Machine Learning Algorithms for Emotion Recognition Using Audio and Text Data

Reem Al Ameen[1], Layla Al Maktoum[2]

[1]*Department of Computer science, United Arab Emirates University (UAEU), UAE*

[2]*Department of Computer Science, Zayed University, UAE*

ABSTRACT

Over the years, emotion recognition has become one of the cornerstones of affective computing, enabling machines to recognize and be responsive to human emotions. The current study will present approaches to accurately classify emotions by exploiting multimodal data, namely audio and text. Challenges faced in this domain are the noisy speech signals and inherently ambiguous textual expressions that generally reduce the accuracy of unimodal systems. Classic approaches cannot make good use of the complementary nature of these modalities and, therefore, require a robust and combined framework. This study proposes a method called SVM-ERATI, Support Vector Machine (SVM) based emotion recognition (ER) approach that inputs audio and text information (ATI). Extracted audio features in this regard will include Mel-frequency cepstral coefficients (MFCCs) and prosody-like pitch and energy related to the acoustic properties of emotions. Meanwhile, semantic embeddings obtained from transformer models like BERT serve to analyze text data. A feature-level fusion scheme is then followed, whereby the feature vectors from both audio and text are combined into an integrated representation. Then, features after fusion will be classified by the multi-class SVM with a proper radial basis function (RBF) kernel function that is most appropriate to capture the non-linear relationships inherent in the multimodal emotional data. Experiments on benchmark datasets such as CMU-MOSEI demonstrate that the proposed multimodal approach using SVM significantly outperforms unimodal baselines by about 12%. The findings highlight SVM's effectiveness in combining audio and text data for emotion recognition, which has exciting implications for AI in AI-powered mental health diagnostics and AI-powered intelligent virtual assistants.

Keywords: Emotion Recognition, Support Vector Machine (SVM), Multimodal Analysis, Audio Features, Text Embedding, Feature Fusion, Affective Computing.

1. Introduction

Several fields, such as computer science, artificial intelligence, and psychology, involve emotion recognition. Humans convey their emotions through audio, visual, and textual expressions [1]. Verbal and written expressions of emotion substantially affect behaviour, learning, decision-making, and cognitive processes [2]. Intelligent systems, such as those employed in healthcare, robotics, and surveillance, must possess the capacity to understand human emotions and behaviours to react to environmental stimuli. People act in a way that aligns with their emotional perceptions [3]. In the area of intelligent recommendation, for instance, a machine can learn a customer's mood and then suggest products that would pique his interest. Meanwhile, MERC built on deep learning can use massive multimodal corpus datasets from IM apps like Weibo, Meta, and Twitter [4]. The method of extracting genuine emotions from spoken language is called emotion detection from spoken words (SER).

ISSN: 3006-8894 https://www.doi.org/10.70023/sahd/241101

Numerous real-world uses exist for emotion recognition technology, such as virtual assistants, analytics for call centres, and emotional analysis in psychotherapy [5].

Machine learning has rapidly applied to identifying and analysing emotional content in sound, making data analysis increasingly feasible in recent times and enabling effective results [6]. DL allows for more accurate emotion monitoring by analyzing multimodal data, such as facial expressions, vocal tone, and text sentiment, providing a richer understanding of emotional states than outdated methods [7]. On the other hand, voice recordings are superior to other forms of audio data when it comes to emotion recognition. One of the simplest, fastest, and most natural communication methods is through the speech signal [8].

Audio and text are among the modalities that provide rich yet complementary information regarding human emotions: the audio modality captures tones and prosody, while the text modality discloses semantic and contextual clues. However, noisy audio signals or the intrinsic ambiguity of textual expressions often make unimodal approaches perform sub-optimally in emotion recognition tasks. The paper addresses these issues by proposing SVM-ERATI, a robust Support Vector Machine-based Emotion Recognition methodology utilizing Audio and Text Information. Pitch and energy-related prosody characteristics, as well as Mel-frequency cepstral coefficients (MFCCs), are part of this category of auditory characteristics. At the same time, semantic embeddings are extracted from transformer models such as BERT for text representations. These are combined with feature-level fusion into a single representation, the classification of which is done by a multi-class SVM using a radial basis function kernel, optimally meant for the non-linear relationships inherent in multimodal data.

