Online at: https://researchtrendsjournal.com **ISSN No:** 2584-282X
Indexed Journal **Peer Reviewed Journal**

INTERNATIONAL JOURNAL OF TRENDS IN EMERGING RESEARCH AND DEVELOPMENT

Volume 3; Issue 1; 2025; Page No. 135-141

Received: 01-10-2024
Accepted: 04-12-2024

To construct a multimodal method for recognizing emotions that incorporates facial expression analysis, physiological signals, and contextual data

[1]Amit Ghosh and [2]Dr. Santanu Sikdar

[1]Research Scholar, P.K University, Shivpuri, Madhya Pradesh, India
[2]Professor, P.K University, Shivpuri, Madhya Pradesh, India

DOI: https://doi.org/10.5281/zenodo.15706963

Corresponding Author: Amit Ghosh

Abstract

In particular, in order to ascertain if interpolation yields superior outcomes, the researchers wish to conduct an experimental assessment of the interpolation technique's effectiveness in upscaling lower-resolution pictures. This By methodically examining its effects on picture quality and model performance, the study seeks to empirically demonstrate how well interpolation works to increase the accuracy of models used in image processing tasks, particularly for face emotion recognition. To demonstrate that interpolation increases the precision of models used for image processing tasks, especially the identification of facial emotions, this research will examine its effects on model performance and picture quality in a systematic way.

Keywords: Human emotions, physiological signs, emotion recognition, emotional intelligence

1. Introduction

Emotional intelligence and management are seen as fundamental to success in the modern workplace, business, and consumer goods [Norman, 2004; Bunting, 2004] [16, 17]. In terms of usability, users care about more than just features; they want experiences that cater to their expectations, emotions, and the purposes of their interactions. So, how end users perceive a system's usability is the deciding element, which presents ongoing difficulties for HCI researchers.

Emotional intelligence and management are seen as fundamental to success in the modern workplace, business, and consumer goods [Norman, 2004; Bunting, 2004] [16, 17]. In terms of usability, users care about more than just features; they want experiences that cater to their expectations, emotions, and the purposes of their interactions. So, how end users perceive a system's usability is the deciding element, which presents ongoing difficulties for HCI researchers.

Emotion, according to the theories, is a multi-faceted concept. Everything from our thoughts and actions to our decisions and interactions with others is greatly impacted by our emotions. Insights like this may motivate us to take action, help us communicate with others, and reveal important details about our environment. Our ability to remember things, focus our attention, and make judgments even ones fueled by emotion are all impacted by our emotional states.

When people are feeling emotionally stable, they are better able to control their reactions, make sound judgments, and manage social situations. Emotional disturbances and disorders may have a devastating effect on a person's ability to function and overall happiness, thus they are also vital to mental health and wellness. In an effort to decipher the nature, function, and history of emotion, a plethora of competing hypotheses have evolved. Many modern researchers continue to rely on classic ideas as initial frameworks, despite the fact that they have been the target of criticism.

By integrating physiological arousal with cognitive interpretation, the Schachter-Singer model improved upon the James-Lange hypothesis. Individuals have physiological arousal in reaction to stimuli, which is then interpreted by contextual signals and labelled emotionally, according to this paradigm. Affective Events Theory, which delves into the temporal dimension of emotional responses, and contemporary Cognitive Appraisal Theories like Cognitive-Mediational Theory put forward by Richard Lazarus are examples of more recent contributions. The impact of cognitive assessment on the formation of emotional experiences is emphasized by these theories.

Emotions, feelings, and beliefs create sentiments, which are complex and long-lasting emotional attitudes or views regarding people, things, or ideas. One aspect that highlights their social origins and effect is the presence of long-lasting social ties that elicit emotions. Dimensions of emotion include valence and arousal. The term "arousal" describes the degree of mental and physical activity linked with an emotional state, which may range from relaxed to agitated. The positive or negative aspect of an emotion is represented by its valence, which may go from pleasant to unpleasant.

2. Literature Review

Rajneesh Kumar *et al*. (2020) [1] advocated the use of a voice audio stream for emotion detection. In order to train SVM, they extracted features using the MFCC, or Mel frequency cepstral coefficient, and the delta MFCC approach. Their accuracy rate using this approach was 83% for all audio files, including those including both male and female voices, and 85% for individual voices.

