EmAIT

http://publisher.uthm.edu.my/ojs/index.php/emait
e-ISSN : 2773-5540

Emerging Advances in Integrated Technology

Fuzzy Layered Convolution Neutral Network for Feature Level Fusion Based On Multimodal Sentiment Classification

Onasoga Olukayode Ayodele[1], Nor Hazlyna Harun[2*], Nooraini Yusoff [3]

[1]School of Computing, College of Arts and Sciences,
Universiti Utara Malaysia (UUM), Sintok, Kedah, 06010, MALAYSIA

[2]College of Arts and Sciences,
Universiti Utara Malaysia (UUM), Sintok, Kedah, 06010, MALAYSIA

[3]Department of Data Science,
Universiti Malaysia Kelantan (UMK), Kota Bharu, Kelantan, 16100, MALAYSIA

*Corresponding Author

DOI: https://doi.org/10.30880/emait.2022.03.02.007
Received 28 September 2022; Accepted 22 December 2022; Available online 31 December 2022

Abstract: Multimodal sentiment analysis (MSA) is one of the core research topics of natural language processing (NLP). MSA has become a challenge for scholars and is equally complicated for an appliance to comprehend. One study that supports MS difficulties is the MSA, which is learning opinions, emotions, and attitudes in an audio-visual format. In order words, using such diverse modalities to obtain opinions and identify emotions is necessary. Such utilization can be achieved via modality data fusion; such as feature fusion. In handling the data fusion of such diverse modalities while obtaining high performance, a typical machine learning algorithm is Deep Learning (DL), particularly the Convolutional Neutral Network (CNN), which has the capacity to handle tasks of great intricacy and difficulty. In this paper, we present a CNN architecture with an integrated layer via fuzzy methodologies for MSA, a task yet to be explored in improving the accuracy performance of CNN for diverse inputs. Experiments conducted on a benchmark multimodal dataset, MOSI, obtaining 37.5% and 81% on seven (7) class and binary classification respectively, reveals an improved accuracy performance compared with the typical CNN, which acquired 28.9% and 78%, respectively.

Keywords: Fuzzy, deep learning, CNN, MSA, fusion

1. Introduction

Sentiment Analysis (SA), also known as opinion mining, has become an essential human task. This is evident as the expression of opinions by humans is on the rise due to the invention of the internet where videos, vlogs, audio, and pictures serve as a medium for such opinion expression. SA typically involves the development of a model which classifies opinion into labeled polarities such as positive, negative, and or neutral classes [1]. These varied polarities necessitate that the raw data availabilities be utilized for mining opinions while also identifying their sentiments, as prior literature focuses on textual data, which may fail to generate accurate results [2]. An ideal source of such multimodal information is a video, which provides visual frames and information on spoken language's acoustic and textual representation [2, 4]. The integration of these varied data is referred to as Multimodal Sentiment Analysis (MSA). The use of more than a modality is known as multimodalities, such as bimodal, which uses two modalities, or trimodal which uses three modalities, learning articulated representation of multiple features [5], a crucial research direction to human-computer interaction [6].

The respective modality possesses exclusive features that can be mined collectively to obtain sentiments and opinions about the entity. In other words, multimodal may help enhance the performance of getting more precise results

*Corresponding author: hazlyna@uum.edu.my

by harmonizing relationships in modalities [7, 8]. The latest developments in MSA are seen in the data fusion of diverse modalities, hence generating better accuracy than a single modality [9]. The data fusion of these modalities common to MSA includes early or feature-level fusion and late or decision-level fusion. The feature level fusion extracts the different modality features and combines them to obtain a single feature vector. The decision level utilizes data on separate modalities and afterward fuses the individual modality results, giving a comprehensive decision at the end.

Additionally, MSA expressed by humans is multi-sensory and information processed via various areas in the cerebral cortex while being correlated by other brain areas. Thus, decisions made on sentiment/reviews are usually jointly dependent and not a unit-modal sensation. The learning process is similar to neural sensors, where weights are altered in a supervised neural network training strategy during the learning process. However, information distortion or fuzzy information can hinder the success rate for SA such that sentiment categories present confusion in the uni-modality category and in situations where there is no distinct boundary between categories. Hence, this study demonstrates the data fusion of multimodalities, integrating a fuzzy layer to improve model accuracy and a better generalization pattern. In addition, this study helps address the deficiency of vague information via integrating a fuzzy layer. Hence, the overlapping of sentiment clusters is avoided such that learned knowledge in uni-modality is crystal and prevents corrupt cluster boundaries which can be subjective. Furthermore, the humans who express their sentiments usually exhibit unsatisfactory knowledge, such as not making a significant agreement with the author's labeling of such expressed sentiment/opinion based on human voting on the datasets. However, this paper explores the feature-level fusion deploying methodologies fusion in providing better classification accuracy via the integration of fuzzy layer, which addresses issues on fuzzy information prominent in uni-modality and its subjective sentiment category boundaries.

Such improved accuracy, ascribed to Machine Learning (ML), involves automated learning from data in building analytical models. The specific attribute is learned from the training dataset and subsequently automates the performance of the task via test data. In the field of ML, Deep Learning (DL) has shown considerable success, making it the most prominent research trend. DL utilizes simultaneous transformations and graph technologies in building multi-layer models while the line between representation learning and predictive modeling is blurred [10, 11]. The Convolutional Neutral Network (CNN) is the fundamental model building block in DL. The main advantage of CNN is detecting significant features without human supervision compared to its predecessors.

Furthermore, the integration of methodologies has shown an improvement in model accuracy [12, 13, 14, 15, 16]. However, these approaches are either simultaneous or sequential paradigm learning [17]. In other words, the methods are in successive or parallel ways, which does not integrate the two methods utilizing their advantages [18]. These methods are indicative of DL fusion methodologies, in present times state of the art, deploy strategies at the decision level in aggregating pre-trained model outputs. Nonetheless, a proximity approach was conducted by [19], where a preliminary study was on fuzzy layer exploratory into DL with an emphasis on semantic segmentation via per-pixel classification.

