A Multimodal system for detecting life threatening emotions in social media using AI

Abstract

Catastrophic emotions like anxiety, fear, anger, and suicidal phrases are more shortly experienced as a result of social and psychological issues occurring after COVID-19. With the tremendous amount of content in various formats on social media, we present a multimodal AI system that detects indicators of early stages of the emotional distress based on speech, vision and text analysis. It uses the latest deep learning models, spectrogram-based LSTMs for speech, facial recognition for vision and sentiment analysis for text to improve detection and timely report generation. As a result of filling the gaps left by other single-modal approaches, our framework can provide timely interventions for these patients, help mental health professionals, and raise the quality of patients' treatment. Its scale and its ethical structure provides a strong framework for delivering intelligent analytics into healthcare, thereby gaining lives of those in danger and allowing for mental health crisis intervention. The findings of this study also show how AI can be used to revolutionize other mental health issues across the world.

Chapter 1: Introduction

Life threatening emotions like anxiety, fear, anger and suicidal thoughts are increasingly becoming common in today's world. The reasons for this phenomenon could be several factors, for e.g., rapid evolution of technology in past few decades and its impact on the inter-personal relationships with family and the society. Notwithstanding the reasons, these emotions have dangerous effects on people. But there is still hope. The social media gives a unique platform for early detection of these emotions so as to possibly help the affected people and give them the required medical assistance. In the past, there have efforts to detect life threatening emotions using photos, audio, video or text datasets.

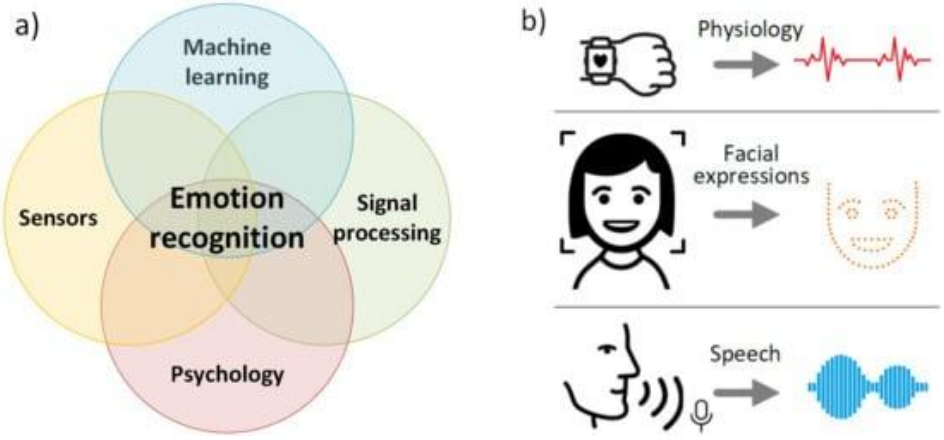

Figure 1: An example of multimodal system that we are intending to use (**Source inspired from: MDPI**)

As seen in the above figure, multimodal deep learning in netizen sentiment recognition shows the relevance of mixing text and visual data to improve classification performance, strengthening public security early warning systems (**Nan & Yao, 2024**). Bayesian networks and cross-attention mechanisms in models like the MCABN employ visual and verbal content's consistency and complementarity to improve emotion recognition (**Wang et al., 2023**). These advances show that multimodal AI systems

can improve emotion recognition and enable early intervention and therapy for at-risk individuals in mental health and public safety (**Ezerceli & Dehkharghani, 2024**).

1.1 Research Motivation

Early Detection and Intervention: The proposed approach can be included into mental health care systems to offer early indicators of suicidal intent. This power may save lives by facilitating prompt interventions from mental health professionals who can observe social media for indicators of distress.

Assistance for Mental Health Practitioners: Employing this model enables psychologists and mental health practitioners to improve their capacity to recognize individuals at risk. The model's efficacy in identifying depression and suicidal ideation can provide more effective treatment approaches, hence enhancing patient outcomes.

1.2 Research Aim

This project aims to utilize photos, text, speech data in social media from publicly accessible audio data and social media content to derive comprehensive insights into mental health indicators. This comprehension will facilitate early diagnosis and intervention, sufficient provision of psychological information, appropriate support measures, advancement of mental health care. The study seeks to establish an extensive framework for identifying threatening emotions via social media analysis, employing sophisticated machine learning methodologies to enhance early detection and intervention efforts in mental health treatment.

1.3 Research Objective

The primary objective is to develop and implement a cutting-edge mental health analytics system utilizing speech data from public, social media platforms, or various datasets to analyse the intent of spoken language. This technology can identify early indicators of life-threatening emotions using machine learning and predictive analysis. The strategic components of this initiative encompass data collection, algorithm development, forecasting and assessment, solution execution.

1.4 Research Questions

This study will explore the following research questions

RQ1: How can the predictive analytics models be optimised to provide reliable early signs of life-threatening emotions with the retrieved data?

RQ2: Which implementation challenges may occur because of extending the developed mental health analytics solution into typical healthcare systems or platforms and how can those challenges be resolved?

Chapter 2: Literature Survey

In order to identify emotional states that may suggest severe mental health difficulties, such as suicidal ideation, the creation of a multimodal system for detecting life-threatening emotions in social media using artificial intelligence involve the integration of a variety of data sources and complex machine learning algorithms. This method utilizes the extensive user-generated content accessible on social media platforms to provide timely interventions and maybe save lives. The subsequent sections will outline the essential components and methodologies pertinent to this system.

2.1 Multimodal Emotion Recognition

The study by **Omar et al. (2023)** introduced MM-EMOR, a multimodal emotion identification system that proficiently identifies and analyzes emotions from audio and textual data in social media environments. MM-EMOR exhibits substantial enhancements in emotion detection accuracy by employing Mel spectrogram and Chromagram features for audio processing via a Mobilenet CNN, in conjunction with an attention-based Roberta model for text analysis. The system's performance is validated across three datasets, with accuracy improvements of up to 18%, demonstrating its potential for identifying important emotional states that may indicate life-threatening scenarios.

Madhura et al. (2024) emphasized the prospect of multimodal emotion identification systems, which combine textual, visual, and audio input to improve the precision of emotion classification. The system utilizes advanced deep learning models, such as BiLSTM for text, CNNs for images, and LSTMs for audio, to proficiently scan social media content for the detection of life-threatening emotions. This method enhances the comprehension of emotional states and provides substantial applications in mental health diagnostics and human-computer interaction, rendering it an essential instrument for detecting vital emotional signals in digital interactions.

Wang et al. (2024) concentrated on multimodal emotion recognition (MER) utilizing audio and visual data, highlighting the significance of precisely comprehending emotional states for applications such as human-computer interaction. Although it does not explicitly focus on identifying life-threatening emotions in social media, it introduces an innovative method employing Transformers and a multi-attention mechanism to improve the accuracy of emotion recognition. This model may be applicable for recognizing significant emotional states in social media environments; however, the research does not specifically investigate this use.

Nan & Yao (2024) investigated the utilization of multimodal deep learning for the recognition of netizen emotions through the integration of textual and visual data. The study employs BERT for textual analysis and VGG-16 for image processing, developing a fusion model augmented by a multi-head attention mechanism. This method enhances classification efficacy, attaining an accuracy of 0.73 for the multimodal fusion model, surpassing that of the individual modalities. This methodology can be pivotal in identifying life-threatening emotions on social media, providing significant insights for public safety and emotional assessment.

2.2 AI for Detecting Life-Threatening Emotions

Malhotra and Jindal (2020) present a multimodal deep learning architecture that examines users' social media content, encompassing text, photos, and videos, to identify suicidal tendencies. It employs advanced methodologies such as VGG-16 for image feature extraction, word2vec for text processing, and Faster R-CNN for video frame analysis. By integrating these multimodal representations, the system produces a weighted average score for classification, facilitating real-time identification of depression and suicidal behavior, thereby overcoming the shortcomings of current systems that depend on manual reporting.

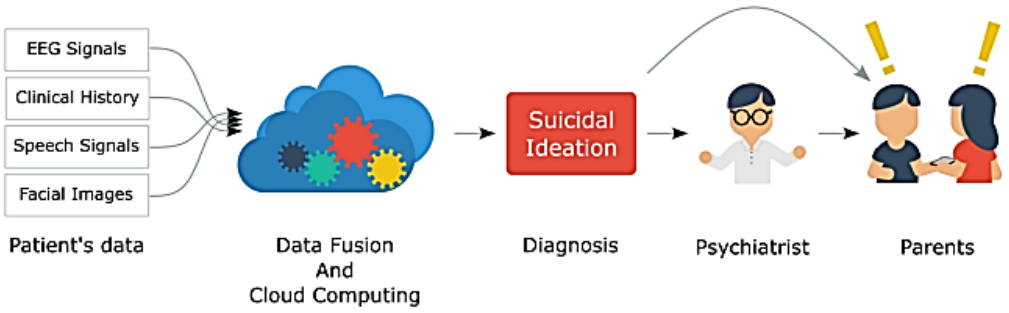

Figure 2: Typical scheme of detection of Suicidal ideation in Individuals (**Barua et. al. (2024)**)

Diana et al. (2020) propose a multi-modal deep learning methodology for identifying suicidal tendencies through the analysis of textual, visual, relational, and behavioural data from social media users. It employs techniques including bag of words, n-grams, and image analysis to construct prediction models. The research indicates that integrating both modalities markedly enhances detection accuracy relative to the exclusive use of text-based models. The methodology was assessed using a dataset of 252 users, demonstrating the efficacy of this comprehensive approach in identifying at-risk individuals.

Varsha et al. (2022) propose a multi-modal methodology for identifying suicidal tendencies through three complementing techniques: Facial Gesture Recognition, Voice Pattern Recognition, and Text Pattern Recognition. These methodologies are included into an Android mobile application, facilitating a thorough evaluation of an individual's mental condition. This strategy seeks to evaluate diverse data sources to notify family and friends of potential suicidal intent, thus facilitating prompt intervention possibilities. The research underscores the significance of integrating various modalities for efficient detection.

