Optimizing Emotional Insight through Unimodal and Multimodal Long Short-term Memory Models

Hemin F. Ibrahim[1]*, Chu K. Loo[2], Shreeyash Y. Geda[3] and Abdulbasit K. Al-Talabani[4]

[1]Department of Information Technology, Tishk International University,
Erbil, Kurdistan Region - F.R. Iraq

[2]Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, Universiti Malaya,
Kuala Lumpur, Malaysia

[3]Indian Institute of Technology Roorkee, Roorkee,
Uttarakhand, India

[4]Department of Software Engineering, Faculty of Engineering, Koya University,
Danielle Mitterrand Boulevard, Koya KOY45, Kurdistan Region – F.R. Iraq

***Abstract*—The field of multimodal emotion recognition is increasingly gaining popularity as a research area. It involves analyzing human emotions across multiple modalities, such as acoustic, visual, and language. Emotion recognition is more effective as a multimodal learning task than relying on a single modality. In this paper, we present an unimodal and multimodal long short-term memory model with a class weight parameter technique for emotion recognition on the CMU-Multimodal Opinion Sentiment and Emotion Intensity dataset. In addition, a critical challenge lies in selecting the most effective fusion method for integrating multiple modalities. To address this, we applied four different fusion techniques: Early fusion, late fusion, deep fusion, and tensor fusion. These fusion methods improved the performance of multimodal emotion recognition compared to unimodal approaches. With the highly imbalanced number of samples per emotion class in the MOSEI dataset, adding a class weight parameter technique leads our model to outperform the state of the art on all three modalities — acoustic, visual, and language — as well as on all the fusion models. The challenges of class imbalance, which can lead to biased model performance, and using an effective fusion method for integrating multiple modalities often result in decreased accuracy in recognizing less frequent emotion classes. Our proposed model shows 2–3% performance improvement in the unimodal and 2% in the multimodal over the state-of-the-art achieved results.**

***Index Terms*—Multimodal emotion recognition, Long short-term memory model, Class weight technique, Fusion techniques, Imbalanced data handling.**

ARO-The Scientific Journal of Koya University
Vol. XII, No. 1 (2024), Article ID: ARO.11477. 7 pages
Doi: 10.14500/aro.11477
Received: 25 January 2024; Accepted: 25 May 2024
Regular research paper: Published: 09 June 2024
Corresponding author's e-mail: hemin.ibrahim@tiu.edu.iq

I. Introduction

Technology in the 21st century has become widespread, dramatically transforming and significantly revolutionizing our way of life. Artificial intelligence (AI) has the potential to perform complex tasks that humans cannot do as quickly or as precisely as machines. However, adopting soft skills such as empathy, creativity, kindness, and caring for one another is just beginning (Paiva, et al., 2017).

Furthermore, emotions can significantly play an important role in various aspects of human daily life, including communication, comprehension, mutual assistance, and sometimes even decision-making. However, real human emotions are challenging to categorize, recognize, and analyze due to the differences among situations, cultures, and individuals (Angelov, et al., 2017).

The vocal modulations (speech) and facial expressions as visual data from the videos, in addition to textual data, provide significant cues to better identify emotions. Nonetheless, managing heterogeneous data in multimodal analysis necessitates a robust fusion strategy (Vijayaraghavan, et al., 2024) (Chen, et al., 1998).

The heterogeneity of multimodal data makes it challenging to build models that achieve information and not only capture complementary information. This explains why implementing proper fusion techniques is necessary to increase accuracy, learn the importance of each modality, and increase the reliability of estimation. This multimodal approach helps close the gap between technology and human interaction, potentially improving applications in customer satisfaction, mental health, and human-computer interfaces that utilize different modalities. By combining different data sources, this approach enhances the ability to understand and respond to complex emotional cues.

In this work, we present four different multimodal fusion techniques and evaluate their performance in detecting binary-class-based emotion recognition, using the CMU-Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset. To the best of our knowledge, the investigation of these three models under the proposed fusions has not been investigated for the emotion recognition application. In addition, we present the results from each modality separately and compare them with one another. The analysis aims to identify which fusion technique is most effective for this specific dataset. Our findings provide valuable insights for researchers and practitioners interested in multimodal emotion recognition.

Most classifications are facing difficulties in getting efficient accuracy and performing poorly when the training dataset is significantly imbalanced (Crangle, et al., 2019). Classification of binary imbalanced data, such as the CMU-MOSEI dataset, is a significant challenge in the field of machine learning (Yang and Wu, 2006).

