applied sciences

MDPI

Article

Model for Determining the Psycho-Emotional State of a Person Based on Multimodal Data Analysis

Nataliya Shakhovska [1,2,*], Oleh Zherebetskyi [1] and Serhii Lupenko [3,4]

[1] Department of Artificial Intelligence, Lviv Polytechnic National University, 79013 Lviv, Ukraine
[2] College of Engineering, Design and Physical Sciences, Brunel University, London UB8 3PH, UK
[3] Faculty of Electrical Engineering, Automatic Control and Informatics, Opole University of Technology, 45758 Opole, Poland
[4] Institute of Telecommunications and Global Information Space, National Academy of Sciences of Ukraine, 02000 Kyiv, Ukraine
* Correspondence: nataliya.b.shakhovska@lpnu.ua

Abstract: The paper aims to develop an information system for human emotion recognition in streaming data obtained from a PC or smartphone camera, using different methods of modality merging (image, sound and text). The objects of research are the facial expressions, the emotional color of the tone of a conversation and the text transmitted by a person. The paper proposes different neural network structures for emotion recognition based on unimodal flows and models for the margin of the multimodal data. The analysis determined that the best classification accuracy is obtained for systems with data fusion after processing each channel separately and obtaining individual characteristics. The final analysis of the model based on data from a camera and microphone or recording or broadcast of the screen, which were received in the "live" mode, gave a clear understanding that the quality of the obtained results is highly dependent on the quality of the data preparation and labeling. This is directly related to the fact that the data on which the neural network is trained is highly qualified. The neural network with combined data on the penultimate layer allows a psycho-emotional state recognition accuracy of 0.90 to be obtained. The spatial distribution of emotion analysis was also analyzed for each data modality. The model with late fusion of multimodal data demonstrated the best recognition accuracy.

Keywords: multimodal data; late fusion; convolution neural network; emotional state; multi-modal emotion recognition

check for updates

Citation: Shakhovska, N.; Zherebetskyi, O.; Lupenko, S. Model for Determining the Psycho-Emotional State of a Person Based on Multimodal Data Analysis. *Appl. Sci.* **2024**, *14*, 1920. https://doi.org/10.3390/app14051920

Academic Editors: Yue Wu, Mingyang Zhang, Qiguang Miao and Kai Qin

Received: 4 February 2024
Revised: 22 February 2024
Accepted: 23 February 2024
Published: 26 February 2024

1. Introduction

More than a decade ago, the realm of human–machine interaction began gaining traction, progressing toward increasingly user-friendly and adaptable machine interfaces. With the emergence of machine learning, the capability to discern human emotions became feasible [1].

Examining interpersonal interaction underscores the importance of comprehending the psycho-emotional state of conversational partners. Consequently, the integration of emotion recognition into machine–human interaction processes has surged in popularity [2].

Traditionally, humans perceived machines as emotionless entities primarily executing intricate computations. However, equipping machines with the ability to decipher facial expressions, intonation, gestures, and other non-verbal cues promises enhanced rapport between machines and individuals [3].

Many global car brands often introduce systems that warn the driver about his drowsiness [4] or feeling unwell. Popular brands are also willing to pay money to evaluate human emotions while viewing advertisements [5]. A human behavior recognition system is often helpful in predicting crimes [6]. And, there are many other areas in which understanding the emotions of consumers would play the role of an essential push in development [7].

Some reports suggest that the emotion detection and recognition market was worth about $12 billion in 2018 and could grow to more than $90 billion by 2024 [8]. The Israeli startup Beyond Verbal (Tel Aviv-Yafo, Israel) [9] is developing intelligent emotion analytics that look for signs of anger, anxiety and excitement by analyzing a person's voice—not the specific words, but intonations. The proposed approach may be helpful for mental health assessments, market research or even to help call centers improve phone customer relations [10].

Amazon (Bellevue, WA, USA) is also working to improve the ability of its own Alexa AI assistant [11] to interact with users by detecting emotions in their voices. Reference [12] proposed the dataset MOISEI to examine a data fusion model. However, the first research on the multimodal analysis of human emotions was conducted based on two modal parts of information. For example, the usage of a combination of sound and speech modalities is analyzed in [13]. Also, an implementation with a combination of sound and speech in systems with a deep neural network is given in the article in [14]. Based on these works, it was concluded that the fusion of two modal signals provides better accuracy than any unimodal system.

There are also many algorithms combining different types of information signals into multimodal data at various stages (early, late) and using other methods of intermodal information extraction. All these implementations are not easy to understand, but they will eliminate the possible loss of information, which is the basis of multimodal emotion data, and give perfect accuracy.

At this stage of development, models are studied and improved based on intermodal information in multimodal recognition of the psycho-emotional state.

The research aims to develop a model and information system for recognizing human emotions in a stream obtained from a PC or smartphone camera, using different methods of merging channels of image, sound and text, and analysis of the results.

To achieve this goal, it is necessary to solve the following tasks:

- Live generation of data received from a camera and microphone;
- Live generation of data obtained from a recording in the device's memory;
- Live generation of data obtained from the broadcast screen;
- Processing live data to a form that neural networks will accept;
- A mechanism for generating new data to supplement ready-made data sets and create new ones;
- Three neural networks for separate processing of image, sound and text streams;
- Two neural networks for multimodal recognition of emotions;
- Three information systems for recognizing emotions with different ways of merging data streams;
- Interface for ease of use of information systems.

Having solved all these problems, we will be able to conduct a qualitative analysis of the three selected ways to merge data streams and draw appropriate conclusions.

The objectives of this research are:

- The formalization of the process of the multimodal data merging and analysis;
- Each data modality analysis and neural network architecture development;
- The development of the methods of late fusion of multimodal data to achieve the best recognition accuracy;
- Proving the adequacy and demonstration of the functionality of the created model.

The main contribution of the research is given below:

1. The deep learning-based late fusion of multimodal information for human emotion classification of sequential images, audio and text is developed. To achieve this, five neural network structures based on CNN and LSTM for single-mode and multimode emotion recognition are proposed. The number of emotions is one of the model's parameters and can be easily changed. Bidirectional LSTM with CNN where the first layer is wrapped with a TimeDistributed function is used for face emotion recognition.

2. The comparative simulation experiment of single-mode and multimode emotion recognition of emotional features in artificial intelligence information systems is conducted. The best accuracy in the multimodal emotion recognition (MMER) network is 0.9245 with merging in the end (late fusion). However, due to the low accuracy of speech emotion recognition (SER) and text emotion recognition (TER)—0.5611 and 0.64, respectively, this system shows high quality in recognition based on multimodal data obtained in the "live" mode.

The rest of the paper is organized in the following way: Section 2 outlines some recent studies and their analyses, including all the papers that use the dataset examined in this study; Section 3 starts with a detailed description of the data preparation and ways of data merging and follows with the developed model; Section 4 presents the obtained results; Section 5 is the related discussion; and Section 6 is the concluding section.

2. Related Works

Over time, systems began to use fusion at different stages, such as early (at the function level) and late (weighted average of unimodal network results). In [12], a method for processing three types of multimodal data is proposed, which considers the delay and time shift of multimodal data in the time dimension and the relationship between embedding in multimodal data. The evaluation is based on the MOSEI dataset. Experimental results show that the method effectively improves the correlation of relevant information between multimodal signals.

Combining features at the input level and combining deep neural networks for emotion analysis based on multimodal data is described in [15]. In [16,17], the author first encodes each modality separately and then uses feature-level fusion in the middle layer to obtain a multimodal embedding, which is a static feature-level fusion. In contrast to the method described in previous articles, the author in [18,19] uses dynamic fusion at the feature level, the essence of which is the simultaneous encoding of unimodal data and fusion at the feature level. Each information modality is given the same weight in the methods described above. However, this approach is only partially correct because during further research in articles [20–23], researchers found that different modal information has other importance for the final result, and the degree of importance differs. For example, in [23], the author noted that verbal data would be more influential than non-verbal data, leading to a change in the word's meaning. The author takes the verbal modality as dominant and the non-verbal modality as an auxiliary for multimodal fusion.

The papers [24,25] analyze separated modalities for mood detection and modality and fusion. The multimodal-fusion approach in [25] uses a strategy referred to as "late-fusion" that involves the combination of unimodal model outputs as inputs of the decision engine.

During the inception of this technology, there needed to be better quality of emotion determination. This problem has an explanation. Emotions reflect a person's subjective attitude to various phenomena. Emotion is a person's specific response to situations that affect his/her interests. There are emotional manifestations that cannot be controlled and that can be recognized. Under the influence of the same external factors, different people may have different emotions [26]. Emotions are manifested in different ways and in combination with each other. Recognizing emotions from facial expressions is a complex mental process. A person can recognize another person's emotions by their face, gait and gestures. Ekman, a psychologist in the 1960s, believed that there were only six basic emotions expressed by facial emotions [27,28]. But further research has shown much more variability regarding the number of emotional states and how they are expressed. This varies in different cultures [29], in different situations, and even within the same day.

One of the companies offered police software to detect emotions [30]. The ability to detect emotions such as anger, stress or anxiety provides law enforcement with additional information when conducting a large-scale investigation. Ultimately, it is believed that the responsible use of this technology will be a factor that will make the world safer.

Many job candidates have benefited from HireVue's technology, which aims to help eliminate the very significant human bias in the existing hiring process [31]. Cogito has developed voice analysis algorithms for call center personnel to help them detect when customers are experiencing problems. However, people believe that before emotion detection can make automated decisions, the industry needs more evidence that machines can detect human emotions effectively and consistently [32].

The field of recognition of a person's psycho-emotional state by a machine gained popularity with the advent of machine learning and neural networks. This is directly related to the fact that to perform this task, it is necessary to analyze the image by extracting specific features from it; the same should be performed with the data of other modalities, namely sound and text. Considering unimodal data, the first development stage was using simple fully connected neural networks. However, such data as image, sound, and text carry much more information if we consider not one pixel (for an image) but also its surroundings. This is characteristic of the next stage, namely convolution neural network (CNN) [33]. This neural network performed exceptionally well in emotion recognition based on all three data types, but separately.

However, as mentioned above, data for recognizing a person's psycho-emotional state should be taken with a specific perspective concerning time. For this, at the stage of development of neural network structures, networks CNN long short-term memory (LSTM) [34] were distinguished. Such a structure considers data in a specific environment and at a particular time interval, which is very suitable for recognizing human emotions.

In addition to the abovementioned types of neural networks, there are many others because new ideas and implementations appear every day in machine learning. A lot of those models are very complex but provide excellent accuracy. Unfortunately, complex systems are hard to use for real-time data analysis.

