Computer Modeling in Engineering & Sciences

Tech Science Press

Doi:10.32604/cmes.2025.075119

ARTICLE

Neuro-Symbolic Graph Learning for Causal Inference and Continual Learning in Mental-Health Risk Assessment

Monalisa Jena[1], Noman Khan[2,*], Mi Young Lee[3,*] and Seungmin Rho[3]

[1]Department of Computer Science, Fakir Mohan University, Balasore, 756019, Odisha, India
[2]Department of Architecture and Architectural Engineering, Yonsei University, Seoul, 03722, Republic of Korea
[3]Department of Industrial Security, Chung-Ang University, Seoul, 06974, Republic of Korea

*Corresponding Authors: Noman Khan. Email: noman@yonsei.ac.kr; Mi Young Lee. Email: miylee@cau.ac.kr

Received: 25 October 2025; Accepted: 19 December 2025; Published: 29 January 2026

ABSTRACT: Mental-health risk detection seeks early signs of distress from social media posts and clinical transcripts to enable timely intervention before crises. When such risks go undetected, consequences can escalate to self-harm, long-term disability, reduced productivity, and significant societal and economic burden. Despite recent advances, detecting risk from online text remains challenging due to heterogeneous language, evolving semantics, and the sequential emergence of new datasets. Effective solutions must encode clinically meaningful cues, reason about causal relations, and adapt to new domains without forgetting prior knowledge. To address these challenges, this paper presents a Continual Neuro-Symbolic Graph Learning (CNSGL) framework that unifies symbolic reasoning, causal inference, and continual learning within a single architecture. Each post is represented as a symbolic graph linking clinically relevant tags to textual content, enriched with causal edges derived from directional Point-wise Mutual Information (PMI). A two-layer Graph Convolutional Network (GCN) encodes these graphs, and a Transformer-based attention pooler aggregates node embeddings while providing interpretable tag-level importances. Continual adaptation across datasets is achieved through the Multi-Head Freeze (MH-Freeze) strategy, which freezes a shared encoder and incrementally trains lightweight task-specific heads (small classifiers attached to the shared embedding). Experimental evaluations across six diverse mental-health datasets ranging from Reddit discourse to clinical interviews, demonstrate that MH-Freeze consistently outperforms existing continual-learning baselines in both discriminative accuracy and calibration reliability. Across six datasets, MH-Freeze achieves up to 0.925 accuracy and 0.923 F1-Score, with AUPRC $\geq$ 0.934 and AUROC $\geq$ 0.942, consistently surpassing all continual-learning baselines. The results confirm the framework's ability to preserve prior knowledge, adapt to domain shifts, and maintain causal interpretability, establishing CNSGL as a promising step toward robust, explainable, and lifelong mental-health risk assessment.

KEYWORDS: Catastrophic forgetting; causal inference; continual learning; deep learning; graph convolutional network; mental health monitoring; transformer

1 Introduction

Mental-health disorders such as depression, anxiety, and suicidal ideation are rising globally, posing serious risks to individuals and society [1,2]. With the growth of online platforms and clinical records, vast amounts of unstructured text now capture personal experiences and distress signals [3,4]. Automatically analyzing this text for early detection of psychological risk has become an urgent research problem. However, it remains highly challenging due to the ambiguity of natural language, and the subtlety of psychological cues [5,6]. Timely detection enables early intervention and targeted allocation of limited mental-health

resources, and real-world deployment demands models that are not only accurate but also interpretable and well-calibrated so that clinicians and moderators can act on predictions with confidence [7].

In many real-world scenarios, posts and interviews carry implicit cues (e.g., hopelessness, insomnia, self-harm) that are difficult to capture with surface features alone [8]. Purely neural approaches can learn powerful representations but often lack interpretability and causal grounding; purely symbolic methods offer transparency but struggle with linguistic variability and generalization [9]. Moreover, data arrive over time from different communities and collection protocols, creating domain shift and exposing models to *catastrophic forgetting* when retrained sequentially [10,11]. These factors emphasize an integrated approach that can (i) structure free text into clinically meaningful graphs, (ii) model directional relations among risk factors, and (iii) learn continually across datasets without erasing earlier competencies. In addition, class imbalance, where rare but critical signals such as suicidal ideation are underrepresented, biases predictions and reduces reliability. Addressing these issues requires models that are both *interpretable* and *adaptable*, while retaining stability across evolving datasets.

Continual learning (CL), also referred to as *lifelong* or *incremental learning*, aims to enable models to acquire new knowledge over time without forgetting previously learned information [12]. Unlike traditional retraining approaches that require access to all past data, CL supports sequential learning across tasks or domains by reusing shared representations and adapting to new inputs efficiently. This paradigm is particularly valuable in mental-health applications, where new linguistic trends, populations, and annotation protocols continuously emerge. A robust continual learning mechanism ensures that models remain up to date while preserving earlier competencies, enabling sustainable and realistic deployment in evolving digital health environments.

To address these challenges, a *Continual Neuro-Symbolic Graph Learning (CNSGL)* framework is proposed in this work for causal inference and continual learning in mental-health risk assessment. CNSGL represents each post as a symbolic graph in which a post node connects to tag nodes derived from a risk lexicon. Beyond simple co-occurrence, graphs are enriched with directional edges using a variant of point-wise mutual information to reflect likely precursors and consequents among risk factors. A two-layer Graph Convolutional Network (GCN) propagates information over this structure, and a lightweight Transformer attention pooler, anchored by a learnable `[CLS]` token, aggregates node embeddings while producing tag-level importances for interpretability.

To enable continual adaptation, the proposed framework employs a Multi-Head Freeze (MH-Freeze) strategy that freezes the shared encoder after the first dataset and incrementally attaches lightweight task-specific heads for subsequent datasets. Here, "task-specific head" refers to a small linear-sigmoid classifier attached to the shared embedding for each dataset. This form is adopted to enable lightweight adaptation on a fixed embedding space: the frozen GCN-Transformer encoder produces a stable representation, and the head maps it directly to a calibrated probability via binary cross-entropy (BCE) loss. This keeps updates simple and efficient, reduces the risk of cross-task gradient interference, and preserves calibration. In contrast, deeper or non-linear heads introduce extra trainable layers that can overfit to a single dataset and reintroduce interference with previously learned tasks. Each dataset is treated as a separate task (T1-T6), allowing systematic evaluation of domain transfer and retention. Our evaluation further incorporates both discrimination and calibration metrics (AUROC, AUPRC, Brier score, and Expected Calibration Error) to quantify predictive reliability under domain shift and sequential learning conditions.

1.1 Motivation

In mental health risk assessment and monitoring, systems are deployed in dynamic and evolving scenarios. In many real world scenarios, diverse signals (clinical notes, social media text, speech, and

wearable biosignals) are used to detect risk states such as depression, anxiety, self-harm intent, and acute stress. However, most pipelines are trained on static datasets with fixed labels and vocabularies. When new expressions, populations, or risk patterns appear, performance is often degraded. In practice, full retraining on new data is often required, while incremental updates risk catastrophic forgetting that overwrites previously learned knowledge [13]. Therefore, continual learning is increasingly regarded as important for mental health tasks. Rather than retraining from scratch, new risk categories or domains can be incorporated as they arise while preserving recognition of earlier ones [14].

1.2 Research Gaps

Despite rapid progress, Several unresolved challenges continue to hinder dependable mental-health risk detection from text:

- Missing causal structure: Most models treat symptoms as flat labels; they do not encode directed tag→tag influences or use them during message passing.
- Limited interpretability: Explanations are often post-hoc for text tokens, not concept-level (symbolic tags) nor pathway-level (causal paths).
- Catastrophic forgetting: Models struggle to retain prior knowledge as new datasets arrive, while simple, deployable continual-learning solutions are still lacking.
- Opaque pooling: Mean/max pooling blurs which symbolic tags matter per post; attention over concept nodes is rarely leveraged.

In light of the above, we propose a continual neuro-symbolic framework that builds per-post symbolic tag graphs with directed links, encodes them via a two-layer GCN, and uses a lightweight attention head to form calibrated, interpretable post representations. For sequential datasets, we adopt a frozen-encoder, multi-head protocol to prevent forgetting while keeping adaptation lightweight. The key contributions of this work are summarized below:

- A neuro-symbolic graph learning framework is proposed that combines symbolic reasoning, causal inference, graph-based neural encoding, and continual learning for mental-health risk assessment.
- Symbolic graphs are constructed from text using risk-related tags, ensuring interpretability by grounding predictions in clinically meaningful indicators.
- A causal-aware enrichment mechanism introduces directed tag–tag edges, capturing potential causal influences among symptoms rather than simple co-occurrence.
- A graph convolutional encoder is employed to propagate symbolic and causal features, followed by a lightweight Transformer-based attention head weights the post–tag embeddings and a classifier that outputs binary risk predictions through probability estimation and thresholding.
- A continual learning strategy (multi-head frozen-encoder) is implemented to preserve knowledge across datasets while enabling adaptation to new domains, mitigating catastrophic forgetting.
- Extensive experiments are conducted on multiple datasets, and results are compared against strong continual learning baselines, demonstrating improved robustness, interpretability, and adaptability.

The remainder of this paper is organized as follows: Section 2 presents an extensive survey of related research and methods relevant to this work, Section 3 presents the proposed CNSGL framework, including symbolic graph construction, causal enrichment, the GCN–Transformer encoder, and the MH-Freeze continual-learning strategy. Section 4 describes the experimental details, and evaluation metrics. Section 5 presents the ablation study analyzing the contribution of each component. Section 6 concludes the paper and highlights future directions.

2 Related Work

Detecting mental-health risk from text is a challenging and active area with direct applications to screening, risk stratification, and clinical decision support. This section reviews related research across the categories listed in the subsections, aligning each category of methods pertinent to this work.

2.1 Mental-Health Risk Detection from Text

This subsection reviews key efforts on detecting mental-health signals from social-media text, ranging from traditional machine learning (ML) to deep learning (DL). Hemmatirad et al. [15] showed that lexicon and handcrafted features paired with support vector machine or logistic regression classifiers can distinguish high-risk users using linguistic and emotional cues. With the advent of contextual embeddings, hybrid models such as BERT+BiLSTM have been proposed by Zhou and Mohd [16] to better handle informal language, emojis, and sequential patterns in depression-related posts.

Prior surveys consolidate the literature and shed light on persistent challenges. Garg [17] reviewed 92 studies, introduced an updatable suicide-detection repository, and emphasized the need for real-time, responsible AI. Skaik and Inkpen [18] surveyed NLP/ML approaches for public mental-health surveillance, summarizing data collection strategies, modeling tools, and remaining gaps. Other studies explore emotion-aware and efficiency-focused systems. Benrouba and Boudour [19] proposed an emotion-aware content-filtering framework that classifies posts into basic emotions and compares them with an "ideal" lexicon to flag potentially harmful content. Ding et al. [20] compared ML models (logistic regression, random forest, LightGBM) with DL models (ALBERT, GRU) for binary and multi-class mental-health classification, finding that ML methods offer better interpretability and efficiency on medium-sized datasets, whereas DL models better capture complex linguistic patterns.

2.2 Causal Inference

Causal reasoning in language means modeling directional influence ($A \rightarrow B$) so that changing A would change the likelihood of B, beyond simple correlation. This has been studied using temporal precedence and directional association measures, causal discovery on event graphs, and counterfactual analyses. However, in social-media risk assessment, such causal structure is rarely embedded within the encoder itself. Choudhury and Kiciman [21] examined the causal impact of online social-support language in Reddit mental-health communities on future suicidal-ideation risk. Using human assessments within a stratified propensity-score framework to form comparable cohorts, they estimated treatment effects of support types and found that esteem and network support significantly reduce subsequent risk, with implications for tools that enhance support provision.

