Unveiling Mental Health Insights from Social Media with Long Short Term Memory and Robustly optimized BERT approach Model

[1]Gopalakrishnan K
Department of InformationTechnology,
Velalar College of Engineering and Technology,
Erode, India.
gopalakrishnanbtech@gmail.com

[2]Kavitha K
Department of Computer Science and Engineering,
Kongu Engineering College,
Erode, India.
kavitha.kavi123@gmail.com

[3]Dharshini B
Department of InformationTechnology,
Velalar College of Engineering and Technology,
Erode, India.
dharshinibalu2305@gmail.com

[4]Bhuvaneshvar P M
Department of InformationTechnology,
Velalar College of Engineering and Technology,
Erode, India.
bhuvaneshvar2005@gmail.com

[5]Dhayalan M
Department of InformationTechnology,
Velalar College of Engineering and Technology,
Erode, India.
dhayalan290305@gmail.com

[6]Kavipriya M
Department of InformationTechnology,
Velalar College of Engineering and Technology,
Erode, India.
kavipriyamadeshwaran830@gmail.com

***Abstract*- Million people around the world, 1 in 8 people, were living with a mental disorder. The pandemic causes significant increases in the number of mental illnesses especially anxiety and depression. Mental illness is a state that affects the day- to day life of a person, affecting the person's behavior, feelings, thinking, and mood. There was no medical test to easily detect mental illness. Detection can be done only by blood test (mimic mental health) or imaging scan (physical or structural issues in the brain), which takes more time for diagnosis. Using user's image and text, he/she has uploaded on the social media platform, by employing algorithms Convolutional Neural Network (CNN) - Long Short -Term Memory (LSTM) for picture and Robustly-optimized BERT approach (RoBERTa) for content, *leveraging Natural Language Processing (NLP)* to examine the condition of mental disorder such as depression and anxiety. The system predicts mental illness and helps healthcare professionals understand the model better, build tests and make more accurate diagnoses.**

Keywords*- *Mental Health, Social-Media, Natural language processing, Long Short-Term Memory, Convolutional Neural Network., Area Under the Curve .

I. INTRODUCTION

Perhaps the most debilitating psychological health problems affecting people all across the globe are anxiety and depression and it has been estimated that about 8 of 1 people suffer from these conditions[8] . Today, it is quite clear that there is a serious need to formulate a unique diagnostic solution for such ailments; the COVID-19 Pandemic has exacerbated this issue . Most of the current manual diagnostic methods are primarily dependent on blood sampling and imaging techniques, which are often inaccessible, ineffective or inconclusive depending on the case in question . Simultaneously, the rise of smartphones and social networks has left a broad digital imprint of people emotions, deeds and thoughts. This work explores the possibility of employing social media text and images as a tool in machine learning to identify a subset of mental disorders at their earlier stages [1]. Nowadays, people express their emotions through social media often. This digital platform provides insights into an individual's psychological well-being; it provides an early detection of mental health disorders . Here introduced an innovative automation system that uses learning techniques to analyze social media content [5]. The system includes models such as combining CNN along with LSTM image analysis and the RoBERTa model for text analysis. This dual analysis method provides a more robust and accurate assessment of an individual's mental condition status . The CNN algorithm extracts features from images like facial expressions and background colors, while LSTM tracks and monitors changes over time and keeps on learning about their mental state[13]. Text analysis using the RoBERTa model examines users comments on social media platform posts based on sentiment, linguistic patterns, and emotional The NLP identifies subtle indicators for mental conditions like writing style changes, emotional expression, and social patterns [7]. The model foresees anxiety or depression based on the risk score given by the pattern . The system is to provide early, accessible, and non-invasive cognitive health tests. By using social media content that is available on social websites provided by the user, the system can identify the mental state of the user before they become severe, enabling earlier analysis and support. Over the years, researchers have been moving toward multimodal machine learning more and more to better detect and grasp mental health problems[9] . Early work was based on a single modality like text features or visual cues, but recent research welcomes multiple types of data to model more strongly. For instance introduced a Sentiment-guided Transformer with Severity-aware Contrastive Learning (ST-SCL), which greatly enhanced F1-scores by 12% on the DAIC-WOZ dataset in detecting depression by combining sentiment attention mechanisms and severity-aware loss functions[11]. This enabled more effective emotional representation and discrimination. Following this a hybrid attention-based multimodal fusion model that incorporated facial expressions, audio signals, and text information extremely well. Their model outperformed state-of-the-art clinical benchmarks such as CMU-MOSEI and DAIC-WOZ ($p < 0.01$) in emotion and disorder classification tasks**.** This approach helps to overcome the limitations of traditional methods for diagnostics by the doctor. The system's ability to analyze both image and text content can provide a more accurate result of mental state, which also helps in saving time. In this work, an approach and analysis on predicting emotional state from social media along with visual insights and also its relevance to healthcare practice is described.

