


# A Novel Multimodal Fusion Algorithm for Non-Intrusive Anxiety Detection

**Mahir Shadid**

(International Islamic University Chittagong, Chittagong, Bangladesh

 <https://orcid.org/0000-0001-7317-1714>, mahir.shadid@gmail.com)


**Mushfiqus Salehin Afnan**

(International Islamic University Chittagong, Chittagong, Bangladesh

salehinafnan@gmail.com)


**Rashed Mustafa**

(University of Chittagong, Chittagong, Bangladesh

 <https://orcid.org/0000-0001-5123-194X>, rashed.m@cu.ac.bd)

**M. Jamshed Alam Patwary**

(Chittagong University of Engineering & Technology, Chittagong, Bangladesh

 <https://orcid.org/0000-0001-5110-4625>, jamshed@cuet.ac.bd)

**Abstract:** Early detection of anxiety disorders in a non-intrusive manner is crucial, as these conditions can profoundly impact an individual's health and daily functioning. Traditional approaches relying solely on unimodal data often fall short, potentially introducing bias and inaccuracies. TI-Fusion is a novel late multimodal fusion technique that integrates text and image data for a unified reliable outcome, overcoming limitations in existing methods. The primary advantage of TI-Fusion is its non-intrusive nature, ensuring patient comfort by avoiding invasive methods while still delivering robust diagnostic capabilities. The study utilizes six advanced machine learning algorithms (Gaussian Naive Bayes, XGB Classifier, K-Neighbors, SVM, Decision-Tree, and RandomForest) for data classification, pattern recognition, and predictive accuracy. Concurrently, image data from the KDEF and CK+ datasets was processed through a Convolutional Neural Network (CNN) enhanced with a Real Gabor filter, which is particularly adept at capturing textures, edges, and complex visual patterns necessary for precise image analysis and recognition. By employing a late multimodal fusion approach, TI-Fusion integrates the outcomes of models trained on distinct data modalities, yielding a more comprehensive and accurate prediction than unimodal methods. This technique not only surpasses existing multimodal approaches but also achieves a commendable final accuracy rate of 92.38%, demonstrating its effectiveness in enhancing the early detection of anxiety disorders.

**Keywords:** Late Fusion, Anxiety Disorder, Convolutional Neural-Network, GridsearchCV, Feed-forward NN

**Categories:** H.3.1, H.3.2, H.3.3, H.3.7, H.5.1

**DOI:** 10.3897/jucs.127703

## 1 Introduction

Early and accurate assessment of anxiety disorders is critical for effective treatment and preventing long-term mental health consequences. Machine learning has proven valuable

in this domain, yet relying on a single data modality, such as text or images, can lead to less reliable results. For instance, individuals with anxiety may struggle to express their feelings verbally, leading to potential inaccuracies when relying solely on textual data. Moreover, unimodal approaches risk false positives or negatives, as highlighted by Haas and Hüllermeier's research [Haas and Hüllermeier 2023].

Another significant barrier to effective mental health care is the intrusiveness of traditional diagnostic methods, which can overwhelm patients and deter them from seeking help. The proposed model integrates both image and text data to produce a unified and more accurate output. Importantly, this model is entirely non-intrusive, making it more accessible and patient-friendly.

The proposed model leverages supervised learning, utilizing image data from CK+ [Lucey, Cohn, Kanade, Saragih, Ambadar, and Matthews 2010] and KDEFS (<https://kdef.se>), alongside text data from the DASS-21 questionnaire. By combining these modalities, the model offers a more reliable and accurate method for anxiety detection compared to traditional unimodal approaches.

The key contributions of this study are as follows:

1. A Text-Image Fusion (TI-Fusion) approach based on multimodal learning is proposed, which employs an innovative algorithm that produces output by fusing image and textual data.
2. Ensure the entire anxiety detection procedure is convenient, non-intrusive, and easy to deploy, with a specific focus on patient ease and comfort.
3. Conduct practical experiments to assess the effectiveness of the suggested methodology evaluating Efficiency, accuracy rates, and practicality on a variety of people, offering empirical proof of the approach's capability and its incorporation into actual anxiety detection technologies.

This paper is structured as follows: Section 2 reviews recent research on mental disorders, Section 3 details the proposed methodology, Section 4 presents the study results, Section 5 discusses limitations and future work, and Section 6 concludes the research.

## 2 Literature Review

Anxiety disorder is an underestimated mental issue, but it has a significant impact on an individual's daily functioning, such as causing quick fatigue, uncontrollable emotions, and reactions. To improve early detection for better diagnosis, research in recent years has focused on two categories:

### 2.1 Unimodal Approaches:

Many research investigations have explored the application of machine learning algorithms in identifying anxiety. For example, Osman et al. [Osman, Tabassum, Patwary, Imteaj, Alam, Bhuiyan, and Miraz 2022] performed a study analyzing past studies on mental health issues using various machine learning and deep learning methodologies. These publications shed light on the novel methodologies used to examine the role of machine learning in mental health. Priya et al. [Priya, Garg, and Tigga 2020] used machine

learning models to estimate anxiety, depression, and stress severity levels utilizing the DASS-21 survey in another study. They achieved an average accuracy of approximately 80% across models such as Decision Tree, Random Forest, Naive Bayes, Support Vector Machines, and K-Nearest Neighbor. Ahmed et al. [Ahmed, Sultana, Ullas, Begom, Rahi, and Alam 2020] employed algorithms such as Convolutional Neural Networks, Support Vector Machines, Linear Discriminant Analysis, KNN, and Linear Regression to predict anxiety with a high 96% accuracy. These models have the potential to lead to early mental health detection and intervention, which in turn will reduce the number of self-harm incidents.

Despite the promising outcomes, these approaches can be improved. Notably, as indicated in psychological studies [Scherer 2009, Lensvelt-Mulders 2012], reliance on surveys and questionnaires may lack emotional depth and induce socially desirable responses. Furthermore, because mental feelings cannot be witnessed externally, relying simply on visual sources such as images and videos may not adequately represent the complex nature of anxiety disorders.