Emotion classes in the CMU-MOSEI dataset used in this work are highly imbalanced; for example, the surprise emotion has a ratio of 1:10 (true class: false class), indicating that classification accuracy may not be efficient. To address this issue, we used the class weight parameter in our model, which assigns weight to the imbalanced classes during training by multiplying the loss of each sample by a specific factor depending on the class. The class weight technique helps focus the model's attention on samples from an underrepresented class. The use of class weights in the long short-term memory (LSTM) model demonstrated considerable efficacy, facilitating the model's ability to learn from underrepresented class samples without overly penalizing them, consequently improving performance across all classes.

II. Related Work

Over the last few years, the use of a multimodal approach has increased. In their review, (Jiang, et al., 2020) indicate that feature-level fusion is a common approach in multimodal models, where features from different types of data, such as text, audio, and images, are combined to create a single, unified representation. The processing of fusion between heterogeneity data and the variety of signals to support different channels through supplementary information fusion are significantly improved (Huang, et al., 2017). The survey paper by (Gladys and Vetriselvi, 2023) offers a detailed summary of emotion recognition that focuses on unimodal, such as visual, auditory, and linguistic, as well as multimodal emotion recognition. The paper examines the different ways to combine and represent information from multiple modalities.

Most previous works have primarily focused on combining features (concatenate features) from different channels into a single input to the network, which is called early fusion (EF) (Churamani, et al., 2018). In their research work, (Zadeh, et al., 2017) proposed a tensor fusion (TF) method to combine speech, facial expressions, and text, using a supervised learning method on the CMU-MOSI sentiment analysis dataset. In addition, the authors have proposed an end-to-end fusion method for sentiment analysis that explicitly represents unimodal, bimodal, and trimodal interactions between different channels. In their work, (Zhang, et al., 2024) investigate the current multimodal datasets for emotion recognition, looking into both hand-crafted and deep learning-based algorithms for extracting features from audio, visual, and text data. In addition, they discuss different methods for combining features, in addition to approaches that fuse at the feature level, decision level, or model level.

Majority voting, classification scores, and Borda count fusion have been used by Griol et al. in different multimodal approaches for both acoustic and language channels, using a multilayer perceptron for the acoustic modality and an extreme learning machine for the language modality (Griol, Molina and Callejas, 2019).

Comparing late fusion (LF) and EF (as called by features-level and decision-level fusion) is studied in (Busso, et al., 2004) by using a dataset that has been recorded from an actress. The study showed that the performance of the multimodal emotion classifier was better than each of the unimodal systems. In addition, the study concluded that both fusion techniques have similar performance.

Some datasets have skewed class distributions, and most classifiers perform poorly on highly unbalanced datasets. Therefore, researchers are addressing this problem by oversampling and using the Synthetic Minority Oversampling Technique (Lotfian and Busso, 2019). Furthermore, other studies proposed Cycle Generative Adversarial Networks to generate extra data for minority classes in the training set for solving the imbalanced data (Zhu, et al., 2018). Another technique, that can be used for the same purpose is the class weight, which helps improve domain classification performance (Ahmed and Green II, 2024; Kim and Kim, 2018).

As a summary of the literature, studies have exposed that the performance of emotion recognition can be developed by using multimodal information fusion. Various techniques for fusion modalities refer to having different performances. Furthermore, there are numerous methods for addressing imbalanced data, which normally improve accuracy results in a more efficient manner.

III. Methods

A. Dataset

For our experiments, the CMU-MOSEI dataset (Zadeh, et al., 2018) has been adopted, which contains 23,453 annotated video segments from 1,000 distinct speakers and 250 topics that were all collected from different channels on YouTube.

The CMU-MOSEI dataset was annotated for both sentiment and six different emotions. The dataset is gender balanced by using the data provided by the judges (57% male to 43% female) and contains different topics from different personalities. The dataset contains the following emotion categorical labels: happiness, sadness, anger, fear, disgust, and surprise.

The dataset itself is split into a train and test set, where the train contains 15290 utterances and the test contains 4832 utterances and provides the {0,1} binary labels for each emotion class. However, all emotion classes in CMU-MOSEI except happiness are highly imbalanced. Table I shows that the emotion class labels are imbalanced, which affects efficiency and accuracy. That is why calculating the unweighted accuracy (UA) is necessary.