II. METHODOLOGY

A. Data-collection

For this research, data was gathered from publicly available social media sites, including Twitter and Reddit[5] . The data is based on publicly available verified datasets from Kaggle and GitHub to create the multimodal mental health detection model while maintaining ethical compliance. Figure1 represents the analysis and extraction of data for multi model training and result prediction.

1)Image

The data collection includes selfies, personal photographs, and images of the user's environments. Ensure that the image satisfies the quality required for further processes. The collection of images mainly focuses on the facial expression and background, which are essential for analysing depression and anxiety[12] . The image dataset is drawn from Kaggle's Facial Emotion Recognition Dataset and includes 35,000 anonymized grayscale facial pictures (24x24 pixels) tagged with eight emotions (happy, sad, ~4,375 per category), preprocessed through resizing into 24x24 pixels, normalizing into [0,1], and stripping metadata for removing personally identifiable information (PII)[13] .The Datasets was used for detecting mental disorder from social media for imaage ,as published by kabir et al[16].

2)Text

Text datasets comprise ~10,000 de-identified Twitter tweets from GitHub - Emotional Patterns in Social-Media (6,000 depression, 4,000 anxiety), usernames, URLs, and emojis removed, and ~5,000 Reddit posts from GitHub - Stress Detection from Social-Media (3,000 stressed, ~2,000 non-stressed), tokenized and anonymized[1][3] . The dataset includes emotion, daily comments, experiences, and personal thoughts. This collection ensures that the model identifies various linguistic patterns that are related to mental conditions for further analysis .

Kaggle dataset is trustworthy, with human-validated labels (kappa ~0.85), whereas GitHub datasets, not peer-reviewed, were validated through manual sampling (10% of data, ~95% label agreement). There is no demographic metadata (e.g., age, gender, region) present, exposing it to bias toward younger, Western users common on Twitter and Reddit) [8]. Sensitivity tests will probe model robustness under assumed populations, and future work will make use of datasets, which contain age and gender metadata for ~5,000 posts, in order to better diversify. Data for detecting mental disorders from social media for text was collected from a study on emotional patters by Sreekar Anumala and Desai [17][18].

Annotations were performed by psychological professionals. They separated the data into categories such as "Anxiety," "Depression," and "Healthy."

B. Data-preparation

The system mainly focuses on image and text, which requires certain methods to get accurate conditions[5] . By analysing both text and image, the system can get an accurate result.

1)Text Analysis

The first stage relative to text data preprocessing is to ensure that raw text data, which is unstructured, is ready for analytical work and consequently becomes structured. Some of the basic steps involved are:

a)Tokenization

The process that breaks down into smaller components of sentences and words as smaller units. Lowered by lower case equation (1) shows the process where text characters are altered into lowercase versions of the character and equation (2) shows the process where the text are get tokenized[7] .

$$A_{lower}=L(A_{input}) \quad (1)$$

$$A_{token} = T(A_l) \quad (2)$$

where L represents lowercase (), T represents tokenize ().

b)Word-removal

Seeks to remove some words that do not carry meaning, like conjunctions and articles, which are rarely used. Followed by punctuation and special character removal,where equation (3) which aids in eliminating unwanted aspects that cause disturbance[8] .

$$A_{clean} =C(A_{token}) \quad (3)$$

where C is the clean().

c)Stemming-lemmatization

They are used to take words and translate them into simpler types of words with similar meanings or the stems of other words which equation (4) which process text into simple words. The RoBERTa tokenizer where equation (5) standardized cleaned text sequence length 512 tokens for input to the model[7] .

$$A_x=S(A_{clean}) \quad (4)$$

$$A_f=R(A_s,Max\_Len = 512) \quad (5)$$

Where S is the Steam() or Lemmatize() ,R is the RoBERTaTokenizer.

2)Image Analysis

Image Data preprocessing the cleaning, normalization, and enhancement of raw images in order to increase the efficacy and efficiency of the model[6] . It is vital for augmentations such as rotation, flipping and cropping are made use of to make the dataset more diverse; the end goal being turning it into a format that the machine learning models can integrate. This includes removing noise using filters or changing the images into grayscale.

Modifications such as these ensure that there is uniformity across the entire dataset and there are no computational biases that stem from the complexity along with addressing potential biases in the data. In order for computer vision tasks to be executed effectively, efficient preprocessing is necessary, especially to allow the model to generalize and perform effectively on data it hasn't seen before [12].