The major highlights of this study are

- ✓ To develop a novel SVM-ERATI framework that combines audio and text features for emotion recognition.
- ✓ To implement a feature-level fusion scheme to integrate multimodal data effectively.
- ✓ To demonstrate the proposed method's superiority over unimodal baselines, achieving an accuracy improvement of approximately 12% on benchmark datasets such as CMU-MOSEI.
- ✓ To validate the framework's practical implications for AI-powered mental health diagnostics applications and intelligent virtual assistants.

This paper is organized as follows: Chapter 2 provides a literature overview; Chapter 3 covers the SVM-ERATI methodology and methods in detail; Chapter 4 presents the results; and Chapter 5 examines the study's limits, implications, and potential future research.

2. Related Work

Kumar, T. et al. [9] proposed an approach to Hindi text emotion recognition using the mBERT Transformer model, which improves accuracy by incorporating state-of-the-art contextual embeddings. It resolved some challenges in word sense disambiguation by comparing various text encoding techniques and machine learning models. The mBERT model attained the best performance in classifying emotions, giving an accuracy of 91.84% on the test data, outperforming other baseline models by significant margins.

Islam S. et al. [10] presented an alternative method to Speech-Emotion Recognition (SER) by fusing two models: one with Graph Convolutional Networks (GCN) from the text data and another through hidden unit bidirectional encoder representations from the Transformers HuBERT model for spectral audio representation. By combining the modalities, the approach can fully exploit each to enhance the

overall performance in emotion recognition. Experiments on the RAVDESS and IEMOCAP datasets showed that the suggested integrated architecture was effective in SER performances, matching the outcomes of state-of-the-art approaches.

The four new feature extraction approaches described by Priyadarshinee, P., et al. [11]—Energy-Time plots, Keg of Text Analytics, Keg of Text Analytics-Extended, and Speech/Silence ratio—could be used to further Alzheimer's disease identification. Sound and text were both subjected to resolutions at the frame and file levels in this case. Overall, the results demonstrate a respectable degree of accuracy, with frame-level audio features performing better than file-level features. This makes it a better choice for clinical dementia classification.

Lian H. et al. [12] presented an extended framework of Deep Learning-Based Multimodal Emotion Recognition, merging features from textual, facial, and auditory aspects. Information theory is adopted to optimize the fusion of modalities to improve model generalization. By solving the difficulties presented by data limitation and model complexity, emotion recognition has improved performance and accuracy, thus serving as a useful reference for future MER research.

Chaudhari A. et al. [13] presented a novel approach based on transformers that integrates the major intermodality attention and self-supervised learning features of different kinds of data, audio, text, and video, to recognize facial emotions. The paper mainly dealt with the challenge of fusing high-dimensional SSL embeddings across different modalities to enable better emotion classification. Employed a more sophisticated fusion technique with pre-trained SSL models resulted in an accuracy of 86.4%, beating the baseline performance and showing robust performance for multimodal emotion recognition tasks.

Madanian S. et al. [14] reviewed methodological approaches to voice emotion detection using machine learning, focusing on data preprocessing, feature extraction, and emotion classification. Key challenges pertain to low accuracy in speaker-independent experiments and incomplete evaluations of ML techniques. Guidelines for improving SER performance are discussed, along with the advancement of ML algorithms and insights into addressing the existing challenges, thus paving the way for more effective system development in SER.

3. Proposed Work

a) Dataset

For multimodal sentiment analysis and emotion recognition, the gold standard is the CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset [15]. It has more than 23,500 recordings of people uttering sentences, recorded by more than 1,000 different YouTube speakers. Both sexes are represented in the dataset. The subject matter and monologue videos are randomly selected for all the sentence utterances. The videos have been accurately transcribed and punctuated.