Nevertheless, the approach deployed in this study is different as the emphasis is on multimodal data and its fusion for sentiment classification via the assimilation of fuzzy layer(s) within the CNN architecture. In addition, approaches taken via fuzzy and DL for various fusion application strategies/techniques are at the decision level in aggregating results from the state-of-the-art trained models [19]. This paper centers on the analytical comparison in the integration of fuzzy layer(s) within the CNN architecture for sentiment classification, with accentuates on model accuracy. Convolutions are performed on multimodal inputs on the CNN with and without the fuzzy layer. Fig. 1 illustrates the proposed method framework.

The rest of this study is structured as follows. Section 2 highlights correlated works explored using fuzzy and DL fusion strategies/techniques for classification improvement. The intuition of fuzzy layer integration with the CNN architecture is presented in section 3, while section 4 reveals experimental results conducted on a prominent multimodal dataset, and section 5 concludes this paper.

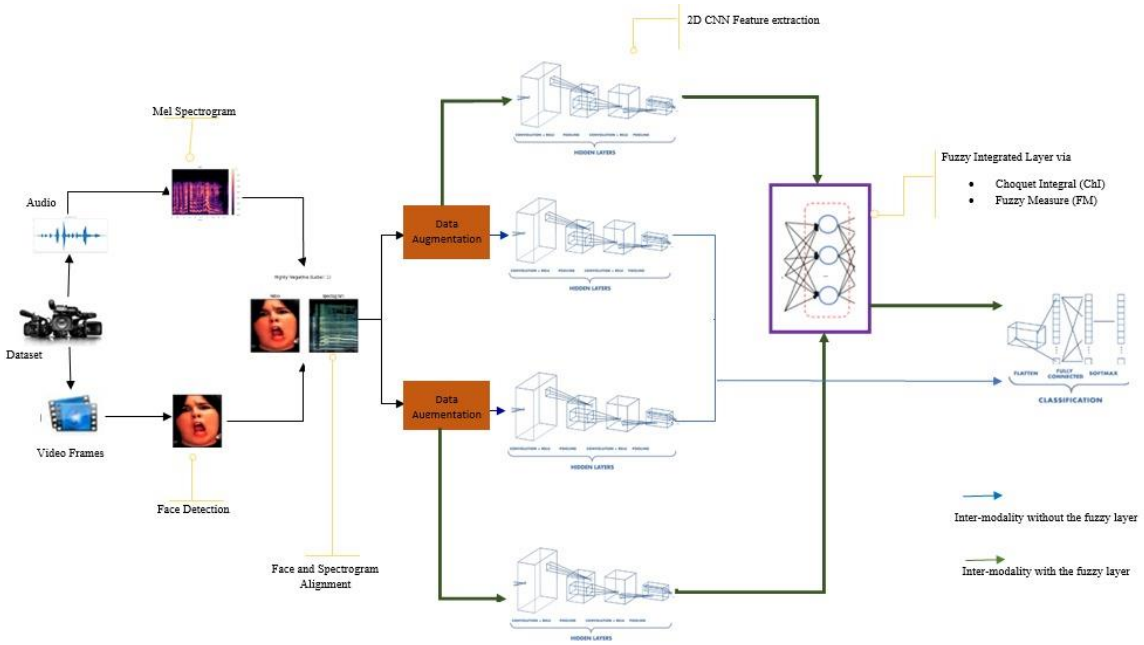

Fig. 1 - Proposed method framework

2. Related Work

Recently, DL has been regarded as the baseline standard for classification tasks. Results obtainable from DL are that of hyperparameter evaluation and network architecture deployed within specific data. Nevertheless, the fusion technique can be utilized in alleviating the comprehensive assessment to which the output of multiple DL models is combined hence utilizing the different DL model strengths. This data fusion involves the information combination from various sources having different extracted features and decisions. The commonly associated fusion strategy uses a fuzzy technique, for instance, Fuzzy logic [19, 20].

The deployment of fusion strategies has been utilized in providing improved classification performance. However, fuzzy strategies about DL architecture are limited [19, 21]. Though, a recent study by [19] presented an introduction of a fuzzy layer to DL architecture. The fusion of techniques involves comprehensively evaluating classification outputs from the classifiers. Thus, classifier strengths are merged by utilizing their advantages [22], where the optimum solution may not necessarily require ideal model parameters but is employed to aggregate models for improving classification performance. Research works conducted via technique fusion strategies, including strategies using fuzzy logic for improved classification performance, are presented in Table 1.

Table 1 - Summary of related works

Authors & Title	Year	Objective	Model
[23]	2021	Improved the quality of multimodal fusion utilizing the multimodal dataset	Multimodal contextual fusion model
[24]	2021	Improved fusion accuracy on text-related modality pairs, TV (text-visual) and TA (text acoustic)	Bi-Bimodal Fusion Network (BBFN on pairwise modality representations
[25]	2021	To model the collaboration of multiple pairs of bimodal and achieve the emotional prediction of multimodal features	A multi-tensor fusion network
[38]	2021	Enhanced sentiment analysis results on textual data	Fuzzy convolutional neural network model.

[15]	2021	Breast ultrasound (BUS) image semantic segmentation	Fully automatic segmentation algorithm.
[26]	2020	To provide means of extracting sentiment features from visual, audio, and text in terms of accuracy.	Multimodal sentiment analysis model via multi-head attention mechanism
[27]	2020	Incorporating fuzziness with DL for improved sentiment prediction on textual data.	Sentiment prediction via deployed LSTM, a type of Recurrent Neural Network (RNN) and Fuzzy Logic.
[12]	2019	Healthcare image data classification	fuzzy hybridized convolutional neural network (FCNN) model
[19]	2019	Fuzzy layer's introduction in the DL architecture based on road image semantic segmentation using per-pixel classification	Convolutional neural network (CNN, VGG16)
[28]	2019	To capture the dynamic nature of nonverbal intents by shifting word representations based on the accompanying nonverbal behaviors	Recurrent Attended Variation Embedding Network (RAVEN)
[29]	2019	Improve the classification capability when dealing with overlapped data	Based on Deep Neural Networks Fuzzy C-means clustering, fuzzy membership grades model
[16]	2019	Multiclass classification	Convolutional Fuzzy Neural Networks
[30]	2019	Enhanced image classification	Conventional Convolutional Neural Network (CNN)
[31]	2019	A method for improved control of non-linear systems	Deep fuzzy neural network (DFNN) framework
[32]	2019	Facial image classification of age and gender	Evolutionary-fuzzy-integral-based convolutional neural networks (EFI-CNNs)
[33]	2019	Improve diabetes detection accuracy using CNN and data fuzzification in matrix form	Fuzzy Convolutional Neural Network
[34]	2019	To obtain a topographic lake map based on intelligent sonar data processing	Fuzzy Logic and CNN
[35]	2019	A simultaneous model that depicts the spatial and temporal dependencies for passenger demand prediction	Convolutional Long Short-Term Memory network (ConvLSTM)