Zhang et al. [22] proposed a causal framework based on a counterfactual neural temporal point process (TPP) to estimate the individual treatment effect (ITE) of misinformation on user beliefs and actions at scale, using a neural TPP with Gaussian mixtures for efficient inference. Experiments on synthetic data and a real COVID-19 vaccine dataset showed identifiable causal effects of misinformation, including negative shifts in users' vaccine-related sentiments. Cheng et al. [23] surveyed Event Causality Identification (ECI) and proposed a systematic taxonomy split into sentence-level ECI (SECI) and document-level ECI (DECI) tasks, reviewing approaches from feature/ML methods to deep semantic encoding, event-graph reasoning, and prompt/causal-knowledge pretraining, with notes on multilingual, cross-lingual, and zero-shot large language model settings.

2.3 Neural Encoders: Graph + Attention

GCN encoders capture relational structure for text via message passing on graphs, benefiting settings with explicit concept relations [24]. Hamilton et al. [25] proposed GraphSAGE, an inductive framework that extended GCNs to unsupervised learning and introduced trainable aggregation functions beyond simple convolutions. The method generated embeddings for unseen nodes by sampling and aggregating neighborhood features, leveraging node attributes for generalization. Yao et al. [26] introduced Text GCN, which built a corpus-level graph from word co-occurrence and document-word relations to jointly learn word and document embeddings. Without relying on external embeddings, Text GCN outperformed state-of-the-art methods on multiple benchmarks and showed strong robustness with limited training data.

Transformers, driven by self-attention, excel at weighting inputs and can be used as interpretable pooling over concept embeddings [27]. Vaswani et al. [28] proposed the Transformer, a sequence transduction architecture based solely on attention mechanisms, removing recurrence and convolutions. The model achieved state-of-the-art results on Workshop on Machine Translation 2014 English to German and English to French translation tasks, while being more parallelizable and significantly faster to train than prior approaches.

Devlin et al. [29] introduced Bidirectional Encoder Representations from Transformers (BERT), a bidirectional Transformer-based model pre-trained on unlabeled text by jointly conditioning on left and right context. With simple fine-tuning, BERT achieved state-of-the-art results on eleven NLP tasks. Yang et al. [30] proposed a hierarchical attention network (HAN) for document classification, which reflected the hierarchical structure of documents and applied attention at both word and sentence levels. The model outperformed prior methods on six large-scale benchmarks and provided interpretable document representations by highlighting informative words and sentences.

2.4 Continual Learning for Mental Health

Continual learning addresses time-varying, patient-specific data in mental-health scenarios by incrementally updating models from electronic health records, speech/text, and wearable streams while preserving prior knowledge. Although the CL for mental health literature remains limited, we highlight a few representative systems that show feasibility under practical constraints. Gamel and Talaat [31] proposed SleepSmart, an Internet of Things (IoT) enabled continual learning framework for intelligent sleep enhancement. The system employed wearable biosensors to capture physiological signals during sleep, which were processed via an IoT platform to deliver personalized recommendations. By leveraging continual learning, SleepSmart improved recommendation accuracy over time, and a pilot study demonstrated its effectiveness in enhancing sleep quality and reducing disturbances.

Lee and Lee [32] explored the role of continual learning in medicine, where models adapt to new patient data without forgetting prior knowledge. They emphasized challenges such as catastrophic forgetting and regulatory constraints, but argued that continual learning offers advantages over non-adaptive Food and Drug Administration approved systems by incrementally improving diagnostic and decision-support performance. Li and Jha [33] proposed DOCTOR, a continual-learning framework for multi-disease detection on wearable medical sensors at the edge. The system used a multi-headed deep neural network with replay-based CL, via exemplar data preservation or synthetic data generation to mitigate catastrophic forgetting while sequentially adding tasks with new classes and distributions. In experiments, a single model maintained high accuracy, yielding up to 43% higher test accuracy, 25% higher F1 score, and 0.41 higher backward transfer over naive fine-tuning.

A structured comparison is presented in Table 1 to more clearly contextualize CNSGL within existing work. Prior methods typically incorporate only one or two of the components, symbolic representations,

causal edge modeling, graph-based encoders, Transformer attention mechanisms, or continual-learning strategies, rather than unifying all of them within a single framework. As shown in Table 1, approaches that employ symbolic reasoning rarely integrate GNN encoders or explicit causal edge construction; causal GNN models generally do not use symbolic tag vocabularies or Transformer-based pooling; and continual-learning systems commonly operate without symbolic graphs or causal modeling. In contrast, CNSGL combines directional PMI-derived causal edges, a symbolic tag graph, a two-layer GCN encoder, a Transformer attention pooler, and a multi-head freeze continual-learning strategy within one architecture tailored for mental-health risk detection. This integrated design forms the central novelty of the approach and demonstrates how CNSGL extends beyond existing component-wise methods.

Table 1: Comparison of CNSGL with representative neuro-symbolic, causal-modeling, and continual-learning architectures

Author	Model	Symbolic Representation	Causal Edge Modeling	Graph Encoder (GCN/GNN)	Transformer Attention	Continual Learning	Application Domain
Nie et al. (2022) [34]	Incremental GCN	No (utterances & speakers as nodes; no symbolic tags/lexicons)	No	GC/N	Yes, multi-head attention for utterance correlation	Yes, Incremental fine-tuning with new utterances	Conversation emotion detection
Kaur et al. (2022) [35]	Transformer-based causal categorization	No	Causal labels, but no causal edges	No	Yes	No	Mental-health causal categorization on social-media posts
Kodati & Tene (2023) [36]	Context-based bidirectional gated recurrent unit with multi-head attention and a convolutional neural network	Partial, POS tags + lexicon features (not symbolic graphs)	No	No, CNN used	Yes, Multi-Head Attention + BERT MLM/Self-attention	No	Suicidal-emotion detection on social-media text
Kumar (2023) [37]	Neuro-Symbolic AI framework	Yes, structured knowledge graphs, symbolic reasoning, cognitive theories	General causal reasoning mentioned, but no graph construction method	No	No	No	personalized mental health therapy, computational psychiatry
Tang et al. (2023) [38]	Causality-Driven GCN Framework	No	Yes, causal interventions + invariant prediction principle + causality scoring	GCN	No	No	Automated classification of postural abnormalities in Parkinson's disease
Bhuyan et al. (2024) [39]	Conceptual Neuro-Symbolic AI framework	Yes, symbolic reasoning & discrete logic	No	GNN	No	Yes	General AI/Neuro-Symbolic reasoning
Dalkic (2025) [40]	Context-Aware EEG Emotion Recognition System	No (raw EEG + context embeddings; no symbolic tags or lexicons)	No	No	Yes, Temporal Transformer encoder	Yes, EWC-based continual learning	EEG-based emotion recognition/affective computing
Patanè et al. (2025) [41]	Prompt-based continual learning framework	No (mobile sensing features; no symbolic tags/lexicons)	No	No	Yes, Transformer backbone with task prompts	Yes, Replay buffer + prompt-based adaptation	personalized mental well-being monitoring
Febrinanto et al. (2025) [42]	Causal Graphs for Brains	No	Yes, causal discovery + transfer entropy + curvature-based rewiring	Yes. GNN models refined causal graphs	No	No	Brain disease classification (neuroscience)

(Continued)

Table 1 (continued)

Author	Model	Symbolic Representation	Causal Edge Modeling	Graph Encoder (GCN/GNN)	Transformer Attention	Continual Learning	Application Domain
Gosala et al. (2025) [43]	GCN-LSTM; 12-layer GCN	No (EEG electrodes as nodes, not symbolic tags)	No, edges from cohesion/ phase-locking, not causal)	Yes, GCN + hybrid GCN-LSTM	No	No	Schizophrenia classification from EEG (clinical neuroimaging)
Our Work	Continual Neuro-Symbolic Graph Learning (CNSGL)	Yes, symbolic mental-health tags with clinical relevance	Yes, directional PMI edges encoding causal tendencies between tags	Yes, 2-layer GCN encoder	Yes, Transformer-based attention pooling with CLS-to-tag weights for explainability	Yes, Multi-Head Freeze (encoder frozen after Task-1) to prevent forgetting	Mental-health risk detection from social media posts

3 Proposed Work

A *Continual Neuro-Symbolic Graph Learning (CNSGL) framework* is proposed for causal inference and continual learning in mental-health risk assessment. In this framework, symbolic reasoning, causal graph construction, graph neural encoding, and continual learning are combined within a single architecture. The details of the proposed work are presented in the following subsections.

3.1 Symbolic Graph Construction

Mental-health text from online platforms or clinical records is largely unstructured and often contains implicit cues about psychological conditions that are difficult to analyze directly. To impose structure and enhance interpretability, each post is represented as a *symbolic graph* that captures both semantic content and clinically meaningful indicators. A compact set of ten symbolic tags, sleep, anxiety, depression, stress, anger, lonely, health, fear, coping, and suicidal, was constructed based on well-established constructs in computational mental-health research and their frequent annotation in benchmark datasets. The vocabulary was further validated through manual inspection and an expert-informed review to ensure clinical relevance. A small, consistent set was intentionally maintained to minimize noise and support stable directional PMI estimation during causal graph construction. Let $\mathcal{D}$ denote the dataset of posts, and let $d_i \in \mathcal{D}$ denote a post in this collection. A vocabulary $\mathcal{V}$ of risk-related terms, including *hopelessness*, *insomnia*, and *self-harm*, is predefined. Using this lexicon, a set of symbolic tags is assigned to the post,

$$\tau_i = \{t_1, t_2, \ldots, t_m\}, \quad t_j \in \mathcal{V},$$

where each t_j denotes a tag identified in d_i. These tags provide explicit signals of potential risk factors, linking the unstructured narrative of a post to interpretable constructs grounded in psychology.

The symbolic graph is defined as $G_i = (V_i, E_i, X_i)$ consisting of:

- **Nodes:** $V_i = \{v_0\} \cup \{v_j : t_j \in \tau_i\}$, where v_0 represents the post and v_j denote the symbolic tags.
- **Edges:** $E_i = \{(v_0, v_j), (v_j, v_0)\}$ establish bidirectional links between the post and its tags.
- **Features:** Node attributes encode symbolic and semantic information. Tag node features use Term Frequency-Inverse Document Frequency (TF-IDF) scores [44] of the tag t_j in post d_i. For a tag node v_j,

  $$x_{v_j}[0] = \mathrm{TFIDF}(t_j, d_i),$$

  while the post node v_0 is enriched with a 64-dimensional embedding obtained through truncated singular value decomposition (SVD) of the TF-IDF representation,

$x_{v_0}[1:] = \mathrm{SVD}(\mathrm{TFIDF}(d_i)) \in \mathbb{R}^{64}$.

At this stage, G_i uses undirected post-tag links (implemented as bidirectional pairs) to encode association rather than causality. Directional tag–tag edges are introduced in the subsequent Causal-Aware Graph Enrichment step, where order-sensitive statistics are is used to determine edge orientation.