Moumita et al. (2022) suggest a multi-modal methodology for detecting suicidal intentions through the analysis of online social media content. It delineates six feature categories that comprise clinical suicidal symptoms and online behaviours. A meticulously annotated dataset from Reddit and Twitter was developed to train the model. The research attained an accuracy of 87% with a Logistic regression classifier, illustrating that proficient feature selection and amalgamation markedly improve performance in detecting suicidal ideation on social media platforms.

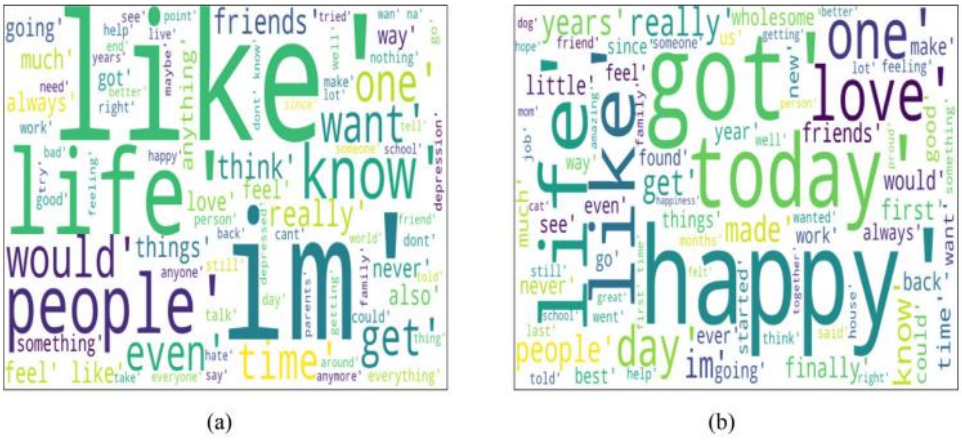

Figure 3: Word Cloud (a) Suicidal data (b) Non-Suicidal data (**Varsha et al. (2022)**)

Joshua's(2023) This clinical cross sectional <u>study</u> evaluates a multimodal dialog system (MDS) named Tina, which utilizes speech, language, and facial movement biomarkers to indicate a client's suicidal risk. MFI values depicted that participants described as at risk for suicide had a greater percentage of pause time during interchanges. The maximum ACKSU score was achieved via decision fusion of all the models, with the AUC measurement equal to 0.76. This implies that, the inclusion of various modes enhance the assessment of suicidal predispositions in clinical sample.

A study carried out by **Toliya & Nagarathna (2024)** presented an application of mDL for early risk prediction of suicide ideation from online social content. News articles are processed by modern methods like deep neural networks and attended relation networks to properly identify suicidal ideologies in user-generated content. Combining content analysis and feature engineering enhance measures of extracting lethal sentiments with the aim of enhanced early interventions for suicide prevention particularly among vulnerable populations including the youth.

Raja & Nagarajan (2024) present a paper where they discuss how postmodern AI methods, especially an LSTM-Attention-RNN model, can be used to approach social media posts to identify signs of suicidal tendencies. The emphasis that is laid on the analysis of texts could be complemented with the attempt to include other types of data which often accompany online emotive responses, e.g., a photo or a video. The model obtained the overall cross-validated accuracy of 93.70% and the F1-score of 95.80%, and the findings affirm its capability in identifying the lethal sentiments, which points for further utilization of proposed solution in prevention of suicides.

2.3 Early Detection and Intervention

AI systems may accurately identify early indicators of mental health crises, like depressed or anxiety episodes, facilitating prompt interventions. A study by **Mansoor & Ansari (2024)** created a multi-modal deep learning model that combines natural language processing and temporal analysis to identify early indicators of mental health crises in social media posts. This AI-driven approach shown significant precision in detecting critical emotional states, including suicidal thoughts and depressive episodes, with a mean advance notice of almost 7 days prior to expert human recognition. Essential digital indicators encompassed linguistic patterns, behavioural modifications, and temporal trends, demonstrating the model's efficacy across various languages and platforms, while underscoring the necessity for ethical considerations in its implementation.

Biswas (2023) examines the detection of suicidal ideation on social media with an AI model, as opposed to a multimodal approach for life-threatening emotions. It utilizes text analysis and GloVe for word embedding, attaining a recall of 0.93 and a precision of 0.94 in detecting suicidal content. Although it tackles the vital concern of early identification of suicide attempts, it fails to examine other perilous emotions or multimodal strategies. Consequently, it is confined to the realm of detecting suicidal ideations.

In a study by **Saraf et al. (2024)**, a hybrid deep learning methodology is utilized, integrating Convolutional Neural Network (CNN) and Bidirectional Gated Recurrent Unit (Bi-GRU) models to identify suicidal tendencies from social media data. This multi-modal approach employs Natural Language Processing (NLP) and Sentiment Analysis, examining multi-source datasets from Reddit and Twitter. The BiGRU-CNN model exhibited exceptional performance, with accuracies of 93.07% and 92.47% on the two datasets, proficiently detecting emotional distress signals associated with suicidal ideation.

2.4 Summary of Literature

Table 1: Research Summary for all the research papers analysed

Paper Name	Author Names	Dataset Used	Algorithms Used	Research Results
Leveraging Online Social Content for Early Detection of Suicidal Ideation: A Multi-Modal Deep Learning Approach	Nikita Paras Toliya, N. Nagarathna	User-generated text from online social content	Deep neural networks and attentive relation networks	By content-analysing suicide-related literature, the study sets a bar for binary classification of suicidal thoughts and provides significant insights into individuals' mental states.. The research uses advanced methods like deep neural networks and attentive relation networks to improve suicidal ideation early detection tools, promoting mental health and suicide prevention among vulnerable populations, particularly youths.
A multimodal dialog approach to mental state characterization in clinically depressed, anxious, and suicidal populations	Joshua Cohen, Vanessa Richter, Michael Neumann, David P. Black, Allie Haq, Jennifer Wright-Berryman, V. Ramanarayanan	The study has 73 sessions with 68 online health registry participants. Conventional screening equipment detected depression (20.6%), anxiety (28.8%), and suicide risk (35.6%) during these sessions.	Machine learning models using extracted features from speech, language, and facial movement data to classify mental states related to depression, anxiety, and suicide risk	Participants with depression and suicide risk had a larger speech pause time, whereas those with anxiety had lower lip movement than healthy controls. Speech characteristics were best for depression (AUC = 0.64), face features for anxiety (AUC = 0.57), and text features for suicide risk (AUC = 0.65). Decision fusion of all models yielded the best suicide risk identification AUC of 0.76. Participants found the multimodal dialog system comfortable, proving its suitability for remote patient monitoring in clinical populations with depression, anxiety, and suicidal tendencies.
Suicide ideation detection from online social media: A multi-modal feature based technique	Moumita Chatterjee, Piyush Kumar, Poulomi Samanta, Dhrubasish Sarkar	Reddit and Twitter posts with suicidal ideas were used to construct a well-labeled suicide thinking dataset. The dataset included six feature groups that included clinical suicidal symptoms and social media activities to detect suicidal thoughts.	Logistic regression classifier	The research achieved an accuracy of 87% in detecting suicidal thoughts on social media using a Logistic regression classifier, which outperformed other baseline models.
Detection of Suicidal Ideation on Social Media: Multimodal, Relational, and Behavioral Analysis.	Diana Ramírez-Cifuentes, Ana Freire, Ricardo Baeza-Yates, Joaquim Puntí, Pilar Medina-Bravo, Diego Velázquez, Josep M. Gonfaus, Jordi Gonzàlez	The study included 252 users' social media data, including tweets, connections, and behavior. Clinicians annotated this dataset to identify suicidal users. A classifier was trained using 90,000 Instagram photographs for training and 60,000 for validation for image analysis.	The study tested random forest, multilayer perceptron, logistic regression, and support vector machines. Each method was tested using 10-fold cross-validation to select the best categorization algorithm. A convolutional neural network (CNN) was used to embed models, especially for picture data, along with classical classifiers..	The study found substantial textual and behavioral differences between suicidal ideation risk group users and both control groups. In particular, focused control users had 578.5 friends and 16 words in their tweets, compared to 372.0 and 13 words, respectively. Combining textual, visual, relational, and behavioral data outperformed models using each modality alone, improving accuracy by up to 8% when identifying risky users from both types of control users. This suggests that the investigated features are essential for identifying suicidal users.

A Deep Learning-Based Sentiment Classification Approach for Detecting Suicidal Ideation on Social Media Posts	Pradeep Kumar, Dilip Singh Sisodia, Rahul Shrivastava	The dataset used in the study is the Suicide and Depression Detection dataset, which is available on Kaggle and contains posts from the social media platform Reddit.	Based on social media posts, the study classifies individuals' suicidal and non-suicidal intent using CNN and LSTM models.	The proposed LSTM model achieved the highest accuracy of 93% in classifying suicidal and non-suicidal intent among users, demonstrating superior performance compared to baseline models from other authors.

2.5 Research Gap

Despite significant advancements in multimodal emotion recognition and AI-driven detection systems for identifying life-threatening emotions on social media, key gaps remain. Current methodologies often focus on specific modalities, such as text or audio, with limited integration of diverse data types like visual, relational, or behavioral cues. Few studies address the scalability and real-time application of these models, particularly across multiple languages and cultural contexts. Additionally, ethical considerations, including user privacy, data bias, and the potential misuse of such systems, are insufficiently explored.

2.6 Research Outcome

To help address personal mental health concerns aggravated by COVID-19, we outline an early warning multimodal AI system that incorporates speech, vision and text analytics. This system integrate deep learning algorithms such as speech emotional recognition, face emotions and textual sentiment analysis in real-time to detect ongoing emotional crisis and suicide risk. Accepted by itself, data from one modality may not be sufficient, but integrating them helps to pick up the signs of a deteriorating situation and act quickly. The system has the potential to close the gap and provide a scalable and ethical solution for diagnosing, saving lives, and helping those in need in a post-COVID world that has seen the need for proper mental health help.

Chapter 3: Methodology

3.1 Data Acquisition

3.1.1 Image to Emotion - FER - 2013

The FER-2013 (https://www.kaggle.com/datasets/msambare/fer2013) dataset categorizes faces by emotion, using 48x48 pixel grayscale images. The FER-2013 dataset, includes 35,887 micro expressions and classifies each face according to the emotion revealed in the facial expression into one of the given seven categories (Angry=0, Disgust=1, Fear=2, Happy=3, Sad=4, Surprise=5, Neutral=6). The faces have been automatically aligned to ensure they are about centered and occupy a similar spatial proportion in each image.