Multimodal data analysis for emotion recognition has made significant strides in recent years, but there are still several gaps and challenges that researchers are working to address. Some of these gaps include:

- While there has been progress in analyzing multiple modalities, integrating these modalities effectively remains a challenge. Combining information from different modalities in a coherent and complementary manner is essential for accurate emotion recognition.
- Emotion data often suffer from class imbalance, where certain emotions are underrepresented compared to others. Additionally, emotions are highly subjective and context-dependent, leading to variability in how they are expressed and perceived across different individuals and cultures.
- Many applications of emotion recognition, such as human–computer interaction and affective computing, require real-time processing of multimodal data. Achieving low-latency, real-time performance while maintaining high accuracy poses a significant technical challenge.

For real-data analysis with different modalities, it is worth accepting something in the middle, in order to implement faster online prediction on a client-server architecture. In particular, transformer structures that include an encoder and a decoder are well suited for that task. These systems are also very user friendly as they are often designed to learn to produce convenient tape output. It is this structure that we will explore further in the work. And also, LSTM is explored for sound and image modality.

In the paper, CNN + LSTM was chosen for our model, because this method is easy to understand and implement with an appropriated level of accuracy. Late fusion in various implementations was chosen as the method of multimodal data fusion. Number of emotions is one of the model's parameters. Bidirectional LSTM with CNN where the first layer is wrapped with a TimeDistributed function is used for face emotion recognition.

3. Materials and Methods

3.1. Data Preparation

The solution implemented in the paper is based on the combination of simple unimodal methods for emotion recognition into one multimodal method. This solution allows a reasonably accurate result to be obtained, for example, in recognizing a person's psycho-emotional state while using the banking system [25]. Also, developing such an information system will make it possible to understand the advantages of multimodal recognition of emotions.

The most straightforward 3 options of the model and information system that can be implemented (Figure 1) are the following:

- Merger of three information streams at the stage when the probability is obtained for three models separately. In this case, we take a constant weighted sum of all three results (Figure 1a).
- Stream merging at the stage when each model has formed a certain set of characteristics for its data stream. In this case, we combine all these characteristics into one array and submit it for further processing with a typical output (Figure 1b).
- Data merging at the beginning of the neural network immediately after pre-processing (Figure 1c).

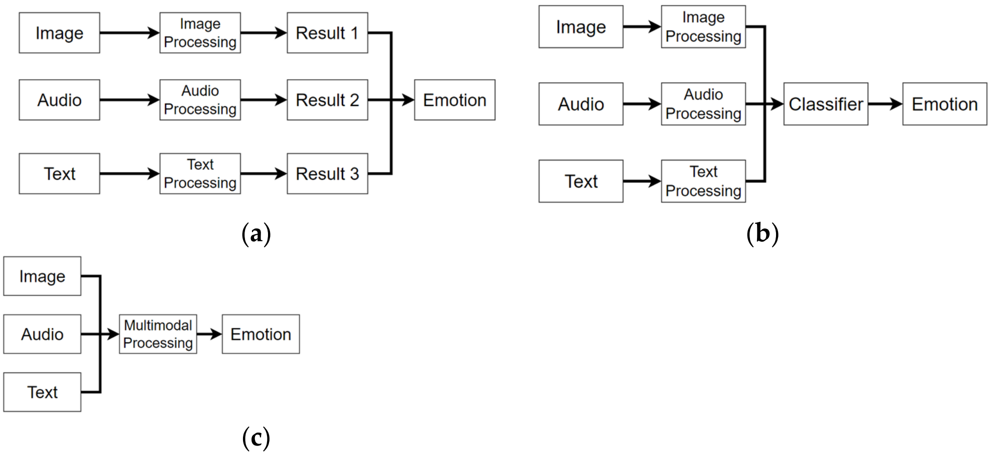


Figure 1. Diagrams of systems with different methods of merging (**a**) merging in the end; (**b**) merging in the middle; (**c**) merging in the start.

Multi-modal emotion recognition (MMER) system merges text, sound and images. The data flow in the system is realized not only from the camera and microphone, but also from the saved recording in the device memory and the screen broadcast. Based on the analyzed requirements, we will use CNN [33] and LSTM [34] models of neural networks for three types of data and different ways of merging them.

We can identify three points where new data arrives. The first moment is when we read data from the camera and microphone (or from a recording or broadcast of the screen) and submit it for processing. The second point is when the data is fed to the input of the neural network for the learning process. The third point is when pre-processed data is fed to a trained network to predict human emotions.

Based on characteristic of supervised learning in [35], we have to enter two data streams, x and y. In our case, y is the probability vector that the emotion belongs to a certain category. And x is either data of one type or an array of data of different types. This is also called the classification problem.

Consider the data for a text emotion recognition (TER) model, i.e., to recognize emotion in text. The data in these datasets are taken from the Kaggle resource [36]. There are about 5–15 words in one fragment of text. The total number of such fragments in the initial data set is 47,291.

Next, consider a similar folder for speech emotion recognition (SER). This is a RAVDESS dataset taken from the Kaggle resource [37]. Files are organized in folders based on different actors who voice emotions. The average recording time in these files is 3–6 min. The original data set consists of 1440 files.

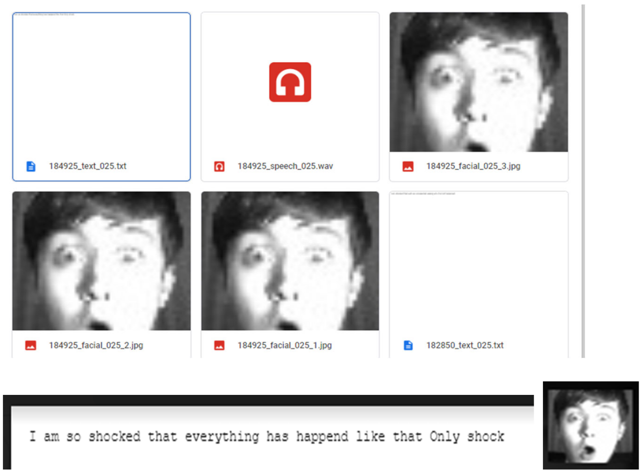
Now let us look at a similar folder for training the neural network facial emotion recognition (FER). This is a data set CK + 48 with 5 emotions [38] taken from the resource Kaggle. Faces are manually cropped in that dataset [39]. Each folder has a sequence of 3 images, which is required for proper training of our LSTM model. Images contain the faces of different people. The size of the image is 48*48 in grayscale. The total number of images in this initial data set is 765. The Fer13 dataset [40] contains 8 emotions. This dataset can be used for testing the efficiency of the proposed model.

The amount of data in the sets can be easily increased due to the implemented functionality of data regeneration from other sets or from the live data stream from the camera and microphone (or from recording or broadcasting the screen). The preprocessing is used for that. At the output of the camera, we have an image in three RGB colors. We need to convert it to grayscale, find and cut out the human face, and reduce the image to 48*48 dimensions, as in the dataset we trained the neural measurer on.

In the end, unimodal datasets consist of:

- FER: 765 images from CK + 48 with surprise, sad, happy, fearful and angry expressions; calm emotion was added from 162 manually cropped images from live-stream (camera and YouTube (San Mateo, CA, USA)); in addition, the dataset was augmented using data augmentation generative adversarial network (DAGAN [41]) with 186 images;
- SER: 1440 files from RAVDESS with surprise, sad, happy, fearful, angry, calm and disgust expressions; disgust emotion files were eliminated;
- TER: 47,291 short texts with surprise, sad, happy, fearful, angry and calm emotions.

Unimodal datasets are used for unimodal neural network training. Then, the pre-trained neural networks are used in merged neural networks. To train a multimodal system with neural networks, the functionality of processing and combining data of three types from these three data sets was implemented. In the first step, data from TER and SER were combined according to the same meaning of the text and the meaning of the emotion. In the second step, a sequence of 3 images with the same emotion value was added. This sequence is generated from 5–10 s video as 3 separated images with some difference between images. The dissimilarity is found using the package imagehash. The final dataset consists of 1167 FER + SEP + TER triplets and 162 triplets generated from live streaming and YouTube. The division into folders is implemented similarly to CK + 48 by emotions. An example of one package consisting of a text file, an audio recording and three images of an emotional face is shown in Figure 2.



I am so shocked that everything has happend like that Only shock

Figure 2. An example of one dataset for the emotion of surprise (https://doi.org/10.6084/m9.figshare.23596362.v1, accessed on 1 September 2022).

Of the dataset, 70% is used for training and 30% for testing.

It is possible to feed each type of data separately to its corresponding neural network. There is also a similar implementation in the case when this network is part of a large network with a merging in the middle (Figure 3a). Another presentation option is data fusion right at the beginning with image preprocessing for the adequacy of this fusion, and already a joint presentation at the entrance to the network (Figure 3b).

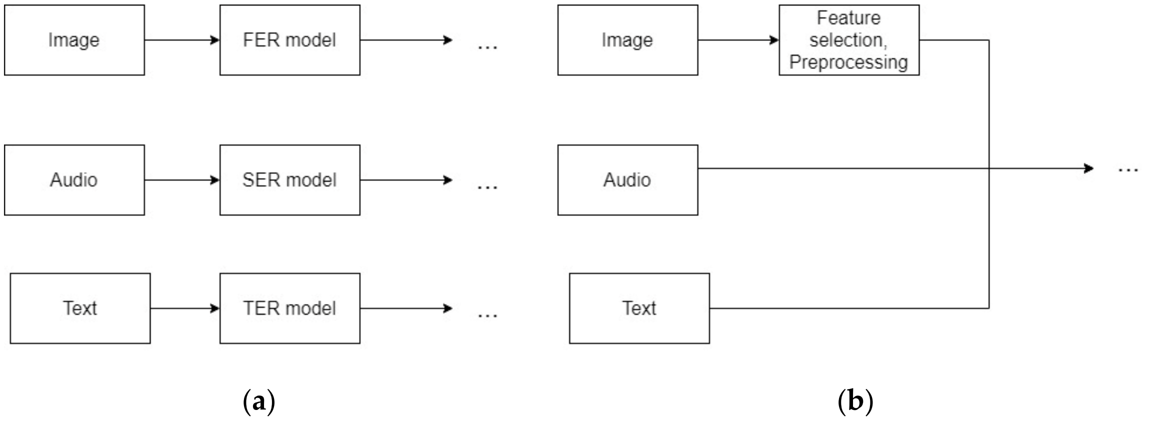


Figure 3. Input different types of data into their respective models: (**a**) to feed each type of data separately to its corresponding neural network; (**b**) data fusion right at the beginning with image preprocessing for the adequacy of this fusion.

3.2. Model Development

There are three cases of multimodal data analysis. In the first case, the multimodal emotion recognition model with late fusion is the structure organized as follows: the results of three modalities are obtained from the corresponding unimodal networks and then added with the same coefficient of influence on the final result (0.33). In the second realization, the multimodal emotion recognition model with middle fusion is the structure realized as follows: the characteristics of the three modalities obtained in the corresponding unimodal networks are added to the softmax layer and processed in several fully connected layers, which allows a well-balanced joint result to be obtained. The latter option, multimodal emotion recognition with early fusion, is organized as a structure where different types of data streams merge as early as possible and are processed as one common unimodality.

