International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 3, May-Jun-2026, ISSN (Online): 2395-566X

Towards Fine-Grained Depressive Symptom Recognition in Memes via Multimodal Transformer-CNN Fusion

Mrs. J. Annie Jennifer[1], Dr. R. Gunasundari[2]
[1]Research Research Scholar Department of Computer Science Karpagam Academy of Higher Education Coimbatore, Tamil Nadu, India
[2]Professor, Department of Computer Applications Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India

Abstract- The mental health indicators can be found in memes, and it is quite complex since memes consist of both text and images, and one must analyze both elements to understand their meaning. This research proposes a novel deep learning technique named Multi-CNN. Its aim is to detect depression-related signs by analyzing their linguistic and visual content simultaneously in memes . The technology uses both the BERTweet model for natural language processing and ResNet18 features for images from a neural network. It was assessed using a dataset of internet memes annotated according to eight depression indicators. Early stopping, data augmentation, and others helped improve its performance, while results were estimated by means of a weighted F1 score. As the study shows, it is more effective to use linguistic and visual components simultaneously than to employ the model based only on language or solely on image analysis for identifying the presence of depressive signs in memes. The multimodal approach resulted in a weighted F1 score of 0.6846, while the language-based model received 0.6716. Using just the picture is ineffective when it comes to recognizing depression-related memes. The study's findings indicate that visual information and text together create strong cues for investigating mental health issues. Besides, the results point to fresh techniques and technologies that can handle the intricate heterogeneous datasets found in social media.

Keywords- Meme Classification, Mental Health Analysis, Deep Learning, Social Media, Computer Vision, Natural Language Processing (NLP).

I. INTRODUCTION

Nowadays, memes are one of the most popular ways to express oneself online. Although memes are associated with entertainment, research proves they can indicate serious issues related to depression[1][2]. This intersection may inspire further investigation of methods for automated identification of mental health problems based on Internet-posted memes.
Detecting mental health signals from memes is particularly challenging because meme content is a mixture of text and graphics [3][4]. Furthermore, the text component often features slang and sarcasm, making it difficult to interpret the messages conveyed in memes.

We introduce MultiT-CNN, a model that aims to identify depression-related memes. In our model, we merge two modes of data: textual data obtained using BERTweet[5] and visual data obtained using ResNet18[6]. Our model classifies the memes into eight classes corresponding to depression-related symptoms. Our contributions include:

• Multimodal Model
The deep learning model, MultiT-CNN, merges textual features from a transformer-based text classifier (BERTweet) and visual features from a convolutional image encoder (ResNet18). It classifies eight depression-related symptoms. The model is robust to sarcasm, humor, and metaphors in memes.

• Preprocessing Framework
To deal with problems in real-world data, such as a lack of images and noise in the text, our model implements an extensive preprocessing framework for handling imbalanced classes.

• Benchmarking Performance with Baselines
An evaluation was carefully conducted comparing our multimodal architecture with its counterparts using only text

and images. Conclusion: Text-only systems perform competitively, but multimodal integration is yet to be fine-tuned to improve classification performance (Weighted F1: 0.6846 in ours vs. 0.6716 in the baseline), and the latter falls short compared to image-only methods.

• Class-Specific Training Technique
Class-specific loss function and early freezing and unfreezing of the BERTweet encoder were used during optimization.

• Application in Mental Health Surveillance
Our study provides computational insights into analyzing social media memes that convey information regarding mental well-being and offers a scalable framework for developing surveillance technology in mental health care.

The experimental results show that the multimodal approach, combining text and images, achieved good results that are better than the text-only model and the image-only model. This suggests that although recognizing depressed symptoms in memes using the text is the most important factor, adding visual information could nevertheless be helpful. The model offers a more complete view by using both types of data and expresses how depression is expressed in memes, which could help in monitoring mental health on social media platforms.

This paper is organized into sections as follows. Section 2 introduces related work with research problem. Section 3 explains detailed design of our model framework. Section 4 gives the evaluation results in predicting depression from meme post with text-only and image-only models. Section 5 provides a conclusion for entire paper with future work.

II. RELATED WORK

Mental Health Detection in Social Media
Mental health signals via social media monitoring is becoming one of the essential topics in the area, with an emphasis on detecting depression, anxiety, and suicidality among users' messages on different platforms (e.g., Reddit and Twitter)[7][8]. Earlier studies relied heavily on language processing and sentiment analysis based on tools such as LIWC. To improve contextual knowledge and increase classification precision, recent studies opted for transformer-based models like BERT[9]. However, all these techniques tend to neglect the role of the visual component, while it is extremely important to analyze memes, for instance, on Instagram and Tumblr.

Wei et al.[10] suggested CANAMRF, a cross-attention multimodal reasoning framework that incorporates both textual and visual inputs along with hybrid transformers for evaluating mental states. The contribution of Moon and Bhattacharyya[11] in the detection of mental health signals on social media vlogs based on a commonsense-aware large language model and behavior analysis should be highlighted. According to the authors, combining multimodal cues increases the efficiency of clinical understanding of mental health signals, as shown in their F1 score of 67.8%. Despite not being focused on meme analysis, these findings are relevant to the topic of the paper.