## 2.2 Multimodal Approaches:

To overcome the mentioned constraints, it is critical to employ methodologies that combine many forms of data to get a cohesive result. Text, photos, music, video, and sensory data are all examples of multimodal data. Recent research has explored combining multimodal approaches with machine learning to improve the identification of anxiety disorders. Naderi et al. [Naderi, Soleimani, Rempel, Matwin, and Uher 2019] developed a multimodal deep neural network system that automatically identifies mental disorders by combining speech and language data from clinical interviews. Guo et al. [Guo, Fu, Pan, Zhang, and Hu 2020] used a combination of electroencephalogram (EEG) and eye movement (EM) data to achieve accurate anxiety detection using the K-GSCCA approach, reaching an accuracy rate of 87.47%. Additionally, Xie et al. [Xie, Wang, Lin, Luo, Chen, Xu, Liang, Liu, Wang, Luo, et al. 2022] used a CNN-LSTM-based multimodal model to determine depression and anxiety, with video-based strategies attaining an 83.78% classification accuracy. Cao et al. [Cao, Wu, Huang, Patwary, and Wang 2022] presented a Multi-modal Feature Fusion (MFF) technique for dealing with Generalized Zero-Shot Learning (GZSL). This strategy improves the quality of pseudo-sample/feature generation and increases the consistency between the created features and the prior semantic information. Asma et al. [Asma, Mostafa, Akter, Mahmud, and Patwary 2022] proposed a fuzziness-utilized semi-supervised deep learning approach for multimodal image categorization, dubbed FSSDL-MIC. This method outperforms current multimodal image categorization algorithms. Furthermore, Patwary et al. [Patwary, Cao, Wang, and Haque 2022] used FSSL-PAR, a multimodal semi-supervised learning method, reducing the reliance on labeled data and domain-specific ability. In the study [Jang, Choi, Kim, Yu, Jeon, and Byun 2023] (2023), physiological data from recovery phases, including EDA, ECG, PT, and RESP, were evaluated to investigate machine learning's ability to differentiate between individuals with symptoms of panic disorder (PD), various anxiety-related conditions, and healthy controls. The importance of ECG and PT features during stress-recovery periods as critical predictors of PD was underlined in this study. The study found that the multilayer perceptron model with all 33 ECG features achieved the highest accuracy (75.61%) when using logistic regression, k-nearest neighbor (k-NN), support vector machines (SVM), random forest (RF), and multilayer perceptron (MLP) methods combined with layered cross-validation. The model outperformed peers that relied on a more limited range of ECG characteristics. A similar approach that provides

support to the method of fusing text and image was presented by Lai et al [Lai and Li 2023]. Their method achieved F1-score of 70.96%. The goal of the study by Vaz et al. [Vaz, Summavielle, Sebastião, and Ribeiro 2023] is to use physiological data acquired from an impartial sample of healthy volunteers to classify anxiety as an imbalanced classification issue. The study used EMG, ECG, as well as Electrodermal Activity (EDA) readings from the publicly available WESAD dataset to do this. By combining machine learning models with ensemble approaches and a proposed pipelining methodology, the whole strategy demonstrated strong performance. Similarly, A machine learning model for pre-sleep anxiety identification was developed in the study of [Cai 2024] utilizing ambulatory ECG together with T-ACC data from college students. QDA, SVM, KNN, DT, and LDA classifiers were utilized in the model, which produced an accuracy of 66.67%.

As highlighted in [Hansson, Lexén, and Holmén 2017, Rubeis and Steger 2019], the process of invasive detection may create sentiments of stigma, violation, and powerlessness, potentially culminating in emotions such as shame, embarrassment, and discomfort.

### 3 Proposed Methodology

In the proposed approach (Figure 1), text and image data are initially prepared, split into training and testing sets, and test data sizes are equalized for late fusion. Late fusion combines predictions from various modality-trained models and uses them to train a general feed-forward neural network. This "TI-Fusion" method, or Text-Image Fusion, employs CNN and ML models. The late fusion combines data by anxiety label into a single data-frame, enabling accurate anxiety detection.

The description on the key concepts of the approach are:

- **Common-key Merging:** A crucial aspect of the proposed approach is the 'common-key' technique, which ensures that the anxiety labels of text and image data are synchronized when merging them into a single dataframe for the fusion model, facilitating accurate fusion based on anxiety predictions.
- **Multiplication of Predicted Data Dependent Variables:** The Late Fusion Data for training and testing is formed by combining both text and image dependent variables to create a single final dependent variable: Anxiety or Not Anxiety, while excluding the mixed condition (Anxiety \* Not Anxiety).
- **Late Fusion Model (Feed-Forward NN):** The feed-forward neural network used is a two-layered NN model, with the first layer consisting of a dense layer with 64 units. Many aspects are balanced in the neural network layer for the selection of 64 units for handling multimodal fusion data. Limiting this dimensionality to 64 units greatly reduces computational cost without sacrificing important characteristics. It captures the most important patterns and features of the image, achieves faster training and inference times, and helps prevent overfitting by lowering the model's capacity. Additionally, it sets negative input values to zero for ReLU, which increases sparsity and can enhance model performance. L2 regularization reduces overfitting by lowering the model's complexity and addresses data imbalances and selected the optimal ones by experimenting with various parameters for these functions.

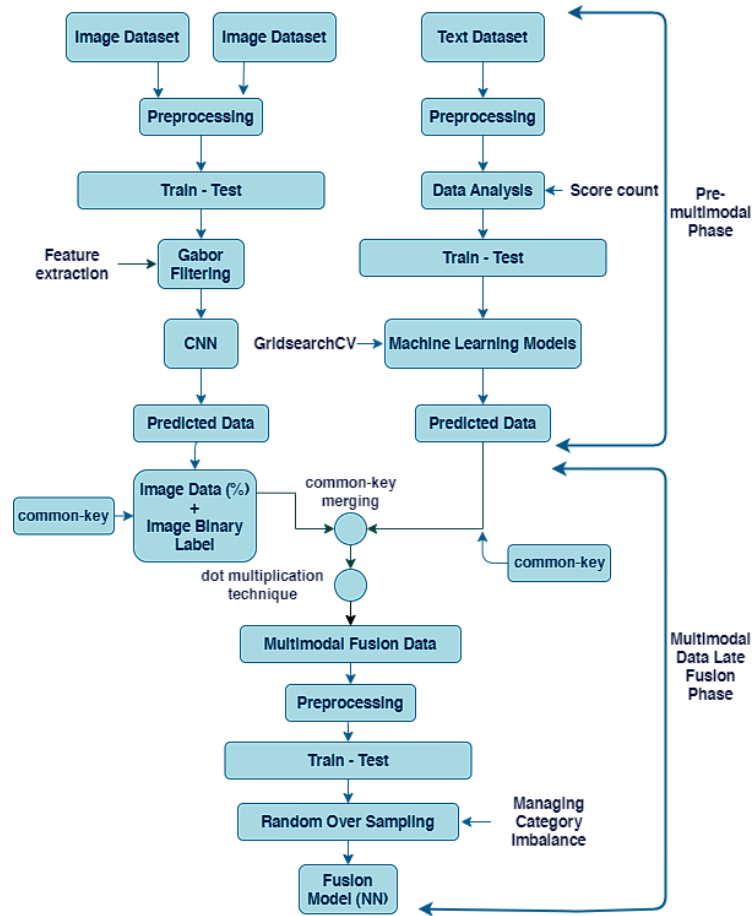


Figure 1: TI-Fusion approach workflow

The second layer is a dense layer with a sigmoid activation function. By using the sigmoid activation function, each real number is mapped to a range of values from 0 to 1, which is extremely beneficial for binary classification tasks that require a probability score in the output. The result for anxiety prediction can be interpreted as the probability that an individual experiences anxiety (class 1) or not (class 0). The output of the sigmoid function is commonly understood as the probability that the input belongs to a certain class in binary classification. For example, a sigmoid output close to 1 in the context of your anxiety detection model indicates a high probability of anxiety in the individual, while a value near 0 signifies a low probability.

