Int. J. Appl. Math. Comput. Sci., 2026, Vol. 36, No. 2, 347–367
DOI: 10.61822/amcs-2026-0023

AN EXPLAINABLE HYBRID MODEL FOR DECODING SILENT MENTAL HEALTH SYMPTOMS THROUGH SOCIAL MEDIA INTERACTION AND TEXTUAL WITHDRAWAL PATTERNS

AYODEJI OLUSEGUN IBITOYE [a,*], OLADOSU OYEBISI OLADIMEJI [b], TEMITAYO MATTHEW FAGBOLA [c]

[a]School of Computing and Mathematical Sciences
University of Greenwich
Old Royal Naval College, Park Row, SE10 9LS, London, UK
e-mail: a.o.ibitoye@greenwich.ac.uk

[b]Faculty of Engineering and Design
Atlantic Technological University
Ash Lane, Sligo, F91YW50, Ireland

[c]Centre of Excellence for Data Science, AI and Modelling
University of Hull
Cottingham Road, Hull, HU6 7RX, UK

Mental health (MH) disorders, particularly depression and social withdrawal, represent critical global challenges, undermining both individual well-being and societal productivity. Social media provide a unique lens to capture digital behaviours that may serve as early indicators of MH states. While conventional diagnostic methods are often subjective and resource-intensive, digital behaviour analysis offers scalable, non-invasive alternatives. Yet, existing studies frequently isolate emotional, relational, and temporal dimensions, limiting predictive accuracy and interpretability. This study proposes digital behavioural continuum theory (DBCT), a framework integrating these dimensions to model MH states holistically. A hybrid machine learning architecture is developed, combining graph neural networks (GNNs) for relational structures, recurrent neural networks (RNNs) for temporal sequences, and Valence Aware Dictionary and Sentiment Reasoner (VADER) for the sentiment analysis technique to extract affective signals from user-generated content. Model transparency is ensured through Shapley additive explanations (SHAP), enabling identification of the most influential behavioural markers. Results demonstrate that emotional features (e.g., sentiment scores, sad reaction ratios) exert the greatest predictive influence, followed by temporal signals such as posting frequency and response latency, while relational attributes contextualise social withdrawal. The proposed model achieves an F1-score of 90.4%, a precision of 89.7%, and a recall of 91.2%, significantly surpassing baseline approaches. Importantly, the datasets analysed were not clinically diagnosed but were curated to reflect real-world social media behaviours associated with potential mental health signals. By advancing an interpretable, data-driven framework, this research bridges theoretical innovation with practical application, enhancing digital MH monitoring and supporting early, scalable interventions.

Keywords: mental health prediction, digital behavioural continuum theory, social media analytics, temporal behaviour analysis, social withdrawal, graph neural networks, recurrent neural networks.

1. Introduction

Social media platforms have transformed the approach through which people communicate, form relationships, and express their emotional states. Research has emerged proposing that user activity patterns on these platforms may serve as behavioural indicators of MH issues, specifically depressive symptoms and social withdrawal (Dhelim *et al.*, 2023; Kim *et al.*, 2025). Building on this, prior research has demonstrated that user behaviours

*Corresponding author

on social media, such as posting frequency, interaction rates, and response latency, can act as digital proxies for underlying psychological states (De Choudhury *et al*., 2013; Tsakalidis *et al*., 2018). For example, reduced posting frequency and increased response latency may indicate social withdrawal or cognitive fatigue, both associated with depressive symptoms. Similarly, diminished interaction rates can reflect declining social engagement, a known behavioural marker in mental health diagnosis (Guntuku *et al*., 2017). These passive signals offer a scalable, non-invasive means to infer mental health states, particularly in contexts where clinical assessment is impractical.

Further studies have also found that individuals experiencing depressive symptoms often demonstrate decreased posting frequency as compared to their more mentally resilient counterparts (Brailovskaia and Margraf, 2022). This decline is frequently coupled with a reduction in the quality of content shared, which may reflect feelings of hopelessness or a lack of interest (Kmetty and Bozsonyi, 2022). Similarly, users withdrawing from social interaction might reduce their overall presence on platforms like Instagram, leading to an absence of self-expression and connectivity, illustrating the internal struggles they might be enduring (Adeyanju *et al*., 2021). Interaction rates, defined as the frequency of engagement with posts through likes, comments, and shares, further illuminate behavioural patterns indicative of social withdrawal, where lower interaction rates can signify depressive symptoms and a retreat from social engagement (Zhang *et al*., 2021). A systematic review highlighted that individuals with depression are likely to interact less with friends and family on social media platforms, reinforcing their feelings of isolation (Brailovskaia *et al*., 2021). Occasionally, individuals may engage only minimally, e.g., by viewing posts without liking or commenting, which is often referred to as passive social media use. This behaviour has been substantially linked to heightened feelings of anxiety and depression, suggesting that the nature of the interaction is as important as the frequency (Marengo *et al*., 2022).

In contrast, increased interaction rates may not always signal healthy engagement. Social media can facilitate compulsive behaviours, leading to negative emotional states. For instance, excessive likes or comments may be driven by fear of missing out (FOMO), propelling users to maintain an active presence while struggling with underlying depressive symptoms (Moore and Craciun, 2021). The dual role of interaction rates as both a symptom and a trigger of distress underscores the need to understand the link between social media dynamics and mental health. Another critical behavioural indicator is response latency, i.e., the time taken to respond to messages or comments. Extended response latency may indicate social withdrawal, a common behaviour observed among users with depressive symptoms (Giuntini *et al*., 2021). Thus, individuals may often feel overwhelmed by social interactions, leading to delays in engagement that reflect their emotional state (Nazmunnahar *et al*., 2023). Conversely, prompt responses may showcase an attempt to maintain social connections, albeit sometimes driven by anxiety or compulsion rather than genuine interest. Further complicating this dynamic, the isolation experienced by individuals can manifest itself in their online interactions, perpetuating a cycle of decline in social engagement. Social media trends during the COVID-19 pandemic demonstrated a significant rise in depressive symptoms as lockdowns limited physical and social interactions (Chemnad *et al*., 2023). Research suggested that users became more reliant on digital interaction, yet many reported feeling lonelier, demonstrating how an increase in social media engagement does not necessarily correlate with improved MH (Brand *et al*., 2024).

The rise of social media as a dominant communication medium, combined with advancements in ML and data analytics, offers unprecedented opportunities to identify behavioural indicators of MH issues. Patterns in user interactions, posting habits, and emotional expressions have been shown to correlate with states such as depression and social withdrawal (Marengo *et al*., 2022). This research leverages these insights to tackle the challenge of identifying and predicting mental health issues in a timely and non-invasive manner. While traditional methods, such as self-reported surveys and clinical interviews, are often time-consuming, subjective, and reactive, they detect problems only after significant distress has occurred. In contrast, analysing social media activity provides an opportunity for proactive mental health assessment by uncovering early indicators through digital footprints. However, the relationship between these behaviours and mental health is multifaceted, involving relational, temporal, and emotional dimensions. Current approaches through ML (Arowosegbe and Oyelade, 2023; Ibitoye *et al*., 2021) often fail to integrate these aspects into a unified framework, resulting in fragmented analyses. This research aims to answer the following question: *How can relational, temporal, and emotional patterns in social media behaviours be integrated into a unified framework to predict mental health states accurately and ethically?*

To address current gaps in mental health prediction research, this study develops digital behavioural continuum theory (DBCT) to analyse mental health states using advanced machine learning techniques. It explores how digital behaviours on social media, specifically relational, temporal, and emotional patterns, can be used to detect early signs of depression and social withdrawal. The research is guided by four core objectives: (1) to develop a multi-dimensional behavioural framework

(DBCT) that unifies relational, temporal, and emotional signals, (2) to implement an interpretable hybrid model that integrates graph neural networks (GNNs), recurrent neural networks (RNNs), and natural language processing (NLP) for predictive analysis, (3) to evaluate the model's capacity for detection of mental health risks while ensuring ethical data usage through anonymised, publicly available, and non-invasive input, and (4) to apply explainable AI methods (specifically SHAP) to interpret model predictions and identify the most influential features, enhancing the transparency and trustworthiness of results.

The broader impact of this research lies in its potential to inform real-time monitoring strategies, shape the design of socially accountable technologies, and lay the groundwork for data-driven mental health policies. Furthermore, to enhance the interpretability of the model's predictions, the study incorporates SHAP (Shapley additive explanations), an explainable AI technique that reveals how different features, emotional, temporal, and relational, contribute to mental health inferences. This layer of transparency is essential for building trust in predictive systems, particularly in sensitive domains like mental health.

The rest of the paper is structured as follows. Section 2 reviews the literature on digital behaviour analysis, ML for MH prediction, and ethical concerns in social media data. Section 3 details the methodology, including data collection, preprocessing, and model architecture. Section 4 presents results, assessing model performance and key behavioural features. Section 5 examines theoretical, managerial, and policy implications, while Section 6 discusses limitations and future research. Finally, Section 7 summarises the contributions and the study's impact on digital mental health interventions.

2. Literature review

The growing intersection between digital behaviour and mental health has prompted a surge of research exploring how individuals' interactions on social media platforms may reflect underlying psychological conditions. As digital platforms become embedded in daily life, they generate rich streams of behavioural data, including user engagement, posting habits, social connectivity, and emotional expression that can provide insights into mental health states. However, while existing studies have demonstrated the potential of digital behaviour analysis, many approaches focus narrowly on isolated features, such as sentiment in text or network size, without capturing the dynamic interplay between social, temporal, and emotional factors. This fragmented perspective limits the predictive power and practical application of existing models, especially in identifying early or subtle signs of distress. This review examines prior research on the use of social media data for mental health prediction, beginning with theoretical frameworks of digital behaviour and their implications on mental health, before focusing on three core dimensions: relational behaviours (Section 2.1), temporal dynamics (Section 2.2), and machine learning methodologies (Section 2.3). It highlights the strengths and limitations of existing studies and lays the foundation for the development of digital behavioural continuum theory, a comprehensive framework that integrates these dimensions to enhance predictive accuracy and interpretability.

2.1. Theoretical frameworks of digital behaviour theory and its implications for mental health. The intersection of digital technology and mental health has prompted the development of various theoretical frameworks aimed at clustering (Ibitoye *et al*., 2025b), understanding and improving digital behaviour interventions (Naslund *et al*., 2017). This aligns with the results of Voorheis *et al*. (2023), who claim that understanding theories, models and frameworks in the change of digital health behaviour is crucial for effective design. Their qualitative analysis indicates that systematic application of theoretical frameworks can lead to more targeted and successful interventions. Šmahel *et al*. (2018) propose a distinct theoretical framework that tackles the interaction between digital technology and health behaviours. Their work highlights the multifaceted nature of the interactions on digital health, suggesting that an understanding of user involvement is essential for the development of effective interventions. Also, Ball *et al*. (2025) strengthens this notion through a review of scoping that examines how various theories have been operational within digital health services for people with serious mental health problems, indicating both the diversity and the potential effectiveness of these approaches in practical contexts.

