashboard autonomously supports case selection without facilitation.

The main usability issue identified in the initial view was related to the meaning of the priority score. P1 explicitly asked what the "model priority score" meant and continued to request clarification after the initial explanation. This suggests that, despite the introduction provided, the concept of priority was not fully self-explanatory within the dashboard itself.

"Yes, but what does score mean, but what does priority score mean?" (P1, translated)

After clarification, P1 appeared to understand the intended meaning and proceeded to the detailed case view, but suggested including an explanation directly in the interface.

"Perhaps I would put here information, an information ['i'] or something like that, and the person would click and read it." (P1, translated)

This result does not necessarily indicate that the interface did not present any explanation of the score, since the first view already included an explanatory note. However, P1's doubt suggests that this explanation may not have been sufficiently salient or directly associated with the "priority score" column. Thus, a possible improvement would be to bring the explanation closer to the score label itself, for example through an information icon, tooltip, or reformulation of the column heading.

The transition between the dashboard and the detailed case view appeared to be understandable. Both P1 and P2 explicitly identified the possibility of opening the case detail.

"And now I can click here on detail." (P1, translated)

In the convergent high-priority case, the detailed view was generally perceived as clinically useful. P1 considered the self-report dimensions presented to be intuitive and interpreted the case as clinically relevant due to the expressiveness of the symptoms.

"I think, I think so, I think it is intuitive once again." (P1, translated)

"It seems to me that, very probably, this is a person, of course, with depression or minor or more or I do not know, or with symptoms, at least already very evident." (P1, translated)

It is important to emphasize that P1 did not interpret the dashboard as providing a diagnosis. Instead, she framed the information as useful for deciding whether additional assessment would be necessary.

"It is very useful, without a doubt, I would apply other scales, even diagnostic ones. To understand whether it is confirmed or not." (P1, translated)

When asked about the possibility of including more scales directly in the dashboard, P1 distinguished between the role of the dashboard as support for clinical review and the assessment process conducted during consultation.

"Not here, no. I would say that in my clinical practice I would apply scales, since this is not diagnostic, correct, it does not serve as a diagnosis. I would apply scales that would support my diagnosis effectively and, of course, clinical interview, etcetera." (P1, translated)

P2 also reacted positively to the detailed view, especially to the structured and visual organization of the information. She valued the presence of a participant description or clinical summary and suggested that the visual arrangement made the information easier to interpret than an exclusively textual report.

"Yes, I actually think it makes sense [...] To have this, to have here the participant description, right? The clinical summary [...] I think visually it becomes easier for the person to understand [...] If it were all written, it would become a bit more boring, I would say." (P2, translated)

These comments suggest that the detailed view may support clinical review by organizing the case around a concise summary, self-report information, the model output, and

explanatory/contextual elements. However, both participants also identified the need for clearer visual semantics. P1 requested that subscale values include denominators or maximum scale values, similarly to what happened with some questionnaire scores presented in other parts of the interface.

"For example, here rumination 26 to. Can you put it like you put it above, for example, 38/42, 30/42. Can you put it here too?" (P1, translated)

P1 also suggested that colors should be explained through a legend or description.

"Another thing [...] here there are these blue colors. In the previous case, it was red and then also green. Perhaps it is important to include a description of these colors." (P1, translated)

She also warned that some elements appeared visually or semantically isolated.

"It appears a little bit loose like this and I fear that the health professional may not be able to understand this." (P1, translated)

P2 made a similar observation, emphasizing that it is not enough to present scores or subscales without explaining their meaning and direction. For some constructs, a high value may be clinically negative, while for others it may represent a protective factor. This would require explicit interpretation through scale descriptions, color coding, or both.

"Because I do not know whether her having very high rumination score is good [...] I already know that it is easier to have the scale and the meaning right away. Yes. Instead of having to consult it separately [...] Or through colors." (P2, translated)

This suggests that the dashboard should not only present clinical or questionnaire-derived information, but also make the meaning of each construct immediately interpretable. In particular, color coding should be semantically consistent and accompanied by legends or short explanatory descriptions, without requiring external consultation.

The passive-driven or discordant case revealed more substantial challenges for clinical action. In this case, the model assigned a high priority score despite less expressive self-reported symptoms. P1 did not reject the alert, but interpreted it cautiously and indicated that she would seek to perform an additional assessment instead of acting directly on the signal.

"If this appeared to me, if I read this report, what I would do would once again be to apply effective diagnostic assessment tests. I find it a bit strange." (P1, translated)

P1 also emphasized that signals related to mouse use could reflect lack of familiarity with the device and not necessarily a pattern associated with depression. In terms of dashboard design, this implies that the interface should include contextual metadata about digital familiarity or mouse use, or at least warn that passive interaction patterns may be influenced by non-clinical factors.

"If it is related to the mouse, it only indicates one thing to me: that the person has no experience using a mouse." (P1, translated)

P2 also identified the discrepancy between self-reports and the model score as clinically relevant. She recognized the potential value of the system for identifying people who do not report symptoms, but also indicated that a high score without expressive symptoms would make her hesitate before contacting the person for an assessment.

"She does not assess as having depression, nor anxiety, nor stress. Normal. But she is above the [threshold] and quite a high [threshold]." (P2, translated)

"*Even people who do not identify the depressive symptom can be identified as having depressive symptoms and [...] need clinical evaluation."* (P2, translated)

"*So, it has 94% but no symptoms. Then perhaps I would question whether I would really call the person to schedule the assessment or not."* (P2, translated)

This indicates that the dashboard may be useful as a triage mechanism, but that alerts based predominantly on passive signals require additional support. Thus, the action supported by the dashboard does not correspond to a definitive clinical decision, but rather to a decision to review, verify, contextualize, or follow up the case, as initially proposed.

This interpretation was also reinforced by P2, who framed triage as a preliminary step that should be complemented by clinical interview/anamnesis, within the discussion of the borderline case.

"Yes. Even if after triage the ideal would be an interview to effectively understand [...] The anamnesis, what we call anamnesis, which is the clinical interview." (P2, translated)

Both participants emphasized the importance of additional contextual information before acting on ambiguous or discordant cases. P1 highlighted the need for information about mouse-use experience, while P2 identified several types of clinical context that would make the interface more useful and safer: medication, suicidal ideation, suicide attempts, sleep, physical activity, and diet.

"Whoever uses this tool in the future has to indicate their experience with the mouse." (P1, translated)

"Only in suicidal ideation it could be divided between suicide attempts and suicidal thoughts, because they are different things […] In terms of diet, perhaps it could be a factor. Physical activity [...] physical exercise is very important in people with depression." (P2, translated)

In the passive-driven case, medication was particularly important for P2, since it could explain why a person would present few self-reported symptoms despite belonging to a clinically relevant profile.

"But is she medicated? [...] One thing is a person [...] being medicated, she may respond differently than if she were not medicated […] Is she currently medicated? And then even having here in the program the medication she is taking [...] If I saw the medication, if I saw [...] attempts and thoughts, sleep, physical activity, it could help..." (P2, translated)

These comments show that the dashboard was perceived as more useful when it could be integrated into broader clinical reasoning. The current scope of the prototype was intentionally limited to the variables available in the study, but the formative evaluation, similarly to the insights from the initial requirements elicitation interviews, indicates that future versions should include or integrate contextual fields related to medication, sleep, suicidality, physical activity, and data collection conditions.

The final open-ended feedback reinforced these requirements. P1 indicated that she had already presented her suggestions orally, while P2 explicitly listed several improvements: medication, sleep, physical activity, suicidal ideation and attempts, colors, explanations of scales, access to individual questionnaire responses, and prior familiarity with the

questionnaires. P2 also stated that access to individual responses would support clinical conversation and the interpretation of results.

"It would be the issue of medication, sleep... Physical exercise. The colors...Suicidal thoughts, suicide attempts..." (P2, translated)

"Perhaps it would be useful to have here a definition of each scale, of what it assesses." (P2, translated)

"Each variable, like the mouse variable, the time interval." (P2, translated)

"To have access, literally, question by question, and how much the person answered in each question [...] I like to talk with people about each question and about the answer to each question." (P2, translated)

Insights from the initial interviews suggested that system integration would differ in hospitals and private clinics: in hospitals, there may be a formal screening consultation, whereas in private clinics the first contact tends to occur by phone, email, informal referral, or internal referral, as was the case with interviewee P2.

This prompted a brief constructive discussion at the end of this session. In this flow, it is concluded that requiring the patient to travel to the clinic solely to complete questionnaires on a standardized computer could introduce friction and reduce the operational feasibility of the system. Nevertheless, P2 recognized clinical value in prior access to information organized by the dashboard, particularly to better prepare for the initial consultation and understand the case before clinical contact. Possible compromise solutions were discussed, such as completing the questionnaires in the waiting room, an initial triage consultation, or integrating the assessment into the first session itself. Furthermore, the interview suggested that, in the future, it may be useful to allow the clinician to select the questionnaires most appropriate to the case, aligned with different styles of clinical assessment. This insight suggests that future iterations of the dashboard should consider not only the clarity of the interface but also the practical context of data collection, the necessary equipment, and the most appropriate time to integrate the assessment into the clinical pathway.

Overall, the RQ4 results suggest that the dashboard prototype was perceived as usable, visually clear, and potentially useful for supporting triage and clinical review of depression-related cases. The initial case queue and detailed view were generally well received, and the SUS results were positive. However, the evaluation also identified several design requirements for future iterations. First, the priority score would benefit from a more salient explanation embedded in the interface. Second, subscales, colors, and score directions should be accompanied by definitions, denominators, and legends. Third, passive or discordant alerts require additional contextual information before clinicians can act with confidence. Fourth, the dashboard should support clinical review through access to individual questionnaire responses and relevant contextual fields, such as medication, sleep, suicidality, physical activity, and data collection conditions.

Thus, Artifact C appears suitable as a formative prototype for organizing cases and supporting clinical review, but not yet as an autonomous operational tool. Its main contribution consists of structuring information in a way that may help professionals identify cases requiring attention or inform referrals or therapeutic aspects to explore, while simultaneously preserving the need for clinical judgement, additional assessment, and contextual interpretation.

5 Conclusions, Future Work, and Limitations

5.1 Discussion and synthesis of the results

This dissertation aimed to design, develop, and evaluate, from a Design Science Research perspective, an explainable system for clinical decision support for prioritizing cases related to depression, combining psychological self-reports with passive human-computer interaction signals. The work sought to address a gap identified in the literature: although digital phenotyping based on smartphones and wearables has been extensively explored, signals derived from mouse/cursor interactions remain less studied in clinical contexts, particularly when integrated with self-reports, explainability, and prototypes oriented to clinicians' workflow. The dissertation was organized around four research questions.

RQ1 examined whether mouse dynamics within each session were associated with intra-individual fluctuations in the severity of depressive symptoms in a small longitudinal clinical sample. This analysis addressed the need to understand whether passive signals captured during the digital completion of questionnaires could contain behavioral information relevant to symptom progression. Results provided no evidence that within-person deviations in these session-level kinematic features were associated with depression severity at the following session in this longitudinal subset. Besides, an ICC of 0.755 indicates that stable differences between participants dominate the variance in depressive severity, leaving only a limited intra-individual signal available for detection. Thus, the results of RQ1 suggest that the isolated intra-individual passive signal is insufficient for robust longitudinal screening in sparse clinical data, but clarify relevant boundary conditions for the use of mouse dynamics in mental health contexts.

The underlying idea of the second research question was to understand whether a hybrid machine learning approach, combining self-reports and passive mouse dynamics features, could support a prioritization task related to depression, and whether it outperformed setups based solely on questionnaires or solely on HCI features. This stage corresponded to Artifact A of the dissertation. The final selected model, a hybrid Random Forest evaluated through nested cross-validation, achieved an AUROC of 0.856 in the outer loop, surpassing the DASS-only baseline (0.772) and the passive-only configuration (0.790). The results indicated that the combination of self-reported information with passive features may help to distinguish individuals with clinical depressive symptom profiles from a control group, which allowed for the construction of a promising prioritization model for the studied context. However, the model output was interpreted as an operational prioritization and clinical review score, not as an automated diagnosis.

The third research question focused on the explainability of the hybrid model. For a model to be plausible in a clinical context, it is not enough to produce a classification or probability; it is necessary that its outputs can be interpreted, questioned, and integrated into professional reasoning. Artifact B addressed this need through an explanation layer based on TreeSHAP, producing local and global attributions for the selected model. The technical metrics indicated high fidelity and completeness of the Top-5 explanations, with a fidelity of 0.993, completeness of 0.980, and sparsity of 5.0, suggesting that a compact representation of five

factors almost entirely preserves the predictive behavior of the model. The formative evaluation with psychology professionals indicated that this type of explanation can support the clinical review of cases, particularly when the dominant factors correspond to clinically familiar self-reports. However, in cases where passive interaction patterns take greater weight in the explanation, clinical interpretability becomes less immediate and may raise doubts about possible confounders, such as familiarity with the mouse, digital experience, medication, completion conditions, or prior knowledge of the questionnaire, potentially prompting additional clinical review or the collection of supplementary context. Accordingly, the dissertation shows that explainability in clinical CDSS based on hybrid features, combining self-reports and cursor-based features, should not be understood solely as technical fidelity to the model, but also as semantic translation, clinical readability, contextualization of passive signals, appropriate user guidance, and integration of complementary clinical information.

RQ4 addressed the design and formative evaluation of a professional-oriented dashboard. Artifact C translated the outputs of the model and the explainability layer into a prioritization queue and a detailed case view, combining priority score, self-reports, explanatory factors, and additional context. The formative evaluation suggested an overall positive perception of usability, visual clarity, and potential utility for clinical screening/review.

The initial interviews had suggested that the usefulness of a CDSS of this type would depend on its ability to support different stages of screening and referral, from preparing the first consultation to identifying cases that require further evaluation or guidance to more appropriate professionals. By cross-referencing the evidence from the literature with data from the initial requirement elicitation interviews and the practical insights gathered in this final assessment, this work allowed for the consolidation of a set of preliminary guidelines for the design of CDSS based on digital phenotyping with mouse interaction dynamics (e.g., preference for having more false positives than false negatives). The evaluation suggests that self-report metrics are more quickly interpretable, whereas passive features derived from mouse interaction require explicit contextualization or previous guidance to avoid generating clinical ambiguity. The participants identified additional information that would make the signaling more actionable, including medication, sleep quality, digital literacy, and risk indicators, signaling that the dashboard should be understood as a tool to support clinical preparation and review, and not as a substitute for the interview or professional judgment.

Thus, the value of the dashboard lies less in automating decision-making and appears to emerge more in the organization of relevant information to support preparation, screening, and clinical review, depending on the service context and the style of different users. Its potential usefulness may vary according to the service context: in hospital settings, it may be integrated with formal moments of screening or referral; in private settings, it may support the preparation for the first consultation or the interpretation of information collected before or during the initial contact. In both cases, the main contribution of Artifact C is to demonstrate how outputs from a hybrid model can be presented in a more interpretable, contextualized, and compatible way with clinical autonomy, in line with CDSS design principles that emphasize workflow suitability, reduction of cognitive load, and preservation of the professional's role and autonomy (Abell et al. 2023; Trinkley et al. 2020).

In this way, the dissertation positions itself between predictive research in digital signals and sociotechnical translation for clinical decision support. Compared to the reviewed literature, it

distinguishes itself from works focused solely on the performance of passive sensors, often lacking a clinical interface, formative validation with professionals, and in-depth consideration of workflow integration and support-oriented design. Thus, the developed system should be understood as a formative prototype to support clinical review and prioritization, not as a tool ready for operational implementation or automated diagnosis.

5.2 Limitations

Despite these contributions, the dissertation has important limitations. The first is related to the size and composition of the sample. The dataset is relatively small, comes from a specific local context, and includes groups recruited in different settings: clinical participants from the Psychology service of ULS Trás-os-Montes and Alto Douro and control group participants from a local factory. This difference in context may introduce confounding factors related to collection environment, digital familiarity, motivation, sociodemographic characteristics, or conditions under which it was completed. Another potential limitation is prior familiarity with the questionnaires. As the instruments had already been used previously in the clinical service in paper format, some clinical participants may have been more familiar with the items or the response procedure than the participants in the control group. This could affect reading time, hesitation, and interaction patterns with the mouse regardless of mental state, and therefore should be documented in future data collections.

The second limitation concerns longitudinal density. Although RQ1 explored intra-individual associations between mouse dynamics and fluctuations in depressive symptoms, the actual data structure contained a limited number of repeated observations per participant. This constrained the robustness of longitudinal inferences and prevented the application of more complex temporal models.

The third limitation relates to the proximity between clinical diagnoses and self-reported predictors. Although the variable target was established within the clinical process, the DASS-21 scores were also used in the hybrid model and represent clinically relevant information. This relationship requires interpretative caution, but it is not deterministic, since some clinical participants had normal or non-clinical scores. Following Davis et al., the removal of clinically relevant predictors could reduce the clinical validity of the model; therefore, the hybrid model was interpreted primarily as an assessment of the incremental value of passive features relative to the self-report baseline, and not as an independent diagnostic tool.

The fourth limitation concerns the formative evaluation of the prototype. This was conducted with a small number of participants, using mockups and representative scenarios, rather than a system implemented in a real context. Thus, the collected feedback allows for the assessment of comprehensibility, perceived usefulness, and design requirements, but does not demonstrate impact on actual clinical decisions. Moreover, the asynchronous format of the responses received for the initial requirements gathering does not allow the same degree of clarification and in-depth exploration as synchronous semi-structured interviews.