To make the approach modular and to trace the data flow easily, the method is built in two phases: the Pre-multimodal phase, which trains separate models, and the Multimodal phase, which combines data into a single form and trains a single model on the combined data to generate output. The problem formulation, followed by these phases and the algorithm, is detailed in the subsection *Problem Formulation*.

### 3.1 Problem Formulation

Conventional multimodal algorithms often struggle due to three crucial requirements: matching dimensions, common information, and identical dependent variables. The absence of these parameters in multimodal data leads to issues of dimension conflicts and unmatched data fusion. Since distinct modalities (such as text and images) naturally have different dimensionalities and formats, traditional multimodal algorithms sometimes have difficulty aligning data from these sources. Advanced preprocessing and feature extraction techniques are necessary for effective alignment to ensure that data from various sources can be beneficially merged. Finding and utilizing common knowledge across modalities is another challenge. Multimodal data fusion requires integrating complementary information from each modality to enhance overall classification performance. Understanding how several modalities interact and contribute to the desired result can be challenging. Additionally, ensuring that the dependent variables—the target labels—are the same for all modalities is crucial. Maintaining uniform labeling and annotation across all data sources is essential, as inconsistencies can negatively impact model performance. To address these challenges, TI-Fusion, an algorithm that seamlessly handles these requirements, is proposed. It only requires uniformly expanded or trimmed predicted datasets after pre-multimodal training. The key difference between conventional methods and the suggested approach is its common-key merging. Conventional methods require various algorithms to create a fusion model, which demands more computational power, analysis, and time. In contrast, the suggested method merges data before training the late fusion model using common-keys, which preserves the input data according to their target labels and reduces the need for additional fusion algorithms.

### 3.2 Pre-Multimodal Phase

To provide a more comprehensive evaluation of the method, priority has been given to the separate models known as "Pre-multimodal phase models," as depicted in Figure 3. As previously mentioned in Section-III, image classification was conducted using Convolutional Neural Networks (CNN). Since pixels are organized into a pattern like a grid, images are by nature spatial. CNNs utilize convolutional layers to collect local patterns as well as relationships among pixels by applying filters. This allows them to take advantage of this spatial information.

The text data uses Machine Learning (ML) algorithms. The CNN model is enhanced with Real Gabor Filtering to enhance image recognition with 64 units. This model accepts standardized and Gabor Filtered numpy arrays as input, featuring crucial layers, including a dropout layer to mitigate overfitting. The ML models used (Gaussian Naive Bayes, XGB Classifier, K-Neighbors, SVM, Decision-Tree, and RandomForest) undergo hyperparameter tuning through GridsearchCV. For text classification, classic machine learning methods that are used in this study work well. Given that textual data is organized and frequently represents features as vectors in a space with high dimensions, these algorithms are capable of handling it with ease. Following preprocessing, the dataset is fitted to these models.

Model: "sequential"

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 48, 48, 6)	156
max_pooling2d (MaxPooling2D)	(None, 24, 24, 6)	0
dropout (Dropout)	(None, 24, 24, 6)	0
conv2d_1 (Conv2D)	(None, 24, 24, 16)	2416
max_pooling2d_1 (MaxPooling2D)	(None, 12, 12, 16)	0
dropout_1 (Dropout)	(None, 12, 12, 16)	0
conv2d_2 (Conv2D)	(None, 10, 10, 64)	9280
max_pooling2d_2 (MaxPooling2D)	(None, 5, 5, 64)	0
dropout_2 (Dropout)	(None, 5, 5, 64)	0
flatten (Flatten)	(None, 1600)	0
dense (Dense)	(None, 128)	204928
dropout_3 (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 2)	258

Figure 2: Summary of the CNN Model

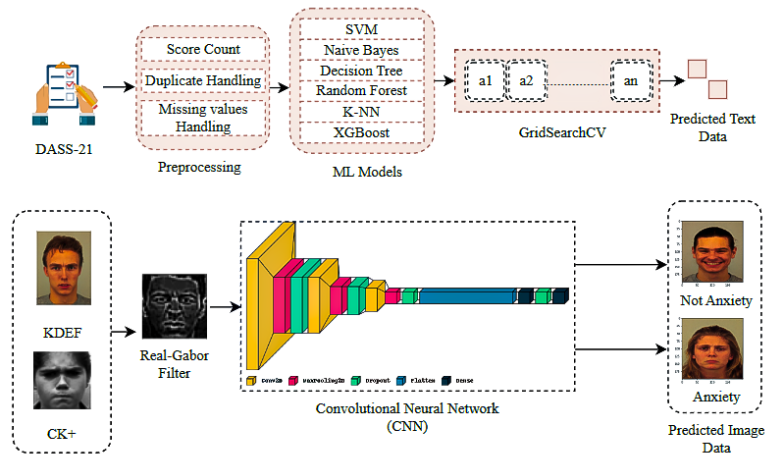


Figure 3: The Pre-multimodal Phase Structure

### 3.3 Multimodal Late Fusion Phase

This phase dissects multimodal late Fusion algorithm (Algorithm-1) and structure (Figure 4).