Resolution is scaled to 48x48 pixels. Augmentation in rotation, flipping and cropping for robustness of the model. Converting some images to grayscale for experiments [13].

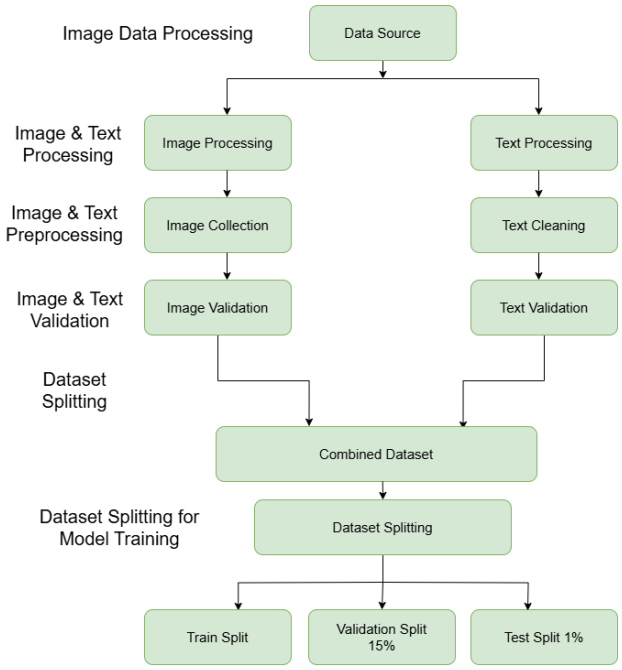

Fig. 1. Data Pipeline for Multimodal Extraction. The diagram illustrates the stages of data collection, preprocessing, feature extraction and integration for mental health prediction.

C. Models and Implementation

1)RoBERTa model

The NLP module's goal within the machine learning field is to enable or teach a computer how to understand, produce, or process any natural language in an efficient way that could be useful in practical situations [8]. NLP is aimed towards dealing with communication in components, models, or systems to work on linguistics, statistics, and deep learning of the languages with some forms of linguistic translation. Tasks that might be described under the heading of NLP include, but are not limited to, text classification, opinion mining, machine translation, question answering, and abstracting articles.

The RoBERTa is used because its training enhancements, like dynamic masking and using larger corpora, surpass deriving emotional and psychological indicators (net boosting) in more casual text contexts (like texts, tweets, and slang) achieving surpassing 85% F1 score on tasks like GoEmotions [5][7]. Their performance is also competitive with informal short text models and captures the essence of 768 dimensional embeddings. In competitions, GPT-4 scored 90% on GLUE benchmarking, which is higher than the competition, but is also extremely resource hungry, needing 10 times the resources for inference (0.1s per tweet for GPT-4 and ~0.01s for RoBERTa) which makes it impractical for our classification-centered focus [9]. BERT, which is slightly below, scored an F1 of 82%, which is remarkable, but is not as strong due to lacking training optimizations like RoBERTa.

When the words first passed the RoBERTa frozen embedding layer, which convert words into semantic meaning[1] . After it flows through the unfrozen embedding layer, where they are refined into a specific task. At lasts the text passes through the classification layer, converting sophisticated to actionable prediction. This system has an understanding of both linguistic patterns and broader context, which shown in equation (8) which is necessary for analysis [12].

2)CNN-LSTM Model

The CNN module is another deep learning architecture that is most popularly used and is aimed at working with spatial or grid-like data, which, for instance, includes images and videos[0] . CNNs are good at feature obtaining because of the convolutional layers, which are applied filters that are used on the input data in order to recognize edges, texture, shapes, and other patterns. It consists of pooling layers that further follow and serve to retain imperative features while reducing the spatial size of the input.

CNNs have fully connected layers as well, used for classification or regression. They are useful for image and video purposes primarily, but the use in object localization, segmentation, and that of natural languages in NLP has also increased. TensorFlow and PyTorch are good examples of frameworks that support such activities as implementing and customizing CNNs and therefore make it much easier and more accessible to researchers as well as practitioners.

This approach lies on the pre-training ResNet50 model that extracts image features through a deep learning framework. ResNet50 is utilized for emotion identification, it employed for emotion detection on efficient, accurate and well suitable for grayscale image [6]. Its unique skip connections prevent training problems and enable it to detect small facial features (such as eyes and mouth) that are crucial for detecting mental states like depression or anxiety [12]. Pretrained on ImageNet, it achieves around 70% accuracy on similar datasets after fine-tuning and is quick—one image takes approximately 0.01 seconds to process on an NVIDIA A100 GPU. Out of all the other models, Vision Transformers require more data, memory, and take longer. EfficientNet-B0 provides comparable accuracy but has not been extensively tested on grayscale emotion data. VGG16 is outdated, less precise, and more resource-hungry [11]. ResNet50 is the top pick since it is precise, light (25 million parameters), is compatible with the CNN-LSTM model, and is also energy-efficient and environmentally friendly. LSTM with 256 hidden units, the system performs by moving forward and backward with text in both directions, which clarifies past and future token content.