5.3 Future Work

Future work should begin with collection and validation in larger, more diverse, and more balanced samples. Future studies should include multiple clinical centers, greater

sociodemographic diversity, repeated collections as well in the control group, and more documentation of contextual variables, such as device used, resolution, layout, screen size, experience with mouse, digital literacy, filling conditions, familiarity with questionnaires, and possible interruptions. This information would allow better distinction between clinically relevant behavioral patterns and technical or contextual variations. Furthermore, future pipelines should explore normalization and standardization strategies that make the models more robust to different devices, layouts, and screen sizes, as well as more task-agnostic, in order to accommodate different questionnaires or digital tasks without compromising the comparability of passive features.

A second direction consists of deepening longitudinal modeling. With greater temporal density, it would be possible to explore models oriented towards intra-individual trajectories, symptomatic change, detection of deviations from a personal baseline, and prediction of worsening or improvement. This direction aligns with recent work showing that repeated measurements can improve the ability to infer changes in mental state. However, such development would require more regular data collection, consistent protocols, and careful validation to avoid overfitting in short time series.

An important future direction is to develop dimensional or multi-target models, instead of limiting the task to binary prioritization between clinical and control participants. Future versions could predict specific dimensions like depression, anxiety, stress, rumination, or emotional regulation patterns, producing outputs that could be more useful for prioritization by specialty, and defining therapeutic targets. This need was reinforced by clinical feedback collected in late interview and assessment cycles, but integrating it would require redefining the modeling problem, preparing new targets, revalidating the pipeline, and adapting the interface, so it was treated as a future requirement rather than a change to the final predictive model.

Regarding the explainability layer, the generation of narrative explanations based on the use of language models (LLMs) to produce textual explanations in natural language may be more suited to clinical reasoning, and this could also be explored, as long as accompanied by mechanisms for verification, fidelity, and control of hallucinations (Schneider 2024).

Finally, the most important development would be a prospective evaluation in a real service context. Before any deployment, the prototype should be refined based on more co-design cycles with psychologists and other professionals, followed by usability testing, clinical simulation studies, and, subsequently, controlled studies in real workflow. This evaluation should measure not only acceptance and usability, but also impact on prioritization, review time, cognitive load, confidence, safety, equity, and recommendation quality. Only after this type of validation would it be possible to discuss the transition from a CDSS-oriented prototype to a tool effectively integrated into the service.

5.4 Scientific publications associated with this research project

This dissertation was developed within the context of a broader research initiative DeepVybe at INESC TEC. By utilizing the shared clinical dataset collected during this project, the research ecosystem fostered parallel exploratory studies, resulting in the following co-authored publications:

Mendes, Pedro, Leonor Pinto, Francisco Andrade, André Thiago Netto, Dennis Paulino, Paulo Pimentel, and Hugo Paredes. 2026. *DepressionSense Project: Digital Phenotyping Feature Engineering from Mouse Dynamics*. Paper accepted for the *International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion (DSAI 2026)*. In press.

Pinto, Leonor, Pedro Mendes, Francisco Andrade, André Thiago Netto, Dennis Paulino, Paulo Pimentel, and Hugo Paredes. 2026. *DepressionSense Project: Behavioral Digital Phenotyping for Depression Severity Discrimination in a Clinical Population*. Paper accepted for the *International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion (DSAI 2026)*. In press

References

Abbas, Qaiser, Woonyoung Jeong, and Seung Won Lee. 2025. "Explainable AI in Clinical Decision Support Systems: A Meta-Analysis of Methods, Applications, and Usability Challenges." *Healthcare* 13 (17): 2154. https://doi.org/10.3390/healthcare13172154.

Abell, Bridget, Sundresan Naicker, David Rodwell, et al. 2023. "Identifying Barriers and Facilitators to Successful Implementation of Computerized Clinical Decision Support Systems in Hospitals: A NASSS Framework-Informed Scoping Review." *Implementation Science* 18 (1): 32. https://doi.org/10.1186/s13012-023-01287-y.

Abras, Chadia, Diane Maloney-Krichmar, and Jenny Preece. 2004. *1. Introduction and History*.

Aldrich, Bernard, Yang Liu, May Almousa, and Mohd Anwar. 2023. "Translating Keystroke and Mouse Dynamics Data to Classify Human Mood." *2023 Fifth International Conference on Transdisciplinary AI (TransAI)*, September 25, 79–86. https://doi.org/10.1109/TransAI60598.2023.00023.

Arias, Daniel, Shekhar Saxena, and Stéphane Verguet. 2022. "Quantifying the Global Burden of Mental Disorders and Their Economic Value." *eClinicalMedicine* 54 (December). https://doi.org/10.1016/j.eclinm.2022.101675.

Bai, Ran, Le Xiao, Yu Guo, et al. 2021. "Tracking and Monitoring Mood Stability of Patients With Major Depressive Disorder by Machine Learning Models Using Passive Digital Data: Prospective Naturalistic Multicenter Study." *JMIR mHealth and uHealth* 9 (3): e24365. https://doi.org/10.2196/24365.

Banholzer, Nicolas, Stefan Feuerriegel, Elgar Fleisch, Georg Friedrich Bauer, and Tobias Kowatsch. 2021. "Computer Mouse Movements as an Indicator of Work Stress: Longitudinal Observational Field Study." *Journal of Medical Internet Research* 23 (4): e27121. https://doi.org/10.2196/27121.

Bayor, Andrew A., Jane Li, Ian A. Yang, and Marlien Varnfield. 2025. "Designing Clinical Decision Support Systems (CDSS)—A User-Centered Lens of the Design Characteristics, Challenges, and Implications: Systematic Review." *Journal of Medical Internet Research* 27 (June): e63733–e63733. https://doi.org/10.2196/63733.

Bell, Andrew, Malcolm Fairbrother, and Kelvyn Jones. 2019. "Fixed and Random Effects Models: Making an Informed Choice." *Quality and Quantity* 53 (March): 1051–74. https://doi.org/10.1007/s11135-018-0802-x.

Bennabi, Djamila, Pierre Vandel, Charalambos Papaxanthis, Thierry Pozzo, and Emmanuel Haffen. 2013. "Psychomotor Retardation in Depression: A Systematic Review of Diagnostic, Pathophysiologic, and Therapeutic Implications." *BioMed Research International* 2013: 158746. https://doi.org/10.1155/2013/158746.

Bergomi, Laura, Giovanna Nicora, Marta Anna Orlowska, et al. 2025. "Which Explanations Do Clinicians Prefer? A Comparative Evaluation of XAI Understandability and Actionability in Predicting the Need for Hospitalization." *BMC Medical Informatics and Decision Making* 25 (1): 269. https://doi.org/10.1186/s12911-025-03045-0.

Braund, Taylor A., Bridianne O'Dea, Debopriyo Bal, et al. 2023. "Associations Between Smartphone Keystroke Metadata and Mental Health Symptoms in Adolescents: Findings From the Future Proofing Study." *JMIR Mental Health* 10 (May): e44986. https://doi.org/10.2196/44986.

Brooke, John. 1996. *SUS -- a Quick and Dirty Usability Scale*.

Brysbaert, Marc. 2018. *Power Analysis and Effect Size in Mixed Effects Models: A Tutorial*. https://doi.org/10.31234/osf.io/fahxc.

Bzdok, Danilo, and Andreas Meyer-Lindenberg. 2018. "Machine Learning for Precision Psychiatry: Opportunities and Challenges." *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 3 (3): 223–30. https://doi.org/10.1016/j.bpsc.2017.11.007.

Chi, Yueh-Yun, Deborah H. Glueck, and Keith E. Muller. 2019. "Power and Sample Size for Fixed-Effects Inference in Reversible Linear Mixed Models." *The American Statistician* 73 (4): 350–59. https://doi.org/10.1080/00031305.2017.1415972.

Choi, Adrien, Aysel Ooi, and Danielle Lottridge. 2024. "Digital Phenotyping for Stress, Anxiety, and Mild Depression: Systematic Literature Review." *JMIR mHealth and uHealth* 12 (May): e40689. https://doi.org/10.2196/40689.

Costantini, Luigi, Cesira Pasquarella, Anna Odone, et al. 2021. “Screening for Depression in Primary Care with Patient Health Questionnaire-9 (PHQ-9): A Systematic Review.” *Journal of Affective Disorders* 279 (January): 473–83. https://doi.org/10.1016/j.jad.2020.09.131.

De France, Kalee, and Tom Hollenstein. 2017. “Assessing Emotion Regulation Repertoires: The Regulation of Emotion Systems Survey.” *Personality and Individual Differences* 119 (December): 204–15. https://doi.org/10.1016/j.paid.2017.07.018.

DeepVybe. 2026. *DeepVybe Project*.

Dhanda, Sumit Singh, Deepak Panwar, Chia-Chen Lin, et al. 2025. “Advancement in Public Health through Machine Learning: A Narrative Review of Opportunities and Ethical Considerations.” *Journal of Big Data* 12 (1): 154. https://doi.org/10.1186/s40537-025-01201-x.

Dwyer, Dominic B., Peter Falkai, and Nikolaos Koutsouleris. 2018. “Machine Learning Approaches for Clinical Psychology and Psychiatry.” *Annual Review of Clinical Psychology* 14 (1): 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037.

Epp, Clayton, Michael Lippold, and Regan L. Mandryk. 2011. “Identifying Emotional States Using Keystroke Dynamics.” *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, May 7, 715–24. https://doi.org/10.1145/1978942.1979046.

Fadul, Ruba, Aamna AlShehhi, and Leontios Hadjileontiadis. 2024. “Robust Remote Detection of Depressive Tendency Based on Keystroke Dynamics and Behavioural Characteristics.” *Scientific Reports* 14 (November): 28025. https://doi.org/10.1038/s41598-024-78489-x.

Fried, Eiko I., and Randolph M. Nesse. 2015. “Depression Is Not a Consistent Syndrome: An Investigation of Unique Symptom Patterns in the STAR*D Study.” *Journal of Affective Disorders* 172 (February): 96–102. https://doi.org/10.1016/j.jad.2014.10.010.

Gopalakrishnan, Abinaya, Raj Gururajan, Xujuan Zhou, Revathi Venkataraman, Ka Ching Chan, and Niall Higgins. 2024. “A Survey of Autonomous Monitoring Systems in Mental Health.” *WIREs Data Mining and Knowledge Discovery* 14 (3): e1527. https://doi.org/10.1002/widm.1527.

Gutiérrez-Colosía, M. R., L. Salvador-Carulla, J. A. Salinas-Pérez, et al. 2019. “Standard Comparison of Local Mental Health Care Systems in Eight European Countries.” *Epidemiology and Psychiatric Sciences* 28 (2): 210–23. https://doi.org/10.1017/S2045796017000415.

H. Birk, Rasmus, and Gabrielle Samuel. 2020. “Can Digital Data Diagnose Mental Health Problems? A Sociological Exploration of ‘Digital Phenotyping.’” *Sociology of Health & Illness* 42 (8): 1873–87. https://doi.org/10.1111/1467-9566.13175.

Harrell , Frank E. 2015. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer Series in Statistics. Springer International Publishing. https://doi.org/10.1007/978-3-319-19425-7.

Hashia, Shivani, Chris Pollett, and Mark Stamp. n.d. *ON USING MOUSE MOVEMENTS AS A BIOMETRIC*.

Hibbeln, Martin, Jeffrey Jenkins, Christoph Schneider, Joseph Valacich, and Markus Weinmann. 2017. “How Is Your User Feeling? Inferring Emotion Through Human-Computer Interaction Devices.” *MIS Quarterly* 41 (March). https://doi.org/10.25300/MISQ/2017/41.1.01.

Holzinger, Andreas, André Carrington, and Heimo Müller. 2020. “Measuring the Quality of Explanations: The System Causability Scale (SCS). Comparing Human and Machine Explanations.” *KI - Künstliche Intelligenz* 34 (2): 193–98. https://doi.org/10.1007/s13218-020-00636-z.

Holzinger, Andreas, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. 2019. “Causability and Explainability of Artificial Intelligence in Medicine.” *WIREs Data Mining and Knowledge Discovery* 9 (4): e1312. https://doi.org/10.1002/widm.1312.

Hopkin, Gareth, Holly Coole, Francesca Edelmann, et al. 2025. “Toward a New Conceptual Framework for Digital Mental Health Technologies: Scoping Review.” *JMIR Mental Health* 12 (February): e63484–e63484. https://doi.org/10.2196/63484.

Huckvale, Kit, Svetha Venkatesh, and Helen Christensen. 2019. “Toward Clinical Digital Phenotyping: A Timely Opportunity to Consider Purpose, Quality, and Safety.” *Npj Digital Medicine* 2 (1): 88. https://doi.org/10.1038/s41746-019-0166-1.

Insel, Thomas R. 2017. “Digital Phenotyping: Technology for a New Science of Behavior.” *JAMA* 318 (13): 1215. https://doi.org/10.1001/jama.2017.11295.

Jain, Sachin H., Brian W. Powers, Jared B. Hawkins, and John S. Brownstein. 2015. "The Digital Phenotype." *Nature Biotechnology* 33 (5): 462–63. https://doi.org/10.1038/nbt.3223.

Kamath, Jayesh, Roberto Leon Barriera, Neha Jain, Efraim Keisari, and Bing Wang. 2022. "Digital Phenotyping in Depression Diagnostics: Integrating Psychiatric and Engineering Perspectives." *World Journal of Psychiatry* 12 (3): 393–409. https://doi.org/10.5498/wjp.v12.i3.393.

Kaye, Jeffrey, Nora Mattek, Hiroko H. Dodge, et al. 2014. "Unobtrusive Measurement of Daily Computer Use to Detect Mild Cognitive Impairment." *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* 10 (1): 10–17. https://doi.org/10.1016/j.jalz.2013.01.011.

Khan, Simon, Charles Devlen, Michael Manno, and Daqing Hou. 2024. "Mouse Dynamics Behavioral Biometrics: A Survey." *ACM Computing Surveys* 56 (6): 154:1-154:33. https://doi.org/10.1145/3640311.

Lamba, Kamini, Shalli Rani, and Mohammad Shabaz. 2026. "Explainable Machine Learning for Mental Health Prediction from Social Media Behavior: A Nested Cross-Validation Study with SHAP and LIME Interpretability." *Discover Mental Health* 6 (1): 25. https://doi.org/10.1007/s44192-026-00373-z.

Lee, Juwon, Megan Lam, and Caleb Chiu. 2019. *Clara: Design of a New System for Passive Sensing of Depression, Stress and Anxiety in the Workplace*. https://doi.org/10.1007/978-3-030-25872-6_2.

Liang, Yunji, Xiaolong Zheng, and Daniel D. Zeng. 2019. "A Survey on Big Data-Driven Digital Phenotyping of Mental Health." *Information Fusion* 52 (December): 290–307. https://doi.org/10.1016/j.inffus.2019.04.001.

Louis, Raphael. 2022. "The Global Socioeconomic Impact of Mental Health." *SocioEconomic Challenges* 6 (2): 50–56. https://doi.org/10.21272/sec.6(2).50-56.2022.

Lundberg, Scott, and Su-In Lee. 2017. *A Unified Approach to Interpreting Model Predictions*. https://doi.org/10.48550/arXiv.1705.07874.

Lundberg, Scott M., Gabriel Erion, Hugh Chen, et al. 2019. "Explainable AI for Trees: From Local Explanations to Global Understanding." arXiv:1905.04610. Preprint, arXiv, May 11. https://doi.org/10.48550/arXiv.1905.04610.

Maas, Cora, and Joop Hox. 2005. "Sufficient Sample Sizes for Multilevel Modeling." *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 1 (January): 86–92. https://doi.org/10.1027/1614-2241.1.3.86.

Mienye, Ibomoiye Domor, George Obaido, Nobert Jere, et al. 2024. "A Survey of Explainable Artificial Intelligence in Healthcare: Concepts, Applications, and Challenges." *Informatics in Medicine Unlocked* 51: 101587. https://doi.org/10.1016/j.imu.2024.101587.

Mohr, David C., Francisca Azocar, Andrew Bertagnolli, et al. 2021. "Banbury Forum Consensus Statement on the Path Forward for Digital Mental Health Treatment." *Psychiatric Services* (Washington, D.C.) 72 (6): 677–83. https://doi.org/10.1176/appi.ps.202000561.

Monaro, Merylin, Andrea Toncini, Stefano Ferracuti, et al. 2018. "The Detection of Malingering: A New Tool to Identify Made-Up Depression." *Frontiers in Psychiatry* 9 (June): 249. https://doi.org/10.3389/fpsyt.2018.00249.

Moshe, Isaac, Yannik Terhorst, Kennedy Opoku Asare, et al. 2021. "Predicting Symptoms of Depression and Anxiety Using Smartphone and Wearable Data." *Frontiers in Psychiatry* 12: 625247. https://doi.org/10.3389/fpsyt.2021.625247.

Nakagawa, Shinichi, and Holger Schielzeth. 2013. "A General and Simple Method for Obtaining R2 from Generalized Linear Mixed-Effects Models." *Methods in Ecology and Evolution* 4 (2): 133–42. https://doi.org/10.1111/j.2041-210x.2012.00261.x.

National Human Genome Research Institute. 2026. "Phenotype." January 4. https://www.genome.gov/genetics-glossary/Phenotype.

National Institute for Health and Care Excellence: Guidelines, National Institute for Health and Care Excellence: Guidelines. 2022. *Depression in Adults: Treatment and Management*. National Institute for Health and Care Excellence: Guidelines. National Institute for Health and Care Excellence (NICE). http://www.ncbi.nlm.nih.gov/books/NBK583074/.