Md. Zia Uddin *et al*. (2020) [2] suggested using audio signals for emotion analysis. They said that because there is no visual data of human faces, emotion analysis is a crucial task. Mel-frequency cepstral coefficients (MFCC) characteristics were used to train fast neural structured learning (NSL) for an emotion identification system that is subject independent. In order to reduce the inter-class scatterings and maximize the inner-class scatterings, discriminant analysis was performed to the MFCC feature prior to network training. When MFCC, discriminant analysis, and NSL were combined, the resulting recognition rates were greater than those of other conventional methods.

Zhang Z *et al*. (2020) [3] emphasized how crucial facial expression detection is for identifying emotions and the difficulty of the job owing to individual expression correlation and ambiguity. They suggested a learning model called the correlation emotion label distribution to tackle these problems; this approach assigns different emotions to each expression based on how similar they are. They utilized a deep learning technique called CNN and the cosine discrete transform to determine how comparable expression features were. The 'Oulu CASIA' database was used for the experiments.

A Joseph *et al*. (2020) [4] displayed facial expressions as a means of emotion detection. The first step in picture enhancement was to combine DWT with fuzzy logic. After that, the altered method for the lips and eyes was used to ascertain the face's geometry. Lastly, a Tensor flow was used in conjunction with a neural network to facilitate categorization. The datasets 'Karolinska Directed Emotional Faces' (KDEF), 'Oulu-CASIA,' and 'CK+'. for the assessment.

Wei Tao *et al*. (2020) [5] shown how to use electroencephalography (EEG) to identify emotions. Their suggested remedy was the attention-based convolutional recurrent neural network (ACRNN). They proved that raw EEG signals include spatial information by highlighting the time dependence across temporal slices and the intrinsic link across different channels. The suggested ACRNN first learned the spatial features of multichannel EEG using convolutional layers, and then it examined the temporal aspects using long-short term memory (LSTM) networks. features of distinct temporal slices. The DEAP and DREAMER databases, which are accessible to the public, were used for the experiments

3. Martials and Methods

3.1 Design and Methods

A feature extraction and classification strategy for facial emotion identification based on the VGG16, Google Net, and ResNet50 models is suggested in this study. In this research, seven Various feelings, including joy, sorrow, rage, surprise, fear, contempt, and neutrality, were attributed to images from the dataset.

3.2 Gathering and preparing data

In order to conduct this study, we will be using structured databases to collect trend datasets. Choosing For instance, face datasets would have unnecessarily complicated the face detection method, even if the main objective of this study is not facial identification. Following a thorough analysis, the trend VGG16 and ResNet50 models were chosen for this study's loading onto DCNN. These models provided labeled photos with a noise-free backdrop and the key face expressions.

3.3 Face Detection

In the processing pipeline, face detection was the first and most important phase. We could not go further with the analysis without first identifying the face, even though our photographs only include frontal facial expressions. we settled on a face detector that used the Dlib library's Oriented gradient histogram (HoG). A combination of Support Vector Machines (SVM) and HoG descriptors is needed for precise face recognition in images worked well.

3.4 Pre-Trained Models

An annual event is the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). that showcases various photo categorization models developed over the years. The components will either operate as a feature extraction process or use the weights of the identified features to formulate new interpretation problems. Keras is great since it's easy to use, has pre-trained CNNs available, and you can acquire model weights for free from their libraries.

Table 1: Keras Models Available for Use.

Model	Size (MB)	Top-1 Accuracy	Top-5 Accuracy	Parameters	Depth
Xception	88	79.0%	94.5%	22.9M	81
VGG16	528	71.3%	90.1%	138.4M	16
VGG19	549	71.3%	90.0%	143.7M	19
ResNet50	98	74.9%	92.1%	25.6M	107
ResNet50V2	98	76.0%	93.0%	25.6M	103
ResNet101	171	76.4%	92.8%	44.7M	209
ResNet101V2	171	77.2%	93.8%	44.7M	205
ResNet152	232	76.6%	93.1%	60.4M	311
ResNet152V2	232	78.0%	94.2%	60.4M	307
InceptionV3	92	77.9%	93.7%	23.9M	189
InceptionResNetV2	215	80.3%	95.3%	55.9M	449
Mobile Net	16	70.4%	89.5%	4.3M	55
MobileNetV2	14	71.3%	90.1%	3.5M	105
DenseNet121	33	75.0%	92.3%	8.1M	242
DenseNet169	57	76.2%	93.2%	14.3M	338
DenseNet201	80	77.3%	93.6%	20.2M	402
NASNet Mobile	23	74.4%	91.9%	5.3M	389
NASNet Large	343	82.5%	96.0%	88.9M	533
EïcientNetV2B0	29	78.7%	94.3%	7.2M	-

Keras applications make use for prediction, feature extraction, and fine-tuning deep learning models. When the model is constructed, all required weights are immediately added to the/.keras/models folder.