The conceptualisation of commitment with the interventions of change of digital behaviour is another critical area, as indicated by Perski *et al*. (2017). Their systematic revision underlines that understanding the involvement of users is fundamental for the success of the interventions and can guide future research in the optimisation of digital health strategies. Interestingly, Mohr *et al*. (2014) introduce the model of behavioural intervention technology (BIT), which integrates both conceptual and technological elements for the eHealth and mHealth solutions, thus advancing the field by providing a complete picture for the design of interventions. Subsequently, the picture of ideas proposed by Mummah *et al*. (2016) offers a structured approach for the development of effective digital interventions aimed at changing health behaviours. This framework underlines the importance of integrating design, evaluation and sharing strategies, which can improve the effectiveness

of digital health interventions. Collectively, these theoretical paintings provide valuable information on the implications and effectiveness of digital behaviour theories in relation to mental health, guiding future research and intervention strategies.

2.2. Relational behaviours in social media. The proliferation of social networks (SNs) has introduced complex MH challenges in the digital age, particularly among adolescents. Valkenburg *et al*. (2022) conducted a general review that highlights the impact of SNs on the MH of adolescents, which suggests both positive and negative results. Social networks offer platforms for social connection; however, they can simultaneously contribute to anxiety and depression (Naslund *et al*., 2020). The dual nature of these interactions underlines the need for a deeper understanding of specific behaviour patterns influenced by the use of SNs. Relational behaviours refer to the observable patterns of digital interaction between users and their networks, including factors such as engagement reciprocity, clustering within social circles, centrality in interaction graphs, and the density of ego-networks (Jackson, 2008; Burke *et al*., 2010). These attributes provide a structural and behavioural context for identifying social isolation or withdrawal online. This need for contextual investigation was reinforced by Karim *et al*. (2020), who provided a review that reinforces the association between MH results and the use of SNs, emphasising the need for specific context investigation. Coyne *et al*. (2020) extended this narrative, exploring how the duration of exposure to SNs can significantly affect mental health over time.

In addition, Chen *et al*. (2021) examined the purposes of SNs related to the health, discovering that, although they can facilitate the dissemination of health information, there are risks of misinformation that could exacerbate psychological anguish. The evolution of digital communication has led to widespread engagement with social media platforms, significantly influencing mental health outcomes by generating vast amounts of user-generated text. In response to this, Ibitoye *et al*. (2025a) proposed a contextual emotional transformer-based model, which demonstrated notable improvements in predictive accuracy by integrating emotional attention mechanisms and contextual embeddings, achieving up to 94.5% accuracy with Roberta, outperforming traditional transformer models in analysing such text. The myriad of relational behaviours observed online, such as social support, interactive exchanges and the pursuit of validation, constitutes a fundamental aspect of the way users digitally browse their social worlds.

Relational behaviours on social media can manifest themselves as intricate models that reflect and influence the mental health of users, underlining the need for a rigorous academic examination of these phenomena (Lin *et al*., 2020). Uban *et al*. (2021) provide convincing evidence for the interaction between online interactions and cognitive-emotional states. Their study highlights how the relational behaviours of users, such as the response to the positions of peers or engaging in discussions in various social media contexts, can deeply model their cognitive assessments and emotional responses. This research illuminates the critical nature of social exchanges, suggesting that users who actively participate in these relational dynamics tend to experience improved emotional well-being.

On the other hand, insufficient involvement or negative interactions on these platforms can contribute to negative results on mental health, such as an increase in feelings of solitude or depression. In addition, the search for social validation emerges as a prevalent behaviour on social media, accentuating the need for an understanding of its impact on mental health. According to Chancellor and De Choudhury (2020), the metrics of involvement, such as likes, shares and comments, are indicators of social validation with direct for implications for the self-esteem of users. Their critical revision indicates that individuals often evaluate their value based on these digital statements, leading to psychological branches that can strengthen or undermine well-being.

The study by Valdez *et al*. (2020) accentuates the need to employ predictive techniques that use these digital metrics not only to evaluate mental health conditions but also to adapt interventions involving users constructively within SNs. The articulation of emotional expressions on social media platforms further improves the understanding of mental health dynamics. According to Gauthier *et al*. (2021), the fluctuations of mental health tendencies during the COVID-19 pandemic reveal how the positive and negative expressions served by bartenders for the mental health of the company were rightly captured. Their analysis underlines the rapid spread of emotions on social media and the corresponding implications for public health discourse, particularly in times of crisis. In addition, Yang *et al*. (2020) elaborated on the role of emotional regulatory strategies used by users in response to the emotional panorama of social media, highlighting how these strategies can mitigate the negative results on mental health. Their results suggest that users who effectively browse emotional expressions online can experience better resilience and emotional stability.

2.3. Temporal dynamics and mental health. Ultimately, the interaction of digital behavioural models on social media acts as a fundamental area for investigations, offering critical insights that can inform both the theoretical paintings and the practical applications aimed at improving mental health in the digital era. The temporal

dynamics of the involvement of social media, which include both the frequency and duration of use, have significant implications for mental health. Kim *et al*. (2021) examined the correlation between the time spent on SNs and emotional well-being, discovering that an increase in commitment is associated with high levels of anxiety and depression. Their study suggests that not only the frequency of the use of social media but also the continuous nature of involvement, characterised by longer session duration, can exacerbate the feelings of solitude and social comparison, thus negatively influencing the psychological states of users. Based on this, Meier and Reinecke (2021) explored how the time spent on social media relates to the perceptions of users of their SNs in real life. Their discoveries indicate that excessive interaction on social media can lead to a diluted sense of realisation derived from face-to-face relationships. Participants who engaged with social media for prolonged periods reported a greater impact of relational anxiety, often manifesting as a fear of losing (FOMO), which emerged as a prevalent concern within digital communication settings.

Shannon *et al*. (2022) also articulate these results in their systematic revision of the problematic use of social media among teenagers, underlining that not only excessive use but also a problematic commitment characterised by compulsive control behaviours can lead to results harmful to MH, particularly in the younger demography. In addition, the influence of external events on social media behaviours highlights a complex interaction between temporal dynamics and emotional responses. The studies by Haddad *et al*. (2021) and Nutley *et al*. (2021) in the context of the COVID-19 pandemic illuminate how crises can cause collective emotional responses that, remodel models of social media involvement. Their research indicates that, during the pandemic, greater dependence on digital platforms for social connection led to intense expressions of anxiety and fear among users. These researchers underline that in times of crisis, social media can serve a double purpose. They provide comfort and community, simultaneously amplifying anguish due to the incessant accessability to negative news and social comparisons.

Furthermore, the intersection of the perception of the body and mental health in the context of social media was an area of growing research interest. Merino *et al*. (2024) investigate how concerns about body image are exacerbated by the pervasive representation of the images idealised on social media, leading to harmful impacts on the self-esteem of users and mental health. Their results suggest that frequent exposure to well-cured online identity and lifestyle representations can lead to a dissonance between the perception of users and external standards, further contributing to anxiety and depressive symptoms. The largest implications of dependence on social media on mental health are outlined in the works of Fabris *et al*. (2020) and Huang (2022). Both studies articulate a clear link between the qualities of dependence on the use of social media, characterised by compulsive behaviour schemes and nomophobia (fear of being without a mobile device), and various mental health outcomes. In particular, Fabris *et al*. (2020) show that the compulsive necessity of interacting with social media has been related to the increase in anxiety and depressive symptoms, suggesting that the incessant need for validation and connection in virtual environments can replace the positive effects of social interaction. Huang (2022) echoes this feeling, claiming that a model of dependence could provide a picture to understand the harmful cycle of the use of social media and its effect on mental health, in particular among vulnerable populations.

2.4. Machine learning in behavioural analysis. The graphic-based models have acquired significant traction in the realm of automatic learning, particularly in the applications of behavioural analysis focused on the analysis of social media and the forecast of mental health. The advent of graph neural networks represents a transformative approach to understanding the intricate relationships relating to data on social media, in which user interactions often manifest themselves as complex networks. By modelling these interactions, GNNs can effectively extract information characteristics that improve predictive skills, allowing better identification of users' behaviour models and social dynamics (Dong *et al*., 2023). This ability is particularly salient in contexts in which mental health problems can be indicated through social interactions, suggesting that GNNs have the potential for early detection strategies and intervention. In the meantime, the temporal dynamics of the behaviour of users are crucial for understanding fluctuations in mental health states captured competently by sequential models, in particular those who use RNNs and short-term memory networks (LSTMs). This architecture excels in the processing of sequential data and the maintenance of information for prolonged periods, allowing the modelling of the user's interactions and behavioural variations over time (Chang *et al*., 2022).

The importance of the temporal context in social media is further underlined by the variability of users' expressions and by the evolving nature of online social interactions, which require models that can adapt to these dynamic environments. The integration of GNNs with sequential models such as LSTMs represents a promising hybrid approach that tries to capitalise on the strengths of both methodologies. For example, the research conducted by Liu *et al*. (2020) illustrates how the combination of RNNs with GNNs can produce significant improvements in the predictive accuracy for complex behavioural tasks. This approach facilitates a better understanding of the

user's behaviour by exploiting both relational insights and temporal data, thus facing the faceted nature of social interactions in behavioural analysis.

Further exemplifying this trend, studies by Chen *et al*. (2022) and Murshed *et al*. (2022) highlight the adaptability and effectiveness of hybrid models in identifying the patterns of abnormal behaviour, including, but not limited to, cyberbullying events and mental health crisis indicators on social media platforms. These models not only incorporate their relational structure of user interactions but also monitor the evolution over time, effectively filling the gap between static and dynamic analysis of social behaviour. The synergistic potential of graph-based, sequential, and hybrid models in behavioural analysis underscores their growing importance as powerful tools for understanding and predicting social media behaviours and associated mental health indicators. While research in this sector continues to expand, it is essential to further explore the integration of these methodologies to improve the effectiveness of predictive models in facing the challenges of contemporary society. The integration of hybrid models in behavioural analysis also highlights their application in providing the results of mental health based on data on social media.

A significant contribution is exemplified by the work of Kour and Gupta (2022), who employed a convincing neural network rich in tandem functionality with long-term short-term memory networks to predict user depression tweets. This approach shows marked progress in the methodologies for forecasting mental health, effectively exploiting the impact of the discourse on social media. In addition, the dynamic nature of social media platforms requires models capable of constantly evolving to a panorama of data. This adaptability is highlighted by Zandavi *et al*. (2021), who developed a hybrid model that combines LSTM networks with behavioural components to predict the spread of COVID-19. This model not only highlights the opportunity for timely health responses to public health but also demonstrates an effective adaptation to the uncertainty inherent in data on social media.

In addition to these results, Theodoropoulos *et al*. (2023) studied the applications of GNNs in representing the application of multivariate resources, producing insights on the user's behaviour observable in multiplayer mobile game contexts. Their results indicate that the behaviour captured by these models can indirectly contribute to the understanding of psychological states, thus influencing mental health assessments. The interaction between sequential and graphic-based models has synergistic advantages that make them particularly suitable for behavioural analysis. A practical illustration of this point is found in the work of Lazcano *et al*. (2023), who successfully applied a hybrid model that integrated recurrent neural networks with GNNs for financial time series. The methodologies they explored demonstrate the potential of extension in social media analysis, allowing improved predictive skills regarding users' behaviour and mental well-being. The prospects in this field are actively modelled, as discussed by Jin *et al*. (2024) and further articulated in the complete investigation by Munikoti *et al*. (2023). These works suggest a flourishing interest in exploiting hybrid models to improve predictive power in behavioural analysis, particularly to draw insights into mental health from the rich activity of social media. The juxtaposition of traditional automatic learning approaches with advanced hybrid architectures indicates a transformative potential in effectively acquiring the complexity of behavioural models and psychological indicators within social platforms. Hence, by exploiting relationships based on graphs and sequential data characteristics, researchers are opening the way to more accurate evaluations of mental health, reflecting a critical intersection between technology, psychology and social dynamics.