Netto, André Thiago C., Dennis Paulino, Artur Rocha, Joana Ferreira de Raposo, and Hugo Paredes. 2025. "Analysis of Users' Digital Phenotyping to Infer and Prevent Mental Health: A Work in Progress." *Proceedings of the 11th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-Exclusion* (New York, NY, USA), DSAI '24, July 31, 143–46. https://doi.org/10.1145/3696593.3696648.

Onnela, Jukka-Pekka, and Scott L. Rauch. 2016. "Harnessing Smartphone-Based Digital Phenotyping to Enhance Behavioral and Mental Health." *Neuropsychopharmacology* 41 (7): 1691–96. https://doi.org/10.1038/npp.2016.7.

Ostojic, Dijana, Paris Alexandros Lalousis, Gary Donohoe, and Derek W. Morris. 2024. "The Challenges of Using Machine Learning Models in Psychiatric Research and Clinical Practice." *European Neuropsychopharmacology* 88 (November): 53–65. https://doi.org/10.1016/j.euroneuro.2024.08.005.

Parvandeh, Saeid, Hung-Wen Yeh, Martin P. Paulus, and Brett A. McKinney. 2020. "Consensus Features Nested Cross-Validation." *Bioinformatics* 36 (10): 3093–98. https://doi.org/10.1093/bioinformatics/btaa046.

Racine, Nicole, Brae Anne McArthur, Jessica E. Cooke, Rachel Eirich, Jenney Zhu, and Sheri Madigan. 2021. "Global Prevalence of Depressive and Anxiety Symptoms in Children and Adolescents During COVID-19: A Meta-Analysis." *JAMA Pediatrics* 175 (11): 1142. https://doi.org/10.1001/jamapediatrics.2021.2482.

Rehm, Jürgen, and Kevin D. Shield. 2019. "Global Burden of Disease and the Impact of Mental and Addictive Disorders." *Current Psychiatry Reports* 21 (2): 10. https://doi.org/10.1007/s11920-019-0997-0.

Ribeiro, J. D., J. C. Franklin, K. R. Fox, et al. 2016. "Self-Injurious Thoughts and Behaviors as Risk Factors for Future Suicide Ideation, Attempts, and Death: A Meta-Analysis of Longitudinal Studies." *Psychological Medicine* 46 (2): 225–36. https://doi.org/10.1017/S0033291715001804.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." arXiv:1602.04938. Preprint, arXiv, August 9. https://doi.org/10.48550/arXiv.1602.04938.

Rosenbacke, Rikard, Åsa Melhus, Martin McKee, and David Stuckler. 2024. "How Explainable Artificial Intelligence Can Increase or Decrease Clinicians' Trust in AI Applications in Health Care: Systematic Review." *JMIR AI* 3 (October): e53207. https://doi.org/10.2196/53207.

Roychowdhury, Sneha, Vita Lanfranchi, and Suvodeep Mazumdar. 2025. "Evaluating Explanation Performance for Clinical Decision Support Systems for Non-Imaging Data: A Systematic Literature Review." *Computers in Biology and Medicine* 197 (October): 110944. https://doi.org/10.1016/j.compbiomed.2025.110944.

Sahoo, Soumyashree, Zakir Hossain, Chinmaey Shende, et al. 2025. "Smartphone Data Gathered Early in Depression Treatment Predicts Treatment Outcome." *Proceedings of the ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies*, June 24, 93–104. https://doi.org/10.1145/3721201.3721380.

Schneider, Johannes. 2024. "Explainable Generative AI (GenXAI): A Survey, Conceptualization, and Research Agenda." *Artificial Intelligence Review* 57 (11): 289. https://doi.org/10.1007/s10462-024-10916-x.

Schröer, Christoph, Felix Kruse, and Jorge Marx Gómez. 2021. "A Systematic Literature Review on Applying CRISP-DM Process Model." *Procedia Computer Science*, CENTERIS 2020 - International Conference on ENTERprise Information Systems / ProjMAN 2020 - International Conference on Project MANagement / HCist 2020 - International Conference on Health and Social Care Information Systems and Technologies 2020, CENTERIS/ProjMAN/HCist 2020, vol. 181 (January): 526–34. https://doi.org/10.1016/j.procs.2021.01.199.

Seelye, Adriana, Stuart Hagler, Nora Mattek, et al. 2015. "Computer Mouse Movement Patterns: A Potential Marker of Mild Cognitive Impairment." *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 1 (4): 472–80. https://doi.org/10.1016/j.dadm.2015.09.006.

Shiffman, Saul, Arthur A. Stone, and Michael R. Hufford. 2008. "Ecological Momentary Assessment." *Annual Review of Clinical Psychology* 4: 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415.

Shulha, Michael, Jordan Hovdebo, Vinita D'Souza, Francis Thibault, and Rola Harmouche. 2024. "Integrating Explainable Machine Learning in Clinical Decision Support Systems: Study Involving a Modified Design Thinking Approach." *JMIR Formative Research* 8 (April): e50475. https://doi.org/10.2196/50475.

Snijders, and Bosker. 1999. *Multilevel Analysis : An Introduction to Basic and Advanced Multilevel Modeling*. With Internet Archive. London ; Thousand Oaks, Calif. : Sage Publications. http://archive.org/details/multilevelanalys0000snij.

Sun, David, Pablo Paredes, and John Canny. 2014. "MouStress: Detecting Stress from Mouse Motion." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, April 26, 61–70. https://doi.org/10.1145/2556288.2557243.

Sunjaya, Anthony Paulo, Allison Martin, and Christine Jenkins. 2022. "A Design Thinking Approach to Developing a Clinical Decision Support System for Breathlessness in Primary Care." In *MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation*. IOS Press. https://doi.org/10.3233/SHTI220197.

Tonekaboni, Sana, Shalmali Joshi, Melissa D. McCradden, and Anna Goldenberg. 2019. *What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use*.

Torous, John, Sandra Bucci, Imogen H. Bell, et al. 2021. "The Growing Field of Digital Psychiatry: Current Evidence and the Future of Apps, Social Media, Chatbots, and Virtual Reality." *World Psychiatry: Official Journal of the World Psychiatric Association* (WPA) 20 (3): 318–35. https://doi.org/10.1002/wps.20883.

Torous, John, Mathew V. Kiang, Jeanette Lorme, and Jukka-Pekka Onnela. 2016. "New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research." *JMIR Mental Health* 3 (2): e16. https://doi.org/10.2196/mental.5165.

Trinkley, Katy E., Michael G. Kahn, Tellen D. Bennett, et al. 2020. "Integrating the Practical Robust Implementation and Sustainability Model With Best Practices in Clinical Decision Support Design: Implementation Science Approach." *Journal of Medical Internet Research* 22 (10): e19676. https://doi.org/10.2196/19676.

Ware, Shweta, Chaoqun Yue, Reynaldo Morillo, et al. 2020. "Predicting Depressive Symptoms Using Smartphone Data." *Smart Health* 15 (March): 100093. https://doi.org/10.1016/j.smhl.2019.100093.

Weilnhammer, Veith, Jefferson Ortega, and David Whitney. 2025. "Human-Computer Interactions Predict Mental Health." arXiv:2511.20179. Preprint, arXiv, December 17. https://doi.org/10.48550/arXiv.2511.20179.

Wilimitis, Drew, and Colin G. Walsh. 2023. "Practical Considerations and Applied Examples of Cross-Validation for Model Development and Evaluation in Health Care: Tutorial." *JMIR AI* 2 (December): e49023. https://doi.org/10.2196/49023.

World Mental Health Report: Transforming Mental Health for All. 2022. 1st ed. World Health Organization.

Yamauchi, Takashi. 2013. *Mouse Trajectories and State Anxiety: Feature Selection with Random Forest*. In *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*. https://doi.org/10.1109/ACII.2013.72.

Yamauchi, Takashi, and Kunchen Xiao. 2018. "Reading Emotion From Mouse Cursor Motions: Affective Computing Approach." *Cognitive Science* 42 (3): 771–819. https://doi.org/10.1111/cogs.12557.

Zantvoort, Kirsten, Jennifer J. Matthiesen, Pontus Bjurner, et al. 2025. "The Promise and Challenges of Computer Mouse Trajectories in DMHIs – A Feasibility Study on Pre-Treatment Dropout Predictions." *Internet Interventions* 40 (June): 100828. https://doi.org/10.1016/j.invent.2025.100828.

Zhan, Yuting, Hongwei Liu, and Yu Wang. 2026. "Digital Phenotyping of Depression: A Multi-Modal Passive Sensing Approach to Identifying Novel Behavioral and Physiological Markers of Treatment Response." *Journal of Psychiatric Research* 194 (March): 40–50. https://doi.org/10.1016/j.jpsychires.2025.12.036.

Zhan, Yuting, and Yu Wang. 2026. "Network Dynamics between Digital Media Use and Depressive Symptomatology: A Computational Phenotyping Approach Using Ecological Momentary Assessment." *Psychiatry Research* 356 (February): 116896. https://doi.org/10.1016/j.psychres.2025.116896.

APPENDIX A: Structured Review Strategy and Expanded Evidence Matrix

Structured Search and Selection Criteria

A structured review was carried out to identify primary empirical studies that implement and evaluate models based on passive digital signals for mental health assessment and monitoring, using relevant clinical outcomes (e.g., depression, anxiety). The evidence matrix was constructed to support the formulation and refinement of the research questions, particularly by examining the extent to which empirical studies had evaluated passive HCI-based interaction dynamics in longitudinal analyses or in relation to clinician-facing decision-support workflows for depression-related assessment. The evidence matrix was constructed from iterative searches in ResearchRabbit, Scopus, Elsevier/ScienceDirect and PubMed up to March 6, 2026. The search strategy was guided by keyword groups from two main domains: behavioral sensing modalities (e.g., mouse dynamics, keyboard or keystroke dynamics, cursor tracking, human–computer interaction, digital phenotyping) and mental health–related terms (e.g., depression, major depressive disorder, mental health prediction). Machine learning-related terms were also occasionally used to capture studies implementing predictive models. Forward and backward citation tracking was also used to identify additional relevant studies.

Inclusion criteria were as follows: (i) primary empirical studies reporting an implemented and evaluated model (excluding reviews or purely conceptual papers); (ii) use of passive or hybrid digital behavioral signals (at least mouse, keyboard, or smartphone/wearable sensors); (iii) clinical outcome related to mental health (depression, anxiety, stress, or distress) utilizing validated psychiatric scales (e.g., PHQ-9, QIDS, MADRS), structured clinical interviews (e.g., DSM-5 criteria), or objective clinical endpoints (e.g., treatment dropout) as the target variable for their predictive models; (iv) quantitative results reported (metrics, N, performance). While priority was given to peer-reviewed publications, highly relevant preprints were also considered when they provided substantial empirical evidence (e.g., the MAILA framework).

Criteria for Longitudinal Analysis Classification: To systematically evaluate the predictive temporal capacity of the models, a strict three-tier decision rule was applied to classify studies in the "Longitudinal Analysis" column. A study was only flagged as *YES* if it successfully met all three of the following conditions: (1) a longitudinal design with at least 2 data collection timepoints per participant; (2) an analytical approach that explicitly modeled *within-person change* or trajectories (either by extracting intra-individual mathematical deltas between periods or by utilizing natively sequential models like RNNs/LSTMs); and (3) a temporal clinical target, meaning the algorithm was trained to predict a future clinical state, a treatment response trajectory, or a mood fluctuation, rather than diagnosing a concurrent cross-sectional state at the end of a monitoring window.

Exclusion Criteria:

1. **Self-Reported Digital Behavior:** Studies evaluating digital behavior solely through active surveys (e.g., asking users to estimate their screen time or social media usage) were excluded to avoid recall bias.
2. **Lack of Clinical Validation:** Studies lacking a robust clinical ground truth, such as those relying on unvalidated online surveys to determine mental health risk in general populations, were excluded.

3. **Overlapping Cohorts:** To prevent duplication bias, when multiple papers reported on the exact same patient cohort or research project, only the most methodologically advanced or comprehensive iteration was included.
4. **Out of Scope:** Studies focusing primarily on neurological or age-related conditions (e.g., mild cognitive impairment in geriatrics) rather than psychiatric populations were excluded. Studies relying strictly on artificially stress-induced tasks in laboratory settings were also excluded.

Criteria for *Workflow Integration / CDSS* Classification: For the purpose of this matrix, "research-only" refers to studies that implemented and evaluated predictive or explanatory models without considering anything related with CDSS design. Studies were considered "prototype/CDSS-oriented" when they explicitly explored decision-support design, even if no real-world deployment occurred. "Workflow-integrated" was reserved for systems evaluated within a real operational or trial-based clinical workflow.

Limitations of the Evidence Matrix: Despite the systematic effort, this matrix has important limitations: (1) the search prioritized terms focused on mouse/keyboard/cursor dynamics and digital phenotyping combined with *depression / major depressive disorder / predict mental health/ machine learning* — an exhaustive scan with specific terms exclusively related to clinical decision support systems (CDSS) was not conducted (so we cannot state with 100% certainty that no operational CDSS exists in this domain); (2) the inclusion of preprints (e.g., MAILA) and small-sample studies increases the risk of publication bias and overfitting; (3) methodological heterogeneity (different sensor modalities, varied clinical targets, and non-homogeneous evaluation metrics) limits direct head-to-head comparisons; and (4) language and journal restrictions may have excluded relevant studies in other languages or localized technical reports.

Consequently, the resulting evidence matrix provides strong support for the presence of a research gap—particularly regarding the absence of (longitudinal) clinical decision support systems based specifically on mouse dynamics evaluated in clinical populations—but it does not constitute definitive proof of the nonexistence of all possible implementations. A complementary search focusing explicitly on clinical decision support terminology was also conducted using Google Scholar; no operational longitudinal CDSS based on mouse or keyboard dynamics were identified. The most closely related work found was the conceptual framework **CLARA**, which proposes an architecture integrating HCI signals with repeated DASS-21 assessments but has not yet been empirically validated as a functioning clinical decision support system.

Table 11 Table A.1: Expanded evidence matrix of digital phenotyping studies.

Study	*Design / sample*	*Longitudinal Analysis*	*Modality*	*Target / ground truth*	*Methods*	*Main result*	*Main limitation*	*CDSS / workflow status*
Fadul et al., 2024	Cross-sectional clinical / longitudinal sensing; N=24	No	Smartphone keystroke dynamics	Depressive tendency; PHQ-9 cutoff ≥5	Gradient Boosting; Mutual Information feature selection	AUC = 0.98; ACC = 95.8%; hold-time features most predictive	Very small sample; single self-report label; Android-only	Predictive Modeling Only
Bai et al., 2021	Prospective longitudinal multicenter, 12 weeks; N=334 MDD outpatients	Yes	Smartphone + wearable sensing	Mood stability vs. swings from biweekly PHQ-9	L1 and tree-based feature selection; KNN and Random Forest best	ACC = 76.7%; recall = 90.4%; subgroup ACC >80%	Class imbalance; Android-only; limited sample size	Predictive Modeling Only
Moshe et al., 2021	Longitudinal observational, 4 weeks; N=55	Yes	Smartphone GPS/usage + Oura ring + EMA	Depression, anxiety and stress symptoms; repeated DASS-21	Multilevel models with random intercepts and slopes	Location variance predicted lower depression; sleep/HRV features associated with symptoms; combined EMA+sensing best fit	Small non-clinical sample; observational; temporal aggregation limits high-frequency dynamics	Predictive Modeling Only
Zhan et al., 2026	Longitudinal, 1–12 weeks; N=183 MDD starting treatment	Yes	Smartphone + wearables	Treatment response; ≥50% MADRS reduction at 12 weeks	Linear mixed-effects models; k-means with DTW; Random Forest	ACC = 76.4%; AUC = 0.84	Single geographic region	Predictive Modeling Only
Zhan & Wang, 2026	Longitudinal, 60 days; N=1642 young adults	Yes	Smartphone metrics + EMA	Depressive symptom dynamics; EMA momentary reports	mlVAR; PCA; Gaussian Mixture Modeling; mixed-effects growth models; TVEM; Random Forest / Gradient Boosting	Four network phenotypes identified; prognostic $R^2 \approx 0.39$	Young adult sample; observational; self-report	Predictive Modeling Only
Sahoo et al., 2025	Longitudinal observational, 12 weeks; N=87 depression patients starting medication	Yes	Smartphone GPS/WiFi + EMA + weekly medication surveys	Treatment response; ≥50% QIDS reduction	Time-series deep learning; GRU-D; BRITS	F1 = 0.68–0.73 with smartphone data; max F1 = 0.77 with combined data	Small sample; high missing location data; predominantly female; self-reported labels	Predictive Modeling Only

Zantvoort et al., 2025	Feasibility study, baseline assessment; N=183	No	Mouse trajectories from desktop web tracker	Treatment dropout; <6 or <7 modules completed	Random Forest; 1D-CNN	Hand-crafted features AUC = 0.58; 1D-CNN AUC = 0.50	Small sample; mobile users excluded; single questionnaire;	Predictive Modeling Only, although tested within a clinical trial context (SOPHIA)
Braund et al., 2023	Cross-sectional baseline in prospective cohort; N=934 adolescents	No	Smartphone keystroke metadata from active typing tasks	Symptoms using PHQ-A, CAS-8, DQ-5, ISI	Correlations; Elastic Net; SVM; Neural Nets; RF; XGBoost	Very weak associations; keystrokes alone failed to predict symptoms	Cross-sectional; active tasks; distance between keys not collected	Predictive Modeling Only
Ware et al., 2020	Longitudinal field study, 8–10 months; N=182 students	No	Smartphone location and/or campus WiFi metadata	Depressive symptoms and severity; PHQ-9, QIDS sub-scores, DSM-V interview	SVM-RBF; SVM-RFE; SVR	F1 up to 0.86 for individual depressive symptoms; WiFi comparable to smartphone sensing	Student sample; missing location data; unbalanced classes; small depressed subgroup; self-reports	Predictive Modeling Only
Monaro et al., 2018	Cross-sectional laboratory experiment; N=87	No	Desktop mouse dynamics in active double-choice task	Malingered vs. genuine depression; clinical diagnosis + instructed condition	CFS; Naive Bayes; SMO; Logistic Model Tree; Random Forest	3-class ACC up to 96.3%; liars vs. depressed ACC up to 94.4%	Small sample; artificial instructed malingering; cognitive confounding; no clinician comparison	Predictive Modeling Only
Weilnhammer et al., 2025 / MAILA	Cross-sectional and longitudinal; ~9000 participants	Yes	Cursor + touchscreen recordings	Latent mental states; depression classification; symptom change	LSTM autoencoder; K-means; SVR; SVC	AUC = 0.64 for depression classification; R = 0.48 for longitudinal change prediction; AUC = 0.73 for improvement vs. worsening	Preprint status; dataset heterogeneity; replication needed	Predictive Modeling Only

APPENDIX B: Original Study Protocol

Protocolo do Estudo / Projeto de Investigação

Título do Estudo: "Rastreio da Depressão com Recurso a Algoritmos de Inteligência Artificial na Análise da Interação Pessoa-Computador"

1. Introdução

Este projeto de investigação propõe-se a explorar o potencial da inteligência artificial no rastreio de perturbações mentais, com foco na depressão. A investigação baseia-se na análise de padrões de interação com o computador — nomeadamente movimentos do rato e utilização do teclado — durante o preenchimento de instrumentos de avaliação psicológica validados, com o objetivo de identificar indicadores digitais de estados depressivos.