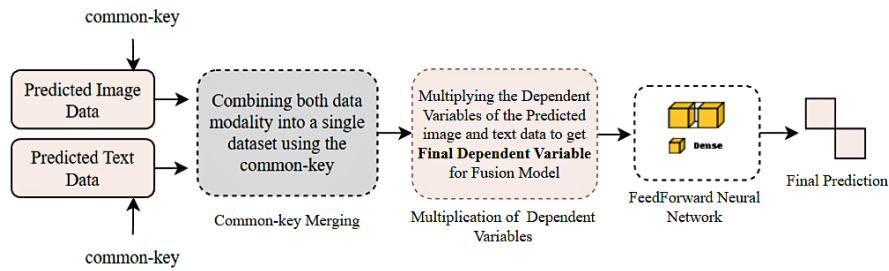


Figure 4: Multimodal Late Fusion Phase Structure

**Input:** Predicted Image-Data

**Input:** Predicted Text-Data

**Output:** Prediction of anxiety (0-1)

**Step 1: Initialize**

Initialize variables;

**Step 2: Assign Common-Key**

for  $i = 1$  to  $Length$  of Dataset do

$common-key \leftarrow$  ('not anxiety' if x is 1 else 'anxiety');

end

**Step 3: Merge Predicted Data**

if  $Common-Key (Text) = Common-Key (Image)$  then

$merged-data \leftarrow$  Merge Predicted Text and Image Data;

end

**Step 4: Fuse dependent variables**

$fusion-data \leftarrow$  Merged Data;

$fusion-data \leftarrow$  Fuse dependent variables of Merged Data into final "Anxiety Label";

**Step 5: Preprocess and Split**

**Step 6: Build Late Fusion Model**

$fusion-model \leftarrow$  Build Feed-Forward Neural Network;

$fusion-model \leftarrow$  Train Model with Fusion Data;

**Step 7: Return Prediction**

return Prediction of anxiety

**Algorithm 1:** Multimodal Late Fusion Phase Algorithm

The algorithm shows the mechanism of data fusion. Predicted data coming from the trained ML and CNN models are the inputs of the late fusion phase algorithm. The common-key is a common string that is added to the predicted datasets to combine them in a single dataset. These common-keys act like checkpoints to fuse data. Next, to train the fusion model, a dependent variable is needed as the proposed method is based on supervised learning. This variable is created by multiplying the dependent variables of the predicted image and text data, where the multiplication acts like the AND operator to provide priority to both text and image dataset. Finally, the fusion model is trained on the fusion data after necessary data preprocessing like Random oversampling because

when the target variable is imbalanced, the model's results are more heavily influenced by the majority class [Niaz 2022]. The ROS has been done on the training data with the help of the imblearn ROS library to create balance with the random seed 32 which ensures reproducibility of results. The output provides a value between 0 and 1 to show the level of severity.

### 3.4 Interpretability of the Machine Learning Models

The interpretation of predictions is greatly impacted by the combination of text and images in machine learning models, especially when dealing with complex tasks as anxiety detection. Image data, particularly facial expressions, provides visual indications that are essential for understanding non-verbal communication, while text data gives a binary classification of anxiety. Combining these modalities allows the model to generate predictions that are more thorough and meaningful. For example, in order to ensure accuracy and reliability in its predictions, the model cross-references the classification of anxiety with matching facial expressions when predicting anxiety. By providing an improved understanding of the data and allowing users to link predictions to non-verbal as well as verbal inputs, this multimodal method improves interpretability.

- **The Significance of Facial Expressions in Accurate Classification:** Facial expressions are essential for increasing the accuracy of classification models, particularly when it comes to conditions involving emotions such as anxiety. Machine learning algorithms have the ability to collect and analyze slight differences in facial movements, which are frequently undetectable to the human eye, in order to identify underlying emotions. Facial expressions allow the model to capture an individual's hidden as well as visible emotional states, which greatly improves the model's accuracy in classifying anxiety. This aspect is essential for both enhancing performance and guaranteeing that the model's predictions are supported by observable, physiological data, which increases its validity and reliability.
- **Contribution to Fusion Model's Overall Performance:** When features like facial expressions are included in a fusion model, improvements in performance are significant. This is so that the model can produce more complex and precise predictions. Facial expressions give an extra layer of data that supplements textual data. The collaborative impact of merging various data sources is advantageous to the fusion model in the setting of anxiety detection, where non-verbal as well as verbal indicators are crucial. With an expanded representation of the emotional state generated by the model due to this multimodal approach, classification accuracy improves and the results are easy to understand.

### 3.5 Dataset

The CK+ and KDEF were acquired from their respective authors [Lucey, Cohn, Kanade, Saragih, Ambadar, and Matthews 2010] and website (<https://kdef.se>). This combined image dataset comprises approximately 3900 photos of size 48x48. Both image datasets combinely contain eight emotion descriptors (neutral, contempt, anger, disgust, fear, sadness, and happiness), which, during implementation, underwent transformation into two categories: anxiety and not anxiety.

The text data (DASS-21) was collected from the authors of [Priya, Garg, and Tigga 2020] containing 580 rows of responses from the subjects who took DASS-21 test having

21 features. To maintain the authenticity of the research, the dataset was obtained from them. To ensure consistent dimensions when merging the data, an equal number of sets were used for testing, which contributed to stability during the training of the late fusion model. The remaining data was employed for training purposes.

### 3.6 Data Preprocessing

The critical aspect of text data analysis lies in scoring determinants after implementing necessary preparation approaches, such as managing null values and handling duplicates. To collect data from the subjects, researchers cited in [Priya, Garg, and Tigga 2020] utilized Google Forms. Among the 24 features present in the DASS-21 dataset, seven (Table 1) can be effectively utilized for anxiety detection. Classification of anxiety is based on the score count. If an individual's score falls within the range of 0 to 7, a "not anxious" classification is assigned, while scores outside of this range indicate an "anxious" classification. The score count can be calculated using the predefined DASS-21 equation:

$$\text{score} = \text{sum of each determinant's rating points} * 2$$

Determinants of Anxiety
1 Dryness in the mouth
2 Breathing difficulty
3 Experience trembling
4 Worried about panic and make a fool of myself
5 Closer to panic
6 Aware of the heart's action in the absence of physical exertion
7 Felt scared without any valid reason

Table 1: Seven Anxiety Determinants of DASS-21

Probable ratings from the subjects are as follows: 0 = did not apply to the individuals; 1 = applied to the individuals to some degree; 2 = applied to the individuals to a considerable degree; 3 = applied to the individuals very much.

In the context of image data analysis, the feature extraction process occurred subsequent to the transformation of the images into numpy arrays using the Gabor filter, but before that, the images of the KDEP were turned into grayscale format and resized to 48x48 pixels to make them compatible with the CK+ dataset. In the function of Gabor Filter, a standard normalization was done by dividing the Gabor features with 255.0, which transformed each feature within a range of 0 to 1 for better prediction. The designations "anxiety" and "not anxiety" were used to represent positive and negative emotions [Surcinelli, Codispoti, Montebanocci, Rossi, and Baldaro 2006].

## 4 Results and Discussion

In this section, the method's experimental results are thoroughly evaluated and compared against various techniques to validate the approach. Efficiency is demonstrated through

a comparison chart and metrics such as confusion matrices, accuracy, training curves, and loss curves. The model's ability to predict anxiety disorder using self-collected data is also assessed and compared with alternative methods.

Furthermore, after completing the essential steps, an experiment was conducted to evaluate the proposed methodology's viability, focusing on assessing invasiveness, accuracy, and testing in real-world scenarios.

