International Journal of Reconfigurable and Embedded Systems (IJRES)
Vol. 14, No. 1, March 2025, pp. 254~264
ISSN: 2089-4864, DOI: 10.11591/ijres.v14.i1.pp254-264

Multimodal recognition with deep learning: audio, image, and text

Ravi Gummula, Vinothkumar Arumugam, Abilasha Aranganathan
Department of Electronics and Communication Engineering, Dr. M.G.R. Educational and Research Institute, Chennai, India

Article Info

Article history:

Received Aug 1, 2024
Revised Sep 11, 2024
Accepted Sep 28, 2024

Keywords:

Convolutional neural network
Efficacy
Emotion detection
Evaluation criteria
Long short-term memory
Term frequency-inverse document frequency vectorizer

ABSTRACT

Emotion detection is essential in many domains including affective computing, psychological assessment, and human computer interaction (HCI). It contrasts the study of emotion detection across text, image, and speech modalities to evaluate state-of-the-art approaches in each area and identify their benefits and shortcomings. We looked at present methods, datasets, and evaluation criteria by conducting a comprehensive literature review. In order to conduct our study, we collect data, clean it up, identify its characteristics and then use deep learning (DL) models. In our experiments we performed text-based emotion identification using long short-term memory (LSTM), term frequency-inverse document frequency (TF-IDF) vectorizer, and image-based emotion recognition using a convolutional neural network (CNN) algorithm. Contributing to the body of knowledge in emotion recognition, our study's results provide light on the inner workings of different modalities. Experimental findings validate the efficacy of the proposed method while also highlighting areas for improvement.

Corresponding Author:

Ravi Gummula
Department of Electronics and Communication Engineering, Dr. M.G.R. Educational and Research Institute
Chennai, Tamil Nadu, India
Email: ravi.gummula@gmail.com

1. INTRODUCTION

Many intelligent systems might benefit greatly from automated emotion recognition, including those used in online gaming, digital advertising, healthcare and consumer feedback collection. For instance, by adapting the game interface according to the user's emotional state, an emotion recognition function might potentially enhance player engagement in online gaming. In a similar way a live emotion detection module may provide the selling firm immediate emotional feedback when a customer buys online, allowing them to present the customer new offers. Healthcare providers may better monitor their patients' physical and mental health via emotion detection, which in turn helps them provide the most appropriate medication or therapy [1].

Intelligent conversational systems, smart cities, affect-aware e-health, affect-aware learning, and travel recommendation systems are just a few of the many applications that are using emotionally-aware intelligent systems [2]. A lot of systems rely on textual or emotion-based inputs. For example, there have been suggestions for emotion-aware e-health systems that search for certain keywords in patient input in order to identify emotions [3]. We have developed trip recommendation algorithms that are based on context or emotion and affect-aware learning technologies [4]–[6]. Improving the quality life of people might be achieved by an affect-aware smart city's ability to recognize and show emotions through the use of hashtags, keywords, and emotions [7]. Each of these systems primarily relies on text or emotions to detect emotions [8].

Video, audio, short words, emotions, long texts, short messages, and facial expressions are some of the inputs that may be utilized to detect emotions. For these inputs, applications employ a variety of formats. For instance, whereas video is often used in gaming systems, short messages, and emotions are more common on social networking. Additionally, systems that can identify emotions from electroencephalogram (EEG) data have been introduced lately [9], [10]. However, wearing an EEG hat is invasive and uncomfortable for the wearer. A literature review found that when it comes to emotion identification, single-modal input often fails to meet the necessary accuracy standards [11], [12].

A technique for audio-visual emotion recognition based on deep network feature extraction and fusion is presented in this study [13]. Finally, non-linear feature fusion is ensured by using support vector machines (SVMs), which are networks. The accuracy of deep learning (DL) models is dependent on the available data and the structure of the model, although it has seen extensive usage in image, video, and audio processing [14]. There are three things that this research adds: the proposed system takes use of a large emotion dataset for training, a three-dimensional convolutional neural network (CNN) with an intricate key frame selection technique for video signals and a variety of informative patterns for feature extraction, including gray-scale key frame images, local binary pattern (LBP) images, and interlaced derivative pattern (IDP) images.

2. LITERATURE SURVEY

As highlighted by Kudiri *et al.* [15] most research in this field has utilized asynchronous data and unimodal or multimodal systems. As a result, incorrect synchronization has become a common issue, increasing system complexity, and decreasing response time. To address this, a unique method has been developed to anticipate human emotions from speech and facial expressions. This method employs two feature vectors: relative bin frequency coefficient (RBFC) for voice data and relative sub-image based (RSB) coefficient for visual data. The fusion strategy between the two modalities is based on feature-level categorization and utilizes a SVM with a radial basis kernel.

According to Wang *et al.* [16] estimating human emotions using a computer have proven challenging during conversational breaks. Their study uses a hybrid approach combining speech and facial expressions to measure fundamental emotions. The approach employs RBFC for audio data and RSB features for visual data, with classification performed by radial basis kernel SVMs. The findings indicate that the proposed feature extraction method significantly enhances the emotion recognition system. It demonstrates that the bimodal emotion recognition system, utilizing both speech and facial expressions, outperforms unimodal methods.

According to Khalil *et al.* [17] recognizing emotions in spoken language is challenging part of human computer interaction (HCI). As an alternative to conventional machine learning (ML) methods, DL approaches are introduced. This article surveys various approaches and examines current research on emotion identification using DL. This review covers databases used, enhancements to speech emotion identification and its limitations.