3.2 Causal-Aware Graph Enrichment

Edges based solely on co-occurrence capture statistical associations but cannot distinguish whether one factor precedes or influences another. For example, *sleep deprivation* may frequently appear with *stress*, yet in many cases it precedes and contributes to *suicidal ideation*. To incorporate such directional relationships, graphs are enriched with causal edges in addition to co-occurrence links. To quantify how strongly tags co-occur in the input space, point-wise mutual information (PMI) is used here. Consider the tags *sleep* and *anxiety*. These tags may frequently appear together in posts, which would yield a symmetric, undirected edge in a standard co-occurrence graph. Directional PMI instead focuses on ordered pairs and estimates whether one tag is more likely to appear before the other. If ordered counts show that mentions of *sleep* problems systematically precede *anxiety* indicators more often than the reverse, then $\mathrm{PMI}_{\mathrm{dir}}(\text{sleep} \rightarrow \text{anxiety})$ exceeds $\mathrm{PMI}_{\mathrm{dir}}(\text{anxiety} \rightarrow \text{sleep})$, and the graph includes the directed edge *sleep* → *anxiety*. In this way, directional PMI encodes asymmetric, precedence-aware relationships that cannot be represented by undirected co-occurrence edges alone. PMI between two given tags t_a and t_b is calculated in Eq. (1) [45]:

$$\mathrm{PMI}(t_a, t_b) = \log \frac{P(t_a, t_b)}{P(t_a)P(t_b)} \tag{1}$$

Probabilities were estimated by counting how often each tag and each tag pair appeared within the same post and then normalizing by the total. A small smoothing constant was applied so that rare tags didn't get zero probability. Because PMI treats a pair the same in either order, it captures association only and does not encode direction. Let $c(t_a \prec t_b)$ be the number of posts in which t_a occurs before t_b, and let N_{pairs} be the total number of ordered tag pairs considered. Eq. (2) shows the mathematical definition of the directional PMI:

$$\mathrm{PMI}^{\mathrm{dir}}(t_a \rightarrow t_b) = \log \frac{P(t_a \prec t_b)}{P(t_a)P(t_b)}, \tag{2}$$

where, $P(t_a \prec t_b) \approx \frac{c(t_a \prec t_b)}{N_{\mathrm{pairs}}}$ is the probability of ordered event. A directional PMI threshold δ is applied to filter out weak or noisy associations. In practice, δ was selected through validation by sweeping values in $\{0.0, 0.05, 0.1, 0.2\}$ and choosing the smallest threshold that removed spurious edges while preserving clinically meaningful relations. The framework was most stable for $\delta = 0.1$, which we adopt for all experiments. Sensitivity analysis showed that the graph structure remained consistent for thresholds within a small neighborhood (0.05–0.15), indicating that results are not overly sensitive to the exact choice of δ.

A directed edge, $t_a \rightarrow t_b$ is introduced whenever $\mathrm{PMI}^{\mathrm{dir}}(t_a \rightarrow t_b) > \delta$. This indicates that the occurrence of t_a increases the likelihood of t_b, suggesting a causal tendency rather than a mere correlation. After adding the causality factor, the resulting enriched graph becomes:

$$G_i = (V_i, E_i^{\mathrm{co}} \cup E_i^{\mathrm{causal}}, X_i),$$

where E_i^{co} denotes post–tag associations, and E_i^{causal} represents directed tag–tag relations. For example: If mentions of *sleep problems* are frequently followed by references to *anxiety*, and the directional PMI exceeds the threshold δ, an edge $v_{sleep} \rightarrow v_{anxiety}$ is added. In this way, the representation allows the model to reason not only about what terms appear together, but also about which factors may act as potential precursors of others. During the execution of the GCN, the directed causal edges are converted into an undirected form (with self-loops) so that message passing remains symmetric. Throughout this work, the directed edges are treated as precedence-aware statistical associations rather than definitive cause–effect links; accordingly, the term "causal" is used in an operational sense to describe directional risk-tendency patterns observed in mental-health text.

The proposed CNSGL framework is depicted as an end-to-end architecture in Fig. 1, emphasizing the left-to-right progression from symbolic/causal structuring to representation learning and, then to sequential adaptation. The diagram marks where the shared encoder is frozen and where dataset-specific heads are attached, making clear how prior knowledge is preserved while new tasks are added. The Transformer Encoder used for graph-level pooling in the encoding stage is described separately in Fig. 2.

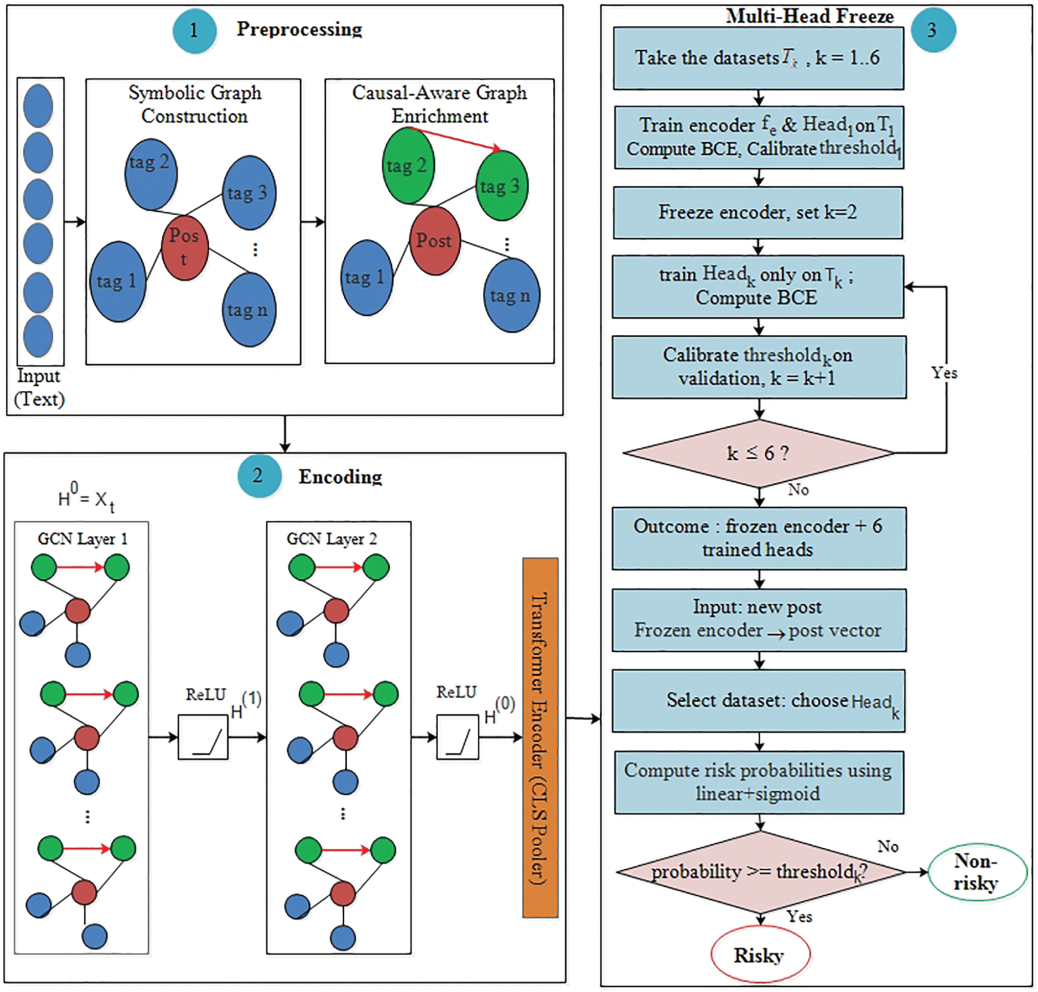

Figure 1: Proposed CNSGL framework consisting of three key components: (**1**) Preprocessing, where a post-tag graph is built and enriched with directional causal edges (red), and causal tag nodes (green). (**2**) Encoding, consisting of two-layer GCN followed by a light-weight Transformer `[CLS]` pooler, produces a post vector and tag importances. (Continued)

Figure 1: (continued) (**3**) Continual learning (MH-Freeze): the encoder and Head_1 were trained on T_1, after which the encoder was frozen; for subsequent datasets, a small linear-sigmoid head was attached and trained, with a dataset specific threshold calibrated. At inference, a post was encoded once, the appropriate head was selected by dataset, and the resulting probability was thresholded to yield *risky*/*non-risky* posts

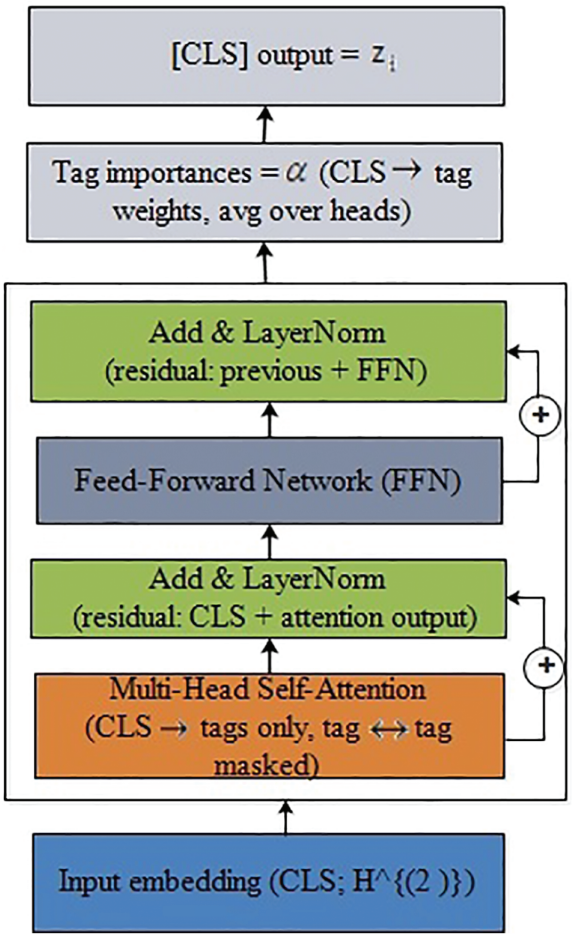

Figure 2: Architecture of the encoder-only Transformer used as an attention pooler. The CLS token and tag embeddings are fed into a single Transformer block, where multi-head attention computes CLS → tag attention scores while masking tag-to-tag interactions. Residual connections, layer normalization, and a feed-forward network refine the CLS representation, which becomes the final pooled embedding for classification

3.3 GCN Encoder

The symbolic graphs enriched with causal relations are processed by a two-layer GCN. The GCN propagates information across connected nodes so that each representation reflects both its own features and those of neighboring nodes. In this way, a post node aggregates signals from its tags, while tag nodes incorporate both symbolic and causal context. Eq. (3) shows how the node representations are updated at layer l [46].

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\,\tilde{A}\,\tilde{D}^{-\frac{1}{2}}\,H^{(l)}\,W^{(l)}\right), \tag{3}$$

where $\tilde{A} = A + I$ is the adjacency with self-loops, $\tilde{D}$ is its degree matrix, $W^{(l)}$ are trainable weights, and σ is an element-wise ReLU activation function. The input is $H^{(0)} = X_i$, the node-feature matrix of graph G_i. Since causal edges are directed, their weights are first assembled in a directed matrix W and then symmetrized to form A (e.g., $A = \frac{1}{2}(W + W^{\top})$) prior to normalization.

Although the final GCN uses a symmetrized adjacency matrix for stable message passing, the causal interpretability of the framework is retained because symmetrization occurs only after causal tendencies have been encoded in the edge-selection stage. Directional PMI determines which tag pairs are connected and the strength of those connections, thereby shaping the underlying causal structure even if the GCN operates on an undirected form of the graph. The interpretability comes from this directed edge construction and from the subsequent analysis of causal paths and attention weights, whereas symmetrization serves primarily as a computational requirement of the canonical GCN rather than a removal of causal information. After two layers, the node embeddings $H^{(2)}$ are passed to a learned classification token [CLS] as a global query over the tag nodes to produce a graph-level representation z_i with tag-level importances. This pooled vector $z_i \in \mathbb{R}^{d_z}$ encodes symbolic, semantic, and causal structure and is used for classification.

3.4 Transformer-Based Attention

While mean pooling provides a simple mechanism for aggregating node embeddings into a graph-level representation, it treats all nodes equally and fails to highlight which risk factors are more influential in a particular post. To address this limitation, the node embeddings produced by the GCN are passed through a Transformer encoder to perform attention-based pooling [47]. Fig. 2 illustrates the Transformer encoder block employed as an attention pooler over GCN-derived node embeddings. The [CLS] token attends to tag embeddings to produce a pooled representation while simultaneously providing interpretable tag-level importance scores through attention weights.

