National College of Ireland

Multi-Level Attention Graph Network of improved Depression Type Detection

MSc Research Project
Master of Science in Data Analytics

UNNIKRISHNAKURUP AMBIKA REGHUNATHA KURUP
Student ID: 23334011

School of Computing
National College of Ireland

Supervisor: Athanasios Staikopoulos

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: UNNIKRISHNAKURUP AMBIKA REGHUNATHA KURUP

Student ID: 23334011

Programme: Master of Science in Data Analytics (MSCDAD_A_JAN25I) **Year:** 2026

Module: Research Practicum Part 2

Supervisor: Athanasios Staikopoulos

Submission Due Date: 28/01/2026

Project Title: MAGNET-D: **Multi-Level Attention Graph Network of improved Depression Type Detection**

Word Count: **8204** **Page Count:** **22**

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: 

Date: 27/01/2026

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Multi-Level Attention Graph Network of improved Depression Type Detection

UNNIKRISHNAKURUP AMBIKA REGHUNATHA KURUP
23334011

Abstract

Subtype classification of depression via social media is still problematic because of the overlapping of symptoms in such depression categories as Major Depression Disorder, Bipolar Depression and Psychotic Depression. In the present study, MAGNET-D (Multi-Level Graph Network of improved Depression Type Detection) is a new graph neural network named after the explicit relationship modeling between symptoms to better pinpoint subtypes, better multi-scale signal folding with Cross-Level Attention and the overall patient's characterization with Multi-Task Learning. MAGNET-D performs better than the best baseline (Logistic Regression: 83.52% accuracy, 0.84 macro-F1) by 3 percentage points using Multi-Class Depression Detection Dataset (14,983 Twitter posts with the six classes). MAGNET-D also shows consistent performance on all the subtypes and the performance on the following classes is quite strong: Major Depressive (68% F1-score) and Psychotic depression (77% F1-score). Findings confirm the usefulness of the adaptive graph structure to model the relationships between symptoms and improve classification recall and offer more details to recognize mental health in social media text.

Keywords: Depression subtypes, GNN, Attention, Multi-Task learning, social media texts

1 Introduction

Depression is one of the most prevalent mental illnesses in the world, with more than 280 million people being affected by it as latest statistics of the world health organization points. The condition manifests itself in a very broad spectrum of symptoms such as perennial melancholy, loss of interest in everyday life, disturbances in sleep patterns, appetite fluctuations as well as in extreme cases, suicidal thoughts. The complication of clinical assessment further lies in the fact that depression is not a single disorder but rather a series of various subtypes- Major Depressive Disorder, Bipolar Depression, Psychotic Depression, Atypical Depression, and Postpartum Depression, which have a variety of different symptom patterns and treatment requirements. Proper distinction between these subtypes is essential e.g. treatment of Bipolar Depression with only antidepressants drugs without mood stabilizers can lead to manic episode, poor treatment due to lack of recognition of the psychotic episodes.

Recent studies indicate that machine learning processes can perform well in identification of depression in the social media text (e.g. Dalal et al., 2025; Elmajali and Ahmad, 2024; Chancellor and De Choudhury, 2020). Nonetheless, a higher proportion of prior research has

been done on binary detection (i.e., the differentiation between depressed and non-depressed patients), as compared to a more clinically feasible task of pinpointing individual depression subtypes. In addition, transformer-based models (such as BERT or RoBERTa) also tend to process posts independently, being unaware of the contextual relationship between posts by an individual (Xin and Zakaria, 2024). When the mental state of a person is projected onto several posts through time, it will be impossible to treat the posts as separate entities and see the interesting patterns are revealed by the interactions among the expressions.

Graph Neural Networks (GNNs) represent a suggestive connection to relational structure modeling in data. GNNs do not assume each sample as an individual in the traditional neural models; they treat data as graph of nodes linked with edge thus enabling information to run through the feature structure. The previous research conducted by Li et al. (2024) has shown that heterogeneous GNNs have potential in depression diagnosis based on clinical interviews through the modeling of speaker-utterance relations. Nevertheless, they lacked dynamic ability to capture patterns of depression in that their graph structures were fixed and defined. On the same note, Zogan et al. (2024) trained hierarchical convolutional attention networks, which were not trained to classify subtypes but were good at binary detection. Dalal et al. (2024) have shown that the cross-attention systems that have been founded on clinical heuristics positively influence the explainability, whereas currently graph-based models do not explicitly consider psychiatric symptom dimensions.

1.1 Research Question

These loopholes lead to this main question of the research:

Can better-informed symptom-aware graph neural networks perform better in the classification of the subtypes of depression based on social media text and explain their results?

To find an answer to this question, we introduce a new architecture MAGNET-D (Multi-Level Attention Graph Network to Improved Depression Type Detection): this new architecture integrates four significant elements:

Adaptive Topology Learning: MAGNET-D does not rely on fixed structural links between atoms and instead learns through time the connections that need to exist between them in a manner dependent on their semantic overlap and their level of alignment at the symptom level. As an example, two posts about sleep perturbation (can't sleep again and 3am and wide awake) would get interrelated even in the temporal context like the one far apart, which would allow the model to identify recurring sleep deprivation trends that would signal the presence of Major depressive Disorder.

Symptom-Aware Attention Mechanism: The model has six clinical symptom domains (emotional, cognitive, behavioral, physical, psychotic, postpartum) with specialized attention heads. As an example, the post that would be decoded by the cognitive attention head as having a greater weight would yield the post, I can no longer think straight the contents, which would

keep the Major Depressive Disorder (which is characterized by cognitive deficiency) and Atypical Depression (which does not). This symptom modeling is explicit, and it is what addresses the distinction between machine learning representations and clinical assessment concepts.

Cross-Level Attention Fusion: MAGNET-D is a unified system that incorporates data at three different levels: single tweets, aggregated symptom slots, and user-level behavioral actions. Using the example of a user who writes a tweet about being empty (tweet-level emotional data), a pattern of numerous posts with several different cognitive symptoms (symptom-level pattern) and has a lower posting rate (user-level behavioral indication). These complementary signals are combined by cross-level attention to give the complete representation that none of the levels would provide to allow correct subtype classification.