Building on the existing body of research, it is evident that, while prior studies have made significant strides in analysing social media behaviour for mental health prediction, most approaches remain limited by their focus on isolated behavioural dimensions. The absence of a unified framework that integrates relational, temporal, and emotional features restricts the ability to capture the complex interplay between social connectivity, behavioural changes over time, and emotional expressions. This study builds on emerging digital behaviour theories that suggest users' online interactions can serve as behavioural correlates of internal psychological states (Naslund *et al*., 2016; Reece *et al*., 2017). DBCT addresses these gaps by providing a holistic approach to mental health prediction, leveraging advanced ML techniques to analyse multi-dimensional social media patterns. To operationalise this framework, the following section details the methodology, including data collection, preprocessing, feature engineering, and the development of a hybrid model that integrates GNNs, RNNs, and NLP for comprehensive behavioural analysis.

3. Methodology

This section presents a clear and practical framework for implementing the proposed digital behavioural continuum theory. It streamlines the process while preserving the core elements necessary to analyse relational, temporal, and emotional dimensions of digital behaviour. A computational approach was adopted to operationalise DBCT, utilising sophisticated ML techniques: specifically, GNNs for relational modelling and RNNs for temporal analysis. These methods enable a comprehensive examination of social media

Algorithm 1. DBCT pipeline for multimodal mental health prediction.

Require: Dataset $D = \{\text{posts}, \text{interactions}, \text{reactions}\}$, labels Y
Ensure: Predictions $\hat{Y}$, SHAP explanations S
CleanedText $\leftarrow$ CleanText(posts)
$G \leftarrow$ ConstructGraph(interactions)
$X \leftarrow$ MergeByUser(CleanedText, reactions, G)

EmotionalFeat $\leftarrow$ ComputeVADER(CleanedText)
TemporalFeat $\leftarrow$ ExtractTemporal(post_times, responses)
RelationalFeat $\leftarrow$ ComputeGraphMetrics(G)

$R \leftarrow$ GNN(RelationalFeat)
$T \leftarrow$ LinearProjection(LSTM(TemporalFeat))
$E \leftarrow$ LinearProjection(EmotionalFeat)

$Z \leftarrow$ Concatenate(R, T, E)
$Z \leftarrow$ Transform(Z) {Dense + ReLU + Dropout}
$\hat{Y} \leftarrow$ Softmax(Z)

Train($\{R, T, E\}, Y$)
Evaluate using stratified 5-fold CV and LOUO-CV

$S_{\text{pre}} \leftarrow$ ComputeSHAP($\{R, T, E\}$)
$S_{\text{post}} \leftarrow$ ComputeSHAP(Z)
$S \leftarrow \{S_{\text{pre}}, S_{\text{post}}\}$
return $\hat{Y}, S$

engagement patterns and their relationship to mental health status. To consolidate the workflow and present the end-to-end implementation succinctly, Algorithm 1 offers a high-level pseudocode representation of the full multimodal pipeline, encompassing data preprocessing, feature extraction across all behavioural modalities, modality-specific modelling, fusion, prediction, evaluation, and explainability. This end-to-end pipeline reflects the operationalisation of the DBCT, translating its conceptual dimensions (relational, temporal, and emotional) into measurable and modelled signals for mental health prediction.

3.1. Digital behavioural continuum theory. The proposed digital behavioural continuum theory offers a conceptual lens through which silent indicators of mental health distress may be understood and operationalised. Building upon foundational concepts in social withdrawal theory (Rubin *et al*., 2009), digital phenotyping (Insel, 2017), and behavioural signal processing (Narayanan and Georgiou, 2013), DBCT assumes that changes in relational, emotional, and temporal behaviours in digital environments reflect underlying psychological states. These dimensions form a continuum of behavioural engagement, where deviations such as reduced posting frequency, delayed responses, or emotionally muted interactions signify potential withdrawal or distress. Therefore, DBCT posits that mental health states, such as depression and social withdrawal, manifest through patterns in three interconnected dimensions of digital behaviour:

1. *Relational dimension.* The relational dimension focuses on the structure and intensity of social connections and interactions within digital networks. It examines how users engage with their social circles through likes, comments, shares, and direct interactions to capture levels of social engagement or isolation. Relational features provide a structured understanding of users' social connections and interactions within a network. These features are derived from Graph Neural Networks (GNN) embeddings, capturing the dynamics and influence of a user's position in their social graph.

   - *Node degree (d)*: represents the total number of direct connections a user has, such as friends or followers. A high node degree indicates a well-connected user, while a low degree may signify social isolation.
   - *Clustering coefficient (c)*: measures the density of connections in a user's local network. It reflects how closely a user's connections are interconnected, offering insights into the cohesion of their social environment.
   - *Betweeness centality (b)*: highlights a user's influence within the network by calculating how often they act as a bridge between other nodes. Users with high betweenness centrality can significantly affect the flow of information within their network.

   The aggregate relational feature (R) is computed as

   $$R = g(d, c, b), \tag{1}$$

   where $g(\cdot)$ represents a weighted combination of these metrics given as $R = w_d d + w_c c + w_b b$, with w_d,w_c,w_b being adaptive learnable weights based on user-specific network dynamics optimised using a GNN. In this way, the social features which matter most for detecting withdrawal are learnt. Together, these relational features provide critical insights into the user's level of social engagement and influence, contributing to the model's ability to detect signs of social withdrawal.

2. *Temporal dimension.* This dimension captures time-dependent patterns, such as posting frequency,

response latency, and behavioural trends over time. It works with the assumption that changes in these patterns may reflect shifts in mental health states. These features are modelled using RNNs to analyse sequential data.

- *Average response latency (L)*: represents the average delay in a user's responses to interactions, such as replying to comments or liking posts. Longer response times can signal reduced engagement or emotional withdrawal.
- *Variance in posting frequency (σ_p^2)*: reflects the consistency or irregularity in a user's activity. High variance may suggest instability in behaviour, while low variance indicates steadiness.
- *Abrupt changes in interaction rates (Δ)*: measures sudden increases or decreases in user engagement, such as sharp drops in the number of comments or likes received. These shifts often correlate with changes in psychological well-being.

The aggregate temporal feature (T) is computed as

$$T = h(L, \sigma_p^2, \Delta), \tag{2}$$

where $h(\cdot)$ is a temporal aggregation function. Temporal features enable the model to detect behavioural changes that unfold over time, providing early warnings of potential mental health issues.

3. *Emotional dimension.* This dimension encompasses the emotional tone of digital interactions, expressed through reactions (e.g., sad, angry) and the sentiment of textual content. Emotional indicators provide insight into a user's psychological and emotional state as expressed through social media content.
   - *Proportions of specific emotional reactions ($P_{\text{sadness}}, P_{\text{anger}}$)*: represent the relative frequency of specific emotional reactions, such as sadness or anger responses, compared to the total number of reactions. A higher proportion of negative reactions can signify emotional distress. Sadness and anger were selected based on their documented association with depressive or withdrawal-related content in prior studies (Carver and Harmon-Jones, 2009; Joormann and Stanton, 2016; Vidal-Ribas and Stringaris, 2021). Other emotions (e.g., fear, disgust) were excluded to prevent label dilution but may be considered in future extensions of the model.
   - *Sentiment scores (S)*: are computed using Valence Aware Dictionary and Sentiment Reasoner (VADER) techniques (Hutto and Gilbert, 2014) and measure the emotional polarity of textual content, ranging from negative to positive. It is indicative that negative sentiment scores are associated with depressive states. VADER was selected due to its superior performance on short-text formats and informal language typical of social media, as validated in prior comparative sentiment evaluations. While alternative lexicons such as NRC and AFINN provide broader emotion coverage, their performance on sentiment detection in social media contexts is generally lower than that of VADER (Ribeiro *et al*., 2016; He *et al*., 2022).

The aggregate emotional feature (E) is computed as

$$E = k(P_{\text{sadness}}, P_{\text{anger}}, S), \tag{3}$$

where $k(\cdot)$ is a weighted combination of these metrics. Emotional features provide a direct view into the user's psychological state, offering crucial insights into the affective dimension of mental health.

Hence, DBCT synthesises these dimensions into a unified model for predicting mental health states defined as follows:

$$M = f(\alpha R + \beta T + \gamma E + \alpha\beta RT + \beta\gamma TE + \alpha\gamma RE + \epsilon), \tag{4}$$

where

- M is the mental health state, modelled as a binary,
- R stands for rational features,
- T denotes temporal features,
- E stands for emotional features,
- α, β, γ are weights that determine the relative contribution of each dimension to the mental health state,
- RT, TE and RE capture higher-order dependencies between the behavioural dimensions,
- ϵ denotes a residual error or noise in the prediction (bias).

This equation illustrates how the dimensions interact dynamically to provide a holistic view of digital behaviour and its relationship with mental health. DBCT integrates these dimensions into a unified theoretical model, asserting that their combined analysis offers a holistic view of digital behaviour and its correlation with mental

health. This theory provides the grounding for hybrid model design, where each subsystem targets a distinct behavioural lens and the fusion reflects their interplay. By grounding the model in DBCT, the framework transitions beyond technical integration to a theoretically-informed system for decoding digital mental health symptoms—the foundation for designing the computational methodology, as described in subsequent sections.

3.2. Dataset collection. The dataset was constructed by integrating two publicly available Kaggle sources: (1) a behavioural dataset containing social media engagement metrics (Isik, 2023) and (2) a textual corpus of user-generated posts labelled by psychological status (Namdari, 2023). The behavioural dataset included post-level interaction features such as reaction counts (e.g., likes, sads, angry), comments, shares, and content types (e.g., status, photo, video). The mental health textual comments were labelled into two classes: at-risk (1) and healthy (0), based on linguistic cues related to psychological distress (e.g., anxiety, depression). These labels provided the target variable for binary classification.

A unified sample of 2,680 records was created via stratified random sampling, preserving the class distribution of the original labelled corpus. Although the datasets lacked common user identifiers, integration was performed at the feature level, resulting in synthetic multimodal records that combine behavioural and linguistic signals. This design enabled joint modelling of relational, temporal, and emotional dimensions within the DBCT framework. While the data are not clinically diagnosed, they reflect real-world social media behaviour, supporting scalable, naturalistic analysis of early mental health risk indicators.

Many other features were derived from the original dataset during the feature engineering phase. These include the following:

- *sad reaction ratio (num_sads)*: computed as the proportion of num_sads reactions out of total reactions, emphasising emotional expression;
- *response latency*: derived from timestamps, representing the average delay in engagement (e.g., comment or like response times);
- *posting frequency trends*: posting frequency was computed daily but also aggregated into weekly intervals to capture both short- and medium-term behavioural trends;
- *node degree and clustering coefficient*: extracted from the interaction graph using graph analysis techniques to quantify users' social context and isolation;
- *sentiment scores*: computed from textual data using VADER to measure the polarity of user-generated content.