3.1.2 Speech to emotion – Toronto emotional speech set (TESS)

The speech dataset used is the TESS dataset (https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess). It was modelled on the Northwestern University Auditory Test No. 6. A set of 200 target words were spoken in a carrier phrase. The carrier phrase was "Say the word _," spoken by two female actors (26 and 64 years old) conveying each of seven emotions (anger, disgust, fear, pleasure, pleasant surprise, sorrow, and neutral). There are 2800 audio data files (.WAV format).

3.1.3 Text to emotion (WASSA-2017)

Twitter possesses a substantial and varied user demographic, resulting in an abundance of textual content that includes non-standard language elements such as emojis, hash tagged terms (#luvumom) and emoticons, artistically altered spellings (happee). Tweets are frequently utilized to express emotions, opinions, and positions (Saif et al. et al., 2017). Consequently, the automatic detection of emotion intensities in tweets is particularly advantageous for applications including monitoring brand and product perception, gauging support for causes and policies, assessing public health and well-being, and managing disasters or crises. The tweets are annotated for ***Anger, fear, joy and sadness***.

3.2 Deep Learning Models for Vision based emotion detection

3.2.1 VGG19

VGG-19 (Fig. 3) is a convolutional neural network consisting of 19 weight layers, which include 16 convolutional layers and 3 fully connected layers. The architecture sticks to a simple and repeating structure, facilitating comprehension and implementation.

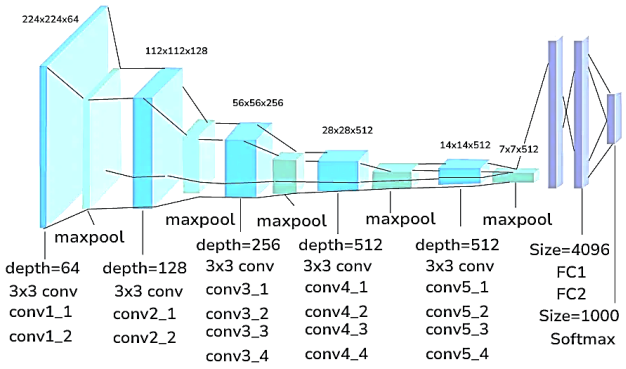

Figure 4: VGG-19 Architecture (**Source: Geeks for Geeks**)

3.2.2 Densenet121

DenseNet's CNN (Fig. 5) connection pattern solved feature reuse, vanishing gradients, and parameter efficiency problems, revolutionizing computer vision. In contrast to standard CNN architectures, DenseNet directly connects all layers within a block. Each layer receives feature maps from all previous levels due to its tight connectedness, promoting information flow throughout the network. Densenet121 having 121 layers is known for balanced computational efficiency and accuracy. It is suitable for intermediate computational jobs.

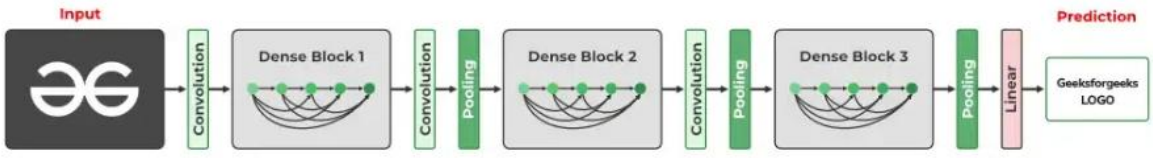

Figure 5: Densenet architecture (**Source: Geeks for Geeks**)

3.2.3 EfficientNet

EfficientNet uses a coefficient to equally scale depth, width and resolution. Real-time applications and resource-constrained contexts like mobile devices and edge computing platforms benefit from

EfficientNet's efficiency. Its tiny but powerful architecture optimizes performance without computational load.

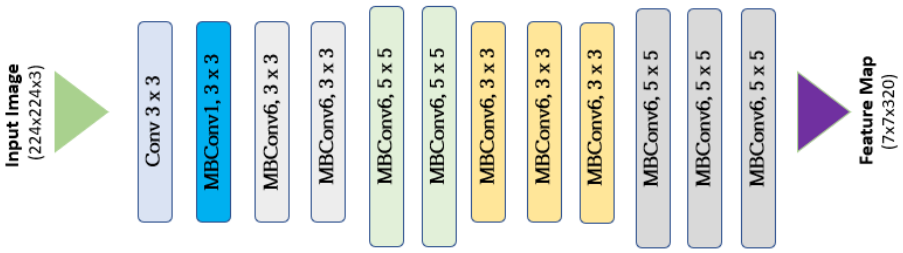

Figure 6: EfficientNet architecture (**Source: Wisdom ML**)

3.2.4 MobileNet

This deep learning architecture effectively processes and classifies photos on resource-constrained smartphones and embedded systems. MobileNet, developed by Andrew G. Howard et al. in 2017, uses depth-wise separable convolutions to simplify convolution processing. MobileNet factors a typical 3x3 convolution to a depth-wise convolution, which applies a single convolutional filter per input channel, and a point-wise convolution, which combines output channels using 1x1 convolutions. This separation of spatial and channel-wise information makes MobileNet lightweight and parameter-efficient without losing accuracy.

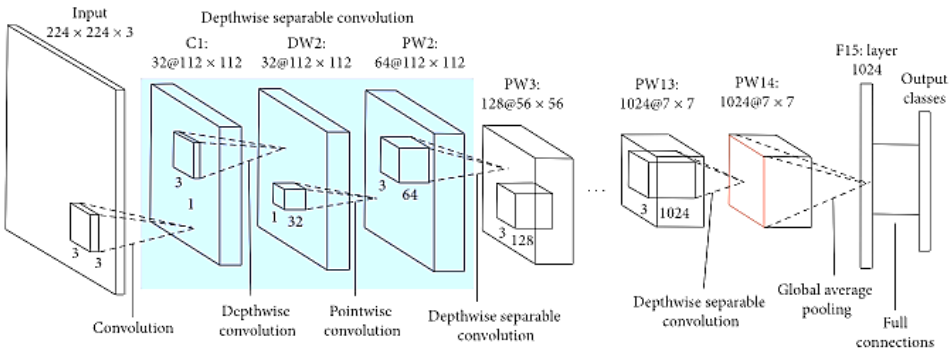

Figure 7: Mobilenet architecture (**Shoaib et. al., 2023**)

3.2.5 ResNet50

ResNet50's main innovation is residual blocks, which let the network learn residual functions instead of direct mappings. Multiple convolutional layers and skip connections propagate input from one layer to the output of a later layer in each residual block. ResNet50 has 50 layers with 1x1, 3x3, and 1x1 convolutions, global average pooling, and a fully connected classification layer. ResNet50's skip connections allow deeper network training without vanishing gradients, improving picture recognition accuracy.

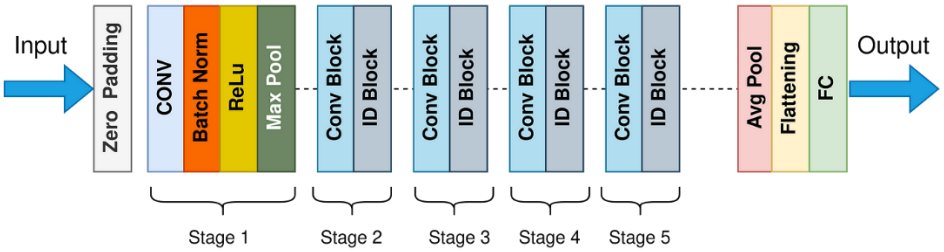

Figure 8: RESNET50 architecture (**Source: Analytics Vidhya**)

3.3 Speech to emotion – Feature Extraction

Fast Fourier Transform: The Fast Fourier Transform (FFT) is an essential analytical method in audio and acoustics measurement. The Fast Fourier Transform converts a signal into its spectral components to elucidate its frequency content.

Chroma Features: Chroma features, also known as chroma vectors or chromagrams, represent the twelve pitch classes (semitones) of the musical octave regardless of octave. Harmonic and melodic features of audio signals are essential for chord recognition, key detection, and harmonic analysis. Chroma characteristics are based on how humans perceive pitch, where notes separated by octaves sound similar. This makes the color representation resistant to timbre and instrumentation variations. Audio signals' short-time Fourier transform (STFT) is used to compute chroma characteristics. The spectral components are then mapped onto the twelve chroma bins representing the musical octave's twelve semitones. To highlight frequency perception, this method generally uses a logarithmic transformation. The two-dimensional chromagram shows time and the twelve chroma bins.

Mel features: Mel-Frequency Cepstral Coefficients (MFCCs) are frequently employed in audio signal processing and speech recognition (**Tiwari, 2010**). They show the short-term power spectrum of a sound stream, including timbral and phonetic properties. MFCCs are developed from the Mel-scale, a perceptual scale of tones listeners perceives as equal distance apart. This scale mimics the human ear's response to different frequencies, making MFCCs useful for speech and audio processing.

Spectrogram: A spectrogram shows a signal's frequency spectrum across time (**Wyse, L., 2017**). It captures temporal and frequency information to show how an audio signal's spectral content evolves over time. A spectrogram is created by dividing the audio stream into brief, overlapping frames and Fourier transforming each frame. The magnitude spectra are then plotted as a function of time, with time on the horizontal axis, frequency on the vertical, and each frequency component's amplitude represented by the plot's intensity or color. Depending on frequency axis scaling, spectrograms are linear, logarithmic, or Mel. Spectrograms indicate harmonic content, formant structures, and transient occurrences in audio sources. Speech analysis uses them to visualize phonemes, pitch, and prosody. Spectrograms identify musical notes, chords, and rhythms.

3.4 Text to Emotions – Features Extraction

Preprocessing: The links and special characters in the dataset are removed and using regular expression (regex statements). The citations, tickers, punctuations., quotes, RTweets, Line break, tab return, blanks, white space, hashtags, emojis, empty rows, stop-words are also removed. Finally, all the text is converted to lower case.