B. Feature Extraction

We adopted the same features for acoustic, visual, and language modalities that were provided in (Zadeh, et al., 2018). The extracted features for channels are as follows:

- Acoustic: The COVAREP software (Degottex, et al., 2014) has been used to extract 74 rich acoustic features. For each utterance, a set of acoustic features is extracted, including 12 Mel-frequency cepstral coefficients, pitch, energy, peak slope, maxima dispersion quotients (Kane and Gobl, 2011), and glottal source parameters (Drugman, et al., 2012).
- Visual: facial expression is one of the most important sources for detecting emotions (Ekman, Friesen, and Ancoli, 1980). The library Emotient Facet (Krosschell, 2017) is used to extract 35 visual features, including facial action units, facial landmarks, head pose, eye gaze, and head orientation features (Zhu, et al., 2006).
- Language: The Glove word embeddings (Pennington, Socher, and Manning, 2014) are used to extract word vectors from transcripts. The timing of word utterances is extracted and aligned with audio at a sound level using P2FA (Yuan and Liberman, 2008), which enables alignment between audio, video, and text.

C. LSTM and Class Weight

LSTM, as a deep learning model for supporting time series data is applied to both proposed unimodal and multimodal approaches to learning and predicting emotions. The LSTM technique is one of the most popular and powerful deep learning methods for time-series data, and unlike the traditional RNN, it can capture long-term dependencies (Sherstinsky, 2020). LSTM gets better performance as a part of recurrent neural networks (RNNs) by having memory cells and gates that protect the information for the long term (Li, Abdel-Aty, and Yuan, 2020).

The main advantages of the LSTM are the memory cells in its hidden layers, which are organized into memory blocks rather than traditional neuron nodes. A memory cell has four main elements: An input gate (i_t), forget gate (f_t), self-connected memory cells (g_t), and an output gate (O_t), as shown in Fig. 1 (Li, Abdel-Aty and Yuan, 2020). The cell state at time is computed using the following equations:

$$i_t=\sigma(W_{xi}x_t+W_{hi}h_{t-1}+W_{ci}c_{t-1}+b_i)$$

$$f_t=\sigma(W_{xf}x_t+W_{hf}h_{t-1}+W_{cf}c_{t-1}+b_f)$$

$$o_t=\sigma(W_{xo}x_t+W_{ho}h_{t-1}+W_{co}c_t+b_o)$$

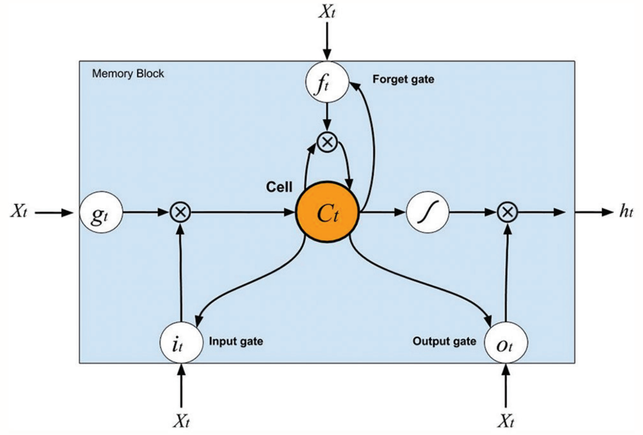

Fig. 1. The structure of long short-term memory cell (Li, Abdel-Aty, and Yuan, 2020).

$$g_t=tanh(W_{xc}x_t+W_{hc}h_{t-1}+b_c)$$

$$c_t=f_t\odot c_{t-1}+i_t\odot g_t$$

$$h_t=o_t\odot tanh(c_t)$$

Where σ is the gate activation function by using the sigmoid function, W represents weight matrices, c_{t-1} is the previous cell state, and $\odot$ denotes element-wise multiplication. However, the CMU-MOSEI dataset has some levels of class imbalance, except happiness class, as shown in Table I. To solve this issue, we used class weight as an additional parameter for weighting the loss function, which gives a weight to imbalanced classes by multiplying the loss of each sample by a certain factor based on their class. Class weights will be given by the following equation:

$$wi=\frac{n}{k\times ni}$$

Where w_i is the weight to class i, n is the number of samples, ni is the number of samples in class i, and k is the total number of classes.

This dictionary weight can be used directly to modify the loss function during the training time, by giving more class weights to the minority class and less class weights to the majority class, so that the learning dynamics of both classes remain the same. If we are assuming that L_1 and L_2 represent losses for the true and false classes, respectively, the total loss (L) can be calculated as follows:

$$L=\alpha\times L_1+\beta\times L_2$$

Where

L_1=-$y\times log(f(x))$

$L_2=-(1-y)\times log(1-f(x))$

$f(x)$ is the output of the final prediction in the dense layer and y is the class label. Binary cross entropy was used as loss for binary emotion classification. σ and β are class weights of the true and false classes, respectively.

TABLE I
Highly Emotion Classes Imbalanced in CMU-MOSEI Dataset, Only Happy Class is Balanced.