3.5 Implementation Tools
Choosing the proper high-level programming language to develop the system was difficult since there are now over 2000 of them accessible [16]. Fortunately, this problem was made simpler by the project requirements. The neural network implementation and GUI design should be done in a straightforward and easy-to-understand programming language. However, learning C++ is infamously hard. and much more so to optimize, despite its reputation as the top option when performance is paramount. As a stand-in, Mat Lab is fast, simple to learn, and performs adequately.

3.6 Evaluation Methods
In order to calculate these values, the detection metrics were used. A model's performance for each class may be learned from the detection metrics. Using the value of the confusion matrix, the following metrics were created

3.6.1 Accuracy Rate
The frequency with which the model predicts the class accurately, that is, with a normal face image and an appropriate emotion picture, is addressed in the following explanation. It is calculated using this formula. In most cases, it is a metric for the classifier's accuracy rate.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

3.7 Facial emotion recognition using multimodal signals
After that, features are extracted from the processed picture using LDTOOP, a combination of LOOP and LDTP. Similarly, features like wavelet coefficients and spectrum parameters including PSD, spectral flux, tonal power ratio, spectral skewness, and spectral centroids are used to get the EEG and physiological signals ready for analysis. extracted.

3.7.1 Gathering of input video
An important aspect of emotional computing systems is emotion recognition. A person's emotional state is multi-faceted, including not just their thoughts and actions but also their physiological responses to external stimuli. Physiological signals, video, and electroencephalogram (EEG) data are common components of facial emotion detection systems.

3.7.2 Image extraction for the face
Here we'll assume a dataset A and use the variable to indicate the number of movies included inside it.

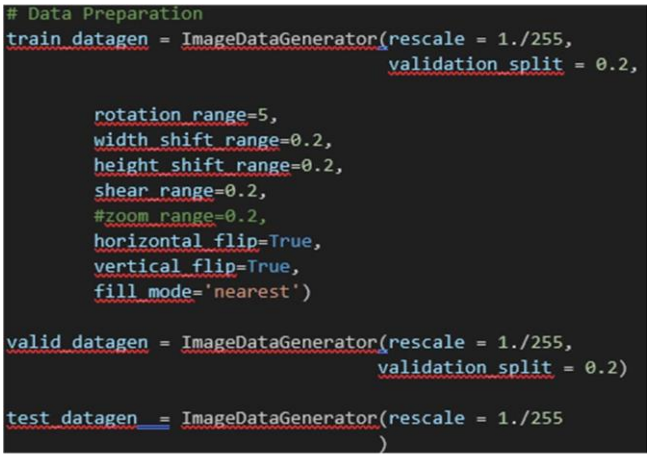
$$H = \{A_1, A_2 \ldots. A_i \ldots\ldots A_n\} \tag{1}$$

Here, H stands for the dataset and n for the number of movies. Face shots are included in each film from different angles. In dataset H, the ith movie contains facial pictures, which may be represented as follows:

$$A_i = \{X_1, X_2, \ldots. X_i \ldots. X_n\} \tag{2}$$

The ith facial picture from the ith video is represented by A in this context. There is a unique location for each individual's face picture, and each face image belongs to a certain person.