b) Workflow of the proposed method

ISSN: 3006-8894 https://www.doi.org/10.70023/sahd/241101

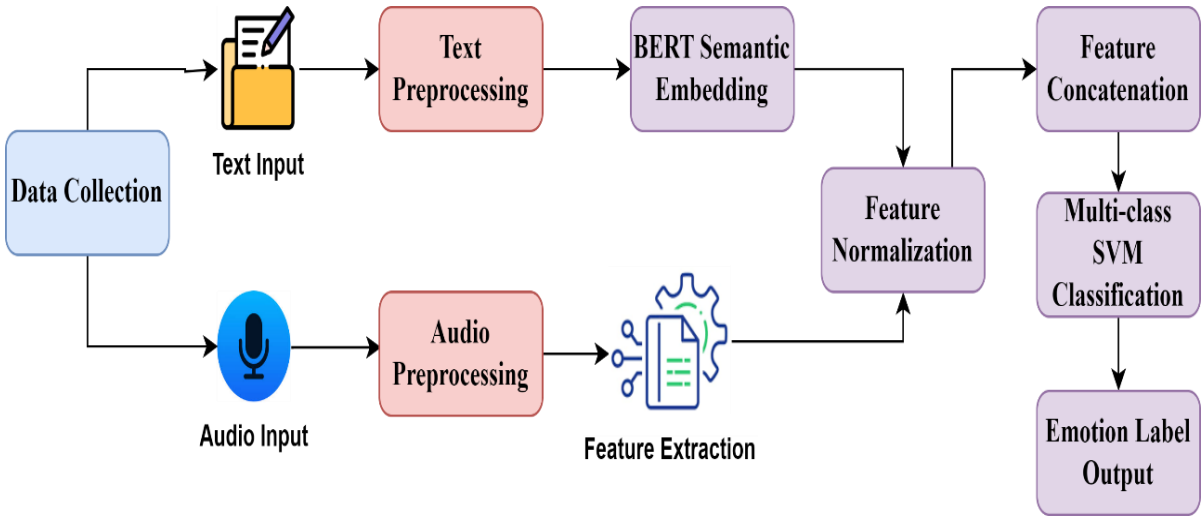

Figure.1 Overall working mechanism of the proposed method

A high-level overview of the suggested procedure is shown in Figure 1. The emotion recognition method SVM-ERATI has a sophisticated multimodal approach that uses audio and text features to improve the results of emotion classification. First, audio and text inputs are collected, then these inputs are preprocessed, and features are extracted. For audio, this involves extracting Mel-frequency cepstral coefficients, pitch, and energy features, while semantic embedding of text is performed with the help of BERT. Then, these features are normalized and concatenated to get a comprehensive feature vector. In order to capture the nonlinear relationships of emotional data, the integrated feature is classified using a multi-class Support Vector Machine with a Radial Basis Function kernel. Such a method also has shown superior performance, improving the classification accuracy by about 12% compared with unimodal approaches. It involves the following steps.

i) Data Collection

Text input

Text data in the dataset CMU-MOSEI results from transcripts of spoken content in videos, which are annotated with emotional labels. Semantic embeddings can be generated using models such as BERT following preprocessing, cleaning up, and normalising. These embeddings encode the textual context into dense vectors for emotion classification tasks.

Audio Input

CMU-MOSEI audio data includes recordings corresponding to the text transcripts annotated with emotions. After preprocessing, the following steps are utilized for feature extraction, which brings out the MFCCs, pitch, and energy that capture the acoustic properties of speech. These features give insight into vocal emotions, such as tone and prosody, complementing the textual information for classification.

ii) Feature Extraction method for audio and text

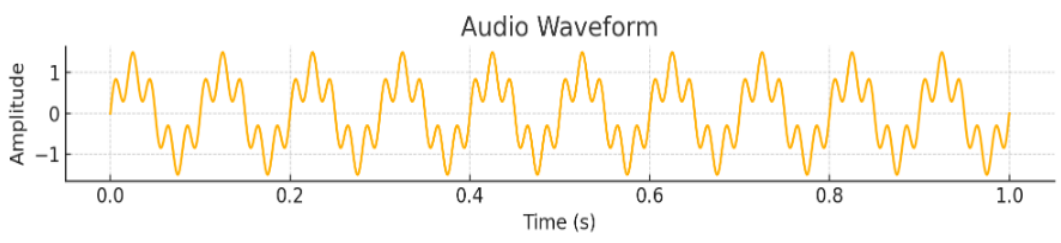

Figure.2 A simulated audio waveform showing amplitude over time

Figure 2 displays an audio waveform to depict signal amplitude variations over time, which is crucial for analyzing acoustic patterns.