All previously mentioned options correspond to the late merger in the concept of methods of multimodal recognition of the psycho-emotional state, without modifications that take into account intermodal connections. As a result of the system, we expect to have a vector of probabilities that the emotion belongs to a certain category, or classification problem.

Once we have the data, we need to have neural networks. In our case, there are 5 neural networks for 3 systems for recognizing a psycho-emotional state. The implementation of those networks is presented below.

In the first system, we have 3 separate neural networks, each of which is responsible for its own type of input. But at the output, the prediction results of 3 neural networks merge into one.

The first network is FER—a neural network for facial emotion recognition. The architecture of this neural network is based on CNN and bidirectional LSTM (Figure 4) where the first layer is wrapped with a TimeDistributed function. The TimeDistributed layer requires a minimum of 3 dimensions. The objective is to consider the sequences of three images as a single input data because every sequence contains necessary information. Six emotions are taken into account.

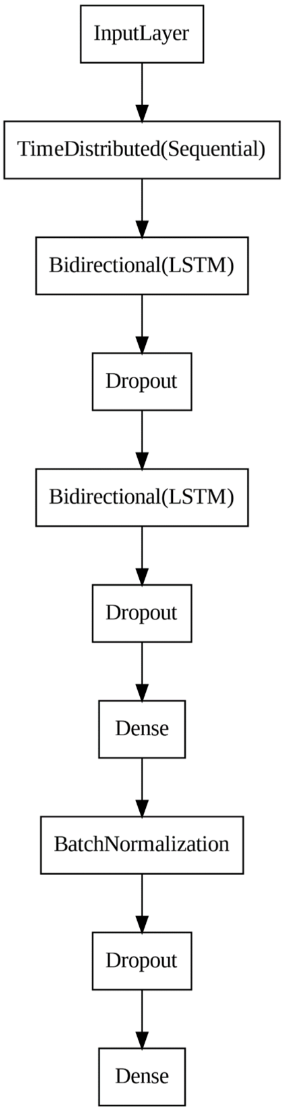


Figure 4. The structure of the FER neural network.

In this and all other neural networks, the size of the input and output layers of the network are parametric. This allows you to quickly change the number of emotions with which the neural network works.

The next speech emotion recognition (SER) network is a network for recognizing emotions by the sound a person makes when talking. This network has the simplest and small architecture. It allows you to quickly and effectively train and predict emotion based on the data received in the "live" mode. CNN is used here. The structure of the network is shown in Figure 5. Emotions are characterized by a different set of jumps in the spectrogram and different acoustic features.

Mel-spectrogram analysis is used for recognition of emotions. According to [42], a context window of 64 frames is adopted to extract the static Mel-spectrogram. The frame shift size of 32 frames is used to generate overlapping segments of the Mel-spectrogram. The overlapping segments is the key to speech segmentation. As a result, the static Mel-spectrogram is obtained at the size of 64*64. Therefore, 64 cores help to find such fluctuations.

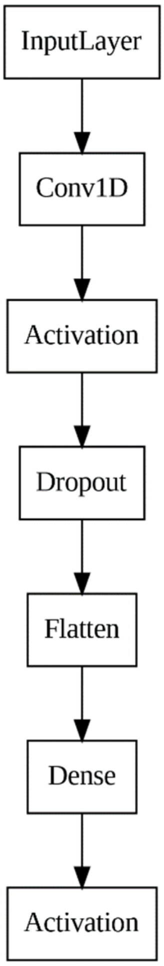


Figure 5. The structure of the SER neural network.

The last network in this system is TER—a network for recognizing emotion from text. This network has a very complex and bulky architecture. This is due to the general difficulty of determining emotions by text. CNN and LSTM are used here. The structure of this network is presented in Figure 6. A maxpooling layer is used after vertical concatenation.

The next system with a merger in the middle is implemented according to the principle of combining the 3 neural networks mentioned above with beige "softmax" layers (MMERm—multimodal emotion recognition with merge at the middle). That is, at the moment of obtaining certain characteristics. CNN LSTM is used here, accordingly. The structure of this neural network is cumbersome and is shown in Figure 7.

Merge at the end (MMERe—multimodal emotion recognition with merge at the end) is developed as the structure when the obtained results of three modalities, in the corresponding unimodal networks, are added with the same coefficient of influence on the final result (0.3 (3)). To do this, we use the previous models (FER, TER and SER).

The last version of the system is implemented according to the principle of merging the neural network immediately at the input (MMERs—multimodal emotion recognition with merge at start). A set of fully connected layers is used here. The idea of such a structure was to merge the data into one row at the very first step and highlight hidden dependencies and characteristics with the help of several fully connected layers. First of all, this network will be used as an experiment with various structures that can only be invented. The structure of this neural network is shown in Figure 8.

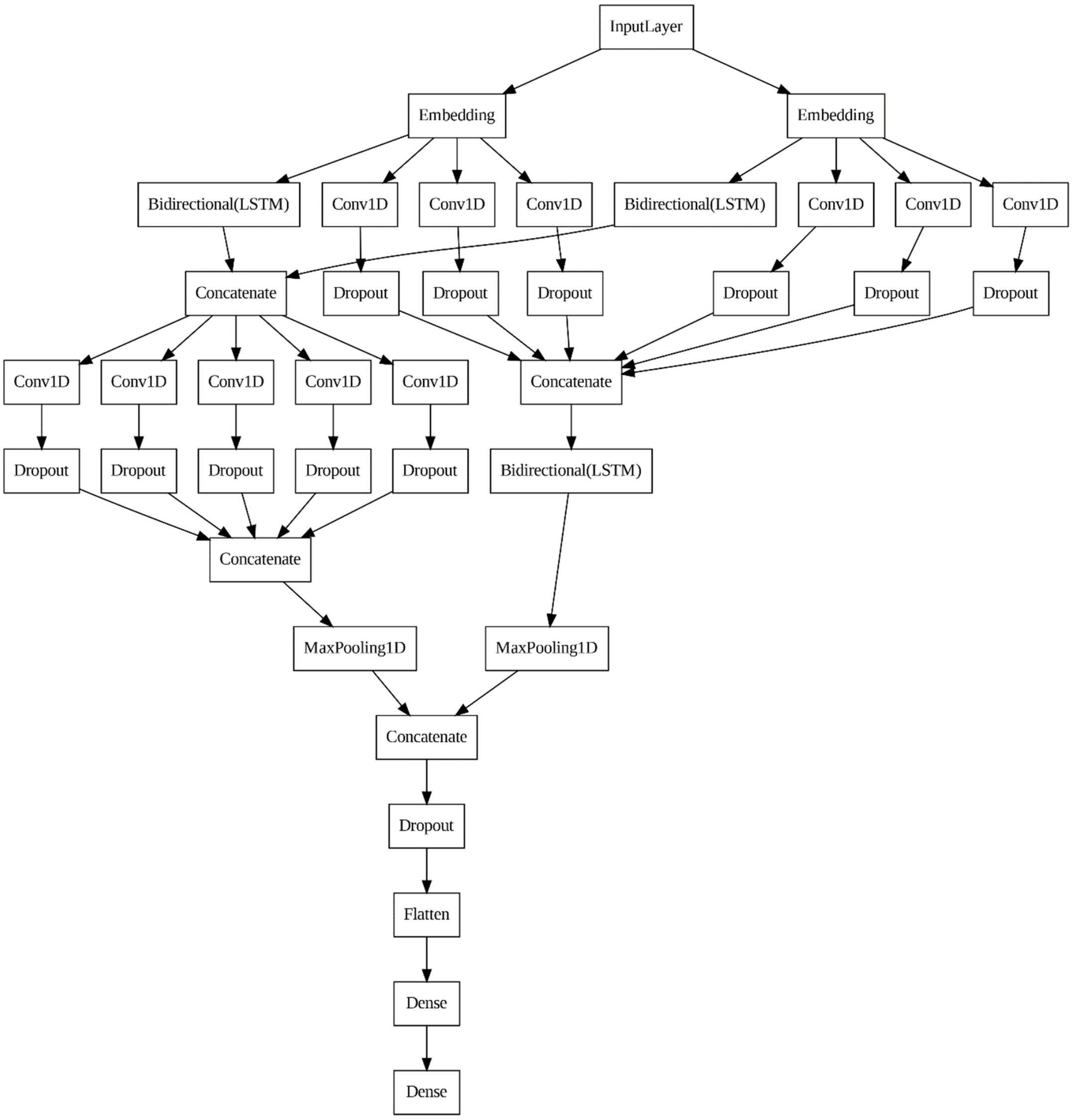


Figure 6. The structure of the TER neural network.

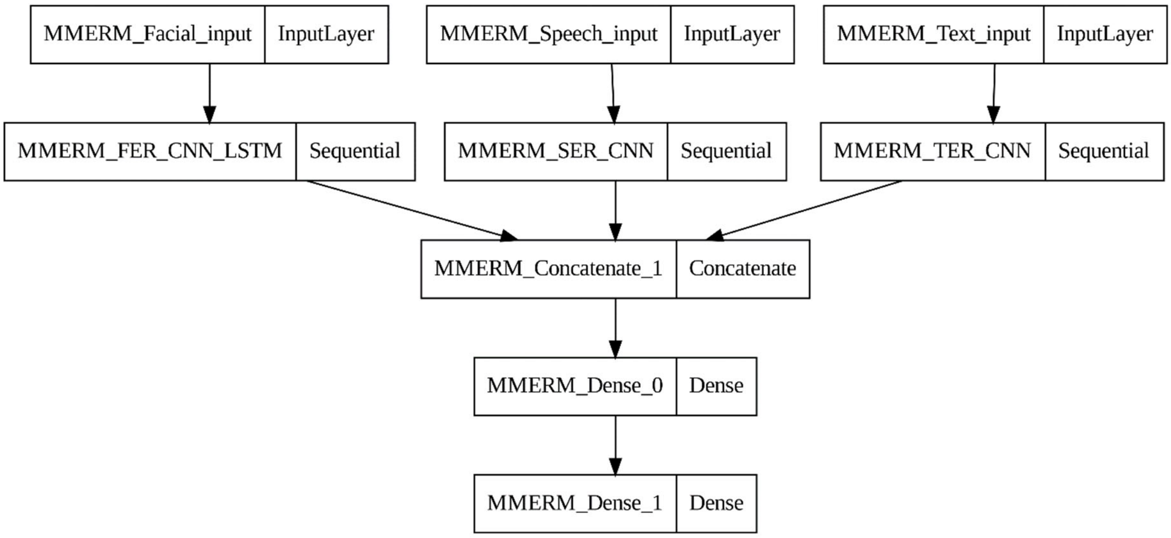

MMERM_Facial_input | InputLayer
MMERM_Speech_input | InputLayer
MMERM_Text_input | InputLayer
MMERM_FER_CNN_LSTM | Sequential
MMERM_SER_CNN | Sequential
MMERM_TER_CNN | Sequential
MMERM_Concatenate_1 | Concatenate
MMERM_Dense_0 | Dense
MMERM_Dense_1 | Dense

Figure 7. The structure of the MMERm neural network.