2. Objetivos

Anualmente, cerca de 30% da população da União Europeia é afetada por perturbações mentais e comportamentais, incluindo a depressão (Liu et al., 2017). Esta é uma das condições mais prevalentes a nível mundial (Vos et al., 2019), afetando cerca de 280 milhões de pessoas (Costantini et al., 2021). É comum nos cuidados de saúde primários (Fekadu et al., 2022) e caracteriza-se por manifestações heterogéneas, afetando domínios emocionais, cognitivos e comportamentais, podendo em casos graves conduzir ao suicídio (Gopalakrishnan et al., 2024).

Essa variabilidade sintomática torna o diagnóstico complexo (Rykov et al., 2021). Tradicionalmente, o rastreio baseia-se em questionários de autorrelato (Asare et al., 2021) — como a Escala de Ansiedade Depressão e Stress – (DASS-21), Padrões de Regulação das Emoções (RESS) e o Working Alliance Inventory - Short Revised (WAI-SR)— e entrevistas clínicas. Contudo, estes instrumentos podem estar sujeitos a imprecisões e enviesamentos (Filho et al., 2021; Chandrasekaran et al., 2021).

Com o avanço da capacidade computacional e dos algoritmos, a aplicação de machine learning (ML) à saúde mental tem ganho força (Dwyer et al., 2018). Uma abordagem promissora é o Digital Phenotyping (DP), que propõe a recolha e análise de dados digitais, incluindo aqueles derivados da interação com a máquina (Torous et al., 2016; Jain et al., 2015). Estes dados permitem inferir padrões relacionados com a saúde física e mental (Liang et al., 2019).

Este estudo tem como objetivo avaliar se algoritmos de ML, treinados com base em padrões de interação pessoa-computador recolhidos durante o preenchimento digital de questionários clínicos, podem melhorar a qualidade dos dados usados no rastreio da depressão.

Objetivos específicos:

1. Recolher dados de interação (movimentos do rato e digitação) durante o preenchimento dos questionários validados (Asare et al., 2021).
2. Treinar e testar modelos de ML para identificar padrões associados à depressão.
3. Ajustar métricas de avaliação, otimizando o desempenho dos modelos.

4. O processo de recolha de dados e treino busca a elaboração de um modelo de inteligência artificial (IA) capaz de inferir depressão durante a interação entre o utilizador e o computador, ao responder a questionários de excelência.

Referências

• Asare et al. (2021). JMIR mHealth and uHealth, 9(7), e26540. https://doi.org/10.2196/26540

• Chandrasekaran et al. (2024). npj Mental Health Research, 3, 62. https://doi.org/10.1038/s44184-024-00107-5

• Costantini et al. (2021). J. Affective Disorders, 279, 473–483. https://doi.org/10.1016/j.jad.2020.09.131

• Dwyer et al. (2018). Annual Review of Clinical Psychology, 14, 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

• Fekadu et al. (2022). Systematic Reviews, 11(1), 21. https://doi.org/10.1186/s13643-022-01893-9

• Filho et al. (2021). IEEE CBMS. https://doi.org/10.1109/CBMS52027.2021.00076

3. Metodologia

Esta investigação insere-se no campo da informática, onde a metodologia Design Science Research (DSRM) é amplamente reconhecida como abordagem válida (Lee et al., 2015). Por essa razão, o presente estudo adotará essa estrutura metodológica.

O DSRM destaca-se pela utilização do conhecimento científico na conceção de artefactos (Cross, 2001), desenvolvidos com o objetivo de resolver problemas de forma mais eficiente ou de propor soluções inovadoras para questões ainda não solucionadas (Hevner et al., 2004). Neste projeto, será seguido o modelo estruturado em seis etapas, conforme delineado por Peffers et al. (2007), cujas especificidades serão apresentadas:

Plano de trabalho com base nas seis atividades do DSRM:

1. 1 Identificação do problema e motivação – Existe uma necessidade de maior precisão nos dados e escalonamento das soluções que apoiam o diagnóstico da depressão, condição que afeta cerca de 280 milhões de pessoas no mundo (Costantini et al., 2021).

2. Definição de objetivos para a solução

a. Mapear as interações com rato e teclado durante o preenchimento de questionários de rastreio da depressão;

b. Desenvolver um algoritmo de IA capaz de analisar os registos de interação e identificar padrões indicativos de depressão.

3. Design e desenvolvimento

a. Desenvolvimento de uma biblioteca JavaScript para monitorizar movimentos do rato, digitação e trocas de janela durante os questionários;

b. Desenvolvimento de Software com instrumentos de avaliação (WAI-SR, DASS-21 e RESS) e integração com a biblioteca desenvolvida;

c. Treino e validação de modelos de aprendizagem automática baseados nesses dados.

4. Demonstração

a. Os resultados serão apresentados em formato gráfico com uso da biblioteca D3.js, ilustrando padrões e previsões do modelo.

5. Avaliação

a. Com o modelo já treinado, será realizado testes com outra amostra diferente da utilizada para treino para verificar a acurácia e precisão.

6. Comunicação

a. A difusão dos resultados será feita paralelamente às demais etapas, através da publicação de artigos científicos em conferências e revistas especializadas.

4. Participantes

A amostra envolverá utentes do serviço de Psicologia Clínica daUnidade Local de Saúde de Trás-os-Montes e Alto Douro (ULSTMAD). Pessoas maiores de 18 anos, de língua portuguesa.

O tamanho da amostra assume um papel determinante no desempenho e na capacidade de generalização dos algoritmos de aprendizagem automática. Amostras reduzidas tendem a originar modelos com elevada variância, susceptíveis ao sobreajuste (overfitting), ao passo que conjuntos de dados mais extensos conduzem, em geral, a estimativas mais estáveis e fidedignas dos parâmetros do modelo (Halevy, Norvig & Pereira, 2009). Além disso, em alguns contextos, mais dados podem ser mais eficazes do que algoritmos mais sofisticados na melhoria do desempenho preditivo, sobretudo em tarefas de elevada complexidade (Banko & Brill, 2001). É possível verificar na literatura que o aumento do tamanho da amostra está fortemente correlacionado com melhorias na acuidade dos modelos, até a um ponto de saturação (Sun et al., 2017).

5. Recrutamento e triagem

A forma de recrutamento dos participantes no estudo decorre conforme descrito abaixo:

I. Os utentes que integram o Serviço de Psicologia Clínica da ULSTMAD são encaminhados para este serviço no seguimento de avaliação clínica interna ou por iniciativa própria, no âmbito da procura de apoio psicológico. Após marcação da consulta, são avaliados por um profissional de saúde qualificado. Sempre que, no decurso da prática clínica, for identificado um quadro sintomatológico que justifique a aplicação de instrumentos de avaliação psicológica padronizados, os utentes poderão ser convidados a participar no presente estudo.

II. O diretor do Serviço de Psicologia Clínica da ULSTMAD, Dr. Paulo Pimentel, procederá à triagem. Os utentes convidados a participar receberão um consentimento informado que será igualmente explicado de forma oral. O processo será conduzido de forma clara, direta e acessível, com ênfase na natureza voluntária da participação, na possibilidade de desistência a qualquer momento e na inexistência de qualquer repercussão clínica ou institucional em caso de recusa. Considerando o perfil da população-alvo, o procedimento de explicação será adaptado para garantir compreensão e evitar qualquer sensação de dúvida, coação ou influência indevida.

III. O recrutamento será feito pelo profissional de saúde, no contexto das atividades habituais do serviço. Cada utente poderá realizar a triagem mais do que uma vez, considerando que os dados recolhidos em cada sessão poderão ser relevantes para o acompanhamento clínico pelo hospital, além do objetivo de investigação científica.

Os dados pessoais identificáveis dos utentes não serão, em caso algum, divulgados. Os dados utilizados para o estudo dizem respeito exclusivamente aos padrões de interação com rato e teclado, os quais serão associados a códigos genéricos, sem qualquer vínculo direto com a identidade dos participantes. A informação tratada terá natureza anonimizada ou pseudonimizada, respeitando os princípios do RGPD.

IV. A população proposta será composta por utentes do Serviço de Psicologia Clínica da ULSTMAD, maiores de 18 anos, falantes de português, e com capacidade de compreensão e consentimento informado. Estes utentes serão encaminhados para participação no estudo após indicação clínica e triagem prévia conduzida por profissional qualificado (Dr. Paulo Pimentel).

Critérios de inclusão:

a. Utentes com idade igual ou superior a 18 anos;

b. Utentes que realizam avaliação psicológica no ULSTMAD;

c. Utentes capazes de compreender o procedimento e assinar o consentimento informado;

d. Utentes que concordem voluntariamente em participar no estudo.

Critérios de exclusão:

a. Utentes com perturbações cognitivas severas que impeçam a compreensão ou execução da tarefa;

b. Utentes com diagnóstico clínico confirmado de depressão maior ou outra perturbação psiquiátrica grave no momento da triagem;

c. Utentes não fluentes em português;

d. Utentes que, mesmo após explicação, não se sintam confortáveis ou recusem participar.

V. O armazenamento e a segurança dos dados ficarão a cargo do Dr. Paulo Pimentel, que procederá à gravação dos dados ao final de cada sessão de triagem. Os dados serão guardados num computador sem ligação à internet, e, diariamente, esse equipamento será armazenado em cofre físico no interior do hospital, garantindo a proteção contra acessos não autorizados.

6. Procedimento

I. Toda a recolha será realizada no Serviço de Psicologia Clínica da ULSTMAD, local onde diariamente ocorrem diversas consultas com utentes que procuram apoio psicológico. Após diagnóstico clínico que justifique a realização dos instrumentos de avaliação selecionados, os utentes serão convidados a participar no estudo. Os instrumentos utilizados serão:

a. Escala de Ansiedade, Depressão e Stress (DASS-21): Avalia três dimensões psicológicas (ansiedade, depressão e stress), com 21 itens (Ribeiro et al., 2004).

b. Padrões de Regulação das Emoções (RESS): Avalia estratégias adaptativas e desadaptativas de regulação emocional (De France & Hollenstein, 2017).

c. Working Alliance Inventory – Short Revised (WAI-SR): Mede a qualidade da aliança terapêutica entre terapeuta e utente (Munder et al., 2010).

Os instrumentos de avaliação psicológica utilizados neste estudo foram previamente disponibilizados pelo Dr. Paulo Pimentel em formato PDF (Apêndice 1.pdf), tal como eram aplicados no serviço clínico — em suporte papel, preenchido manualmente pelos utentes. A equipa de investigação procedeu à digitalização integral desses instrumentos, desenvolvendo uma aplicação informática em HTML, CSS e JavaScript, que permite a sua realização em formato digital, através de computador. Esta transição para o formato digital facilita significativamente o armazenamento e gestão dos dados no contexto clínico, enquanto, para efeitos de investigação, possibilita a recolha de métricas detalhadas de interação do utilizador com o sistema (como movimentos do rato, padrões de digitação e trocas de ecrã) recolhidas por uma biblioteca JavaScript específica (Behavioral Digital Phenotyping Library1) durante o preenchimento dos questionários.

II. A aplicação será feita através de um software desenvolvido para este fim, instalado num computador portátil de uso exclusivo do ULSTMAD. O psicólogo Dr. Paulo Pimentel acompanhará cada sessão e será o responsável por apresentar o procedimento aos participantes.

III. O procedimento terá início com o acolhimento do utente, que será convidado a sentar-se frente ao computador. Antes de iniciar os testes, será apresentado e explicado o Termo de Consentimento Informado, que o participante deverá ler e aceitar. A seguir, o participante completará, de forma autónoma, os três instrumentos digitais, clicando em “Continuar para a Avaliação Seguinte” entre cada um, e finalizando com o botão “Finalizar Avaliação”.

IV. Durante a realização dos testes, caso algum participante apresente dúvidas, o psicólogo prestará apoio, e tal assistência será registada. A recolha de dados será feita automaticamente, incluindo cliques, tempo de resposta, teclas pressionadas e movimentos do cursor. A cada participante será atribuído um código aleatório (ex.: “part387”), não sendo armazenados dados que permitam identificar diretamente os utentes. O resultado clínico dos testes ficará apenas no sistema interno do serviço, sem qualquer ligação com os dados utilizados na investigação. Estes dados clínicos continuarão a ser utilizados exclusivamente pelo psicólogo responsável, como já acontecia anteriormente nos registos em papel.

V. Os dados serão armazenados num computador sem ligação à internet. Ao final de cada dia, o equipamento será guardado em cofre físico sob responsabilidade do Dr. Paulo Pimentel. Nenhuma gravação de áudio ou vídeo será realizada. O início da recolha está previsto para o primeiro semestre de 2025, com sessões de aproximadamente 30 minutos, podendo ser repetidas para o mesmo participante.

VI. Não haverá gravações audiovisuais.

7. Métodos e instrumentos de recolha de dados:

• Não há recolha de dados sensíveis;

Durante a realização dos questionários, será feita a recolha automática dos seguintes dados:

• Registos da Interação do Rato e teclado

• Alterações de foco entre janelas.

8. Privacidade e Confidencialidade:

A privacidade dos participantes será assegurada através de diversas medidas técnicas e éticas. Não existe qualquer ligação entre os dados clínicos recolhidos durante a triagem e os dados utilizados para investigação. Nenhuma informação pessoal identificadora (como nome, número de processo clínico ou contacto) é recolhida no âmbito do estudo. Para efeitos de organização dos dados, será atribuído um código pseudonimizado a cada participante (ex.: "part387"), que não permite, por si só, a reidentificação dos utentes, sendo a chave de correspondência mantida exclusivamente sob responsabilidade do psicólogo Dr. Paulo Pimentel. Este procedimento assegura que a identificação dos participantes é altamente improvável, respeitando os princípios do Regulamento Geral sobre a Proteção de Dados (RGPD).

9. Considerações Éticas

O estudo segue os princípios da Declaração de Helsínquia, da Convenção de Oviedo e está alinhado com o Regulamento Geral sobre a Proteção de Dados (RGPD).

Todos os participantes serão informados sobre a natureza do estudo e os seus direitos, e assinarão um termo de consentimento informado. A participação é voluntária e poderá ser interrompida a qualquer momento, sem prejuízo para o utente. O estudo foi previamente avaliado pela Comissão de Ética da UTAD.

APPENDIX C: Engineered Feature Set and Feature Definitions

Table 12 Table C.1: Engineered mouse dynamics feature set and definitions.

Feature	*Raw source*	*Simple formula / definition*	*Sources*
total_time	time.totalTime	total session seconds	Banholzer, Sun
time_on_page	time.timeOnPage	visible-page seconds	Banholzer
click_count	clicks.clickCount	total clicks in session	Banholzer
distance_travelled	mouseMovements	sum of Euclidean distances between consecutive points	Yamauchi, Aldrich, Khan
average_cursor_speed	mouseMovements	distance_travelled / movement_time	Monaro, Aldrich, Banholzer, Zantvoort
acceleration_mean	mouseMovements	mean change in segment speed over time	Monaro
direction_changes	mouseMovements	count sign changes in movement direction	Monaro
x_flip	mouseMovements	direction changes on x-axis	Monaro
y_flip	mouseMovements	direction changes on y-axis	Monaro
jitter	mouseMovements	relative excess of raw path length over a smoothed-path length (Khan-inspired jitter proxy)	Khan, Zantvoort
context_change_count	contextChange	total number of hidden / visible transitions in session	logger-derived; compatible with event structure
days_since_baseline	sessionDate	difference in days between current session and the first session for same participant	RQ1 longitudinal design / mixed-effects rationale
dass_depression_t	dass_linked / DASS subscales	depression score of the current session (t)	RQ1 design; DASS-21 session-level self-report
dass_anxiety_t	dass_linked / DASS subscales	anxiety score of the current session (t)	RQ1 wording; DASS-21 session-level self-report
dass_stress_t	dass_linked / DASS subscales	stress score of the current session (t)	RQ1 wording; DASS-21 session-level self-report
click_density	clicks.clickCount, time.totalTime	click_count / total_time	logger-derived; compatible with event structure
mouse_move_count	mouseMovements	total number of mouse movement events	logger-derived; compatible with event structure
mouse_dt_mean	mouseMovements	mean time between consecutive valid mouse-event intervals (dt)	logger-derived; Zantvoort and Monaro inspired
mouse_dt_max	mouseMovements	maximum time between consecutive valid mouse-event intervals (dt)	logger-derived; Zantvoort and Monaro inspired

APPENDIX D: Clinical Requirements Elicitation Interview Protocol

Forms Link:

https://docs.google.com/forms/d/e/1FAIpQLSeYgQbxDkdTTryAMmj2mF9KHvQE6PykXSroiVWMw9R4xuUMJw/viewform?usp=dialog

Excel from all interview results:

https://docs.google.com/spreadsheets/d/e/2PACX-1vQyR0GF3usmH2HRaugOB-WVu1tMJUR7s62sBLrqZHXvLLYksn60BAJLRjRY5Bp7PYGkNUGVCnUP7MRt/pub?output=xlsx

Introduction

The purpose of this interview is to understand your clinical context and gather perspectives on the potential usefulness of a clinical decision-support system for depression, developed within the scope of the master's dissertation.