Let $H^{(2)} = [h_1, \ldots, h_{|V_i|}]$ be the node embeddings (post + tags) for post i after the 2-layer GCN. The input sequence formed is presented in Eq. (4):

$$X = [\mathrm{CLS};\, H^{(2)}] \tag{4}$$

where `[CLS]` is a learnable pooling token. The encoder computes linear projections

$$Q = XW_Q, \qquad K = XW_K, \qquad V = XW_V,$$

with trainable W_Q, W_K, W_V and per-head key dimension d_k. We apply a *pooler mask* so that only the `[CLS]` row of Q issues queries (tag↔tag attention is masked). Let q_{CLS} denote the `[CLS]` query. The attention weights from `[CLS]` to all tokens are calculated in Eq. (5):

$$\alpha = \mathrm{softmax}\left(q_{\mathrm{CLS}} K^{\top} / \sqrt{d_k}\right) \tag{5}$$

These are then used to form a weighted sum of the values for `[CLS]` (computed per head, concatenated, and projected). Each encoder block applies Add&LayerNorm around multi-head attention and a Feed-Forward Network (FFN). The final `[CLS]` vector is taken as the pooled embedding

$$z_i \in \mathbb{R}^d.$$

The weights α (averaged over heads) serve as *tag importances*, providing a transparent summary of which tags influenced the decision. The pooled vector z_i is passed to a linear sigmoid head (p_i) to obtain

the risk label, explained in the subsequent section in detail. The attention pooler improves aggregation of symbolic and causal information while preserving interpretability via tag-level weights.

The pooling module is implemented as a single Transformer-style encoder block with multi-head self-attention and a position-wise feed-forward network (FFN). In practice, the CLS attention Pooler uses an embedding dimension of $d = 128$, four attention heads, and an FFN hidden size of 256 with dropout 0.1. A learned `[CLS]` token is prepended to the node embeddings, and its final output vector is used as the graph-level representation, while the CLS-to-tag attention weights provide interpretable tag importances. The same configuration is used across datasets to maintain consistency in the continual-learning setup.

To illustrate how the Transformer attention pooler provides qualitative interpretability, two example posts are shown below. In each case, the model highlights the most influential tags and their causal relations when producing a risk prediction.

`Example 1 (High-risk post):` *"I have not slept properly for days, and the constant anxiety is making everything feel overwhelming. Lately I keep thinking that things would be easier if I just disappeared."* The attention pooler assigns high importance to the tags *sleep*, *anxiety*, and *suicidal*, with a strong causal pathway *sleep* → *anxiety* →*suicidal*. These attended tags correspond to clinically salient risk indicators, leading the classifier to assign a high-risk label.

`Example 2 (Low-risk post):` *"Feeling a bit stressed about exams next week, but talking to friends has helped and I'm trying to stay positive."* The model focuses primarily on *stress*, with low attention weights on other tags and no causal escalation toward *depression* or *suicidal*. The attention pattern reflects a non-escalatory emotional state, leading to a low-risk prediction.

3.5 Continual Learning Strategy

In real-world applications, data arrive in stages $\{\mathcal{T}_1, \ldots, \mathcal{T}_K\}$ with evolving language, populations, and even label definitions. Sequential training on $\mathcal{T}_k$ risks *catastrophic forgetting* of knowledge learned on earlier datasets. Continual learning is incorporated to address the sequential arrival of mental-health datasets and the risk of catastrophic forgetting. This ensured that knowledge acquired from earlier domains was preserved while adaptation to new sources was achieved, thereby enhancing robustness and practical applicability of the framework. A simple, effective continual learning technique, MH-Freeze is used, which preserves a shared encoder while adding a small task-specific head per dataset. Let G_i be a post graph and let f_θ denote the encoder mapping G_i to a pooled embedding. The pooled embedding is computed in Eq. (6).

$$z_i = f_\theta(G_i) \in \mathbb{R}^d \tag{6}$$

where f_θ comprises causal-aware graph construction, two GCN layers, and the lightweight Transformer encoder. Thus z_i integrates symbolic structure, causal links, and attention-based tag weighting.

For dataset $\mathcal{T}_k = \{(G_i^{(k)}, y_i^{(k)})\}$ with a *single* binary risk label $y \in \{0,1\}$, the task head g_{ϕ_k} is a linear–sigmoid classifier unit that produces a probability

$$\hat{p}_i^{(k)} = g_{\phi_k}(z_i) = \sigma\left(w_k^\top z_i + b_k\right) \in [0,1],$$

with parameters $\phi_k = \{w_k \in \mathbb{R}^d, b_k \in \mathbb{R}\}$ and sigmoid $\sigma(\cdot)$. A threshold τ_k (calibrated on a small validation split) yields the binary decision. At inference, predictions are labeled as positive if $p \geq \tau$.

MH-Freeze

At first, the shared encoder f_θ and the initial linear–sigmoid head g_{ϕ_1} are jointly trained on $\mathcal{T}_1$ by minimizing the average BCE loss, which is the negative log-likelihood of a Bernoulli target and directly trains

calibrated probabilities $\hat{p} \in [0,1]$ from the sigmoid head. Eq. (7) presents formulation of BCE for a single label $y \in \{0,1\}$ and predicted probability $\hat{p} \in [0,1]$ is [48]:

$$\mathrm{BCE}(y,\hat{p}) = -\big[\, y \log \hat{p} \;+\; (1-y)\log(1-\hat{p}) \,\big]. \tag{7}$$

The shared encoder f_θ and the initial linear–sigmoid head g_{ϕ_1} are then jointly trained on $\mathcal{T}_1$ by minimizing the average BCE, as shown in Eq. (8):

$$\min_{\theta,\phi_1} \frac{1}{|\mathcal{T}_1|} \sum_{(G,y)\in\mathcal{T}_1} \mathrm{BCE}\big(y,\ \hat{p}^{(1)}(G)\big),$$

$$\hat{p}^{(1)}(G) = \sigma\big(w_1^\top f_\theta(G) + b_1\big). \tag{8}$$

After convergence, the encoder parameters θ are frozen, so that $z_i = f_\theta(G_i)$ provides a fixed representation for all subsequent heads. Freezing is applied after Task 1 because the first dataset provides the broadest and most diverse distribution of symbolic tags, allowing the encoder to learn generalizable representations before domain-specific heads are introduced. We also examined variants where the encoder is frozen after Task 2 or Task 3. These alternatives showed higher forgetting on earlier datasets and reduced overall stability, as the encoder continued adapting toward the later-task distributions and drifted away from the symbolic-causal structure learned initially. Freezing after Task 1 therefore offered the best balance between preserving prior knowledge and supporting effective multi-head adaptation.

For each subsequent dataset $\mathcal{T}_k$ $(k \geq 2)$, a lightweight head is instantiated and only its parameters are optimized while keeping the encoder fixed, as mentioned in Eq. (9):

$$\min_{\phi_k} \frac{1}{|\mathcal{T}_k|} \sum_{(G,y)\in\mathcal{T}_k} \mathrm{BCE}\big(y,\ \hat{p}^{(k)}(G)\big),$$

$$\hat{p}^{(k)}(G) = \sigma\big(w_k^\top f_\theta(G) + b_k\big), \quad \nabla_\theta \mathcal{L}_k = 0. \tag{9}$$

As θ is fixed, adaptation reduces to fitting task-specific *linear separators* in the common embedding space z, which avoids cross-task interference and sharply limits forgetting. Intuitively, the shared encoder captures domain-general structure (symbolic and causal relations plus attention-based tag weighting), while each head accounts for dataset-specific prevalence, wording, or scope. Given dataset, the corresponding head g_{ϕ_k} is selected and its probability $\hat{p}^{(k)}$ is thresholded to yield the label:

$$\text{risky if} \hat{p}^{(k)} \geq \tau_k, \qquad \text{non-risky otherwise.}$$

4 Experimental Results

The experiments were conducted on a high-performance workstation equipped with an AMD Ryzen Threadripper 2950X (16 cores, 3.50 GHz) and 32 GB RAM. The experiments were implemented in Python 3.11 using key libraries such as PyTorch 2.2, PyTorch Geometric 2.5, NumPy, Scikit–learn, and Matplotlib. All codes were executed in a Jupyter Notebook environment configured on Windows 11, ensuring a consistent and reproducible experimental setup.

4.1 Dataset Description

To evaluate the effectiveness and generalizability of the proposed Continual Neuro-Symbolic Graph Learning framework, six diverse datasets were used spanning Reddit-based mental health discourse and

clinician-guided interviews. A brief summary is provided in Table 2. These corpora differ in annotation protocols, linguistic style, and risk indicators, enabling a comprehensive assessment of both the symbolic reasoning components and the graph-based learning modules.

Table 2: Summary of datasets used as continual-learning tasks, showing source, text type, and size

Dataset	Source	Type	Size
DASH-2020	Zenodo	Reddit posts	3151
Dereaddit	Kaggle	Reddit posts	3553
Kaggle MH	Kaggle	Reddit posts	5957
SWMH	Zenodo	Reddit posts (split)	54,412
Go_emotions	Kaggle	Reddit posts	58,011
E-DAIC	USC-ICT	Clinical dialogues	418 transcripts

- Data Analytics for Smart Health (DASH-2020) [49]: It consists of reddit posts annotated for substance use, addiction, and recovery. For our binary setup, we merged all recovery-related categories into a single non-addicted class, while posts explicitly labeled as addicted are retained as the positive class.
- Go_emotions [50]: It is a reddit-based dataset annotated with 27 fine-grained emotion categories plus neutral. It contains about 58,000 unique comments collected from diverse subreddits. For binary mental-health risk classification in our work, all emotion categories associated with distress (e.g., sadness, anger, fear, anxiety) were grouped as risky, while the rest were treated as non-risky.
- Kaggle Mental Health [51]: This dataset sourced from Kaggle repository, contains Reddit posts labeled across five mental health conditions. For binary classification, all risk-associated categories were merged into a single risky class, while the remaining category was treated as non-risky.
- Dreaddit [52]: This dataset also sourced from Kaggle, consists of reddit corpus for stress detection across five community categories. The authors collected around 190 K posts and crowd-sourced stress labels for around 3.5 K text segments. The public release provides official splits (≈2838 train/715 test) with roughly balanced stress vs. non-stress. In our setup, we used the provided binary label (1 = stressful/risky), and the official train/test.
- Reddit SuicideWatch and Mental Health Collection (SWMH) [53]: This is a Reddit-derived dataset released via Zenodo, combining posts from the SuicideWatch subreddit and other mental health communities. Posts from SuicideWatch are categorized as the risky class, while those from broader mental health forums are assigned to the non-risky class.
- Extended DAIC (E-DAIC) [54]: E-DAIC is an extended version of the original Distress Analysis Interview Corpus with Wizard-of-Oz (DAIC-WOZ) corpus [55]. The dataset, sourced from University of Southern California-Institute for Creative Technologies (USC-ICT), includes semi-structured interviews conducted by a virtual agent named Ellie, controlled either by a human wizard or an autonomous AI system. It contains transcribed clinical interviews annotated using PHQ-8 scores.

4.2 Continual Learning Baselines

The proposed MH-Freeze framework is compared with several representative continual learning techniques, each reflecting a distinct strategy for mitigating catastrophic forgetting in sequential task scenarios.