Emotions	Train set		Test set	
	True	False	True	False
Happy	53.3%	46.7%	52.2%	47.8%
Angry	22.5%	77.5%	20.1%	79.9%
Sad	25.5%	74.5%	27.6%	72.4%
Fear	8.6%	91.4%	6.9%	93.1%
Disgust	17.8%	82.2%	19.1%	80.9%
Surprise	10.2%	89.8%	9.9%	90.1%

In the validation label, we used UA (Tong, et al., 2017) which represents the actual performance, especially when the data areimbalanced in terms of sample sizes per emotion class. UA is computed as follows:

$$\text{Unweighted accuracy} = \frac{1}{|C|}\sum_{i=1}^{|C|}\frac{TP_i}{TP_i + FN_i}$$

Where i = 1,2,…, C introduces the number of emotion classes used, true positive refers to the number of positive samples that were recognized correctly as positive samples from the classification model.

IV. Unimodal and Multimodal

In this work, unimodal and multimodal approaches are proposed and evaluated using the specified dataset. First, we tried to check all the modalities, acoustic, visual, and language, separately and input their features into a specified LSTM for each modality, as shown in Fig. 2. In this study, we adopted various fusion approaches for multimodal emotion recognition using acoustic, visual, and language, such as EF, LF, deep fusion (DF), and TF.

EF: EF (Churamani, et al., 2018) as shown in Fig. 3, refers to the simple concatenation of acoustic (A), visual (V), and language (L) features and feeding them to the deep learning model. Features are fused at an early stage, hence the name EF. Concatenation was performed, on feature dimensions (35+74+300 = 409) and then the 409 features were used as input to the LSTM model. The training procedure was kept consistent with the unimodal approach to enable a comparison of the relative strengths of the models.

LF: LF (Busso, et al., 2004) as shown in Fig. 4, refers to the concatenation of 32 features that are extracted from the last layer of each LSTM unimodal for acoustic, visual, and language channels. The concatenated features feed two fully connected layers with several neurons equal to 32 and 1, respectively, which is shown in Fig. 4. This proposed fusion makes unimodal learn intramodal relations, while dense learns intermodal relations. The class weight is used for weighting the loss function during the training model.

DF: The DF approach (Fig. 5) is applied based on the work of (Nojavanasghari, et al., 2016). First, the pre-trained features from all three channels are used to give final prediction values, which are combined with their complementary values, then fed to the new dense, and classification is performed. Furthermore, the class weight used to tell the model takes more attention to these samples from an underrepresented population.

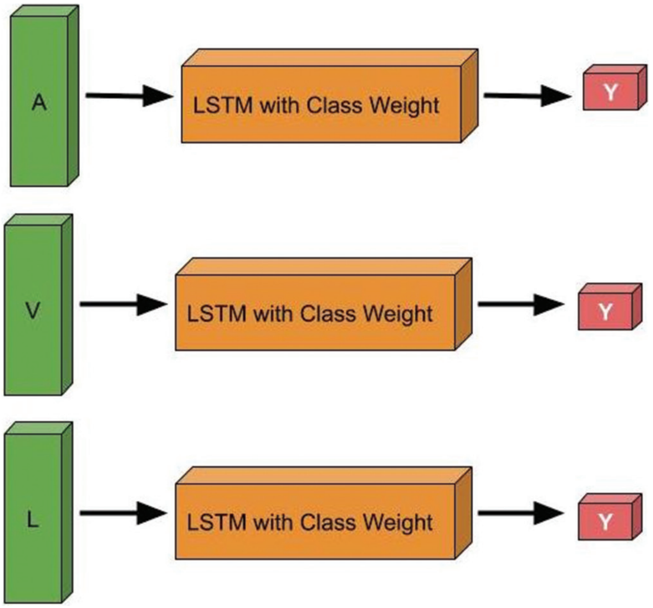

Fig. 2. The features from each unimodal (Acoustic [A], Visual [V] and Language [L]) are fed to the long short-term memory (LSTM)+CW using a 3-layer LSTM model with 1 dense layer and prediction.