The system under development explores the combination of self-report information, namely depression, anxiety, and stress dimensions assessed by DASS-21/DASS, with digital behavioral data collected during the completion of the RESS questionnaire, such as mouse use patterns, clicks, and interaction times. These data are analyzed through modeling and machine learning techniques. The potential outputs of the system may include a score or priority ranking and the signaling of potential cases to be clinically reviewed, accompanied by explanatory information about the factors considered by the model. The goal is not to replace clinical judgment or produce autonomous diagnoses, but to investigate ways of supporting triage processes, follow-up, and clinical contextualization.

The interview seeks to validate clinical needs, points of integration in the workflow, preferences regarding the type of output, interpretability requirements, and possible adoption limitations. The system is still in the design phase, so the information gathered will serve to guide its design and delimit its most appropriate use. There are no right or wrong answers; the focus is on understanding your practice and clinical perspective.

Since the answers are open-ended, it is natural that some questions may not fit perfectly with your work setting. If that is the case, you may simply skip the question or focus your answer on what reflects your experience and your general knowledge of the field.

Note: Although the predefined space is short, each text response has no word limit, so you may share your perspective with the level of detail you consider most appropriate.

PART A — Current Clinical Workflow

- What is your work setting (e.g., hospital care, clinic)?
- How many years of experience do you have?
- "In what types of cases do you most frequently work?"

• “Where does depressive/anxious symptomatology usually appear in those cases?”
• “How is the process carried out from the moment a patient arrives until the first intervention?”
• “Which instruments do you use to assess symptomatology? (e.g., DASS-21, PHQ-9) What time horizon do they assess (e.g., previous week)?”
• “How does that information influence clinical decisions?”
• “Are there formal prioritization criteria, or is it more based on clinical judgment?”
• “Do patients undergo periodic reassessments? How often?”
• “Is there a current system to monitor deterioration between consultations?”

PART B — Purpose & Functionality

• “At which points in your workflow would additional decision support make the most sense?”
• “Do you see more usefulness in:”
o (a) prioritizing new patients (initial triage)
o (b) monitoring follow-up patients
o (c) both
• “Why would that option be more useful in your context?”
• “What type of information would be most useful to you in practice?”
o Continuous score
o Categories
o Ranking
• “In which situations would that information lead to a concrete action?”
• “What level of risk would need to be reached to justify immediate intervention?”
• “What would you do with a case classified as ‘high risk’?”

PART C — Human Responsiveness & Integration

C1. Timing & Cadence

• “How often would you consult this system? (daily, weekly, on-demand?)”
• “Information would be more useful:”
o Immediately after the patient completes the DASS (real-time)
o In a daily review of the queue (batch)
o When planning appointments (on-demand)

C2. Responsiveness

• “What response time would be acceptable after identifying elevated risk?”
• “Who would be responsible for follow-up? (psychologist, administrative staff, automated reminder?)”
• “Would this fit naturally into your workflow, or would it require changes?”

C3. Alerts vs Dashboard

• “Would you prefer:”
o Passive dashboard (consult when needed)
o Notifications for high-risk cases (email, SMS)
o Hybrid system (dashboard + alerts for critical thresholds)

PART D — Trust & Explainability

D1. Trust in an Algorithmic System

• “What information would you need to trust a prioritization recommendation?”
• “Would it be useful to see which features/patterns contributed to the risk score?”

D2. Behavioral Features

• “Does it make clinical sense to use mouse/keyboard patterns? (hesitation, motor slowness)”
• “Are there digital behaviors that you would associate with depressive symptoms?”

PART E — Risk Trade-offs & Limitations

• “In a system like this, what do you consider more problematic:”
o False positives (signaling risk unnecessarily)
o False negatives (failing to detect at-risk cases)
• “Which of these errors would be more acceptable in your context?”

PART F — Adoption & Barriers

• “What could prevent the adoption of this type of system in practice?”
• “Are there concerns at the level of data, privacy, or infrastructure?”
• “What would make this system useful in day-to-day practice?”

References

Hopkin, G., Coole, H., Edelmann, F., Ayiku, L., Branson, R., Campbell, P., Cooper, S., & Salmon, M. (2025). Toward a New Conceptual Framework for Digital Mental Health Technologies: Scoping Review. JMIR Mental Health, 12, e63484. https://doi.org/10.2196/63484

Shulha, M., Hovdebo, J., D’Souza, V., Thibault, F., & Harmouche, R. (2024). Integrating Explainable Machine Learning in Clinical Decision Support Systems: Study Involving a Modified Design Thinking Approach. JMIR Formative Research, 8, e50475. https://doi.org/10.2196/50475

Trinkley, K. E., Kahn, M. G., Bennett, T. D., Glasgow, R. E., Haugen, H., Kao, D. P., Kroehl, M. E., Lin, C.-T., Malone, D. C., & Matlock, D. D. (2020). Integrating the Practical Robust Implementation and Sustainability Model With Best Practices in Clinical Decision Support Design: Implementation Science Approach. Journal of Medical Internet Research, 22(10), e19676. https://doi.org/10.2196/19676

Bayor, A. A., Li, J., Yang, I. A., & Varnfield, M. (2025). Designing Clinical Decision Support Systems (CDSS)—A User-Centered Lens of the Design Characteristics, Challenges, and Implications: Systematic Review. Journal of Medical Internet Research, 27, e63733. https://doi.org/10.2196/63733

APPENDIX E: Hyperparameter Grids for the RQ2 Nested Cross-Validation Pipelines

Table 13 Table D.1: Hyperparameter grids for the nested cross-validation pipelines.

Pipeline	*Algorithm*	*Hyperparameter*	*Grid values*
DASS-only	Random Forest	`n_estimators`	`[200, 500]`
DASS-only	Random Forest	`max_depth`	`[None, 4, 8, 12]`
DASS-only	Random Forest	`min_samples_split`	`[2, 5, 10]`
DASS-only	Random Forest	`min_samples_leaf`	`[1, 2, 4]`
DASS-only	Random Forest	`class_weight`	`[None, "balanced"]`
DASS-only	Random Forest	`max_features`	`[3]`
HCI_RESS-only / Hybrid	Random Forest	`selector__threshold`	`[0.7, 0.8]`
HCI_RESS-only / Hybrid	Random Forest	`n_estimators`	`[200, 500]`
HCI_RESS-only / Hybrid	Random Forest	`max_depth`	`[None, 4, 8, 12]`
HCI_RESS-only / Hybrid	Random Forest	`min_samples_split`	`[2, 5, 10]`
HCI_RESS-only / Hybrid	Random Forest	`min_samples_leaf`	`[1, 2, 4]`
HCI_RESS-only / Hybrid	Random Forest	`class_weight`	`[None, "balanced"]`
HCI_RESS-only / Hybrid	Random Forest	`max_features`	`[0.5, "sqrt"]`
HCI_RESS-only / Hybrid	XGBoost	`selector__threshold`	`[0.7, 0.8]`
HCI_RESS-only / Hybrid	XGBoost	`n_estimators`	`[200, 500]`
HCI_RESS-only / Hybrid	XGBoost	`max_depth`	`[3, 5, 7]`
HCI_RESS-only / Hybrid	XGBoost	`learning_rate`	`[0.05, 0.1]`
HCI_RESS-only / Hybrid	XGBoost	`subsample`	`[0.8, 1.0]`
HCI_RESS-only / Hybrid	XGBoost	`colsample_bytree`	`[0.8, 1.0]`
HCI_RESS-only / Hybrid	XGBoost	`min_child_weight`	`[1, 3]`
HCI_RESS-only / Hybrid	XGBoost	`gamma`	`[0, 0.5]`
HCI_RESS-only / Hybrid	XGBoost	`reg_lambda`	`[1, 5`

APPENDIX F: Clinical Requirements Synthesis and Design Requirements Matrix

Table 14 Table E.1: Clinical requirements synthesis from qualitative interviews

D — Human Responsiveness & Integration	Momento de consulta do sistema	Consulta no dia anterior à consulta ou automático; revisão semanal/quinzen houver alertas.
D — Human Responsiveness & Integration	Resposta a high-risk e workflow	Ideação suicida: avaliação no próprio máximo 24h; mudança de workflow s positiva.
D — Human Responsiveness & Integration	Formato do Sistema	Valorizou a visualização explicativa/S considerando o gráfico visual útil, e re explicação textual pode ajudar a interp output; mostrou abertura a alterações workflow desde que organizadas e úte
E — Trust & Explainability	Informação necessária para confiar	Confiança construída com desempenh observado e confirmação clínica contí
E — Trust & Explainability	Utilidade de fatores explicativos / XAI	Considerou gráfico SHAP 'excelente'; t apoiar leitura; separar informação do contexto.
E — Trust & Explainability	Features comportamentais e interpretação clínica	Lentificação digital plausível, mas iden familiarity bias: clínicos podem respon depressa por repetição/saturação/des
F — Risk Trade-offs & Limitations	Falsos positivos vs. falsos negativos	Prefere claramente falsos positivos: fa negativo pode deixar doente sem cuid
G — Barriers & Adoption	Limitações clínicas/operacionais	Adesão do paciente e terapeuta; otimi terapeuta; Necessidade de adesão; rej comunicar diretamente com paciente; abertas só se pré-definidas.
G — Barriers & Adoption	Condições para utilidade diária	Briefing compacto pré-consulta e evol eficiência sem sobrecarga.

Forms Domain	Subtheme	E1 — Dr. Paulo	E2 — Dr. Joana (colleague)	E3 — Forms 20/05	E4 — Forms 29/05	Cross-sectional synthesis

A — Professional profile and clinical context	Practice context and population	Hospital context; follow-up of selected clinical population; articulation with psychiatry.	Independent/private team practice; mostly adults; triage to professionals with different specialties.	Private clinic; adults; frequent depressive/anxious symptoms.	Hospital care and private clinic; addictive, personality, and mood disorders; frequent depressive/anxious symptomatology.	Contexts vary, but all deal with relevant symptomatology; system utility depends on the clinical environment.
B — Current clinical process and triage	Entry, assessment and instruments	Internal/psychiatry referral → consultation → assessment; DASS-21, RES and WAI; reassessment at each consultation.	Initial form and manual referral; applies instruments between the 2nd and 4th consultation; occasional reassessments.	Contact with clinic and referral by availability/complaints; instruments and history as guidance.	First psychiatry consultation and subsequent referral; mainly clinical interview; occasional tests.	Triage is currently human and contextual; instruments support decisions, but are not used in isolation.
B — Current clinical process and triage	Monitoring/reassessments	Systematic fortnightly or monthly assessment (at each consultation); values dimensional evolution throughout treatment.	Values periodic reassessment and already monitors evolution; utility at specific moments.	Assessment considered constant during the therapeutic process.	Reassessments at distinct phases in addiction treatment programs; no current deterioration system.	There is cross-sectional interest in evolution, although frequency and applicability vary.
C — Purpose & Functionality	Main use of the system	Prioritize urgency, but above all support monitoring and identification of therapeutic targets ('precision psychotherapy').	Triage/referral of the professional and occasional monitoring; organization by urgency and day management.	Initial triage; support for more conscious referral of the professional.	Prioritize new patients in the initial phase and decide referrals/specialties; recognizes limitation of initial adherence to tests.	Triage/prioritization is a common use case, although monitoring also receives positive validation; Dr. Paulo adds dimensional therapeutic monitoring.
C — Purpose & Functionality	Output format	Prefers dimensional categories (mood/emotional regulation, etc.) and evolution; ranking useful to compare priority between patients.	Ranking/severity accompanied by context(sleep, suicidal ideation); summary with option to go deeper.	Prefers categories.	Does not close answer: depends on the case.	Priority and contextual information seem more acceptable than an isolated score.
D — Human Responsiveness & Integration	Moment of consultation of the system	Consultation on the day before the consultation or automatic alert; weekly/biweekly review if there are no alerts.	On-demand when request arrives or before consultation; critical alerts.	Daily / when planning scheduling; on-demand.	Immediately after filling; prefers notification for high-risk.	The system must support on-demand use and critical alerts, adapted to the workflow.
D — Human Responsiveness & Integration	Response to high-risk and workflow	Suicidal ideation: evaluation on the day itself or at most 24h; workflow change would be positive.	Alerts for critical risk; therapist availability and relational limits need to be protected.	Multidisciplinary team intervention; system would require changes.	Psychiatry/coordinator/meeting/supervision; follow-up by the psychologist on the day itself; would require significant changes.	High-risk requires fast human referral; adoption implies time and responsibility management.
D — Human Responsiveness & Integration	System Format	Valued the explanatory/SHAP visualization, considering the visual graph useful, and mentioned that the textual explanation can help interpret the output; showed openness to workflow changes as long as organized and useful.	Preference for a mixed model with centralized dashboard and alerts/notifications for critical cases, keeping contact and decision dependent on the availability and judgment of the psychologist.	The Forms point to preference for a hybrid system (dashboard + alerts for critical thresholds)	Notifications for high-risk cases (email, SMS)	Develop a clinician-facing interface with prioritization dashboard, model score, visual explanation of the factors, short narrative and additional context; avoid direct automatic communication to the patient
E — Trust & Explainability	Information needed to trust	Trust built with observed performance and continuous clinical confirmation.	Values context, essential information and evolution; questionnaires do not substitute clinical understanding.	Trust depends on the case/degree of risk;	Requires very concrete risk criteria.	Trust depends on transparency, understandable criteria and must be accompanied by contextual info.
E — Trust & Explainability	Utility of explanatory factors / XAI	Considered SHAP graph 'excellent'; text can support reading; separate model information and context.	Values score/ranking + behavioral and qualitative signals; highlighted changes/incoherence in the responses.	Considers it useful to see indicators; accepts mouse patterns.	Useful to see features if succinctly; no opinion on mouse patterns.	There is support for compact explanations; the interpretation of passive signals needs caution.
E — Trust & Explainability	Behavioral features and clinical interpretation	Digital slowing plausible, but identifies familiarity bias: clinicians may answer faster due to repetition/saturation/discomfort.	Changes in responses and possible minimization are relevant; interest in erased/rewritten responses.	Finds it clinically relevant to use mouse patterns.	Cannot answer regarding mouse patterns.	Passive signals are promising, but should not be converted into causal or diagnostic statements.
F — Risk Trade-offs & Limitations	False positives vs. false negatives	Clearly prefers false positives: false negative can leave patient without care.	Prefers false positives; flagged case can be clarified clinically.	Considers false negatives more problematic.	Considers false negatives more problematic, because the added value is detecting risk.	Total convergence in favor of higher sensitivity.
G — Barriers & Adoption	Clinical/operational limitations	Adherence of the patient and therapist; therapist optimism; Need for adherence; rejects AI communicating directly with patient; open questions only if pre-defined.	Risk of repeated use for attention seeking; GDPR and unauthorized access; security/authentication; gradual adoption.	GDPR and human resources time.	Some patients are not receptive to tests in the initial phase; time and resources; does not miss the system in own practice.	There may be risks of adherence, context and operational load. Privacy and workload are recurring barriers; autonomous AI for patient is not accepted in the available clinical feedback.
G — Barriers & Adoption	Conditions for daily utility	Compact pre-consultation briefing and visual evolution; efficiency without overload.	Utility is more in referral within the team and knowing how bad the person is.	More intuitive, fast and complete process; optimization of HR.	Does not identify need in current practice; system would only have value if it did not create high load.	Utility may depend on providing brief, actionable information adapted to the context.

Table 15 Table E.2: Consolidated design requirements matrix.