- Elastic Weight Consolidation (EWC) [56]: EWC addresses catastrophic forgetting in sequential learning by estimating the importance of each parameter for previously learned tasks (via a Fisher-based approximation) and selectively slowing changes to those important weights when learning a new task. This preserves prior expertise while allowing plasticity on less critical parameters.
- Gradient Episodic Memory (GEM) [57]: GEM uses an episodic memory of past tasks and projects the current gradient to satisfy inequality constraints that do not increase loss on stored past-task examples. This enforces update compatibility with earlier tasks and can yield positive backward transfer when gradients align. In our experiments, we adopt the efficient A-GEM variant with the same memory protocol as ER and apply projection at every step before the optimizer update.
- Learning without Forgetting (LWF) [58]: LWF adapts a network to new tasks using only new-task data while preserving prior capabilities via knowledge distillation: the current model is trained to match the frozen previous model's outputs on the new data, alongside the new-task loss. This avoids storing old datasets, competes with multitask training that has access to old data, and often outperforms plain feature extraction or finetuning when old and new tasks are similar.
- Experience Replay (ER) [59]: It mitigates forgetting by maintaining a small episodic memory of past-task examples and interleaving them with current-task batches during training. This simple rehearsal stabilizes prior decision boundaries while preserving plasticity on new data, yielding a strong, low-complexity baseline. We have kept a fixed-size, class-balanced buffer. Each minibatch mixes current-task samples with buffer samples at a fixed ratio. Buffer size and ratio are tuned on validation.
- Finetuning: The finetuning (Sequential Learning) across tasks without any anti-forgetting mechanism serves as a lower-bound baseline [60]. In this work, a single shared head and encoder are updated sequentially across tasks under the same optimizer/schedule and validation-based thresholding; no replay or regularization terms are added.
- Synaptic Intelligence (SI) [61]: This CL technique assigns an 'importance' score to each weight based on how much it contributed during training on a task. At the end of a task, those importance scores are retained as a summary of what mattered most. When the next task arrives, SI adds a lightweight penalty that discourages large changes to previously important weights while leaving the others free to adapt. We applied SI to the encoder (and pooler), snapshot parameters at each task boundary, and tune the overall regularization strength and a small stabilizer constant on the validation split.

4.3 Evaluation Metrics

The effectiveness of the proposed framework and the baseline continual learning techniques is evaluated using multiple performance metrics. These metrics capture not only classification accuracy but also robustness under class imbalance and calibration of probabilistic outputs.

4.3.1 Accuracy

Accuracy measures the proportion of correctly classified instances, as computed in Eq. (10). It reflects how effectively each continual-learning method distinguishes risky posts from non-risky ones across sequential mental-health datasets:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative instances, respectively.

4.3.2 F1-Score

It is the harmonic mean of precision and recall, rewarding models that balance both low false positives and low false negatives, as shown in Eq. (11). Precision is the proportion of predicted positives that are correct, recall is the proportion of actual positives that are correctly identified. F1 score reflects how well each continual-learning method maintains balanced risky vs. non-risky decisions across sequential datasets.:

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

4.3.3 Area under ROC Curve (AUROC)

The AUROC evaluates the trade-off between true positive rate (TPR) and false positive rate (FPR) across varying thresholds. It is defined as the probability that a randomly chosen positive is ranked higher than a randomly chosen negative.

4.3.4 Area under Precision-Recall Curve (AUPRC)

The AUPRC integrates the precision- recall curve, which is more informative under class imbalance. It summarizes the trade-off between precision and recall across thresholds.

4.3.5 Brier Score

The Brier score evaluates the accuracy of probabilistic predictions by measuring the mean squared error between predicted probabilities p_i and true labels y_i. It can be computed using Eq. (12):

$$\text{Brier} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2. \tag{12}$$

4.3.6 Expected Calibration Error (ECE)

ECE measures the alignment between predicted probabilities and observed accuracy. Predictions are partitioned into M bins according to confidence, and the weighted average gap between accuracy and confidence is reported. It can be computed using Eq. (13). In our work, ECE is computed per task (dataset) to assess whether MH-Freeze and the baselines produce well-calibrated risk probabilities after threshold calibration in the continual-learning sequence.

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\text{acc}(B_m) - \text{conf}(B_m)\right| \tag{13}$$

where B_m is the set of samples in bin m, $\text{acc}(B_m)$ the empirical accuracy, and $\text{conf}(B_m)$ the mean confidence.

4.3.7 Matthews Correlation Coefficient (MCC)

MCC quantifies how well the classifier balances positive/negative decisions across sequential tasks and shifting, imbalanced class distributions, penalizing asymmetric error patterns that F1 or accuracy may hide. A higher value of MCC indicates better predictive performance. The MCC is computed in Eq. (14) [62].

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{14}$$

4.4 Results and Discussions

The proposed MH-Freeze framework demonstrates strong continual-learning behavior across heterogeneous datasets as seen in Table 3. MH-Freeze performs continual learning by freezing a shared GCN-Transformer encoder and training lightweight, task-specific heads as new tasks arrive. MH-Freeze is compared against six continual-learning baselines across six tasks, where each task corresponds to a different dataset in a fixed sequential order. The six tasks correspond to distinct datasets: T1 = DASH, T2 = Dreaddit, T3 = SWMH, T4 = Go_emotions, T5 = DAIC-WOZ, and T6 = Kaggle-MH. For brevity and consistency, these datasets are referenced as T1-T6 throughout the remainder of the paper. In all experiments, tasks are encountered in the fixed order T1-T6; balanced mini-batches are used to counter dataset-level imbalance, and the MH-Freeze architecture prevents dominance by any single dataset due to differing label distributions. The discriminative capability remains uniformly high, with AUPRC $\geq$ 0.934 and AUROC $\geq$ 0.942 throughout the task sequence, indicating robust separability between risk and non-risk classes. Both Accuracy (0.898 to 0.925) and F1-Score (0.886 to 0.923) follow a steady upward trajectory from T1 (DASH) to T6 (Kaggle-MH), reflecting positive forward transfer without degradation of earlier competencies. The MCC also improves from 0.829 to 0.873, confirming balanced predictive behavior under label imbalance. In parallel, Brier and ECE scores decline from 0.069 to 0.060 and 0.023 to 0.014, respectively, demonstrating progressive improvement in probability calibration. These metrics affirm that MH-Freeze effectively preserves prior knowledge while adapting to new domains, achieving well-calibrated, generalizable predictions with minimal catastrophic forgetting. It maintains a clear advantage in both discrimination and calibration.

Table 3: Continual learning performance of MH-Freeze across all datasets

Task	Accuracy	F1-Score	AUPRC	AUROC	MCC	Brier	ECE
T1 (DASH)	0.898	0.886	0.934	0.942	0.829	0.069	0.023
T2 (Dreaddit)	0.906	0.893	0.94	0.949	0.836	0.065	0.020
T3 (SWMH)	0.912	0.898	0.939	0.948	0.841	0.066	0.021
T4 (Go_emotions)	0.918	0.905	0.947	0.955	0.849	0.061	0.016
T5 (DAIC-WOZ)	0.923	0.911	0.945	0.954	0.852	0.062	0.018
T6 (Kaggle-MH)	0.925	0.923	0.947	0.965	0.873	0.060	0.014

Among the CL baselines used for comparison, GEM performs better, attaining moderately high AUROC (0.94–0.95) and balanced MCC values, though it gains plateau beyond mid-sequence tasks. LWF exhibits reasonable F1-scores but suffers from high calibration error and inconsistent reliability across datasets. Experience Replay and Synaptic Intelligence provide stable yet lower performance, with AUROC typically below 0.91 and limited robustness to domain shifts. EWC achieves comparable mid-range results but shows greater sensitivity to task transitions, while Finetuning performs worst overall, displaying rapid accuracy decay (0.73–0.76) and high Brier/ECE values indicative of severe forgetting. In contrast, MH-Freeze sustains near-optimal metrics across all six tasks, confirming that its frozen encoder with task-specific heads yields superior retention, adaptation, and calibration in continual-learning environments. The detailed results for each continual-learning baseline are presented in Tables 4–9, providing a comprehensive comparison across all tasks. Experience Replay and Synaptic Intelligence show early performance saturation, as seen in Tables 6 and 7, respectively. This likely reflects limited forward transfer and calibration instability under

domain shift. A plausible cause is that ER's small replay buffer cannot adequately represent later datasets, while SI's weight-importance penalty restricts the flexibility needed to adapt. Consistently higher Brier and ECE on the final tasks (T4-T6) reinforce this interpretation, indicating less reliable probabilities and weaker calibration as the data distribution changes. While MH-Freeze exhibits a monotonic increase in MCC from T1 to T6, GEM stabilizes at a slightly lower range (Table 4). Brier and ECE generally decline for MH-Freeze, indicating progressively better calibration, whereas replay and regularization-based baselines (ER, EWC, SI) show smaller or inconsistent reductions.

Table 4: Continual learning performance of GEM across all datasets

Task	Accuracy	F1-Score	AUPRC	AUROC	MCC	Brier	ECE
T1 (DASH)	0.888	0.876	0.929	0.937	0.819	0.072	0.025
T2 (Dreaddit)	0.896	0.883	0.935	0.944	0.826	0.068	0.022
T3 (SWMH)	0.902	0.888	0.934	0.943	0.831	0.069	0.023
T4 (Go_emotions)	0.908	0.895	0.942	0.950	0.839	0.064	0.018
T5 (DAIC-WOZ)	0.913	0.902	0.940	0.949	0.842	0.065	0.02
T6 (Kaggle-MH)	0.913	0.903	0.942	0.955	0.863	0.063	0.016

Table 5: Continual learning performance of LWF across all datasets

Task	Accuracy	F1-Score	AUPRC	AUROC	MCC	Brier	ECE
T1 (DASH)	0.865	0.854	0.906	0.915	0.791	0.181	0.172
T2 (Dreaddit)	0.889	0.877	0.926	0.934	0.821	0.17	0.129
T3 (SWMH)	0.875	0.864	0.921	0.928	0.809	0.273	0.23
T4 (Go_emotions)	0.882	0.87	0.928	0.935	0.816	0.191	0.127
T5 (DAIC-WOZ)	0.895	0.883	0.932	0.941	0.829	0.172	0.123
T6 (Kaggle-MH)	0.902	0.888	0.931	0.94	0.834	0.112	0.091

Table 6: Continual learning performance of Experience Replay across all datasets

Task	Accuracy	F1-Score	AUPRC	AUROC	MCC	Brier	ECE
T1 (DASH)	0.845	0.835	0.887	0.895	0.77	0.199	0.139
T2 (Dreaddit)	0.858	0.845	0.893	0.905	0.782	0.195	0.136
T3 (SWMH)	0.864	0.848	0.892	0.908	0.784	0.184	0.123
T4 (Go_emotions)	0.872	0.857	0.902	0.913	0.795	0.182	0.119
T5 (DAIC-WOZ)	0.875	0.862	0.902	0.915	0.792	0.171	0.117
T6 (Kaggle-MH)	0.867	0.852	0.895	0.912	0.783	0.182	0.121

Table 7: Continual learning performance of Synaptic Intelligence across all datasets

Task	Accuracy	F1-Score	AUPRC	AUROC	MCC	Brier	ECE
T1 (DASH)	0.866	0.845	0.891	0.905	0.785	0.085	0.038
T2 (Dreaddit)	0.845	0.833	0.878	0.892	0.776	0.088	0.042
T3 (SWMH)	0.851	0.843	0.877	0.911	0.795	0.085	0.035
T4 (Go_emotions)	0.848	0.832	0.851	0.895	0.772	0.087	0.039
T5 (DAIC-WOZ)	0.858	0.842	0.887	0.903	0.781	0.083	0.037
T6 (Kaggle-MH)	0.851	0.839	0.877	0.901	0.762	0.081	0.036

Table 8: Continual learning performance of EWC across all datasets

Task	Accuracy	F1-Score	AUPRC	AUROC	MCC	Brier	ECE
T1 (DASH)	0.856	0.841	0.863	0.902	0.780	0.173	0.126
T2 (Dreaddit)	0.854	0.839	0.886	0.907	0.785	0.084	0.138
T3 (SWMH)	0.859	0.852	0.885	0.918	0.803	0.081	0.132
T4 (Go_emotions)	0.859	0.848	0.895	0.921	0.771	0.167	0.133
T5 (DAIC-WOZ)	0.866	0.851	0.895	0.910	0.792	0.179	0.124
T6 (Kaggle-MH)	0.874	0.854	0.898	0.927	0.793	0.151	0.135

Table 9: Continual learning performance of Finetuning across all datasets

Task	Accuracy	F1-Score	AUPRC	AUROC	MCC	Brier	ECE
T1 (DASH)	0.725	0.775	0.765	0.775	0.645	0.311	0.265
T2 (Dreaddit)	0.763	0.741	0.797	0.805	0.681	0.267	0.251
T3 (SWMH)	0.755	0.739	0.785	0.803	0.677	0.263	0.257
T4 (Go_emotions)	0.735	0.735	0.775	0.799	0.665	0.325	0.311
T5 (DAIC-WOZ)	0.748	0.745	0.781	0.785	0.655	0.322	0.255
T6 (Kaggle-MH)	0.731	0.722	0.745	0.798	0.672	0.326	0.267

Methods that incorporate explicit memory mechanisms or parameter regularization, such as GEM and EWC, demonstrate better retention than Finetuning or LWF across all six tasks, confirming that constraining weight drift mitigates forgetting. However, these approaches still exhibit limited calibration stability, as indicated by elevated Brier and ECE values across late tasks. Synaptic Intelligence achieves moderate balance between accuracy and calibration, but its adaptation saturates beyond mid-sequence datasets, revealing difficulty in scaling to domain shifts. In contrast, MH-Freeze consistently maintains high discriminative accuracy while achieving the lowest calibration errors.