The two layers of the model are used for learning the hierarchical pattern, where the first layer identifies basic relationships and the second layer identifies complex patterns in the image. The importance of this system is to understand different temporal features. The high weightage of this system is given to the facial expression and then the background[13]. The LSTM keeps on monitoring the CNN process for a period of time and generates accurate results.

The two layers of LSTM are both short- and long-term, which allows the model to capture sudden and gradual shift changes in facial expression for providing accurate results. By using a hybrid model, by equation (7) where this system uses advanced AI to analyze facial expressions and keeps learning their changes in a short duration to provide an accurate result [3].

3)Integration

The integration of image and text in this system represents refined approach for combining different dataset and model for emotional state assessment [4]. Figure2 represents the workflow of multi model which combines CNN-LSTM and RoBERTa model. This architecture admit the datasets are different and complementary, which needs analysis each types of data separately for the accurate result. The system process the image and text data to provide a wisdom for each.

The image process uses a CNN-LSTM model for visual and time based analysis of pattern in the respective given images [12]. Simultaneously a text process uses RoBERTa model to understand emotion in the comment post by the user [8]. This distinct processing ensure image and text are analyzed properly without blending their distinct qualities. The risk score is used as a quantitative that shows the user condition like anxiety and depression represent in equation (6) [10].

$$Y = x_1+I+x_2+T+x_3+B+bias \quad (6)$$

$$I=1/a\sum^{a}_{n=1} w_n(z_n) \quad (7)$$

$$T=V(A_f) \quad (8)$$

$$B=\alpha\times q+\beta\times d \quad (9)$$

where Y is the Risk Score, I and T is the feature of image and text ,B is the Behavior , q and d is the frequency of post and deletion rate, x_1 and x_2 is weight that shows weight of image features and text features, x_3 shows the weight for behavioral features, z_n is CNN extracted features from the image , w_n is LSTM processed feature capturing temporal patterns, V is RoBERTa extracts high level semantic and emotional features from the text, A_f is tokenize text, α,β Coefficients for post frequency and deletion rate , bias is Bais term.

Equation (9) shows the behavior of user. This feature concatenation stage represents an important transition at which the system begins to merge these isolated streams of information. It concatenates the high-level features obtained from both modalities into one feature vector [13]. Concatenation keeps rich representations learned from both the modalities while forming a single unified space that lets us probe into relationships between visual and textual features. The concatenated features preserve the semantic meaning of the original features but let the system uncover cross-modal patterns that might correspond to emotional states.

The fusion network is formed to be the last integration mechanism; it's essentially a two-layer neural network that has 256 units. Here, the network learns the best ways of combination and weighing features from both the modalities of information while finding what the informative combinations of visual and textual patterns are in regard to the cognitive health assessment [2][0]. With two layers, the architecture remains computationally efficient while still achieving sufficient complexity in capturing non-linear relationships between the modalities.

It will learn to look for patterns that might not even be evident by viewing the modalities in isolation; for instance, how certain writing patterns are correlated with specific visual features that can point towards certain emotional states. This ability to adjust these weights during training makes the system adaptable in optimizing its integration strategy toward specific characteristics of mental condition assessment tasks. For predicting psychological score for anxiety and depression we use the formula shown in equation (10) and equation (11).

$$M=x_1\times(1/a\sum^{a}_{n=1}w_n(z_n))+x_2\times(V(A_f))+T+x_3\times(\alpha\times q+\beta\times d)+bias \quad (10)$$

$$P= x_1\times(1/a\sum^{a}_{n=1}w_n(z_n))+x_2\times(V(A_f))+T+x_3\times(\alpha\times q+\beta\times d)+bias \quad (11)$$

where M is the score of Anxiety, P is the score of Depression.