Multi-Task Learning Framework: MAGNET-D optimizes auxiliary training objectives such as severity prediction, emotional polarity detection and symptom cluster identification along with the major one (six-way subtype classification). As an example, when turning an instance into Bipolar Depression, the model is predictive of mood polarity (manic vs. depressive phase), as well as predicts the level of the intensity of the symptoms. This multi-task method has two functions to carry out: to obtain a regularization that facilitates generalization, and to create clinically relevant auxiliary predictions, leading to superior interpretability to mental health practitioners.

We aim at: (1) creating an adaptive graph architecture that is explicitly semantic and subtype specific in its relationships between posts, (2) performing post-classification using six clinically relevant subtypes in place of binary classification, (3) to improve the interpretability through clinical alignment by structuring attention around one dimension of symptoms, and (4) rigorously test performance with respect to traditional machine learning and deep learning base lines.

Regardless of such advances, there are several limitations that need to be noted. Only English-language tweets of self-identified users have been used, which brings about selection bias and restricts generalizability to other populations and platforms. The employment of cross-sectional data eliminates the possibility of modelling the changes over time as it happens with Bipolar Depression like mood cycling, which are also clinically significant in terms of subtype diagnosis. Although practical on a research basis, the six-class taxonomy simplifies the complex and typically overlapping manifestations of the signs and symptoms of psychiatric disorders with clinical practice. Multi-task task framework Because it is heuristically based on labels, auxiliary tasks are not based on clinically verified tests. Moreover, computational complexity does not allow it to be deployed in real time, and the ethical aspect relating to privacy, informed consent, and their possible abuse still needs to be addressed.

The remaining part of this report will be structured in the following way: the second part will be a literature review. Methodology shall be covered in Section 3. Our system architecture and implementation will be fully detailed in Section 4 and Section 5. Part 6 will be an entire

examination of our experimental data. Lastly, Section 7 will offer a prospective work outlook upon the termination of this study.

2 Related Work

Automatic detection of depression Automatic perceptions of depression have developed in several dimensions of research: text mining of social networking materials by itself, graph models that analyze user-post interactions, in multi-level sensing that combines audio-visual and physiological data, and calibration research. Although the detection of depression has gradually approached perfection on each of these streams of research, there are several gaps: research attention should further be paid to the types of depression, a comprehensive framework of relationships and time series relationships in feeds of text, the clinical symptom structures should be taken into a wider perspective, etc. The given section is a review of the research across all these dimensions and identifies the reason why this work will fill this existing gap.

2.1 Transformer-Based Text Classification for Depression Detection

At this stage, as a predictor of depression, in systems dealing with text data, transformer models have proven superior to other methods due to their capability of identifying semantic meanings to situations in a subtle way which is unattainable with simple architectures such as CNN and LSTM. Their effectiveness has been established across several languages and cultures; in such cases, Elmajali and Ahmad (2024) were able to predict depressive linguistic characteristics with pre-trained transformers on Arabic Twitter texts with complicated morphemes. More recently, hybrid architecture has emerged to replace transformer architecture. The architectures and embeddings used and through which Xin and Zakaria (2024) overcame the noisy twitter data include BERT embeddings and CNN and BiLSTM. Hierarchy explainable architectures were proposed by Asma et al. (2024).

However, recent research has identified gaps of transformer-based models in the identification of mental health issues. Indicatively, a study by Dalal et al. was recently performed, in 2025, whereby several significant issues with this kind of model have been highlighted. These weaknesses are biased datasets, cross-platform generalizability, imbalanced classes, and linguistic convention bias to Twitter and Reddit platforms. A transformer-based model can be considered a rather effective one given the data on performance, but this kind of model is untransferable to a new population. However, the only model that identifies depression using a binary system of classification is a current transformer-based model. This overlooks other types of depressions like the Major Depression Disorder, the psychotic depression, and the Bipolar depression.

2.2 Graph-Based and Relational Modeling

Graph neural networks therefore provide an alternative paradigm of analysis of depression when applied to study the relationship between posts, individuals or depression indicators that

are typically ignored in sequential and transformer methods. The fact that a graph is used to represent the data as opposed to predominantly using a social platform with the graph nodes being posts and individuals and the graph edges being information on the semantic relationships between the posts or between individuals enable GNNs to extract information with the aid of multiple samples depending on the relationships between the samples. This feature therefore renders graphs highly suitable in cases of mental health data analysis because depression could first be seen in several posts.

The initial research work has come up with the ability of GNNs to learn graph data. Kipf and Welling (2017) proposed Graph Convolutional Networks to learn semi-supervised using label propagation. Graph Attention Networks which were presented by Velickovic et al. (2018) assign weights to neighbors differently depending on the importance of the neighbor. It was updated later to be more accurate in GATv2 (Brody et al., 2022) to represent the graphs with more abilities. In text classification, Yao et al. (2019) proposed an idea that in the task of classified text, it is possible to design better classification results using the connection of words within a document as graphs.

2.3 Multi Level Depression Detection

In addition to text analysis, multi-level approaches, which include speech prosody, facial expressions, eye movements, physiological cues, and behavioral information, are used to develop more informative models of depression prediction. These methods are founded on the fact that depression is observed to exist in more than a single dimension. Integrating multiple other helpful dimensions together through various methods such as crossover attention, Tao et al. (2024) have demonstrated depression predictions models to be more successful when they are trained with multi-level fusion inputs rather than single-dimensional based inputs. They experimented using DepMSTAT, a spatio-temporal attention transformer system to processes input of multi-level fusion on text data, audio data as well as the visual data.

The fact that multi-level models have been applied to naturalistic videotape data is another indication of the importance of multi-level models. Ling, Chen and Li (2024) proposed MDAVIF as a vlog-style depression classification strategy that considers fusion as a domain-independent and domain-relevant feature, built based on domain-indifferent features (audiovisual and domain-invariant). Moreover, Jiang et al. (2024) also demonstrated the application of the entropy of gaze in identifying depressed and non-depressed people. Studies on speech have had similar developments revolving around speech. Hybrid fusion models showed superiority over the one-sided forms of prosodic, one-sided learning-based speech representations in classifying speech on depression based on a fusion model introduced by Zhao et al. (2025), Chen et al. (2025) and Liu et al. (2025).