4. Results
4.1 Data Preparation
An effective "Image Data Generator," a Keras tool for real-time data augmentation during neural network training, is often used to fictitiously increase the size of the training dataset and improve the generalizability of the model. The picture sets used for training and validation were both preprocessed and improved using Image Data Generator objects, which were used for data augmentation in this work.



```
# Data Preparation
train_datagen = ImageDataGenerator(rescale = 1./255,
                                   validation_split = 0.2,

        rotation_range=5,
        width_shift_range=0.2,
        height_shift_range=0.2,
        shear_range=0.2,
        #zoom_range=0.2,
        horizontal_flip=True,
        vertical_flip=True,
        fill_mode='nearest')

valid_datagen = ImageDataGenerator(rescale = 1./255,
                                  validation_split = 0.2)

test_datagen  = ImageDataGenerator(rescale = 1./255
                                  )
```

Fig 1: Preparation of data

4.1.1 Load the Validation and Training Datasets
Using the 'Image Data Generator' class in Keras to set up for testing, validation, and training a convolutional neural network (CNN). In deep learning, this is a typical way to handle picture data.

```
# Training Dataset
train_dataset  = train_datagen.flow_from_directory(directory = 'input/fer2013/train',
                                                   target_size = (48,48),
                                                   class_mode = 'categorical',
                                                   subset = 'training',
                                                   batch_size = 64)
# Validation Dataset
valid_dataset = valid_datagen.flow_from_directory(directory = 'input/fer2013/train',
                                                  target_size = (48,48),
                                                  class_mode = 'categorical',
                                                  subset = 'validation',
                                                  batch_size = 64)

# Test Dataset:
test_dataset = test_datagen.flow_from_directory(directory = 'input/fer2013/test',
                                                target_size = (48,48),
                                                class_mode = 'categorical',
                                                batch_size = 64)
```

Fig 2: Distribution of classes for the training and validation sets, respectively

It is quite probable that Train datagen is an Image Data Generator object with data augmentation settings. The output will be a set of resized images with dimensions of (48, 48) taken from the 'input/fer2013/train' directory. With the 'categorical' class mode, One-hot encoding is used for the labels. The subset used for training is labeled as "training." There is a batch size of 64.

4.1.2 Define the Architecture of the Model

Each of the emotional models' detection has its own unique architecture. To demonstrate, we look at several pre-training instances of popular model architectures, such as VGG16, InceptionV3, and ResNet50. With these models, the weight "ImageNet" is introduced. Because the weights were trained for a particular input configure ration, It is crucial to realize that the architecture of the model cannot be altered.

```
# VGG16 Model Architecture
model = Sequential()
model.add(base_model)
model.add(GlobalAveragePooling2D())
model.add(Dense(256, activation='relu'))
model.add(BatchNormalization())
model.add(Dropout(0.5))
model.add(Dense(128, activation='relu'))
model.add(BatchNormalization())
model.add(Dropout(0.5))
model.add(Dense(7, activation='softmax'))
```

Fig 3: Architecture of the VGG16 Model

The first stage of the ResNet50 transfer learning model, similar to the VGG16 model, is to load the ResNet50 model that has already been trained, without the top layer.

4.2 Results of the experiment

4.2.1 matrix of confusion

A 7x7 grid that contrasts the anticipated and actual names for various emotions is called a confusion matrix. The matrix's values indicate the proportion of times a predicted emotion was correct relative to the actual emotion.

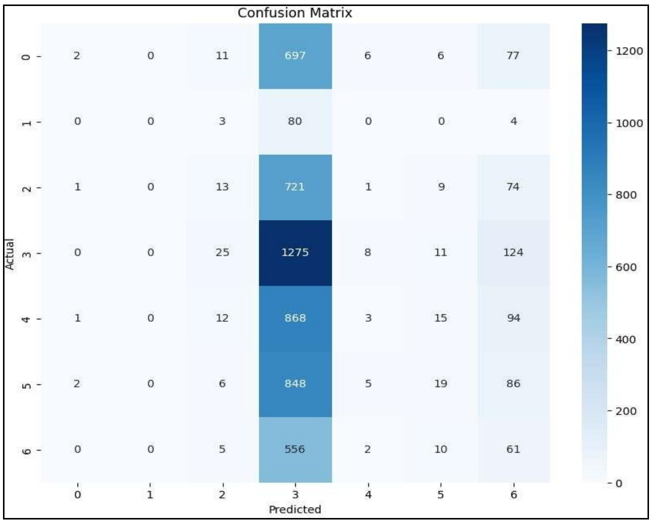


Fig 4: The pre-trained Resnet50 model's average confusion matrix across 10 epochs.

4.2.2 Loss and accuracy

To improve the model's performance, further investigation and adjustment may be necessary, especially if certain classes are more difficult or uncommon in the dataset.