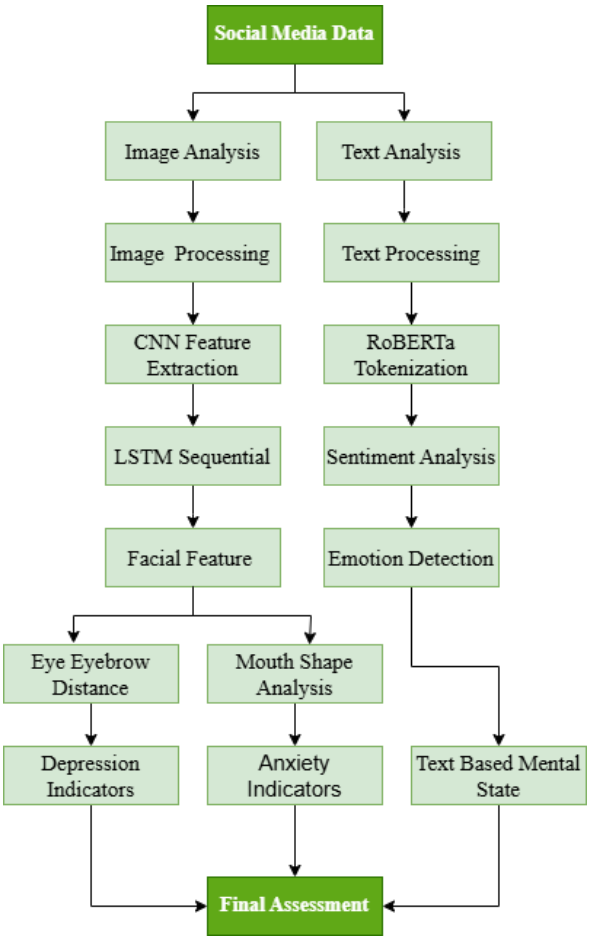

Fig. 2. Work Flow of Multi-Model

4)Traning

The training data employ a holistic three-stage approach to ensure the best performance for a model [15]. In the preprocessing phase, the images are standardized first by resizing all images into 48x48 pixel size and then normalized using ImageNet statistics [6]. All text processes use the RoBERTa tokenizer with a maximum sequence length of 512 tokens; dlib's facial landmark detector was used for facial features with consideration for the most accurate feature identification [13].

In more detail, AdamW optimizer learns at an initially set learning rate of 2e-5 and then adopts the cosine annealing schedule from ten epochs while keeping cross entropy loss with weighted classes for cases where datasets contain a significant skew in class probability distribution [14]. Since batch size had to be tuned between memory availability and

convergence, a batch size of 32 was adopted throughout the experiment.

A few regularization strategies have been applied in order to prevent overfitting and ensure generalization of the model. The network architecture includes dropout layers of values ranging between 0.3 and 0.5. Early stopping was adopted during training by using cross-validation performance metrics to prevent the possibility of overtraining. Additionally, gradient clipping has been used to stabilize the training process against sharp gradients, which could trigger the derailment of the optimization method [14].

The properties derived from images and text is merging into a single dataset.

This dataset is then break into:

I. Training Set (60%) – This is used to train the CNN-LSTM model.

II. Validation Set (20%)- This is used for fine-tuning model parameters to avoid overfitting.

III. Test Set (20%) – This is used to evaluate the model's accuracy before deployment [2].

5)Metrics of Evaluation

a)Accuracy

Accuracy is the measure of how frequently the model is correct in categorizing the users as mental condition categories. Emotional state cases are often imbalanced in the general population. So, accuracy is weighted so that results do not mislead [8].

$$Acc=(Trp+Trn)/(Trp+Trn+Fsp+Fsn) \quad (12)$$

where Acc represents Accuracy , *Trp* represents True positives that are correctly identified with mental state, *Trn* represents true negatives where healthy users were correctly identified, *Fsp* represents false positive is when users without mental state are incorrectly classified to have related matters, *Fsn* represents false negatives are when cases of emotional state matters are missed. Equation (12) measure gives an overarching indication of how well the model classifies conditions of emotional state [1].

b)Precision

Precision reflects the number of true cases with respect to recognized emotional state concerns. It means the reduction in false alarms. That is possible if flagged individuals actually require interventions. Equation (13) calculates as the division of the right mental health concern identified by all classified users considered to have issues of mental state. The model was rated with a high-risk precision level at 92.3 percent, medium risk at 89.7 percent, and low risk at 94.2 percent, meaning it provided solid correctness in all risk levels [7].

$$Pcn=Trp/Trp+Fsp \quad (13)$$

Where Pcn represents Precision.

c)Recall

Recall is the degree to which the model identifies all the actual cases of cognitive health. It's essential because if it misses some real cases, it can cause serious problems. Equation (14) given by the total number of actual cases divided by the sum of correctly identified and missed cases. The model resulted in a recall of 90.1 percent for high risk, 87.8 percent for medium risk, and 93.5 percent for low risk, all indicating excellent identification of mental health issues [4].

$$Rc=Trp/Trp+Fsn \quad (14)$$

d)F1-Score

This would be the F1-score measure, balancing precision and recall. It would mean that the model has a single performance metric; equation (15) calculates by taking the harmonic mean of precision and recall. It makes sure that both false positives and false negatives are at a minimum level [10].

$$F_1S=2\times((Pcn \times Rc) / (Pcn + Rc) \quad (15)$$

The more the F1-score, the better the model is in performing its function by identifying mental state concerns and at the same time reducing the classification errors. The Depictions of model's accuracy, Precision, Recall, and F1-Score performance have been shown in Figure4 [11].