3. PROBLEM STATEMENT

This study proposes a hybrid approach that uses speech and facial expressions to find fundamental emotions in an arsonist during a conversational break [18]. For the audio and visual data respectively RBFCs and RSB tires are used. For classification a SVM with a radial basis kernel is used. This study's findings indicate that, in conjunction with the fusion approach, the most significant factor influencing the emotion recognition system is feature extraction through speech and facial expression. While some factors may have an impact on the system's ability to identify emotions, this impact is rather small [19]. Through intentional facial expressions, it was shown that the unimodal emotion identification system outperforms the bimodal emotion detection system in terms of performance. A proper database is employed to address the problem [20]. The suggested emotion recognition system outperformed the others in terms of fundamental emotional classes.

4. PROPOSED MODEL

Emotions are important to human perception and communication and have a substantial influence on actions and choices. Consequently, emotion detection has been trending recently, with a focus on human emotion recognition and classification across several media types (e.g., images, text, and audio) [21]. In our research, we adopt a multi-modal approach to emotion recognition, encompassing text, image, and audio modalities. Our goal is to identify the seven primary emotions: anger, fear, disgust, neutral, happy, sorrow, and surprise across different data sources, leveraging CNN architectures for image and audio data, and long short-term memory (LSTM) networks for textual data. Similarly, for audio-based emotion recognition, we employ CNN architectures trained on datasets like the RAVDESS audio dataset. These networks analyze spectrograms or other representations of audio signals to identify emotional cues conveyed through speech.

For text-based emotion analysis, we employ LSTM networks due to their ability to capture sequential dependencies in textual data. We preprocess conversational text and feed it into LSTM networks for emotion classification. Additionally, we utilize the term frequency-inverse document frequency (TF-IDF) vectorizer to transform textual conversations into numerical representations, capturing the importance of individual terms in expressing emotions within the corpus. Through this integrated approach, we aim to leverage the strengths of each modality to enhance the accuracy and robustness of emotion recognition systems [22]. By experimenting with various CNN and LSTM architectures tailored for image, audio, and text modalities, our research strives to advance the state-of-the-art in multi-modal emotion recognition, offering valuable insights into the complexities of human emotions expressed through language, facial expressions, and speech [23].

- The proposed multi-modal system integrates text, image, and audio modalities, enhancing accuracy and robustness in emotion recognition.
- It offers a comprehensive understanding of human emotions, improving versatility, and potential for real-world applications.

For the proposed emotion recognition system, we have devised an LSTM architecture tailored for text data. The LSTM model depicted in Figure 1 consists of multiple layers designed to capture sequential dependencies in textual inputs effectively. Figure 1 outlines the detailed architecture specifications of the LSTM model.

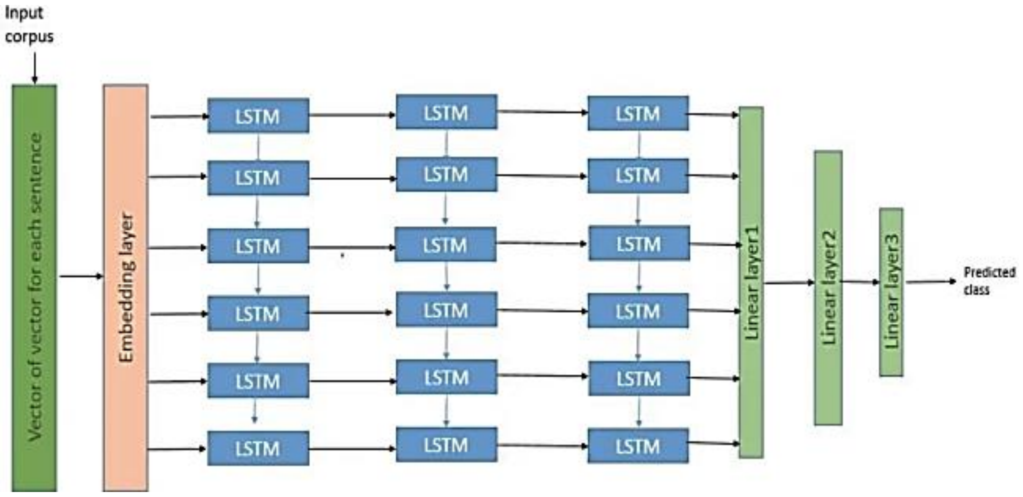

Figure 1. LSTM model

Following the LSTM layers, a fully connected neural network with hidden layers is employed to further process the learned features. A SoftMax function is then applied to the output of the fully connected layer to generate probability distributions over the target emotion classes. Subsequently, the output of the SoftMax layer is passed through a classifier for final emotion classification.

However, conventional ML techniques exhibit limitations in accurately recognizing emotions from text data. Enhancing the accuracy of emotion recognition in text-based inputs remains a significant objective in the domain of artificial intelligence and DL algorithms. Traditional text-based emotion recognition datasets often achieve suboptimal accuracy rates. To address these limitations and improve emotion recognition accuracy, this study proposes the following innovations.

Integration of a supervised encoder for feature extraction from textual inputs. Utilization of LSTM networks to analyze textual inputs and perform emotion recognition. Implementation of neural networks for information collection and visualization of emotional cues conveyed through text. Deployment of a CNN algorithm for analyzing emotions from text data within intelligent learning environments. Additionally, classification enhancement algorithms based on ML techniques are employed for improved accuracy in text-based emotion recognition. Compared to traditional text-based emotion recognition approaches, the LSTM architecture presented in this study demonstrates superior accuracy and yields better recognition results.