From T1 (DASH) to T6 (Kaggle-MH), most baselines show mild fluctuations in F1-Score and AUROC due to changing dataset characteristics and label imbalance. However, MH-Freeze exhibits smooth performance progression, achieving improvements in accuracy and F1-Score compared with the strongest baseline (GEM). This trend demonstrates strong forward transfer and minimal backward interference. Moreover, the consistently low Brier (≈0.06) and ECE (≈0.014) values emphasize its reliability in producing well-calibrated probabilities, critical in sensitive applications such as mental-health risk prediction, where overconfident misclassifications can have severe consequences.

Fig. 3 shows accuracy trends for all methods across the six datasets in sequence. The accuracy curves show a clear and persistent margin for the proposed MH-Freeze approach on every task. From $T1$ to $T6$, the accuracy of MH-Freeze rises from 0.898 to 0.925 (an increase of 3.0%), which indicates that knowledge gained on earlier tasks is retained while useful information from later tasks is added. In contrast, Finetuning changes only slightly (0.725 to 0.731; 0.8% increase) and shows signs of forgetting as new tasks are introduced. Regularization methods yield smaller gains, EWC improves by 2.1% and SI decreases by 1.7%, suggesting limited ability to adapt to the domain shifts in this sequence. Replay (ER), distillation (LWF), and gradient-projection (GEM) make training more stable, but they still perform worse than the proposed method. The highest final accuracy of 0.925 and the steady rise across datasets suggest that freezing the shared GCN-Transformer encoder and using separate heads for each dataset reduces interference and supports reliable forward transfer.

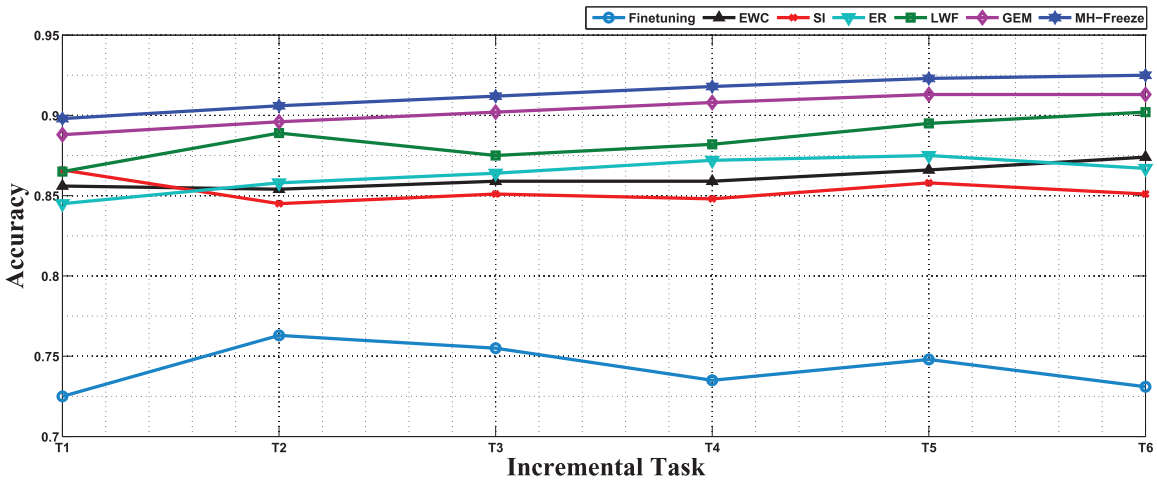

Figure 3: Performance comparison of Continual Learning techniques in terms of Accuracy

Similarly, a comparison based on F1-Score is presented in Fig. 4. This metric is informative under class imbalance and helps assess whether decisions remain balanced as new datasets are introduced. The F1-Score of MH-Freeze increases from 0.886 at $T1$ to 0.923 at $T6$, a 4.2% increase, indicating that earlier decision boundaries are preserved while new patterns are learned. The baselines follow a consistent ordering: Finetuning decreases from 0.775 to 0.722 (a 6.8% decrease), reflecting forgetting; EWC shows a small improvement of 1.5%; SI decreases slightly by 0.7%; ER and LWF achieve moderate increases of 2.0% and 4.0%, respectively; and GEM improves by 3.1% but remains below MH-Freeze. These results suggest that the multi-head freeze strategy maintains balanced predictions across datasets while enabling steady gains, which is the intended behavior in a continual-learning environment.

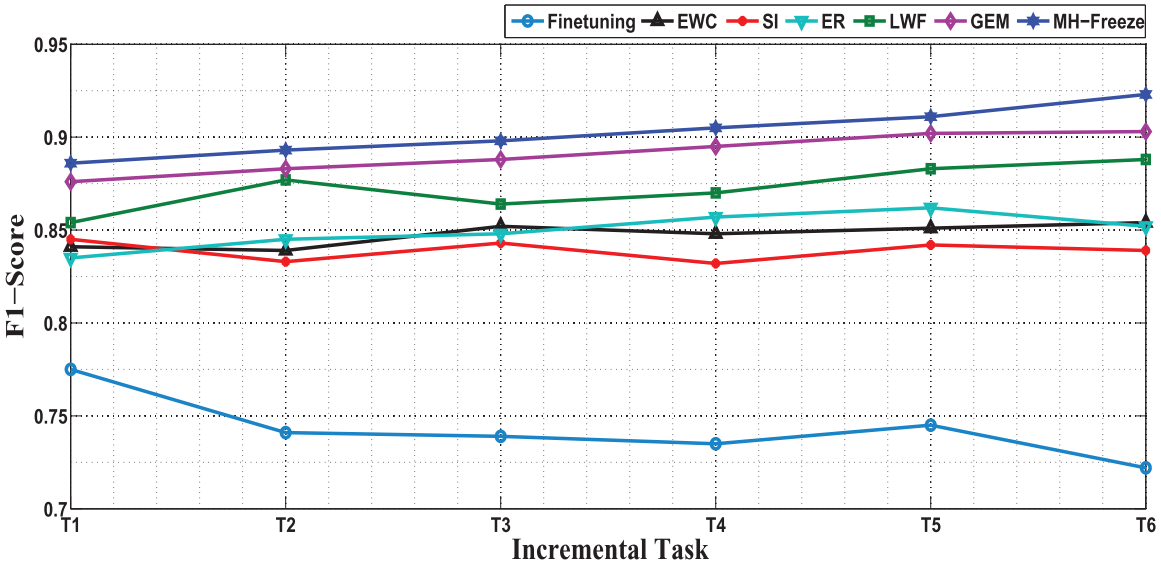

Figure 4: Performance comparison of Continual Learning techniques in terms of F1-Score

In addition, ROC curves are presented to provide an intuitive visualization of classification trade-offs across varying decision thresholds. Unlike single-value metrics, ROC curves reveal how each model balances the true-positive and false-positive rates, offering a deeper understanding of their discriminative behavior. The ROC curves shown in Fig. 5 illustrates the comparative classification performance of all continual-learning baselines and the proposed MH-Freeze model across six sequential tasks (T1–T6). Each subplot corresponds to a specific task, where the x-axis represents the False Positive Rate (FPR) and the y-axis represents the True Positive Rate (TPR). The ROC trajectories of all models are plotted within each panel, while only the area under the ROC curve (AUC) of the MH-Freeze model, representing the top-performing method, is explicitly annotated. The solid black curve corresponds to MH-Freeze and demonstrates its consistently superior separability across all tasks. In contrast, baseline models such as Finetuning, EWC, and SI are depicted with thinner colored curves to provide visual benchmarking and highlight relative performance differences. This visualization clearly demonstrates that MH-Freeze maintains stable and near-optimal discriminative capability across all incremental tasks, confirming its strong resistance to catastrophic forgetting and enhanced adaptability in continual-learning environments.

4.5 Computational Efficiency and Model Complexity

The CNSGL framework is designed to remain computationally efficient while still using the same encoder as the continual-learning baselines. The shared encoder, comprising a two-layer GCN (hidden size 128) and a single Transformer attention block with four heads (d_{model} = 128, FFN size 256), contains approximately 482 k learnable parameters, which are used in all models, including the proposed MH-Freeze. Each task also has a linear–sigmoid classifier head with 129 parameters, giving a total of about 482 k+129 parameters per model. The major difference is not in how many parameters exist, but in how many are updated during each new task. In the proposed framework, MH-Freeze freezes the encoder after first task (T1), and trains only the 129-parameter head, whereas baseline models continue to update all 482 k+129 parameters for every new task. Training on T1, where the encoder and head are jointly optimized, takes about 19 min, while training on subsequent tasks (T2–T6), where only the head is updated, completes in roughly

2 to 4 min per task. A single forward pass requires approximately 1.47 ms per post and about 1.37 million floating point operations per second (FLOPs).