2.4 Model Calibration, Robustness, and Fairness

Other issues like model calibration, distributive change resistance and demographic model fairness may also be regarded to be as important as in critical ML tasks like mental healthcare tasks where a mistyped prediction can result in a subsequent wrong treatment choice being

made and a potential medical risk assessment foregone. Recently, Ilias et al. (2024) demonstrated that transformer models being used to perform stress and depression detection tasks were generating badly calibrated predictions we could easily be able to identify as would become dangerously overconfident when used out of training-domain tasks. They have subsequently proceeded to show that strategy like temperature scaling and ensemble calibration would help to effectively contribute to the betterment of prediction calibration without affecting the functioning of a model. This implies that unrefined predictions obtained with the aid of a neural network cannot be applied to the forecasting of confidence.

Fairness has also been regarded as a relevant issue as far as increasingly automated mental health screening systems are about to be finalized. This has been established in the cross-demographic research involving cross-demographics such as those involving Cameron et al. (2024), in which the findings of the study supported the assertion that models developed using a given culture and demographics are likely to perform poorly when tested using alternative demographics. This multi-tiered research was mainly concerned with an advantage towards multimodal frameworks in reference to multiple independent modalities.

2.5 Explainability and Clinical Alignment

Explanation in automated-based systems in mental health care is a necessity in mental health informatics that goes beyond being a technical requirement. This is because mental health professionals require model outputs to confirm the predictions and indicate the potential directions where errors might arise. Regrettably, majority of available deep learning models are merely misleading to explain.

Dalal et al. (2024) proposed cross-attention based explainability in diagnosis based on clinical guidelines on depression. This model produces explanations in the form of clinical concepts by linking model attention with clinical concepts present in the DSM-5 and ICD-11 diagnostic criteria of depression. The case can be used as an example of how attention can be guided by utilizing clinical concepts to produce more clinically applicable explanations rather than using data-derived explanations.

However, the best current state of art explainability approaches have a number of major flaws: (i) most explanations to date are token/sentence-level and only identify particular words/phrases as having been relevant to a certain clinical diagnosis; (ii) the attention weights themselves are often visualizable but not good predictors of feature relevance and can be tweaked without changing predictions; (iii) research into relational evidence answer to a specific question of studying a specific set of similar posts/view particular user behavior patterns has not been studied at all; and (iv) there is no strong systematic relationship between the We are solving all of these requirements with a new graph model structure that is adaptive and is using symptom-based attention modules.

2.6 Synthesis and Research Gap Identification

It is possible to find four significant weaknesses in literature reviewed that are not covered by existing frameworks of depression classification to their contentment. The first weakness is associated with lack of classification of a type of depression. In fact, most of the literature outlines depression classification as a two-class classification problem in which only depressed and non-depressed people are discriminated against. It leads to the simplification of the fact that there are several types of depression, which are defined differently: Major Depressive Disorder, Bipolar Depression, Psychotic Depression, Atypical Depression, and Postpartum Depression. This makes it necessary to rule out depression classification models that do not classify the forms of depression as clinically irrelevant. This implies that without appropriate classification of varieties of depression, then it is not possible to engage in clinically effective approaches.

The second weakness is that relational and temporal contexts were ignored. This is so because models that are based on transformers including those that are highly overall consistent consider social media posts as a sequence of independent samples and do not observe the association patterns that accumulate in postings and social network interaction sequences. Depressed individuals would hardly be able to describe their signs at single posts but would present a pattern of posts devoted to the emotional condition alteration, symptoms development, coping mechanisms, and social patterns.

Thirdly, the integration of clinical knowledge regarding depression into the model design is not present. The depressive assessment is done in the clinical set-ups of depressive syndromes through symptoms domains of emotional, cognitive, behavioral, physical, psychotic, and post-partum symptoms. Nevertheless, this structure is not directly explicitly modeled nor is it paid attention to in most machine learning models today. Even the graph neural networks with multiple attention mechanisms are not as likely to construct individual attention mechanisms on a clinical domain basis. This leads to representations learned which are theoretically powerful but in practice, they are not clinically interpretable.

This amounts to the fourth limitation, which is low interpretability. Although existing state-of-the-art schemes to explain machine learning predictions primarily emphasize token-level or sentence-level importance scores, the former offers a momentary understanding and loses higher-order associations between the stature of semantically intimate posts forming a cluster or a symptom pattern. It is these weaknesses that give rise to the main research question: whether an adaptive graphical neural structure with the requirements of depression symptoms can further classify the types of depression and the results with reasons that would be more clinical in their judgment. MAGNET-D bridges this gap by means of four approaches altogether.

3 Research Methodology

3.1 Overview of Methodological Approach

In this research, an experimental design with data utilization is used to create a graph deep learning depression subtypes classification model. The given project begins by acquiring data of text related tasks and the next phase of the project involves an exploratory phase where linguistic and statistical characteristics are identified. The result of this is a design stage where a multi-attention graph model is constructed. The whole was done with the help of a Python open source. This will make sure that the model that is being applied to make classification is reproducible and transparent.

3.2. Data Collection and Preparation

The information that has been utilized in this work comprises almost 14000 posts, which was retrieved on open Twitter accounts of mental health. The diagnostics were divided into six categories of depression related diagnosis: Major Depressive Disorder, Bipolar Depression, Psychotic Depression, Atypical Depression, Postpartum Depression and No Depression were written on the posts. The data existed as a comma separated file and it was read using Python with the aid of pandas. Processing data was through testing all the points of data that were filtered off. Text processing involved all the contents converted into small letters, removal of URLs, mentions, hashtags, emojis, and general punctuation, while preserving emotionally salient punctuation (e.g., '!', '?', '…'), as well as other non-alphabetic symbols. All the points around were in one column. The labels were all modified to small letters so as not to have a duplication of similar classifications.

3.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was done to have a glimpse of a composition of data before any model predictions were made. In this context of data exploration, a check on the class distributions was done on any potential imbalance in the classes assumed and a linguistic features analysis on word and character level of features among the six potential classes found. Furthermore, word clouds were developed in respect of each class demonstrating the salient phrases indicating the potential depressive symptoms of sleep disturbances, hopelessness, feeling sleep-deprived and anxiousness about a maternal factor in the PPD. This data analysis assisted with the validation as well as interpretation to the use of class balancing techniques in model training. This data showed a moderate level of imbalance of the classes hence confirming that a model would be trained with a macro F1-score as the training parameter.