[36]	2019	Features hand-crafted based on fuzzy weighted multi-resolution depth motion maps (FWMDMMs) and Deep Learning.	Spatial-temporal Human action recognition (HAR) model
[37]	2019	Short-term load forecasting model development	A short-term load forecasting (STLF) method based on the fuzzy time series (FTS) and convolutional neural networks (CNN)
[38]	2019	To understand intelligent agents' decision explanation	A convolutional neuro-fuzzy network model

Although the approaches obtained improved accuracy, they fall within the simultaneous and sequential paradigm learning [17], i.e., models are in successive/parallel ways, which does not integrate the two models utilizing their advantages [18]. In contrast, this study embeds within the CNN architecture a fuzzy layer, which can efficiently handle imprecise and uncertain information.

3. The Model

The model deployed for this research is the two-dimensional (2D) CNN which learns features from the input in an image format. The concept of CNN was the inspiration of Hubel and Wiesel in 1962 via the hierarchical representation of neurons through their research based on the stimuli study of the visual cortex in cats. The study presented a fundamental breakthrough in understanding the visual cortex working in humans and animals. CNNs are stacks of multiple layers that learn features from input data to the desired output in a process termed End-to-End learning [10, 39, 40]. This learning process is such that the problem's prior knowledge task significantly lessens while minimizing engineering efforts, i.e., hand-crafted features. Thus, better performances are obtained from learned features rather than engineered ones (hand-crafted features).

3.1 CNN Architecture

A CNN architecture (see Fig. 2) is multiple layers stacked together, possessing distinct layers, including convolution, pooling, and fully connected layers. Feature extraction is the unique nature of the convolution layer that utilizes convolution operation in extracting high-level features via local receptive filters. Hence, a resultant feature map is obtained, which is completed via the convolution layer if the conforming filter to a particular feature map is slithered over the entire input matrix.

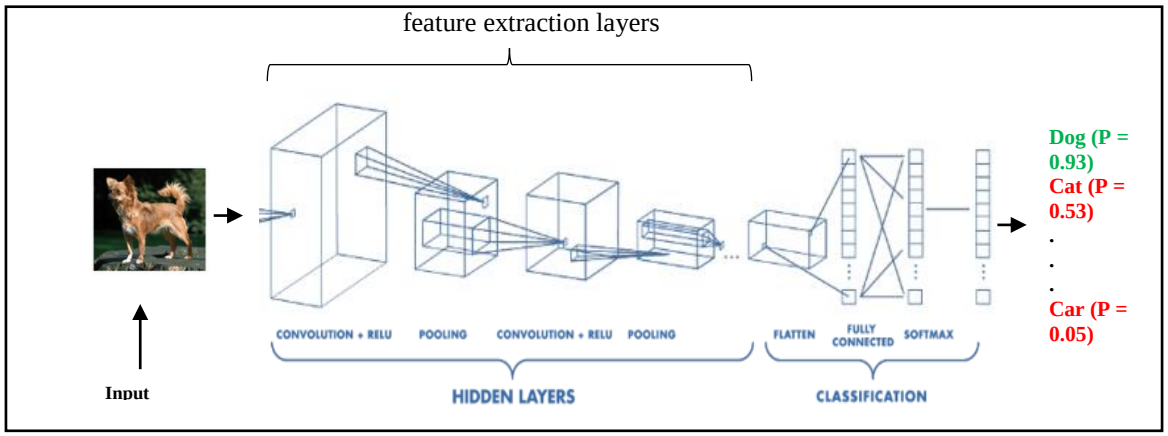

Fig. 2 - Typical CNN architecture

An operation referred to as downscaling proceeds after convolution by the pooling layer where the feature map spatial size is reduced. Downscaling helps in feature robustness against distortion and noise within the input signal. The primarily deployed downscaling function is the Max-pooling, which allows maximum activation of the local receptive filter to the ensuing layer. Hence, reducing the parameters learned by the network.

In this study, MSA will be explored in providing input fusion modality dynamics via the integration of FL in determining enhanced sentiment accuracy. The spectra of the enormous high dimensionality of multimodal data present the inevitability to investigate the usage of feature extraction. This helps to abridge data dimensionality and ultimately facilitates understanding which aspects of human expression (audio-visual) lead to a sentiment polarity and how they relate.

3.2 Convolution Layer

This layer is peculiar to the CNN algorithm and is devised to extract features from the input data, encompassing several convolution kernels. Every convolution kernel has a weight coefficient and bias vector akin to a neural network. The convolution kernel scans the input features consistently while multiplying and summing the matrix element of the input features in the receptive field and superimposes the bias value. The equation is as indicated in (1).

$$v^l = [v^{l-1} \otimes w^l] + b^l \tag{1}$$

where v^l and v^{l-1} is the output and input of the *l*th convolution layer respectively, b^l and w^l being the bias vector and weight coefficient of the lth convolution layer respectively. A typical convolution operation is shown in Fig. 3.