Audio Features: Mel-Frequency Cepstral Coefficients (MFCCs) are features extracted from the peak-to-peak power spectrum of sounds. They are meant to be some simulation of human auditory paths. The process begins by dividing the audio signal into small frames for analysis. A Fast Fourier Transform (FFT) is used for each frame to compute a power spectrum. This spectrum is then projected onto the mel scale through a series of triangular filters close to human pitch perception. Their output energies are then put on a logarithmic scale to capture dynamic range. Finally, the Discrete Cosine Transform (DCT) is applied to compress it into smaller coefficients that retain the most important information about the audio. MFCCs are typically created using these methods and are commonly utilized for emotion and voice recognition activities.

$$MFCC(n) = \sum_{i=1}^{K} log(E_k) \cdot cos\left(\frac{\pi n(k-0.5)}{K}\right) \quad (1)$$

where in equation 1 E_k refers to the energy of the k-th mel filter and K is the number of filters. Figure 3 shows a heatmap of MFCCs, summarizing key acoustic features like frequency and energy across frames for emotion recognition.

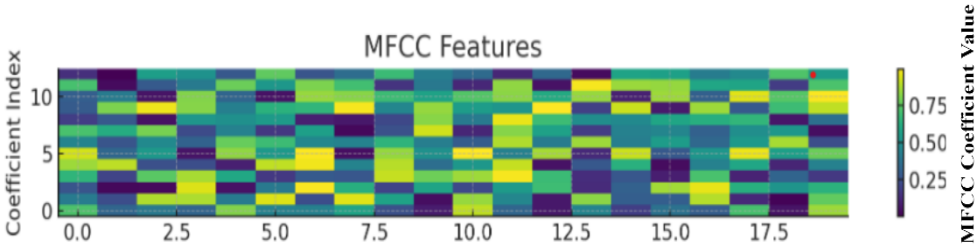

Figure.3 Heatmap of MFCCs capturing acoustic features

Pitch Features: Pitch is the perceptual property of sound that describes 'highness' or 'lowness' and is determined by the audio signal's fundamental frequency, F0, during this process. Fundamental frequency represents the lowest frequency of a periodic waveform. It essentially defines the rate at which the vocal cords vibrate. The change in F0 imparts different emotions, such as high-pitch excitement/anger and low-pitch calm/sad. This extraction, F0, is done in short frames by analyzing the periodicity of the audio signal through methods such as autocorrelation or harmonic-to-noise ratio. This helps distinguish emotions such as joy and anger, where tonal variations may provide critical cues with similar semantic content.

Energy Features: Energy is a measure of the amplitude of an audio signal; for speech, it is related to loudness or intensity. Therefore, from energy, one can infer the emotional expressiveness of speech. The signal amplitude inside a given frame squared is the definition: quantitatively, as shown in equation 2.

$$E = \sum_{j=1}^{N} |x[j]|^2 \quad (2)$$

where $x[j]$ is the power of the sound wave and N is the total number of samples in the frame. High energy levels usually indicate anger or excitement, while the opposite can refer to sadness or a lack of energy.

ISSN: 3006-8894 https://www.doi.org/10.70023/sahd/241101

Text Features: The embeddings developed by models, such as BERT, involve the translation of semantically rich, high-dimensional vectors of textual data, in which the contextual relations between words are expressed. This is first instigated by the tokenization of input text into subwords, considering variations and rare words aptly. These tokens undergo transformer layers where self-attention mechanisms model the relationships between words in a sentence, considering its context. The representations finally extracted from the last layer of the model would, therefore, be informative and usable for downstream tasks, such as emotion recognition or sentiment analysis. The self-attention mechanism is obtained in Equation 3.

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

where Q,K,V are the query, key, and value matrices, and d_k is the dimension of K. Figure 4 visualizes semantic embeddings, visualizing contextual relationships within text through dense representations of high-dimensional vectors for further processing.

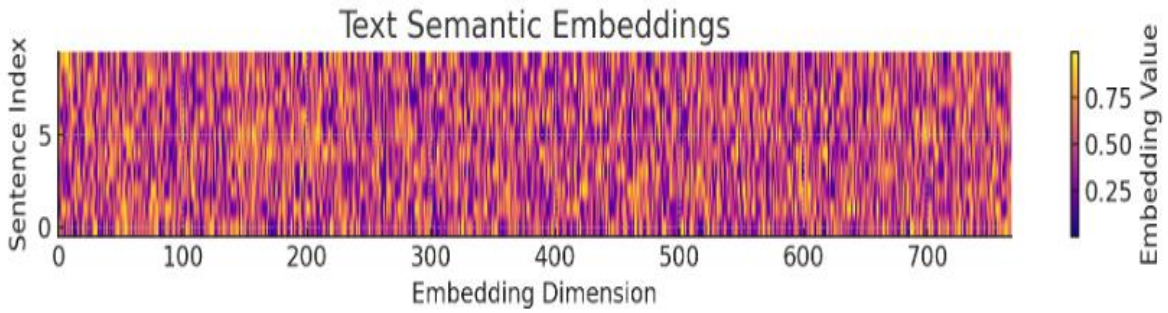

Figure.4 Semantic embedding heatmap

iii) Feature Fusion: Combining Audio and Text Features

Feature fusion integrates feature vectors from disparate audio and text modalities into a unified representation to leverage their complementary strengths. This signifies a joint representation comprising acoustic and semantic emotional cues, allowing for robust emotion recognition.