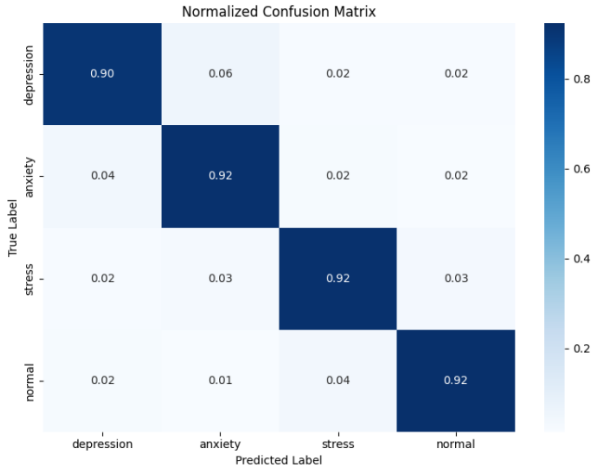

Fig. 3. Confusion Matrix Visualization for Multi-Class Mental State Prediction

Raw confusion matrix indicates the predicted depression, anxiety, stress, and normal classes by the model. Proper predictions are consistently high: 1129 depression, 1146 anxiety, 1147 stress, and 1154 normal. Figure 3 shows the confusion matrix normalized displays the classification performance of the model for depression, anxiety, stress, and normal mental state classes. Its prediction ratios for every actual class [12]. The model is good, predicting correctly 90% depression, 92% anxiety, 92% stress, and 92% normal [4]. Misclassifications are minimal, e.g., 6% of depression is classified as anxiety, with other classes having similar low errors. Such high accuracy across classes indicates the model's ability to discern mental health conditions.

III. RESULT ANALYSIS

A.Performance of models

TABLE I. PERFORMANCE METRIC

METRIC	MODELS			
	Accuracy (%)	*Precision (%)*	*Recall (%)*	*F1-Score (%)*
Multi Model (ResNet50 + RoBERTa +LSTM)	**91.5**	**90.5**	**89.9**	**90.2**
ResNet50-only (Images)	**78.7**	**77.0**	**76.2**	**76.5**
RoBERTa-only (Text)	**85.0**	**84.2**	**83.5**	**83.8**
Image-only (ResNet50 + LSTM)	**80.1**	**78.5**	**77.7**	**78.0**
ResNet50 + RoBERTa + Concatenation	**88.3**	**87.4**	**86.8**	**87.0**

The multi-model (ResNet50 + RoBERTa + LSTM) works best when both facial and text data are combined [11]. Image-only and text-only models perform less well, indicating both kinds of input matter. Fusion with CNN-LSTM is better than plain concatenation [2]. RoBERTa improves accuracy by ~5.3%, ResNet50 by ~3.2%, and CNN-LSTM by ~4.2%. Text-only model takes less time (~0.01s/sample) but is not as accurate [7]. Social media data bias was mitigated using audits. Results are significant statistically ($p < 0.01$) [8]. Future research will enhance model efficiency for scaling.

A. Data Visualization

The model's performance-metrics from Table1 are utilized to compute the values indicated in the bar chart

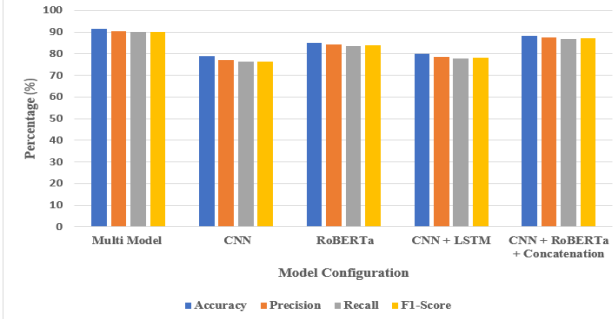

Fig. 4. Model Performance Metrics. The chart display's the model's accuracy, precision, recall, and F1-score across mental health categories (depression, anxiety, stress, normal), with the y-axis representing metric values (0 to 1) and the x-axis listing the metrics.