3.3. Data preprocessing. In this study, silent symptoms are operationalised as behavioural patterns and linguistic markers that deviate from a user's typical digital engagement without direct self-disclosure. These include consistent reductions in posting frequency, increased response latency, increased usage of negative affect words, and emotional withdrawal (e.g., higher sad-to-total reaction ratios). Several preprocessing steps were carefully undertaken to ensure the dataset was of high quality and suitable for use in ML models. First, incomplete or irrelevant records were systematically identified and removed from the dataset to ensure accuracy and relevance. Metrics representing user activity levels, such as engagement counts, were normalised to reduce variability. This step was critical for creating a consistent dataset, allowing fair comparisons across users with differing activity levels. Then, to model relationships and interactions between users, a user interaction graph was constructed, where each node represents an individual user. Edges between nodes signified interactions, such as comments or shares between users. These edges were further enhanced by assigning weights, calculated from engagement metrics like num_comments and num_shares, to represent the strength of each connection.

Nodes were enriched with attributes, such as num_reactions, to provide user-specific contextual data. This graph-based representation allowed understanding user interactions and their significance. Recognising the importance of temporal patterns, the data was further transformed into a time-series format. Metrics such as num_likes and num_comments were segmented into weekly intervals, creating structured chronological sequences. This segmentation facilitated the analysis of trends over time and prepared the dataset for temporal modelling, enabling insights into how user behaviour evolved. Also, emotional features such as reaction proportions and sentiment scores were extracted and formatted for integration into the hybrid model. By combining these preprocessing steps, i.e., cleaning, graph construction, and time-series segmentation, the dataset was transformed into a structured and multifaceted format. This approach ensured that key data types like engagement metrics (num_likes, num_comments), node attributes (num_reactions), and edge weights (num_comments, num_shares) were meticulously preserved and utilised, creating a robust foundation for ML applications.

3.4. Graph neural network implementation. To analyse relational behaviours, a GNN (Li *et al.*,

2024) was implemented. Users were represented as nodes, and their interactions formed edges with weights reflecting engagement levels. A 2-layer graph convolutional network was employed to generate node embeddings, capturing each user's relational patterns within the network. The model's hyperparameters included a learning rate of 0.01, a dropout rate of 0.5 to prevent overfitting, and the Adam optimiser for efficient convergence. In addition to Graph Convolutional Network (GCN)-derived embeddings, several classical graph-theoretic features were computed to enrich the relational representation. These included the following:

- *node degree* $d_i = \sum_j A_{IJ}$, where A is the weighted adjacency matrix;
- *clustering coefficient* $C_i = \frac{2T_i}{k_i(k_i-1)}$, where T_i is the number of closed triplets through node i and k_i the degree;
- *betweenness centrality* $b = \sum_{s\neq i\neq t} \frac{\sigma_{st}(i)}{\sigma_{st}}$, where σ_{st} is the number of shortest paths from node s to t_i and $\sigma_{st}(i)$ are those passing through i. Betweenness values were normalised by $\frac{(n-1)(n-2)}{2}$ for comparability across nodes within the network.

All relational features, including the node degree, the clustering coefficient, and betweenness centrality, were computed using weighted edges (Opsahl and Panzarasa, 2009) to reflect the intensity of interactions such as likes, comments, and shares. The final relational embedding was constructed using a weighted combination $R = g(d, c, b) = \alpha d + \beta c + \gamma b$, where α, β, γ was learnt during training.

This fused relational embedding was concatenated with the GCN output to form a comprehensive relational feature set representing both structural and intensity-based aspects of social connectivity, which serves as input for the hybrid model. This approach ensured that both the structure and strength of social ties were captured in the relational embeddings. These measures ensured that both structural and intensity-based aspects of social connectivity were accurately represented.

3.5. Recurrent neural network implementation. Temporal behaviours were modelled using a 2-layer long short-term memory (LSTM) network (Malhotra *et al.*, 2015). Time-series data, including daily posting frequency and response latency, was fed into the Recurrent Neural Network (RNN) to capture sequential dependencies and behavioural trends over time. Response latency L was defined as the average time (in minutes) between a user's post and the first received reaction. Posting frequency f_t was calculated as the mean number of posts per user per day, aggregated into weekly trends. Abrupt changes in engagement Δ_t were computed as $\Delta_t = |f_t - f_{t-1|}$, indicating behavioural volatility. Hyperparameter tuning was conducted using grid search over validation loss. Parameters included hidden units ϵ 64,128, a learning rate ϵ 0.001,0.005 and a dropout rate ϵ 0.3,0.5. The RNN's outputs provided temporal features (T), which were integrated into the hybrid model. The LSTM architecture was chosen over a basic RNN due to its ability to model long-range dependencies and handle sparse, irregular sequences, which are common in social media data. Response latency (measured in minutes between a user's post and first interaction) and posting frequency (averaged daily and aggregated weekly) were selected as representative indicators of user engagement. This design supports the capture of both short-term and long-term behavioural dynamics, which are critical for detecting digital withdrawal or fluctuations in interaction patterns.

3.6. Hybrid model integration. To operationalise the DBCT framework, a hybrid deep learning model was developed to integrate relational (R), temporal (T), and emotional (E) dimensions. As shown in Fig. 1, the architecture consists of three specialised branches: a graph neural network for social connectivity, a long short-term memory network for temporal behaviour modelling, and an emotional module based on reaction and sentiment features. To ensure consistency across modalities, outputs from all three modules, including emotional features, were passed through a linear projection layer to obtain a shared 64-dimensional latent representation. Thus, relational embeddings (64-d), temporal embeddings (initially 128-d and reduced to 64-d), and emotional features (projected to 64-d) all occupy a uniform vector space before fusion. The projected representations were concatenated into a unified multimodal feature vector (192-d), which was passed through a fully connected layer with ReLU activation, followed by a dropout layer (rate $= 0.4$), and finalised with a softmax classifier for binary prediction of mental health status The model in Fig. 1 was trained using binary cross-entropy loss and optimised via the Adam optimiser. To ensure balanced contribution across modalities, feature representations were normalised and jointly fine-tuned during training. This multimodal approach enabled the model to learn cross-dimensional patterns indicative of digital withdrawal and social disengagement. By synthesising the outputs of relational, temporal, and emotional subsystems, the architecture embodies the DBCT's principles and offers a scalable, interpretable framework for mental health prediction in online contexts.

Importantly, the architecture implicitly captures higher-order dependencies, interactions across relational, temporal, and emotional features that may co-vary. While not explicitly modelled using cross-feature terms or attention mechanisms, the fusion of GNN embeddings

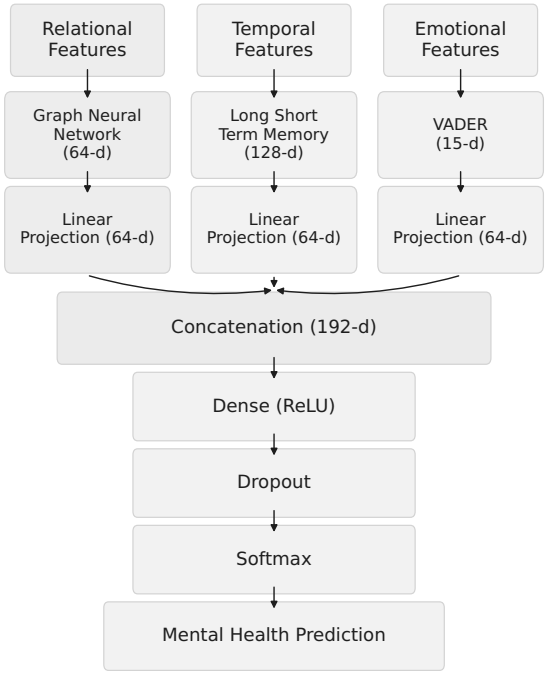

Fig. 1. Multimodal feature fusion architecture for mental health prediction.

(R), LSTM outputs (T), and emotional metrics (E) into a shared latent space enables the learning of non-linear relationships across modalities. This facilitates richer pattern recognition, allowing the model to detect complex digital behaviours reflective of mental health states. Future research will, among others, explore explicit modelling of these interactions through mechanisms like tensor fusion or attention-based architectures to further enhance interpretability and predictive performance.

4. Model evaluation

The model was trained on a labelled dataset using a stratified 5-fold cross-validation strategy to ensure robustness and generalisability across different data partitions. Each fold comprised 80% training, 10% validation, and 10% testing subsets. The Adam optimiser was used to fine-tune model parameters and accelerate convergence. To further ensure that the results were not biased due to between-user variability, we conducted an additional leave-one-user-out cross-validation (LOUO-CV) evaluation. In this setting, each user's entire data was held out for testing in turn, while the model was trained on the remaining users. This protocol ensured that no user appeared in both training and test sets, providing a more rigorous assessment of the model's generalisability to unseen users. Performance metrics, including accuracy, precision, recall, and F1-score, were averaged across all user-based folds and are reported in Tables 1 and 3. The SHAP analysis was integrated at two stages of the pipeline. Initially, a preliminary model was used to identify and eliminate low-impact features, reducing noise and computational overhead. Feature selection was conducted independently within each training fold to prevent data leakage. SHAP was then recomputed on the final hybrid model post-fusion, with attributions mapped back to their original modalities; relational, temporal, and emotional. This interpretability analysis not only reinforced the theoretical foundations of DBCT by highlighting the predictive value of relational and temporal dynamics but also enabled clear visualisation of feature contributions. SHAP beeswarm and violin plots were included to summarise global importance and modality-specific effects. Following model optimisation, predictive performance was evaluated using standard classification metrics, i.e., accuracy, precision, recall, and F1-score, to provide a comprehensive and balanced assessment of model effectiveness.

The model was evaluated under three experimental settings: (1) a baseline comparison with logistic regression to provide a foundational benchmark using a conventional, non-specialised classifier, (2) an ablation study in which the relational (GNN), temporal (RNN), or emotional components were systematically removed from the hybrid architecture to assess their individual contributions to model performance, and (3) benchmarking against state-of-the-art models from recent literature, with all models evaluated on the same dataset using identical 5-fold cross-validation splits, feature sets, and evaluation metrics to ensure fair and consistent comparison.

4.1. Baseline analysis. First, the predictive strength of the proposed hybrid model, as presented in Table 1, can be partially attributed to the extracted emotional features. To ensure the reliability of these affective signals, VADER sentiment scores were validated against a subset of human-annotated sentiment labels (15% of the dataset). The comparison yielded high correspondence, with an accuracy of 87.2%, a precision of 85.4%, a recall of 88.1%, and an F1-score of 86.7%. Furthermore, Cohen's Kappa coefficient was calculated at 0.81, indicating substantial agreement between automated and human sentiment assessments. These results provide empirical justification for the use of VADER as a valid tool for affective tone extraction in social media contexts. The alignment between sentiment validity and model performance further underscores the contribution of emotional features to the overall predictive capacity of the hybrid architecture. Baseline models were incorporated to establish a foundation for evaluating the

Table 1. Model performance outcome.