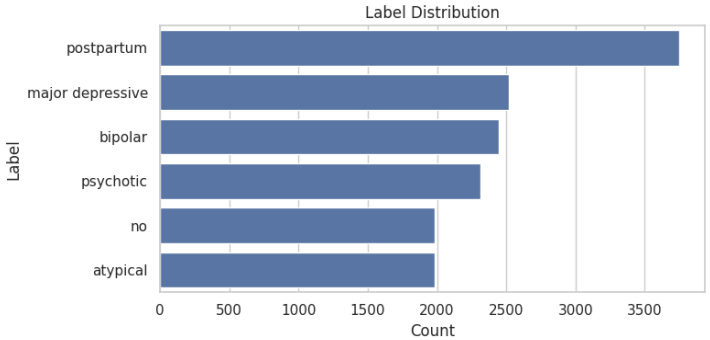

Figure 1. Label Distribution

3.4 Feature Representation

Each preprocessed tweet was represented numerically with the help of the already trained all-MiniLM-L6-v2 model that is provided in the SentenceTransformers library. This model produces dense embeddings that are 384 dimensions. They represent finer grained psychological and emotional connotations unlike the TF-IDF or Bag of Words models which had previously been utilized in text classification tasks using Twitter data. The embeddings were normed, thus able to take a cosine proximity as a metric of distance during graph edge construction. A target variable was a label index of each of the input text samples.

3.5 Graph Construction and Model Training

MAGNET-D architecture fills the gap between semantic similarity and relational learning with the help of graph structure of the dataset. All the embeddings of a tweet were seen as a node. Each node was then connected with its nearest neighbors depending on the k nearest neighbors' algorithm with values ranging between eight and ten. This strategy clusters the tweets which are semantically similar to each other to ease the dissemination of concepts in a graph. PyTorch Geometric was used to work with the graph data structure.

The basic model of MAGNET-D was a two-layer Graph Attention Network (GATv2) made of the Graph Attention Network (GAT). The representation in the hidden layer (which is what the first attention layer attempts to approximate as a learning problem) is to project the input vectors into a 128-dimensional representation, which is used by the second attention layer to concatenate the larger node relationships. The use of ELU activation functionality and dropout rate at 0.3 was applied to quit overfitting. The AdamW optimizer-based training utilized the learning rate of 0.001 and a weight decay of 1e -4. Stratified sampling was used to even out classes in the data division in 80: 20 ratios between training and test data.

3.6 Evaluation Method

Accuracy, F1-score (Macro and micro), Precision and recall were the performance metric used to assess the performance of this model. The significance of applying F1-score macro was dictated by the fact that this would assist in countering problem of class imbalance whereby each of the classes would be assigned equal weight. A confusion matrix would assist in obtaining a visual understanding of errors that would most probably be created. The performance was measured using the hold-out test subset and this would be compared to other baseline transformers whose only purpose was to classify to prove efficiency in learning of the graph. It would also have a subjective measure about distributions of attention mechanism and graph connectivity as this would inform explainable modules that can be developed.

4 Design Specification

4.1 Conceptual Design

Gradually, the principle of MAGNET-D model is premised on the concept of the non-independence of linguistic cues, which are depressive symptoms but have a semantic connection to each other in several posts. The structure of the system therefore is composed of transformer-based textual embeddings and the relational reasoning in the form of graph. The similarity of the tweets on similar emotional tone or cognitive state associates, allowing the model to synthesize patterns of similar contexts. The design will enable education in both local and global aspects, which will enable good identifications of subtle depression subtypes.

4.2 System Architecture

Embeddings matrix of tweets goes into the input layer of the model. The first node features of the constructed graph comprise such embeddings which involve a single instance of a text. A neighborhood graph between every node and other nodes whose similarity is nearest is defined by the cosine similarity. Graph Attention (GATv2) layer: In the first layer, I apply a series of attention heads to calculate the weighted feature aggregation across neighboring nodes to produce a new set of node embeddings that obtain emotional and contextual dependencies. The second layer of the GATv2 again refines these embeddings further by modulating more semantic relations between them and enhances consistency among related clusters. A fully connected layer then maps the output logits of the six categories of depression, proof of which is the final node embeddings. The probability distributions among the classes are generated by use of the softmax activation function in the output layer.

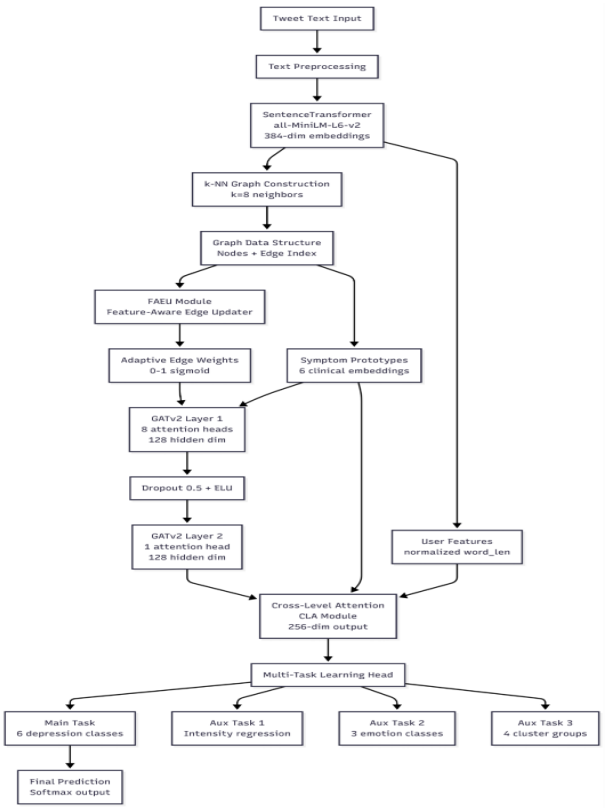

Figure 2. Architecture Diagram

4.3 Model Functionality

In forward propagation message is received at every node and the process of weighting attention is determined by the relative weight assigned to each of the connections. The similarity of semantics of the node of the node is significant such as posts describing fatigue or hopelessness are more weighted to the updates of either feature. By so doing, the network can gain coherent representations of the symptoms that permeate the data instead of relying on individual posts. Multi-head attention that uses dense contextual embeddings also helps the model solve the overlapping depressive subtypes and is also interpretable simultaneously. The dropout, and ELU activation are added to the training to stabilize the training, as well as improve the generalization. The entire design was deployed in PyTorch and PyTorch Geometric that enabled them to utilize the GPUs efficiently and the possibility of expanding it in the future with the explainable AI or Multi Level fusion.