#### 4.1 Pre-multimodal Phase

For **Image Data**, findings were obtained by investigating the accuracy of selected image data. Randomly selected images were input into the CNN model, and predictions were generated. To assess overfitting in the CNN model, training and loss accuracy curves were analyzed (Figure 5). Overfitting typically occurs when training loss and accuracy continue to improve while validation loss rises and validation accuracy decreases after a certain point. However, this behavior was not observed in the CNN model.

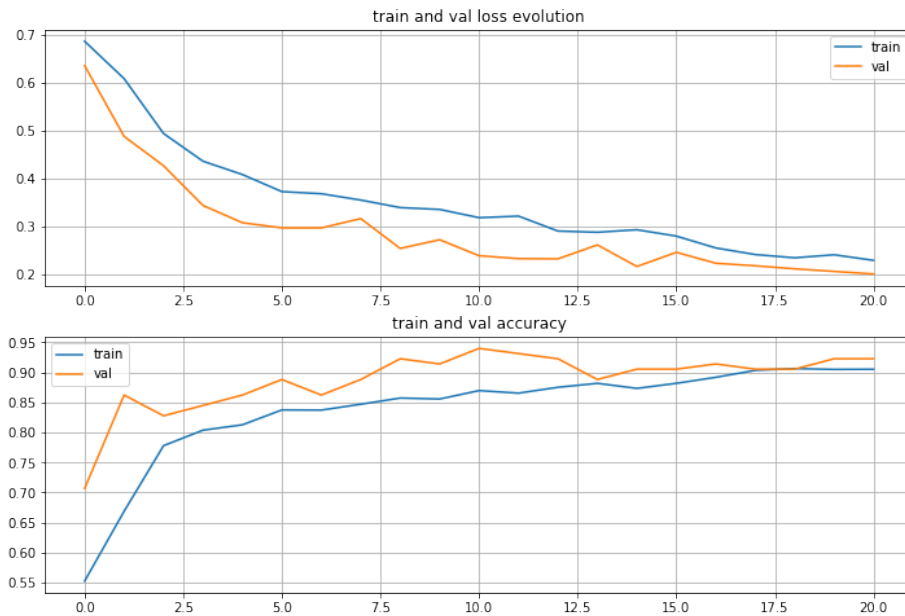


Figure 5: Training Loss and Accuracy Curve (CNN)

Overall accuracy of the CNN-model is **91%** which is better than [Eng, Ali, Cheah, and Chong 2019]. Metrics like precision, recall and f1-score are **91%, 89%, 90%** respectively. The CNN-model applied on the dataset image (Figure 6) is performing very well achieving the accuracy of 100%.

For **text data**, the models were trained using the DASS-21 dataset, and two stages were examined: one with hyperparameter tuning and another without hyperparameter tuning. The overall average accuracy of the ML-GridSearchCV model is **88%**. After

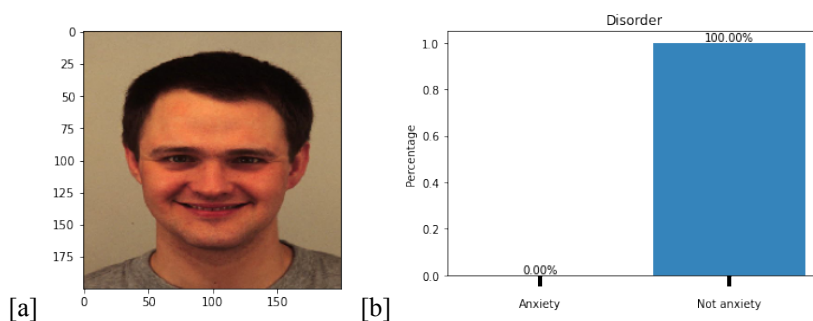


Figure 6: Testing CNN on Dataset Image. (a) Happy face; (b) CNN prediction

applying appropriate hyperparameter tuning, the SVM classifier achieved **100%** accuracy, precision, recall, and F1 Score.

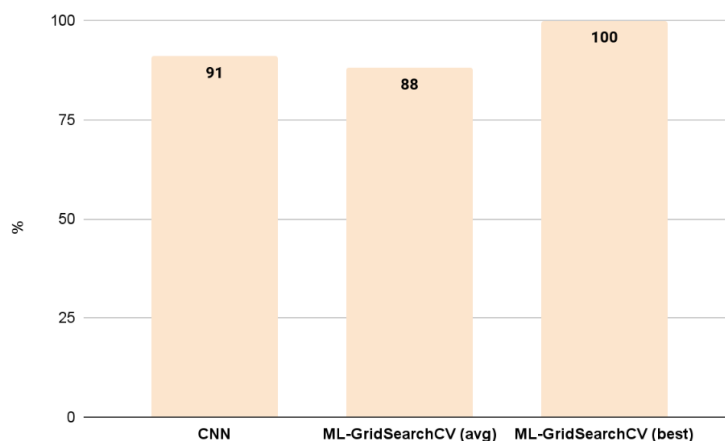


Figure 7: Pre-multimodal Model Performances

Anxiety Labels: 0 = Normal; 1 = Mild; 2 = Moderate; 3 = Severe; 4 = Extremely severe. In this approach, the designation for Normal (0) is "not anxious (1)," while for other levels (1-4), it is "anxious (0)." And the text data output is 0 (Anxiety) or 1 (Not anxiety). Despite the fact that the SVM model was put together without taking database characteristics into particular consideration, it is still effective. The simplicity of employing binary classification (0/1) balances SVM's advantages, since it performs best in situations requiring distinct divisions between two groups.

Creating synthetic data, a series of tests were conducted on the machine learning models. The results were accurate as anticipated, as demonstrated in Table 2. The score count is calculated as  $(2 + 2 + 2 + 2 + 3 + 3 + 3) = 17 * 2 = 34$ , and  $34 > 7$ , indicating the presence of anxiety.

Determinants	Answer	Model Prediction
1 Dryness in the mouth	2	Anxiety (0)
2 Breathing difficulty	2	
3 Experience trembling	2	
4 Worried about panic and make a fool of myself	2	
5 Closer to panic	3	
6 Aware of the heart's action in the absence of physical exertion	3	
7 Felt scared without any valid reason	3	

Table 2: ML-GridsearchCV Model Testing

As observed, the score count calculated mathematically aligns with the prediction made by the ML-GridsearchCV model, indicating the model's satisfactory performance.

#### 4.2 Multimodal Late Fusion Phase

The three-layered sequential feed-forward fusion model performed very well on the test dataset prepared by the pre-multimodal phase models touching the accuracy of **92.38%**. Other metrics such as precision, recall and F1-score are **77%**, **100%** and **87%** respectively. (Figure 8).