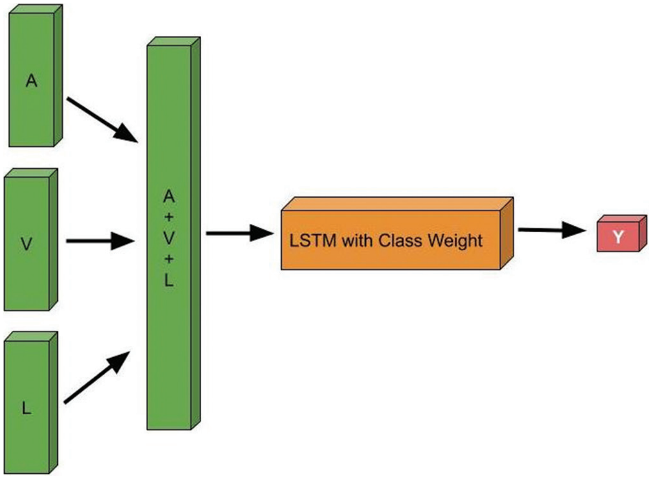

Fig. 3. Early fusion technique which refers to the direct concatenation of acoustic (A), visual (V) and language (L) features and feeding them to the long short-term memory+CW model.

After pre-training A, V, and L channels are represented as confidence scores of unimodal classifiers, and 1-A, 1-V, and 1-L are complementary values.

TF: Most of the fusion methods discussed so far use concatenation for fusion, but TF Network (Zadeh, et al., 2017) presents a new type of fusion method where they use the outer product of unimodally extracted features instead of direct concatenation, as shown in Fig. 6.

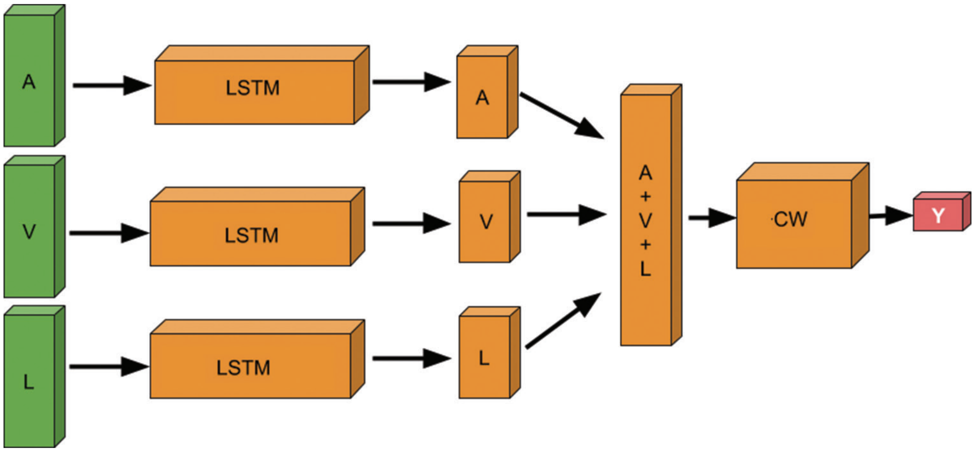

Fig. 4. Late fusion of the concatenation of features extracted from unimodal models for acoustic (A), visual (V) and language (L) and utilized a dense layer using these features for classification.

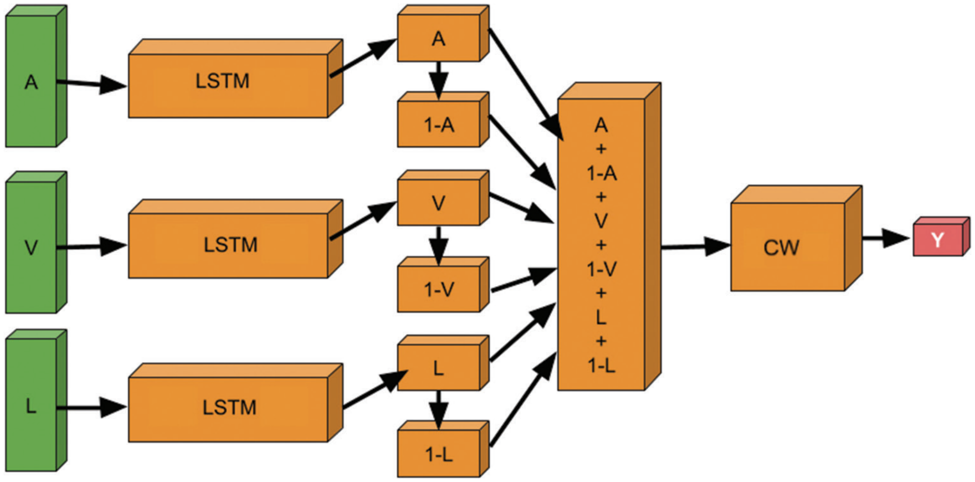

Fig. 5. Deep fusion, final prediction values from unimodal and complementary values are fed to dense.