Memes and Multimodal Depression Detection
First, memes have the characteristic feature of multimodality, combining images and text, which complicates the analysis of the presence of mental health-related signs, particularly depressive symptoms. Previous literature has focused on the link between memes and psychology, as well as psychological factors indicating emotions such as anxiety[12][13][14]. Moreover, several approaches have been developed using visual-textual fusion techniques for finding abusive content. Specifically, Saha et al.[15] emphasized the importance of using multimodal reasoning for detecting mental health-related cues in memes; however, not many models were found capable of fine-tuning depressive symptom types, which is the gap addressed by this research.

To address that, Yadav et al.[16] have created a specialized dataset for detecting depression signs in memes called RESTORE. The use of orthogonal representation learning in this method allows recognizing and interpreting distinct text and image information within memes, providing the possibility of a finer analysis of depression symptom types by dividing them into eight clinically valid categories. On the other hand, Sharma et al.[17] proposed an ALFRED framework that includes a gated attention module for grounding emotions in multimodal sentiment detection for memes.

Multimodal Learning Architectures
Regarding the categorization of depressive memes, multimodal deep learning integrates data from various sources, such as text and imagery, to enhance the predictive capability of the model[18]. These approaches can be categorized as early fusion and late fusion. In early fusion, the fusion occurs at the

feature level prior to processing, whereas late fusion combines the output of different models using the attention mechanism[19][20][21].

The latter proves advantageous in vision-language joint modeling but requires significant computational power and is challenging to tune with smaller amounts of data. Extending from there, MFFNC, proposed by Li and Xiao[22], utilizes a multi-feature fusion approach where text representations obtained via MacBERT are fused with image embeddings using cross-modal attention.

There are two basic fusion concepts that involve early fusion, which is a combination of features prior to any processing operation, and late fusion, which is the combination of outputs after any processing operation. The above two basic methods of multimodal fusion prove inadequate in cases where multimodal data like memes are involved. Attention-based, orthogonal, and symbolic methods can be successfully used to increase performance and explainability in multimodal tasks[23][24]. Multimodal models like BERTweet and ResNET18 employ feature-level fusion based on pre-trained encoder features.

III. DATASET

We employ the RESTORE dataset, which M. Yadav [16][25]compiled as a carefully curated multimodal dataset designed to analyze mental health cues in internet memes. Figure 1 and 2, shows examples of depression in memes.

Fig. 1. Sleeping Disorder Meme

Fig. 2. Self-Harm Meme

In the present research paper, we apply the MultiT-CNN model to detect depressive symptoms, which involves a whole process including the training, validation, and testing steps. Each meme in our dataset has 3 crucial components:

- Image: the picture part of the meme.
- OCR text: the text in an image that is detected by Optical Character Recognition.
- Labels: one or more depressive symptom categories from clinical psychology, with 8 classes in total.

We highly regard the RESTORE dataset due to the realistic nature of the Internet memes included in this corpus, which is full of sarcasm and metaphors. In this case, we made use of 8,814 samples from the RESTORE dataset. Here is the division:

Table 1: Shows data statistics for the depression dataset

Classes	**Count**
Train	6169
Validation	1322
Test	1323
Total	**8814**

Training data: we have 8,814 samples to train the model and optimize the hyperparameters. Data augmentation and class weighting were employed to improve the generalization ability of the model.

Validation data: there are 8,814 samples used to monitor the training progress and pick up the best checkpoints according to the weighted F1 score criterion.

Nothing was trained or tuned on this dataset. This partitioning ensures that our model remains robust and unbiased because the same distribution is maintained in all three sets. The statistics of the depression dataset are illustrated in Table 1 above. Figure 3 illustrates the pie charts of these statistics. More precisely, the training data statistics are illustrated in Figure 4, the validation statistics in Figure 5, and the test statistics in Figure 6.