4.1. Data set

It is impossible to overstate the significance of data while developing DL models. In order to learn, generalize and draw educated conclusions, algorithms need data. This data is used to fuel developments in many different industries. The amount, diversity, and quality of the data used for training play a significant role in our emotion recognition work. We have developed a trustworthy and highly effective emotion recognition model using a large dataset that include voice, image, and text data. This section provides a detailed description of how we gather and process data. The selected emotion classes are placed across all 3 modules.

4.2. Data preprocessing and analysis

Data preparation was crucial to the research since it enabled us to properly use our unique data gathering. Here we took the steps to prepare the audio, picture, and textual data for model analysis and training. In order to provide a solid groundwork for creating accurate and dependable emotion recognition models, we used systematic data preparation to increase the dataset's representativeness, balance, and quality.

4.2.1. Text

The text module has a robust dataset consisting of 32,500 records, each of which has two essential columns: "text" and "emotion". The "text" column lists all the materials that our model needs for its rigorous training, and the "emotion" column shows which emotion category each text item belongs to. Despite some initial setbacks, we eventually located and selected several text sources that met our established criteria and allowed us to accomplish our research goals, allowing us to compile a sizable corpus of text data. We took great effort to select reliable and authentic sources despite the challenges we had while gathering data in order to ensure that our dataset remained accurate and applicable. Emotion expressions in real-world data sources are naturally distributed and variable, making it difficult to achieve a perfect balance for all emotion classes. Nevertheless, we were able to capture a wide range of patterns by intentionally crafting the dataset with an emphasis on inclusion and variety. With seven distinct emotions represented by the encoded classes in the dataset, the records in each category exhibit a wide range of distributions. Our subsequent research and model development meticulously addressed any potential class imbalances to ensure the text emotion recognition (TER) model's unbiased performance. The extensive and detailed text data gathered from sources allows for a full exploration of textual expressions for emotion recognition.

4.2.2. Image

We used a dataset that found on Kaggle for the image module to construct our study on face-based emotion recognition. In all, these datasets include around a thousand images which have been classified into 7 distinct groupings according to various emotional states as shown in Figure 2. To fix any potential data imbalances and increase the image data's range, we applied data augmentation techniques. Augmentation enhances our model's generalizability by enhancing the dataset with manufactured versions of the existing pictures.

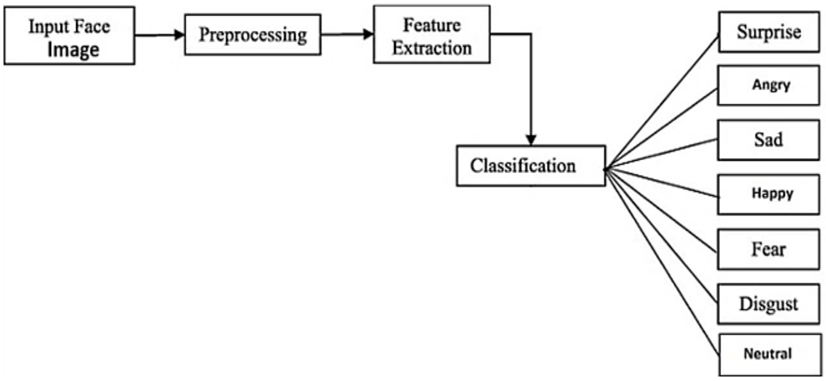

Figure 2. Image model

4.2.3. Audio voice preprocessing

The audio files must be prepared before anything else. An individual's distinct character in the file's third digit, which represents the emotion key, allows the sound to recognise different emotions. Seven distinct emotions: happy, disgust, scared, surprised, sad, angry, and neutral are part of the data set. Transform the unprocessed audio samples into a suitable format (such as WAV. files). Edit the audio samples such that Mel-frequency cepstral coefficients (MFCCs) are not present. To obtain a Mel-spectrogram from a speech signal, follow these steps and process is shown in Figure 3.

- Frame the signal: divide the speech signal into 40-millisecond frames with a 50% overlap between consecutive frames.
- Apply a window function: multiply each frame by a hamming window to reduce spectral leakage.
- Compute the fast Fourier transform (FFT): perform an FFT on each windowed frame to convert it from the time domain to the frequency domain.
- Filter the frequencies: apply 25 band-pass filters to the frequency-domain signal. These filters are designed to match the critical bandwidths of human auditory perception, with their center frequencies distributed according to the Mel scale.
- Logarithmic compression: apply logarithm function to the outputs of filters to suppress the dynamic range.
- Construct the Mel-spectrogram: arrange the logarithmically compressed filter outputs from each frame to form the Mel-spectrogram of the signal.

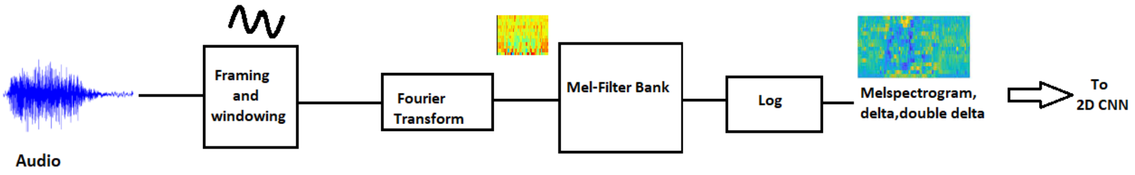

Figure 3. MFCCs model

5. METHOD

Our goal is to develop a multi-input model capable of processing text, images, and audio, while simultaneously analyzing the existing emotion recognition models in detail. The first steps of a fundamental technique were data collection and preprocessing. This research collects datasets from a variety of reputable sources, including academic journals, repositories, and organizations like hugging face, Inc., in order to validate the data's authenticity and relevancy. Previous researchers contributions are greatly appreciated. The citation of all sources used in this work is of the utmost importance to maintain academic honesty. Following the data collection phase, meticulous data cleaning and preparation took place. To ensure data integrity and quality, significant updates were made to the databases. The data preparation techniques utilized will be discussed in further detail in the portions that follow.