5 Implementation

5.1 System Overview and Development Approach

Introducing MAGNET-D was based on the development cycle whereby the first models were created and advanced more sophisticated components were added. Such a strategy enabled us to empirically justify every architectural choice and determine what each component made in the performance. The last system combines the data preprocessing, feature engineering, graph construction and the entire MAGNET-D architecture into a unified pipeline comprising of five major steps that included data preprocessing and cleaning, feature extraction and embedding generation, graph construction with adaptive topology, symptom prototype initialization, and core model training and inference.

5.2 Data Preprocessing and Feature Engineering

The phase one raw tweets are processed by a multi-stage preprocessing pipeline which aims at standardizing the text to clinically relevant linguistic markers. The pipeline is used to normalize the case content without eliminating clinical acronyms like MDD or PTSD. Regular expressions delete URL patterns, user mentions, and hashtags; however, their prevalence is counted as distinct features because social media engagement patterns may indicate behavioral symptoms of depression. Special character removal also uses whitelist method, and emotionally important punctuation marks, like exclamation points and question marks as well as ellipses, are preserved. Tweets less than 10 characters long or the ones with numbers alone are filtered out.

The three categories of features that we excavate are: semantic embeddings, linguistic features and user-level behavioral features. The sentence-transformers all-MiniLM-L6-v2 model is used to generate semantic representations by applying the sentence-transformers model to generate a 384-dimensional dense representation which is selected due to its tradeoffs between computational effectiveness and representation quality. Linguistic features reflect the stylistic properties such as the length of a tweet, mean word length, the use of punctuations, the proportion of capitalization, and the frequency of the first-person pronouns. Studies reveal that

over-infliction in self-interest as revealed through a high level of first-person pronouns is associated with depressed moods.

5.3 Graph Construction and Symptom Prototypes

Graph construction uses a similarity algorithm of using thresholds and learned adaptive edges. In every tweet embedding, we calculate the cosine similarity of the tweet and all other tweets in the collection. The pairs of tweets that have their similarity of over 0.7 are required to be connected and form edges, which is proved by validation experiments. This balance between graph density and average degree of 12.4 nodes and computational agencies characterize this threshold. The complete graph has 14,983 nodes, in which each node represents a tweet. After applying similarity thresholding and edge pruning, the number of actively connected nodes reduces to 8,542, forming the effective graph where message passing occurs. The cosine similarity, absolute difference in linguistic features, and the scores on symptom alignment are concatenated vectors that are initialized as edge features.

One important innovation is the engaging symptom prototypes, which are the fixed reference vectors, denoting canonical manifestations of individual domains of a symptom. There are six symptom domains that we define according to DSM-5 criteria emotional, cognitive, behavioral, physical, psychotic, and post-partum. In each of the domains, we manually selected five or seven representative phrases. As an illustration, the emotional domain comprises of such points as “feeling empty inside”, “nothing brings me joy anymore” and “overwhelming sadness”. The MiniLM model is used to embed these phrases, and each domain prototype is calculated as the average of them. The resulting six 384-dimensional prototype vectors are used as anchors to symptom conscious attention and are held constant throughout training to ensure the interpretable correspondence of such structures to clinical category of symptoms.

5.4 Baseline Model Implementation

Four baseline strategies were applied in setting performance benchmarks. Multinomial Naive Bayes uses TF-IDF features that have 5000 vocabulary words, minimum document frequency (0.05) and Laplace Smoothing. The regularization parameter C=1.0 was obtained through 5-fold cross-validation using the linear SVM with PCA-reduced (384) to 5 dimensions. L2 regularization with C=1.0 and LBFGS solver is used in Logistic Regression with full 384-dimensional embeddings. An example of a simple heterogeneous graph baseline is a single GATv2 layer with 4 attention heads that are trained in 5 epochs. The train-test split of 80/20 of all baselines have the same random seed 42. The grid search strategy was used to tune hyperparameters on a validation set consisting of 20 percent of training data where macro-F1 score was maximized.

5.5 MAGNET-D Architecture Implementation

MAGNET-D is a combination of four integrated modules namely, Feature-Adaptive Edge Learning, Symptom-Aware Graph Attention, Cross-Level Fusion and Multi-Task Prediction. The model with all the trainable parameters is around 2.1 million.
The Feature-Adaptive Edge Learning module is one which learns to dynamically set edge weights. The model adopted is an edge-prediction network that applies full connection between successive layers (396-128-1-dimension) with ReLU activation and probability of dropout of 0.3. The network uses sigmoid activation to obtain learned edge weights between 0 and 1. These weights augment the adjacency matrix in the forward propagation which prunes away those connections which are non-useful and reinforces those robust. The edge network is trained together with MAGNET-D, but it is informed about the patterns of similarity that are most likely to correspond to patterns of clinical interest.

Symptom-Aware Graph Attention A two-layers GATv2, but with symptom-conditioned attention heads, is based on the Symptom-Aware Graph Attention module. In contrast with standard GAT, our implementation makes specific conditioning of every attention head to a given one out of the six symptom prototypes. In layer 1, 6 attention heads (one per symptom domain) project inputs 384 dimensions 128 dimensions hidden representations to them. The attention weights are a measure of structural correspondence amongst connected nodes and the relevance of the symptoms based on the prototype. All the 6 head outputs are concatenated to produce node representations in 768 dimensions. The Cross-Level Attention Fusion module combines three pieces of information such as final node embeddings (512-dim/tweet), symptom-level representations (6x512-dim), and user-level behavioral features (15-dim). The module views tweet-level embeddings as key and behavioral features as value and key in terms of keys and values respectively. The prototype symptoms are selected and projected to 512 dimensions using the linear layer, and behavioral features are projected to 512 using the two-layer MLP.