Fig. 3. Pie chart of the depression dataset

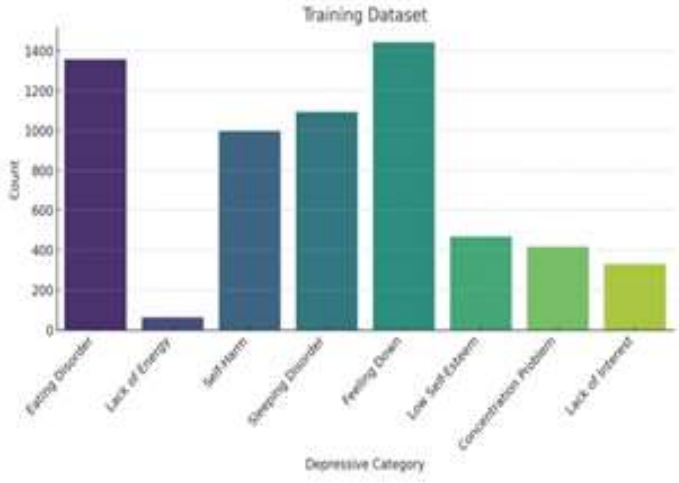

Fig. 4. Data statistics for training data

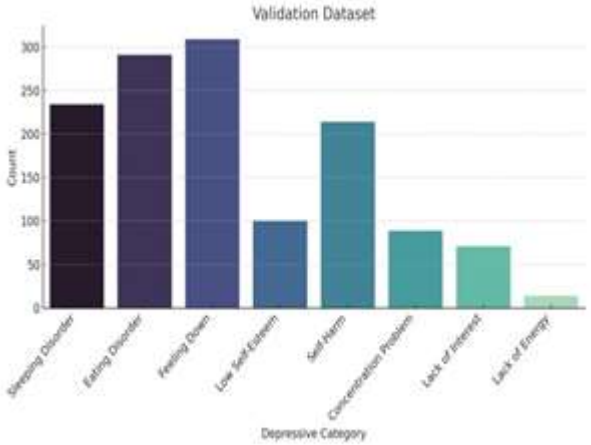

Fig. 5. Data statistics for validation data

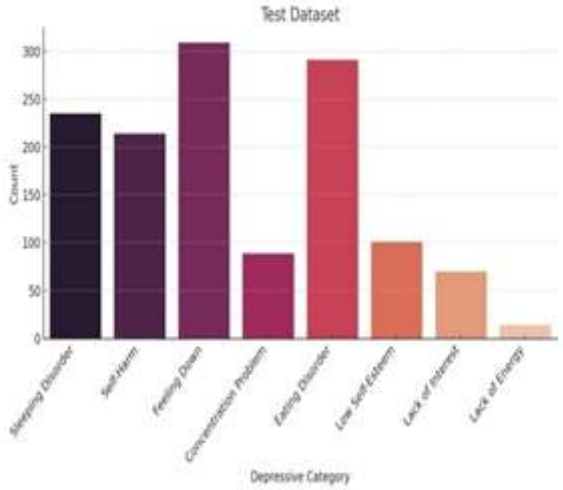

Fig. 6. Data statistics for test data

Sleeping Disorder: The presence of insomnia, hypersomnia, or non-restorative sleep is characteristic of a sleep disorder and serves as an additional symptom of a mental disorder. Sleep disturbances also include a tendency to stay awake and restlessness, which may be accompanied by irritability and nervousness. The sleep disorder symptom is not particularly important, but rather represents a trivial occurrence that suggests the existence of a more serious mental state. This symptom is not only a sign of a disease but also a risk factor and a predictor of major depressive disorder.

For instance, one may take the following quote: "I have not slept in months after Money, but what am I supposed to do? There is always a reason why coffee lasts so long." These are words from a dataset that describe years of sleepless nights

driven by coffee consumption. It indicates the experience of emotional trauma or plunges a user back into a period of stress. Combined with a suitable image, such as a picture of someone restrained or a person sitting behind a desk in exhaustion, it becomes a powerful expression of despair. The combination of visual and textual data explains the rationale for developing MultiT-CNN.

Self-Harm: Self-harm, which is alternatively termed self-injury or Non-Suicidal Self-Injury (NSSI), involves the intentional infliction of physical harm on oneself in order to cope with overwhelming emotional pain. Although it may not necessarily involve suicidal tendencies, self-harm is indicative of grave mental health issues in several categories, including depressive symptoms, anxiety disorders, borderline personality disorders, and post-traumatic stress disorders. Some common ways include cutting, burning, scratching, or indulging in self-abusive tendencies. These actions are often kept secret.

In the RESTORE corpus, memes associated with self-harm may include:

- Visual symbols: razors, bandages, or sketches depicting suffering or wounds
- Metaphoric descriptions: for instance, "I come alive when it hurts" or "pain makes me become a real person"
- Dark humor: e.g., "No need to be worried about those scars, that old pal of mine"

The memes' text could sometimes exceed the literal meaning by using metaphors or geochemical symbolism to express self-harm as a mirror of blocked therapy, social suffering, or internal chaos. The image in the meme could illustrate:

- A lone individual
- Drawings of an arm with scars or red marks
- Darkness and shadow features.

These kinds of examples clearly involve understanding both text and image. Single modal fails to implement deeper meaning thus supporting the need of multi-modal fusion (MultiT-CNN)

Feeling Down: Feeling down is characterized by a persistent, uncomfortable feeling of absence of happiness with no apparent external cause. The feeling correlates clinically with subthreshold depression and the onset of a depressive episode.

Symptoms include:

- decreased vitality or enthusiasm;
- feeling dull, lifeless, heavy;
- emotional exhaustion;
- avoiding thinking about committing suicide

In terms of clinical psychology, “feeling down” is a very common yet underreported condition experienced mostly by teens and young adults. It may not be classified as a major depressive disorder, but it certainly is emotional distress which may get worse when ignored. This state is characterized by trying to pretend that everything is well, while being unable to break out of a mental loop—a type of meme which highlights an internal conflict. The archetype in question would be a downer:

- Not graphic or violent;
- Featuring a smiling facade masking sadness behind closed eyes, or an animated cartoon figure slouching on a couch.
- Without adequate context, it would be difficult to understand the meaning of the meme.