The method integrates CNN for image and audio data and LSTM with TF-IDF vectorization for text data aligning with the proposed system's multi-modal approach [24]. For image data CNN architectures are employed, leveraging datasets such as the emotion facial expression photographs dataset. These CNN models are trained to capture spatial features and patterns in facial expressions, facilitating accurate emotion classification from images.

Similarly, for audio data CNN architectures trained on datasets like the RAVDESS audio dataset are utilized. These models analyze spectrograms or other audio representations to identify emotional cues conveyed through speech. For text data the method utilizes LSTM networks coupled with TF-IDF vectorization. LSTM networks excel at capturing sequential dependencies in textual data, while TF-IDF vectorization transforms conversational text into numerical representations, capturing term importance. This combined approach enables the accurate classification of emotions from textual content.

The method ensures meticulous data preprocessing to enhance data quality and consistency across all modalities. Techniques such as data cleansing and enhancement are employed to address challenges associated with emotion recognition datasets. In the modeling phase, individual ML models are developed for each data modality, tailored to leverage the unique characteristics of text, image, and audio data. This includes designing CNN architectures for image and audio data and implementing LSTM networks with TF-IDF vectorization for text data.

A key feature of the research is the development of a multi-input model capable of accommodating text, image, and speech inputs. This model seamlessly directs inputs to appropriate sub-models based on their modality, enhancing emotion recognition accuracy across different data types. Efficiency and performance considerations are carefully balanced throughout the research design, ensuring accuracy while maintaining practicality. Overall, the method aligns closely with the proposed system's objectives and approach, facilitating a comprehensive study of emotion recognition across text, image, and audio modalities [25]. By contributing to the development of emotion detection algorithms that have a deeper understanding of human emotions, our work opens door to more sophisticated applications in affective computing and HCI.

6. IMPLEMENTATION

Building advanced emotion detection models capable of reliably identifying and classifying emotions across different modalities was the primary goal of our study. In this section, we detailed the steps involved in creating our model from selecting appropriate architectures and algorithms to fine-tuning hyperparameters. The goal of our research was to develop a set of models that could accurately anticipate how people will feel based on their interpretation of text, images, and audio inputs, utilizing cutting-edge technology and industry standards. Achieving state-of-the-art performance was the major objective of our modeling effort. This would propel the development of emotion recognition technology and its game-changing applications across several sectors.

6.1. Text emotion recognition

We looked at DL architectures as well as conventional ML models extensively for TER. As a foundation for TER, we began with logistic regression, a straight forward model that was later extended to handle multiclass problems. Our next stop was to CNN models, which are well-known for accurately depicting nuanced decision-making limits. We ran tests using a wide range of kernel functions and regularization parameters. We explored decision trees and random forests as part of our ensemble learning for TER. Next, we delved into DL to show how artificial neural networks (ANNs) can adapt to identify emotions. We used the TF-IDF vectorizer to convert textual input into numerical representations so that our models could process and assess the talks. This method was useful for highlighting key terms within the text's context, which led to the identification of seven emotions: surprise, anger, fear, disgust, neutral, joy, sorrow, and neutrality. Word cloud visualization helped us grasp the textual data better by graphically representing the most prevalent phrases associated with each emotion.

The conventional ML models included in the Scikit-learn package quickly proved to be underwhelming. First, we covered DL and its many models. The classic ANN is excellent at collecting complex data correlations, so we started with it. 1D CNNs are great at finding little patterns in text, so we looked at them next. Utilizing LSTM networks, we were able to capture sequential dependencies. The lack of computational efficiency provided by gated recurrent units (GRUs) in comparison to LSTM models was counterintuitive to our expectations. Bidirectional long short-term memory (Bi-LSTM) and other bidirectional models have shed light on the significance of contextual bidirectional processing. We decided to try out hybrid architectures after our research revealed that all neural networks functioned equally. The idea behind these bidirectionally integrated models was to capitalize on their complementary strengths. Furthermore, a complex model including CNNs, LSTMs, and Bi-LSTMs was produced as a consequence of the merging of these two types of models, which sought to extract textual features that were both local and global. But these hybrid designs whereas effective as the original versions. Consistent results were achieved using neural network models that were trained and optimized for TER. On average, these models have a 79% success rate when it comes to detecting complicated patterns in textual data and subtle emotional indicators. Accuracy, precision, recall, and F1-score, among other industry-standard metrics, were used to conduct a comprehensive evaluation of the model. Visual representations of these results allowed for an understandable comparison assessment of the models' performance in the complicated task of emotion identification from textual input. The findings of this study will pave the way for fruitful discussions and decisions about text emotion identification techniques.

6.2. Word embedding

The goal of word embedding is to create real-valued vector representations of words in a specified vector space. Compared to such a dense representation, sparse word representations are plainly worse due to their high dimensionality. The word embedding architecture is shown in Figure 4.