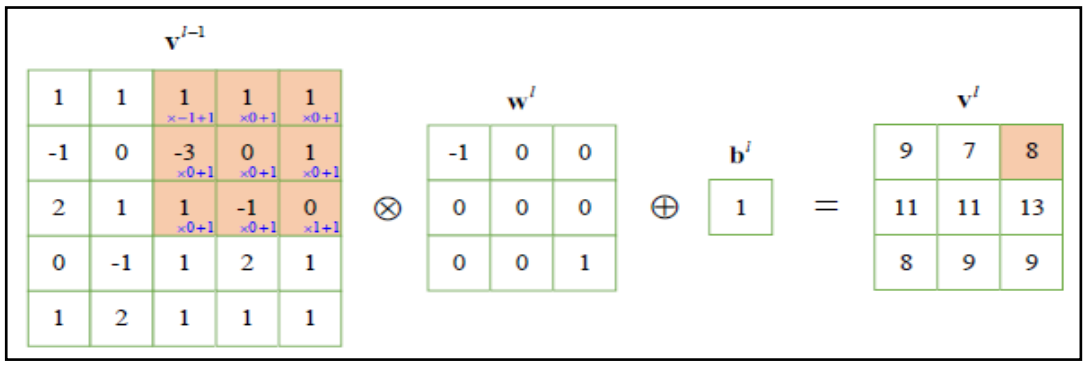

Fig. 3 - Convolution operation process

3.3 Pooling Layer

Also referred to as the sub-sampling layer, the pooling layer is usually after the convolution layer. When the convolution layer performs feature extraction, the output feature maps are conveyed to the pooling layer for information filtering and feature selection while maintaining equality between the input and output maps. However, the size of input feature maps is reduced owing to the sub-sampling operation represented by equation 2.

$$v^{l+1} = sub\ (v^l\ ) \tag{2}$$

where v^l and v^{l+1} are the input and output of the (l +1) *th* pooling layer, respectively and *sub* is the sub-sampling operation. The sub-sampling operation usually deployed in the CNN algorithm is the max-pooling or average-pooling operation (see Fig. 4.).

3.4 Fully Connected Layer

This layer is responsible for a non-linear sequence of extracted features from the convolution and pooling layer. However, the fully connected layer does not extract features but utilizes higher-order features in completing the learning objective. The fully connected layer is stated as presented in equation 3.

$$v^{l+1} = ful\ (h^{l+1}\ ) \tag{3}$$

where h^{l+1} and v^{l+1} are input and output of the (l +1) *th* fully connected layer and *ful* is the activation function in the layer.

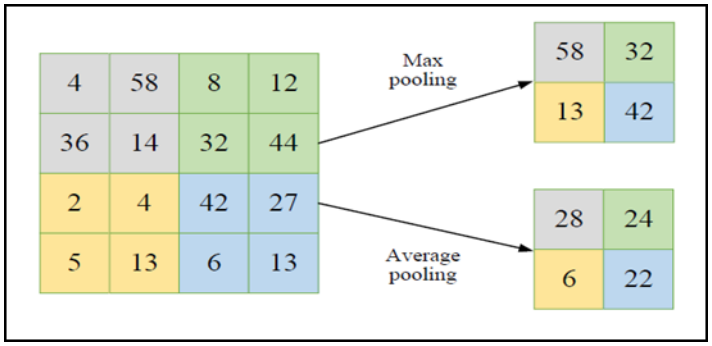

Fig. 4 - Pooling operation - max and average

The capacity and ability of CNN is the learning of apt features, via backpropagation, in reference to raw inputs in a supervised manner [40, 41]. Hence, optimal features are learned automatically while enabling generality capabilities in the perspective of emotional expression. To ensure an increase in the ability for stochastic gradient descent convergence, the ReLU layer is utilized as the activation function, unlike the sigmoid function which possesses non-saturating characteristics [42].

Presented in Table 2 are the architectural parameters. In addition, for emotion recognition about learned features, the SoftMax classifier was used. For an input $x(i,j)$ into a 2D convolution layer, the resultant $z(i,j)$ is achieved via the signal $x(i,j)$ convolved by the kernel $w(i,j)$ of size $a \times b$ as

$$z(i,\ j) = x(i,\ j) \times w(i,\ j) = \sum_{s=0}^{a-1} \sum_{t=0}^{b-1} x(s,\ t) \cdot w(i-s,\ j-t) \tag{4}$$

4. Fuzzy Layered 2D CNN Architecture

The fuzzy layered 2D architecture has two components; the CNN which is utilized for feature extraction from the input and the fuzzy components which learn complex intrinsic rules between the feature map and the resultant classification/sentiment results. In other words, the fuzzy component only processes feature maps and not raw input data to avoid variations that could occur in the input data which also considers various uncertainties. The architectural-end structure for the FMCNN is shown in Fig. 5. as compared to Fig. 2.

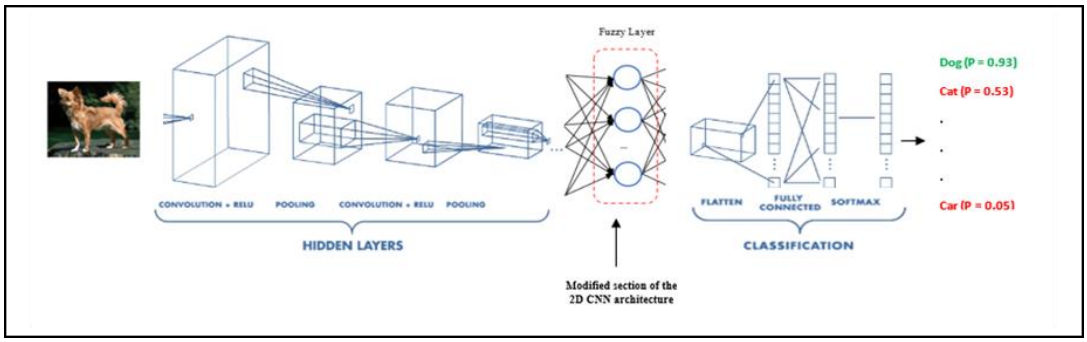

Fig. 5 - Architectural-end structure for the Fuzzy 2D CNN

The extracted feature at the convolutional layer serves as the input to the fuzzy layer, whose primary role is learning the intricate intrinsic relationship between the feature maps and the subsequent class labels. The corresponding video frame and spectrogram are paired and stacked in training the network. This training ensures that the network aligns extracted features side by side, as illustrated in Fig. 6.

Table 2 and Table 3 presents differences in the architectural parameters between the typical 2D CNN and the Fuzzy layered 2D CNN. The fuzzy layer utilized for this work as its focuses on the Choquet Integral (ChI) [43] a non-linear aggregation function that is parameterized by Fuzzy Measures (FM). Hence, parameterized ChI with FM provides an avenue for the combination of information encoded by FMs with the evidence of the input data.