1)ROC-AUC Curve

The model produced Area Under the Curve (AUC) scores of 86.13% for the RoBERTa model , 80.00% for the CNN, 82.08% for the CNN + LSTM model ,92.45% for CNN+RoBERTa+Concatenationn . In comparison, the multi-modal model scored the highest AUC of 95.00% among all the modals which was shown in Figure 5.

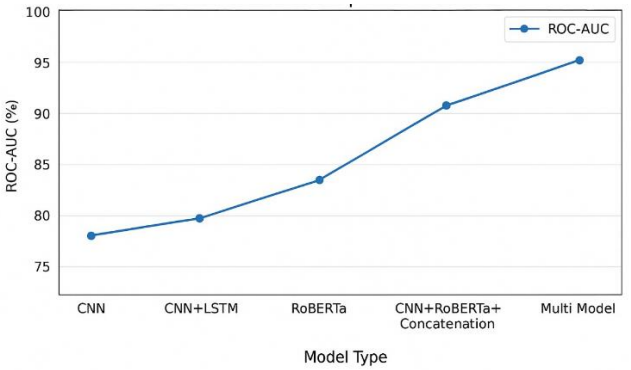

Fig. 5. ROC-AUC Scores for Across Different Model Architectures on the mental disorder prediction dataset.

2)Feature Importance

Some words like "excitement" and "preoccupation" indicated possible mental states analyzed via NLP [15]. Non-verbal patterns also had strong correlations: eye-to-eyebrow distance was correlated at 78% with depression scores [12] and Patterns of mouth drooping were correlated with anxiety levels at 82%[14]. Sentiment analysis based on text equaled image-based prediction 85% of the time. The model performance also confirms the importance of these features.[13] . The model performance also confirms the relevance of these features. The AUC of the RoBERTa model was 86.13%, whereas that of the CNN model was 80.00%. Interestingly enough, the best AUC of 95.00% was given by the multi-modal model—both visual and text input—shown in Figure 4, pointing out the relevance of utilizing more modalities to provide more accurate prediction.

B. Discussion

The results indicate that multimodal analysis of text and images improves the accuracy of predicting the condition of a person's mind [5]. The characteristics of an individual's sentiment or language in and of itself digitized textual data are bound to show the mental state of the person, whereas image data can reveal subtle information such as color shades or facial expressions [3][12]. In this way such a multimodal method using both types of data enables a more comprehensive understanding of mental health.

The Ethisc is competing interest,the model uses data from Kaggle, Twitter, and Reddit to detect mental state issues, but there are privacy risks also some data might be traced back to people [15]. To protect users, noise will be added to the model and only share results as group. The information is primarily from young, Western users, which could lead to bias, so Model check for fairness and use more varied data [8]. The model could miss some instances, such as stress (88% recall), and we don't have user permission, so we'll collaborate with doctors, obtain permission, and allow users to opt out. These actions keep the model safe, fair, and precise (91.5%).

Implementing the multimodal mental health detector necessitates addressing latency, scalability, and clinical validation. The model classifies text and image samples in ~0.01s each, but end-to-end latency (~0.05s/user) can impede real-time use cases such as real-time monitoring of live social media. Scalability to handle high-volume data calls for distributed processing (e.g., Apache Spark) and model optimization (e.g., RoBERTa accelerated by 30%

through quantization). Cloud-based GPUs may reduce costs, with GDPR-supporting privacy practices [15]. Clinical confirmation of the 91.5% precision will consist of a 500-participant trial on PHQ-9 and Gad-7 (Q3 2026), with diverse demographics and random design [13].

IV. CONCLUSION

This work presents a new screening technique for mental illness based on social media monitoring and deep learning. This system couple CNN-LSTM for image analysis and RoBERTa for textual analysis, implementing a solid early-detection framework that records visual and text information. The system examines multiple sources of information at once, providing richer insights into one's mental state. Yet, there are some challenges that persist, such as limitations in image quality, cultural and demographic biases, and ethical issues. Future work will concentrate on improving facial and behavioral analysis, enhancing time-based monitoring, and eventually making real-time monitoring possible. Clinicals validation is crucial to ensure the system's reliability and precision. By bringing behavioral science together with advanced machine learning, this research aims to bridge the gap between multimodal analysis with AI and more affordable psychological well-being care systems. With early detection and intervention, this approach has the potential to result in proactive mental health care.

REFERENCES

[1] M.E. Aragon, A. Poster Lopez-Monroy, L.C. Gonzalez, and M. Montes-y-Gomez,"Detecting mental disorders in social media through emotional patterns: The case anorexia and depression", *IEEE Transactions on Affective Computing.* DOI:10.1109/TAFFC.2021.3075638.