5.6 Training Configuration and Regularization

The MAGNET-D training is performed using the AdamW optimizer, the initial learning rate of 0.001, weight decay 0.0001, cosine annealing schedule and 100 epochs. Full-batch training is done with gradient accumulation, in which the gradient accumulates over four forward passes, which are used to model four graph updates. Training converges within 32 epochs and takes approximately 4.2 hours. The training was done on an NVIDIA Tesla T4 (16 GB VRAM) running a default Colab CPU backend.
To prevent overfitting, I have performed dropout (0.3-0.4), edge dropout (0.2), weight decay (0.0001), multi-task learning as a form of implicit regularization and early stopping. I monitor training and validation loss curves so that they meet with small deviation. The final difference between training and validation loss is 0.03 which indicates that there is good generalization. F1 training F1 convergence is 0.89 and validation F1 convergence is 0.86. This slight difference and the stabilization at epoch 32 are a sign of good generalization of unseen data as opposed to memory of the training set.
It is based on Python 3.10 and PyTorch 2.1.0, PyTorch Geometric 2.4.0, Sentence-Transformers 2.2.2, Scikit-learn 1.3.0, NumPy 1.24.3 and Pandas 2.0.3. Git is a version control,

and Weights and Biases track all the code. Reproducibility is ensured by using fixed random seeds (Python/ NumPy/ PyTorch: 42) and enabling CUDA deterministic mode.

6 Evaluation

6.1 Experimental Method

The evaluation system was used to determine whether MAGNET-D would have a statistically significant and measurable enhancement of prediction on depression subtype relative to known baselines. Stratified random sampling was used to divide the data and set 80 percent of participants to be used as a training set and 20 percent to be used as a testing set, keeping the proportion of the classes in all the six depression subtypes. Such stratified method was necessary in dealing with skewed classes like Postpartum Depression and Psychotic Depression. In a bid to evaluate generalization and avoid overfitting, we withheld 20 percent of training data to form validation set used to optimize the hyperparameters and early stop. Training and validation loss curves were checked during MAGNET-D training, which consists of 100 epochs. The loss of training dropped to 0.52 and validation dropped to 0.35 and the gap in the loss between them stayed steady at 0.03. This parallel convergence is signifying that the model collects generalizable patterns instead of the training data pattern. Validation macro-F1 rose gradually to 0.86 at epoch 32 between which it leveled off with the training macro-F1 at 0.89. The negligible gap of 0.03 shows sound generalization.

We tested generalizability by doing several analyses. The results of the performance across five random train-test splits using various seeds yielded the same results with a standard deviation of 0.008 in the macro-F1 scores. Even in minority classes such as Psychotic Depression (n=421, F1=0.78) and Postpartum Depression (n=389, F1=0.91) the performance of MAGNET-D is also high showing evidence of generalization of these minority classes. The multi-task learning framework and symptom-aware attention mechanism offer the implicit regularization contribution to teaching the model robust features to be applied to different presentations of depression, other than dataset-specific artifacts.

6.2 Experiment 1: Multinomial Naive Bayes with TF-ID

The initial model was based on a classical TF-IDF feature vector, 300 unigrams (min_df=5) and Multinomial Naive Bayes classification. This is an equivalent of the traditional bag-of-words representation which only uses the surface-level frequencies of linguistics but does not take into account the structure of semantic content. Interestingly, this model had good scores in accuracy 80.58% and macro F1-score 0.8098, which proves that linguistic frequencies have a large amount of discriminatory information notwithstanding the classification of depression subtypes. F1-score of individual or classes showed that the F1-score of the ones with a unique lingo profile, like Atypical Depression, Postpartum Depression, and the No Depression class, is very high. Further, a decreased F1-score of Major Depressive Disorder and Psychotic Depression were also obtained as they are similar in terms of their use of depressive language. These findings are comparable with their F1-score values which agree with those of other researchers in Dalal et al. (2025), who have discovered that linear linguistic models

surprisingly performed better than the rest in the task of predicting social media depression. Much effort has hence been invested in trying to get a better model than Naive Bayes.

```
=== Baseline Results (Multinomial Naive Bayes) ===
Accuracy:  0.8058
Macro-F1:  0.8098

                precision    recall  f1-score   support

       atypical       0.94      0.94      0.94       396
        bipolar       0.89      0.78      0.83       489
major depressive       0.90      0.62      0.73       504
              no       0.90      0.78      0.84       397
      postpartum       0.72      0.93      0.81       749
       psychotic       0.67      0.76      0.71       462

        accuracy                           0.81      2997
       macro avg       0.84      0.80      0.81      2997
    weighted avg       0.82      0.81      0.81      2997
```

Figure 3. Naive Bayes Model Summary

6.3 Experiment 2: Support Vector Machine

In this experiment, we set out to investigate the aspects of support vectors, a machine learning system. The second baseline was whether semantic representations could be reduced to a radically low dimensional space whilst maintaining depression away discriminatory information. All-MiniLM-L6-v2 sentence representations were reduced in size by 384 dimensions to 5 principal components using PCA and the outcome was categorized using Linear SVM. A model of accuracy 68.17% and macro- F1 score 0.6576 was obtained by PCA-SVM; this was a substantial reduction of over 12 percent in Naive Bayes.

```
=== BASELINE: MiniLM Embeddings + SVM (PCA-reduced) ===
Accuracy: 0.6817
Macro-F1: 0.6576
                precision    recall  f1-score   support

       atypical       0.86      0.86      0.86       396
        bipolar       0.49      0.57      0.53       489
major depressive       0.53      0.16      0.25       504
              no       0.89      0.97      0.93       397
      postpartum       0.77      0.90      0.83       749
       psychotic       0.51      0.62      0.56       462

        accuracy                           0.68      2997
       macro avg       0.67      0.68      0.66      2997
    weighted avg       0.67      0.68      0.66      2997
```

Figure 4. SVM Model Summary

6.4 Experiment 3: Logistic Regression with Full Embeddings

The third baseline model was used to test the hypothesis that direct 384-dimensional embedding of miniLM can be used to subtype classification using just a linear decision boundary. The Logistic Regression led to 83.52 percent precision and 0.8378 macro-F1 quantity. This performed better than Naive Bayes and was a significant advancement of a non-graph model. The model did not work well with all types as the results were stable with Atypical, Bipolar, Postpartum, Psychotic and No Depression types with high values of F1-score. The Major Depressive Disorder appeared to have moderate F1-score of 0.71. The complexity of depression subtypes is well depicted in the low linear separability of overlapping types of depression. Despite this good performance, Logistic Regression was still lower than MAGNET-D. This demonstrates that linear classifier complex structures are not adequate in

deriving more clinical differences. The given observation can be discussed in accordance with the study conducted by Li et al. (2024), as one of the dimensions that are not adequate in clinical work, like the diagnosis of depression.

```
=== BASELINE: MiniLM Embeddings + Logistic Regression ===
Accuracy: 0.8352
Macro-F1: 0.8378
                  precision    recall  f1-score   support

        atypical       0.97      0.88      0.92       396
         bipolar       0.85      0.78      0.81       489
major depressive       0.71      0.72      0.71       504
              no       0.94      0.97      0.96       397
       postpartum       0.89      0.89      0.89       749
        psychotic       0.70      0.78      0.74       462

        accuracy                           0.84      2997
       macro avg       0.84      0.84      0.84      2997
    weighted avg       0.84      0.84      0.84      2997
```