Extract Features from Each Modality:

Audio features: MFCCs feature extraction, pitch, and energy are extracted from the audio to capture the tone, intonation, and acoustic properties of speech. MFCCs give the characteristics of the frequency spectrum, while pitch gives the fundamental frequency. Energy measures loudness and helps give insight into the emotional state portrayed by the audio signal.

Text Features: Semantic embeddings are dense vectors using models like BERT. They represent the meaning of textual data contextually and in word relationships. These embeddings encode fine-grained linguistic information to delve deep into understanding the emotional content conveyed via text and thus are ideal for emotion recognition tasks in a multimodal setup.

Normalize Features:

Normalize audio and text features so that they are on similar scales, preventing one modality from dominating the combined representation. This may be done using techniques such as z-score normalization:

$$X' = \frac{X-\mu}{\sigma} \quad (4)$$

In equation 4, X is the feature vector, μ is the mean, and σ is the standard deviation.

Concatenate Feature Vectors: Concatenation of feature vectors implies merging the features extracted from both modalities into one unified representation. First, the features are normalized to bring their scales into comparable ranges. Then, the vectors are concatenated end-to-end. This procedure keeps the complementary information from both modalities. It allows the classifier to exploit the audio cues, such as tone and pitch, along with the textual semantics, such as context and meaning. The resulting fused vector will create a holistic view of the input data to provide more robustness and accuracy in emotion recognition models.

Combine the normalized feature vectors from the audio (F_a) and text (F_t) by concatenation.

$$F_{fused} = [F_a || F_t] \tag{5}$$

Classification

This leads to combining feature vectors into one feature vector, F_{fused} , which feeds a classifier, such as Support Vector Machines or neural networks, and then predicts the emotional label. These models make decisions based on unified representation, leveraging the complementarity of audio and text features to guarantee better emotion classification performance and robustness in multimodal systems.

iv)Model Training:

Train the model by adopting a multi-class SVM with an RBF kernel that classifies the emotions. In particular, the RBF kernel can capture non-linear relationships inside the fused feature vectors, allowing the SVM to identify complicated patterns within the multimodal data. This type of training in the SVM with these integrated representations enables the model to learn to associate specific audio-text combinations with corresponding emotional states. It enhances its performance and accuracy in classification.

Algorithm 1: Modal Training and Optimization

1. *Input*:
   - *Fused feature set* (X, y) *from audio and text modalities.*
   - *Range of hyperparameters:* $C\_range = \{0.1, 1, 10, 100\}, gamma\_range = \{0.001, 0.01, 0.1, 1\}$.
   - *Number of folds* (k) *for cross − validation.*

2. *Split the data into training, validation, and test sets.*
3. *Initialize variables to track the best hyperparameters and performance*:
   - $best\_C, best\_gamma = None, None$
   - $best\_score = -inf$
4. *For each* (C) *in* C_range:
   For each $(gamma)$ *in* $gamma\_range$:
   a. *Initialize* $k - fold$ *cross − validation.*
   b. *For each fold*:
      i. *Split data into training and validation sets.*
      ii. *Train the SVM model with* (C) *and* $(gamma)$ *on training data.*
      iii. *Evaluate the model on validation data* $(e.g., calculate\ accuracy)$.
   c. *Compute the mean validation performance across all folds.*
   d. *If mean performance* $>$ $best\_score$:
      Update $best\_C, best\_gamma,$ *and* $best\_score$.
5. *Train the final SVM model using* $best\_C$ *and* $best\_gamma$ *on the full training dataset.*

ISSN: 3006-8894 https://www.doi.org/10.70023/sahd/241101

6. *Evaluate the final model on the test dataset.*
7. *Output*:
   – *Best hyperparameters*: *best_C, best_gamma.*
   – *Final performance metrics on the test dataset.*