NLP Feature Extraction- CounterVectorizer: CountVectorizer is a class in scikit-learn that converts a set of text documents into a matrix representing word or token counts. This class has several parameters that facilitate text preparation tasks, including elimination of stop words, thresholds of word count (both maximum and minimum), vocabulary limitations, n-gram generation, and additional functionalities.

3.5 Machine Learning Models

Random Forest Classifier: A random forest is an ensemble model that constructs multiple decision tree classifiers on different sub-samples of the dataset and utilizes the process of averaging to enhance predicted accuracy and reduce over-fitting.

Gradient Boosting Classifier: Gradient Boosting iteratively chooses a function that tends toward a weak hypothesis or negative gradient to minimize a loss function. The loss function measures the accuracy of model predictions using available data. The outcome depends on the topic. A weak learner can categorize data but has a high error rate. These are usually decision trees. An additive model is a mathematical model that predicts or explains an event by combining multiple components or elements. To develop a complete model, components representing different factors are added.

Naïve bayes: Bayes' Theorem enables Naive Bayes' statistical classification. This is one of most simple supervised learning algorithms. Naive Bayes classifiers are fast, accurate, and reliable. When applied on large datasets, Naive Bayes classifiers are accurate and efficient.

Logistic Regression: A logistic regression model classifies binary data. It predicts an instance's class membership using the sigmoid function to ensure the output between 0 and 1. The model computes a linear combination of input features and maps it using the sigmoid and optimizes its coefficients using gradient descent algorithm to minimize log loss. These coefficients determine the class decision boundary. Logistic Regression is frequently employed in machine learning for binary outcome issues due to its simplicity, interpretability, and adaptability across domains. Overfitting can be avoided via regularization.

Support Vector Machines: Support Vector Machine (SVM) is a supervised machine learning method utilized for classification and regression applications. The fundamental principle of SVM is to determine the best boundary (or hyperplane) that separates the data into various categories.The boundary is selected to maximize the margin, defined as the distance between the boundary and the nearest data points from each class. The nearest data points are referred to as support vectors. Support Vector Machines (SVMs) can be employed for non-linear classification through a method known as the kernel trick. The kernel trick transforms the input data into a higher-dimensional space, rendering the data linearly separable. Prevalent kernels comprise the radial basis function (RBF) and the polynomial kernel.

LGBN Classifier: Light Gradient Boosting Machine Classifier (LGBN) uses decision tree algorithms for machine learning tasks like ranking and categorization. LGBMClassifier provides a novel Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) approach for huge datasets that is virtually as fast as the original algorithm but uses less memory. Classic gradient boosting approaches train the model with all the data, which takes time for large datasets. GOSS of LightGBM keeps all instances with big gradients and randomly samples a subset with tiny gradients. Due to their difficult fit, instances with big gradients provide additional information. GOSS uses a constant multiplier on data instances with minor gradients to avoid sampling-induced information loss.

XGBoost Classifier: XGBoost is a distributed, high-performance gradient boosting system utilized in machine learning for various applications. It provides parallel tree boosting and is the most widely utilized machine learning library for regression, classification, and ranking. It encompasses supervised machine learning, decision trees, ensemble learning, and gradient boosting.

LSTM: LSTM recurrent neural networks (RNNs) use control signals to retain short-term and long-term memory. LSTMs may learn long-term dependencies in time series and sequential data by retaining and processing information over multiple time steps. Selective memory loss or acquisition is feasible with LSTMs, rendering them less vulnerable to the vanishing gradient issue typical of standard RNNs.

3.6 Evaluation Method

The models are evaluated using standard classification criteria, offering insights into many facets of performance. These measurements encompass overall accuracy and class-specific indicators such as precision, recall, and F1-score. Moreover, confusion matrices are employed to illustrate the models' efficacy across several emotion categories, providing an in-depth analysis of categorization strengths and shortcomings. This comprehensive methodology offers a solid framework for addressing emotion classification challenges with multimodal data. It includes the complete process from raw data to model evaluation, providing a thorough methodology that may be modified for analogous classification issues in the field of speech and audio analysis.

Chapter 4: Implementation

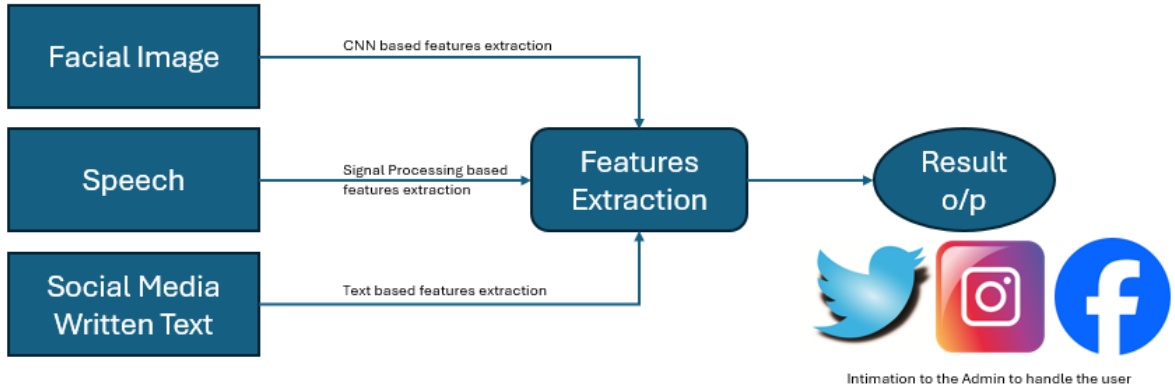

Figure 9: Implementation flow of the multi modal system

4.1 Image to Emotion

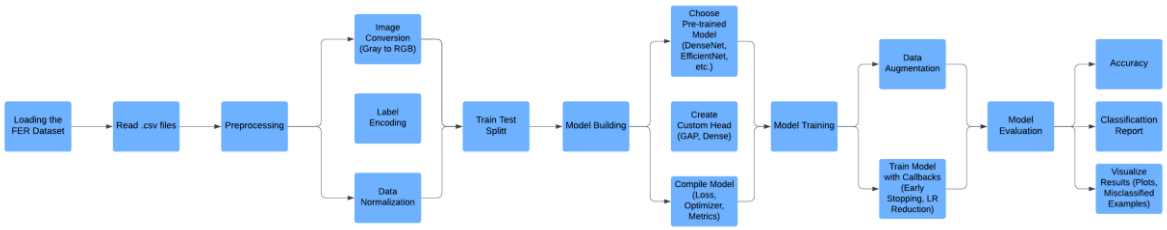

Figure 10: Image based emotion detection

4.1.1 Data Loading

Data Loading and Conversion: The fer2013 dataset was used, which includes pixel of images. Each image is represented by a string of pixel values. The values were split, reshaped them into 48x48 arrays and converted them into numpy arrays for easier manipulation. To ensure the images matched the input requirements of my pre-trained models, the grayscale images were converted to RGB format.

Figure 11: FER images to check the emotion detection

Label Encoding and Splitting: The emotions from the dataset were encoded into numerical labels and converted them into categorical format using one-hot encoding. Then the data underwent splitting into training and validation sets using a stratified approach, maintaining the original distribution of emotion labels across both sets.

Normalization: The image pixels had to be normalized to the range [0, 1]. This normalization step helped improve the accuracy and performance of the neural networks during training by ensuring that all input values were on a similar scale.

4.1.2 Model Building

For the detection of emotion, more default models were employed, which are DenseNet121, VGG19, MobileNet, and EfficientNetB7. All the models were trained specifically on the given dataset.

Data Augmentation and Callbacks: To improve the stability of the model, data augmentation was used on the training data in the form of ImageDataGenerator. This included either random transformations like depth, rotation, width and height shifts, shear, zoom, horizontal flip, etc or random transformations like depth etc. Further, callback was also set for early stopping and reducing the learning rate when it reaches to plateau for better training.

Model Training: The model was trained by an augmented data generator. This involve the process used to train the model and then checking on the efficiency of the model in the validation set. The early stopping callback that was utilized earlier to avoid overfitting did the work of stopping the training process when there was no improvement on the validation accuracy next epoch after previous epochs

4.1.3 Evaluation

For evaluation of the performance during training the training and the validation accuracy and loss was graphed against the epochs. These plots were useful in pointing out any of the dangers of overfitting or underfitting. The matrix of confusion helped to identify the accuracy of the model in identifying each emotion, the classification report added information such as precision, recall, and F1-score for each class.

4.2 Text to Emotion

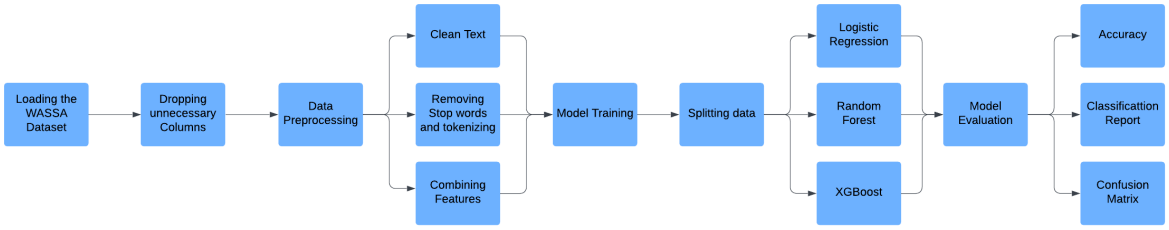

Figure 12: Text based detection process flow

4.2.1 Data Preparation and Preprocessing

Data Loading and Cleaning: Hence, the dataset was provided from the WASSA-2017 competition. The first step in data pre-processing was to read the data set and exclude unnecessary columns like 'rid' 'tid'. To preprocess each tweet text, mentions, links and special characters were removed from the text using regexion.

Text Preprocessing: The cleaned tweets were processed again through Natural Language Toolkit (NLTK) for further pre-processing. This involved splitting the text into individual words or tokens, eliminating the stop words and keeping all the tokens that were alphanumeric in nature to definitely eliminated the noise words.