Problem of Concentration: A problem where one struggles to maintain focus and finish tasks that involve mental work. This symptom tends to be seen commonly in cases of MDD, GAD, and ADHD comorbidity. In depression, it manifests itself as distraction, inability to complete mental tasks, zoning out, racing thoughts, forgetting basic information, and a foggy mind. Example: “Me: Time to focus. Me: What was that embarrassing presentation moment seven years ago?” This concept is not a mere joke but an example of how poor concentration can be expressed indirectly through intrusive thinking. Visualize yourself sitting at your desk, looking out into space, with messy surroundings, or doodling pictures that look like brains full of spaghetti strands. The text is intended to imply cognitive dysfunction, while the picture acts as a visual clue of the same.

Eating Disorder: It is a mental illness with inappropriate eating behaviours which may consist of binging, restricting, food aversion or guilt related to food. Concerning memes on the internet, this disorder tends to go through figurative or indirect types of disengaging like:

- Jokes about lunch skipping or binging
- Our body image is a form of dissatisfaction
- Mindful Eating Guilt

Humor as a coping mechanism for deeper self-image problems such as me: I need to eat healthy eats whole cake by myself at

2 am. The caption contrasts the mention of eating clean with reality: nighttime eating at night.

Low Self-Esteem: A poor self-esteem, a feeling of worthlessness and helplessness, a lack of confidence, and constant criticism. It is strongly linked to both depression, social withdrawal, and cognitive distortions, including self-blame. The representative thoughts in internet memes are expressed indirectly or ironically, so that their direct interpretation makes them difficult to detect by the analysis, which fails to understand caption-like and context-sensitive language. For instance, "why bother with friends (natural) because you will let them down sooner or later".

Lack of Interest: Often referred to as anhedonia, it is the loss of pleasure and interest where once-loved activities no longer bring joy. On the internet, this theme manifests as:

- Diminished engagement in life events
- Humorous nods that shield the social, academic, and personal aspects of life

The language leans towards emotional numbness or disconnection. Consider, for instance, the phrase "Friends: So do you want to meet up? me: Can't, ignoring life too" that was extracted from the data set. This meme describes the depletion of social motivation, wrapped up in dark humor, with the emotional withdrawal associated with the symptom summed up in one statement: the "ignore file."

Lack of Energy: It is the feeling of fatigue or psychomotor retardation, a defining characteristic of depression. One drags oneself through days weighed down by exhaustion, lacking any motivation, barely performing basic activities. For instance, lying in bed pondering if it is worth waking up and engaging in simple activities such as urination, which feels like a chore. It represents the physical inability to engage in activities and fulfill basic duties. However, this is only part of the story; beneath the surface lies mental or physical burnout.

IV. METHODOLOGY

In this section, we will present the complete process of methodological design involved in developing and validating the MultiT-CNN model that combines Transformer based text feature extraction mechanism with CNN based visual analysis for classification of memes with depressive elements. This includes the process of framing the classification task and collection of a specific dataset for depressive memes, along with details about preprocessing of text and images. Further, we elaborate on the model architecture that integrates BERTweet embeddings with features from ResNet18. Additionally, the training procedures adopted by us and how class imbalance was mitigated using loss functions are discussed, along with the comparison of our work with baseline methods.

Problem Definition

With the emergence of memes as an expression medium on the Internet, scientists have been pushed further down the path of analyzing the hidden psychological and emotional information contained within. Memes combining images with text create a more difficult challenge for computational mental health assessment, especially when it comes to detecting signs of depression, which can be expressed indirectly through sarcasm and cultural images, among other things. The current research aims to construct a new classification model that is capable of dealing with this complexity. It views the problem as multi-class supervised classification, where each meme is classified into one of eight clinically meaningful depressive symptoms.

Text Preprocessing

For preprocessing of textual data, the BERTweet tokenizer (a type of pretrained language model dedicated to social media data and based on transformer architecture) was used to tokenize captions into subword units, limiting token sequence length to 128 tokens for uniform input dimensions I This involves padding and truncation to create a fixed length input across batches which returns two key tensors: input_ids and attention_mask that will be fed to the BERTweet encoder. The new shape using the squeeze() method enables a more efficient processing for batch training in Pytorch. And for detecting depressive symptom in a meme, it helps to handle multiple types of input.

Image Preprocessing

Multimodal deep learning techniques are utilized to identify the presence of depression in memes. Images are pre-processed using the PIL library and torchvision transforms to fit into the input size of 224x224 needed for ResNet18. To increase generalization and minimize overfitting, the process involves data augmentation through random horizontal flip, random color jitter, and random rotation. These steps enhance the accuracy of detecting depression in multimodal images.
The normalization parameters adopted include

μ=[0.485, 0.456, 0.406] and σ=[0.229, 0.224, 0.225],

which correspond to the pre-trained weights of the ResNet18 model.

Addressing Missing/Corrupted Images: In case of errors or pixel loss, the program resorts to a default image. In particular, an image that is corrupted or missing is substituted with a black-colored image measuring 224x224 pixels generated using Image.new('RGB', (224, 224)).