Model	Accuracy	Precision	Recall	F1-score
Logistic regression	81.5%	79.8%	77.2%	78.5%
Standalone GNN	86.7%	84.2%	85.6%	84.9%
Standalone RNN	87.3%	85.4%	86.1%	85.8%
Hybrid model (GNN+RNN)	92.4%	89.7%	91.2%	90.4%

Table 2. Hyperparameter tuning results.

Model	Parameter	Default value	Tuned value	Impact
GNN	Learning rate	0.001	0.01	Improved convergence
GNN	Dropout	0.3	0.5	Reduced overfitting
RNN	Hidden units	64	128	Enhanced temporal modeling
RNN	Sequence length	30 days	14 days	Better computational efficiency

hybrid model's performance. These methods, including logistic regression, a standalone GNN, and a standalone RNN, allowed an independent analysis of relational, temporal, and emotional dimensions. Specifically, the logistic regression model was trained on the full feature set without structured relational or sequential learning, serving as a naive baseline. The standalone GNN was used to model relational behaviours alone (R), while the standalone RNN focused solely on temporal patterns (T). This modular setup enabled a direct comparison of each behavioural modality's predictive contribution. While baseline models achieved moderate success, their inability to integrate multiple dimensions highlighted the necessity of the hybrid approach.

For example, the standalone GNN effectively captured relational features, achieving an F1-score of 84.9%, while the standalone RNN excelled in modelling temporal trends, reaching an F1-score of 85.8%. However, these models failed to account for the interaction between dimensions, such as how social isolation (relational) and declining activity (temporal) combine to signal depressive symptoms. By integrating these dimensions, the hybrid model achieved an F1-score of 90.4%, demonstrating its ability to operationalise DBCT and provide actionable insights into mental health states. The impact of hyperparameter optimisation on these models is further detailed in Table 2.

Overall, tuning improved the F1-score from 84.2% (default hyperparameters) to 90.4% (optimised hyperparameters). To assess generalisability to entirely new users, we applied LOUO-CV. This user-level evaluation reflects real-world deployment, where the model must predict on individuals not seen during training. As expected, performance was slightly lower than in the 5-fold CV due to the stricter generalisation requirement. Nonetheless, the hybrid model maintained a strong F1-score of 85.5%, confirming its robustness across diverse user profiles.

As shown in Table 3, all models experienced a moderate performance drop under LOUO-CV due to increased user-level variability and reduced training data per fold. Logistic regression yielded the lowest F1-score of 73.3%, reaffirming its limited capacity to model complex behavioural interactions. Both the standalone GNN and RNN models remained reasonably okay, achieving F1-scores of 80.8% and 81.6%, respectively. Importantly, the proposed hybrid model maintained a strong F1-score of 85.5%, i.e., a marginal 4.9% decrease from its 5-fold score (90.4%). This stability reinforces the model's ability to generalise to previously unseen users by leveraging the complementary strengths of relational, temporal, and emotional cues. Such consistency validates the architectural integrity of DBCT and supports its practical applicability in detecting silent mental health patterns across diverse user profiles.

4.2. Benchmarking against related models. To further validate the reliability and generalisability of the proposed hybrid model, we benchmarked its performance not only against baseline standalone models but also against foundational, recent and prominent studies from the literature that have addressed mental health prediction using social media data. These selected studies span a range of methodologies, including transformer-based models, multimodal architectures, and hybrid learning frameworks, ensuring a comprehensive and up-to-date evaluation. The comparative studies included are the following:

- De Choudhury *et al.* (2013): applied logistic regression on linguistic and behavioural features to predict depressive symptoms;
- Bokolo and Liu (2024): evaluated transformer models (e.g., RoBERTa, DeBERTa) for depression and suicide detection in tweets, showing their superiority over traditional ML classifiers;

Table 3. Model performance outcome under LOUO-CV.

Model	Accuracy	Precision	Recall	F1-score
Logistic regression	77.3%	74.5%	72.1%	73.3%
Standalone GNN	82.6%	80.4%	81.2%	80.8%
Standalone RNN	83.4%	81.3%	82.0%	81.6%
Hybrid model (GNN+ RNN)	87.9%	85.1%	86.0%	85.5%

Table 4. Benchmark performance comparison and statistical significance.

Model	Mean F1-score (%)	Precision (%)	Recall(%)	p-Value (vs. hybrid model)
De Choudhury *et al*. (2013)	82.1	80.3	84.0	0.021
Bokolo and Liu (2024)	88.9	88.0	89.5	0.009
Khan and Ali (2024)	86.5	85.7	87.1	0.015
Bucur *et al*. (2023)	88.2	87.4	88.9	0.008
Pourkeyvan *et al*. (2024)	87.6	86.8	89.9	0.008
Ilias *et al*. (2023)	87.6	86.8	88.3	0.013
Proposed DBCT	90.4	89.7	91.2	NA

- Khan and Ali (2024): conducted an extensive review of ML-based detection of mental health disorders using online social media, identifying the most promising techniques and challenges;
- Bucur *et al*. (2023): introduced a time-enriched multimodal transformer model incorporating temporal features for state-of-the-art depression detection;
- Pourkeyvan *et al*. (2024): demonstrated the predictive power of Hugging Face transformer models for mental disorder classification, achieving high accuracy with minimal feature engineering;
- Ilias *et al*. (2023): focused on transformer model calibration, integrating linguistic features to enhance depression and stress detection in social media posts.

Each benchmark model was simulated and evaluated on the study dataset using the same 5-fold cross-validation splits, feature sets, and evaluation metrics as the proposed model, to ensure consistency and a valid comparison. While original training configurations (e.g., learning rate, architecture) were preserved where applicable, minor adjustments (e.g., input size, batch size) were made only when necessary to ensure compatibility with the dataset and computational constraints. This simulation approach ensured consistency, parameter fidelity, and valid comparison across all benchmarked models. The proposed hybrid model, which integrates graph neural networks for relational features, recurrent neural networks for temporal dynamics, and natural language processing for emotional analysis, achieved a mean F1-score of 90.4%, outperforming all selected benchmarks. The benchmarked model performance and its statistical significance are presented in Table 4.

To assess the significance of these performance differences, paired t-tests across five evaluation runs were conducted. Results confirmed that the performance improvements of the proposed hybrid model over each benchmarked model were statistically significant, with p-values < 0.01 in all pairwise comparisons. These results affirm that the integration of relational, temporal, and emotional dimensions through the DBCT framework provides meaningful and measurable advantages over prior models.

4.3. Feature importance analysis. To interpret the hybrid model's predictions, Shapley additive explanations, an explainable AI technique, was employed to assess the relative contributions of relational, temporal, and emotional features. This analysis provided critical insights into which characteristics most significantly influenced the prediction of mental health states. To preserve interpretability across modalities. The SHAP analysis was first applied to the pre-fusion feature sets, enabling direct attribution to their respective categories: relational, temporal, and emotional. This allowed semantic clarity in identifying which dimensions influenced predictions the most. To validate consistency, SHAP was also recomputed post-fusion. Feature contributions were mapped back to their original inputs via index alignment, ensuring that interpretability was retained despite vector concatenation.

Figure 2 illustrates the SHAP values for the top 10 features, showcasing their relative contributions to the model's predictions. Emotional features like num_sads and temporal ones such as response latency were the most impactful, followed by relational metrics like the clustering coefficient and the node degree.

The SHAP analysis offers strong validation for

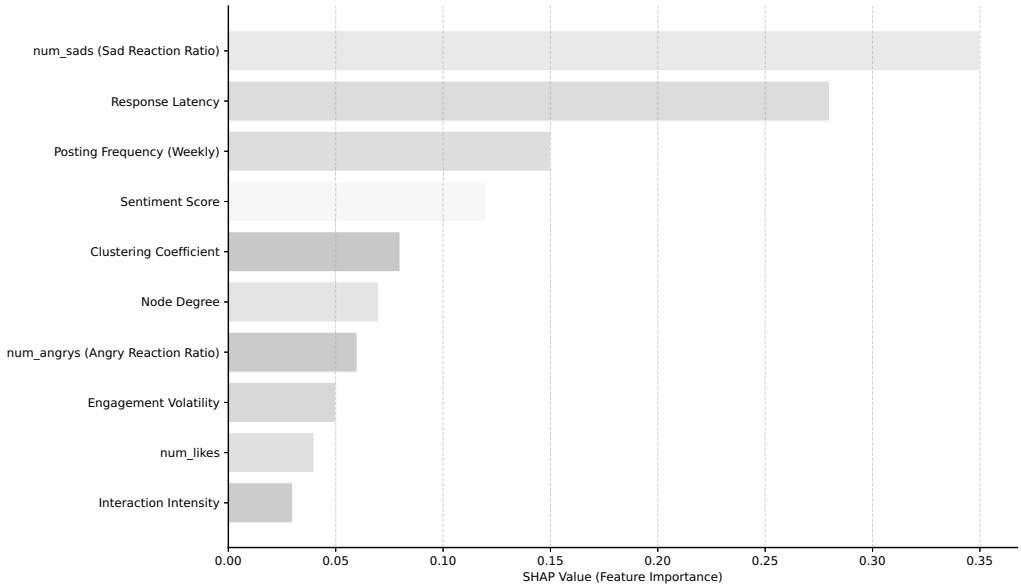

Fig. 2. SHAP analysis of important features.

the theoretical foundations of digital behavioural continuum theory. It highlights the critical role of emotional dimensions in predicting mental health states, emphasising the significance of user affect in identifying depressive behaviours. Features such as the sad reaction ratio and sentiment scores accentuate how expressions of emotion on social media serve as key indicators of psychological well-being. First, temporal dynamics further enhance this understanding by capturing changes in engagement patterns over time. Declining activity levels or increased delays in responding to interactions emerge as early signals of potential behavioural shifts, providing a temporal lens into users' evolving mental states. Then, relational metrics, including measures like the node degree and the clustering coefficient, complement these insights by situating individual behaviours within the broader context of SNs.

These features provide valuable context for understanding how social connectivity, or the lack thereof, intersects with emotional and temporal patterns. Together, these findings validate the hybrid model's ability to integrate relational, temporal, and emotional dimensions into a cohesive framework. This not only strengthens the model's predictive accuracy but also underscores the necessity of combining these dimensions to provide a holistic understanding of mental health states. The analysis ultimately demonstrates the practical and theoretical value of DBCT in advancing mental health research and interventions. Similarly, Fig. 3 enriches this view by illustrating the distribution of SHAP values per feature and their modality-specific contributions, offering insight into how each input influences the model.

Figure 3 presents a SHAP beeswarm plot illustrating the global impact of individual features across the relational, temporal, and emotional modalities. Notably, emotional features, particularly the sadness ratio, the anger ratio, and the sentiment score, emerge as the most influential in shaping the model's output, with higher SHAP values indicating a stronger association with the predicted "at-risk" class. Temporal indicators such as response latency, engagement volatility, and posting frequency also show substantial predictive relevance, affirming their role as behavioural proxies for social withdrawal, cognitive fatigue, and fluctuating engagement. Among relational features, the clustering coefficient, the node degree, and interaction intensity exhibit moderate but consistent contributions, underscoring the influence of social connectivity and user positioning within the network. This visualisation supports the multi-modal architecture of the hybrid model and empirically validates the theoretical contribution of each DBCT dimension to predictive performance. To further contextualise these feature-level insights, Fig. 4 presents the aggregated SHAP value distributions across the three DBCT modalities, revealing the relative dominance and variability of each feature group.