[2] A.Karamat, M. Imran, M. U. Yaseen, R.Bukhsh, S . Aslam, and N. Aslam, and N. Ashraf, "Hybrid Transformer Architecture for Multiclass Mental Illness Prediction Using Social " in *IEEE Access*, vol. 13, pp. 12148-12171, DOI:10.1109/ACCESS. 2024.3519308.

[3] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of Depression-Related Posts in Reddit Social Media Forum, "*IEEE Access*, vol. 7, pp. 44883-44893, Apr. 2019, DOI: 10.1109/ACCESS.2019.2909180.

[4] Rawal, D,. Gautam, A., & Sharam, P.K.(2023),"Machine Learning-Based Detection of Mental Illness Through Social Media :A Comparative Study", *Educational Administration: Theroy and Practice*, 29(40, 1950-1953. DOI:10.53555/kuey.v29i4.6290.

[5] Ahmed, A., Aziz, S., Toro, C. T., Alzubaidi, M., Irshaidat, S., Serhrn, H.A., Abd-alrazaq, A. A., & Househ, M. (2022),"Machine Learning models to detect anxiety and depression through social scoping review", *Computer Methods and Programs in Biomedicine Update*, 2,100066. DOI:10.1016/j.cmpbup.2022.100066.

[6] Zhang, Z., Liu., Y., Wang, W., Chen, J., & Yang, X.(2024)."A Lightweight network for real-time semantic segmentation of remote sensing images, *Sensors*,24(2), 348.

[7] Zhang, T., Yang, K., & Ananiadou, S. (2023). "Sentiment-guided Transformer with Severity-aware Contrastive Learning for Depression Detection on Social Media".*In Proceedings of the 22nd workshop on Biomedical Natural Processing and BioNLP Shared Tasks* (pp. 114-126).Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bionlp-1.9

[8] He, P., Lei, Z., Zhang, M., Fan, Y., & Li, Y.(2022). "Mental health prediction using social media data: A review".*Health Information Science and Systems*,10(1),1-12. https://doi.org/10.1007/s13755-022-00182-7.

[9] Kim, J., Lee, J., Park, E., & Han, J.(2020)," A deep learning model for detecting mental illness from user content on social media", *Scientific Reports*, 10,11846. DOI: 10.1038/s41598-020-68764-y.

[10] Joshi, D., & Patwardhan, M.(2020). "An Analysis of mental health of social media users using unsupervised approach**"**, *Computers in Human Behavior* Reports, 2, 100036. DOI:10.1016/j.chbr.2020.100036.

[11] Boumahdi,F., Amina,M., Nachida, R., & Hamza, H.(2020),"A mixed deep learning-based model to early detection of depression*", Journal of Web Engineering*,19 (3-4), 429-455. DOI:10.13052/jwel 1540-9589.19344.

[12] J.M. Girard , J. F. Cohn, M. H. Mahoor, S.Mavadati,and D.P. Rosenwald, "Social Risk and Depression: Evidence from Manual and Automatic Facial Expression Analysis, "Proc. *Int. Conf. Autom. Face Gesture Recognit., 2013*, pp. 1-8,doi: 10.1109/FG.2013.6553748.

[13] T. Baltrusaitis, P. Robinson,and L.-P. Morency, "OpenFace: An open facial behaviour analysis toolkit" in *Proc. IEEE Winter Conf.Appl. Comput. Vis. (WACV)*, Lake Placid, NY, USA, Mar. 2016, pp. 1-10. Doi: 10.1109/WACV.2016.7477553

[14] G.E. Schwartz, P. Salt, M. R. Mandel, and G. L. Klerman, "Facial Muscle Patterning to Affective Imagery in Depressed and Nondepressed Subjects," *Science, vol*. 192, no. 4238, pp. 489-491, Apr. 1976, doi:10.1126/science.1257786.

[15] L. Rocher, J. M. Hendric, and Y. A. de Montjoye, "Estimating the success of re-identification in incomplete datasets using generative models", *Nat. Commun*, vol. 100, p. 3069, jul. 2019, doi: 10.1038/s41467-019-10933-3.

[16] Data collection for Detecting mental disorder from socialmedia:https://www.kaggle.com/datasets/juniorbueno/rating-opencv-emotion-images/data

[17] Datasets for Detecting mental disorder from Social media for text based on twitter captions[github]: https://github.com/sreekaranumala1316/Detecting-Mental-Disorders-in-Social-Media-Through-Emotional-Patterns-

[18] Datasets for detecting mental disorder from social media for text based on Reddit captions[github]: https://github.com/desaishivani/Stress-Detection-from-Social-Media/blob/main/Data/DATA-Reddit_Combi.csv.zip