Class Imbalance and Class Weights: The second problem with the dataset utilized in the detection of depression is class imbalance. For example, the symptoms of hopelessness and worthlessness are present in larger numbers than emotional numbness and suicide ideations. The computation of class weights in Scikit-learn's compute_class_weight (balanced) enables penalization of mistakes on less common classes and plays a vital role in modality training.

Proposed Model Architecture

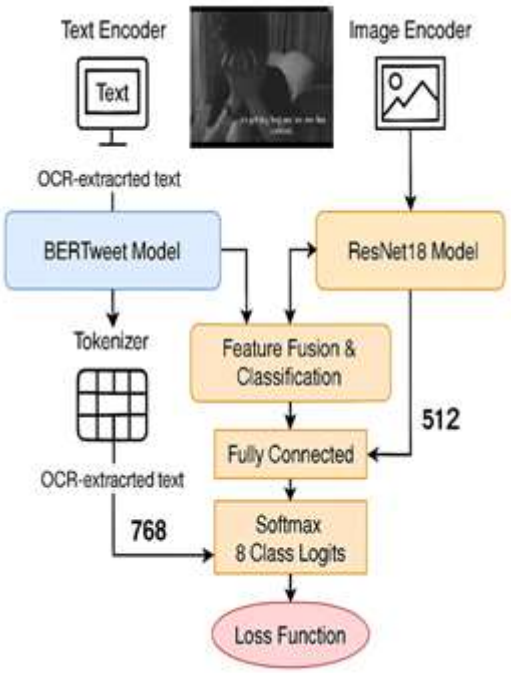

Fig. 7. The Architecture of the Proposed Model

The suggested neural network model, MultiT-CNN, is a powerful multimodal network that can recognize depression symptoms expressed in internet memes. MultiT-CNN utilizes the advantages of transformers for the text modality and CNNs for images. The design of this model considers the complex semantics between the text and image modalities of memes and incorporates them in two separate encoders: text and image, which are then aggregated for classification.

The text modality is represented by BERTweet-base, a RoBERTa-based transformer pre-trained on a massive English-language tweet dataset. BERTweet-base is an excellent choice for understanding the informal language used in meme captions. The encoding of the text modality, obtained from the meme through OCR, is performed with a tokenizer, resulting in a 768-dimensional embedding.

Regarding the image modality, ResNet18, pre-trained on ImageNet, acts as an image encoder. The output of the ResNet18 encoder results in a 512-dimensional pooled image embedding.

$$h = [h_{utc_{hie}}] \in R^{1280} \qquad (1)$$

Here, text and image cues are fused and fed to a feed-forward classifier that is trained to detect signs of depression in memes. Specifically, the 1280-dimensional feature vector produced by the fusion of text and image cues is mapped to 256-dimensional space via a linear mapping followed by a ReLU activation function. In order to prevent overfitting due to the presence of class imbalance, a dropout layer with p = 0.3 is introduced before the final linear layer that outputs logits corresponding to eight classes of depression symptoms.

$$z = \mathrm{ReLU}(W_{1h} + b_1),\ \hat{y} = \mathrm{Softmax}(W_{2z} + b_2) \qquad (2)$$

As such, the network is able to examine the visual and textual components of a meme to capture relevant information about depression. As demonstrated through multimodal learning, MultiT-CNN has the tendency to yield more accurate classification results at the symptom level.

Feature Fusion and the Classification: Under this method, the feature fusion process takes place at the level of representation, where the high-level semantic features extracted from the text and image of the memes are combined. While for the text, we use BERTweet and get a vector of size 768 dimensions using OCR Text, for the image component, we apply ResNet18 to get a vector of 512 dimensions using meme image. Next, these two vectors are concatenated together to create a fused feature vector of 1280 dimensions. This fused feature is mapped to

eight outputs via an 8-dimensional fully connected layer with ReLU and Dropout activation.

Loss Function and the Optimization: In regard to handling imbalanced data to detect depression symptoms, the loss function implemented in the model is the weighted cross-entropy. The function considers the number of instances of each symptom class using Scikit-learn's compute_class_weight function. Its mode is "balanced." This strategy avoids biases towards larger classes. During training, the optimizer used is AdamW with a fixed learning rate of 2e-5 that ensures convergence without being too fast, leading to unstable training. Focal loss is not used because of the instability associated with handling smaller classes.

Training Approach: The training procedure involves mini-batch gradient descent, with a batch size of 16, a maximum of 15 epochs, early stopping based on patience of 5 epochs based on the validation F1 metric. In the first two epochs, the BERTweet backbone is kept frozen so that it trains only the classifier component of the model. Starting with the third epoch, the BERTweet backbone is unfrozen for end-to-end training. Checkpoints are created when the F1 score increases at each epoch using PyTorch's torch.save() function. See Figure 8 to view the training results of the F1 score of the proposed model.