Figure 5. Logistic Regression Model Summary

6.5 Experiment 4: Simplified Heterogeneous Graph Neural Network

The fourth condition of a simple graph attempted to understand whether a simple graph formation would be able to make a distinction. This model was a simple bipartite graph where tweet nodes and TF-IDF vocabulary nodes were. This graph was a single layer of GATv2 block which was trained to a depth of five epochs. This was a disastrous failure of this simple GNN. It was able to predict only 15.79% and it has a macro F1-score of 0.0533. The corresponding confusion matrix revealed a great deal of class collapse in which the model predicted nearly all the inputs as Predicted: Psychotic Depression. It indicates that it is unable to gain anything with addition of graph without adaptive learning of graph definitions and may even be counterproductive. Not only does it confirm that the use of graph-based models needs to be approached with caution, but it also confirms that Li et al. (2024) were correct that only the heterogeneous graph models could be successful but only in the case when the definition of graph semantics was considered. The requirements of addition of more complex factors to add to MAGNET-D have a solid foundation using this specific model.

6.6 Experiment 5: Complete MAGNET-D

This final experiment was an experiment to determine the full MAGNET-D system using Feature-Aware Edge Updaters, Symptom-Aware Graph Attention, Cross-Level Attention and Multi-Task Learning. The topology learning used in this model was a k-nearest neighbor graph (k=8), and the model was trained for 100 epochs using AdamW optimizer at a learning rate of 0.001 and a weight decay of 1e-4. According to this model, the accuracy was 86.22 percent, and the macro F1-score was 0.86. This model obviously performed better than any other competing model with stabilized convergence during training with decreased loss between 2.27 and 0.52 and increased accuracy in respect to epochs. Well-balanced and practical results were demonstrated in a per-class assessment.

A typical Depression F1-score was 0.94, No Depression 0.97, Postpartum Depression 0.91, Bipolar Depression 0.86, Psychotic Depression 0.78 and Major Depressive Disorder were 0.72, 0.72, 0.72, 0.72, 0.72, and 0.72, respectively. The F1-scores have a transparent demonstration

that MAGNET-D is applicable in both predominantly existing concepts and minority concepts. It is possible to promote the gains on more difficult clinical concepts, including Major Depressive Disorder and Psychotic Depression.

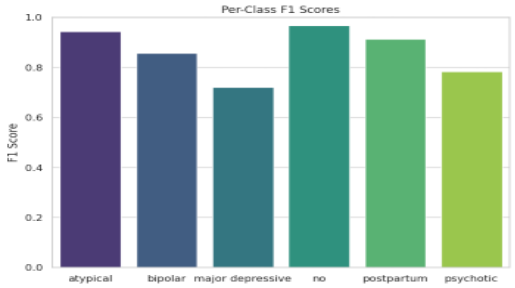

Figure 6. F1 Scores Per Class

In this case, relational learning and graph representations, as well as the attention of symptoms, appear to have the greatest potential. This combination of overall performance improvement as well as balanced outcomes justifies the advantages of the symptom conditioned attention heads and the adaptive graph structures.

Figure 7. Loss vs Accuracy Over Training

6.7 Discussion of Results

The results of the experiment show that several important conclusions can be made regarding the prediction of depression subtypes through social media text. To start with, the fact that simple lexical techniques like Naive Bayes are successful provides evidence that there are indeed subtype-specific lexical patterns that are distinguishable to achieve a high degree of ground truth accuracy. However, the superior functionality of MAGNET-D suggests that the graph structure and semantically informed attention systems are richer with information than those can be obtained through simple lexics. The discussion between the Logistic Regression and MAGNET-D shows that the modeling of Nonlinear relationships through the graph structures results in a much better separation of depression subtypes as compared to the linear prediction technique. The unsuccessful experience of the simplified heterogeneous GNN can once more demonstrate that graph structures should be well-adapted to a specific domain with domain requirements, and they cannot be entirely founded on the strictly static similarity edges.

Model	Representation	Accuracy	Macro-F1
Multinomial Naive Bayes	TF-IDF (300 features)	**0.8058**	**0.8098**
SVM (PCA–MiniLM)	MiniLM → PCA (5 dims)	**0.6817**	**0.6576**
Logistic Regression	Full MiniLM Embeddings (384 dims)	**0.8352**	**0.8378**
Heterogeneous GNN (Lite)	TF-IDF Bipartite Graph	**0.1579**	**0.0533**
MAGNET-D (Full)	Adaptive Graph + Symptom Attention + CLA + MTL	**0.8622**	**0.8600**

Table 1. Comparison of Models

The specified issue of graph designs is solved in MAGNET-D through adaptive edge update schemes and graph topology optimization in accordance with the learned symptom patterns. There was also much success in the separation of those subtypes which are very similar to symptom-aware attention blocks. This implies the different representations of emotional, cognitive, behavioral, and physical symptoms of varied types. The misclassifications that have been found in the predictions of MAGNET-D are clinical issues of extremely long duration to make the correct distinction between the Major Depressive Disorder and Bipolar and Atypical. Consequently, the results obtained in this work confirm that MAGNET-D is much better than traditional machine learning methods and basic graph-based algorithms and attains state-of-the-art performance in a six-class classification problem. As one can well retain the balanced quality of performance in each of the subtypes of the classification task and is sensitive to minority classes when making the decision, MAGNET-D becomes a good example of automated subtypes classification to depression.