Forms Domain	***Consolidated requirement***	***Operationalization in the mockup***	***Current decision***
A — Professional profile and clinical context	Adaptability to different clinical contexts	Language supporting triage/clinical review, without assuming automatic diagnosis.	Implemented in the framing / partially represented in the mockup
B — Current clinical process and triage	Consider longitudinal evolution	Not implemented; scope limited to the first digital assessment.	Future work / outside the current scope
C — Purpose &	Present signaling/priority	Priority score and label	Implemented

Functionality	instead of diagnosis	"Above threshold" / "Below threshold".	
C — Purpose & Functionality	Priority-ordered queue	Initial prioritization queue ordered by priority score.	Implemented
D — Human Responsiveness & Integration	Support human review and eventual follow-up	Actions "View detail", "Consult responses" and "Mark for review".	Partially represented / future operational integration
E — Trust & Explainability	Succinct visual explanation	Top-5 contributing factors and aggregation of "other factors".	Implemented
E — Trust & Explainability	Complementary explanatory text	Short narrative with interpretation of the signaling and final caution.	Implemented
E — Trust & Explainability	Separate self-report, HCI and additional context	Separate blocks for self-report, interaction, model explanation and additional context.	Implemented
E — Trust & Explainability	Avoid overinterpretation of HCI	Warnings supporting clinical review, without automatic diagnosis or causal inference.	Implemented
F — Risk Trade-offs & Limitations	Prioritize sensitivity in signaling	Operational threshold oriented toward recall/F2 and signaling for human review.	Implemented
G — Barriers & Adoption	Summary first, detail later	Overview with metrics/queue; detail view with self-report, explanation and context.	Implemented
G — Barriers & Adoption	Supervision, privacy and operational burden	Human supervision made explicit; privacy/operational burden discussed in the evaluation.	Partially represented / discuss in the evaluation and limitations
Additional emergent dimension	Incoherence or discrepancy between self-report and interaction patterns	Contextual indicator "possible discrepancy to explore".	Contextual placeholder / future work

APPENDIX G: Formative Sessions Thematic Analysis

Protocol Synthesis-P1:

0. Initial preamble / framing

Element	Extraction
Quote P1	"Yes, yes, of course."
Paraphrase	P1 authorizes recording/annotation.
Quote P1	"I think not."
Paraphrase	After the preamble, P1 indicates that she has no relevant doubts about the initial assumptions.
Initial codes	Consent; general understanding of the assumptions; absence of initial doubts; delimitation of accepted non-diagnostic use.
Artifact	Both, but especially the general framing of B/RQ3 and C/RQ4.
Cross-synthesis	Only P1. The evidence suggests that the initial framing was accepted without explicit objection, but this does not mean that all concepts were understood, because doubts arise later about the "priority score."

---

1. Task 1 — Prioritization dashboard

1.1 First impression of the dashboard

Element	Extraction
Quote P1	"It seems very intuitive."
Paraphrase	The overall first impression of the dashboard is positive.
Quote P1	"The graphs, perhaps the subdued colors."
Paraphrase	What first draws attention are the visual elements, especially the graphs and color palette.
Initial codes	Positive first impression; visual attractiveness; prominent graphs; subdued colors; initial visual legibility.
Artifact	Artifact C/RQ4.
Cross-synthesis	There is only P1. The evidence is positive for the dashboard's initial usability, especially at the visual level, but it does not yet test deep conceptual understanding.

1.2 Clarity of the ranking

Element	Extraction
Quote P1	“I think it is fine. I think it is easy to understand.”
Paraphrase	P1 considers the ranking understandable.
Quote P1	“I also like the font.”
Paraphrase	The typography is evaluated positively.
Quote P1	“Yes, some words, I would like them to be a little bigger, but everything was fine.”
Paraphrase	Legibility is acceptable, although some words could be larger.
Initial codes	Ranking understandable; appropriate typography; need for larger text; overall acceptable legibility.
Artifact	Artifact C/RQ4.
Cross-synthesis	P1 validates the organization of the ranking, but points to a minor accessibility improvement: slightly increase the size of some words.

1.3 Understanding the “priority score”

Element	Extraction
Quote P1	“Priority score of the model. What is that? What does that mean?”
Paraphrase	The term “priority score” is not immediately clear.
Quote P1	“Yes, but what does score mean, what does priority score mean?”
Paraphrase	Even after an initial explanation, P1 continues to ask for clarification about the meaning of the score.
Quote P1	“I understand, but I do not know what priority is, what is priority here for you?”
Paraphrase	The problem is not only “score,” but the operational concept of “priority.”
Quote P1	“Maybe I would put some information here, an IA or something like that, and the person would click and read.”
Paraphrase	P1 suggests an informational icon/tooltip for the score.
Initial codes	Terminological ambiguity; need for operational definition; informational tooltip/icon; on-demand contextual explanation; risk of insufficient interpretation of the score.
Artifact	Mainly Artifact C/RQ4, with a link to Artifact B/RQ3 because interpretation of

Element	Extraction
	the score depends on explanation of the model.
Cross-synthesis	The dashboard is visually intuitive, but the central concept of "priority score" needs explicit explanation in the interface. P1's suggestion points to a concrete improvement: an information icon next to the score/term "priority."

1.4 Case that would be opened first / initial action

Element	Extraction
Quote P1	"And now I can click here for detail."
Paraphrase	P1 understands the possibility of opening the case detail.
Initial codes	Navigation to detail; action affordance; dashboard → individual case flow.
Artifact	Artifact C/RQ4.
Cross-synthesis	The action of opening the detail seems understandable. However, the transcript does not contain an explicit justification of the type "I would open case X because...," so the evidence is limited for active prioritization.

1.5 Distribution by signaling

Element	Extraction
Quote P1	No clear direct evidence.
Paraphrase	The participant does not explicitly comment on whether the distribution by signaling helps to interpret the queue.
Initial codes	Absent evidence; distribution not directly assessed.
Artifact	Artifact C/RQ4.
Cross-synthesis	There is not enough basis to conclude whether this specific component helped or did not help interpretation of the queue.

---

2. Task 2 — High-priority convergent case

2.1 Relationship between self-reports, score, and clinical interpretation

Element	Extraction
Quote P1	"I think, I think yes, I think it is intuitive once again."
Paraphrase	The self-report dimensions seem understandable/intuitive.
Quote P1	"It seems to me that, very probably, it will be a person, of course, with

Element	Extraction
	depression or minor or more or I do not know, or with symptoms, already at least very evident."
Paraphrase	P1 interprets the case as clinically relevant, with evident depressive symptoms.
Quote P1	"What is that very useful, without a shadow of a doubt, I would apply other scales, even diagnosis. To see whether it is confirmed or not."
Paraphrase	The information is considered useful to trigger additional assessment, not as a final diagnosis.
Initial codes	Self-reports intuitive; usefulness for triage; strong clinical signal; support for decision to reassess; need for diagnostic confirmation.
Artifact	Both: C/RQ4 for the clinical usefulness of the detail view; B/RQ3 for the relation to the score explanation.
Cross-synthesis	P1 sees clinical value in the convergent case: the information supports the decision to carry out complementary assessment. The utility appears as a trigger for clinical review, not as a substitute for professional judgment.

2.2 Explanation of the factors / SHAP

Element	Extraction
Quote P1	"I like this; I do not know whether this is very intuitive here for everyone, for all professionals."
Paraphrase	P1 personally understands/likes the section, but questions whether all professionals would understand it.
Quote P1	"Well, I know, but I do not know whether..."
Paraphrase	P1 recognizes the technical meaning, but anticipates difficulties for other users.
Initial codes	Potentially useful explanation; comprehensibility depends on literacy; risk of low intuitiveness for professionals; need for simplification.
Artifact	Artifact B/RQ3.
Cross-synthesis	The factor explanation is not rejected, but is seen as potentially demanding. This suggests that the SHAP/factors layer should be retained, but accompanied by simpler language and interpretive support.

2.3 Textual narrative / summary interpretation

Element	Extraction
Quote P1	"This interpretation is more for you, right? Or can it also be [...] for the professional?"

Element	Extraction
Paraphrase	P1 questions whether the summary interpretation is intended for the researchers or the clinician.
Quote P1	“Well, I do not know if it is simple enough to understand for the professional.”
Paraphrase	The narrative may not be sufficiently simple for clinical use.
Quote P1	“For example, ‘no factors stood out among the top 5 contributions.’ I know what that is, but it is not very intuitive.”
Paraphrase	A concrete formulation of the narrative is considered not very intuitive.
Initial codes	Ambiguous target audience; overly technical narrative; need for simple clinical language; unnatural explanation; textual reformulation.
Artifact	Artifact B/RQ3.
Cross-synthesis	The textual narrative is a critical area for improvement. P1 does not consider it useless, but indicates that it should be written explicitly for the professional, with more natural phrases and less dependence on terms such as “top 5 contributions.”

2.4 Distinction between data used by the model and additional context

Element	Extraction
Quote P1	“Exactly, that makes sense, yes.”
Paraphrase	P1 validates the idea of flagging inconsistencies/additional context.
Quote P1	“For example, here the rumination 26 a. It can be put as they put it above, for example, 38/42, 30/42. It can be put here too.”
Paraphrase	P1 suggests presenting values with denominator/max scale also in the subdimensions, to facilitate interpretation.
Quote P1	“Another thing [...] there are these blue colors here. In the previous case, it was red and then also green. Maybe it is important to add a description of these colors.”
Paraphrase	P1 suggests a legend for the colors used in the categories/levels.
Quote P1	“It appears a little detached and I fear that the health professional will not be able to understand THIS.”
Paraphrase	Certain elements seem isolated/decontextualized and may not be understood by professionals.
Initial codes	Additional context useful; need for denominators; color legend; risk of detached information; need for visual/semantic integration.

Element	Extraction
Artifact	Both: B/RQ3 if the distinction helps interpret the explanation; C/RQ4 for the organization of the detail view.
Cross-synthesis	P1 supports the inclusion of additional context, but asks for greater visual explicitness: full scales, color legends, and better integration of the elements. This reinforces the need to clearly separate and explain what is model output, what are self-reports, and what is complementary context.

---

3. Task 3 — Discordant case / passive signal dominant

The protocol planned to test a case in which the model assigns high priority despite less expressive self-reports, with the explanation based mainly on mouse interaction patterns.

3.1 Usefulness or doubts regarding discordant signaling

Element	Extraction
Quote P1	“If it appeared to me, if I read this report, what I would do would be, once again, apply actual diagnostic assessment tests.”
Paraphrase	The signaling would lead to additional assessment, but would not be accepted in isolation.
Quote P1	“I think it is a bit strange.”
Paraphrase	The discordant case generates strangeness/doubt.
Initial codes	Signaling useful as an alert; clinical doubt; need for confirmation; non-automatic acceptance; caution toward discordance.
Artifact	Both: B/RQ3 for the explanation of the signaling; C/RQ4 for usefulness in the triage flow.
Cross-synthesis	P1 does not reject the alert, but interprets it cautiously. The system’s usefulness lies in signaling cases for reassessment, not in replacing clinical assessment.

3.2 Interpretation of mouse interaction patterns

Element	Extraction
Quote P1	“It is not related to the mouse, it only indicates one thing to me: that the person has no experience using the mouse.”
Paraphrase	P1 interprets the mouse patterns as possibly explained by lack of mouse experience, not necessarily by depression.
Quote P1	“I noticed that right at the beginning [...] people who had no experience with mice [...] are very slow.”

Element	Extraction
Paraphrase	The participant brings empirical collection experience: some participants were slow because they were not familiar with the mouse.
Quote P1	“Whoever uses this tool in the future has to indicate their experience with the mouse.”
Paraphrase	P1 recommends collecting/marking mouse-use experience.
Quote P1	“Digital literacy is different [...] it is very different from mouse use.”
Paraphrase	P1 distinguishes general digital literacy from the specific skill of using a mouse.
Initial codes	Motor/digital skill confounder; mouse experience; validity of passive signals; need for a control variable; distinction between digital literacy and mouse use; interpretive caution.
Artifact	Mainly Artifact B/RQ3, because it affects the interpretability and trust in the passive factors; also C/RQ4 if the dashboard needs to display this context.
Cross-synthesis	This is perhaps the strongest finding of the session: mouse-based passive signals can be clinically difficult to interpret without information about mouse familiarity. For P1, these signals may reflect technological/motor limitation, and not necessarily depressive symptomatology.

3.3 Clarity that the system is not diagnostic

Element	Extraction
Quote P1	“Since THIS is not for diagnosis, right, it is not a diagnosis.”
Paraphrase	P1 understands that the tool is not diagnostic.
Quote P1	“I would apply scales that would serve as support for my diagnosis, of course, and clinical interview, etcetera.”
Paraphrase	P1 frames the tool as preliminary support before diagnostic instruments and clinical interview.
Initial codes	Clear non-diagnostic boundary; auxiliary role; confirmation by scales/interview; preserved clinical judgment.
Artifact	Mainly Artifact C/RQ4, with a link to clinical safety.
Cross-synthesis	The participant seems to understand well the system’s positioning as triage/prioritization. This supports the appropriateness of the non-diagnostic framing, although the interface should continue to reinforce it.

---

4. Final questionnaire

Element	Extraction
Quote P1	“What is the code?”
Paraphrase	P1 asks for operational clarification to fill in the form.
Quote P1	“Profile, professional profile, what should I put?”
Paraphrase	P1 asks for help classifying her professional profile.
Quote P1	“What does unnecessarily complex mean?”
Paraphrase	A SUS item generates semantic doubt.
Quote P1	“I think I already gave it, right? In other words, the changes. I think I just said some, I already did, I already gave, I already gave the suggestion.”
Paraphrase	The participant considers that the open suggestions were already given orally.
Quote P1	“Congratulations.”
Paraphrase	Positive closing of the session.
Initial codes	Doubts in the form; ambiguity of SUS item; suggestions already verbalized; overall positive evaluation.
Artifact	Explainability items → B/RQ3; SUS/usability → C/RQ4.
Cross-synthesis	Without the numerical values of the form, it is only possible to code comments during completion. The doubt about “unnecessarily complex” suggests that some standardized items may require clarification, but this says more about the form than about the interface.

Overall synthesis by artifact

Artifact C / RQ4 — Dashboard, prioritization, usability and clinical utility
P1’s evidence is globally favorable toward the dashboard: the interface is described as “very intuitive,” the ranking is considered “easy to understand,” and the graphs/subdued colors attract attention positively. The participant also understands the action of opening the detail. However, there are concrete clarity issues: “priority score” is not self-explanatory, some words could be larger, the colors need a legend, and certain elements appear “detached.” Clinical utility is framed as support for triage and the decision to reassess, not as diagnosis.

Artifact B / RQ3 — Explanation, factors, narrative and clinical action
The explanation by factors is potentially useful, but the participant anticipates difficulties for professionals who are less familiar with it. The textual narrative needs to be more clearly directed at the clinician and use less technical language. The phrase about the “5 main contributions” is explicitly identified as not intuitive. The strongest finding is the caution with mouse-based passive signals: P1 considers that they may reflect lack of experience using the mouse, requiring metadata/context for safe interpretation.

Potential emerging themes separated from the protocol-organized findings

1. Strong initial visual usability, but central concepts need embedded explanation
   The dashboard seems intuitive, but the term "priority score" needs a tooltip/information icon.
2. Explainability must be translated into simple clinical language
   SHAP/factors may be useful, but the narrative needs to avoid technical or ambiguous formulations.
3. Mouse-based passive signals require control of confounders
   Experience using the mouse, technological difficulties, and collection context should be collected or made visible.
4. The tool is perceived as triage, not diagnosis
   P1 would use the information to decide to apply additional scales and carry out a clinical interview.
5. Legends, complete scales, and context improve interpretive safety
   Color legends, "x/y" values, and local explanations can reduce misinterpretations.

Protocol Synthesis-P2:

1. Initial preamble / framing

Element	Extraction
Quote P2	"One thing is for me to be in a sadder and more depressed phase, another thing is for me to have a depressive presentation."
Paraphrase	P2 distinguishes transient symptomatology or self-report from a more structured clinical depressive assessment.
Quote P2	"If I assess a person who does not have [...] but is going through [...] a depressive moment, even a relatively intense one, the questionnaire score will be higher."
Paraphrase	P2 recognizes that questionnaires may reflect recent states and not necessarily a stable clinical condition.
Quote P2	"Patients do not have access to this, they only have access to the questionnaire, right?"
Paraphrase	P2 seeks confirmation that the dashboard is only for the professional, not for the patient.
Quote P2	"Because sometimes results frighten people."
Paraphrase	P2 warns that showing results directly to the patient may have negative effects or cause anxiety.

Element	Extraction
Quote P2	“I just wanted to understand whether it is only for adults over 18 years old...”
Paraphrase	P2 identifies the need to clearly delimit the target population of the system.
Initial codes	Distinction between momentary state vs. clinical condition; limitations of self-reports; safety in presenting results; dashboard only for professionals; target population; scope delimitation.
Artifact	Both. Mainly Artifact C/RQ4, because it affects context of use, safety, and target audience; also Artifact B/RQ3, because it conditions interpretation of the score/model.
P2 synthesis	The preamble seems to have been understood, but P2 adds important nuances: self-reports may capture recent states, results should not be shown directly to the patient, and the target population should be explicit.

---

1. Task 1 — Prioritization dashboard

1.1 First impression of the home page / queue

Element	Extraction
Quote P2	“Yes, it seems fine to me.”
Paraphrase	In the recap of the home page/queue, P2 globally validates the dashboard.
Additional evidence	The facilitator summarizes that “the home page”, “the queue” and “the graphs” seemed “ok”, and P2 confirms.
Initial codes	Positive overall assessment; understandable queue; acceptable graphs; absence of strong criticism of the initial view.
Artifact	Artifact C/RQ4.
P2 synthesis	The evidence regarding the initial view is positive but limited, because the initial exploration of the dashboard is not as detailed as in the first interview. The safest conclusion is that P2 did not identify relevant problems in the queue or graphs, but there is no rich verbalization about ranking, distribution, or first visual attention.

1.2 Clarity of the ranking

Element	Extraction
Quote P2	There is no direct and developed quote about the ranking itself.
Paraphrase	The participant appears to accept the queue/ranking during the recap, but does not explain in detail how she interprets the ordering.

Element	Extraction
Initial codes	Weak evidence; ranking apparently acceptable; absence of explicit problem.
Artifact	Artifact C/RQ4.
P2 synthesis	It should not be strongly claimed that the ranking was clearly understood, only that no explicit criticism of the queue emerged.