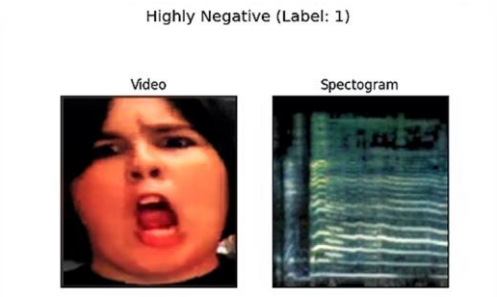

Fig. 6 - Feature alignment for extraction

Table 2 - Typical 2D CNN architectural layer parameters

Model: "sequential"

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 126, 116, 32)	896
batch_normalization (BatchNo	(None, 126, 116, 32)	128
max_pooling2d (MaxPooling2D)	(None, 63, 58, 32)	0
dropout (Dropout)	(None, 63, 58, 32)	0
conv2d_1 (Conv2D)	(None, 61, 56, 64)	18496
batch_normalization_1 (Batch	(None, 61, 56, 64)	256
max_pooling2d_1 (MaxPooling2	(None, 30, 28, 64)	0
dropout_1 (Dropout)	(None, 30, 28, 64)	0
conv2d_2 (Conv2D)	(None, 28, 26, 128)	73856
batch_normalization_2 (Batch	(None, 28, 26, 128)	512
max_pooling2d_2 (MaxPooling2	(None, 14, 13, 128)	0
dropout_2 (Dropout)	(None, 14, 13, 128)	0
conv2d_3 (Conv2D)	(None, 12, 11, 256)	295168
batch_normalization_3 (Batch	(None, 12, 11, 256)	1024
max_pooling2d_3 (MaxPooling2	(None, 6, 5, 256)	0
dropout_3 (Dropout)	(None, 6, 5, 256)	0
flatten (Flatten)	(None, 7680)	0
dense (Dense)	(None, 128)	983168
batch_normalization_4 (Batch	(None, 128)	512
dropout_4 (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 64)	8256
batch_normalization_5 (Batch	(None, 64)	256
dropout_5 (Dropout)	(None, 64)	0
dense_2 (Dense)	(None, 7)	455

Total params: 1,382,983
Trainable params: 1,381,639
Non-trainable params: 1,344

Table 3 - 2D Fuzzy layered CNN architectural layer parameters

Model: "sequential"

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 126, 116, 32)	896
batch_normalization (BatchNo	(None, 126, 116, 32)	128
max_pooling2d (MaxPooling2D)	(None, 63, 58, 32)	0
dropout (Dropout)	(None, 63, 58, 32)	0
conv2d_1 (Conv2D)	(None, 61, 56, 64)	18496
batch_normalization_1 (Batch	(None, 61, 56, 64)	256
max_pooling2d_1 (MaxPooling2	(None, 30, 28, 64)	0
dropout_1 (Dropout)	(None, 30, 28, 64)	0
conv2d_2 (Conv2D)	(None, 28, 26, 128)	73856
batch_normalization_2 (Batch	(None, 28, 26, 128)	512
max_pooling2d_2 (MaxPooling2	(None, 14, 13, 128)	0
dropout_2 (Dropout)	(None, 14, 13, 128)	0
fuzzy_layer (FuzzyLayer)	(None, 14, 13, 20)	5120
conv2d_3 (Conv2D)	(None, 12, 11, 256)	46336
batch_normalization_3 (Batch	(None, 12, 11, 256)	1024
max_pooling2d_3 (MaxPooling2	(None, 6, 5, 256)	0
dropout_3 (Dropout)	(None, 6, 5, 256)	0
flatten (Flatten)	(None, 7680)	0
dense (Dense)	(None, 128)	983168
batch_normalization_4 (Batch	(None, 128)	512
dropout_4 (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 64)	8256
batch_normalization_5 (Batch	(None, 64)	256
dropout_5 (Dropout)	(None, 64)	0
dense_2 (Dense)	(None, 7)	455

Total params: 1,139,271
Trainable params: 1,137,927
Non-trainable params: 1,344

Suppose the feature map is denoted by H x W x C in size, H and W as height and width while C signifies the number of channels. Then the aggregation function is expressed as:

Let

$X = \{x_1, x_2, \ldots, x_N\}$, where N is the sources (5)

The aggregation function which maps data to N sources, is indicated by

$h(x_1) \in R$ mapped to data $f(h(x_1), h(x_2), \ldots, h(x_N), \Theta) \in R$ (6)

where Θ are the parameters of f

The FM addresses the ambiguity axis of uncertainty which involves providing answers to how likely various subsets are found in the sources of information.

For a finite N information source, X; FM is a set-valued function such that:

$g: 2^x \longrightarrow [0,1]$ with the conditions

$g(\emptyset) = 0$ and $g(X) = 1$ as the boundary (7)

if A, B ⊆ X with A ⊆ B, then g(A) ≤ g(B) as the monotonicity (8)

The ChI is denoted, for a finite set of N information sources, X, FM g, as

$$\int h \circ g = \sum_{i=1}^{N} w_i h(x_{\pi(i)}) \quad (9)$$

where $w_i = (G_{\pi(i)} - G_{\pi(i-1)})$,
$G_{(i)} = g(\{x_{\pi(1)}, \ldots, x_{\pi(i)}\})$,
$G_{\pi(0)} = 0$, $h(x_i)$ symbolizes the hypothesis strength and π(i) is a sorting on X such that $h(x_{\pi(1)}) \geq \ldots \geq h(x_{\pi(N)})$

The FM is defined by Ordered Weight Average (OWA) [9, 44] which includes max, min, softmax, softmin, and average. Hence, the output from the convolution layer serves as input to the fuzzy layer while performing the ChI, resulting in five fused outputs passed to the next layer of the network. The fuzzy layer presents the ability to capture information while condensing rather large feature maps to an arbitrarily reduced number of feature maps, i.e., a further improvement in learned information utilization from the previous layer. This is evident as the fuzzy layer reduces approximately 18% of the total trainable parameters.