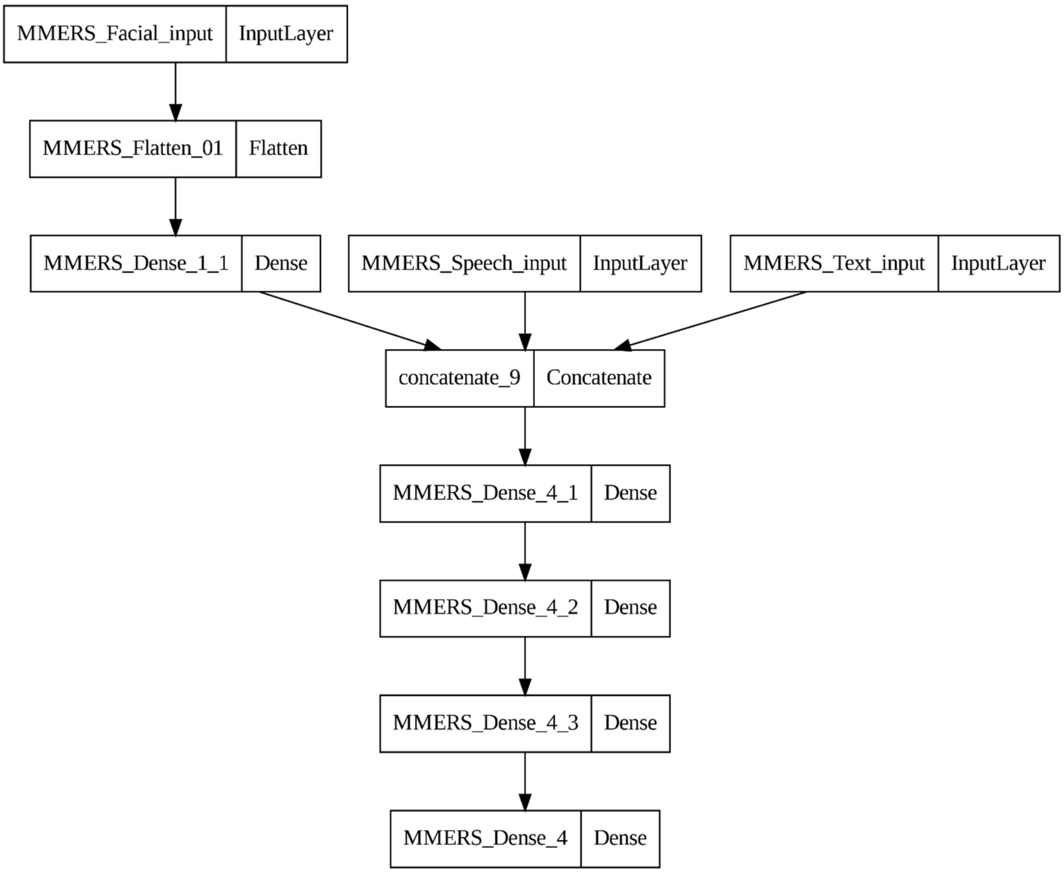

MMERS_Facial_input | InputLayer
MMERS_Flatten_01 | Flatten
MMERS_Dense_1_1 | Dense
MMERS_Speech_input | InputLayer
MMERS_Text_input | InputLayer
concatenate_9 | Concatenate
MMERS_Dense_4_1 | Dense
MMERS_Dense_4_2 | Dense
MMERS_Dense_4_3 | Dense
MMERS_Dense_4 | Dense

Figure 8. The structure of the MMERs neural network.

4. Results

The system was developed based on the following technologies algorithms and frameworks:

- CNN, LSTM
- Dropout, MaxPooling, Flatten, Dense, Reshape, GlobalMaxPool2D, BatchNormalization, Bidirectional, TimeDistributed, Embedding;
- Nadam;
- Relu, softmax, elu;
- EarlyStopping, ReduceLROnPlateau;
- Accuracy, categorical_crossentropy;
- Precision, recall, f1-score, support.

In addition:

- PCA, TSNE;
- OpenCV 4.8.1 (convertScaleAbs, LUT);
- Ffmpeg 0.2.0 -python (ffmpeg.input);
- Librosa 0.10.1 (feature.mfcc);
- SpeechRecognition (speech_recognition.recognize_google);
- Tokenizer (texts to sequences).

For results evaluation, the following metrics are used: accuracy, categorical_crossentropy, precision, recall, f1-score.

Accuracy is the percentage of correctly classified data (*tp*—true positive, *tn*—true negative, *fp*—false positive, *fn*—false negative):

$$\mathrm{A} = \frac{tp + tn}{tp + tn + fp + fn}.$$

Categorical cross-entropy is a loss function that calculates the difference between two probability distributions: the observed distribution and the predicted probability distribution. In the case of classification, the input is fed to the input of the model, which generates a vector of distribution probabilities across all classes. The next step is to compare this vector with the vector of actual values and calculate the cross-entropy between the two vectors.

The basic idea behind categorical cross-entropy is that we penalize a model if its predicted probabilities differ from the actual values. Therefore, the goal of optimization is to reduce this loss by adjusting the model parameters.

Because neural networks are used not only in the learning process, where logits immediately hit the evaluation function but also to show the percentage of predicted emotion, the result normalized by the Softmax layer is more presentable and understandable. If we consider the context of the study of structures on metrics, the softmax layer can be omitted and specify the CrossEntropyloss function with the from_logits = true parameter, in which case the internal code would do the same as the softmax layer. Therefore, this decision has no effect on the result. The separate layer was made for convenience in our context.

Precision is the proportion of relevant instances among the extracted instances:

$$P = \frac{tp}{tp + fp}.$$

Recall is the proportion of relevant instances that were retrieved:

$$R = \frac{tp}{tp + fn}.$$

$F1$-score combines precision and recall and takes their harmonic mean. It is used to compare the effectiveness of classifications:

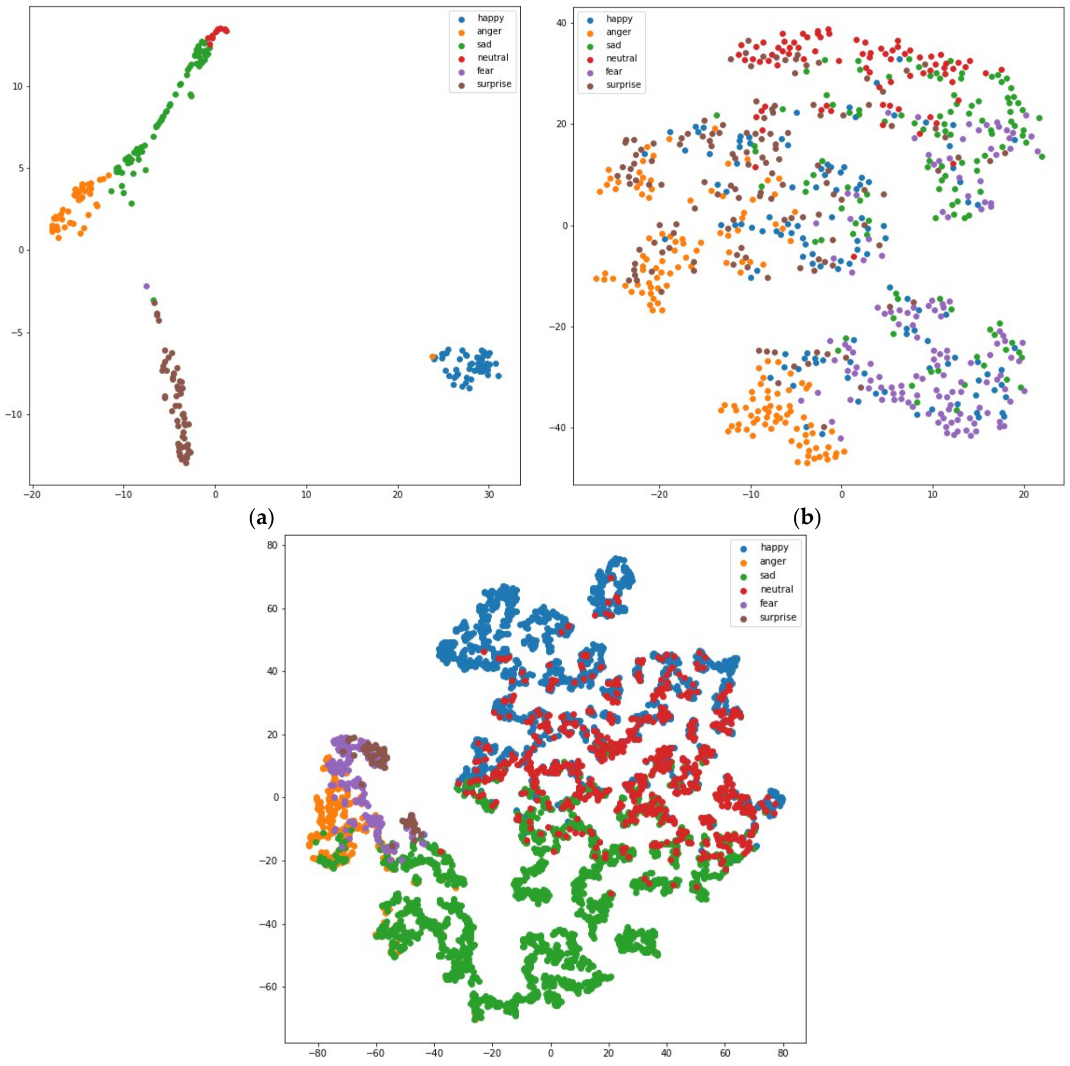
$$F1 = \frac{2 \times (P \times R)}{P + R}.$$

Python 3 Google Compute Engine backend (GPU) was used for experiments performing.

To begin, let us look at detailed results for training and testing using six emotional states.

For qualitative assessment, TSNE derivation is also implemented for the characteristics formed in each network at the penultimate layer of the neural network during prediction. The program uses "PCA" and "TSNE" implementations for this. For example, on the validation data for five neural networks, we will have the graphs shown in Figures 9 and 10. Therefore, the cluster density for FER is the best compared with the rest of the unimodal models. The SER model demonstrates clusters overlapping. MMERm and MMERs demonstrate appropriate cluster density too.