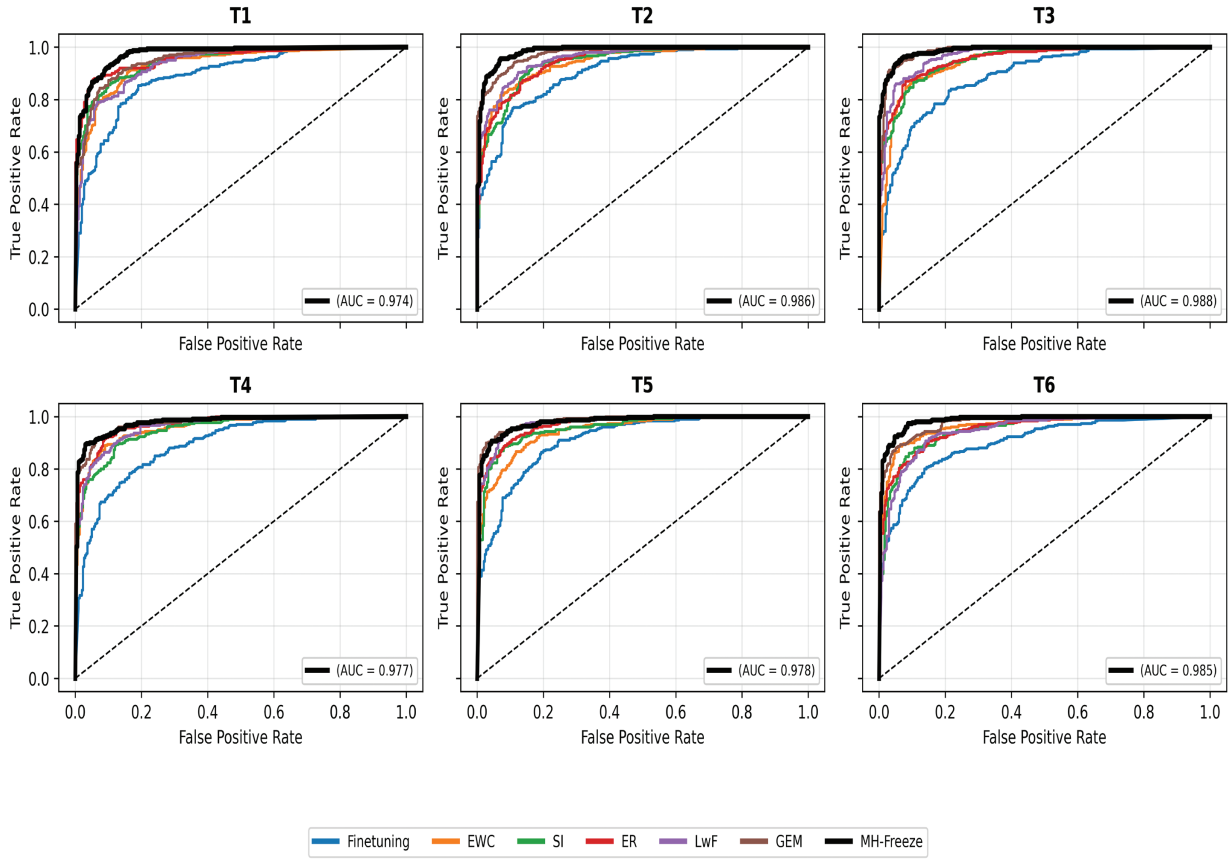

Figure 5: ROC curves across six sequential tasks (T1–T6) for continual-learning baselines and the proposed MH-Freeze model. Each task corresponds to a distinct dataset used in the continual-learning sequence

All continual-learning baselines use the same encoder architecture for a fair comparison, so their inference FLOPs are identical to CNSGL. However, they differ substantially in how many parameters are updated during each new task and in the resulting training time. CNSGL (MH-Freeze) updates only 129 parameters per new task, yielding the lowest per-task training time, whereas baselines must update the full encoder (≈ 482 k+129 parameters), leading to longer training times despite identical inference complexity.

5 Ablation Study

An ablation study was performed to examine individual contribution of each component in the proposed CNSGL framework, where symbolic reasoning, causal enrichment, GCN-based message passing, Transformer attention pooling, and the continual-learning mechanism are removed one at a time. Table 10 summarizes the performance of these variants.

Table 10: Ablation study evaluating the contribution of each component in the CNSGL framework

Model	Accuracy	F1-Score	AUPRC	AUROC	MCC	Brier	ECE
No Symbolic & Causal (GCN+Transformer+CL)	0.919	0.911	0.924	0.932	0.851	0.073	0.021
No GCN (Symbolic-Causal+ Transformer+CL)	0.872	0.861	0.892	0.893	0.824	0.173	0.152
No Transformer (Symbolic-Causal+ GCN+CL)	0.893	0.887	0.913	0.911	0.885	0.085	0.023
No Continual Learning (Symbolic-Causal+GCN+ Transformer)	0.831	0.829	0.853	0.861	0.811	0.195	0.141
Full CNSGL Symbolic-Causal+GCN+ Transformer+MH-Freeze)	0.925	0923	0.947	0.965	0.873	0.060	0.014

Removing symbolic tags and causal edges ("No Symbolic & Causal") yields a model that operates purely on GCN + Transformer embeddings without structured risk concepts. While performance remains reasonably strong (F1 = 0.911, AUPRC = 0.924), a noticeable drop appears compared to the full system, particularly in calibration (Brier = 0.073 vs. 0.060). This confirms that symbolic grounding provides clinically meaningful structure that enhances predictive reliability. When the GCN encoder is removed ("No GCN"), performance declines sharply across all metrics (F1 = 0.861, AUPRC = 0.892), and calibration degrades significantly (ECE = 0.152). This indicates that graph-based message passing is essential for leveraging symbolic–causal structure; replacing it with flat representations harms both accuracy and stability. Removing the Transformer attention pooler ("No Transformer") further demonstrates the role of attention in extracting concept-level importance. Although the model still performs moderately well due to symbolic–causal structure (F1 = 0.887), it shows lower AUROC (0.911) and poorer calibration relative to the full framework.

The performance degrades drastically when continual learning is removed ("No Continual Learning"), where sequential fine-tuning leads to catastrophic forgetting (F1 = 0.829, AUPRC = 0.853, Brier = 0.195). This highlights the necessity of the MH-Freeze strategy; without it, performance on earlier tasks collapses, and calibration becomes unstable. At last, the full CNSGL model, integrating symbolic tags, directional associations, GCN encoding, Transformer pooling, and MH-Freeze, achieves the strongest and most consistent performance across all metrics (F1 = 0.923, AUROC = 0.965, Brier = 0.060, ECE = 0.014). These results confirm that each component contributes meaningfully and that the full architecture offers the best balance of predictive accuracy, stability, and interpretability.

6 Conclusion and Future Work

The proposed Continual Neuro-Symbolic Graph Learning framework successfully integrates symbolic reasoning, causal inference, and continual learning to address the evolving nature of mental-health risk detection. By constructing symbolic graphs enriched with directional causal edges, the framework enables interpretable reasoning about risk factors and their interrelations. The hybrid encoder, comprising a two-layer GCN and a Transformer-based attention pooler, effectively captures both structural and contextual dependencies, producing discriminative yet interpretable graph-level embeddings. The MH-Freeze strategy,

which freezes the shared encoder and attaches task-specific heads, ensures strong retention of prior knowledge while allowing efficient adaptation to new datasets. Experimental results across six tasks (datasets) validate its robustness, showing that MH-Freeze consistently achieves the highest AUROC and F1-Score values, alongside superior calibration metrics (Brier and ECE), compared to all other continual-learning baselines. These findings confirm that MH-Freeze mitigates catastrophic forgetting and sustains stable, generalizable decision boundaries across diverse domains. Ablation analysis further confirms that each component contributes meaningfully to overall performance and calibration, and that the MH-Freeze continual-learning scheme is particularly critical for preserving performance and stability as new tasks are introduced.

Despite its advantages, this work also has a few limitations. The symbolic tag vocabulary is kept intentionally small and manually curated to ensure clarity and cross-dataset consistency. This focused design works well for the current scenario, but future extensions could incorporate richer or domain-specific tags to capture more subtle risk cues in broader clinical or social media data. In addition, the directional PMI module models precedence-based associations rather than fully validated causal relations. Future work can address these points by learning richer tag sets in a data-driven way and by integrating stronger causal discovery or longitudinal validation to refine the directed edges.

Future extensions will broaden the scope of the framework in several ways. Integrating multimodal signals, such as speech, facial expressions, and physiological markers can enrich early detection by complementing text with non-textual cues. Enhancing causal graph enrichment with temporal and counterfactual reasoning can deepen interpretability and strengthen the reliability of causal claims. Adopting federated or other privacy-preserving continual learning schemes can enable secure training across distributed mental-health datasets without direct data sharing. These advances would move CNSGL toward a more explainable, adaptive, and ethically deployable system for real-world mental-health risk assessment.

Acknowledgement: Not applicable.

Funding Statement: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00518960) and in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00563192).

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Monalisa Jena and Noman Khan; methodology, Monalisa Jena; software, Monalisa Jena and Noman Khan; validation, Monalisa Jena and Noman Khan; formal analysis, Mi Young Lee; investigation, Seungmin Rho; data curation, Monalisa Jena; writing—original draft preparation, Monalisa Jena; writing—review and editing, Monalisa Jena and Noman Khan; visualization, Monalisa Jena and Noman Khan; supervision, Mi Young Lee and Seungmin Rho; project administration, Mi Young Lee; funding acquisition, Mi Young Lee and Seungmin Rho. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The SWMH (SuicideWatch and Mental Health), E-DAIC (DAIC-WOZ), and DASH-2020 datasets are available on request from the respective authors due to ethical and access restrictions (SWMH: https://doi.org/10.5281/zenodo.6476178, accessed on 16 May 2025; E-DAIC: https://dcapswoz.ict.usc.edu/wwwedaic/, accessed on 16 May 2025; DASH-2020: https://zenodo.org/record/4278895#.X7T6cgzY2w, accessed on 18 May 2025), whereas the Dreaddit, Kaggle Mental_Health, and GoEmotions datasets are openly available via public repositories on Kaggle (Dreaddit: https://www.kaggle.com/datasets/rishantenis/dreaddit-train-test, accessed on 20 May 2025; Kaggle Mental Health: https://www.kaggle.com/datasets/entenam/reddit-mental-health-dataset?resource=download-directory, accessed on 22 May 2025; GoEmotions: https://www.kaggle.com/datasets/debarshichanda/goemotions, accessed on 22 May 2025).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

Abbreviations

AUPRC	Area Under the Precision-Recall Curve
AUROC	Area Under the Receiver Operating Characteristic Curve
BERT	Bidirectional Encoder Representation from Transformers
CL	Continual Learning
CNSGL	Continual Neuro-Symbolic Graph Learning
DASH	Data Analytics for Smart Health
DL	Deep Learning
EWC	Elastic Weight Consolidation
ECI	Event Causality Identification
ECE	Expected Calibration Error
ER	Experience Replay
FP	False Positive
FPR	False Positive Rate
FN	False Negative
FFN	Feed-Forward Network
FLOP	Floating Point Operations Per second
GEM	Gradient Episodic Memory
GCN	Graph Convolutional Network
HAN	Hierarchical Attention Network
IoT	Internet of Things
LWF	Learning without Forgetting
MCC	Matthews Correlation Coefficient
MH-Freeze	Multi-Head Freeze
ML	Machine Learning
NLP	Natural Language processing
PMI	Point-wise Mutual Information
SI	Synaptic Intelligence
SVD	singular value decomposition
SWMH	SuicideWatch and Mental Health Collection
TF-IDF	Term Frequency Inverse Document Frequency
TPP	Temporal Point Process
TN	True Negative
TP	True Positive
TPR	True Positive Rate

References

1. Li Y, Mihalcea R, Wilson SR. Text-based detection and understanding of changes in mental health. In: Social informatics (SocInfo 2018). Cham, Switzerland: Springer; 2018. p. 176–88. doi:10.1007/978-3-030-01159-8_17.
2. Hossain E, Alazeb A, Almudawi N, Alshehri M, Gazi M, Faruque G, et al. Forecasting mental stress using machine learning algorithms. Comput Mater Contin. 2022;72(3):4945–66. doi:10.32604/cmc.2022.027058.
3. Xu X, Yao B, Dong Y, Gabriel S, Yu H, Hendler J, et al. Mental-LLM: leveraging large language models for mental health prediction via online text data. Proc ACM Interact Mob Wearable Ubiquitous Technol. 2024;8(1):1–32.
4. Hossain MM, Hossain MS, Mridha MF, Safran M, Alfarhood S. Multi-task opinion enhanced hybrid BERT model for mental health analysis. Sci Rep. 2025;15(1):3332. doi:10.1038/s41598-025-86124-6.