5. Dataset – The Multimodal Opinion Level Sentiment Intensity Dataset (CMU-MOSI)

CMU-MOSI dataset constitutes a collection of 2199 opinion videos, each annotated with a sentiment range [-3,3]. The opinion videos resulted from 93 randomly selected videos with 89 distinct speakers, including 41 females and 48 males, with an approximate age range of 20 to 30 years. Speakers are from different ethnic backgrounds but expressed themselves in English as videos originated from the United States of America or the United Kingdom. The videos were recorded in diverse setups; while some users have high-tech microphones and cameras, others with less technicality. Similarly, the distance between the camera, background, and lighting varied between videos. All videos are in the original resolution and recorded in MP4 format, ranging from 2-5 minutes. The videos contained only one speaker primarily looking at the camera. The dataset distribution is presented in Fig. 7.

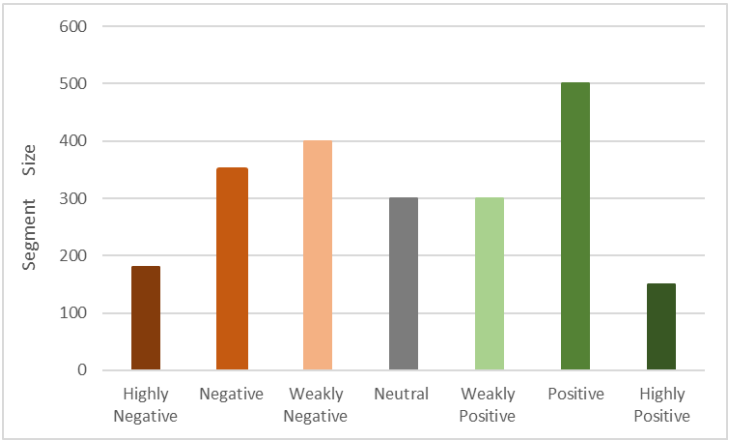

Fig. 7 - Sentiment distribution over the entire dataset

6. Experimental Results

This paper investigates the assimilation of a fuzzy layer in the CNN architecture in exploiting the aggregated properties of fuzzy methodologies explored via fuzzified hidden layers. Experimental results were conducted on seven (7) and binary class classifications using a supervised learning approach. Also, the CNN process was completed in two stages: the training and inference stages. The training stage was conducted on the training samples obtained in a 70:20:10 ratio from the MOSI dataset (training, validation, and testing, respectively). The models were validated via the validation samples, and the prediction of model(s) results was via the testing samples. For training, the model (s) framework was implemented in Python using the Keras library on a CPU for computation. In addition, data augmentation strategies were deployed to increase training data, achieving more robust solutions. Furthermore, the Adam algorithm [45] was employed to provide for better optimization performance, while Dropout [46] was utilized after pooling with a 25% dropout rate.

Hence, accuracy measurement was deployed to evaluate the model's performance for investigation for future works about fuzzy integration with CNN. The accuracy measure demonstrates an overall model evaluation of the projected ratio of precisely classified total occurrences to the number of samples in total. The obtained findings illustrate the model's efficacy in the state-of-the-art MOSI dataset. The quantitative results on the average accuracy

comparison of the models' results are presented in Table 4 for seven (7) and binary classification over five (5) iterations.

Table 4 - Averaged accuracy over five (5) iterations for seven (7) and binary classifications

	Accuracy Comparison			
	Seven (7) Class		Binary Class	
	Typical 2D CNN	Fuzzy layered 2D CNN	Typical 2D CNN	Fuzzy layered 2D CNN
Accuracy	28.9%	37.5%	78.0%	81.0%

A graphic illustration of accuracy results over five iterations is presented below (see Fig. 8.) for seven (7) and binary classifications.

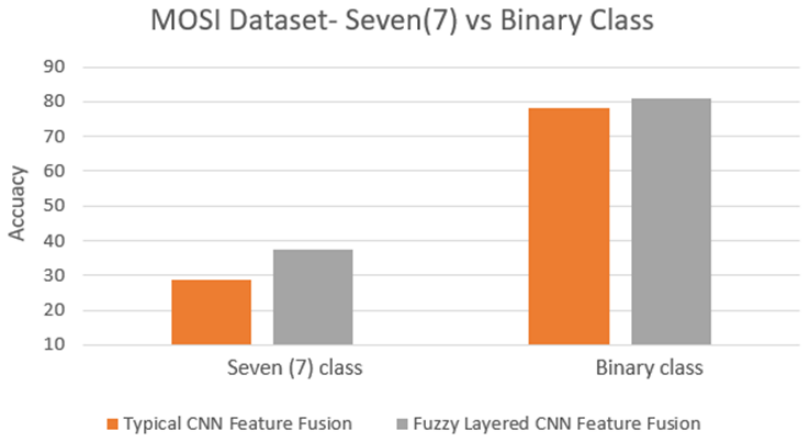

Fig. 8 - Averaged on seven (7) and binary classifications over five (5) iterations

Experimental results obtained above (see Table 3) reveals that fuzzy layer inclusion, which can handle uncertainty/vagueness and complex information, with CNN architecture improves its performance in reference to the evaluation method i.e., accuracy. Furthermore, the inclusion layer reveals an aggregation strategy in the feature maps that provides better classification performance and is unequivocally superior to the typical CNN. In addition, the fuzzy layer offers minimal impact on training time in seconds to a few minutes. Similarly, the inclusion of the fuzzy layer within the CNN architecture reveals that it satisfies the end-to-end learning while establishing a connection between features and the results hence improving the learning capability of the DL algorithm, CNN.

Furthermore, to compare the models (inter-modalities) on their means, a technique called analysis of variance (ANOVA) was deployed. ANOVA is a technique applied in verifying for statistically significant differences amid two or more independent (disparate) groups' mean values. The one-way ANOVA analysis was conducted which tests the null hypothesis:

$$H_o: \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k \quad (10)$$

where μ = group mean and k = number of groups.

However, if the result obtained is statistically significant, the alternative hypothesis (H_A) is accepted. Such that;

$$H_A: \mu_1 \neq \mu_2 \neq \mu_3 \neq \ldots \neq \mu_k \quad (11)$$

That is; statistically, a significant difference exists in at least the two group means. The analysis was computed using Excel 2013 [see; 47] for mathematical interpretation.

Obtained results were contrasted with the ANOVA table (see Table 5) to check for the computed F ratio and the F critical value. The attained F ratio for seven (7) classes is 8.80 with (1,8) degrees of freedom at a significance level of 0.05 which when crossed-verified with the F critical value (the intersection of row 8 and column 1) from the distribution table is 5.32. Similarly, the obtained F ratio for binary classification is 7.58 with (1,8) degree of freedom at 0.05 significance level.