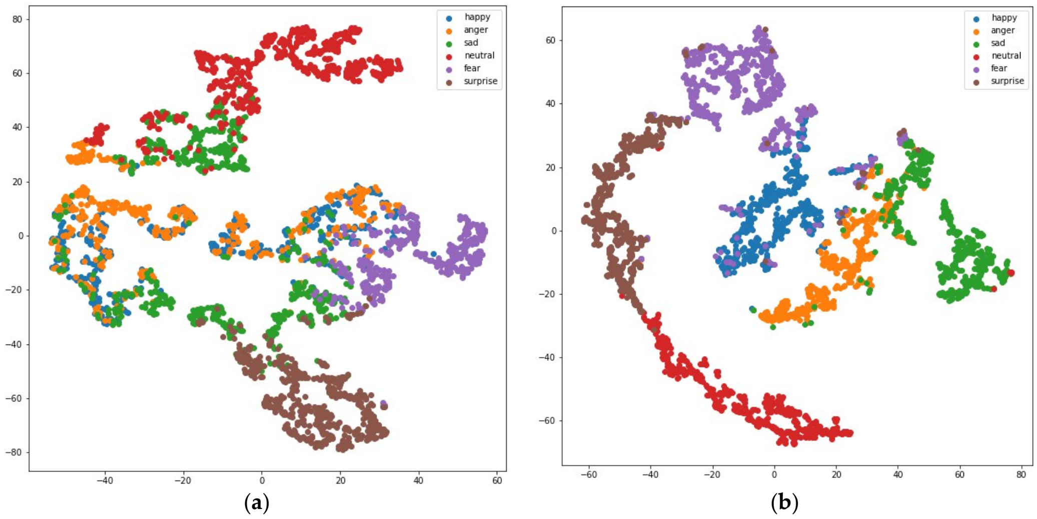
And finally, we visualize the TSNE graphs for three architectures for test dataset (Figures 11 and 12). TSNE converts similarities between data points to joint probabilities and tries to minimize the Kullback–Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data (based on scikit-learn version 0.16.1 documentation [43]). The obtained result is the same. The best result is for MMERm, MMERs and FER. SER presents clusters overlapping. Anger emotion is well recognized based on MMERm.



(**a**) (**b**)

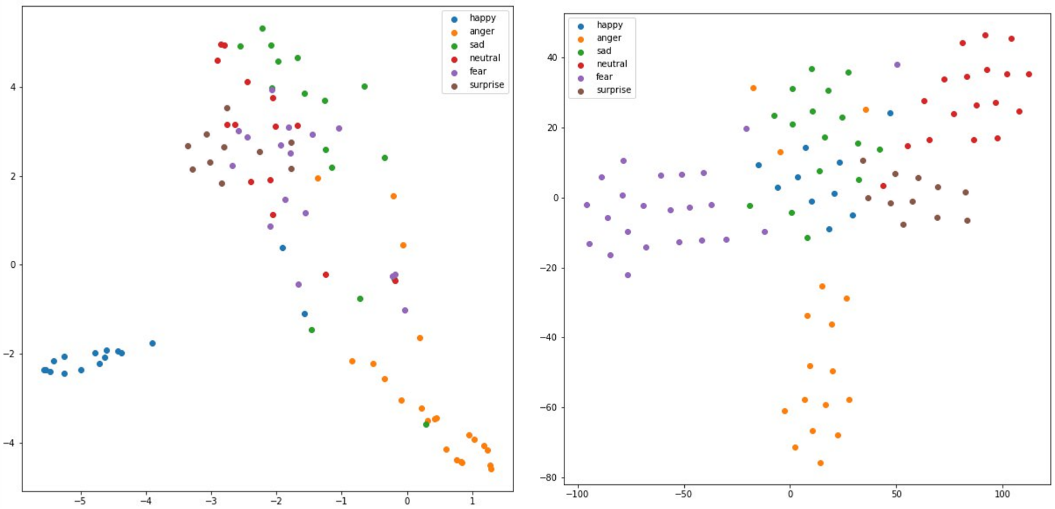
(**c**)

Figure 9. TSNE for FER (**a**), SER (**b**), TER (**c**)—training dataset.



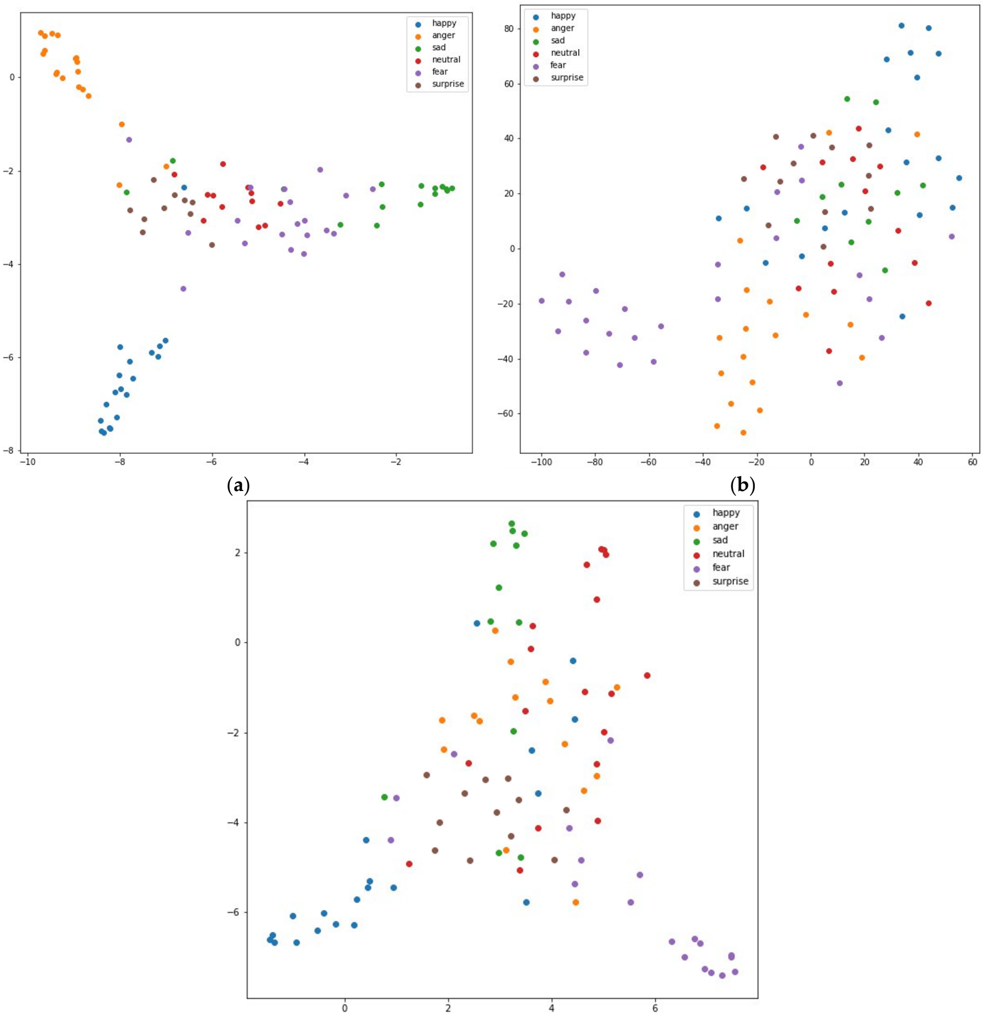
(**a**) (**b**)

Figure 10. TSNE for MMERm (**a**) and MMERs (**b**)—training dataset.



(**a**) (**b**)

Figure 11. TSNE for MMERs (**a**), MMERm (**b**)—testing dataset.



(**a**) (**b**)

(**c**)

Figure 12. TSNE for FER (**a**), SER (**b**), TER (**c**)—testing dataset.

Based on these three conclusions, we are able to qualitatively analyze the results of testing in "live" mode.

First, Table 1 shows the resulting accuracy of five neural networks. The highest validation accuracy based on multimodal data is obtained for MMERS. However, multimodal accuracy is still worse than the accuracy for unimodal FER. FER, SER and TER were trained on a unimodal separated dataset, then the pretrained neural networks are used in merged neural networks.

Table 1. Table of neural network training results.

	MMERe			**MMERm**	**MMERs**
	FER	**SER**	**TER**		
val_accuracy	0.9245	0.4363	0.6416	0.7000	0.7347
val_loss	0.4402	0.9542	0.7048	0.4647	0.6396
train_accuracy	1.0000	0.6544	0.8209	0.9583	0.9774
train_loss	0.0267	0.8468	0.4675	0.0732	0.05832

Table 2 shows the metrics of testing on validation data. Metric values will be presented as "weighted avg". However, metrics "micro avg", "macro avg" and "samples avg" are also available for analysis. The best result is obtained for FER (video modality) and the worst result is obtained for SER (audio modality). The difference in accuracy across modalities is explained by the difference in the quality of the datasets. So, for images, a sequence of three images is analyzed. A convolutional neural network works well for recognizing individual images. Moreover, accuracy increases with combined results of CNN and LSTM for sequential images. Speech recognition is performed for a dataset consisting of audio files 4 s long, due to which the end of the sentence to determine the emotion is not always correct. The same situation exists for the text dataset. At the same time, the combination of three modalities makes it possible to significantly improve the result.

Table 2. Table of test results of networks after training.

	MMERe			**MMERm**	**MMERs**
	FER	**SER**	**TER**		
precision	0.92	0.56	0.64	0.92	0.90
recall	0.92	0.56	0.64	0.90	0.89
f1-score	0.92	0.56	0.64	0.91	0.89
support	106	319	16377	120	720

For a better understanding, we display this data in the form of a bar chart in Figure 13.