5. Omarov B, Narynov S, Zhumanov Z. Artificial intelligence-enabled chatbots in mental health: a systematic review. Comput Mater Contin. 2023;74(3):5105–22. doi:10.32604/cmc.2023.034655.
6. Tejaswini V, Sathya Babu K, Sahoo B. Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model. ACM Trans Asian Low-Resour Lang Inf Process. 2024;23(1):1–20. doi:10.1145/3569580.
7. Vajrobol V, Saxena GJ, Pundir A, Singh S, Gaurav A, Bansal S, et al. A comprehensive survey on federated learning applications in computational mental healthcare. Comput Model Eng Sci. 2025;142(1):49–90. doi:10.32604/cmes.2024.056500.
8. Thekkekara JP, Yongchareon S, Liesaputra V. An attention-based CNN-BiLSTM model for depression detection on social media text. Expert Syst Appl. 2024;249:123834. doi:10.1016/j.eswa.2024.123834.
9. Mazurets O, Tymofiiev I, Dydo R. Approach for using neural network BERT-GPT2 dual transformer architecture for detecting persons depressive state. In: VI International Scientific and Practical Conference; 2024 Nov 15; Bologna, Italy. p. 147–51.
10. Helmy A, Nassar R, Ramdan N. Depression detection for Twitter users using sentiment analysis in English and Arabic tweets. Artif Intell Med. 2024;147:102716. doi:10.1016/j.artmed.2023.102716.
11. Kodati D, Tene R. Advancing mental health detection in texts via multi-task learning with soft-parameter sharing transformers. Neural Comput Appl. 2025;37(5):3077–110. doi:10.1007/s00521-024-10753-7.
12. Yang Y, Zhou J, Ding X, Huai T, Liu S, Chen Q, et al. Recent advances of foundation language models-based continual learning: a survey. ACM Comput Surv. 2025;57(5):1–38. doi:10.1145/3705725.
13. Thuseethan S, Rajasegarar S, Yearwood J. Deep continual learning for emerging emotion recognition. IEEE Trans Multimedia. 2021;24:4367–80. doi:10.1109/tmm.2021.3116434.
14. Han J, Zhang Z, Mascolo C, André E, Tao J, Zhao Z, et al. Deep learning for mobile mental health: challenges and recent advances. IEEE Signal Process Mag. 2021;38(6):96–105. doi:10.1109/msp.2021.3099293.
15. Hemmatirad K, Bagherzadeh H, Fazl-Ersi E, Vahedian A. Detection of mental illness risk on social media through multi-level SVMs. In: Proceedings of the 8th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS); 2020 Sep 2–4; Mashhad, Iran. p. 116–20.
16. Zhou S, Mohd M. Mental health safety and depression detection in social media text data: a classification approach based on a deep learning model. IEEE Access. 2025;13:63284–97. doi:10.1109/access.2025.3559170.
17. Garg M. Mental health analysis in social media posts: a survey. Arch Comput Methods Eng. 2023;30(3):1819. doi:10.1007/s11831-022-09863-z.
18. Skaik R, Inkpen D. Using social media for mental health surveillance: a review. ACM Comput Surv. 2020;53(6):1–31. doi:10.1145/3422824.
19. Benrouba F, Boudour R. Emotional sentiment analysis of social media content for mental health safety. Soc Netw Anal Min. 2023;13(1):17. doi:10.1007/s13278-022-01000-9.
20. Ding Z, Wang Z, Zhang Y, Cao Y, Liu Y, Shen X, et al. Trade-offs between machine learning and deep learning for mental illness detection on social media. Sci Rep. 2025;15(1):14497. doi:10.1038/s41598-025-99167-6.
21. De Choudhury M, Kiciman E. The language of social support in social media and its effect on suicidal ideation risk. In: Proceedings of the International AAAI Conference on Web and Social Media. Palo Alto, CA, USA: AAAI Press; 2017. p. 32–41.
22. Zhang Y, Cao D, Liu Y. Counterfactual neural temporal point process for estimating causal influence of misinformation on social media. Adv Neural Inf Process Syst. 2022;35:10643–55.
23. Cheng Q, Zeng Z, Hu X, Si Y, Liu Z. A survey of event causality identification: taxonomy, challenges, assessment, and prospects. ACM Comput Surv. 2025;58(3):59. doi:10.1145/3756009.
24. Kipf TN. Semi-supervised classification with graph convolutional networks. arXiv:1609.02907. 2016.
25. Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 1025–35.

26. Yao L, Mao C, Luo Y. Graph convolutional networks for text classification. In: Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, CA, USA: AAAI Press; 2019. Vol. 33. p. 7370–7. doi:10.1609/aaai.v33i01.33017370.
27. Subakan C, Ravanelli M, Cornell S, Bronzi M, Zhong J. Attention is all you need in speech separation. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ, USA: IEEE; 2021. p. 21–5.
28. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 6000–10.
29. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Stroudsburg, PA, USA: ACL; 2019. p. 4171–86.
30. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical attention networks for document classification. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Stroudsburg, PA, USA: ACL; 2016. p. 1480–9.
31. Gamel SA, Talaat FM. SleepSmart: an IoT-enabled continual learning algorithm for intelligent sleep enhancement. Neural Comput Appl. 2024;36(8):4293–309. doi:10.1007/s00521-023-09310-5.
32. Lee CS, Lee AY. Clinical applications of continual learning machine learning. Lancet Digit Health. 2020;2(6):e279–81. doi:10.1016/s2589-7500(20)30102-3.
33. Li C-H, Jha NK. DOCTOR: a multi-disease detection continual learning framework based on wearable medical sensors. ACM Trans Embed Comput Syst. 2024;23(5):1–33. doi:10.1145/3679050.
34. Nie W, Chang R, Ren M, Su Y, Liu A. I-GCN: incremental graph convolution network for conversation emotion detection. IEEE Trans Multimedia. 2021;24:4471–81. doi:10.1109/tmm.2021.3118881.
35. Kaur S, Bhardwaj R, Jain A, Garg M, Saxena C. Causal categorization of mental health posts using transformers. In: Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation. New York, NY, USA: ACM; 2022. p. 43–6.
36. Kodati D, Tene R. Identifying suicidal emotions on social media through transformer-based deep learning. Appl Intell. 2023;53(10):11885–11917. doi:10.1007/s10489-022-04060-8.
37. Kumar A. Neuro Symbolic AI in personalized mental health therapy: bridging cognitive science and computational psychiatry. World J Adv Res Rev. 2023;19(2):1663–79. doi:10.30574/wjarr.2023.19.2.1516.
38. Tang X, Guo R, Zhang C, Zhuang X, Qian X. A causality-driven graph convolutional network for postural abnormality diagnosis in Parkinsonians. IEEE Trans Med Imaging. 2023;42(12):3752–63. doi:10.1109/tmi.2023.3305378.
39. Bhuyan BP, Ramdane-Cherif A, Singh TP, Tomar R. Neuro-Symbolic AI: the integration of continuous learning and discrete reasoning. In: Neuro-symbolic artificial intelligence: bridging logic and learning. Singapore: Springer; 2024. p. 29–44. doi:10.1007/978-981-97-8171-3_3.
40. Dalkic H. CognEmoSense: a continual learning and context-aware EEG emotion recognition system using transformer-augmented brain-state modeling. J Brain Sci Ment Health. 2025;1(1):1–9.
41. Patanè G, Sorrenti A, Bellitto G, Palazzo S. Continual learning strategies for personalized mental well-being monitoring from mobile sensing data. In: Proceedings of the International Workshop on Personalized Incremental Learning in Medicine. New York, NY, USA: ACM; 2025. p. 9–17.
42. Febrinanto FG, Simango A, Xu C, Zhou J, Ma J, Tyagi S, et al. Refined causal graph structure learning via curvature for brain disease classification. Artif Intell Rev. 2025;58(8):222. doi:10.1007/s10462-025-11231-9.
43. Gosala B, Singh AR, Tiwari H, Gupta M. GCN-LSTM: a hybrid graph convolutional network model for schizophrenia classification. Biomed Signal Process Control. 2025;105(1):107657. doi:10.1016/j.bspc.2025.107657.
44. Chen L-C. An extended TF-IDF method for improving keyword extraction in traditional corpus-based research: an example of a climate change corpus. Data Knowl Eng. 2024;153(2):102322. doi:10.1016/j.datak.2024.102322.

45. Ma D, Chang KC-C, Chen Y, Lv X, Shen L. A principled decomposition of pointwise mutual information for intention template discovery. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM). New York, NY, USA: ACM; 2023. p. 1746–55.
46. Ghorbani M, Baghshah MS, Rabiee HR. MGCN: semi-supervised classification in multi-layer graphs with graph convolutional networks. In: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). Piscataway, NJ, USA: IEEE; 2019. p. 208–11.
47. Wu X, Lao Y, Jiang L, Liu X, Zhao H. Point Transformer V2: grouped vector attention and partition-based pooling. Adv Neural Inf Process Syst. 2022;35:33330–42.
48. Terven J, Cordova-Esparza D-M, Romero-González J-A, Ramírez-Pedraza A, Chávez-Urbiola EA. A comprehensive survey of loss functions and metrics in deep learning. Artif Intell Rev. 2025;58(7):195. doi:10.1007/s10462-025-11198-7.
49. Ghosh S, Misra J, Ghosh S, Podder S. Utilizing social media for identifying drug addiction and recovery intervention. In: 2020 IEEE International Conference on Big Data (Big Data). Piscataway, NJ, USA: IEEE; 2020. p. 3413–22.
50. Demszky D, Movshovitz-Attias D, Ko J, Cowen A, Nemade G, Ravi S. GoEmotions: a dataset of fine-grained emotions. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). Stroudsburg, PA, USA: ACL; 2020. p. 4040–54.
51. Rani S, Ahmed K, Subramani S. From posts to knowledge: annotating a pandemic-era Reddit dataset to navigate mental health narratives. Appl Sci. 2024;14(4):1547. doi:10.3390/app14041547.
52. Turcan E, McKeown K. Dreaddit: a Reddit dataset for stress analysis in social media. arXiv:1911.00133. 2019.
53. Ji S, Li X, Huang Z, Cambria E. Suicidal ideation and mental disorder detection with attentive relation networks. Neural Comput Appl. 2022;34:10309–19. doi:10.1007/s00521-021-06208-y.
54. Ringeval F, Schuller BW, Valstar M, Cowie R, Kaya H, Amiriparian S, et al. AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition. In: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop. New York, NY, USA: ACM; 2019. p. 3–12.
55. Gratch J, Lucas GM, King A, Morency L-P. The Distress Analysis Interview Corpus of human and computer interviews. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014). Stroudsburg, PA, USA: ACL; 2014. p. 3123–8.
56. Kirkpatrick J, Pascanu R, Rabinowitz N, Veness J, Desjardins G, Rusu AA, et al. Overcoming catastrophic forgetting in neural networks. Proc Natl Acad Sci U S A. 2017;114(13):3521–6. doi:10.1073/pnas.1611835114.
57. Lopez-Paz D, Ranzato M. Gradient episodic memory for continual learning. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 6470–9.
58. Li Z, Hoiem D. Learning without forgetting. IEEE Trans Pattern Anal Mach Intell. 2017;40(12):2935–47. doi:10.1109/tpami.2017.2773081.
59. Rolnick D, Ahuja A, Schwarz J, Lillicrap T, Wayne G. Experience replay for continual learning. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2019. p. 350–60.
60. De Lange M, Aljundi R, Masana M, Parisot S, Jia X, Leonardis A, et al. A continual learning survey: defying forgetting in classification tasks. IEEE Trans Pattern Anal Mach Intell. 2021;44(7):3366–85. doi:10.1109/tpami.2021.3057446.
61. Zenke F, Poole B, Ganguli S. Continual learning through synaptic intelligence. In: ICML'17: Proceedings of the 34th International Conference on Machine Learning; 2017 Aug 6–11; Sydney, NSW, Australia. p. 3987–95.
62. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21(1):6. doi:10.1186/s12864-019-6413-7.