Transforming words into dense vectors, word embedding algorithms (see Algorithm 1) like Word2Vec facilitate various NLP tasks including text classification and emotion detection. Here is the equation of word embedding as shown in (1):

$$\theta_{w,k+1} = \theta_{w,k} - \eta \cdot \frac{\partial L}{\partial \theta_{w,k}} \qquad (1)’$$

where $\theta_{w,k}$ represents the embedding vector of word ‘w’ at iteration k.

'η' is the learning rate.
'L' is the loss function.
'$\frac{\partial L}{\partial \theta_{w,k}}$' is the gradint of the loss function with respect to the embedding vector.

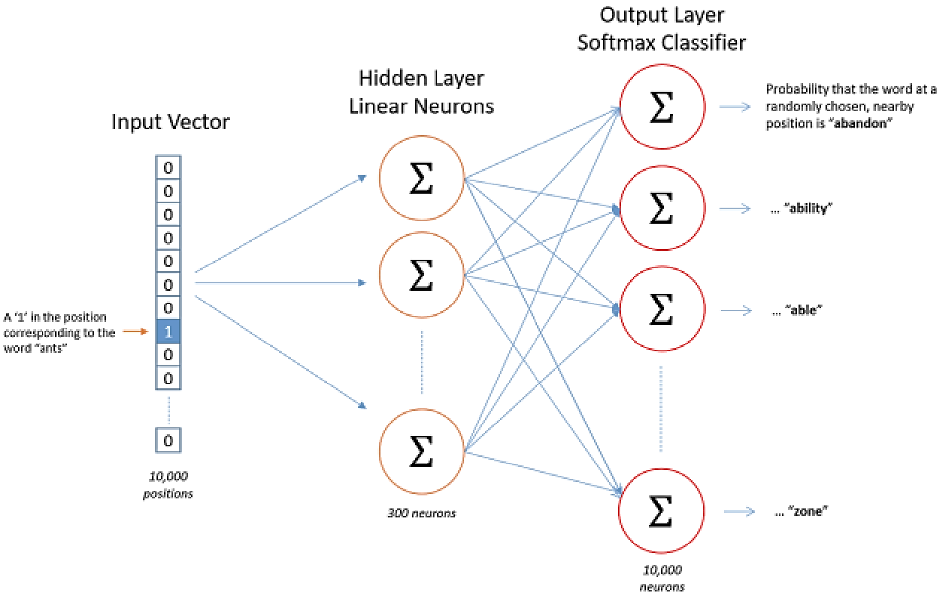

Figure 4. Word embedding model

Algorithm 1. Word embedding algorithm

```
Input: Corpus C; Target word w_I and the contextual word set {w_n}^{N}_{n=1}.
Output: Tailored word embedding of w_I.
Train θ and β with the corpus C.
for each word w_n∈ {w_n}^{N}_{n=1}, n ≠ I do
    for k∈ (1, . . . , K) do
        Update θ_{w_n, k} with Eq. (1).
    end for
end for
for k∈ (1, . . . , K) do
    Update θ_{w_I, k} with Eq. (1).
end for
```

Using these vectors we may find the semantic relationships between words. For example, they pick up on the "male-female" connection between "man" and "woman" as well as "king" and "queen". To top it all off, they can recognize relationships between different tenses of verbs like "walk", "walks", "walked", and "walking".

Our LSTM model's word vectors are trained in conjunction with the neural network and are first constructed at the embedding layer. Since we used an embedding size of 400, each word is represented in a 400-dimensional vector space. To get the text ready, remove any unnecessary characters and replace them with regular characters. This includes tab characters, newline characters, excessive spaces, and other similar characters.

Tokenizing the data is done using the spaCy software. Since spaCy does not have a parallel/multicore tokenizer, we resort to using the fast.ai package. When compared to its serial counterpart, the parallel spaCy tokenizer is much faster because it makes use of all CPU cores. Separating the text into individual tokens and assigning a unique index to each token is the main objective of tokenization. In order for our model to work, we need to convert the text into integer indices. The next step is numericalization, which involves turning tokens into numbers. This process comprises:

- Compiling an ordered list of all the words appearing in the text.
- Changing the word index in that list to match each one.

Our focus is not on the whole vocabulary list since singular words are meaningless and easily missed. A word has to appear twice or more times before it may be spelled properly. The model cannot be taught using infrequent phrases. Stacking architectures in DL, such as LSTM-CNN (see Algorithm 2), leverage the distinct capabilities of sequential data handling and spatial patterns processing, respectively, to combine the characteristics of both LSTM networks and CNN. Here is the equation of LSTM and CNN Algorithm:

Algorithm 2. LSTM and CNN stacking architecture

```
Require: Dataset O={(a1,b1),(a2,b2),...,(aU,bU)}
Base classifiers: LSTM, CNN;
Meta-classifier: Logistic Regression;
Preprocess the data O to form dataset 1D.
Choose both base classifiers, LSTM and CNN as well as the meta-classifier and Logistic
Regression.
Train the LSTM and CNN models using dataset 1D′, employing five-fold cross-validation.
Generate dataset Train using the trained LSTM and CNN models to estimate the probability
values φ^kLSTMand φ^k CNNrespectively.
Construct the stacking model and predict the result at the segment level ^k.
Obtain the result at the utterance level ^k using majority voting.
return k~;
output: The final result k~.
```

7. MODULE DESCRIPTION

This system is intended to identify emotions in three distinct forms of input: text, speech, and facial expressions. The main objective is to use CNN, a kind of deep learning model made to find patterns in data, to train distinct models to correctly identify emotions in these various sources of input. To start, datasets with text, audio, and facial photos representing a range of emotions including happiness, sadness, rage, and disgust are prepared. After preprocessing each dataset to guarantee consistency in size and quality, the data are labeled with the appropriate emotions. When the data is prepared, CNN models are trained on each form of data: text for written emotions, audio for voice, and photos for facial expressions. Following training, any new picture, audio file, or text input from the user can be used by the system to anticipate emotions.