1.3 Meaning of score/threshold

Element	Extraction
Quote P2	“Very close to the threshold [...] 35%, but the model says it has a score of 30%.” / P2 replies: “Right.”
Paraphrase	P2 follows the explanation regarding the case below but close to the threshold.
Quote P2	“Hence this threshold being 30%. It is lower.”
Paraphrase	P2 interprets the lower score as consistent with less rumination, greater relaxation, and lower emotional suppression.
Initial codes	Threshold understood in context; score interpreted comparatively; score–factor relationship; borderline case.
Artifact	Artifact C/RQ4, linked to Artifact B/RQ3 through the explanation of factors that reduce the score.
P2 synthesis	Unlike the first participant, P2 does not appear blocked by the term “priority score.” She is able to follow a near-threshold case when it is explained through the associated factors and self-reports.

1.4 Case that would be opened first / initial action

Element	Extraction
Quote P2	There is no explicit response of the type “I would open this case first because...”.
Paraphrase	The session advances to specific cases guided by the facilitator rather than participant-led selection.
Initial codes	Missing evidence; case selection not directly tested.
Artifact	Artifact C/RQ4.
P2 synthesis	There is insufficient evidence to conclude whether the dashboard supports autonomous selection of the first case to open.

1.5 Distribution by signaling

Element	Extraction
Quote P2	No direct evidence.
Paraphrase	The participant does not explicitly comment on distribution by signaling.
Initial codes	Missing evidence; component not directly evaluated.
Artifact	Artifact C/RQ4.
P2 synthesis	It is not possible to assess from this transcript whether distribution by signaling helped interpret the queue.

---

2. Task 2 — High-priority convergent case

2.1 Relationship between self-report, score, and explanation

Element	Extraction
Quote P2	“Yes, I even think it makes sense.”
Paraphrase	P2 considers that the organization of the section is appropriate.
Quote P2	“Having this, having the description of the participant here, right? The clinical summary...”
Paraphrase	P2 values the presence of a description/clinical summary of the participant.
Quote P2	“I think visually it is easier for the person to understand.”
Paraphrase	The visual presentation facilitates understanding.
Quote P2	“If everything were written, it would be more boring, I would say.”
Paraphrase	P2 prefers a visual/structured representation over a purely textual description.
Initial codes	Clear organization; useful clinical summary; visualization facilitates interpretation; preference for structured visuals; self-report + explanation articulated.
Artifact	Both: Artifact C/RQ4 by the detail view.
P2 synthesis	P2 considers that the relationship between clinical summary, assessment, and explanation is generally understandable. The visual presentation seems to increase legibility and avoid the clinical information becoming too textual.

2.2 Explanation of the factors

Element	Extraction

Element	Extraction
Quote P2	“And then here the explanation. It seems ok to me.”
Paraphrase	P2 considers the explanation section acceptable, but the assessment is brief and not very developed.
Quote P2	“Well, here it already added anxiety.”
Paraphrase	In the borderline case, P2 notices that the model explanation changed and that anxiety entered into the factors presented.
Quote P2	“Yes, it seems a lot to me.”
Paraphrase	After the explanation that some factors reduced the risk and that only the top-5 appears, P2 seems to accept the explanatory logic.
Initial codes	Explanation of the model acceptable; top-5 understood with facilitator support; comparison between cases; factors that increase/reduce score; positive but not spontaneous evidence.
Artifact	Artifact B/RQ3.
P2 synthesis	Direct evidence about the model explanation is moderately positive, but limited. P2 does not make a long analysis of the SHAP/factor explanation in the convergent case; she only says that the explanation “seems ok.” Understanding becomes more observable in the borderline case, when P2 identifies that anxiety entered the explanation and follows the idea of factors that reduce risk.

2.3 Textual narrative / summary interpretation

Element	Extraction
Quote P2	“Yes, yes, yes.” / “Yes.” in response to the idea that the summary interpretation helped.
Paraphrase	P2 confirms that the summary interpretation helps, although the evidence is partly induced by the facilitator’s wording.
Quote P2	“For someone who is not in the machine learning field [...] it may be less intuitive.” / P2 responds: “That’s it.”
Paraphrase	P2 agrees that machine learning elements may be less intuitive for professionals without that knowledge.
Initial codes	Narrative useful; dependence on technical literacy; need for translation for clinicians; explanation should reduce ML barrier.
Artifact	Artifact B/RQ3.
P2	The textual narrative seems useful, but it should be oriented to professionals without machine learning training. The evidence is positive, but not as

Element	Extraction
synthesis	spontaneous as in other points, because part of the assessment comes from confirmation of a question by the facilitator.

2.4 Distinction between information used by the model and additional context

Element	Extraction
Quote P2	“I think it is interesting that it has these [...] factors, right? The assessment...”
Paraphrase	P2 considers it useful to have the factors/assessment presented separately.
Quote P2	“It seems ok to me.”
Paraphrase	The separation between explanation and the remaining elements seems acceptable.
Quote P2	“It is to understand whether it is actually something consistent in the assessment or whether it is something that the person feels in that question but not in the other.”
Paraphrase	P2 validates the usefulness of exploring redundancy/inconsistency in similar items.
Initial codes	Acceptable separation; internal coherence of the questionnaires; inconsistency as clinical data; contextualized self-report; useful clinical summary.
Artifact	Both: Artifact C/RQ4 by the organization of the detail view.
P2 synthesis	P2 considers that the general distinction between self-report, clinical summary, and explanation makes sense. She also validates the idea of using inconsistencies between redundant items as clinically relevant contextual information.

2.5 Scales, colors, and meaning of the subdimensions

Element	Extraction
Quote P2	“I do not know whether having rumination very high is good.”
Paraphrase	P2 needs to know whether high values are positive or negative in each subscale.
Quote P2	“Assuming rumination is higher [...] this would be [...] red.”
Paraphrase	P2 suggests color coding according to the clinical direction of the subscale.
Quote P2	“Suppression [...] in principle is also not good [...] here it would also be red.”
Paraphrase	High emotional suppression would be interpreted as a negative indicator.
Quote P2	“Relaxation [...] maybe it would be [orange] [...] And engagement would be green.”

Element	Extraction
Paraphrase	P2 proposes differentiated color semantics by construct.
Quote P2	“I already know that it is easier to have the scale and the meaning right there. Yes. Instead of having to consult it separately.”
Paraphrase	The interface should show the scale and the meaning, without requiring external consultation.
Quote P2	“Or by colors.”
Paraphrase	Colors may function as interpretive support.
Initial codes	Direction of scales; color legend/semantics; high values do not always equal risk; need for local meaning; reduction of cognitive load; clinical interpretation by construct.
Artifact	Mainly Artifact C/RQ4.
P2 synthesis	This is a strong finding: it is not enough to show subscales; it is necessary to make explicit what a high/low value means in each construct. The interface should include scale, clinical direction, and/or interpretive colors.

---

3. Task 3 — Borderline / below but close to the threshold

The guide planned a case below but close to the threshold, asking whether the proximity to the threshold is relevant, whether the explanation helps, and whether the case should continue to be followed.

3.1 Interpretation of the case below the threshold

Element	Extraction
Quote P2	“Here the rumination is slightly lower than the previous one, so it is a person who ruminates less.”
Paraphrase	P2 compares the case with the previous one and interprets lower rumination as a favorable factor.
Quote P2	“Relaxation is higher, so it is a person who has a greater ability to relax.”
Paraphrase	Greater relaxation is interpreted as a protective indicator.
Quote P2	“Suppresses emotion less.”
Paraphrase	Lower emotional suppression is interpreted as a favorable factor.
Quote P2	“Hence this threshold being 30%. It is lower.”

Element	Extraction
Paraphrase	P2 relates the factors to the score below the threshold.
Initial codes	Comparative interpretation; protective factors; explanation of lower score; clinical reasoning from subscales.
Artifact	Both: Artifact B/RQ3 for the explanation of the score; Artifact C/RQ4 for the borderline case in the dashboard.
P2 synthesis	P2 is able to interpret the case below the threshold when the explanation and subscales are presented comparatively. This suggests that the interface can support understanding of why a case falls below the threshold, as long as the constructs are clear.

3.2 Additional information for triage and risk

Element	Extraction
Quote P2	“Suicidal ideation could be divided into suicide attempts and suicidal thoughts, because those are different things.”
Paraphrase	P2 considers it important to distinguish suicidal thoughts from attempts.
Quote P2	“At the level of food intake, maybe it can be a factor.”
Paraphrase	Diet is suggested as an additional clinical variable.
Quote P2	“Physical activity [...] is very important; exercise is very important in people with depression.”
Paraphrase	Physical activity is seen as relevant for assessment and prognosis/intervention.
Quote P2	“A person [...] has a good capacity for emotional regulation, has good sleep [...] it is much easier for us to help this person...”
Paraphrase	Sleep and emotional regulation are seen as resources that influence intervention and recovery.
Quote P2	“Yes. Even if triage later, ideally it would be an interview to understand effectively [...] the anamnesis, what we call anamnesis, which is the clinical interview.”
Paraphrase	Triage should be complemented by a clinical interview/anamnesis.
Initial codes	Suicidal ideation; attempts vs thoughts; sleep; diet; physical activity; emotional regulation; anamnesis; complementary assessment; resources/prognosis.
Artifact	Mainly Artifact C/RQ4, because it refers to additional context and clinical action; secondarily Artifact B/RQ3 if these data help calibrate model interpretation.
P2	P2 considers the dashboard more useful and safer if it integrates additional

Element	Extraction
synthesis	clinical variables, especially suicidality, sleep, diet, physical activity, and emotional regulation. The tool is seen as triage support, but the clinical interview remains essential.

---

4. Task 3 — Discordant case / passive signal dominant

4.1 Usefulness or doubts about the discordant signal

Element	Extraction
Quote P2	“But she did not report it.”
Paraphrase	P2 immediately identifies the discrepancy between score/model and self-report.
Quote P2	“She does not rate depression, nor anxiety, nor stress. Normal. But she is above the [threshold] and a fairly high [threshold].”
Paraphrase	The discordance is clinically salient: normal self-reports but a high score.
Quote P2	“Even people who do not identify the depressive symptom can be identified as having depressive symptoms and [...] they need clinical assessment.”
Paraphrase	P2 understands the potential usefulness of the passive signal to detect cases not captured by self-report.
Quote P2	“Because I do not understand [...] the mouse questions.”
Paraphrase	Despite recognizing the utility, P2 has low interpretive confidence in the mouse factors.
Quote P2	“So it has 94% but no symptoms. Then maybe it would make me question whether I would actually call the person to schedule the assessment or not.”
Paraphrase	A very high score without self-reported symptoms would generate doubt about the clinical action to take.
Initial codes	Model–self-report discordance; potential usefulness of passive signal; doubt in the face of a high score; low literacy about mouse data; need for contextual support; uncertainty in clinical action.
Artifact	Both: Artifact B/RQ3 for interpretability of passive signals; Artifact C/RQ4 for the decision to contact/schedule assessment.
P2 synthesis	P2 sees value in the passive signal, especially for cases where the person does not report symptoms. However, without additional clinical information, a high score based on the mouse can generate hesitation in the decision to contact or schedule assessment.

4.2 Interpretation of mouse patterns

Element	Extraction
Quote P2	“The program itself also helps identify with the mouse movements.”
Paraphrase	P2 understands that the system uses mouse movements as an additional source of signal.
Quote P2	“Because I do not understand [...] the mouse questions.”
Paraphrase	P2 does not feel comfortable directly interpreting the mouse patterns.
Quote P2	“Hence I would automatically look at the symptoms. [...] anxiety, stress.”
Paraphrase	Faced with an explanation based on mouse data, P2 would tend to return to the self-reported symptoms.
Initial codes	Passive signal generally understood; technical interpretation limited; return to self-report; need to translate mouse patterns into clinical language.
Artifact	Artifact B/RQ3.
P2 synthesis	Mouse patterns are understood as a detection mechanism, but they are not clinically self-explanatory for P2. The explanation needs to translate passive variables into comprehensible clinical implications, or else present them as an alert with uncertainty/context.

4.3 Clarity that it is not diagnosis

Element	Extraction
Quote P2	“They need clinical assessment.”
Paraphrase	P2 frames the score as a reason for assessment, not as a final diagnosis.
Quote P2	“Anamnesis [...] is the clinical interview.”
Paraphrase	The clinical interview is seen as a necessary step for confirmation and case understanding.
Initial codes	Tool as triage; clinical assessment necessary; non-substitution of the clinician; confirmation by anamnesis.
Artifact	Artifact C/RQ4, with implications for clinical safety.
P2 synthesis	The non-diagnostic boundary seems clear to P2. The system is interpreted as a triage/signaling tool that should lead to clinical assessment, not as an automatic diagnosis.

4.4 Additional information before acting

Element	Extraction

Element	Extraction
Quote P2	“But is she [medicated]?”
Paraphrase	P2 identifies medication as critical information for interpreting discrepancy between score and self-report.
Quote P2	“One thing is for a person [...] to be medicated; she may answer differently than if she were not medicated.”
Paraphrase	Medication may alter self-reports and the symptomatic presentation.
Quote P2	“Is she currently medicated? And then even display here in the program the medication she is taking.”
Paraphrase	P2 suggests collecting and showing medication in the system.
Quote P2	“If I saw the medication, if I saw [...] attempts and thoughts, sleep, physical activity, it could help...”
Paraphrase	Medication, suicidality, sleep and physical activity would help decide whether to act on the alert.
Initial codes	Medication; masked/attenuated symptoms; suicidal ideation; attempts; sleep; physical activity; decision to contact; context to resolve discordance.
Artifact	Artifact C/RQ4, because it involves contextual information and action; also Artifact B/RQ3, because it helps interpret the model alert.
P2 synthesis	P2’s main need in discordant cases is additional clinical context. Medication is especially important because it may explain normal self-reports despite a clinically relevant history/symptomatology.

4.5 Bias from familiarity with the questionnaire / equipment

Element	Extraction
Quote P2	“And as they [get to know] the questions, it will be easier for them to answer...”
Paraphrase	P2 agrees that prior familiarity with the questionnaire may alter response behavior.
Quote P2	“If the person has already filled it in once or twice, they are more [...] used to filling it out...”
Paraphrase	Prior completion may introduce bias into the measured behavior.
Initial codes	Familiarity with the questionnaire; behavioral bias; task learning; artificial speed; validity of passive signals.
Artifact	Mainly Artifact B/RQ3; also Artifact C/RQ4 if the dashboard has to show

Element	Extraction
	"previous completion/knowledge of the questionnaire."
P2 synthesis	P2 confirms the relevance of prior knowledge of the questionnaire as a possible confounder of passive signals. This reinforces the need to collect/show whether the person has previously answered the instrument.

---

5. Final questionnaire / open question

The guide ends with explainability items for Artifact B, SUS items for Artifact C, and an open-ended question about what change would make the interface more useful, clearer, or safer.

Element	Extraction
Quote P2	"I do not know what the session identifier code is."
Paraphrase	Operational doubt while completing the form.
Quote P2	"The profile is the profession, is that it?"
Paraphrase	The profile item may need clarification.
Quote P2	"Ah, I have to write everything I have been saying here and I do not remember."
Paraphrase	The open-ended question may create memory burden if the participant has already given extensive oral feedback.
Quote P2	"It would be here the question of medication, sleep... Physical exercise. The colors..."
Paraphrase	P2 summarizes the main changes: medication, sleep, physical activity and colors/legends.
Quote P2	"Suicidal thoughts, suicide attempts..."
Paraphrase	P2 reinforces the distinction between types of suicidal information.
Quote P2	"Maybe it would be good to have a definition of each scale, of what it assesses."
Paraphrase	The interface should explain each scale/construct.
Quote P2	"Each variable, like the mouse, the time interval."
Paraphrase	It would also be useful to define passive variables and time windows.
Quote P2	"Literally, have access to question by question and to how much the person answers each question."

Element	Extraction
Paraphrase	P2 wants access to item-by-item questionnaire responses.
Quote P2	“I like talking to people about each question and about the answer to each question.”
Paraphrase	Item-by-item responses would support the clinical interview and exploration of the subjective meaning of each response.
Initial codes	Form clarification; memory burden in the open question; medication; sleep; exercise; suicidality; colors; scale definitions; passive variable definitions; item-by-item access; support for clinical interview.
Artifact	Both. Artifact C/RQ4 for functionalities and contextual information; Artifact B/RQ3 for scale definitions, variables, and model explanation.
P2 synthesis	The final question consolidates the main improvement requirements: add clinical context, make scales and colors interpretable, explain model variables, and allow access to the full questionnaire responses.

Overall synthesis by artifact — Interview 02

Artifact C / RQ4 — Dashboard, prioritization, clinical utility and workflow
P2’s evaluation is globally positive regarding the visual organization and clinical usefulness of the interface. The detail view is considered easier to understand visually than if it were only text. The dashboard is seen as support for triage, preparation of assessment, and identification of cases that may require a clinical interview. However, P2 reinforces that the interface would be more useful and safer if it included medication, sleep, physical activity, diet, suicidal ideation differentiated between thoughts and attempts, and access to item-by-item responses. There is also a strong requirement regarding the interpretability of the scales: P2 needs to know the meaning of each subscale, whether high values are good or bad, and how to interpret colors/levels. This directly affects the clinical usefulness of the dashboard.

Artifact B / RQ3 — Explanation, factors, narrative and passive signals
P2 accepts the explanation as globally useful and understands the general logic of combining self-reports with passive signals. In the discordant case, she recognizes that the mouse can help identify people who do not report symptoms. However, the mouse-based explanation is not clinically transparent by itself: P2 says she does not really understand “the mouse questions” and that she would first look at symptoms, anxiety, and stress.
Thus, the model explanation needs to be translated further into clinical language, including definitions of variables, scales, time intervals, and possible confounders such as medication, prior familiarity with the questionnaire, and perhaps the equipment/collection conditions.