4.2.2 Feature Extraction and Model Building

Vectorization: The processed tweets were vectorized with CountVectorizer in order to transform the textual data into numerical characteristics. Furthermore, instead of using only the text data the intensity values obtained from set were appended to these text feature set which basically contains both the textual and intensity characteristics of the data.

Label Encoding: Feelings in the dataset were labeled numerically. These labels were specifically used for the training and testing of the developed models.

4.2.3 Machine Learning Modelling

Train-Test Split: The data was partitioned into training and testing data with a ratio 80% training data and 20% testing data.

Model Training: There as several model training for the class of the given input text.

4.2.4 Evaluation

Accuracy and Classification Report: The performance of each model was analyzed using accuracy statistics as well as utilization of classification reports. The reports contained information about accuracy, sensitivity, and specificity in connection with each of the emotion categories.

Confusion Matrix: The confusion matrices were created for every model to compare the results and to look for the patterns of mistakes.

4.3 Speech to Emotion

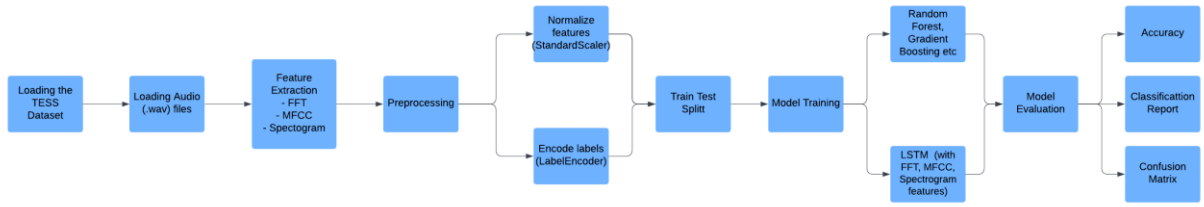

Figure 13: Speech Emotion detection process flow

4.3.1 Data Preparation and Feature Extraction

Data Loading: The series of actions was performed on TESS dataset which has emotional speech data collected from Toronto. The various audio files in the dataset correspond to various types of emotions for instance, neutral, disgust, sad, pleasant surprise, angry, fear and happiness.

4.3.2 Feature Extraction

For each audio file, multiple features were extracted using the librosa library:

Chroma STFT: Chromagram from a short-time Fourier transform, representing the energy of each pitch class.

FFT: Fast Fourier Transform, providing a frequency domain representation of the audio signal.

MFCCs: Mel-frequency cepstral coefficients, which represent the short-term power spectrum of the sound.

Spectrogram: Visual representation of the spectrum of frequencies of the audio signal as it varies with time.

Each feature was calculated, averaged across the time axis, and concatenated to form a comprehensive feature set for each audio file.

4.3.3 Model Building and Training

Data Transformation: The extracted features were flattened down and the dataset was transformed for compatibility in feeding the neural network input. Emotions are such attributes that were input encoded into numerical values through the label encoding technique. The dataset was split into training and testing sets using sklearn's train_test_split api in the ratio of 80:20. StandardScaler was applied to scale the data for features so as to ensure that the model performs well.

Model Architecture: A Long Short-Term Memory (LSTM) neural network was used for emotion classification.

Layers: It was comprised of two LSTM layers and two dropout layers for performances optimization by preventing over fitting. The final flatten layer was followed by a dense layer with ReLU activation useful for feature transformation as the last layer adopted softmax activation helpful in emotion classification.

Compilation: Thus, the model was used with the help of the sparse categorical cross-entropy loss function and the Adam optimizer.

4.3.4 Evaluation

Accuracy and Classification Report: The trained models were as to their accuracy and the forms of classification as the classification reports which have data on precision, recall, the F1-scores for each of the emotion categories.

Confusion Matrix: Confusion matrices were built to every model in order to analyze its performance and see cases when classes were confused.

Chapter 5: Results and analysis

5.1 Case 1: Image to Emotion

5.1.1 DenseNet121 Model

Due to the sensitivity of the task, DenseNet121 was trained for 25 epochs from the FER-2013 dataset with data augmentation. At first, there are high fluctuations, where beginning training accuracy is 30.12% while the accuracy of the model on the validation set is 47.45%. Thus, while undergoing training, the accuracy of our model was steadily beaming and at our final level, it reached up to 71.88% for Training and 63.78% for Validation. Nonetheless, the variation in the loss function was managed by using early stopping and learning rate reduction callbacks to avoid overlearning.

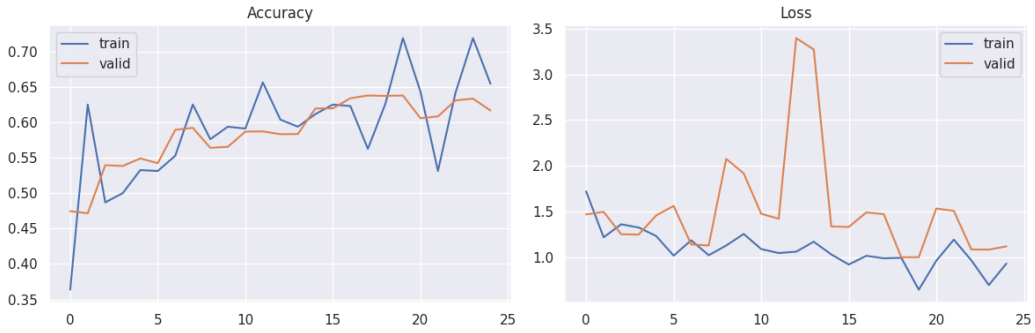

Figure 14: DenseNet training curves

The final performance of the DenseNet121 model on validation set is as following accuracy of 64%. The model's ability to predict emotions varied across different categories, **Best Performance**: The enhancing results of the model could be seen in the aspect of 'Happiness', in which it obtained an F1-score of 0.84. Challenging Emotions: 'Fear' and 'Sadness' were less recognizable; the F1-scores were 0.39 for 'Fear' and 0.50 for 'Sadness'. **Other Emotions**: Person emotions such as 'Anger,' 'Disgust,' 'Surprise' and 'Neutral' based statistics had most moderate to good performance with 'Surprise' having an F1 score of 0.72.

Figure 15: Output prediction of the DenseNet

5.1.2 EfficientNetB7 Model

The FER2013 set was also trained with the EfficientNetB7 model with the data augmentation to generalize the model for 25 epochs. The model began with the training accuracy of 29.23% and a validation accuracy of 45.58%. At the end of the training period, the model train accuracy was at 81.25% while validation accuracy was slightly lower at 66.68 % by the last epoch. Summary: All model weights were set to they value at the end of the best epoch in order to optimize the models.

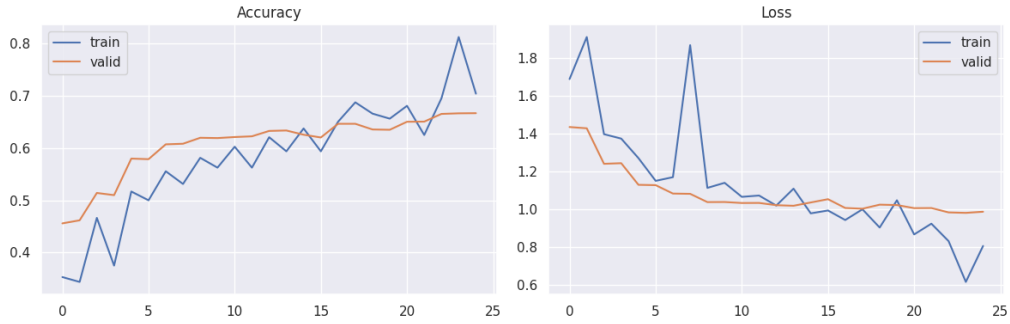

Figure 16: EfficientNetB7 training curves

The final accuracy was 67% on the validation set, Best Performance: The results show that the model is most accurate in predicting 'Happiness' based on an F1-score of 0.86. Challenging Emotions: The F1-scores of 'Fear' and 'Sadness' were a little low at 0.49 and 0.52, respectively." Other Emotions: 'Anger', 'Disgust', 'Surprise' and 'Neutral' emotion categories performed differently; out of which 'Surprise' got an F1-score of 0.76 and 'Neutral' having an F1-score of 0.64.

Figure 17: Output prediction of the EfficientNetB7

5.1.3 MobileNet

The MobileNet model underwent 25 epochs of training with the FER2013 dataset, applying data augmentation techniques to enhance its robustness. The training began with an initial accuracy of 26.42% and a validation accuracy of 29.28%. As the training progressed, the model showed notable improvements. By the end of training, the model achieved a training accuracy of 56.25% and a validation accuracy of 56.73%. The best model weights were restored from epoch 21 to ensure optimal performance.

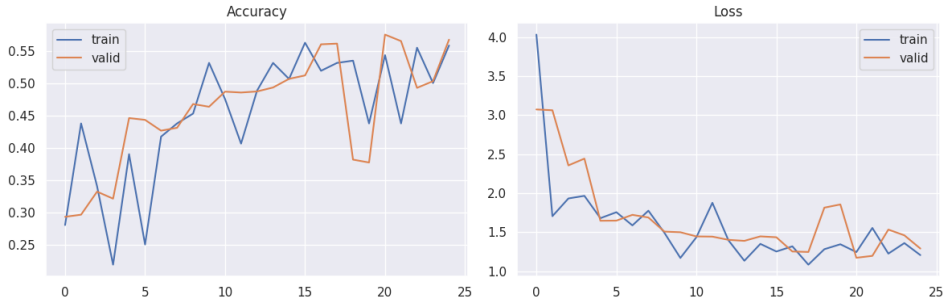

Figure 18: MobileNet training curves

The final accuracy was reached on the validation set which was 58 percent. Best Performance: ‘Happiness’ was accurately predicted by the model using the F1-score of 0.80. Challenging Emotions: ‘Fear’ and ‘Sadness’ were comparatively moredifficult to classify with an F1-score of 0.38 and 0.41 respectively. Other Emotions: Four tests including ‘Anger’, ‘Disgust’, ‘Surprise’, and ‘Neutral’, varied performance, ‘Anger’ had a recall of 0.65 while ‘Surprise’ had F1-score 0.64.