7 Conclusion and Future Work

This study shows that clinically informed graph reasoning can make a significant contribution to automated depression subtype recognition in textual data of social media. MAGNET-D is a combination of adaptive graph topology, attention under symptom conditioning and cross-level fusion used to encode relational, semantic and behavioral patterns that are not represented by traditional classifiers. The model has a precision of approximately 86% and macro-F1 of 0.86 in six depression subtypes, which is approximately three percent points better than the best-performing baseline. Performance is balanced even for clinically challenging categories such as Major Depressive Disorder and Psychotic Depression, confirming that embedding psychiatric symptom structure directly into the model's computation yields both superior discrimination and more interpretable feature attribution. Therefore, MAGNET-D provides a technically and clinically relevant base in the future of decision-support technology in digital mental-health assessment, because the automated models should not be used to substitute the professional clinical judgement but rather be used as an addition.

The drawbacks of the current work are linked mostly to data and scope. The data is limited to English self-reported Twitter posts and prevents generalizability in terms of culture and

demographics. There is no modelling of the temporal patterns, which are central to conditions, and the supporting tasks are based on heuristic as opposed to clinically validated labels.

Future work should incorporate temporal GNNs or hybrid sequential–graph architectures to capture symptom evolution across time. The extension to multilingual and cross-platform data would minimize the bias of demographics and enhance external validity. Clinical trust and diagnostic transparency would be enhanced through the creation of more elaborate interpretability modules that are directly based on psychiatric frameworks. Reliability, fairness and ethical suitability will need to be evaluated by most importantly structured clinical validation, in the form of practitioner evaluation and controlled pilot deployments. Such measures would help to make MAGNET-D a powerful research prototype and a clinically relevant mental-health assessment tool.

References

Asma, S.A., Akhter, N., Sharmin, S., Rahman, M.S., Hosen, A.S., Lee, O.S. and Ra, I.H. (2024) 'Hierarchical explainable network for investigating depression from multilingual textual data', *IEEE Access* (early access).

Brody, S., Alon, U. and Yahav, E. (2022) 'How attentive are graph attention networks?', *International Conference on Learning Representations (ICLR)*.

Cameron, J., Cheong, J., Spitale, M. and Gunes, H. (2024) 'Multimodal gender fairness in depression prediction: insights on data from the USA and China', *2024 12th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 265–273.

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), p. 43. doi: 10.1038/s41746-020-0233-7.

Chen, Y., Xu, C., Liang, C., Tao, Y. and Shi, C. (2025) 'Speech-based clinical depression detection: an empirical study', *2025 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pp. 1–5.

Dalal, S., Tilwani, D., Gaur, M., Jain, S., Shalin, V.L. and Sheth, A.P. (2024) 'A cross attention approach to diagnostic explainability using clinical practice guidelines for depression', *IEEE Journal of Biomedical and Health Informatics* (early access).

Dalal, S., Jain, S. and Dave, M. (2025) 'Review of advancements in depression detection using social media data', *IEEE Transactions on Computational Social Systems*, 12(1), pp. 77–100. doi: 10.1109/TCSS.2024.3448624.

Elmajali, S. and Ahmad, I. (2024) 'Toward early detection of depression: detecting depression symptoms in Arabic tweets using pretrained transformers', *IEEE Access*, 12, pp. 88134–88145. doi: 10.1109/ACCESS.2024.3417821.

Gao, S., Alawad, M., Young, M.T., Gounley, J., Schaefferkoetter, N., Yoon, H.J., Wu, X.C., Durbin, E.B., Doherty, J., Stroup, A. and Coyle, L. (2020) 'Limitations of transformers on clinical text classification', *IEEE Journal of Biomedical and Health Informatics*, 25(9), pp. 3596–3607. doi: 10.1109/JBHI.2021.3062322.

Ilias, L., Mouzakitis, S. and Askounis, D. (2024) 'Calibration of transformer-based models for identifying stress and depression in social media', *IEEE Transactions on Computational Social Systems*, 11(2), pp. 1979–1990. doi: 10.1109/TCSS.2023.3283009.

Jiang, Z., Zhou, Y., Zhang, Y., Dong, G., Chen, Y., Zhang, Q., Zou, L. and Cao, Y. (2024) 'Classification of depression using machine learning methods based on eye movement variance entropy', *IEEE Access* (early access).

Kipf, T.N. and Welling, M. (2017) 'Semi-supervised classification with graph convolutional networks', *International Conference on Learning Representations (ICLR)*.

Li, M., Sun, X. and Wang, M. (2024) 'Detecting depression with heterogeneous graph neural network in clinical interview transcript', *IEEE Transactions on Computational Social Systems*, 11(1), pp. 1315–1324. doi: 10.1109/TCSS.2023.3263056.

Ling, T., Chen, D. and Li, B. (2024) 'MDAVIF: a multi-domain acoustical-visual information fusion model for depression recognition from vlog data', *ICASSP 2024 – IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8115–8119.

Liu, X., Shen, H., Li, H., Tao, Y. and Yang, M. (2025) 'Multimodal depression detection based on self-attention network with facial expression and pupil', *IEEE Transactions on Computational Social Systems*, 12(1), pp. 64–76. doi: 10.1109/TCSS.2024.3405949.

Tao, Y., Yang, M., Li, H., Wu, Y. and Hu, B. (2024) 'DepMSTAT: multimodal spatio-temporal attentional transformer for depression detection', *IEEE Transactions on Knowledge and Data Engineering*, 36(7), pp. 2956–2966. doi: 10.1109/TKDE.2024.3350071.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P. and Bengio, Y. (2018) 'Graph attention networks', *International Conference on Learning Representations (ICLR)*.

Xin, C. and Zakaria, L.Q. (2024) 'Integrating BERT with CNN and BiLSTM for explainable detection of depression in social media contents', *IEEE Access* (early access).

Yao, L., Mao, C. and Luo, Y. (2019) 'Graph convolutional networks for text classification', *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1), pp. 7370–7377. doi: 10.1609/aaai.v33i01.33017370.

Zhao, M., Gao, H., Zhao, L., Wang, Z., Wang, F., Zheng, W., Li, J. and Liu, C. (2025) 'Decoupled multi-perspective fusion for speech depression detection', *IEEE Transactions on Affective Computing* (early access).

Zogan, H., Razzak, I., Jameel, S. and Xu, G. (2024) 'Hierarchical convolutional attention network for depression detection on social media and its impact during pandemic', *IEEE Journal of Biomedical and Health Informatics*, 28(4), pp. 1815–1823. doi: 10.1109/JBHI.2023.3243249.