- Upload facial emotion dataset: gathering images showing different facial expressions like happy, sad, angry, and disgust.
- Preprocess dataset: make sure they are all similar in size and quality. Takes each image a label saying what emotion is being shown (like "happy" and "sad").
- Train facial emotion CNN algorithm: it will learn to recognize patterns in the image that match different emotions. This learning process is called training and gets better accuracy with more training.
- Train audio emotion CNN algorithm: similar to the facial emotion training, but here will use recordings. It will learn to find patterns in sound that match different emotions like happiness and sadness.
- Train text emotion CNN algorithm: it will learn to recognize patterns in the text that match different emotions. This learning process is called training and gets better accuracy with more training.
- Accuracy comparison graph: it is shown in a graph, making it easy to see which accuracy is better.
- Predict facial emotion: in this module user will upload test image to detect emotion.
- Predict speech emotion: in this module user will upload test audio file to detect speech.
- Predict text emotion: in this module user will upload test text data to detect emotion.

8. RESULTS AND DISCUSSION

From sentiment graph shown in Figure 5 the count value for different sentients is: angry: 8,500; disgusted: 4,000; fearful: 2,000; happy: 5,500; neutral: 9,000; sad: 5,000; and surprised: 2,500. The dataset indicates a predominant presence of neutral sentiment, followed by significant counts of sad and angry emotions, reflecting a generally mixed sentiment with a slight leaning towards negative emotions. The blue line representing training accuracy shows how well the model performs on the training data, as time progresses it becomes better as shown in Figure 6. The validation accuracy (red line) demonstrates the model's ability to generalize to new data. It should ideally be very similar to the training accuracy with little to no variance. Monitoring both curves aids in assessing model performance and identifying overfitting.

The training loss curve (blue) represents how well the model fits the training data, decreasing over epochs as shown in Figure 7. The validation loss curve (red) mirrors this trend but indicates how well the model generalizes to unseen data; if it diverges significantly or starts increasing, it suggests overfitting. Comparing the two helps optimize model performance.

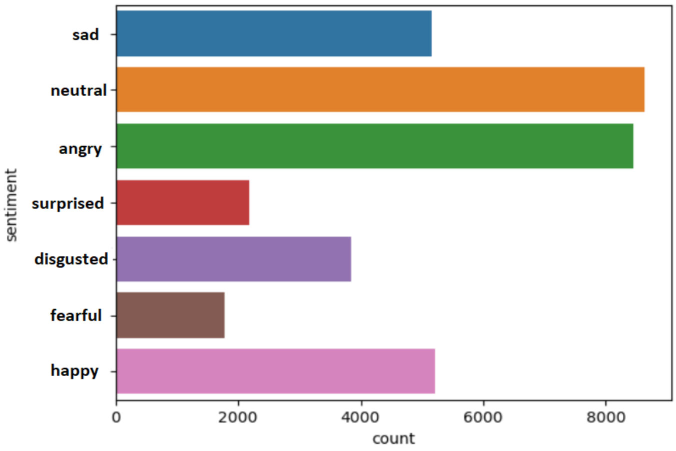

Figure 5. Sentiment graph

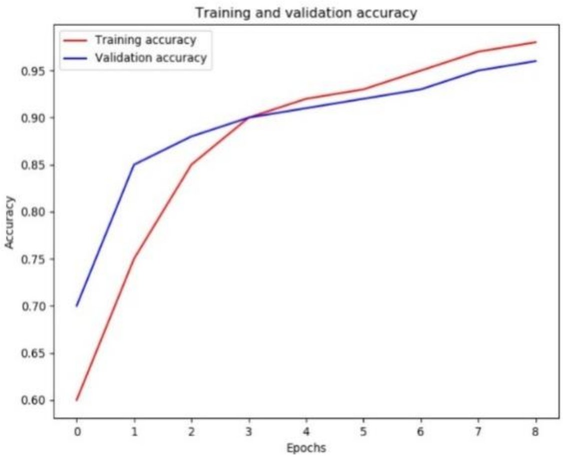

Figure 6. Training and validation accuracy graph

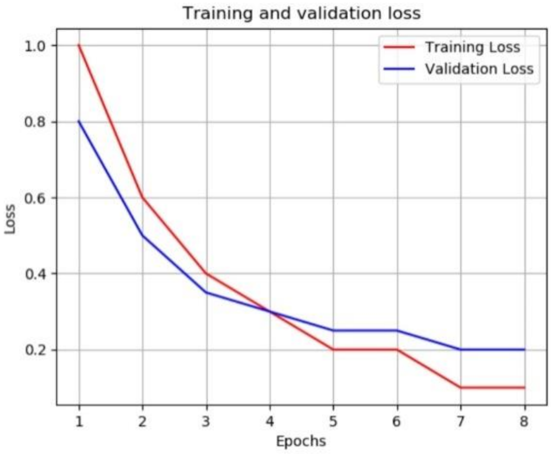

Figure 7. Training and validation loss graph

The confusion matrix provides insights into the model's performance across different categories, presenting precision, recall, F1-score, and support metrics as shown in Table 1. High precision (93%) and recall (92%) indicate accurate identification in one category, while slightly lower precision (95%) and good recall (84%) in another suggest few false negatives. Similarly, high precision (93%) and recall (95%) showcase successful identification, whereas lower precision (82%) and recall (82%) in a category indicate potential challenges. Meanwhile, high precision (96%) and recall (96%) signify excellence, while lower precision (74%) and high recall (91%) imply more false positives but effective identification. Image and audio models achieved the highest accuracy at 99%, show casing superior performance. Text models achieved slightly lower accuracy at 96%, still demonstrating strong performances shown in Table 2.