Figure 19: Output prediction of the MobileNet

5.1.4 VGG19

The training of VGG19 model was done for 25 epochs using dataset named FER2013. First, it gave training accuracy of 33.26% and validation accuracy of 52.77%. From this training alone, there were enhancements as time went on. As for the last iteration of the training, which refers to 40 epochs, the training accuracy of the model reached 69.67%, while its validation accuracy is slightly lower – 65.87%. We restored the weights of the best model from epoch 20 to build and train the loaded model to standard accuracy.

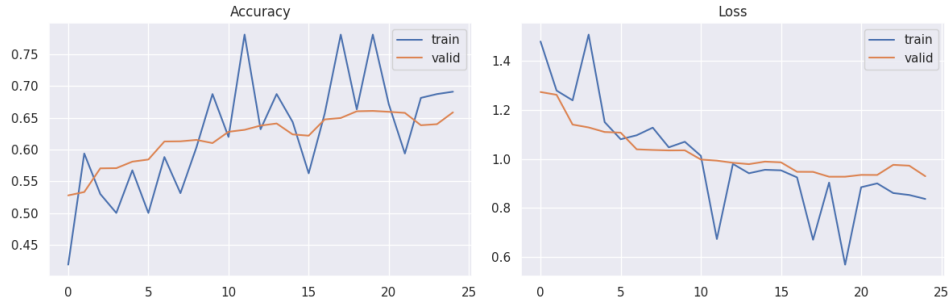

Figure 20: VGG19 training curves

When tested on the validation set, the final accuracy was 66 percent. **Best Performance**: The data was best modelled in predicting ‘Happiness’ with an F1-score of 0.87. **Challenging Emotions**: The confusion matrices of ‘Fear’ and ‘Sadness’ were slightly less easy to categorise with both having F1-scores of 0.45 and 0.57 respectively. **Other Emotions**: The performances of ‘Anger’, ‘Disgust’, ‘Surprise’ and ‘Neutral’ emotional labels were not identical; where ‘Surprise’ classification got 0.77 F1-score while that of ‘Neutral’ resulted to 0.61 F1-score..

Figure 21: Output prediction of the VGG19

5.1.5 ResNet50

Thus, training the ResNet50 model was conducted in 25 epochs using a dataset as FER2013, using augmentation. Training started with accuracy of 29.98% and the validation accuracy was 41.35%. But as the days went by the training the improvement of the model was remarkable. Training accuracy of the proposed model reached 64.26% while the validity accuracy was at a level of 60.69% at the end of the training. In order to get the best model weights, the epoch 24 weights were restored for better performance.

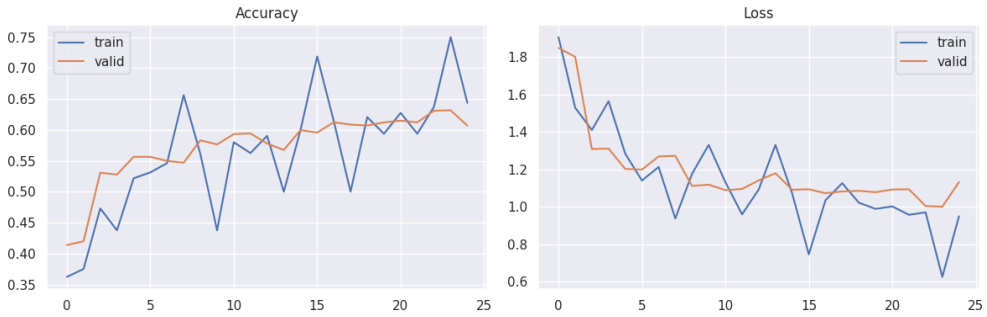

Figure 22: ResNet training curves

On the validation set the final accuracy was 63%. **Best Performance**: MSE and F1 score were evaluated to compare the designed model performance in predicting these indicators—Overall, the model had a remarkably improved accuracy to predict 'Happiness' with F1-score: 0.83. **Challenging Emotions**: 'Fear' was slightly more challenging to classify with an F1-score of 0.39 and 'Sadness' fear had a score of 0.55. **Other Emotions**: Categorical Sentiment Analysis has a fairly good result, Average F1-score = 0.65 Medium F1-score: 'Surprise' = 0.72 Low F1-score: 'Neutral' = 0.59.

Figure 22: Output prediction of the ResNet

5.2 Case 2 – Text to Emotion

The Logistic Regression model gave 87.76% accuracy on the test set. It was most precise and accurate in the 'Joy' target emotion and yielded precision and recall scores of 0.90 & 0.88 and F1 scores of 0.88 in each of the six target emotions. 'Fear' was also recognized with high accuracy of F1-score of 0.89.. The model performed marginally poorer than the other six categories for 'Sadness', with an F1- score of 0.81, although true to the nature of other ten categories, the model is equally accurate and consistent.

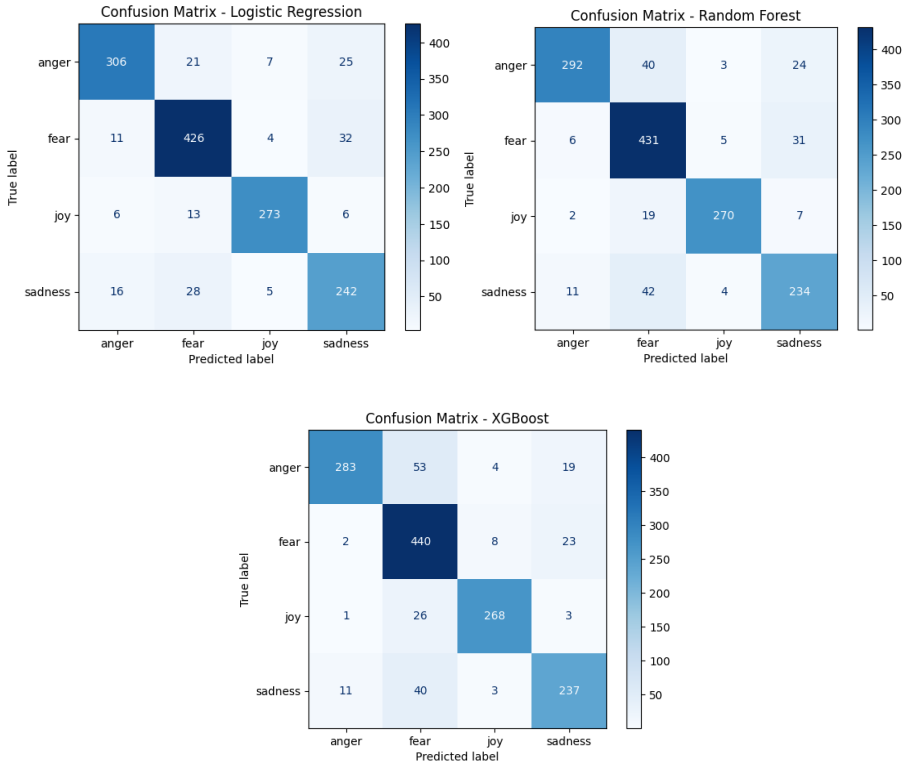

Figure 23: Confusion matrix (a) Logistic Regression (b) Random Forest (c) XGBoost

From Table 4, Random Forest model performed well with high accuracy of 86.35%. It was most accurate in estimating 'Joy' with the F1-score of 0.93 and 'Anger' with the F1-score 0f 0.87. When it came to 'Sadness' and 'Fear' the model was marginally poorer with F1-scores of 0.80 and 0.86, respectively. In general, Random Forest yielded a favourable ratio of precision and recall for all the categories of emotion. From the XGBoost model, the quality of prediction is 86.42%; For 'Joy' it has F1-score of 0.92 and 'Anger' it has F1 of 0.86. It also obtained high F1-scores of 0.83 for 'Sadness', and 0.85 for 'Fear'. The micro average F1 score for the XGBoost is 0.87 which shows that it has a good performance and repeats good classifying performance across the different categories.

5.3 Case 3 – Speech to Emotion

The proposed LSTM model using chroma STFT, FFT, MFCCs, and spectrogram achieved a remarkable emotion classification model. Across the fifty epochs of training, this model very well converged, which can been seen by experimental validation where the model starting with a first epoch validation accuracy of 79.69% and reached a second epoch validation accuracy of 92.86% rapidly. The model constantly enhances, placing close to perfect—achieving the final training accuracy of 99.72% and validating accuracy of 95.31%. The class report highlighted this performance whereby precision, recall and F1 - scores were nearly 1.0 for all emotion classes and overall accuracy stood at 98%.