Figure 4 shows a violin plot of SHAP value distributions across the three DBCT modalities: relational, temporal, and emotional. The results indicate that emotional features exhibit the greatest variability and impact on model predictions, with a broader and more dispersed distribution of SHAP values. This underscores their dominant role in distinguishing "at-risk" from "healthy" users, consistent with the theoretical emphasis on affective signals in mental health detection. Relational features show moderate contribution, reflecting the

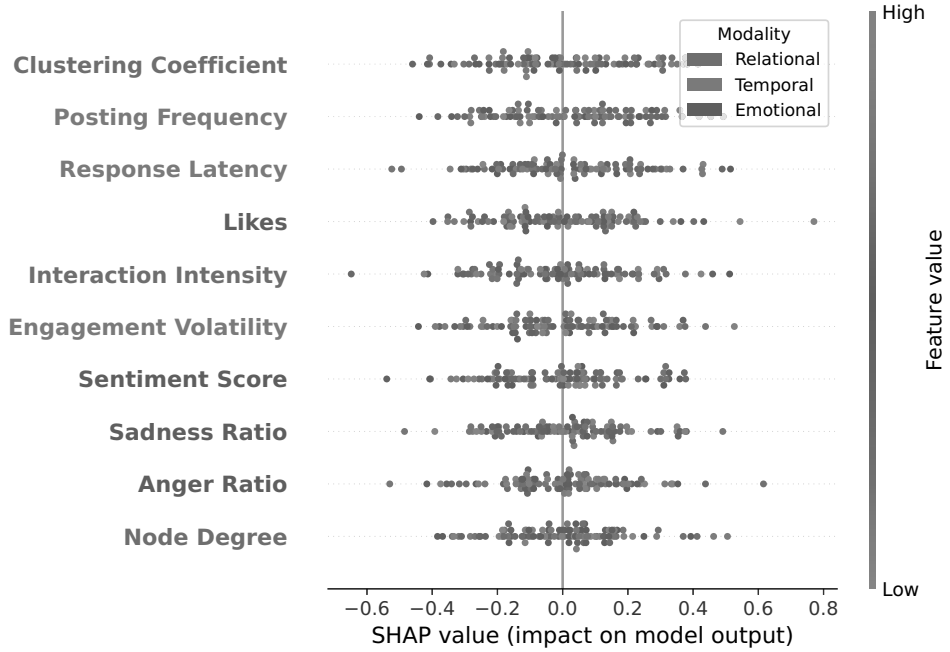

Fig. 3. SHAP beeswarm plot with modality labels.

influence of social connectivity metrics such as the node degree and the clustering coefficient. Temporal features display the narrowest and most centred distribution, suggesting a more stable but less discriminative role. Collectively, these distributions validate the hybrid model's multi-dimensional structure and highlight the complementary roles of emotional, relational, and temporal cues in predicting mental health outcomes.

5. Key findings and discussions

This research yields significant findings across theoretical, managerial, and policy dimensions, demonstrating the value of DBCT and its practical applications. From a theoretical perspective, the study validates the DBCT framework by showcasing how relational, temporal, and emotional dimensions collectively capture digital behaviours indicative of mental health states. The integration of these dimensions provides a more comprehensive understanding than analysing them in isolation. Emotional features, such as the sad reaction ratio and sentiment scores, emerged as the most impactful predictors, highlighting the central role of affective signals in identifying depressive symptoms. Temporal dynamics, including response latency and posting frequency trends, provide critical early signals of behavioural shifts, while relational metrics like the node degree and the clustering coefficient complement these insights by contextualising behaviours within users' SNs. The study also formalises the mathematical framework for mental health states, introducing an equation that quantifies the contributions of these dimensions and demonstrating the model's capability to identify signs of mental health decline weeks before clinical onset.

5.1. Benchmarking existing methodologies. Compared to prior studies such as those by Bokolo and Liu (2024) or Bucur *et al*. (2023), which focus on transformer-based text classification and time-enriched multimodal architectures, respectively, the proposed hybrid model achieved a higher F1-score (90.4%) and offered more transparency through explainability techniques like SHAP. While these studies showed strong performance on text-heavy tasks (with F1 scores ranging from 88.2% to 88.9%), they lacked relational modelling or comprehensive interpretability frameworks. In contrast, the integration of GNNs enabled the model to factor in users' social positions (e.g., node degree, clustering), which proved valuable for detecting social withdrawal, an area underexplored in the cited benchmarks. In addition, the finding that emotional features such as 'sad' reactions and sentiment scores were the most influential predictors aligns with earlier work by De Choudhury *et al*. (2013), who reported the predictive power of negative affect in text. However, the inclusion of temporal indicators, like response latency and interaction variance, adds a behavioural dimension rarely emphasised in those works. This reinforces the hypothesis that declining engagement and delayed interaction are early digital distress markers. Notably, the study by Pourkeyvan *et al*. (2024) highlighted

Table 5. Key behavioural features by dimension with descriptions.

Dimension	Feature	Description
Emotional	Angry reaction ratio	Proportion of angry reactions out of total reactions, reflecting emotional agitation or frustration.
	Sad reaction ration	Proportion of sad reactions out of total reactions, indicating emotional negativity.
	Sentiment score	Polarity of textual posts, reflecting the affective tone of user content.
Temporal	Response latency	Average time delay between receiving and responding to interactions, signalling engagement levels.
	Posting frequency	Captures temporal patterns in user engagement by tracking the number and regularity of posts over time, which may reflect behavioural shifts related to mental health states.
Relational	Engagement volatility	Degree of fluctuation in user activity over time (e.g., posting, reacting, commenting), indicating behavioural consistency or instability.
	Node degree	Number of connections (weighted interactions) a user has in their interaction network.
	Clustering coefficient	Measure of network density, providing insights into users' social cohesion or isolation.
	Betweenness centrality	Measure of network density, providing insights into users' social cohesion or isolation.

the strong performance of Hugging Face transformer models in predicting mental health outcomes. Although those models achieved high accuracy, they relied primarily on text-based features and did not incorporate behavioural or network-related dynamics. In contrast, the proposed model demonstrates the value of integrating multi-dimensional signals, offering a richer behavioural profile that supports early and accurate detection.

5.2. Social and managerial implications. Socially, these results support the notion that behavioural cues on social media reflect not just momentary emotions but ongoing psychological states. The implications are significant: platforms could integrate such models to flag at-risk users in real-time, enabling early interventions. These findings also echo the concerns and recommendations by Khan and Ali (2024), who emphasised the importance of ethically aligned, real-time monitoring frameworks. Managerially, the findings offer a roadmap for implementing predictive systems that enable real-time monitoring of user behaviours on digital platforms. Organisations can leverage these insights to develop tools for identifying at-risk individuals, tailoring interventions, and enhancing user engagement strategies. For instance, monitoring temporal trends like sudden declines in activity or increased emotional negativity can trigger proactive support measures. Also, social media platforms and mental health services can utilise this framework to understand user behaviours more holistically and design platforms that promote positive interactions while minimising emotionally harmful experiences. The hybrid model presented in this study serves as a prototype for leveraging cutting-edge ML techniques, such as GNNs and RNNs, for behavioural analytics, offering managers sophisticated tools to process multi-dimensional data effectively.

5.3. Policy implications. From a policy standpoint, the research underscores the importance of ethical considerations in using digital behavioural data. It highlights the need for strict guidelines to anonymise and secure user data, ensuring privacy and responsible application of predictive models. Policymakers can adopt these insights to establish ethical standards for mental health data collection and processing. In addition, the framework provides an opportunity for governments and health organisations to design data-driven policies for early mental health interventions, emphasising proactive strategies over reactive measures. For example, alerts generated by predictive systems could prompt healthcare providers to reach out to individuals displaying early signs of social withdrawal or depression. Furthermore, social media platforms can be encouraged or mandated to adopt predictive models for monitoring user well-being, ensuring that they prioritise user safety and emotional health. The integration of such systems into public health programs can also aid in monitoring population-wide behavioural trends and allocating mental health resources more effectively. This research advances theoretical understanding through DBCT, offers actionable insights

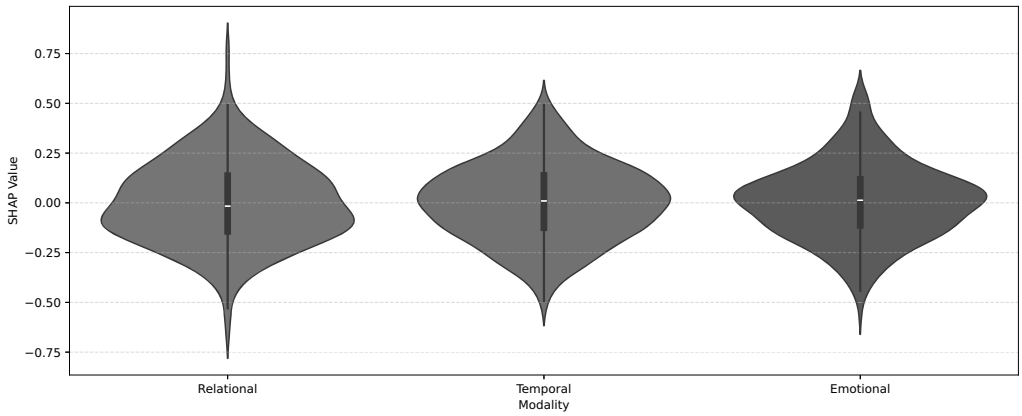

Fig. 4. SHAP value distribution by feature modality.

for managers in designing empathetic, user-centred platforms, and informs policymakers on ethical and scalable approaches for leveraging digital behaviour analysis. By bridging theory and application, it provides a framework for addressing mental health challenges in the digital age, with significant implications for societal well-being.

6. Conclusions and future directions

This research presented a novel approach to understanding mental health states through digital behaviour, introducing DBCT as a comprehensive framework for analysing social media interaction patterns. By integrating relational, temporal, and emotional dimensions, the study demonstrated how digital traces can reveal silent signals of depression and social withdrawal. Importantly, these signals are derived from non-clinical, real-world social media data rather than formally diagnosed patient records, reflecting naturally occurring online behaviour. The findings validate the necessity for a multi-dimensional approach, as emotional expressions, behavioural trends, and social connectivity interact dynamically to reflect mental health conditions. The implementation of a hybrid ML model, combining GNNs for relational analysis, RNNs for temporal pattern recognition, and NLP for emotional sentiment extraction, enables a scalable and data-driven methodology for predicting mental health states. The model achieved high predictive accuracy, with an F1-score of 90.4% and an overall accuracy of 92.4%, demonstrating its effectiveness in identifying at-risk behavioural patterns within socially generated, non-clinical datasets intended to approximate real-world social media use.

From a theoretical standpoint, this research advances the field by operationalising DBCT, bridging the gap between behavioural science and computational modelling. The study confirms that emotional signals, particularly sad reaction ratios and sentiment scores, are the strongest indicators of mental health distress. Temporal indicators, such as response latency and posting frequency variations, provide early warning signs, while relational metrics, such as network centrality and clustering coefficients, offer a crucial context for understanding social isolation. These findings should be interpreted as indicators of potential mental health vulnerability rather than a clinical diagnosis. Beyond theoretical contributions, this research has significant managerial and policy implications. Social media platforms, mental health organisations, and technology firms can leverage these findings to develop real-time monitoring tools, enabling early intervention strategies to support individuals exhibiting concerning behavioural shifts. Policymakers must also consider the ethical use of social media data, ensuring that privacy, security, and consent remain at the forefront of mental health prediction models.