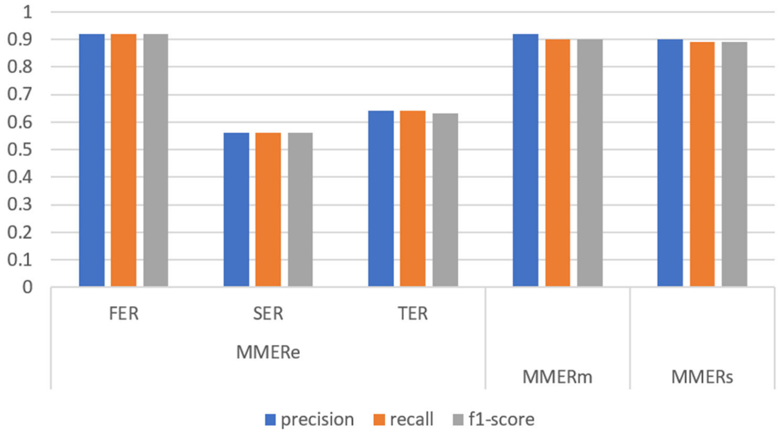
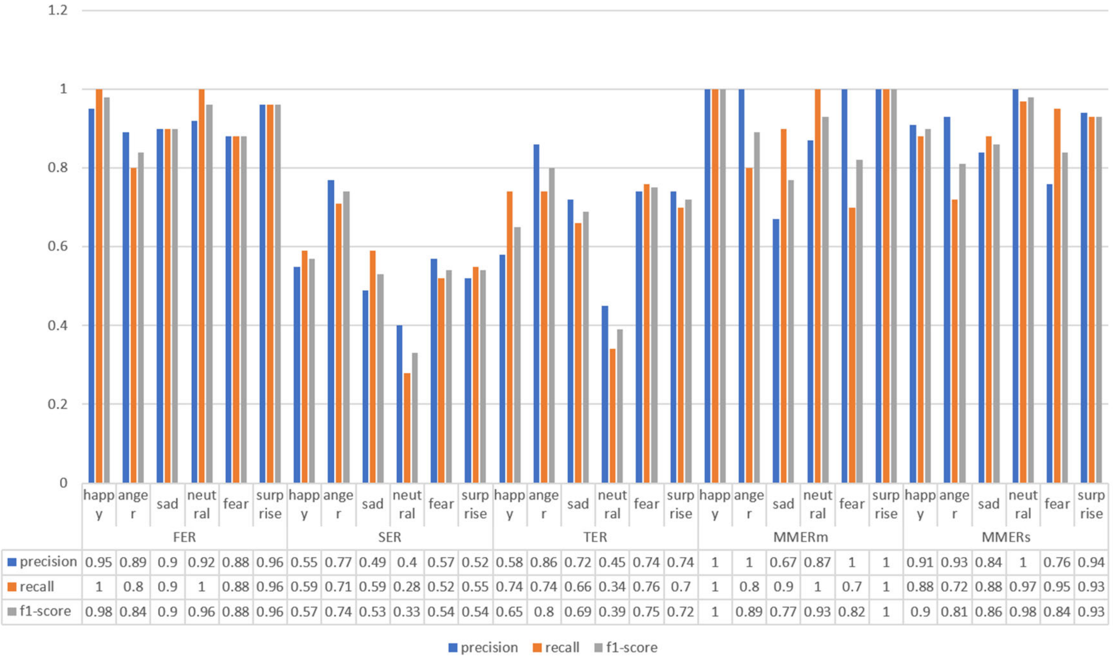


Figure 13. Bar chart of average training results.

The value of such metrics can be obtained for each emotion separately.

For a better understanding, we display this data in the form of a bar chart in Figure 14.



	FER						SER						TER						MMERm						MMERs					
	happy	anger	sad	neutral	fear	surprise	happy	anger	sad	neutral	fear	surprise	happy	anger	sad	neutral	fear	surprise	happy	anger	sad	neutral	fear	surprise	happy	anger	sad	neutral	fear	surprise
■ precision	0.95	0.89	0.9	0.92	0.88	0.96	0.55	0.77	0.49	0.4	0.57	0.52	0.58	0.86	0.72	0.45	0.74	0.74	1	1	0.67	0.87	1	1	0.91	0.93	0.84	1	0.76	0.94
■ recall	1	0.8	0.9	1	0.88	0.96	0.59	0.71	0.59	0.28	0.52	0.55	0.74	0.74	0.66	0.34	0.76	0.7	1	0.8	0.9	1	0.7	1	0.88	0.72	0.88	0.97	0.95	0.93
■ f1-score	0.98	0.84	0.9	0.96	0.88	0.96	0.57	0.74	0.53	0.33	0.54	0.54	0.65	0.8	0.69	0.39	0.75	0.72	1	0.89	0.77	0.93	0.82	1	0.9	0.81	0.86	0.98	0.84	0.93

Figure 14. Bar chart of training results for each emotion.

In addition to validation training, it is important to analyze live test results, as these are important in our system of multimodal recognition of a person's psycho-emotional state in the data stream from the camera and microphone or recording or broadcasting the user's screen.

First, we present the results of various metrics with averaging "weighted avg" (Table 3).

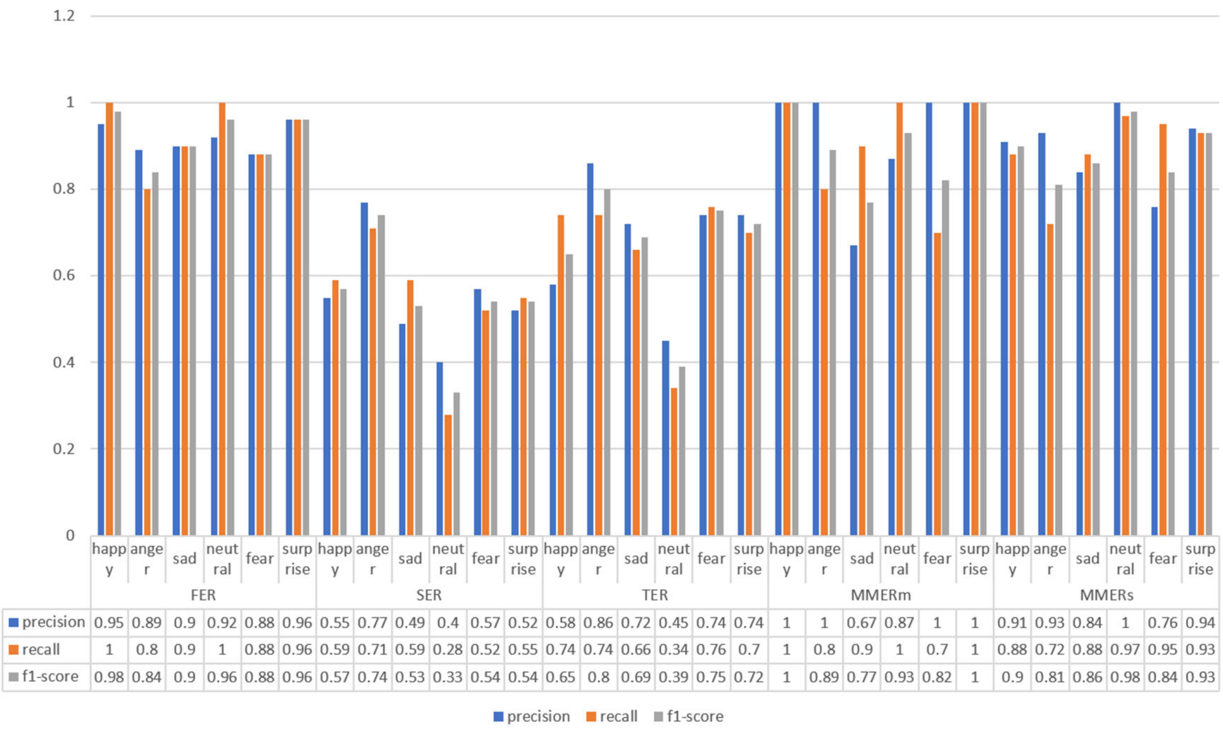
Table 3. Table of test results of networks after testing.

	MMERe			**MMERm**	**MMERs**
	FER	**SER**	**TER**		
precision	0.78	0.54	0.67	0.71	0.62
recall	0.66	0.48	0.63	0.62	0.58
f1-score	0.715	0.51	0.65	0.66	0.60
support	67	31	29	62	52

The results of the same metrics for each emotion separately are given in Figure 15. Happy emotion is the best recognized by FER, SER, MMERm and MMERs. Neutral emotion is the hardest emotion for recognition by SER and TER.

We will also show an alternative version of the visualization of test results for each emotion in Figure 16.

Next, the training time for different numbers of emotions and accuracy were analyzed. All neural networks were trained on the same dataset. Figure 17 presents time dependence of number of emotions. The maximum number of emotions for multimodal data is six because the same number of emotions is used in the FER dataset.



	FER						SER						TER						MMERm						MMERs					
	happy	anger	sad	neutral	fear	surprise	happy	anger	sad	neutral	fear	surprise	happy	anger	sad	neutral	fear	surprise	happy	anger	sad	neutral	fear	surprise	happy	anger	sad	neutral	fear	surprise
■ precision	0.95	0.89	0.9	0.92	0.88	0.96	0.55	0.77	0.49	0.4	0.57	0.52	0.58	0.86	0.72	0.45	0.74	0.74	1	1	0.67	0.87	1	1	0.91	0.93	0.84	1	0.76	0.94
■ recall	1	0.8	0.9	1	0.88	0.96	0.59	0.71	0.59	0.28	0.52	0.55	0.74	0.74	0.66	0.34	0.76	0.7	1	0.8	0.9	1	0.7	1	0.88	0.72	0.88	0.97	0.95	0.93
■ f1-score	0.98	0.84	0.9	0.96	0.88	0.96	0.57	0.74	0.53	0.33	0.54	0.54	0.65	0.8	0.69	0.39	0.75	0.72	1	0.89	0.77	0.93	0.82	1	0.9	0.81	0.86	0.98	0.84	0.93

Figure 15. Bar chart of test results for each emotion.

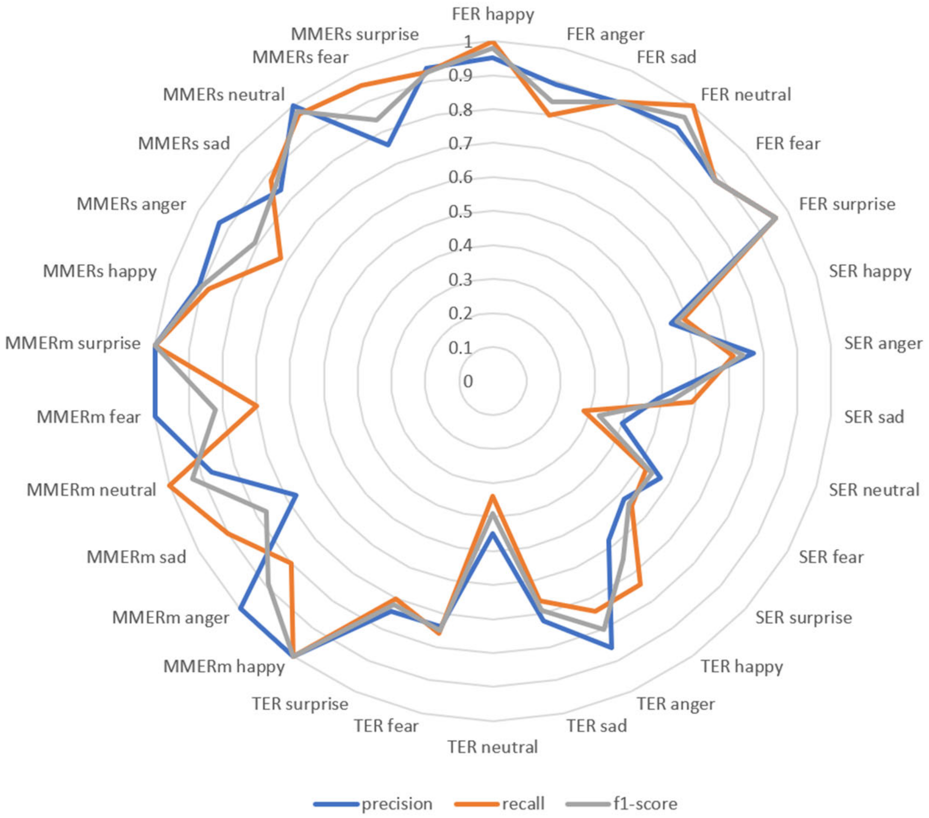


Figure 16. Drawing test results for each emotion.