Table 1. Training performance

S. No.	Precision	recall	F1-score	Support
0	0.93	0.92	0.93	275
1	0.95	0.84	0.89	224
2	0.93	0.95	0.94	515
3	0.82	0.82	0.82	159
4	0.96	0.96	0.96	581
5	0.74	0.91	0.82	66
6	0.95	0.95	0.96	180
Accuracy			0.92	2000
Macro avg	0.89	0.90	0.89	2000
Weighted avg	0.93	0.92	0.92	2000

Table 2. Model and accuracy

No.	Model type	Accuracy
0	Text	0.96
1	Image	0.99
2	Audio	0.99

9. CONCLUSION

We want to extensively study the state-of-the-art emotion detection models and develop a multi-input model capable of interpreting text, images, and audio. The primary objective was to identify and categorize human emotions across several modalities using CNN for visual and auditory input and LSTM networks with TF-IDF vectorization for textual output. The suggested emotion recognition system uses a 2D CNN architecture to handle speech data and incorporates several fusion strategies, one of which is a novel CNN-based fusion strategy. Emotion detection systems are made more accurate and resilient by this method, which uses different ML models for each kind of input and guarantees careful data preparation. Whereas LSTM networks with TF-IDF vectorization were better at classifying emotions in text, CNN models were better at capturing the spatial features and patterns of facial expressions in images and audio recordings. The primary result of our study is a multi-input model that improves emotion recognition accuracy across different datasets by intelligently routing inputs to the appropriate sub-models based on their modality. The single model's adaptability and suppleness to various input types proved better despite some challenges like potential delays caused by several processing pipelines. Our approach is very congruent with the aims of the suggested system, which is to conduct comprehensive studies on emotion recognition in text, image, and audio formats.

In the future, we aim to evaluate the proposed method within an edge-and-cloud computing setup, employing text datasets, and the emotion in the wild challenge databases. Additionally, we are keen on exploring alternative DL architectures to ascertain if they yield improved performance. Taken together, our research paves the way for future emotion recognition algorithms with a deeper understanding of human emotions, which in turn will allow for more advanced uses in affective computing and HCI.

REFERENCES

[1] M. Chen, Y. Zhang, M. Qiu, N. Guizani, and Y. Hao, “SPHA: smart personal health advisor based on deep analytics,” *IEEE Communications Magazine*, vol. 56, no. 3, pp. 164–169, Mar. 2018, doi: 10.1109/MCOM.2018.1700274.

[2] F. Doctor, C. Karyotis, R. Iqbal, and A. James, “An intelligent framework for emotion aware e-healthcare support systems,” in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, Dec. 2016, pp. 1–8. doi: 10.1109/SSCI.2016.7850044.

[3] K. Lin, F. Xia, W. Wang, D. Tian, and J. Song, “System design for big data application in emotion-aware healthcare,” *IEEE Access*, vol. 4, pp. 6901–6909, 2016, doi: 10.1109/ACCESS.2016.2616643.

[4] X. Liu, L. Zhang, J. Yadegar, and N. Kamat, “A robust multi-modal emotion recognition framework for intelligent tutoring systems,” in *2011 IEEE 11th International Conference on Advanced Learning Technologies*, Jul. 2011, pp. 63–65. doi: 10.1109/ICALT.2011.26.

[5] R. A. Calvo and S. D’Mello, “Frontiers of affect-aware learning technologies,” *IEEE Intelligent Systems*, vol. 27, no. 6, pp. 86–89, Nov. 2012, doi: 10.1109/MIS.2012.110.

[6] K. Meehan, T. Lunney, K. Curran, and A. McCaughey, “Context-aware intelligent recommendation system for tourism,” in *2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, Mar. 2013, pp. 328–331. doi: 10.1109/PerComW.2013.6529508.

[7] S. Meng *et al.*, “Privacy-aware factorization-based hybrid recommendation method for healthcare services,” *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, pp. 5637–5647, Aug. 2022, doi: 10.1109/TII.2022.3143103.

[8] M. H. Shahbaz, Zain-Ul-Abidin, K. Mahboob, and F. Ali, “Enhancing contextualized GNNs for multimodal emotion recognition: improving accuracy and robustness,” in *2023 7th International Multi-Topic ICT Conference (IMTIC)*, May 2023, pp. 1–7. doi: 10.1109/IMTIC58887.2023.10178481.

[9] Y. Li, “Research direction of smart home real-time monitoring,” in *2020 International Conference on Computer Engineering and Intelligent Control (ICCEIC)*, Nov. 2020, pp. 220–232. doi: 10.1109/ICCEIC51584.2020.00051.

[10] Y. J. Liu, M. Yu, G. Zhao, J. Song, Y. Ge, and Y. Shi, “Real-time movie-induced discrete emotion recognition from EEG signals,” *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 550–562, 2018, doi: 10.1109/TAFFC.2017.2660485.