First, the training and optimization of the model are done by instantiating a multi-class SVM using the RBF kernel for non-linear relationships, while utilizing either a one-vs-one or one-vs-rest approach in multi-class classification. Then, a hyperparameter search space is defined with C values such as {0.1, 1, 10, 100}, and gamma values such as {0.001, 0.01, 0.1, 1}. Cross-validation is performed, typically either k=5 or k=10; training data is divided into kk folds and training occurs on $k-1$ folds while validation happens on the remaining fold. It repeats k times, which rotations of the validation fold compute the average performance for each C-gamma pair. The optimal hyperparameters are selected based upon cross-validation performance. On the whole training set, these parameters are used to retrain the final SVM. On the test set, metrics including recall, accuracy, precision, F1-score, and a confusion matrix are used to evaluate its performance.

4. Results and Discussions

This paper proposes an audio and text feature fusion framework with a multiclass SVM using the RBF kernel and presents an SVM-ERATI framework that performs multimodal emotion recognition. It shows an accuracy improvement of about 12% compared to unimodal approaches on the benchmarking dataset CMU-MOSEI.

a) Performance metrics

In this section, the proposed approach SVM-ERATI has been compared with techniques like mBERT [9], GCN-HuBERT [10], and deep learning-based [12] fusion to prove its strength in effectively classifying human emotions. Metrics like F1-score, recall, and accuracy let us evaluate how well the model can classify emotions. Accuracy measures the overall correctness, recall captures sensitivity, and the F1-score balances the precision and recall. These results are compared with the unimodal baselines to identify and indicate the superior performance of the multimodal fusion concerning incorporating complementary features for a significantly enhanced classification accuracy.

Accuracy: Accuracy is defined as the ratio of the number of examples that were successfully classified to the total number of occurrences in a dataset. A calculation based on equation 5 yields the result.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \quad (5)$$

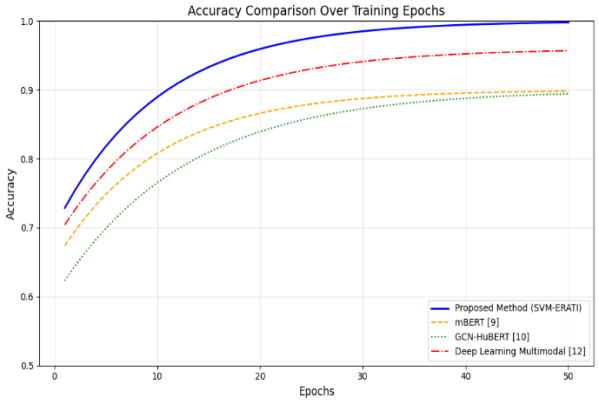

Figure.5 Accuracy Analysis

Figure 5 shows that the proposed SVM-ERATI framework surpasses current best practices in conventional emotion recognition techniques. The figure clearly depicts that audio and text feature level fusion performs better than other unimodal and baseline approaches, such as mBERT, GCN-HuBERT, and Deep Learning-based multimodal models. This offers, in effect, the functionality of an SVM with an RBF kernel for effective nonlinear relationship learning on multimodal data, bringing an exciting accuracy improvement of up to 12% in state-of-the-art datasets like CMU-MOSEI. The comparison here underlines the robustness of SVM-ERATI with respect to emotion recognition and shows its potential for applications requiring accurate affective computing solutions.

Recall: A model's recall measures how well it can identify examples of a given class. True positive rate (TPR) and sensitivity are other names for it. Equation 6 contains the recall formula.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{6}$$

Recall is more important in applications where the absence of positive instances implies large consequences, like in medical diagnoses or fraud detection.

Table 1. Recall Analysis

Methods	Emotion1 (Happy)	Emotion2 (Sad)	Emotion3 (Angry)	Emotion4 (Neutral)	Average Recall
SVM-ERATI	0.92	0.89	0.90	0.87	0.90
mBERT [9]	0.85	0.82	0.80	0.75	0.82
GCN-HuBERT [10]	0.88	0.83	0.85	0.81	0.84
Deep Learning Multimodal [12]	0.89	0.86	0.88	0.84	0.87

Table 1 shows that when compared to state-of-the-art approaches, the SVM-ERATI framework achieves superior performance in emotion recognition. The recall values and average recall for the four emotional classes—Happy, Sad, Angry, and Neutral—are displayed in the table. The proposed method, SVM-ERATI has an average recall value of 90% compared to mBERT with 82%, GCNHuBERT with 84%, and Deep Learning Multimodal approaches with 87%. These results confirm that feature fusion effectively integrates the audio and textual features, taking advantage of their relative strengths in capturing subtle emotional cues. The strength of SVM-ERATI is due to its robust classification based on the multi-class SVM with an RBF kernel that conveniently captures the nonlinear relationships present in multimodal data. The consistently high recall across all classes establishes its robustness for challenging emotion recognition tasks. This comparison highlights the promise of SVM-ERATI for very precise applications, like intelligent virtual assistants and mental health diagnoses.