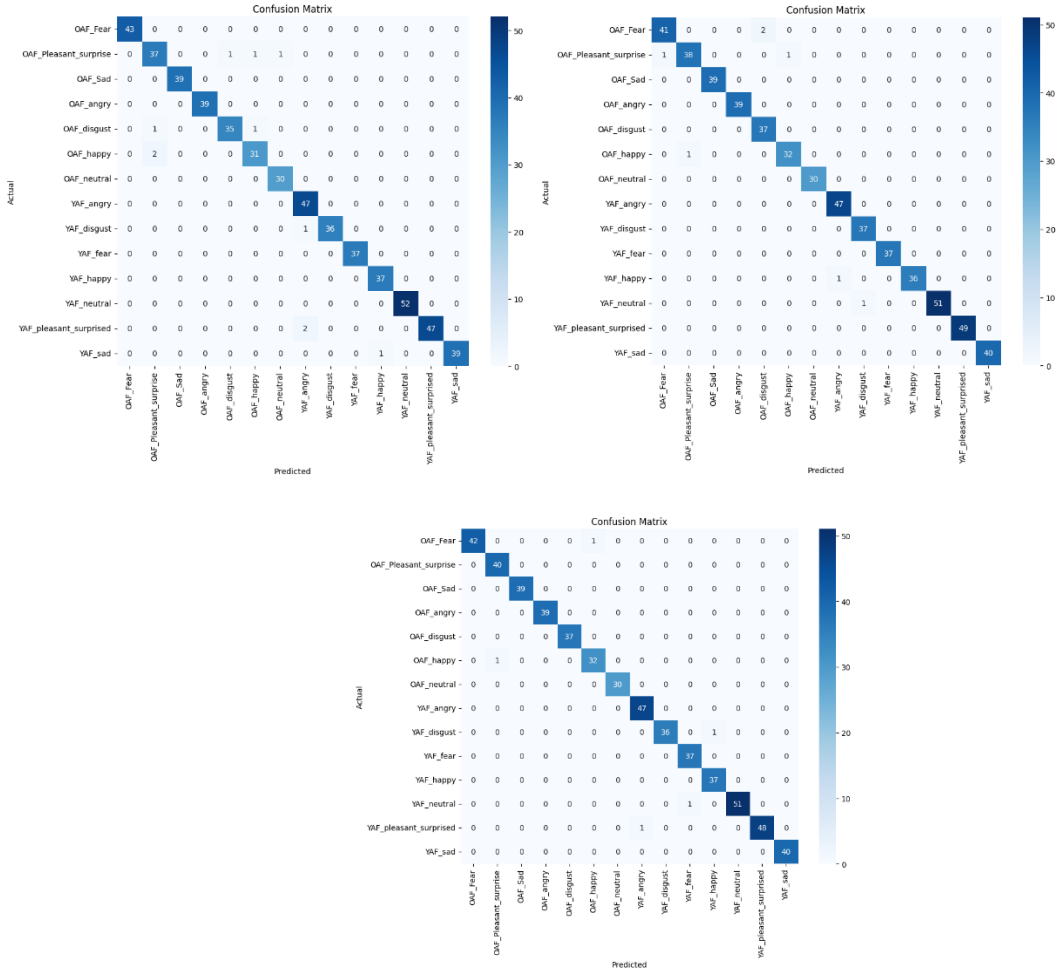

Figure 24: (a) STFT – Based LSTM Model (b) MFCC – Based LSTM Model (c) Spectrogram – Based LSTM Model

The LSTM model developed based on the MFCC also yielded exceptional performances which were mainly due the fact that the features of the MFCC were mainly concentrated on. Training from 24.78% of accuracy in the first epoch, the model increases by leaps and bound; it had a validation accuracy of 87.50% by the third epoch, and 98.66% by the 50th epoch. Such rapid merging indicates the soundness of using MFCC features as a way of interpreting the dynamism of emotions during speech. In these two models, evaluation metrics stayed the same; moreover, all the categories approached to 100% and the total accuracy reached 99%. This further validates the analysis that concluded that MFCC features can be extremely beneficial for detecting emotion.

The Spectrogram-based LSTM model using the spectrogram features also revealed similar performance. Originally, the accuracy of training: 29.72 %, accuracy of validation: 70.98 % at epoch: 1. It rose considerably, as suggested in Figure 8 where the model attained a validation accuracy of 95.98% in the third epoch and remained almost constant throughout the training period reaching 99.11% at the last epoch. Classification report of this model also depicted best results in terms of precision, recall, and F1- score which are almost close to 0.99 and accuracy of almost 99%.

5.4 Case 4 – Real Time Application

The proof of concept (POC) of this application has been deployed on Streamlit and can be accessed by the url: http://multimodal-emotions-detector.streamlit.app. The results of the POC for all three types of input are as follows:

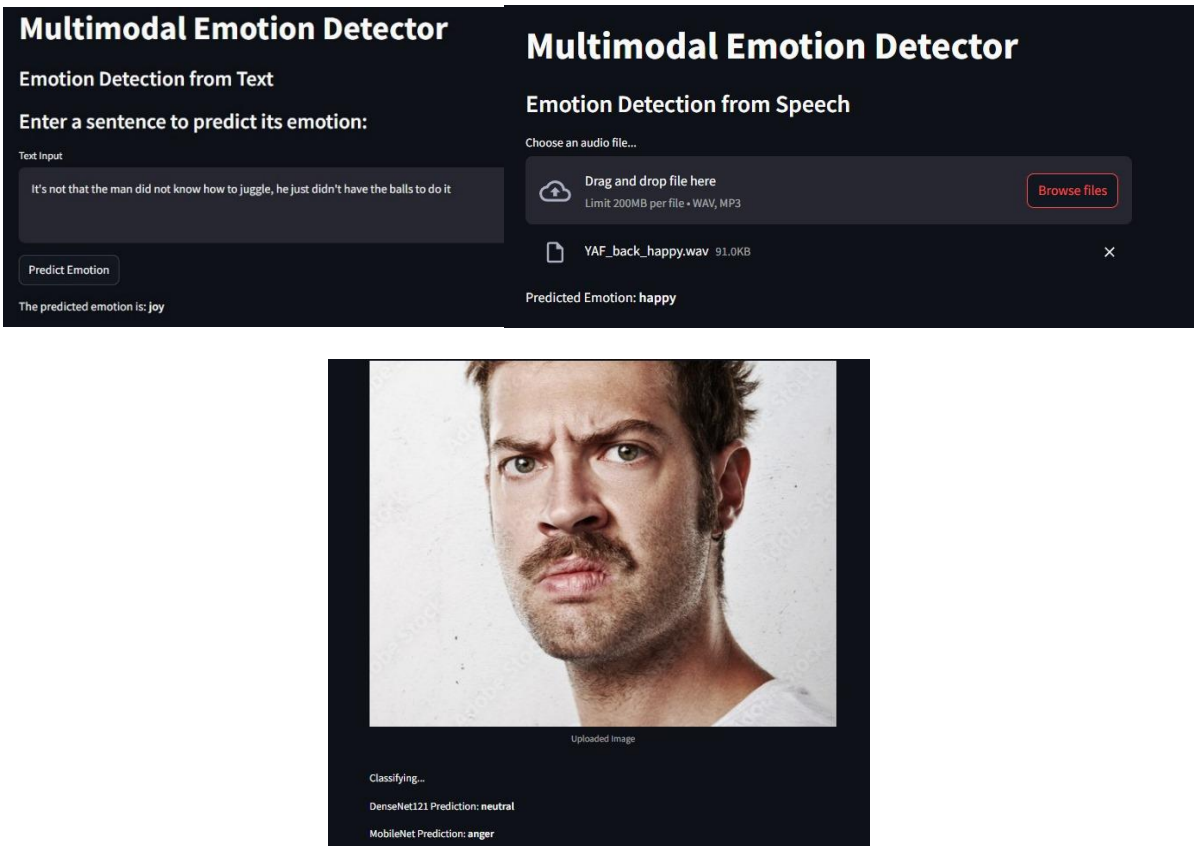

Figure 25: Real time application using the Streamlit application

5.5 Discussion and Analysis

RQ1: How can the predictive analytics models be optimised to provide reliable early signs of life-threatening emotions with the retrieved data?

Table 2: CV based analysis

Model	F1 Score (Anger)	F1 Score (Disgust)	F1 Score (Fear)	F1 Score (Happiness)	F1 Score (Sadness)	F1 Score (Surprise)	F1 Score (Neutral)
DenseNet121	0.58	0.54	0.39	0.84	0.50	0.72	0.61
EfficientNetB7	0.60	0.64	0.49	0.86	0.52	0.76	0.64
MobileNet	0.50	0.16	0.38	0.80	0.41	0.64	0.58
VGG19	0.57	0.38	0.45	0.87	0.57	0.77	0.61
ResNet50	0.55	0.54	0.39	0.83	0.55	0.72	0.59

Comparing all the models it can be ascertained that each model seemed to have its unique forte and downsides when it came to the identification of emotions. DenseNet121 and EfficientNetB7 yielded the best performance with validation accuracies of 64% and 67% respectively and performed very well for

‘Happiness’ and ‘Surprise’. However, the MobileNet and VGG19 models provided quite decent levels of accuracy: MobileNet scored 58% and VGG19— 66% to detect the ‘’Fear’’ and/or ‘’Sadness’’ classes. Nonetheless, ResNet50 is considered a more robust architecture and achieved a final accuracy of approx 63% but the challenges with the particular classes were as same as the previous case. In general, all the models presented certain effectiveness with different levels and proved that DenseNet121 and EfficientNetB7 were the most efficient among all the models, fertile from deeper architecure and more advanced pre-training method. Further research could be performed with the aim of optimising the classification of difficult emotions such as ‘Fear’ and ‘Sadness’ using additional and more ecologically valid training samples and refining the models. Often, the choice of a model is influenced by certain context or particular scenario, as it has been demonstrated in this study with various performance improvements across the different emotions; but, overall, the findings of this assessment stress the necessity of their fine-tuning depending on particular tasks and aims in emotion detection endeavors.

Table 3: Analysis of Speech dataset

Model	Test Accuracy	Joy F1 Score	Fear F1 Score	Sadness F1 Score	Anger F1 Score	Avg Macro F1 Score
Logistic Regression	87.76%	0.93	0.89	0.81	0.88	0.88
Random Forest	86.35%	0.93	0.86	0.80	0.87	0.86
XGBoost	86.42%	0.92	0.85	0.83	0.86	0.87

Based on the evaluation of the text-to-emotion models, the Logistic Regression model demonstrated the highest overall accuracy at 87.76%, closely followed by the XGBoost model at 86.42% and the Random Forest model at 86.35%. Each model showed strengths in predicting 'Joy' and 'Anger', while 'Sadness' and 'Fear' were more challenging. The Logistic Regression model excelled in its precision and recall for 'Joy' and 'Fear', making it the most reliable for these emotions. However, all three models demonstrated a high degree of accuracy and consistency, indicating their effectiveness for text-based emotion detection. Future improvements could focus on enhancing the models' performance for the more challenging emotions like 'Sadness' and 'Fear' through additional training data and further fine-tuning.

RQ2: Which implementation challenges may occur because of extending the developed mental health analytics solution into typical healthcare systems or platforms and how can those challenges be resolved?

The main challenge is to make a proper channel that can help in integrating in the real time analysis. Currently we have tested with the standard datasets such that none of the GDPR issues are there.

Chapter 6: Conclusion and Future Work

As more people are experiencing mental health issues more and more in the present and even more so as a result of COVID-19, there is a critical need for better systems to discern and treat lethal emotions. The proposed multimodal AI system seems to have potential for apps which could apply speech, vision and text analysis for the identification of emotional deterioration. Using deep learning technologies, this framework improves the detection and interpretation of key emotions that will help in the identification of patients in need of early counseling. In addition, due to its ability to grow and apply in real time, resources are beneficial for mental health care employees who strive to enhance the efficacy of therapy and intervention. The system emphasizes the future impact of AI in mental health care, but at the same time emphasizes the need for further studies to solve the ethical issues and difficulties in

introducing the technique as well as dataset diversification. Given these factors, this approach should help to enhance the prospects for better mental health as well as save lives everywhere.