[11] J. Teo and J. T. Chia, “Deep neural classifiers for EEG-based emotion recognition in immersive environments,” in *2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*, Jul. 2018, pp. 1–6. doi: 10.1109/ICSCEE.2018.8538382.

[12] X. Li, X. Zhang, H. Yang, W. Duan, W. Dai, and L. Yin, “An EEG-based multi-modal emotion database with both posed and authentic facial actions for emotion analysis,” in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, Nov. 2020, pp. 336–343. doi: 10.1109/FG47880.2020.00050.

[13] P. Lopez-Otero, L. Dacia-Fernandez, and C. Garcia-Mateo, “A study of acoustic features for depression detection,” in *2nd International Workshop on Biometrics and Forensics*, Mar. 2014, pp. 1–6. doi: 10.1109/IWBF.2014.6914245.

[14] S. Gao, Y. Zhong, and W. Li, “Random weighting method for multisensor data fusion,” *IEEE Sensors Journal*, vol. 11, no. 9, pp. 1955–1961, Sep. 2011, doi: 10.1109/JSEN.2011.2107896.

[15] K. M. Kudiri, A. Md Said, and M. Y. Nayan, “Emotion detection using sub-image based features through human facial expressions,” in *2012 International Conference on Computer & Information Science (ICCIS)*, Jun. 2012, vol. 1, pp. 332–335. doi: 10.1109/ICCISci.2012.6297264.

[16] X. Wang, X. Chen, and C. Cao, “Human emotion recognition by optimally fusing facial expression and speech feature,” *Signal*

Processing: Image Communication, vol. 84, p. 115831, May 2020, doi: 10.1016/j.image.2020.115831.

[17] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: a review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019, doi: 10.1109/ACCESS.2019.2936124.

[18] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Cooperative learning and its application to emotion recognition from speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 1–1, 2014, doi: 10.1109/TASLP.2014.2375558.

[19] E. Mishra, A. K. Sharma, M. Bhalotia, and S. Katiyar, "A novel approach to analyse speech emotion using CNN and multilayer perceptron," in *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, Apr. 2022, pp. 1157–1161. doi: 10.1109/ICACITE53722.2022.9823781.

[20] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, Sep. 2013, pp. 511–516. doi: 10.1109/ACII.2013.90.

[21] J. Deng and F. Ren, "Multi-label emotion detection via emotion-specified feature extraction and emotion correlation learning," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 475–486, Jan. 2023, doi: 10.1109/TAFFC.2020.3034215.

[22] C. Dalvi, M. Rathod, S. Patil, S. Gite, and K. Kotecha, "A survey of AI-based facial emotion recognition: features, ML & DL techniques, age-wise datasets and future directions," *IEEE Access*, vol. 9, pp. 165806–165840, 2021, doi: 10.1109/ACCESS.2021.3131733.

[23] W. Li *et al.*, "A spontaneous driver emotion facial expression (DEFE) dataset for intelligent vehicles: emotions triggered by video-audio clips in driving scenarios," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 747–760, Jan. 2023, doi: 10.1109/TAFFC.2021.3063387.

[24] K. Kasiri, P. Fieguth, and D. A. Clausi, "Self-similarity measure for multi-modal image registration," in *2016 IEEE International Conference on Image Processing (ICIP)*, Sep. 2016, vol. 2016-Augus, pp. 4498–4502. doi: 10.1109/ICIP.2016.7533211.

[25] S. Z. H. Naqvi, S. Aziz, M. U. Khan, N. Asghar, and G. Rasool, "Emotion recognition system using pulse plethysmograph," in *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*, Mar. 2020, pp. 1–6, doi: 10.1109/ICETST49965.2020.9080725.

BIOGRAPHIES OF AUTHORS

Ravi Gummula is a Ph.D. scholar of electronics and communication engineering at the Dr. M.G.R. Educational and Research Institute in Chennai, Tamil Nadu, India. He obtained his M.Tech. in VLSI design in 2012 from Shadan College of Engineering and Technology and his B.E. in electronics and communication engineering from Deccan College of Engineering and Technology in 2007. He is a Life Member in The Indian Society for Technical Education, The Institution of Electronics and Telecommunication Engineers and International Association of Engineers. He has organised several national and international seminars, workshops, and conferences. He can be contacted at email: ravi.gummula@gmail.com.

Dr. Vinothkumar Arumugam is a Professor of electronics and communication engineering at the Dr. M.G.R. Educational and Research Institute in Chennai, Tamil Nadu, India. He obtained his Ph.D. in machine learning in 2017 and his M.Tech. in applied electronics in 2010 from Dr. M.G.R. Educational and Research Institute and his B.E. in electronics and communication engineering from Anna University in 2008. He received an M.Sc. in real estate valuation from Annamalai University in 2016. He is a Chartered Engineer and Member of the Institution of Engineers (India) and he is recognised as an approved Valuer and Member of the Institution of Valuers and a Member of various national and international professional societies. He can be contacted at email: dravinoth@gmail.com.

Abilasha Aranganathan is an Assistant Professor of electronics and communication engineering at the Dr. M.G.R. Educational and Research Institute in Chennai, Tamil Nadu, India. She obtained her M.Tech. in nanotechnology in 2014 and her B.E. in electronics and communication engineering in 2012 from Anna University. She has participated in several national workshops, seminars, and conferences. She has published several national and international journals and reputed publications. She is a Life Member in The Indian Society for Technical Education, The Institution of Electronics and Telecommunication Engineers and International Association of Engineers. She can be contacted at email: vabilasha90@gmail.com.