Thus, since the obtained P-value, 0.017 and 0.024 for seven (7) and binary classifications respectively, is less than the α (significance) value 0.05, then the null hypothesis H_o is rejected. Hence, we conclude that the result is statistically significate. Thus, this implies that,

F value obtained > F critical and P value < alpha value

Table 5 - Table of critical values of F Distribution

Critical values of F for the 0.05 significance level:

	1	2	3	4	5	6	7	8	9	10
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.39	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.97	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.10	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60

7. Conclusion and Future Works

Information models have exploded in the epoch of big data alongside the incessant artificial intelligence development. Similarly, with the existence of multiple modalities, high dimensions with different structures, and vast information redundances, this paper presents a deep learning-based fuzzy integrated layer multimodal fusion sentiment analysis methodology. Evaluations revealed that the proposed fuzzy-layer fusion methodology improves performance on MSA for the CMU-MOSI data as demonstrated experimentally both on binary and multiple (seven-class) classification. Furthermore, it is worth noting that the inclusion of the fuzzy layer in the proposed methodology had minimal impact on the total model training time, which increased at most by a few minutes. Furthermore, though the fusion of DL and fuzzy enhances the prediction or classification accuracy in contrast with conventional CNN models, certain limitations still exist with reverence to computational intricacy, interpretability, and parameter settings [48].

In addition, we investigated the statistical significance of the models by the proposed method. The experiment reveals a statistical significance in both classification categories; seven (7) and binary. The F-ratio computed is greater than the critical F values and the P-value result attained is less than the set alpha value at a 0.05 significance level. Therefore, the null hypothesis is rejected, implying that the sample means are equal (no significant difference). The alternative hypothesis is accepted suggesting that at least one of the sample means differs from the rest of the sample means (a statistical difference).

Additionally, DL has become an indispensable and proven ML tool via its performance characteristics. However, being a black-box model, there exists a difficulty in diagnosing decision-driven aspects of the model's input. Hence, there is a need for further method development and studies that could enable the explanation of DL decisions on model transparency and trustworthiness. Thus, in future works, this study intends to provide explainability aspects of the DL decision via integrating the FL within the DL architecture.

Acknowledgement

The authors wish to express their gratitude to Universiti Utara Malaysia and Universiti Malaysia Kelantan for its support.

References

[1] S. I. R. E. E. S. H. A. Jasti and G. V. Kumar, "Deep sentiment extraction using fuzzy-rule based deep sentiment analysis," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 6, 2022.

[2] A. Gandhi, K. Adhvaryu, and V. Khanduja, "Multimodal sentiment analysis: Review, Application Domains and future directions," *2021 IEEE Pune Section International Conference (PuneCon)*, 2021

[3] D. Ghosal, M. S. Akhtar, D. Chauhan, S. Poria, A. Ekbal, and P. Bhattacharyya, "Contextual inter-modal attention for multi-modal sentiment analysis," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[4] U. Sehar, S. Kanwal, K. Dashtipur, U. Mir, U. Abbasi, and F. Khan, "Urdu sentiment analysis via multimodal data mining based on Deep Learning Algorithms," *IEEE Access*, vol. 9, pp. 153072–153082, 2021

[5] X. Li and M. Chen, "Multimodal sentiment analysis with multi-perspective fusion network focusing on sense attentive language," *Lecture Notes in Computer Science*, pp. 359–373, 2020.

[6] D. Liu, Z. Wang, L. Wang, and L. Chen, "Multi-Modal Fusion Emotion Recognition Method of speech expression based on Deep Learning," *Frontiers in Neurorobotics*, vol. 15, 2021.

[7] G. Chandrasekaran, T. N. Nguyen, and J. Hemanth D., "Multimodal Sentimental Analysis for Social Media Applications: A comprehensive review," *WIREs Data Mining and Knowledge Discovery*, vol. 11, no. 5, 2021.

[8] Z. Cai, H. Gao, J. Li, and X. Wang, "Deep learning approaches on multimodal sentiment analysis," *2022 IEEE International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA)*, 2022.

[9] S. Mai, H. Hu, and S. Xing, "Divide, conquer and combine: Hierarchical feature fusion network with local and Global Perspectives for multimodal affective computing," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[10] O. A. Onasoga, N. Yusof, and N. H. Harun, "Audio classification - feature dimensional analysis," *The Importance of New Technologies and Entrepreneurship in Business Development: In The Context of Economic Diversity in Developing Countries*, pp. 775–788, 2021.

[11] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, "Very deep convolutional neural networks for raw waveforms," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[12] B. Ramasamy and A. Z. Hameed, "Classification of healthcare data using hybridised fuzzy and Convolutional Neural Network," *Healthcare Technology Letters*, vol. 6, no. 3, pp. 59–63, 2019.

[13] C. El Hatri and J. Boumhidi, "Fuzzy deep learning based urban traffic incident detection," *Cognitive Systems Research*, vol. 50, pp. 206–213, 2018.

[14] P. Bedi and P. Khurana, "Sentiment analysis using fuzzy-deep learning," *Proceedings of ICETIT 2019*, pp. 246–257, 2020.

[15] Sugiyarto, J. Eliyanto, N. Irsalinda, and M. Fitrianawati, "Fuzzy sentiment analysis using convolutional neural network," *International Conference on Mathematics, Computational Sciences and Statistics 2020*.

[16] V. V. Borisov and K. P. Korshunova, "Multiclass classification based on the Convolutional Fuzzy Neural Networks," *Communications in Computer and Information Science*, pp. 226–233, 2019.

[17] Y. Deng, Z. Ren, Y. Kong, F. Bao, and Q. Dai, "A hierarchical fused fuzzy deep neural network for Data Classification," *IEEE Transactions on Fuzzy Systems*, vol. 25, no. 4, pp. 1006–1012, 2017.

[18] T.-L. Nguyen, S. Kavuri, and M. Lee, "A fuzzy convolutional neural network for text sentiment analysis," *Journal of Intelligent & Fuzzy Systems*, vol. 35, no. 6, pp. 6025–6034, 2018.