All three models demonstrated robust performance in detecting emotions from speech data, with accuracies consistently above 98%. The Spectrogram-based LSTM model and the MFCC-based LSTM model both achieved slightly higher overall accuracies of 99%, while the combined features LSTM model closely followed with 98%. These results underscore the effectiveness of using advanced LSTM architectures in conjunction with rich feature sets for emotion detection tasks. The high accuracies across all models highlight their capability to generalize well, effectively handling class imbalances and providing reliable emotion classification across diverse speech datasets. This consolidated performance suggests that any of the three models could be reliably employed in real-world applications, with the choice of model potentially being guided by specific computational constraints or feature availability.

In addition, several future works can be performed to improve the outcomes of the proposed multimodal AI system. Increasing data sample to cover increasingly numerous populations, those speaking different languages, and those from different cultures will help in increasing the range of possible uses of the system. A unification of features containing more sophisticated integration schemes of speech, vision, or text can be achieved through more complex fusion techniques including but not limited to graph-based or dynamic fusions. More, the capabilities of real-time deployment in alert and timely identification of emotional distress have to be established. Some of the concerns that comes out include data privacy and data transparency issues which they must meet in order to apply proper ethical practices. Modernization of the system in other words will allow linking with healthcare systems that will align operation of AI tools with mental health specialists. Integrating temporal and behavioral aspects can facilitate risk evaluation procedures, whereas exploring cross-modal conjunctions, for instance, tone face can help in emotion identification. Mechanisms of adaptive learning may help the system change according to the situation occurred in social media platforms. These actions will change the layout into a more powerful and elastic instrument for reaction to public aid and mental illness all over the world.

References

Geethanjali, R., & Valarmathi, A. (2024). A Novel Hybrid Deep Learning IChOA-CNN-LSTM Model for Modality-Enriched and Multilingual Emotion Recognition in Social Media A Novel Hybrid Deep Learning IChOA-CNN-LSTM Model. https://doi.org/10.21203/rs.3.rs-4609288/v1

Makhmudov, F., & Kultimuratov, A. (2024). Enhancing Multimodal Emotion Recognition through Attention Mechanisms in BERT and CNN Architectures. Applied Sciences. https://doi.org/10.3390/app14104199

Y Chai, Emotion Intensity Detection in Online Media: An Attention Mechanism Based Multimodal Deep Learning Approach. (2024). Tehnicki Vjesnik-Technical Gazette. https://doi.org/10.17559/tv-20230628001154

Dai, Z., Fei, H., & Lian, C. (2024). Multimodal information fusion method in emotion recognition in the background of artificial intelligence. Internet Technology Letters. https://doi.org/10.1002/itl2.520

Wang, X., Ran, F., Hao, Y., Zang, H. L., & Yang, Q. (2024). Sequence Modeling and Feature Fusion for Multimodal Emotion Recognition. https://doi.org/10.1109/iccect60629.2024.10546216

Wang, X., Li, M., Chang, Y., Luo, X., Yao, Y., & Li, Z. (2023). Multimodal Cross-Attention Bayesian Network for Social News Emotion Recognition. https://doi.org/10.1109/ijcnn54540.2023.10191298

Multimodal Emotion Recognition Using Heterogeneous Ensemble Techniques. (2022). https://doi.org/10.1109/iccit57492.2022.10054720

R, A. R., & Nagarajan, B. (2024). Cutting-Edge AI Technology to Recognize Signs of Suicidal Thoughts in Social Media Posts. https://doi.org/10.21203/rs.3.rs-4621229/v1

Nan, J., & Yao, T. (2024). Research on netizen sentiment recognition based on multimodal deep learning. https://doi.org/10.1117/12.3032110

Ezerceli, Ö., & Dehkharghani, R. (2024). Mental disorder and suicidal ideation detection from social media using deep neural networks. Journal of Computational Social Science. https://doi.org/10.1007/s42001-024-00307-1

Omar, Adel., Karma, M.Fathalla., Ahmed, Abo, ElFarag. (2023). MM-EMOR: Multi-Modal Emotion Recognition of Social Media Using Concatenated Deep Learning Networks. Big data and cognitive computing, doi: 10.3390/bdcc7040164

Madhura, Prakash, M., S, Meghana., Varshitha, VS., Ashwini, Kodipalli., Trupthi, Rao. (2024). Neural Networks and Emotions: A Deep Learning Perspective. doi: 10.1109/i2ct61223.2024.10544003

Xiaofei, Wang., Feng, Ran., Yanpeng, Hao., H., L., Zang., Qian, Yang. (2024). Sequence Modeling and Feature Fusion for Multimodal Emotion Recognition. doi: 10.1109/iccect60629.2024.10546216

Jun, Nan., Tianhao, Yao. (2024). Research on netizen sentiment recognition based on multimodal deep learning. doi: 10.1117/12.3032110

Nikita, Paras, Toliya., N., Nagarathna. (2024). Leveraging Online Social Content for Early Detection of Suicidal Ideation: A Multi-Modal Deep Learning Approach. doi: 10.1109/icetcs61022.2024.10544279

Anjit, Raja, R., Bhalaji, Nagarajan. (2024). Cutting-Edge AI Technology to Recognize Signs of Suicidal Thoughts in Social Media Posts. doi: 10.21203/rs.3.rs-4621229/v1

R., Biswas. (2023). Suicidal Thoughts Detection from Social Media Using AI. doi: 10.70121/001c.121685

Masab, A., Mansoor., Kashif, Ansari. (2024). Early Detection of Mental Health Crises through AI-Powered Social Media Analysis: A Prospective Observational Study. doi: 10.1101/2024.08.12.24311872

Hamed, Jelodar., Hamed, Jelodar., Rita, Orji., Stan, Matwin., Stan, Matwin., Swarna, Weerasinghe., Oladapo, Oyebode., Yongli, Wang. (2021). Artificial Intelligence for Emotion-Semantic Trending and People Emotion Detection During COVID-19 Social Isolation. medRxiv, doi: 10.1101/2021.01.16.21249943

Varsha, Baby., Dama, Sudheshna., D, Sarvani., Sahithi, Vesangi., Saketh, Reddy, Regatte. (2022). 3. An Integrated Approach for Suicidal Tendency Detection. doi: 10.1109/ICAC3N56670.2022.10074359

Saraf, Anika., Swarup, Dewanjee., Sidratul, Muntaha. (2024). Analyzing Multiple Data Sources for Suicide Risk Detection: A Deep Learning Hybrid Approach. International Journal of Advanced Computer Science and Applications, doi: 10.14569/ijacsa.2024.0150270

Anshu, Malhotra., Rajni, Jindal. (2020). Multimodal Deep Learning based Framework for Detecting Depression and Suicidal Behaviour by Affective Analysis of Social Media Posts. EAI Endorsed Transactions on Pervasive Health and Technology, doi: 10.4108/EAI.13-7-2018.164259

Diana, Ramírez-Cifuentes., Ana, Freire., Ricardo, Baeza-Yates., Joaquim, Puntí., Pilar, Medina-Bravo., Diego, Velázquez., Josep, M., Gonfaus., Jordi, Gonzàlez. (2020). Detection of Suicidal Ideation on Social Media: Multimodal, Relational, and Behavioral Analysis.. Journal of Medical Internet Research, doi: 10.2196/17758

Joshua, Cohen., Vanessa, Richter., Michael, Neumann., David, P., Black., Allie, Haq., Jennifer, Wright-Berryman., V., Ramanarayanan. (2023). A multimodal dialog approach to mental state characterization in clinically depressed, anxious, and suicidal populations. Frontiers in Psychology, doi: 10.3389/fpsyg.2023.1135469

Moumita, Chatterjee., Piyush, Kumar., Poulomi, Samanta., Dhrubasish, Sarkar. (2022). Suicide ideation detection from online social media: A multi-modal feature based technique. International journal of information management data insights, doi: 10.1016/j.jjimei.2022.100103

Pradeep, Kumar., Dilip, Singh, Sisodia., Rahul, Shrivastava. (2024). A Deep Learning-Based Sentiment Classification Approach for Detecting Suicidal Ideation on Social Media Posts. Communications in computer and information science, doi: 10.1007/978-3-031-54547-4_21

Fahim K. Sufi; Ibrahim Khalil (2022). Automated Disaster Monitoring from Social Media Posts using AI based Location Intelligence and Sentiment Analysis. doi: 10.36227/techrxiv.19212105.v1

Barua, P.D., Vicnesh, J., Lih, O.S. et al. (2024). Artificial intelligence assisted tools for the detection of anxiety and depression leading to suicidal ideation in adolescents: a review. Cogn Neurodyn 18, 1–22 https://doi.org/10.1007/s11571-022-09904-0

I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio. Challenges in representation learning: A report on three machine learning contests. Neural Networks, 64:59--63, 2015. Special Issue on "Deep Learning of Representations"

Pichora-Fuller, M. Kathleen; Dupuis, Kate, (2020). Toronto emotional speech set (TESS). , https://doi.org/10.5683/SP2/E8H2MF, Borealis, V1

Saif Mohammad and Felipe Bravo-Marquez. (2017). WASSA-2017 Shared Task on Emotion Intensity. In Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 34–49, Copenhagen, Denmark. Association for Computational Linguistics.

https://www.geeksforgeeks.org/vgg-net-architecture-explained/

Sohaib & Ming, Zhao & Tang, Fengxiao & Zhu, Yusen. (2023). LWSE: a lightweight stacked ensemble model for accurate detection of multiple chest infectious diseases including COVID-19. Multimedia Tools and Applications. 83. 1-37. 10.1007/s11042-023-16432-4.

Tiwari, V., 2010. MFCC and its applications in speaker recognition. International journal on emerging technologies, 1(1), pp.19-22.

Wyse, L., 2017. Audio spectrogram representations for processing with convolutional neural networks. arXiv preprint arXiv:1706.09559.