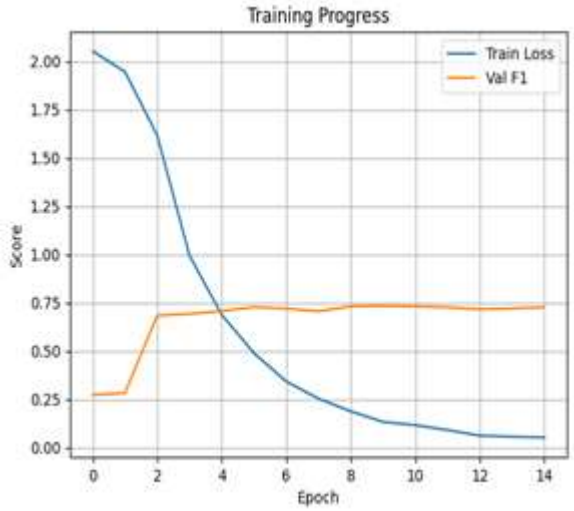

Fig.8. Training loss with F1 score for proposed model

The Evaluation Metrics

A systematic approach for measuring the effectiveness of MultiT-CNN model and its base model based on unimodality, emphasizing on the weighted F2 score metric, by using a set of well-established metrics of classification.

Weighted F1-Score: F1-score is defined as the harmonic mean between the precision and recall for each class and is calculated separately for each class. The weighted F1-score represents an aggregation of all class-specific F1-scores. This metric is an indicator of general performance that considers both accuracy and class imbalance, and it is calculated as

$$\text{F1 weighted} = \sum_{i=1}^{k} \left(\frac{ni}{N} \text{x F1i} \right) \quad (3)$$

Where
K = The total number of classes
ni = Number of true instances for class i
N = Total number of instances (i.e., N =∑ni)
F1i = F1 score for the class i
ni/N = Class weight

Precision, Recall and Accuracy: In addition to weighted F1, we compute:

- Precision: The percentage of correctly predicted is known as precision.
- Recall: Recall refers to the percentage of true positives that are accurately predicted.
- Accuracy: Accuracy is defined as the overall percentage of accurately predicted cases over all predictions.

Precision = TP / (TP + FP) (4)
Recall = TP / (TP + FN) (5)
Accuracy = (TP + TN) / (TP + TN + FP + FN) (6)

where

- TP: True Positives
- FP: False Positives
- FN: False Negatives

However, this may be an overly optimistic estimate due to an imbalance in class distribution.

Confusion Matrix: The confusion matrix will be used to assess the agreement between predicted labels and ground truth labels, helping us understand the performance of each class as well as common errors. It is constructed using the following equation:
CMi,j = Number of samples whose ground truth label is i and prediction label is j.

The confusion matrix will be normalized after calculation and displayed as a heatmap using the Seaborn library with class labels generated from the Label Encoder in training. All evaluation metrics will be returned by Scikit-learn classification_report() and confusion_matrix() functions. Model predictions will be done using the argmax operation on the SoftMax logits. Testing will be done exclusively on test data, using the checkpoint that achieves the highest accuracy on the validation data set. Accuracy is one metric, while precision and recall metrics give better insights into the recognition of each class.

V. EXPERIMENTS AND RESULTS

Design for experiment, evaluation settings, and performance analysis of the proposed model, along with its comparative analysis with text-only and image-only classification models, will be discussed in this section. The aim of all the experiments is to examine the efficiency of the multimodal fusion technique for depressive memes classification.

Experimental Design

All experimental analysis was performed using PyTorch in a GPU-enabled Google Colab environment. The dataset included annotated memes containing both the caption extracted by OCR processing and corresponding image features. Dataset is split into three parts – Training, Validation, and Testing, where 70%, 15% and 15% of the total memes belonged to each group respectively. BERTweet-base pretrained model was chosen for text encoder while ResNet18 model trained on ImageNet was taken as an image encoder.

Baseline Models

For this purpose, two uni-modal baselines will be introduced.
Text-only Baseline: Text-only classifier adopts the BERTweet transformer model, pre-trained on large amounts of tweets and, hence, more adapted for processing meme texts that might contain slang or figurative speech.

- Input: meme text features (input_ids, attention_mask)
- Encoder: the BERTweet encoder produces the [CLS] token embedding of dimensionality 768
- Classifier: Dropout(p=0.3), linear classifier with 8 outputs (logits for 8 depressive classes)
- Output: Softmax probabilities over depressive classes.

Image-only Baseline: For the case when only image information should be used for depression detection, we use ResNet18 architecture. This neural net is fully convolutional without any text processing layers.

- Input: meme images converted into RGB format of 224x224 resolution
- Encoder: the ResNet18 with the fc classification layer removed (fc=Identity), producing an embedding of 512 dimensions
- Classifier: a fully connected layer that maps the 512-dimensional visual embedding into logits for depressive classes
- Output: Softmax probabilities over depressive classes.

These two baselines will serve us as a measure of how much additional value can be obtained from the fusion of modalities. All baselines were trained in identical conditions to the multimodal networks.