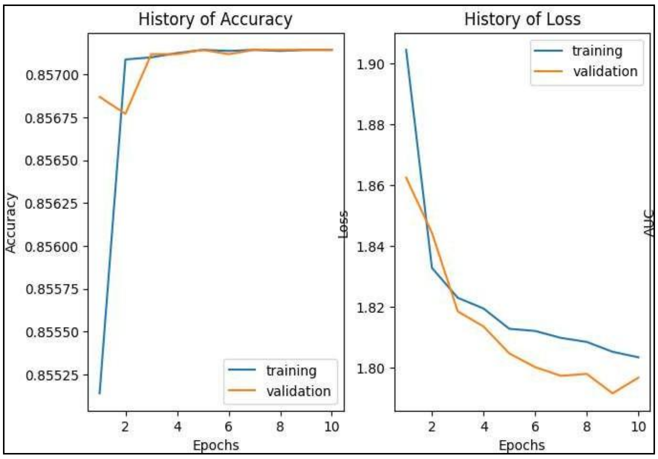


Fig 5: Using the Fer 2013.csv dataset, the pre-trained Resnet50 model's training accuracy and loss were assessed across ten epochs

According to the machine learning model's ultimate training and validation accuracy on both datasets was about 85.7%. This indicates that the model continuously performed better than anticipated on both the training and invisible validation data.

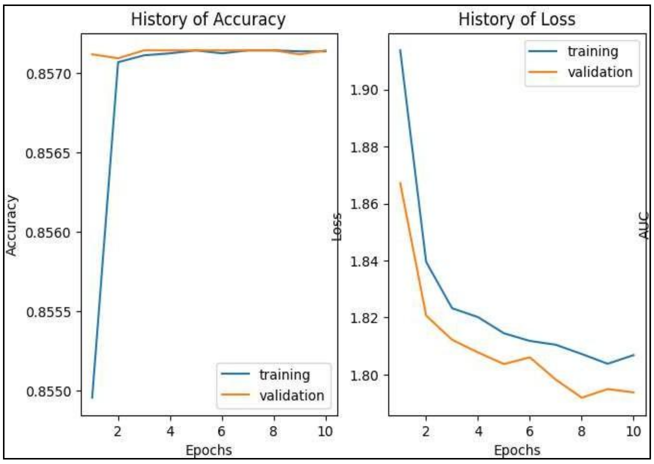


Fig 6: Using the Fer2013.csv dataset, the pre-trained VGG16 model's training accuracy and loss were assessed across ten epochs.

4.3 F1-Score

You may evaluate the classification model's performance across seven different categories utilizing the F1-score values provided for each class. Resnet50 Class 3 is notable for its balanced recall and accuracy, with an F1-score of around 39.3%. However, Class 1 has an F1-score of 0%, which means that the model failed to accurately categorize any occurrences in this class-a major drawback. When all classes' accuracy and recall are averaged out, the resulting F1-score is 0.1225.

F1-scores of 0% for classes 0, 1, 4, and 5 in VGG16 indicate that the model struggled to balance recall and accuracy for these classes. This might be as a result of its inability to minimize false positives and false negatives or to capture true positives. With an F1-score of around 38.3%, Class 3 performs the best overall in terms of recall and accuracy.

4.3.1 Reports on the classification of every class or emotion

In a classification assignment, the classification report offers a thorough synopsis of every class's performance statistic. The accuracy, recall, F1-score, and support values for each class are often included. Here is the report that we prepared for categorization using the accuracy values that were given to us:

Classification Report:

	precision	recall	f1-score	support
0	0.33	0.00	0.00	799
1	0.00	0.00	0.00	87
2	0.17	0.02	0.03	819
3	0.25	0.88	0.39	1443
4	0.12	0.00	0.01	993
5	0.27	0.02	0.04	966
6	0.12	0.10	0.11	634
accuracy			0.24	5741
macro avg	0.18	0.15	0.08	5741
weighted avg	0.21	0.24	0.12	5741

Fig 7: The Resnet50 image displays the classification report for the sequential model.

Class 3 has a greater memory rate (88%) and a substantially higher accuracy rate (25%) than Classes 0, 1, and 4, with Class 0 having the best recall rate but the lowest precision. Both the macro and weighted average F1-scores show that the model typically performs modestly, with an overall accuracy of 24%.