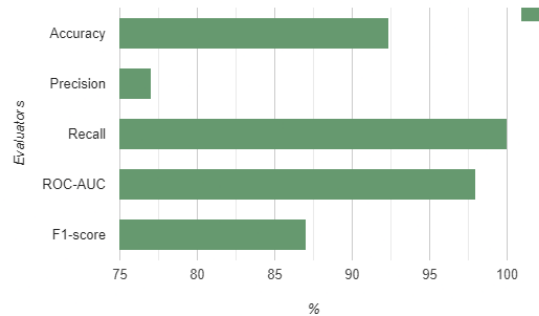


Figure 8: Fusion Model Evaluation

#### 4.3 Stratified K-fold cross-validation

Stratified k-fold cross-validation is beneficial in machine learning where maintaining class balance is crucial. It results in more accurate model evaluation and improved generalization. Although the fusion dataset is quite balanced, after data analysis, stratified K-fold validation is preferred. For k=10, ten-folds stratified cross-validation was used to

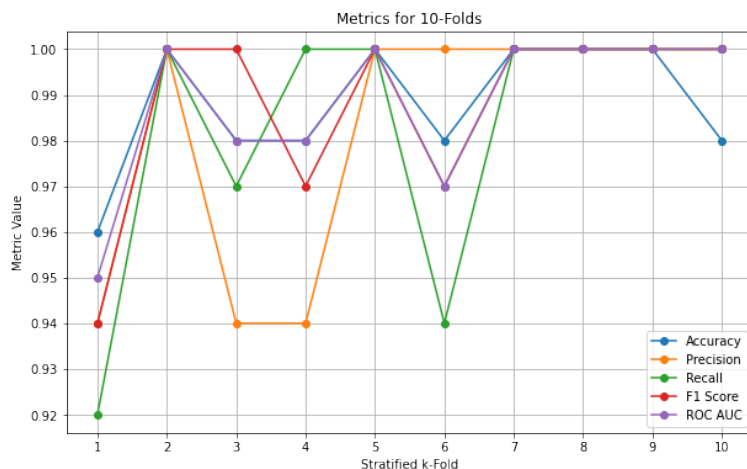


Figure 9: Stratified k-Fold CV

evaluate the fusion model of TI-Fusion, providing the performance measurement of the model, as shown in Figure 9.

The model’s validity was examined in the ten-folds cross-validation analysis using ten different folds of the dataset. The model’s robust ability to accurately classify binary outcomes (0 and 1) was demonstrated by the very consistent findings, which showed high accuracy (varying from 0.92 to 1.00), precision, recall, and ROC-AUC scores. These results imply that the model demonstrated high generalization and predictive power across different data subsets, proving its dependability and suitability for the task.

#### 4.4 Real-World Testing:

In order to anticipate the level of anxiety in actual situations, this study uses text and image data as inputs for the Late Fusion model. According to the probabilistic predictions of the model, scores below 0.5 are the indicators of anxiety, while scores over 0.5 are not. The binary classification system is used where 0 represents ”anxiety” and 1 represents ”no anxiety.”

##### Test-1:

Image Data Input:

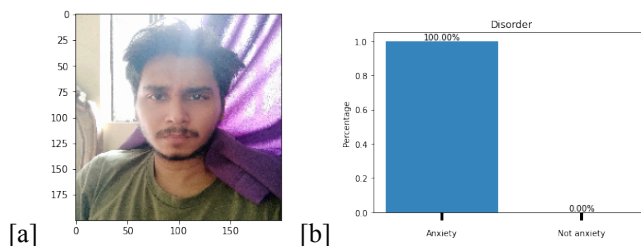


Figure 10: Person-1. (a) Worried face; (b) CNN prediction

Text Data Input:

Determinants	Answers
Dryness in the mouth	2
Breathing difficulty	1
Experienced trembling	1
Worried about panic and making a fool of myself	0
Closer to panic	2
Aware of the heart's action in the absence of physical exertion	2
Scared without any valid reason	0

Table 3: Person-1 Text Data

The score from the text data (Table 3) will be the sum of the responses (8 times 2 = 16), which is higher than the anxiety level's cutoff point of 7. Therefore, it should be 0 (anxiety).

Fusion Model Prediction:

Input	Expected Prediction	Model Prediction	Fusion Model Prediction
Image	Anxiety (0)	Anxiety (0)	$\approx 0.27010867$
Text	Anxiety (0)	Anxiety (0)	(0) Anxiety

Table 4: Person-1 Final Output

### Test-2:

Image Data Input:

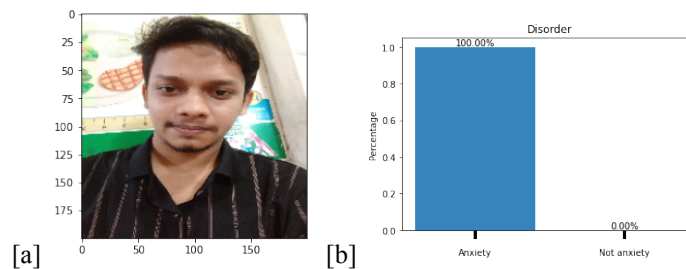


Figure 11: Person-2. (a) Neutral face; (b) CNN prediction

*Text Data Input:*

<b>Determinants</b>	<b>Answers</b>
Dryness in the mouth	1
Breathing difficulty	0
Experienced trembling	0
Worried about panic and making a fool of myself	0
Closer to panic	1
Aware of the heart's action in the absence of physical exertion	1
Scared without any valid reason	0

*Table 5: Person-2 Text Data*

The score from the text data (Table 5) will be the sum of the responses (3 times 2 = 6), which is lesser than the anxiety level's cutoff point of 7. Therefore, it should be 1 (not anxiety).

*Fusion Model Prediction:*

<b>Input</b>	<b>Expected Prediction</b>	<b>Model Prediction</b>	<b>Fusion Model Prediction</b>
Image	Not Anxiety (1)	Anxiety (0)	$\approx 0.5612543$
Text	Not Anxiety (1)	Not Anxiety (1)	(1) Not Anxiety

*Table 6: Person-2 Final Output*

In this case, the CNN model showed the person to be anxious, because it detected the face to be neutral, but based on the survey data, it predicted the person to be not anxious, which proves that the text data and the image data provides validation and prediction support to each other.

The Fusion Model Prediction tables (Table 4 and Table 6) depict the fusion prediction details. Consequently, these results offer a static perspective indicating the accurate prediction of anxiety disorders by the method.

#### **4.5 Comparison and Findings**

The evaluation encompassed an examination of the invasiveness as depicted in Table 7 and the accuracy as illustrated in Table 8.