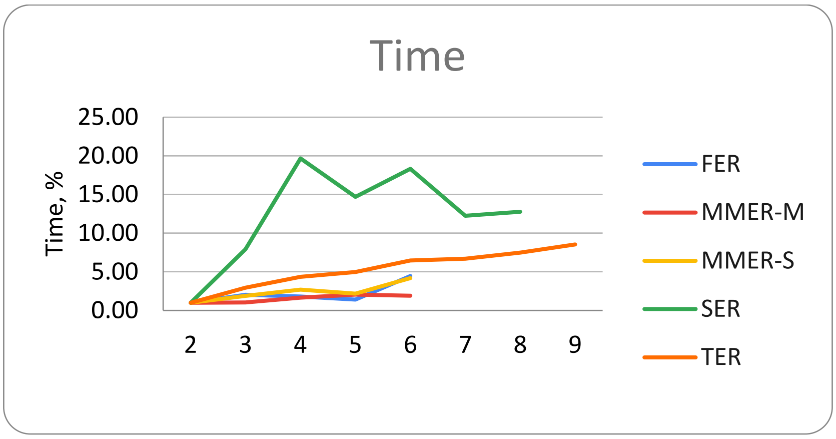


Figure 17. Training time dependence from number of emotions.

More detailed information is presented in Tables 4 and 5.

Table 4. Accuracy and training time for different numbers of emotions (from nine to six emotions).

Model	Number of Emotions											
	9			8			7			6		
	Time (s)	Accuracy (train/val)	Epochs	Time (s)	Accuracy (train/val)	Epochs	Time (s)	Accuracy (train/val)	Epochs	Time (s)	Accuracy (train/val)	Epochs
FER	-	-	-	-	-	-	-	-	-	310.12	1.0/0.92	464
SER	-	-	-	36	0.67/0.47	138	34.74	0.71/0.55	141	52.03	0.72/0.55	250
TER	725.43	0.75/0.66	48	634	0.75/0.65	45	567.93	0.74/0.66	45	547.73	0.74/0.65	45
MMER-S	-	-	-	-	-	-	-	-	-	29.51	1.00/0.97	78
MMER-M	-	-	-	-	-	-	-	-	-	195.6	0.99/0.85	18

Table 5. Accuracy and training time for different numbers of emotions (five to two emotions).

Model	Number of Emotions											
	5			4			3			2		
	Time (s)	Accuracy (train/val)	Epochs	Time (s)	Accuracy (train/val)	Epochs	Time (s)	Accuracy (train/val)	Epochs	Time (s)	Accuracy (train/val)	Epochs
FER	96.8	0.87/0.86	157	125.61	0.99/0.88	222	140.25	1.0/0.85	347	69.57	1.0/1.0	196
SER	41.71	0.73/0.64	235	55.85	0.77/0.66	346	22.45	0.80/0.75	176	2.84	0.59/0.75	21
TER	421.23	0.74/0.65	36	369.32	0.86/0.78	38	249.9	0.97/0.94	40	84.88	0.98/0.95	66
MMERs	15.28	0.96/0.90	43	18.94	1.00/0.94	71	13.32	1.00/0.89	57	7.062	1.00/0.94	39
MMERm	211.58	0.98/0.82	22	170.03	0.99/0.95	24	105.82	1.00/0.95	16	102.76	0.99/1.00	26

The list of emotions is presented below:

9 “neutral”, “happy”, “sad”, “hate”, “anger”, “joy”, “fear”, “love”, “surprise”
8 “neutral”, “happy”, “sad”, “anger”, “joy”, “fear”, “love”, “surprise”
7 “neutral”, “happy”, “sad”, “anger”, “fear”, “love”, “surprise”
6 “neutral”, “happy”, “sad”, “anger”, “fear”, “surprise”
5 “neutral”, “happy”, “sad”, “anger”, “surprise”
4 “happy”, “sad”, “anger”, “surprise”
3 “sad”, “anger”, “surprise”
2 “anger”, “surprise”

From Tables 4 and 5, we can make a conclusion about the dependence of learning time on the number of emotions. Also, the accuracy of recognition slightly decreases with the increase in the number of emotions. Although for six emotions, the accuracy is higher than for five. This situation is explained by the features of the data sets and the number of emotions that were basically presented in them.

The difference in train and validation accuracy for different numbers of emotions can be explained in the following way. Some emotions are similar and that is why it is very hard to recognize them based on only one modality. From the other side, MMERs looks the best in comparison with the rest of the models.

The next experiment was conducted based on size of images. The purpose of the experiment was to test the hypothesis of whether a larger image size would affect the accuracy of emotion classification. To conduct the experiment, two options for changing the data were used. The first option was to artificially increase the image size of the original CK + 48 dataset to 192*192. The second way was to use new data obtained from a real data stream (noted with symbol +) in which the frame size was 192*192. Also, for comparison, the new images were reduced to 48*48. Accordingly, for the analysis of new data, the FER architecture in Figure 4 has also been resized to fit the image. The change in the training time of the FER neural network depending on the image size was also analyzed.

Early stopping is used to increase the efficiency of training neural networks and to avoid overfitting. On other hand, one of the drawbacks of early stopping is that it can be sensitive to the choice of the validation set, the criterion, and the threshold or the patience parameter. To analyze the quality of developed FER NN, three early stopping criteria were implemented based on:

- Validation loss,
- Number of epochs,
- Accuracy.

The first experiment was conducted based on validation loss as the early stopping criteria. Figures 18 and 19 represent validation accuracy and training time, respectively, for different numbers of emotions. The comparison in Figure 18 is presented on an absolute scale. The time percentage in Figure 19 expresses the training time increasing when comparing two emotions. In this figure, the relative time for two emotions is equal to one, the relative time of the rest of the emotions is presented as relative time increasing.

In Figure 19, for three emotions and for six emotions training time is higher for smaller images (48*48) than for bigger images (192*192). One possible explanation for that is lower recognition rate for smaller images. The classification accuracy (Figure 18) on the new 192*192 images is slightly lower than on the original SK + 48 dataset. Obviously, for artificially reduced or artificially enlarged images, the classification accuracy is much lower. The analysis of computational complexity (Figure 19) showed the dependence of time on the number of emotions and the size of images.

The number of epochs was chosen as a second experimental stopping criteria. Due to limited computation resources for multiple experiments, the number of epochs is equal to 20. The validation accuracy and validation loss are given in Figure 20. Based on the presented results, the decision is that the number of epochs is not sufficient to train the models, especially for FER (192*192).

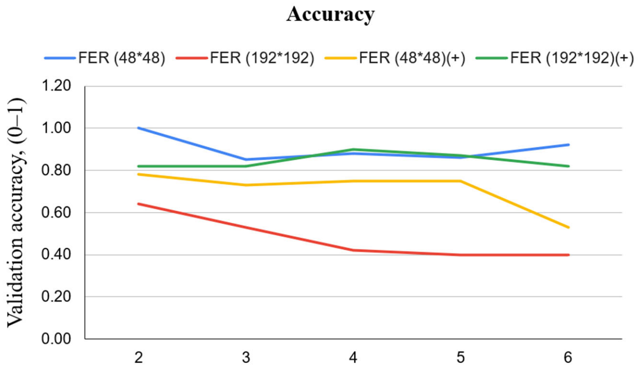


Figure 18. Accuracy dependence from number of emotions and image size for FER.

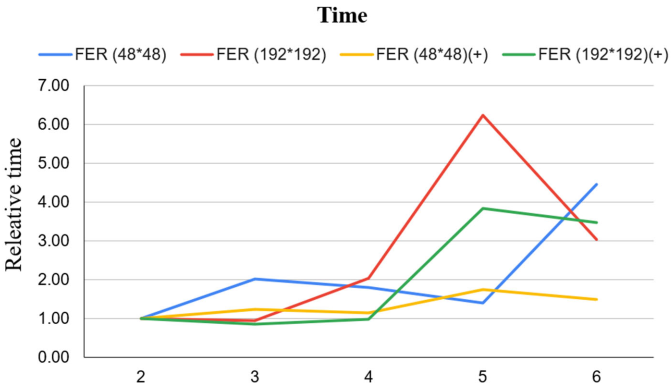
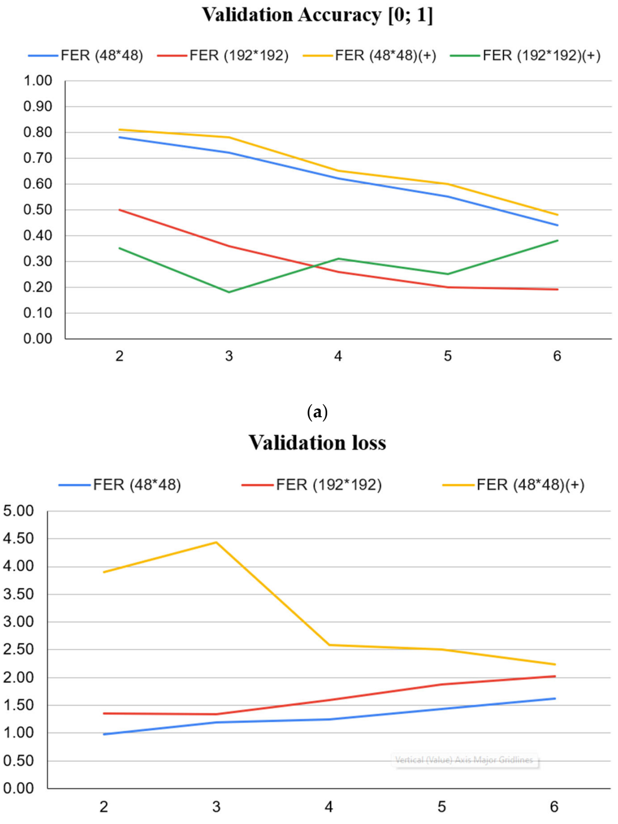


Figure 19. Training time dependence from number of emotions and image size for FER.



(**a**)

(**b**)

Figure 20. Validation accuracy (**a**) and loss (**b**) for 20 epochs.

The next stopping criteria is built on accuracy. The value 0.4 is chosen for experimental purposes only and used for time evaluation and evaluation of the number of epochs (Table 6). Thus, the increasing number of epochs and training time depending on number of emotions and image size is presented.