[19] S. R. Price, S. R. Price, and D. T. Anderson, "Introducing fuzzy layers for deep learning," *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2019.

[20] L. A. Zadeh, "A fuzzy-algorithmic approach to the definition of complex or imprecise concepts," *Systems Theory in the Social Sciences*, pp. 202–282, 1976.

[21] J. J. Buckley and Y. Hayashi, "Fuzzy Neural Networks: A survey," *Fuzzy Sets and Systems*, vol. 66, no. 1, pp. 1–13, 1994.

[22] G. Suhang, F.-L. Chung, and S. T. Wang, "A novel deep fuzzy classifier by stacking adversarial interpretable tsk fuzzy sub-classifiers with smooth gradient information," *IEEE Transactions on Fuzzy Systems*, pp. 1–1, 2019.

[23] M. G. Huddar, S. S. Sannakki, and V. S. Rajpurohit, "Correction to: Attention-based multimodal contextual fusion for sentiment and emotion classification using Bidirectional LSTM," *Multimedia Tools and Applications*, vol. 80, no. 9, pp. 13077–13077, 2021.

[24] W. Han, H. Chen, A. Gelbukh, A. Zadeh, L.-philippe Morency, and S. Poria, "Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis," *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021

[25] X. Yan, H. Xue, S. Jiang, and Z. Liu, "Multimodal sentiment analysis using multi-tensor fusion network with Cross-modal modeling," *Applied Artificial Intelligence*, vol. 36, no. 1, 2021.

[26] C. Xi, G. Lu, and J. Yan, "Multimodal sentiment analysis based on multi-head attention mechanism," *Proceedings of the 4th International Conference on Machine Learning and Soft Computing*, 2020.

[27] P. Bedi and P. Khurana, "Sentiment analysis using fuzzy-deep learning," *Proceedings of ICETIT 2019*, pp. 246–257, 2020.

[28] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 7216–7223, 2019.

[29] R. Dabare, K. W. Wong, M. F. Shiratuddin, and P. Koutsakis, "Fuzzy deep neural network for classification of overlapped data," *Neural Information Processing*, pp. 633–643, 2019.

[30] T. Sharma, V. Singh, S. Sudhakaran, and N. K. Verma, "Fuzzy-based pooling in Convolutional Neural Network for Image Classification," *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2019.

[31] A. Sarabakha and E. Kayacan, "Online deep fuzzy learning for control of nonlinear systems using expert knowledge," *IEEE Transactions on Fuzzy Systems*, pp. 1–1, 2019.
[32] Lin, Lin, Sun, and Wang, "Evolutionary-fuzzy-integral-based convolutional neural networks for facial image classification," *Electronics*, vol. 8, no. 9, p. 997, 2019.
[33] T. Deshmukh and H. S. Fadewar, "Fuzzy deep learning for diabetes detection," *Advances in Intelligent Systems and Computing*, pp. 875–882, 2018.
[34] D. Glukhov, R. Bohush, J. Mäkiö, and T. Hlukhava, "A joint application of fuzzy logic approximation and a deep learning neural network to build fish concentration maps based on Sonar Data," *Computer Modeling and Intelligent Systems*, vol. 2353, pp. 133–142, 2019.
[35] X. Liang, G. Wang, M. R. Min, Y. Qi, and Z. Han, "A deep spatio-temporal fuzzy neural network for passenger demand prediction," *Proceedings of the 2019 SIAM International Conference on Data Mining*, pp. 100–108, 2019.
[36] M. Al-Faris, J. Chiverton, Y. Yang, and D. Ndzi, "Deep learning of fuzzy weighted multi-resolution depth motion maps with spatial feature fusion for action recognition," *Journal of Imaging*, vol. 5, no. 10, p. 82, 2019.
[37] H. J. Sadaei, P. C. de Lima e Silva, F. G. Guimarães, and M. H. Lee, "Short-term load forecasting by using a combined method of convolutional neural networks and Fuzzy Time Series," *Energy*, vol. 175, pp. 365–377, 2019.
[38] T.-L. Nguyen, S. Kavuri, and M. Lee, "A multimodal convolutional neuro-fuzzy network for emotion understanding of movie clips," *Neural Networks*, vol. 118, pp. 208–219, 2019
[39] S. Dieleman and B. Schrauwen, "End-to-end learning for Music Audio," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.
[40] Y. LeCun, Y., Bengio, & G. Hinton, "Deep learning. *nature*," vol. *521*, *no.* 7553, pp. 436-444, 2015.
[41] I. Wieser, P. Barros, S. Heinrich, and S. Wermter, "Understanding auditory representations of emotional expressions with neural networks," *Neural Computing and Applications*, vol. 32, no. 4, pp. 1007–1022, 2018.
[42] D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for Image Classification," *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
[43] J. M. Keller, D. B. Fogel, and D. Liu, Fundamentals of computational intelligence: neural networks, fuzzy systems, and evolutionary computation.
John Wiley & Sons, 2016.
[44] R. R. Yager, "On ordered weighted averaging aggregation operators in multicriteria decision-making," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 18, no. 1, pp. 183–190, 1988.
[45] D.P. Kingma, & J. Ba, "Adam: A Method for Stochastic Optimization," 2015 *CoRR, abs/1412.6980*. 3rd International Conference for Learning Representations, San Diego, 2015 *(ICLR)*. https://doi.org/10.48550/arXiv.1412.6980
[46] N., Srivastava, G. Hinton, A. Krizhevsky, , I. Sutskever, &, R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol *15*, *no.* 1, pp. 1929-1958, 2014.
[47] Admin, "ANOVA formula in statistics with solved example," *BYJUS*, 03-Jan-2022. [Online]. Available: https://byjus.com/anova-formula. [Accessed: 30-Nov-2022].
[48] Y. Zheng, Z. Xu, and X. Wang, "The fusion of deep learning and fuzzy systems: A state-of-the-art survey," *IEEE Transactions on Fuzzy Systems*, vol. 30, no. 8, pp. 2783–2799, 2022.
[49] C. Luna-Jiménez, D. Griol, Z. Callejas, R. Kleinlein, J. M. Montero, and F. Fernández-Martínez, "Multimodal Emotion Recognition on RAVDESS dataset using transfer learning," *Sensors*, vol. 21, no. 22, p. 7665, 2021.