TI-Fusion	Multimodal Approaches
<b>Text + Image</b>	EEG + EM [Guo, Fu, Pan, Zhang, and Hu 2020]
	ECG + EDA + RESP + PD [Jang, Choi, Kim, Yu, Jeon, and Byun 2023]
	EEG + PPG [Zheng, Wong, Leung, and Poon 2016]
	EMG + ECG + EDA [Vaz, Summavielle, Sebastião, and Ribeiro 2023]

Table 7: Invasiveness vs Non-invasiveness

Multimodal Approaches	Accuracy
CNN-LSTM [Xie, Wang, Lin, Luo, Chen, Xu, Liang, Liu, Wang, Luo, et al. 2022]	83.78%
K-GSCCA (ML) [Guo, Fu, Pan, Zhang, and Hu 2020]	87.47%
LSTM [Naderi, Soleimani, Rempel, Matwin, and Uher 2019]	74.57%
MLP-ML [Jang, Choi, Kim, Yu, Jeon, and Byun 2023]	75.61%
ML [Vaz, Summavielle, Sebastião, and Ribeiro 2023]	92.00%
<b>TI-Fusion (ML-CNN-NN)</b>	<b>92.38%</b>

Table 8: Late Fusion Approach Comparison

Table 7 indicates that the approach presented is **non-invasive** and provides individuals with anxiety a more convenient means to identify their anxiety. Moreover, Table 8 demonstrates that the proposed method, TI-Fusion, attains an overall accuracy of **92.38%**.

Furthermore, the DASS-21 dataset obtained from the paper [Priya, Garg, and Tigga 2020] underwent a thorough examination, revealing the significance of hyperparameter tuning, a crucial aspect for enhancing machine learning algorithms. To address this, hyperparameter tuning was performed within GridSearchCV, resulting in a substantial increase in algorithm accuracy, as demonstrated in Table 9. When incorporating the hyperparameters, GridSearchCV provided the flexibility to include additional parameters to identify the most relevant ones. Among these parameters, the best ones were applied to the DASS-21 dataset, as detailed in Table 10.

It's important to understand, however, that contextual factors such as varying lighting

Algorithms	Accuracy (%)		
	Before Hyperparameter Tuning	After Hyperparameter Tuning	The Unimodal method's [Priya, Garg, and Tigga 2020]
SVM	83.6	<b>100.0</b>	67.8
Naive Bayes	69.0	69.0	<b>73.3</b>
Decision Tree	71.0	<b>93.0</b>	73.3
Random Forest	68.0	<b>93.0</b>	71.4
KNN	66.0	<b>79.0</b>	69.8
XGBoost	66.3	<b>94.0</b>	Absent

Table 9: Comparing ML-GridsearchCV with Unimodal Approach [Priya, Garg, and Tigga 2020]

Algorithms	Best Hyperparameters
SVM	{'C': 10, 'gamma': 'scale', 'kernel': 'linear'}
Naive Bayes	None
Decision Tree	{'criterion': 'entropy', 'random_state': 0}
Random Forest	{'criterion': 'entropy', 'n_estimators': 100, 'random_state': 0}
KNN	{'metric': 'euclidean', 'n_neighbors': 5}
XGBoost	{'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 200}

Table 10: List of the Best Hyperparameters

conditions and intended facial expressions may introduce biases, leading to inaccurate positive or negative results. Therefore, maintaining a consistent operating environment and giving careful consideration to these factors is crucial for achieving accurate outcomes from tests. Additionally, it is essential to maintain the integrity of the test results to protect patient privacy.

#### 4.6 Justification

The following key-points validate the proposed TI-Fusion approach:

1. TI-Fusion is an enhancement of the unimodal method to provide a more reliable anxiety disorder assessment. The identical performance on text data in Table 9 before GridSearchCV usage supports this approach. The results of the experiments also demonstrate accuracy on multimodal data.

2. The Late Fusion Data was obtained using a unique algorithm for integrating pre-multimodal phase output data because the fusion involved using data from different modalities and sources. The evaluation of the Late Fusion Model (NN) began by comparing models from the pre-multimodal phase with existing unimodal models [Priya, Garg, and Tigga 2020], [Kim, Kim, Roy, and Jeong 2019] and [Eng, Ali, Cheah, and Chong 2019] based on the dataset. To confirm the effectiveness of the late fusion model, it was compared with traditional methods that use similar modalities (visual, textual) and processing techniques (MLP, ML, NN). This comparison is shown in Table 8.
3. Lastly, an innovative contribution to the discipline is the unconventional fusion algorithm created to combine text and visual data for the purpose of assessing anxiety. This algorithm introduces new mechanisms for efficiently combining different modalities of data, setting it apart from traditional approaches. Its uniqueness lies in its easy-to-deploy common-key merging technique. The empirical studies conducted with this algorithm have consistently delivered reliable results, demonstrating its effectiveness and establishing it as a genuinely novel technique in the field of anxiety disorder assessment.

## 5 Discussion and Future Work

Despite the promising outcomes, it is important to acknowledge certain potential limitations in the study. The dataset utilized may not fully encompass the entire spectrum of anxiety disorders, necessitating further research on more extensive and diverse datasets to validate the generalizability of the approach. Some of the limitations that commonly occur when dealing with mental health data include:

- Scarcity of mental health data: Having access to a larger dataset is crucial for effectively training and testing machine learning or neural network algorithms.
- The dataset used in this study might not fully capture the entire range of anxiety disorders, and exploring datasets with a more comprehensive representation is essential.

This study is primarily focuses on enhancing the detection of anxiety disorders by integrating multimodal data, which encompasses both facial expressions and survey data. The main objective is to improve the accuracy and reliability of these detection systems, thereby addressing a crucial aspect of mental health diagnosis and treatment. To further optimize the performance of the proposed multimodal framework, it is recommended to employ more advanced feature engineering techniques and delve into more complex deep learning architectures. The aspiration is that this research will make a significant contribution to the field of mental health assessment and care, with the ultimate aim of improving the well-being and quality of life for individuals who require these services.

## 6 Conclusion

The research introduces an innovative non-intrusive, multi-modal late fusion technique (TI-fusion). This technique utilizes a unique algorithm (Algorithm 1) that combines facial and survey data to produce a unified output. Specifically, the late fusion model

integrates the common features derived from the predicted data of the machine learning models (SVM, NB, DT, RF, KNN, XGB) and CNN. One of the main advantages of this system is its non-intrusive nature, which enhances its robustness and user-friendliness. Furthermore, it was observed that facial expressions played a significant role in improving the classification accuracy.

The effectiveness of this approach was tested in experiments involving individuals both with and without anxiety disorders. The results demonstrated that our method surpasses several unimodal and multimodal techniques in terms of performance, convenience, and reliability.