Despite its contributions, this study has several limitations. First, the dataset comes from publicly available social media interactions, which omit private or offline behaviour and may not reflect clinically verified diagnoses, limiting both generalisability and medical interpretation. Second, sentiment analysis using VADER provides valuable insights but may not always capture sarcasm, cultural context, or other emotional expressions. Third, while the hybrid model achieves high accuracy, it still relies on labelled mental health datasets, which may introduce bias in classification due to self-reporting limitations.

Future work will focus on expanding the model's applicability to diverse populations and platforms, ensuring that behavioural insights remain valid across different digital ecosystems. Further refinement of real-time intervention strategies will be crucial in translating predictive insights into actionable mental health support mechanisms. In addition, advancements in

explainable AI (XAI) will be integrated into the model to ensure transparency and interpretability, allowing stakeholders such as mental health professionals to trust and act upon model predictions. While SHAP offers valuable global feature importance estimates, it does not capture the temporal ordering inherent in LSTM outputs. This limitation arises when applying SHAP to fused feature vectors or non-sequential model components. Future work would explore explainability methods better aligned with sequential architectures, such as integrated gradients or attention-based saliency mechanisms, to enhance the interpretability of temporal dynamics. In conclusion, this study underscores the potential of digital behaviour analysis as a proactive tool for mental health monitoring. While this paper does not aim to replace clinical assessment, by leveraging the vast data available on social media platforms, DBCT provides a foundation for ethically responsible and scientifically sound mental health prediction models.

References

Adeyanju, G.C., Solfa, R.P., Tran, T.L., Wohlfarth, S., Büttner, J., Osobajo, O.A. and Otitoju, A. (2021). Behavioural symptoms of mental health disorder such as depression among young people using Instagram: A systematic review, *Translational Medicine Communications* **6**(15): 1–13.

Arowosegbe, A. and Oyelade, T. (2023). Application of natural language processing (NLP) in detecting and preventing suicide ideation: A systematic review, *International Journal of Environmental Research and Public Health* **20**(2): 1514.

Ball, H., Eisner, E., Nicholas, J., Wilson, P. and Bucci, S. (2025). How theories, models, and frameworks have been used to implement digital health interventions in services for people with severe mental health problems: A scoping review, *BMC Public Health* **25**(1): 1023.

Bokolo, B.G. and Liu, Q. (2024). Advanced comparative analysis of machine learning and transformer models for depression and suicide detection in social media texts, *Electronics* **13**(20): 3980.

Brailovskaia, J. and Margraf, J. (2022). The relationship between active and passive Facebook use, Facebook flow, depression symptoms and Facebook addiction: A three-month investigation, *Journal of Affective Disorders Reports* **10**: 100374.

Brailovskaia, J., Truskauskaite-Kuneviciene, I., Kazlauskas, E. and Margraf, J. (2021). The patterns of problematic social media use (SMU) and their relationship with online flow, life satisfaction, depression, anxiety and stress symptoms in Lithuania and in Germany, *Current Psychology* **42**(5): 3713–3724.

Brand, C., Fochesatto, C.F., Gaya, A.R., Schuch, F.B. and López-Gil, J.F. (2024). Scrolling through adolescence: Unveiling the relationship of the use of social networks and its addictive behavior with psychosocial health, *Child and Adolescent Psychiatry and Mental Health* **18**(1): 107.

Bucur, A.-M., Cosma, A., Rosso, P. and Dinu, L.P. (2023). It's just a matter of time: Detecting depression with time-enriched multimodal transformers, *European Conference on Information Retrieval, Dublin, Ireland,* pp. 200–215.

Burke, M., Marlow, C. and Lento, T. (2010). Social network activity and social well-being, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Atlanta, USA*, pp. 1909–1912.

Carver, C.S. and Harmon-Jones, E. (2009). Anger is an approach-related affect: Evidence and implications, *Psychological Bulletin* **135**(2): 183.

Chancellor, S. and De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: A critical review, *NPJ Digital Medicine* **3**(1): 43.

Chang, C.-W., Chang, C.-Y. and Lin, Y.-Y. (2022). A hybrid CNN and LSTM-based deep learning model for abnormal behavior detection, *Multimedia Tools and Applications* **81**(9): 11825–11843.

Chemnad, K., Aziz, M., Belhaouari, S.B. and Ali, R. (2023). The interplay between social media use and problematic internet usage: Four behavioral patterns, *Heliyon* **9**(5): e15745.

Chen, A., Fu, Y., Zheng, X. and Lu, G. (2022). An efficient network behavior anomaly detection using a hybrid DBN-LSTM network, *Computers & Security* **114**: 102600, DOI: 10.1016/j.cose.2021.102600.

Chen, J. and Wang, Y. (2021). Social media use for health purposes: Systematic review, *Journal of Medical Internet research* **23**(5): e17917.

Coyne, S.M., Rogers, A.A., Zurcher, J.D., Stockdale, L. and Booth, M. (2020). Does time spent using social media impact mental health?: An eight year longitudinal study, *Computers in Human Behavior* **104**: 106160.

De Choudhury, M., Gamon, M., Counts, S. and Horvitz, E. (2013). Predicting depression via social media, *Proceedings of the International AAAI Conference on Web and Social Media, Cambridge, USA,* Vol. 7, pp. 128–137.

Dhelim, S., Chen, L., Das, S.K., Ning, H., Nugent, C., Leavey, G., Pesch, D., Bantry-White, E. and Burns, D. (2023). Detecting mental distresses using social behavior analysis in the context of COVID-19: A survey, *ACM Computing Surveys* **55**(14s): 1–30.

Dong, G., Tang, M., Wang, Z., Gao, J., Guo, S., Cai, L., Gutierrez, R., Campbel, B., Barnes, L. E. and Boukhechba, M. (2023). Graph neural networks in IoT: A survey, *ACM Transactions on Sensor Networks* **19**(2): 1–50.

Fabris, M.A., Marengo, D., Longobardi, C. and Settanni, M. (2020). Investigating the links between fear of missing out, social media addiction, and emotional symptoms in adolescence: The role of stress associated with neglect and negative reactions on social media, *Addictive Behaviors* **106**: 106364.

Gauthier, G.R., Smith, J.A., García, C., Garcia, M.A. and Thomas, P.A. (2021). Exacerbating inequalities: Social networks, racial/ethnic disparities, and the COVID-19 pandemic in the United States, *The Journals of Gerontology: Series B* **76**(3): e88–e92.

Giuntini, F.T., De Moraes, K.L., Cazzolato, M.T., de Fátima Kirchner, L., Dos Reis, M.d.J.D., Traina, A.J., Campbell, A.T. and Ueyama, J. (2021). Tracing the emotional roadmap of depressive users on social media through sequential pattern mining, *IEEE Access* **9**: 97621–97635.

Guntuku, S.C., Yaden, D.B., Kern, M.L., Ungar, L.H. and Eichstaedt, J.C. (2017). Detecting depression and mental illness on social media: An integrative review, *Current Opinion in Behavioral Sciences* **18**: 43–49.

Haddad, J.M., Macenski, C., Mosier-Mills, A., Hibara, A., Kester, K., Schneider, M., Conrad, R.C. and Liu, C.H. (2021). The impact of social media on college mental health during the COVID-19 pandemic: A multinational review of the existing literature, *Current Psychiatry Reports* **23**(11): 1–12.

He, L., Yin, T. and Zheng, K. (2022). They may not work! An evaluation of eleven sentiment analysis tools on seven social media datasets, *Journal of Biomedical Informatics* **132**: 104142.

Huang, C. (2022). A meta-analysis of the problematic social media use and mental health, *International Journal of Social Psychiatry* **68**(1): 12–33.

Hutto, C. and Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text, *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media, Ann Arbor, USA,* Vol. 8, pp. 216–225.

Ibitoye, A.O., Famutimi, R.F., Olanloye, D.O. and Akioyamen, E. (2021). User centric social opinion and clinical behavioural model for depression detection, *International Journal of Intelligent Information Systems* **10**(4): 69–73.

Ibitoye, A.O.J., Oladimeji, O.O. and Onifade, O.F.W. (2025a). Contextual emotional transformer-based model for comment analysis in mental health case prediction, *Vietnam Journal of Computer Science* **12**(03): 277–299, DOI: 10.1142/S2196888824500192.

Ibitoye, A.O., Oladimeji, O.O. and Afe, O.F. (2025b). Clustering digital mental health perceptions using transformer-based models, *Franklin Open* **11**: 100262.

Ilias, L., Mouzakitis, S. and Askounis, D. (2023). Calibration of transformer-based models for identifying stress and depression in social media, *IEEE Transactions on Computational Social Systems* **11**(2): 1979–1990.

Insel, T.R. (2017). Digital phenotyping: Technology for a new science of behavior, *Jama* **318**(13): 1215–1216.

Isik, M. (2023). Social media engagement: A Comprehensive analysis, *Dataset*, `https://www.kaggle.com/datasets/mehmetisik/livedataset`.

Jackson, M.O. (2008). *Social and Economic Networks*, Vol. 3, Princeton University Press, Princeton.

Jin, M., Koh, H.Y., Wen, Q., Zambon, D., Alippi, C., Webb, G.I., King, I. and Pan, S. (2024). A survey on graph neural networks for time series: Forecasting, classification, imputation, and anomaly detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **46**(12): 10466–10485.

Joormann, J. and Stanton, C.H. (2016). Examining emotion regulation in depression: A review and future directions, *Behaviour Research and Therapy* **86**: 35–49.

Karim, F., Oyewande, A.A., Abdalla, L.F., Ehsanullah, R.C. and Khan, S. (2020). Social media use and its connection to mental health: A systematic review, *Cureus* **12**(6): e8627.

Khan, A. and Ali, R. (2024). Unraveling minds in the digital era: A review on mapping mental health disorders through machine learning techniques using online social media, *Social Network Analysis and Mining* **14**(1): 78.

Kim, J., Lee, D. and Park, E. (2021). Machine learning for mental health in social media: Bibliometric study, *Journal of Medical Internet Research* **23**(3): e24870.

Kim, S., Jang, Y.S. and Park, E.-C. (2025). Associations between social isolation, withdrawal, and depressive symptoms in young adults: A cross-sectional study, *BMC Psychiatry* **25**(1): 1–12.

Kmetty, Z. and Bozsonyi, K. (2022). Identifying depression-related behavior on Facebook—An experimental study, *Social Sciences* **11**(3): 135.

Kour, H. and Gupta, M.K. (2022). An hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM, *Multimedia Tools and Applications* **81**(17): 23649–23685.

Lazcano, A., Herrera, P.J. and Monge, M. (2023). A combined model based on recurrent neural networks and graph convolutional networks for financial time series forecasting, *Mathematics* **11**(1): 224.

Li, J., Wu, L., Du, Y., Hong, R. and Li, W. (2024). Dual graph neural networks for dynamic users' behavior prediction on social networking services, *IEEE Transactions on Computational Social Systems* **11**(5): 6131–6144.