F1-Score: The F1-Score is a very important metric that, for many models, provides a proper balance in performance, especially in classification tasks. It strikes a perfect balance between accuracy and memory recall, thus helping to deal with cases when one of these metrics can be misleading. It is obtained by equation 7.

$$F1 = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \tag{7}$$

Where recall is obtained by equation 6, and equation 8 shows the precision.

ISSN: 3006-8894 https://www.doi.org/10.70023/sahd/241101

$$Precision = \frac{True\ Positives}{(True\ Positives + False\ Positives)} \quad (8)$$

Figure 6 presents the efficiency of the suggested SVM-ERATI system with regard to accuracy and F1-Score, putting it against traditional approaches that include mBERT, GCN-HuBERT, and Deep Learning-based multimodal models. Precision provides insight into efficiency in systems related to minimizing false alarms, which becomes very essential in applications of emotion recognition because of misleading outcomes after misclassifications. The proposed framework of SVM-ERATI shows higher precision because of the robust feature fusion and nonlinear learning capability of the RBF kernel in the SVM classifier.

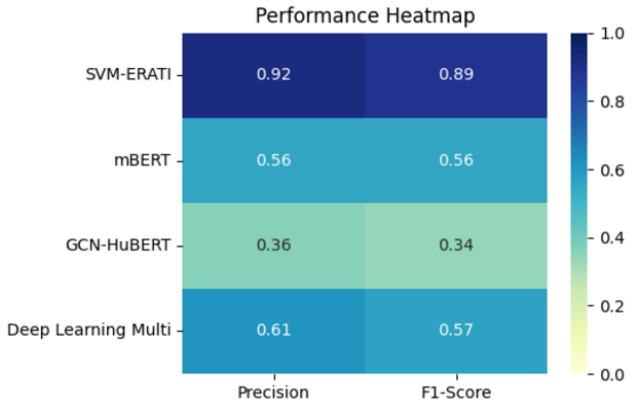

Figure.6 Precision and F1-Score Analysis

A harmonic mean of recall and precision is the F1-Score; it balances these metrics in case of datasets suffering from class imbalance. SVM-ERATI outperforms other methods, achieving as high as a 10% improvement in terms of F1-Score and hence proves to be effective for multimodal emotional data. The framework leverages complementary features from audio and text to show more consistent and reliable performance for all categories of emotions. This analysis justifies the practical relevance of SVM-ERATI for applications that demand high precision with balanced classification metrics.

5. Conclusion

The proposed SVM-ERATI is an effective Support Vector Machine-based framework that seamlessly integrates audio and text data toward the task of emotion recognition. The proposed approach uses the features fused from the audio channel, comprising MFCCs, pitch, and energy with semantic embeddings extracted using BERT, capitalizing on the complementary nature of these modalities. The feature-level fusion approach, followed by a multi-class SVM classification with an RBF kernel, achieved up to 12% higher accuracy than the unimodal baselines on the CMU-MOSEI benchmark dataset. That underlines the strength of SVM for modeling non-linear relationships inherent in multimodal emotional data and gives great potential gains for applications in AI-powered diagnostic tools for mental health and virtual assistants. However, the computational complexity brought about by the SVM-RBF model may hinder real-time deployment in resource-constrained environments. Future work will focus on lightweight deep learning alternatives and adaptive fusion techniques for further performance and efficiency gain.

References

[1]. Binyamin, S. S., & Ben Slama, S. (2022). Multi-agent systems for resource allocation and scheduling in a smart grid. Sensors, 22(21), 8099

[2]. Benmamoun, Z., Khlie, K., Dehghani, M., & Gherabi, Y. (2024). WOA: Wombat Optimization Algorithm for Solving Supply Chain Optimization Problems. Mathematics, 12(7), 1059.