Classification Report:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	958
1	0.00	0.00	0.00	111
2	0.18	0.03	0.05	1024
3	0.25	0.40	0.31	1774
4	0.20	0.00	0.01	1233
5	0.18	0.47	0.26	1247
6	0.11	0.11	0.11	831
accuracy			0.20	7178
macro avg	0.13	0.15	0.11	7178
weighted avg	0.17	0.20	0.14	7178

Fig 8: VGG16 Image displaying the classification report for the sequential model

The model's poor accuracy, recall, and F1-scores in predicting emotional classes are highlighted in the classification report. Precision and recall for classes 0, 1, and 4 are both at 0.00%, showing that the model fails miserably when it comes to identifying positive examples for these classes. The total model accuracy is still poor at 20%, even though Class 3 shows somewhat improved recall at 40% and precision at 25%. Both the total and weighted F1-scores point to the model's generally poor performance.

Classification Report:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	799
1	0.00	0.00	0.00	87
2	0.11	0.00	0.00	819
3	0.25	1.00	0.40	1443
4	0.00	0.00	0.00	993
5	0.00	0.00	0.00	966
6	0.00	0.00	0.00	634
accuracy			0.25	5741
macro avg	0.05	0.14	0.06	5741
weighted avg	0.08	0.25	0.10	5741

Fig 9: Examining V3 An image displaying the classification report for the sequential model

The model's significant flaws are shown by its low F1-score, recall, and accuracy for the vast majority of classes. It is important to note that the accuracy, recall, and F1-score values of classes 0, 1, 2, 4, 5, and 6 are quite low, suggesting that these groups are not being appropriately identified. Class 3's higher recall of 1.00 demonstrates the model's capacity to catch all positive occurrences, but the accuracy and F1-score are still rather small. The total accuracy of 25% highlights the model's inherent shortcomings.

Table 2: After ten epochs, the FER-2013 test set.

Model	Precision	Recall	F1-Score
Vgg16	16.60%	19.87%	14.26%
Resnet50	21.40%	23.92%	12.25%
inspectionV3	7.89%	25.05%	10.10%

4.4 Experimental con-figure ration

Python was used on a Windows 10 operating system with a configure ration that included 4 GB of RAM to perform the experimental setup for the suggested strategy. The main data source for the studies was the DEAP dataset.

4.5 Details of deep dataset

The publicly accessible online self-assessment ratings and the participant evaluations make up the bulk of the dataset. For studies that aim to categorize emotions, it is an invaluable resource. For the purpose of emotion analysis, the collection includes many physiological recordings, including EEG readings and films of faces.

4.6 Strategy analysis

A wide range of tactics, procedures, and approaches may be used into strategies in order to accomplish goals in an efficient and successful manner. This article presents an elementary examination of the prior research on emotion recognition.

Table 3: Analysis of Strategies for Emotion Discovery

Methods	Research study
ML	25
CNN	15
Fusion	13
SM	13
Deep LSTM	12
AE	8
DBN	7
NN	7

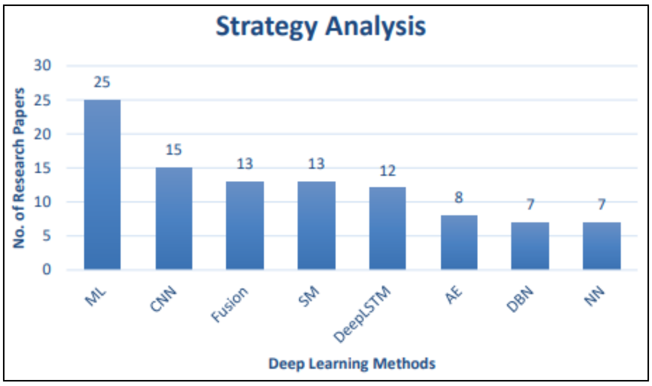


Fig 10: Strategy Analysis for The Existing Methods

4.7 Analysis of tools used for implementation

Using a variety of implementation methods, this investigation explored how literary works have implemented and changed classical methodologies. In the context of emotion identification from multimodal data, there are several software tools that have been optimized for distinct paths. A few of these tools are particularly noteworthy: One popular program for handling voice and audio data is open SMILE. For voice-based applications in particular, it is a well-liked choice for identifying emotions due to its many characteristics and functionalities that extract useful information from audio signals. Its widespread use indicates that it is useful for processing auditory modalities in emotion identification.