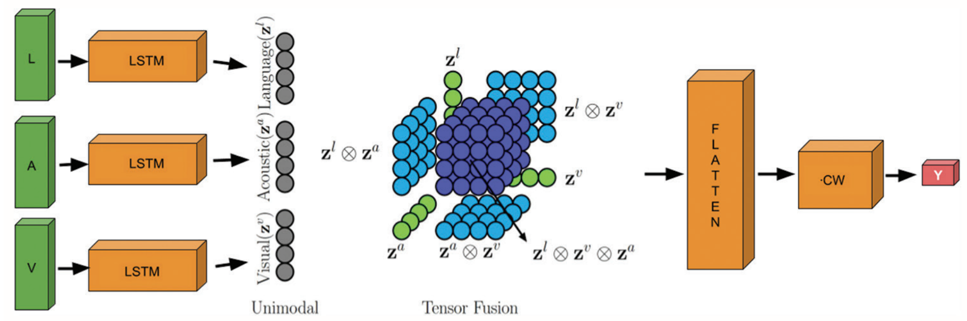
L
A
V
LSTM
LSTM
LSTM
Language($\mathbf{z}^l$)
Acoustic($\mathbf{z}^a$)
Visual($\mathbf{z}^v$)
$\mathbf{z}^l \otimes \mathbf{z}^a$
$\mathbf{z}^l$
$\mathbf{z}^l \otimes \mathbf{z}^v$
$\mathbf{z}^v$
$\mathbf{z}^a$
$\mathbf{z}^a \otimes \mathbf{z}^v$
$\mathbf{z}^l \otimes \mathbf{z}^v \otimes \mathbf{z}^a$
FLATTEN
CW
Y
Unimodal
Tensor Fusion

Fig. 6. Tensor fusion.

In Fig. 6, z^v, $z^{a,}$ and z^l are the pre-trained unimodal and they used the outer product between these features to find a correlation according to the formula:

$$z^m = \begin{bmatrix} z^l \\ 1 \end{bmatrix} \otimes \begin{bmatrix} z^v \\ 1 \end{bmatrix} \otimes \begin{bmatrix} z^a \\ 1 \end{bmatrix}$$

Where z^m is flattened and fed to a dense layer for the classification task. This way, they argue to learn better correlation among different modalities. Furthermore, class weight is used in the loss function to assign a higher value to these instances that are smaller.

V. Results and Discussion

In this section, we present the results of our model for both unimodal and multimodal approaches with different fusion approaches. In addition, to evaluate our results and have valid accuracy, we use UA as our metric. Our model outperforms many baselines and state-of-the-art (SOTA) methods for emotion recognition on the CMU-MOSEI dataset.

A. Unimodal

On the unimodal part, we trained and tested each channel separately, and our binary classification results for different emotions demonstrate better performance compared to SOTA methods. Table II shows the results for all three modalities, indicating that only the sad emotion class in the acoustic modality and the fear class in the visual modality did not outperform the SOTA (Zadeh, et al., 2018) with a difference of 1.55%, and 0.35%, respectively. However, all the other classes and the average of all modalities outperform the SOTA. The last row shows the average value over the six emotion classes, indicating that the proposed LSTM+CW outperforms the work of (Zadeh, et al., 2018) in the unimodal approach.

B. Multimodal and Fusion Approaches

As mentioned in Section III, we used four different fusion approaches with the LSTM and the class weight parameter to address the issue of imbalance. Our model, across all the different fusion approaches, was able to achieve better performance than the SOTA (Zadeh, et al., 2018). However, as shown in Table III, one can see that the EF approach outperforms other models in three emotions in addition to the average of all emotions.

Table III shows that our multimodal (acoustic [A], visual [V], and language [T]) UA using the four suggested fusions for binary classification of different emotions outperforms the SOTA (Zadeh, et al., 2018). The average of EF is higher than that of SOTA and the other approaches; also, TF performs competitively with EF. Three emotion classes, such as happiness, disgust, and surprise, in TF produce better results. The last row shows the average value over six emotion classes. We can clearly observe that DF and LF outperform SOTA, but they are still not performing as well as tensors and EF. Overall, for all three modalities (acoustic, visual, and language), LSTM and class weight for giving weight to the loss function, outperformed SOTA. Furthermore, all comparisons are based on UA metrics, because we handle imbalanced data directly.