Comparative performance

The performance of the multimodal approach is tested against two different unimodal models. One of them is only based on the texts and applies BERTweet with a Dense Classifier. The second one uses the ResNet18 architecture with the final layer modification for images only. For the evaluation, the weighted F1 score, precision, recall, and confusion matrix are considered. The performance metrics for the test sets of texts are shown in Table 2.

Table 2: Model Performance Metrics for proposed model

Model Performance Metrics (Test Set)			
	Text Only	**Image Only**	**Multimodal**
F1 Score	0.672	0.008	0.685
Recall	0.671	0.067	0.680
Precision	0.675	0.005	0.704
Accuracy	0.671	0.067	0.685

Analysis of results

The MultiT-CNN approach proved to have better results in identifying expressive symptoms among memes by achieving a higher weighted F1 score of 0.6846 compared to the text-only model's score of 0.6716. This result shows the importance of the textual part of memes to represent emotions but also the ability of visual cues to increase the effectiveness of classification when memes use visual sarcasm or symbolism. At the same time, the image-only approach has demonstrated very poor results with the F1 score of 0.0085, showing the inefficiency of visual-only methods in detecting depression memes, and proving the need to consider the textual context to

identify complex emotional messages within memes. Table 3 presents the F1 Score for each of 8 classes for baseline and proposed models. Figure 9 shows F1 score performance comparison for Baseline model with proposed model.

Table 3: F1 score for each 8 classes

Classes	Text-only	Image-only	Multimodal (Proposed)
Concentration problem	0.59	0.28	0.67
Eating Disorder	0.80	0.46	0.82
Feeling Down	0.66	0.29	0.66
Lack of Energy	0.58	0.00	0.58
Lack of Interest	0.46	0.21	0.59
Low Self-Esteem	0.65	0.15	0.62
Self-Harm	0.65	0.34	0.68
Sleep Disorder	0.70	0.52	0.74

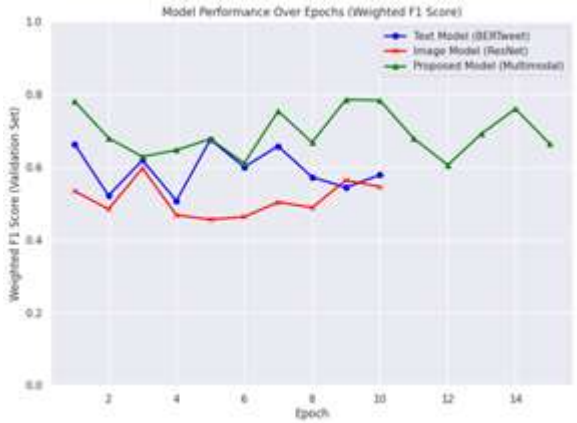

Fig. 9. Performance Comparison of F1 score for baseline model with proposed model

According to the confusion matrix of the multimodal approach, this type of model proves high values for precision and recall for common types of depressive symptoms in the form of worthlessness and hopelessness since these categories can be detected more effectively. See Figure 10, Confusion Matrix of all 8 classes for proposed model. However, the performance of the model declines when rare categories, such as emotional numbness and suicidal ideations, become objects of detection, and hence there is a need to employ techniques such as data augmentation and special loss functions, such as focal loss.

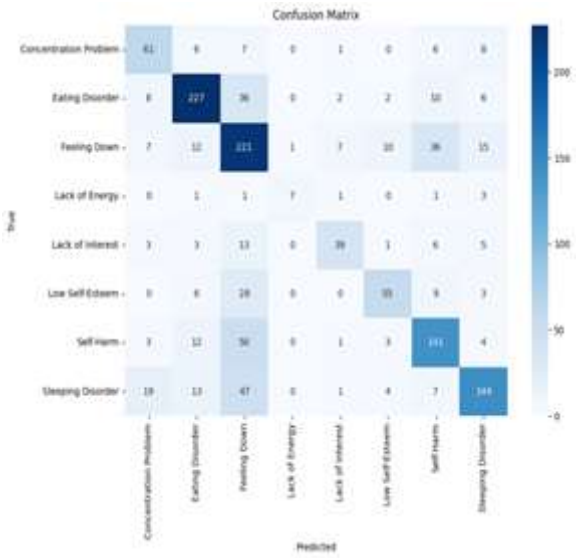

Fig. 10. Confusion Matrix of all 8 classes for the proposed model

Visualization

The plot showing the training loss versus the validation weighted F1 score provides very important metrics that are plotted against the number of epochs used to train the model and determine its learning capability and convergence. The stability in a drop in training loss and rise in validation weighted F1 score proves that the model has learned well and it generalizes well, hence it does not overfit to the training data. Therefore, it validates the importance of the validation F1 score in determining when training should stop using early stopping approach.