Table 4: Methods and Tools Used to Implement Current Approaches

Software	Study
open SMILE	7
TensorFlow	2
HTK	2
LibSVM	3
MATLAB	6
Wavelet	5
Python	4
WEKA	5
HMM	3
Keras	1
NIT	2
ANVIL	4

Hidden Markov Models (HTKs) are often employed in HMM-related tasks. Because HMMs are so effective at modeling temporal sequences, HTK is a useful technique for capturing the dynamics of emotions across time in multimodal data.

5. Conclusion

We demonstrated that while the fundamental concept of information fusion is to combine raw data, the semantic level is more often used due to its computational simplicity compared to data and features level fusions. In terms of accuracy in positive occurrence classification, ResNet50 outperformed the other models with a precision of 21.40 percent. The fact that ResNet50 was able to correctly detect a large percentage of true positive events is shown by its recall of 23.92%. In contrast, InceptionV3 achieved the best recall of 25.05% but the worst accuracy of 7.89% and F1-Score of 10.10%. These findings highlight the necessity of choosing a model that is in sync with particular job needs and goals by highlighting the compromises each model makes between accuracy and recall

6. References

1. Kumar R, Punhani A. Emotion detection from audio using SVM. In: Proceedings of the International Conference on Big Data, Machine Learning and their Applications; c2020. p. 257-265.
2. Uddin MZ, Nilsson EG. Emotion recognition using speech and neural structured learning to facilitate edge intelligence. Eng Appl Artif Intell. 2020;94:103775.
3. Zhang Z, Lai C, Liu H, Li YF. Infrared facial expression recognition via Gaussian-based label distribution learning in the dark illumination environment for human emotion detection. Neurocomputing. 2020;409:341-350.
4. Joseph A, Geetha P. Facial emotion detection using modified eyemap–mouthmap algorithm on an enhanced image and classification with tensorflow. Vis Comput. 2020;36:529-539.
5. Tao W, Li C, Song R, Cheng J, Liu Y, Wan F, *et al.*

EEG-based emotion recognition via channel-wise attention and self-attention. IEEE Trans Affect Comput. 2020;11:1-1.
6. Sajjad M, Nasir M, Ullah FUM, Muhammad K, Sangaiah AK, Baik SW. Raspberry Pi assisted facial expression recognition framework for smart security in law-enforcement services. Inf Sci. 2019;479:416-431.
7. Shen Z, Cheng J, Hu X, Dong Q. Emotion recognition based on multi-view body gestures. In: 2019 IEEE International Conference on Image Processing (ICIP); c2019. p. 3317-3321.
8. Li X, Song D, Zhang P, Zhang Y, Hou Y, Hu B. Exploring EEG features in cross-subject emotion recognition. Front Neurosci. 2018;12:162.
9. Lee HJ, Hong KS. A study on emotion recognition method and its application using face image. In: 2017 International Conference on Information and Communication Technology Convergence (ICTC); c2017. p. 370-372.
10. Saha S, Ganguly B, Konar A. Gesture-based improved human-computer interaction using Microsoft's Kinect sensor. In: 2016 International Conference on Microelectronics, Computing and Communications (MicroCom); c2016. p. 1-6.
11. Wang H, Wu X, Yao L. Identifying cortical brain-directed connectivity networks from high-density EEG for emotion recognition. IEEE Trans Affect Comput. 2020;11:1-1.
12. Alakus TB, Gonen M, Turkoglu I. Database for an emotion recognition system based on EEG signals and various computer games-GAMEEMO. Biomed Signal Process Control. 2020;60:101951.
13. Song T, Zheng W, Lu C, Zong Y, Zhang X, Cui Z. MPED: A multi-modal physiological emotion database for discrete emotion recognition. IEEE Access. 2019;7:12177-12191.
14. Baghdadi A, Aribi Y, Fourati R, Halouani N, Siarry P, Alimi AM. DASPS: A database for anxious states based on a psychological stimulation. arXiv. 2019;1901.02942.
15. Bachmann M, Lass J, Hinrikus H. Single channel EEG analysis for detection of depression. Biomed Signal Process Control. 2017;31:391-397.
16. Norman PM. Knowledge acquisition, knowledge loss, and satisfaction in high technology alliances. Journal of Business Research. 2004;57(6):610-619.
17. Bunting L. Parenting programmes: The best available evidence. Child Care in Practice. 2004;10(4):327-343.