TABLE II
Unweighted Accuracy (in Percentage) of CMU-MOSEI Dataset for Each Modality Based on Different Emotion Classes, the SOTA Refers to the SOTA from the CMU Multimodal Data SDK GitHub (Zadeh, et al., 2018)

Emotions	Acoustic		Visual		Language	
	SOTA	LSTM+CW	SOTA	LSTM+CW	SOTA	LSTM+CW
Happy	61.5	62.73	57.4	64.11	54	63.12
Angry	56.4	61.46	60	62.27	56.6	62.64
Sad	62	60.45	57.7	59.38	54	59.49
Fear	62.7	63.91	64.2	63.85	58.8	63.98
Disgust	60.9	71.26	60.3	70.52	64	68.61
Surprise	54.3	53.81	51.8	53.11	54.3	59.13
Avg	59.63	62.27	58.57	62.21	56.95	62.83

LSTM: Long short-term memory, SOTA: State-of-the-art

TABLE III
Unweighted Accuracy (in Percentage) of CMU-MOSEI Dataset Multimodal for 4 Different Fusion Approaches, such as DF, TF, LF, and EF Compared with the SOTA (Zadeh, et al., 2018)

Emotions	SOTA	DF	TF	LF	EF
Happy	66.3	64.43	67.57	66.48	67.44
Angry	62.6	63.32	64.15	63.71	66.46
Sad	60.4	59.78	61.52	61.61	62.68
Fear	62	64.29	61.87	58.43	65.14
Disgust	69.1	69.54	72.92	71.68	70.53
Surprise	53.7	57.41	58.99	58.45	56.64
AVG	62.35	63.13	64.5	63.39	64.82

DF: Deep fusion, TF: Tensor fusion, LF: Late fusion, EF: Early fusion, SOTA: State-of-the-art

VI. Conclusion and Future Work

In this paper, we present a deep learning architecture using LSTM to introduce an additional parameter, which is the class weight parameter, in the loss function that deals with the class distribution to handle the imbalance from the CMU-MOSEI dataset. LSTM and class weight parameters are used during training the model for unimodal and multimodal approaches with different fusion approaches.

Our model achieves better results in language modality compared to the other modalities, and the average of unimodal outperforms the SOTA UA.

Our study shows that EF in multimodal emotion recognition performs the best compared to other fusion approaches. The difference between all the fusion approaches is 1.5%, and the best-proposed fusion approaches outperform the SOTA by 2.47%. A limitation of this work that needs to be mentioned is that the proposed models have been validated on the binary version of emotions rather than the multiclass approach. For the next step in this work, different methods will be proposed to deal with imbalanced data, such as creating an end-to-end model. Furthermore, our model can be applied to sentiment analysis in the MOSEI dataset, and using more than one multimodal emotion dataset. A more advanced model can be tried (with class weights), such as unsupervised or semi-supervised learning methods, which may help in dealing with the problem of manual labeling of emotional data.

Acknowledgment

This study was supported by the Grand Challenge Grant-HTM (Wellness): GC003A-14HTM from the University of Malaya, IIRG Grant (IIRG002C-19HWB) from the University of Malaya and funding from the International Interfaculty Initiative in Computational Systems Care at Tokyo Metropolitan University, Japan.

References

Ahmed, J., and Green 2nd, R.C., 2024. Cost aware LSTM model for predicting hard disk drive failures based on extremely imbalanced S.M.A.R.T. sensors data. *Engineering Applications of Artificial Intelligence*, 127, 107339.

Angelov, P., Gu, X., Iglesias, J., Ledezma, A., Sanchis, A., Sipele, O., and Ramezani, R., 2017. Cybernetics of the mind: Learning individual's perceptions autonomously. *IEEE Systems, Man, and Cybernetics Magazine*, 3(2), pp.6-17.

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Lee, S., and Neumann, U., 2004. Analysis of Emotion Recognition Using Facial Expressions, Speech and Multimodal Information. In: *Proceedings of the 6th International Conference on Multimodal Interfaces*.

Chen, L., Huang, T., Miyasato, T., and Nakatsu, R., 1998. Multimodal Human Emotion/Expression Recognition. In: *Proceedings 3rd IEEE International Conference on Automatic Face and Gesture Recognition.* Nara, Japan.

Churamani, N., Barros, P., Strahl, E., and Wermter, S., 2018. Learning Empathy-Driven Emotion Expressions using Affective Modulations. In: *Proceedings of the 2018 International Joint Conference on Neural Networks* (*IJCNN*).

Crangle, C.E., Wanga, R., Perreau-Guimaraesa, M., Nguyena, M.U., Nguyena, D.T., and Suppes, P., 2019. *Machine learning for the recognition of emotion in the speech of couples in psychotherapy using the Stanford Suppes Brain Lab Psychotherapy Dataset*. Available from: https://arxiv.org/abs/1901.04110v1

Degottex, G., Kane, J., Drugman, T., Raitio, T., and Scherer, S., 2014. COVAREP - A Collaborative Voice analysis Repository for Speech Technologies. In: *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). Florence, Italy.