Potential emerging themes — separated from the protocol-organized findings

1. The interface is visually promising, but depends on explicit clinical semantics
   P2 values graphs, factors, and summary, but asks for the meaning of the scales, the direction of the scores, and interpretable colors.

2. Passive signal is useful as an alert, but insufficient without clinical context
   The mouse can detect discordance, but P2 would hesitate to act on a 94% score without self-reported symptoms.

3. Medication is a critical variable for interpreting discordance
   Medication may explain low self-reports in clinically relevant people; it should be collected and visible.

4. Safe triage requires risk variables and personal resources
   Suicidal ideation, attempts, sleep, physical activity, diet and emotional regulation appear as clinically useful data.

5. Explainability must include definitions, not only factors
   For P2, it is not enough to present "factors": it is necessary to explain what each scale/variable assesses and how to interpret high/low values.

6. Item-by-item access can support the clinical interview
   P2 wants to see individual questions and responses in order to explore with the patient the meaning of each answer.

7. Familiarity with the questionnaire may bias behavioral signals
   If the patient already knows the questionnaire, they may answer faster, affecting mouse variables and possibly the model.

8. Workflow integration depends on the institutional context
   P2 suggests that the integration of the system would be different in hospitals and private clinics. In hospitals, there may be a formal triage consultation, while in private clinics the first contact tends to occur by phone, email, informal referral, or internal referral.

9. The system may support referral, not only risk prioritization
   In the private-clinic context described by P2, triage involves understanding which professional is most appropriate for each case, considering not only severity, but also areas of expertise, target disorders, and the therapeutic approaches used by the clinicians.

10. Passive signals require more controlled collection conditions
    Because mouse variables may depend on the equipment, screen resolution, collection context, and familiarity with the questionnaire, real implementation may require a standardized computer or controlled collection conditions, at least while the model is not adapted to different devices.

11. Remote application may be difficult in the current version
    Sending the questionnaire home for completion could be useful from a practical perspective, but may introduce technical and behavioral variation that affects cursor variables. This makes remote collection less straightforward, especially if passive features are used by the model.

12. The system seems more viable as a complement to first contact than as a direct substitute
    P2 recognizes value in the system for providing structured data before assessment, but requiring the patient to travel only to fill in questionnaires could create friction. Thus, the tool seems more appropriate as a complement to the initial contact, for example

before or during the first consultation, rather than as a direct substitute for the phone call.

13. Application in the waiting room or during a first consultation may be a plausible integration path
    One emerging possibility is to apply the system in the clinic, before the first consultation or during an initial triage consultation. In this scenario, the dashboard would help the professional prepare the assessment, identify topics to explore, and decide whether other instruments or referrals are needed.

14. Future versions may allow greater clinical control over the questionnaires used
    The interview suggests that, in the future, it may be useful to allow the clinician to select the most appropriate questionnaires for the case, rather than always using a fixed battery. This would make the system more flexible and more aligned with different clinical assessment styles.

APPENDIX H: Formative Evaluation Protocol

Forms Link:
https://docs.google.com/forms/d/e/1FAIpQLSe39WYnxlzRyMploRpQ7BvlNPFyDLnhxkfhJmA0F4uRd6CqTQ/viewform?usp=dialog

Formative evaluation guide for the mockups

1. Session objective
   This session aims to evaluate mockups of a clinical decision-support system for triage/prioritization in a depression context. The evaluation focuses on two artifacts:
   • Artifact B — Explainability engine: how the model's explanatory factors are presented through SHAP contributions and textual narrative.
   • Artifact C — Dashboard/CDSS: how the dashboard organizes the prioritization ranking, scores, self-reports, additional context, and clinical review actions.

The session is formative and exploratory. The objective is not to evaluate the participant's performance, but rather to identify whether the interface is clear, useful, and appropriate for the clinical context.

1. Initial preamble — 2 minutes
   Before showing the mockups, explain:

Imagine that this system has been integrated into an occupational health, clinical, or mental health service context, where workers or users complete questionnaires such as the DASS/EADS-21 and the RESS digitally (for the first time). This interface is a static mockup and does not yet correspond to a production system. The system's objective is to support triage/prioritization and clinical review, not to produce an automatic diagnosis.

In this evaluation, the focus is on the initial use of the system after the first completion of the questionnaires, with the aim of supporting the triage/prioritization queue for clinical review and providing contextual information that may help the professional consider referrals or aspects to explore during assessment. Longitudinal follow-up functionalities, detailed history, or continuous monitoring are not part of the scope of this mockup version.

The score shown results from a machine learning model trained on self-report data and interaction patterns during the digital completion of the questionnaires, having learned to distinguish profiles with clinical symptomatology from a control group. In the interface, this score is used to order cases by review priority. The threshold used for signaling was chosen to favor sensitivity, that is, to reduce the probability of missing potentially relevant cases, accepting the possibility of more false signals.

The available data were limited to the instruments used in the study, namely DASS/EADS-21, RESS, and mouse interaction patterns during the RESS. For this reason, the interface does not include clinical variables that were not collected in this study, such as suicidal ideation, sleep, medication, or detailed clinical history.

The final decision always belongs to the clinical professional. During the session, I would like you to think aloud what you are interpreting, what seems clear, confusing, useful, or

potentially problematic, in an honest and impartial way. There are no right or wrong answers and all feedback is welcome to help improve the prototype.

Ask permission for recording/annotation:
“Do you authorize the session to be recorded/annotated only for academic analysis of the dissertation?”

2. Task 1 — Prioritization dashboard — 3 minutes
   Show View 1: Clinical Prioritization Dashboard.

Think-aloud prompt:
“Imagine that you are reviewing new cases after the first completion of the questionnaires. Please say out loud what you interpret on this page.”

Supporting questions, if necessary:
• What information draws your attention first?
• Is the ranking clear?
• Is the meaning of “priority score” and “above/below threshold” understandable?
• Which case would you open first and why?
• Does the distribution by signaling help interpret the queue?

3. Task 2 — High-priority convergent case — 3 minutes
   Show part900.

Brief explanation:
“This is a case in which the model’s high score is accompanied by self-reports that are also expressive.”

Think-aloud prompt:
“Observe the case detail and say how you would interpret this information to support a possible clinical review.”

Supporting questions:
• Is the relationship between self-report, score, and model explanation clear?
• Does the explanation of the factors help you understand the signaling?
• Does the textual narrative add anything to the graph/list of factors?
• Is the distinction between information used by the model and additional context clear?

4. Task 3 — Borderline or discordant case — 3 minutes
   Preferably show part721 to test interpretation of passive signals, or control_Part546 to test proximity to the threshold.

For part721:
“This is a case in which the model assigns high priority, but the DASS/EADS-21 self-reports are less expressive. The explanation points mainly to interaction patterns during the RESS.”

Supporting questions:
• Would this signaling be useful or raise doubts?
• How would you interpret the mouse interaction patterns in this context?
• Does the system make it clear that this is not diagnosis?
• What additional information would you like to have before acting?

For control_Part546:
“This is a case below, but close to the operational threshold.”

Supporting questions:
• Is the proximity to the threshold relevant to the decision?
• Does the explanation help you understand why it fell below the threshold?
• Should this case continue to be monitored?

5. Final questionnaire — 4 minutes
   Apply in Google Forms.
   Scale: 1 = Strongly disagree; 5 = Strongly agree.

Items adapted from Bergomi et al. — Artifact B / Explainability

1. The explanation presented helps me understand why the model signaled the case with this priority.

2. The explanation presented would be useful to decide whether the case should be reviewed or clinically followed up.

SUS — Artifact C / Overall usability
3. I would like to use this system frequently.
4. I found the system unnecessarily complex.
5. I found the system easy to use.
6. I think I would need technical support to be able to use this system.
7. I found that the various functions of the system were well integrated.
8. I found there was too much inconsistency in this system.
9. I imagine most people would learn to use this system very quickly.
10. I found the system very difficult or cumbersome to use.
11. I felt confident using this system.
12. I needed to learn many things before I could use this system.

Final open-ended question
13. What change would make this interface more useful, clearer, or safer for use in a real-world context?

6. Closing
   Thank the participant for their participation and reinforce that the feedback will be used only to improve and discuss the prototype in the dissertation context.

APPENDIX I: Analytical Dataset Construction Criteria for RQ1 and RQ2

Table 16 Table H.1: Analytical dataset construction criteria for RQ1 and RQ2.

	RQ1 — Longitudinal	***RQ2 — Cross-sectional***
Unit of analysis	Consecutive session pair (t → t+1)	First eligible session per participant
Group	Clinical only	Clinical + control
Longitudinal feasibility	42 participants with ≥2 sessions; 27 with >2; 14 participants with ≥3 sessions; max 8	—
Additional criteria	Same-day ambiguous sessions removed; keyboard-use sessions removed; two layout-	Same-day ambiguous sessions removed; keyboard-use sessions removed; clinical
Passive source	DASS-21	RESS
Final N	37 participants, 91 pairs	152 participants (84 clinical, 68 control)
	RQ1 — Longitudinal	RQ2 — Cross-sectional
Unit of analysis	Consecutive session pair (t → t+1)	First eligible session per participant
Group	Clinical only	Clinical + control
Longitudinal feasibility	42 participants with ≥2 sessions; 27 with >2; 14 participants with ≥3 sessions; max 8	—
Additional criteria	Same-day ambiguous sessions removed; keyboard-use sessions removed; two layout-	Same-day ambiguous sessions removed; keyboard-use sessions removed; clinical
Passive source	DASS-21	RESS
Final N	37 participants, 91 pairs	152 participants (84 clinical, 68 control)

APPENDIX J: Representative Cases and Dashboard Prototype Screenshots Used in the Formative Evaluation

Prototype Link: https://player-notion-19928112.figma.site

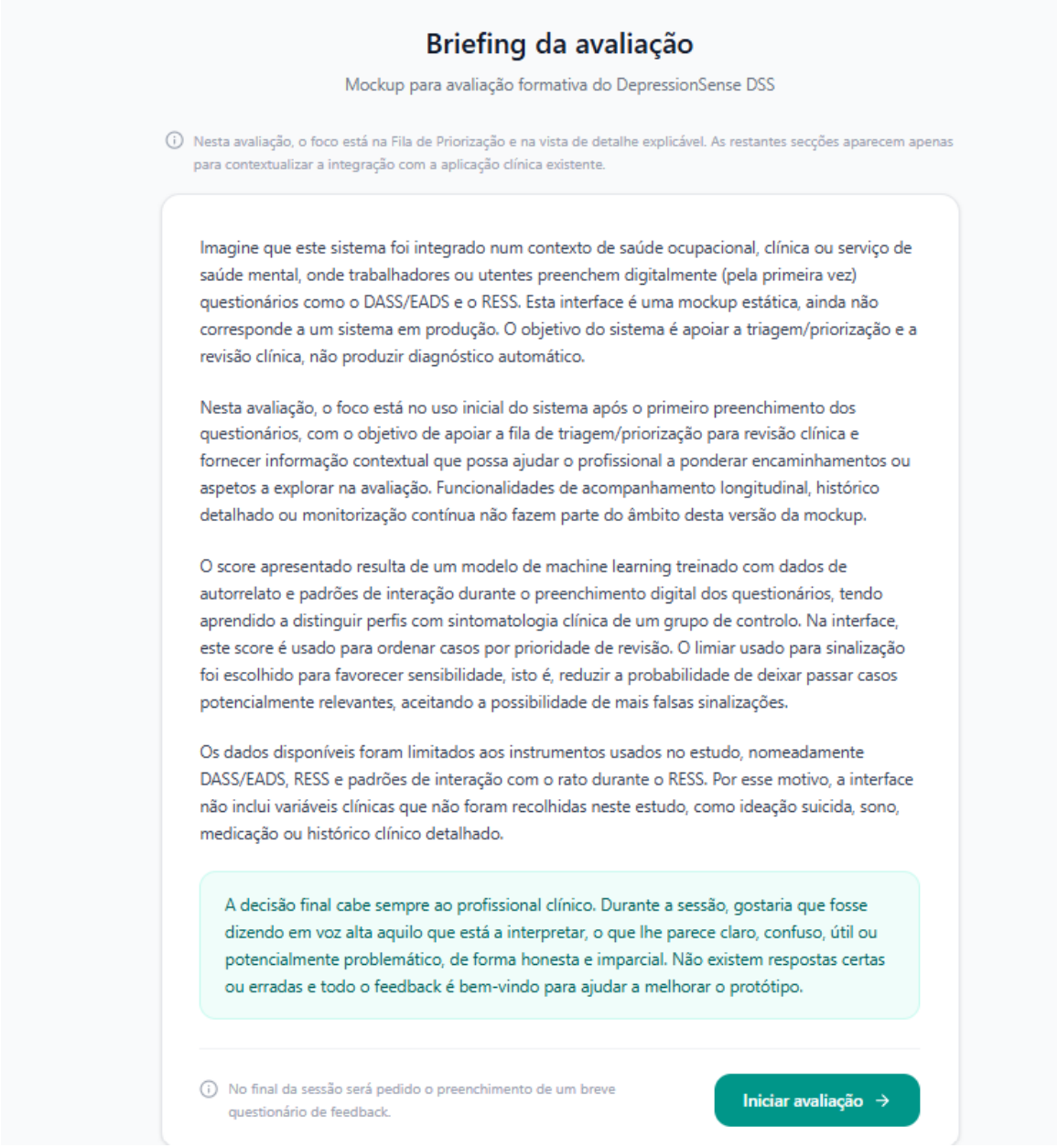

Figure 6 Figure I.1: Preamble view from prototype mockup

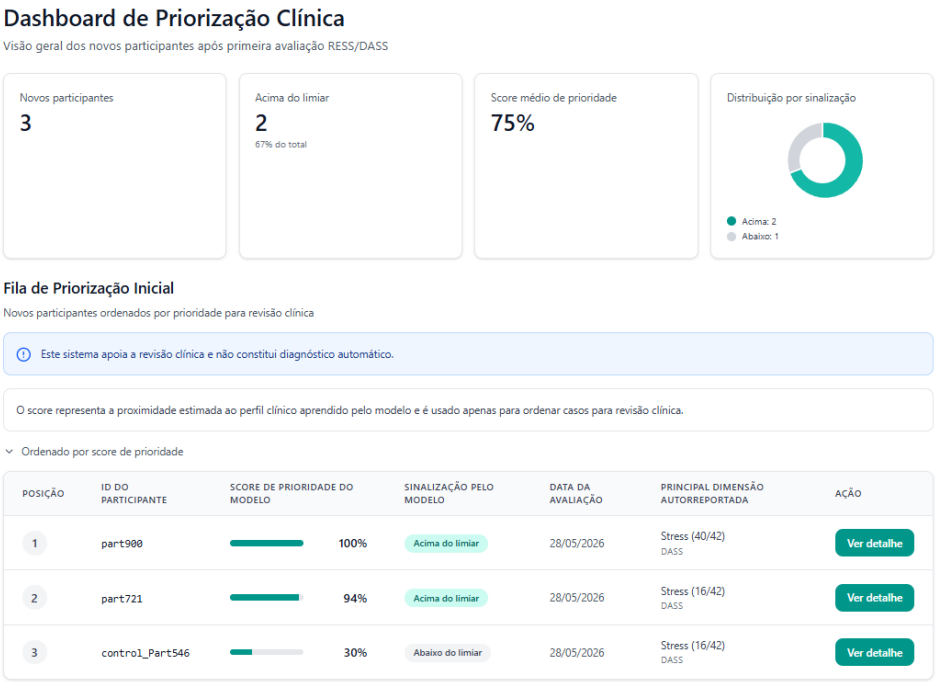

Figure 7 Figure I.2: Dashboard prototype mockup - Prioritization queue

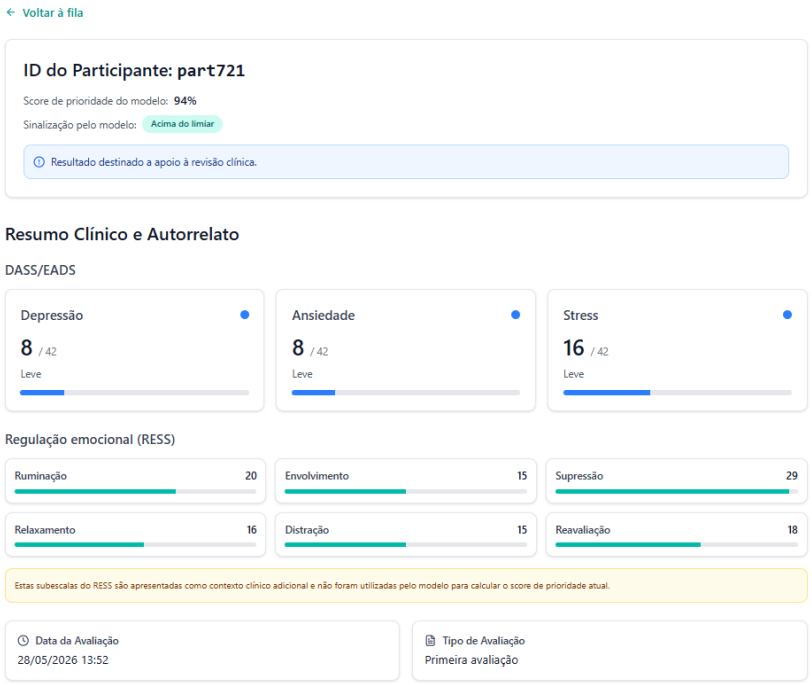

Figure 8 Figure I.3: Case detail mockup (Clinical Context Section) - High-priority ambigiuos case