VI. CONCLUSION

In the current research, a novel multimodal machine learning architecture called MultiT-CNN has been developed, which combines textual input obtained from meme captions using the BERTweet model and visual input extracted using the ResNet18 CNN. Results obtained after extensive experimentation have clearly established the superiority of the multimodal solution in detecting depression-related symptoms through a weighted F1 score of 0.6846 over a single modality text-based approach, where the score was 0.6716 and the image-only based system where the value was 0.0085. These results emphasize the importance of multimodality in depression detection because the image-only solution shows the inability of visual information alone in detecting depression symptoms, thus emphasizing the importance of language use in

meme captioning in depression detection. It can be observed that although mostly textual input carries more significance in conveying depressive sentiment, the interplay between image and text inputs plays an important role in determining the ambiguity of figurative speech. This underscores the importance of conducting additional studies in the future to unravel the intricacies of image-text interactions in social media posts produced by users, especially when analyzing mental health conditions. Future studies involving the MultiT-CNN framework are expected to improve the interpretability and robustness of the model through multimodal attention, which will enable the model to concentrate on salient features in text and images. Moreover, the use of explainability components and commonsense reasoning, based on contextual information, will make it easier to comprehend symptom types. Finally, the integration of clinical psychology knowledge and multi-label classification techniques is likely to increase the accuracy of symptom identification and diagnosis.

REFERENCES

1. E. Andalibi and O. Haimson, “Understanding social media as a platform for mental health expression,” Proc. ACM Hum.-Comput. Interact., vol. 5, no. CSCW2, Art. no. 434, Oct. 2021, doi: 10.1145/3479577.
2. S. Saha et al., "Mental Health Awareness Through Memes: A Computational Study," in Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES), 2021.
3. S. Wu et al., "Humor Detection in Social Media: A Multimodal Approach," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020.
4. A. Ghosh et al., "Detecting Sarcasm in Multimodal Social Media Posts," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.
5. D. Q. Nguyen, T. Vu, and A. T. Nguyen, "BERTweet: A Pre-trained Language Model for English Tweets," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 9–14.
6. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
7. S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt, "Detecting depression and mental illness on social media: An integrative review," Current Opinion in Behavioral Sciences, vol. 18, pp. 43–49, Dec. 2017.
8. A. Yates, A. Cohan, and N. Goharian, "Depression and Self-Harm Risk Assessment in Online Forums," in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 2968–2978.
9. A. Trotzek, S. Koitka, and C. Friedrich, "Utilizing BERT for Emotion Detection from Text," in CLEF 2020 Working Notes, 2020.
10. X. Wei, H. Jin, and T. Liu, "CANAMRF: Cross-modal Attention Network for Adaptive Multimodal Reasoning in Depression Detection," arXiv preprint arXiv:2401.02995, 2024.
11. P. Moon and P. Bhattacharyya, "We Care: Multimodal Depression Detection and Knowledge Infused Mental Health Therapeutic Response Generation," in Proceedings of the 21st International Conference on Natural Language Processing (ICON), 2024, pp. 296–310.
12. N. Andalibi, "What Happens after Disclosure? Examining the Social, Psychological, and Relational Consequences of Disclosing Depression on Social Media," Proceedings of the ACM on Human-Computer Interaction, vol. 4, no. CSCW2, pp. 1–25, Oct. 2020.
13. D. Kiela et al., "The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes," in Advances in Neural Information Processing Systems 33, 2020, pp. 2611–2624.
14. K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, and Y. Liu, "Combating Fake News: A Survey on Identification and Mitigation Techniques," Information Fusion, vol. 52, pp. 278–299, Dec. 2019.
15. S. Saha, D. Sheshadri, and E. Cambria, "Mental Health Awareness Through Memes: A Computational Study," in Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES), 2021, pp. 480–490.
16. S. Yadav, C. Caragea, C. Zhao, N. Kumari, M. Solberg, and T. Sharma, "Towards Identifying Fine-Grained Depression Symptoms from Memes," in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 8835–8849.
17. S. Sharma, Ramaneswaran S., M. S. Akhtar, and T. Chakraborty, "Emotion-Aware Multimodal Fusion for Meme Emotion Detection," arXiv preprint arXiv:2403.10279, 2024.

18. T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423–443, Feb. 2019.
19. J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks," in Advances in Neural Information Processing Systems 32, 2019, pp. 13–23.
20. A. Radford et al., "Learning Transferable Visual Models from Natural Language Supervision," in Proceedings of the 38th International Conference on Machine Learning, 2021, pp. 8748–8763.
21. D. Kiela et al., "The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes," in Advances in Neural Information Processing Systems 33, 2020, pp. 2611–2624.
22. X. Li and J. Xiao, “MFFNC: Multimodal Feature Fusion with Neural Cross Attention for Depression Detection,” arXiv preprint arXiv:2407.12825, 2024.
23. U. Akram and J. Drabble, “Mental health memes: beneficial or aversive in relation to psychiatric symptoms?” Humanities and Social Sciences Communications, vol. 9, no. 1, pp. 1–6, 2022.
24. A. Bhaumik and T. Strzalkowski, “Towards a Generative Approach for Emotion Detection and Reasoning,” arXiv preprint arXiv:2408.04906, 2024.
25. A. Mazhar, Z. H. Shaik, and A. Srivastava, “Figurative-cum-Commonsense Knowledge Infusion for Multimodal Mental Health Meme Classification,” arXiv preprint arXiv:2501.15321, 2025.