Drugman, T., Thomas, M., Gudnason, J., Naylor, P., and Dutoit, T., 2012. Detection of glottal closure instants from speech signals: A quantitative review. *IEEE Transactions on Audio Speech and Language Processing*, 20, pp.994-1009.

Ekman, P., Friesen, W.V., and Ancoli, S., 1980. Facial signs of emotional experience. *Journal of Personality and Social Psychology*, 39, pp.1125-1134.

Geetha, A.V., Mala, T., Priyanka, D., and Uma, E., 2024. Multimodal emotion recognition with deep learning: Advancements, challenges, and future directions. *Information Fusion*, 105, 102218.

Gladys, A.A., and Vetriselvi, V., 2023. Survey on multimodal approaches to emotion recognition. *Neurocomputing*, 556, p.126693.

Griol, D., Molina, J.M., and Callejas, Z., 2019. Combining speech-based and linguistic classifiers to recognize emotion in user spoken utterances. *Neurocomputing*, 326, pp.132-140.

Huang, Y., Yang, J., Liao, P., and Pan, J., 2017. Fusion of Facial Expressions and EEG for Multimodal Emotion Recognition. *Computational Intelligence and Neuroscience*, 2017, p.2107451.

Jiang, Y., Li, W., Hossain, MS., Chen, M., Alelaiwi, A., and Al-Hammadi, M., 2020. A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition. *Information Fusion*, 53, pp.209-221.

Kane, J., and Gobl, C., 2011. Identifying Regions of Non-modal Phonation Using Features of the Wavelet Transform. In: *Proceedings of the Annual Conference of the International Speech Communication Association*.

Kim, J.K., and Kim, Y.B., 2018. Joint Learning of Domain Classification and Out-of-Domain Detection with Dynamic Class Weighting for Satisficing False Acceptance Rates. In: *Proceedings of the Annual Conference of the International Speech Communication Association.*

Stöckli, S., Schulte-Mecklenbeck, M., Borer, S., and Samson, A.C., 2018. Facial expression analysis with AFFDEX and FACET: A validation study. *Behavior Research Methods*, 50, pp. 1446-1460.

Li, P., Abdel-Aty, M., and Yuan, J., 2020. Real-time crash risk prediction on arterials based on LSTM-CNN. *Accident Analysis and Prevention*, 135, p.105371.

Lotfian, R., and Busso, C., 2019. Over-sampling emotional speech data based on subjective evaluations provided by multiple individuals. *IEEE Transactions on Affective Computing*, 12, pp.870-882.

Nojavanasghari, B., Gopinath, D., Koushik, J., Baltrušaitis, T., and Morency, L.P., 2016. Deep Multimodal Fusion for Persuasiveness Prediction. In: *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. New York.

Paiva, A.M., Leite, I., Boukricha, B., and Wachsmuth, I., 2017. Empathy in virtual agents and robots: A survey. *ACM Transactions on Interactive Intelligent Systems*, 7, pp.1-40.

Pennington, J., Socher, R., and Manning, C.D., 2014. GloVe: Global Vectors for Word Representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*).

Sherstinsky, A., 2020. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Physica D: Nonlinear Phenomena*, 404, p.132306.

Tong, E., Zadeh, A., Jones, C., and Morency, L.P., 2017. Combating Human Trafficking with Multimodal Deep Models. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*).

Yang, Q., and Wu, X., 2006. 10 challenging problems in data mining research. *International Journal of Information Technology and Decision Making*, 5, pp.597-604.

Yuan, J., and Liberman, M., 2008. Speaker identification on the SCOTUS corpus. *The Journal of the Acoustical Society of America*, 123, p.3878.

Zadeh, A., Chen, M., Poria, S., Cambria, E., and Morency, LP., 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. Copenhagen, Denmark. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Zadeh, A.B., Liang, P.P., Poria, S., Cambria, E., and Morency, L.P., 2018. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*). Melbourne, Australia.

Zhang, S., Yang, Y., Chen, C., Zhang, X., Leng, Q., and Zhao, X., 2024. Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects. *Expert Systems with Applications*, 237, p.121692.

Zhu, Q., Yeh, M.C., Cheng, K.T., and Avidan, S., 2006. Fast Human Detection Using a Cascade of Histograms of Oriented Gradients. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (*CVPR'06*).

Zhu, X. Liu, Y., Li, J., Tao, W., and Qin, Z., 2018. *Emotion Classification with Data Augmentation Using Generative Adversarial Networks*. Springer, Cham.