## References

- [Ahmed, Sultana, Ullas, Begom, Rahi, and Alam 2020] Ahmed, A., Sultana, R., Ullas, M. T. R., Begom, M., Rahi, M. M. I., Alam, M. A. (2020). A machine learning approach to detect depression and anxiety using supervised learning. In *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*. IEEE, pp. 1–6.
- [Asma, Mostafa, Akter, Mahmud, and Patwary 2022] Asma, A., Mostafa, D. N., Akter, K., Mahmud, M., Patwary, M. J. (2022). Fuzziness based semi-supervised deep learning for multimodal image classification. In *International Conference on Machine Intelligence and Emerging Technologies*. Springer, pp. 91–105.
- [Cao, Wu, Huang, Patwary, and Wang 2022] Cao, W., Wu, Y., Huang, C., Patwary, M. J., Wang, X. (2022). MFF: Multimodal feature fusion for zero-shot learning. *Neurocomputing*, 510, pp. 172–180.
- [Cai 2024] Cai, B., Liu, M., Li, B., & Li, J. (2024, July). Pre-sleep anxiety detection by multimodal signal. In *Third International Conference on Biomedical and Intelligent Systems (IC-BIS 2024)* (Vol. 13208, pp. 414–419). SPIE.
- [Eng, Ali, Cheah, and Chong 2019] Eng, S., Ali, H., Cheah, A., Chong, Y. (2019). Facial expression recognition in JAFFE and KDEF datasets using histogram of oriented gradients and support vector machine. In *IOP Conference series: materials science and engineering*, Vol. 705, No. 1. IOP Publishing, p. 012031.
- [Guo, Fu, Pan, Zhang, and Hu 2020] Guo, Z., Fu, E., Pan, J., Zhang, X., Hu, B. (2020). Anxiety detection with nonlinear group correlation fusion of electroencephalogram and eye movement. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, pp. 2596–2602.
- [Haas and Hüllermeier 2023] Haas, S., Hüllermeier, E. (2023). Rectifying bias in ordinal observational data using unimodal label smoothing. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 3–18.
- [Hansson, Lexén, and Holmén 2017] Hansson, L., Lexén, A., Holmén, J. (2017). The effectiveness of narrative enhancement and cognitive therapy: a randomized controlled study of a self-stigma intervention. *Social psychiatry and psychiatric epidemiology*, 52, pp. 1415–1423.
- [Jang, Choi, Kim, Yu, Jeon, and Byun 2023] Jang, E. H., Choi, K. W., Kim, A. Y., Yu, H. Y., Jeon, H. J., Byun, S. (2023). Automated detection of panic disorder based on multimodal physiological signals using machine learning. *ETRI Journal*, 45(1), pp. 105–118.
- [Kim, Kim, Roy, and Jeong 2019] Kim, J.-H., Kim, B.-G., Roy, P. P., Jeong, D.-M. (2019). Efficient facial expression recognition algorithm based on hierarchical deep neural network structure. *IEEE Access*, 7, pp. 41 273–41 285.
- [Lucey, Cohn, Kanade, Saragih, Ambadar, and Matthews 2010] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I. (2010). The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, pp. 94–101.

- [Lensvelt-Mulders 2012] Lensvelt-Mulders, G. (2012). Surveying sensitive topics. In *International Handbook of Survey Methodology*. Routledge, pp. 461–478.
- [Lai and Li 2023] Lai, S., & Li, Z. (2023, December). Detection of potential anxiety in social media based on multimodal fusion with deep learning methods. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 560-566). IEEE.
- [Naderi, Soleimani, Rempel, Matwin, and Uher 2019] Naderi, H., Soleimani, B. H., Rempel, S., Matwin, S., Uher, R. (2019). Multimodal deep learning for mental disorders prediction from audio speech samples. *ArXiv*, abs/1909.01067. Available at: <https://api.semanticscholar.org/CorpusID:202541260>
- [Niaz 2022] Niaz, N. U., Shahariar, K. N., & Patwary, M. J. (2022, March). Class imbalance problems in machine learning: A review of methods and future challenges. In *Proceedings of the 2nd International Conference on Computing Advancements* (pp. 485-490).
- [Osman, Tabassum, Patwary, Imteaj, Alam, Bhuiyan, and Miraz 2022] Osman, A. B., Tabassum, F., Patwary, M. J., Imteaj, A., Alam, T., Bhuiyan, M. A. S., Miraz, M. H. (2022). Examining mental disorder/psychological chaos through various ML and DL techniques: A critical review. *Annals of Emerging Technologies in Computing (AETiC)*, pp. 61–71.
- [Priya, Garg, and Tigga 2020] Priya, A., Garg, S., Tigga, N. P. (2020). Predicting anxiety, depression and stress in modern life using machine learning algorithms. *Procedia Computer Science*, 167, pp. 1258–1267.
- [Patwary, Cao, Wang, and Haque 2022] Patwary, M. J., Cao, W., Wang, X.-Z., Haque, M. A. (2022). Fuzziness based semi-supervised multimodal learning for patient's activity recognition using RGBDT videos. *Applied Soft Computing*, 120, p. 108655.
- [Rubeis and Steger 2019] Rubeis, G., Steger, F. (2019). A burden from birth? Non-invasive prenatal testing and the stigmatization of people with disabilities. *Bioethics*, 33(1), pp. 91–97.
- [Scherer 2009] Scherer, K. R. (2009). Emotions are emergent processes: they require a dynamic computational architecture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535), pp. 3459–3474.
- [Surcinelli, Codispoti, Montebanocci, Rossi, and Baldaro 2006] Surcinelli, P., Codispoti, M., Montebanocci, O., Rossi, N., Baldaro, B. (2006). Facial emotion recognition in trait anxiety. *Journal of anxiety disorders*, 20(1), pp. 110–117.
- [Vaz, Summavielle, Sebastião, and Ribeiro 2023] Vaz, M., Summavielle, T., Sebastião, R., Ribeiro, R. P. (2023). Multimodal classification of anxiety based on physiological signals. *Applied Sciences*, 13(11), p. 6368.
- [Xie, Wang, Lin, Luo, Chen, Xu, Liang, Liu, Wang, Luo, et al. 2022] Xie, W., Wang, C., Lin, Z., Luo, X., Chen, W., Xu, M., Liang, L., Liu, X., Wang, Y., Luo, H., et al. (2022). Multimodal fusion diagnosis of depression and anxiety based on CNN-LSTM model. *Computerized Medical Imaging and Graphics*, 102, p. 102128.
- [Zheng, Wong, Leung, and Poon 2016] Zheng, Y., Wong, T. C., Leung, B. H., Poon, C. C. (2016). Unobtrusive and multimodal wearable sensing to quantify anxiety. *IEEE Sensors Journal*, 16(10), pp. 3689–3696.