Lin, J., Lin, S., Turel, O. and Xu, F. (2020). The buffering effect of flow experience on the relationship between overload and social media users' discontinuance intentions, *Telematics and Informatics* **49**: 101374.

Liu, S., Li, T., Ding, H., Tang, B., Wang, X., Chen, Q., Yan, J. and Zhou, Y. (2020). A hybrid method of recurrent neural network and graph neural network for next-period prescription prediction, *International Journal of Machine Learning and Cybernetics* **11**(11): 2849–2856.

Malhotra, P., Vig, L., Shroff, G. and Agarwal, P. (2015). Long short term memory networks for anomaly detection in time series, *in* M. Verleysen (Ed.), *Proceedings of the 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2015), Bruges, Belgium*, Vol. 89, pp. 89–94.

Marengo, D., Montag, C., Mignogna, A. and Settanni, M. (2022). Mining digital traces of Facebook activity for

the prediction of individual differences in tendencies toward social networks use disorder: A machine learning approach, *Frontiers in Psychology* **13**: 830120.

Meier, A. and Reinecke, L. (2021). Computer-mediated communication, social media, and mental health: A conceptual and empirical meta-review, *Communication Research* **48**(8): 1182–1209.

Merino, M., Tornero-Aguilera, J.F., Rubio-Zarapuz, A., Villanueva-Tobaldo, C.V., Martín-Rodríguez, A. and Clemente-Suárez, V.J. (2024). Body perceptions and psychological well-being: A review of the impact of social media and physical measurements on self-esteem and mental health with a focus on body image satisfaction and its relationship with cultural and gender factors, *Healthcare* **12**(14): 1396.

Mohr, D.C., Schueller, S.M., Montague, E., Burns, M.N. and Rashidi, P. (2014). The behavioral intervention technology model: An integrated conceptual and technological framework for eHealth and mHealth interventions, *Journal of Medical Internet Research* **16**(6): e146.

Moore, K. and Craciun, G. (2021). Fear of missing out and personality as predictors of social networking sites usage: The Instagram case, *Psychological Reports* **124**(4): 1761–1787.

Mummah, S.A., Robinson, T.N., King, A.C., Gardner, C.D. and Sutton, S. (2016). IDEAS (integrate, design, assess, and share): A framework and toolkit of strategies for the development of more effective digital interventions to change health behavior, *Journal of Medical Internet research* **18**(12): e317.

Munikoti, S., Agarwal, D., Das, L., Halappanavar, M. and Natarajan, B. (2023). Challenges and opportunities in deep reinforcement learning with graph neural networks: A comprehensive review of algorithms and applications, *IEEE Transactions on Neural Networks and Learning Systems* **35**(11): 15051–15071.

Murshed, B.A.H., Abawajy, J., Mallappa, S., Saif, M.A.N. and Al-Ariki, H.D.E. (2022). DEA-RNN: A hybrid deep learning approach for cyberbullying detection in Twitter social media platform, *IEEE Access* **10**: 25857–25871.

Namdari, R. (2023). Mental health corpus, *Dataset*, `https://www.kaggle.com/datasets/reihanenamdari/mental-health-corpus`.

Narayanan, S. and Georgiou, P.G. (2013). Behavioral signal processing: Deriving human behavioral informatics from speech and language, *Proceedings of the IEEE* **101**(5): 1203–1233.

Naslund, J.A., Aschbrenner, K.A., Kim, S.J., McHugo, G.J., Unützer, J., Bartels, S.J. and Marsch, L.A. (2017). Health behavior models for informing digital technology interventions for individuals with mental illness, *Psychiatric Rehabilitation Journal* **40**(3): 325.

Naslund, J.A., Aschbrenner, K.A., Marsch, L.A. and Bartels, S.J. (2016). The future of mental health care: Peer-to-peer support and social media, *Epidemiology and Psychiatric Sciences* **25**(2): 113–122.

Naslund, J.A., Bondre, A., Torous, J. and Aschbrenner, K.A. (2020). Social media and mental health: Benefits, risks, and opportunities for research and practice, *Journal of Technology in Behavioral Science* **5**(3): 245–257.

Nazmunnahar, Nasim, R., Mosharrafa, R.A., Hossain, I., Saima, J., Taher, T., Hossain, M.J., Rahman, M.A. and Islam, M.R. (2023). Association between flaunting behaviors on social media and among the general population in Bangladesh: A cross-sectional study, *Health Science Reports* **6**(11): e1701.

Nutley, S.K., Falise, A.M., Henderson, R., Apostolou, V., Mathews, C.A. and Striley, C.W. (2021). Impact of the COVID-19 pandemic on disordered eating behavior: Qualitative analysis of social media posts, *JMIR Mental Health* **8**(1): e26011.

Opsahl, T. and Panzarasa, P. (2009). Clustering in weighted networks, *Social Networks* **31**(2): 155–163.

Perski, O., Blandford, A., West, R. and Michie, S. (2017). Conceptualising engagement with digital behaviour change interventions: A systematic review using principles from critical interpretive synthesis, *Translational Behavioral Medicine* **7**(2): 254–267.

Pourkeyvan, A., Safa, R. and Sorourkhah, A. (2024). Harnessing the power of hugging face transformers for predicting mental health disorders in social networks, *IEEE Access* **12**: 28025–28035.

Reece, A.G., Reagan, A.J., Lix, K.L., Dodds, P.S., Danforth, C.M. and Langer, E.J. (2017). Forecasting the onset and course of mental illness with Twitter data, *Scientific Reports* **7**(1): 13006.

Ribeiro, F.N., Araújo, M., Gonçalves, P., André Gonçalves, M. and Benevenuto, F. (2016). SentiBench—A benchmark comparison of state-of-the-practice sentiment analysis methods, *EPJ Data Science* **5**(1): 23.

Rubin, K.H., Coplan, R.J. and Bowker, J.C. (2009). Social withdrawal in childhood, *Annual Review of Psychology* **60**(1): 141–171.

Shannon, H., Bush, K., Villeneuve, P.J., Hellemans, K.G. and Guimond, S. (2022). Problematic social media use in adolescents and young adults: Systematic review and meta-analysis, *JMIR Mental Health* **9**(4): e33450.

Šmahel, D., Macháčková, H., Šmahelová, M., Čevelíček, M., Almenara, C.A. and Holubčíková, J. (2018). Digital technology and health: A theoretical framework, *in* D. Šmahel *et al.* (Eds), *Digital Technology, Eating Behaviors, and Eating Disorders*, Springer, Cham, pp. 21–43.

Theodoropoulos, T., Makris, A., Kontopoulos, I., Violos, J., Tarkowski, P., Ledwoń, Z., Dazzi, P. and Tserpes, K. (2023). Graph neural networks for representing multivariate resource usage: A multiplayer mobile gaming case-study, *International Journal of Information Management Data Insights* **3**(1): 100158.

Tsakalidis, A., Aletras, N., Cristea, A.I. and Liakata, M. (2018). Nowcasting the stance of social media users in a sudden vote: The case of the Greek referendum, *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, Turin, Italy*, pp. 367–376.

Uban, A.-S., Chulvi, B. and Rosso, P. (2021). An emotion and cognitive based analysis of mental health disorders from social media data, *Future Generation Computer Systems* **124**: 480–494.

Valdez, D., Ten Thij, M., Bathina, K., Rutter, L.A. and Bollen, J. (2020). Social media insights into us mental health during the COVID-19 pandemic: Longitudinal analysis of Twitter data, *Journal of Medical Internet Research* **22**(12): e21418.

Valkenburg, P.M., Meier, A. and Beyens, I. (2022). Social media use and its impact on adolescent mental health: An umbrella review of the evidence, *Current Opinion in Psychology* **44**: 58–68.

Vidal-Ribas, P. and Stringaris, A. (2021). How and why are irritability and depression linked?, *Child and Adolescent Psychiatric Clinics* **30**(2): 401–414.

Voorheis, P., Bhuiya, A.R., Kuluski, K., Pham, Q. and Petch, J. (2023). Making sense of theories, models, and frameworks in digital health behavior change design: Qualitative descriptive study, *Journal of Medical Internet Research* **25**: e45095.

Yang, Y., Liu, K., Li, S. and Shu, M. (2020). Social media activities, emotion regulation strategies, and their interactions on people's mental health in COVID-19 pandemic, *International Journal of Environmental Research and Public Health* **17**(23): 8931.

Zandavi, S.M., Rashidi, T.H. and Vafaee, F. (2021). Dynamic hybrid model to forecast the spread of COVID-19 using LSTM and behavioral models under uncertainty, *IEEE Transactions on Cybernetics* **52**(11): 11977–11989.

Zhang, Y., Lyu, H., Liu, Y., Zhang, X., Wang, Y. and Luo, J. (2021). Monitoring depression trends on Twitter during the COVID-19 pandemic: Observational study, *JMIR Infodemiology* **1**(1): e26769.

Ayodeji Olusegun Ibitoye is a senior lecturer in data science at the University of Greenwich, London, specialising in machine learning, explainable AI, and responsible AI. He holds a PhD in computer science, an MBA in strategic leadership, and is a Senior Fellow of the Higher Education Academy (SFHEA). He is an active member of various professional bodies, including the British Early Career Researchers' Network and Black in AI, and contributes to the research community as a journal reviewer, keynote speaker, and collaborator on international AI initiatives. His research advances ethical and interpretable AI for social good, with a thematic focus on mental health, digital behaviour analysis, and healthcare decision support. He supervises postgraduate and doctoral research in responsible AI and applied machine learning, and mentors and consults globally to translate research into meaningful societal impact.

Oladosu Oyebisi Oladimeji is a doctoral researcher at the Centre for Mathematical Modelling and Intelligent Systems for Health and Environment (MISHE), Atlantic Technological University, Ireland. His research focuses on medical image analysis for breast cancer screening using artificial intelligence. With a distinguished academic background, including an MSc in computer science (distinction) from the University of Ibadan and a first class honours BSc from Bowen University, he has published over 20 peer-reviewed papers in top-tier journals and conferences. His expertise spans machine learning, deep learning, computer vision, and health informatics. As a recipient of prestigious awards, including a MOCHAS scholarship and WTUN funding, and a winner of the National AI Challenge 2024, he combines cutting-edge research with practical teaching experience. He actively contributes to international research collaborations.

Temitayo Fagbola (PhD, MIEEE, SFHEA, MBCS) is an assistant professor in the Centre of Excellence for Data Science, Artificial Intelligence and Modelling (DAIM) at the University of Hull, UK. His work focuses on medical natural language processing, healthcare-based agentic AI systems, explainable AI for precision cancer prognosis, and responsible AI for social good, with an emphasis on developing interpretable, human-centred models that support healthcare innovation and decision-making. He has authored two books, co-edited special issues, and published more than 30 peer-reviewed articles in international journals and conferences. He actively supervises MSc and PhD researchers, contributes to interdisciplinary and international collaborations across the UK, Asia, Africa, and Europe, and regularly serves as a reviewer and organiser for AI and healthcare-focused conferences and journals.

Received: 14 April 2025
Revised: 1 October 2025
Re-revised: 15 December 2025
Accepted: 29 January 2026