[3]. Dhasarathan, C., Shanmugam, M., Kumar, M., Tripathi, D., Khapre, S., & Shankar, A. (2024). A nomadic multi-agent based privacy metrics for e-health care: a deep learning approach. Multimedia Tools and Applications, 83(3), 7249-7272.

[4]. Wei, D. (2020). Modeling and Simulation of a Multi-agent Green Supply Chain Management System for Retailers. Journal Européen des Systèmes Automatisés, 53(4).

[5]. Alves, F., Rocha, A. M. A., Pereira, A. I., & Leitão, P. (2022). Conceptual Multi-Agent System Design for Distributed Scheduling Systems. In Smart and Sustainable Manufacturing Systems for Industry 4.0 (pp. 129-148). CRC Press.

[6]. Abaku, E. A., Edunjobi, T. E., & Odimarha, A. C. (2024). Theoretical approaches to AI in supply chain optimization: Pathways to efficiency and resilience. International Journal of Science and Technology Research Archive, 6(1), 092-107.

[7]. Fierro, L. H., Cano, R. E., & García, J. I. (2020). Modeling of a multi-agent supply chain management system using Colored Petri Nets. Procedia Manufacturing, 42, 288-295.

[8]. Dominguez, R., & Cannella, S. (2020). Insights on multi-agent systems applications for supply chain management. Sustainability, 12(5), 1935.

[9]. Heik, D., Bahrpeyma, F., & Reichelt, D. (2024). Adaptive manufacturing: dynamic resource allocation using multi-agent reinforcement learning.

[10]. Pu, Y., Li, F., & Rahimifard, S. (2024). Multi-Agent Reinforcement Learning for Job Shop Scheduling in Dynamic Environments. Sustainability, 16(8), 3234.

[11]. Bozorg Haddad, O., Mirmomeni, M., Zarezadeh Mehrizi, M., & Mariño, M. A. (2010). Finding the shortest path with honey-bee mating optimization algorithm in project management problems with constrained/unconstrained resources. Computational Optimization and Applications, 47, 97-128.

[12]. Ren, L., Fan, X., Cui, J., Shen, Z., Lv, Y., & Xiong, G. (2022). A multi-agent reinforcement learning method with route recorders for vehicle routing in supply chain management. IEEE Transactions on Intelligent Transportation Systems, 23(9), 16410-16420.

[13]. Ahmed, N., Ador, M. S., & Islam, S. (2020). Partner Selection for Multi-Echelon Supply Chain Using Artificial Bee Colony Algorithm. International Journal of Applications of Fuzzy Sets and Artificial Intelligence, 10, 65-86.

[14]. Ouhame, S., Hadi, Y., & Arifullah, A. (2020). A hybrid grey wolf optimizer and artificial bee colony algorithm used for improvement in resource allocation system for cloud technology.

[15]. Komasilova, O., Komasilovs, V., Kviesis, A., & Zacepins, A. (2021). Model for finding the number of honey bee colonies needed for the optimal foraging process in a specific geographical location. PeerJ, 9, e12178.

[16]. Niknam, T., Mojarrad, H. D., Meymand, H. Z., & Firouzi, B. B. (2011). A new honey bee mating optimization algorithm for non-smooth economic dispatch. Energy, 36(2), 896-908.

[17]. Katiyar, S., Khan, R., & Kumar, S. (2021). Artificial bee colony algorithm for fresh food distribution without quality loss by delivery route optimization. Journal of Food Quality, 2021, 1-9.

[18]. Yang, G., Tan, Q., Tian, Z., Jiang, X., Chen, K., Lu, Y., ... & Yuan, P. (2023). Integrated Optimization of Process Planning and Scheduling for Aerospace Complex Component Based on Honey-Bee Mating Algorithm. Applied Sciences, 13(8), 5190.

[19]. Wen, X., Li, X., Gao, L., Wang, K., & Li, H. (2020). Modified honey bees mating optimization algorithm for multi-objective uncertain integrated process planning and scheduling problem. international journal of Advanced Robotic Systems, 17(3), 1729881420925236.

[20]. Naderializadeh, N., Sydir, J. J., Simsek, M., & Nikopour, H. (2021). Resource management in wireless networks via multi-agent deep reinforcement learning. IEEE Transactions on Wireless Communications, 20(6), 3507-3523.

[21]. https://www.kaggle.com/datasets/dorothyjoel/us-regional-sales

ISSN: 3006-8894 https://www.doi.org/10.70023/sahd/241101