Table 6. Training time and number of epochs to obtain accuracy equal to 0.4 or high.

Number of Emotions/ Model	FER (48*48)		FER (192*192)	
	Training Time	Number of Epochs	Training Time	Number of Epochs
6	9.55	33	14.84	26
5	8.83	28	8.3	12
4	6.33	13	6.03	6
3	4.62	2	4.66	2
2	4.42	1	4.48	2

Based on experiments presented in Figures 18–20, the best accuracy is obtained using early stopping criteria on loss. The training time depends on number of emotions, size of images and number of epochs.

The next experiment was a study of the dependence of classification accuracy on the number of images in the sequence (sequence length of the frames was used as input of the LSTM) used in FER. As can be seen from Table 7, increasing the number of images does not significantly affect the classification accuracy.

Table 7. Accuracy for different numbers of emotions and numbers of images in the sequence.

Emotion Number/ Validation Accuracy (%)	FER (48*48) + 3 Images	FER (48*48) + 5 Images
6	0.920	0.915
5	0.864	0.865
4	0.882	0.893
3	0.854	0.863
2	1.000	1.000

5. Discussion

From the results shown in Tables 1–3 and Figures 13–17, the following conclusions can be drawn.

Among the individual neural networks of the first MMERe system, the best and fairly good accuracy in the FER network is 0.9245. However, due to the low accuracy of SER and TER—0.5611 and 0.64, this system will show not so good quality in recognition based on multimodal data.

The other two options for merging data, MMERm and MMERs, trained better. The average accuracy of the models is 0.90 and 0.89, respectively.

Number of emotions is one of the model's parameters and can be easily changed.

However, there is a high probability that the networks in the last two systems have been retrained. This is confirmed by the data we received after testing in Table 3. In addition to the problems in the testing process mentioned above, one we must also highlight is the possible impact of fuzzy pronunciation on records, poor lighting, and other details that affect data quality.

The future research will be connected with MOSEI [44] dataset usage and comparison with existing results. The next step will be to find additional computing resources for training networks with the ability to work with images with a higher resolution than 48*48.

6. Conclusions

This work demonstrates the use of various simple methods of late fusion of multimodal data to achieve the best recognition accuracy. Based on the analysis, it was determined that the best among the proposed systems is the system with data fusion after processing each channel separately and obtaining individual characteristics. As the conclusion, we would like to note that marking the data obtained in live testing is a very long and difficult process. First, there is the subjective understanding of what the emotion is, which may not coincide with reality due to the similarity of the manifestations of psycho-emotional states and specifics depending on the author in the video and the circumstances. Secondly, for the amount of data that will allow adequately assessing the results, it is necessary to carry out an extremely long process of marking the data, which must be taken from a very long record of emotional manifestations of different people. Third, as predicted, on live data the accuracy will be much lower than on an almost identical, well-processed data set, which trains the neural networks of multimodal emotion recognition systems.

Therefore, for a correct analysis, it is necessary to conduct a much larger study, after which the correct conclusion can be determined.

The topic covered in this paper is very extensive. It takes a lot of time to cover at least part of this area. This topic is attractive, with its diversity and the results of the work of systems that use the technology of multimodal recognition of emotions based on machine learning.

According to [45], the highest recognition accuracy of 99% for FER is provided by the SVM classifier and it recognizes seven emotions. As obtained in the paper, FER's 92% accuracy is lower. However, the proposed model is tested for "live" video too with 79% accuracy, and that is impossible for SVM. In 2D FER, mostly JAFFE and CK databases are used for more efficient performance than the other databases.

When it comes to results presented in Tables 4 and 5, multimodal neural networks, particularly MMERs, present the best result for all numbers of emotions. It can be explained by training of all networks on the same dataset and increasing the accuracy based on merging streams.

Author Contributions: Conceptualization, N.S. and S.L.; methodology, O.Z.; software, O.Z.; validation, N.S., O.Z. and S.L.; formal analysis, O.Z.; investigation, N.S.; resources, S.L.; data curation, O.Z.; writing—original draft preparation, O.Z.; writing—review and editing, N.S.; visualization, O.Z.; supervision, N.S.; project administration, S.L.; funding acquisition, S.L. All authors have read and agreed to the published version of the manuscript.

Funding: The National Research Foundation of Ukraine funded this research under project number 2021.01/0103 and British academy fellowship number RaR\100727.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting this study's findings are openly available in https://doi.org/10.6084/m9.figshare.23596362.v1 (accessed on 1 September 2022). Datasets are used for different models' performance evaluation, namely: FER: fer2013: https://www.kaggle.com/deadskull7/fer2013 (accessed on 1 September 2022) and CK + 48 five emotions: https://www.kaggle.com/gauravsharma99/ck48-5-emotions (accessed on 1 September 2022); SER: RAVDESS Emotional speech audio: https://www.kaggle.com/uwrfkaggler/ravdess-emotional-speech-audio (accessed on 1 September 2022); TER: Text-Emotion-detection: https://www.kaggle.com/dataset/f10c38f8f356a43b344ca82476b6b32b5d31b99af19276ba1f7846004c0851f2 (accessed on 1 September 2022); Datasets from the Internet inside the project: https://drive.google.com/drive/folders/1ZV3ceCjNND7xcUxbsJb57aitTpUbcYa9?usp=sharing (accessed on 1 September 2022); Videos for tests from YouTube: (1) Biden Delivers Remarks On Inflation_NBC News—https://www.youtube.com/watch?v=ckCOF719atE (accessed on 1 September 2022); (2) Boris Johnson_Ukraine will win war and 'be free'—https://www.youtube.com/watch?v=WPM8Pvgkz7Y (accessed on 1 September 2022); (3) Father's final words to his dying son!—https://www.youtube.com/watch?v=C3hABRHmQoo (accessed on 1 September 2022); (4) Minecraft Warden Update is a NIGHTMARE!—https://www.

youtube.com/watch?v=2osdz9Z5JKY (accessed on 1 September 2022). Video for Live Test: https://drive.google.com/drive/folders/1wAR2CdlGIEtOSjKv7T9e-gQhBHIAiLLM?usp=sharing (accessed on 1 September 2022).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Jena, P.K.; Ghosh, S.; Koley, E.; Mohanta, D.K.; Kamwa, I. Design of AC state estimation based cyber-physical attack for disrupting electricity market operation under limited sensor information. *Electr. Power Syst. Res.* **2022**, *205*, 107732. [CrossRef]
2. Qin, B.; Liu, D.; Chen, G. Formal modeling and analysis of cyber-physical cross-space attacks in power grid. *Int. J. Electr. Power Energy Syst.* **2022**, *141*, 107790. [CrossRef]
3. Cao, G.; Gu, W.; Lou, G.; Sheng, W.; Liu, K. Distributed synchronous detection for false data injection attack in cyber-physical microgrids. *Int. J. Electr. Power Energy Syst.* **2022**, *137*, 107788. [CrossRef]
4. Tahoun, A.H.; Arafa, M. Secure control design for nonlinear cyber–physical systems under DoS, replay, and deception cyber-attacks with multiple transmission channels. *ISA Trans.* **2021**, *128*, 294–308. [CrossRef]
5. Stellios, I.; Kotzanikolaou, P.; Grigoriadis, C. Assessing IoT enabled cyber-physical attack paths against critical systems. *Comput. Secur.* **2021**, *107*, 102316. [CrossRef]
6. Jena, P.K.; Ghosh, S.; Koley, E. Design of a coordinated cyber-physical attack in IoT based smart grid under limited intruder accessibility. *Int. J. Crit. Infrastruct. Prot.* **2021**, *35*, 100484. [CrossRef]
7. Li, L.; Wang, W.; Ma, Q.; Pan, K.; Liu, X.; Lin, L.; Li, J. Cyber attack estimation and detection for cyber-physical power systems. *Appl. Math. Comput.* **2021**, *400*, 126056. [CrossRef]
8. Ding, D.; Han, Q.-L.; Xiang, Y.; Ge, X.; Zhang, X.-M. A survey on security control and attack detection for industrial cyber-physical systems. *Neurocomputing* **2018**, *275*, 1674–1683. [CrossRef]
9. Lima, P.M.; Carvalho, L.K.; Moreira, M.V. Detectable and Undetectable Network Attack Security of Cyber-physical Systems. *IFAC-Pap.* **2018**, *51*, 179–185. [CrossRef]
10. Barrère, M.; Hankin, C.; Nicolaou, N.; Eliades, D.G.; Parisini, T. Measuring cyber-physical security in industrial control systems via minimum-effort attack strategies. *J. Inf. Secur. Appl.* **2020**, *52*, 102471. [CrossRef]
11. Liu, Y.; Deng, L.; Gao, N.; Sun, X. A Reliability Assessment Method of Cyber Physical Distribution System. *Energy Procedia* **2019**, *158*, 2915–2921. [CrossRef]
12. Qi, Q.; Lin, L.; Zhang, R. Feature Extraction Network with Attention Mechanism for Data Enhancement and Recombination Fusion for Multimodal Sentiment Analysis. *Information* **2023**, *12*, 342. [CrossRef]
13. Naseem, M.T.; Seo, H.; Kim, N.H.; Lee, C.S. Pathological Gait Classification Using Early and Late Fusion of Foot Pressure and Skeleton Data. *Appl. Sci.* **2024**, *14*, 558. [CrossRef]
14. Hu, A.; Seth, F. Multimodal sentiment analysis to explore the structure of emotions. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 350–358.
15. Williams, J.; Kleinegesse, S.; Comanescu, R.; Radu, O. Recognizing Emotions in Video Using Multimodal DNN Feature Fusion. In Proceedings of the Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML), Melbourne, VIC, Australia, 18 July 2018; pp. 11–19. [CrossRef]
16. Majumder, N.; Hazarika, D.; Gelbukh, A.; Cambria, E.; Poria, S. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowl. Based Syst.* **2018**, *161*, 124–133. [CrossRef]
17. Poria, S.; Cambria, E.; Hazarika, D.; Majumder, N.; Zadeh, A.; Morency, L.-P. Context-Dependent Sentiment Analysis in User-Generated Videos. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 873–883. [CrossRef]
18. Zadeh, A.; Liang, P.P.; Mazumder, N.; Poria, S.; Cambria, E.; Morency, L.-P. Memory Fusion Network for Multi-view Sequential Learning. *Proc. AAAI Conf. Artif. Intell.* **2018**, *32*, 1. [CrossRef]
19. Liang, P.P.; Liu, Z.; Zadeh, A.; Morency, L.-P. Multimodal Language Analysis with Recurrent Multistage Fusion. *arXiv* **2018**, arXiv:1808.03920.
20. Wang, Y.; Shen, Y.; Liu, Z.; Liang, P.P.; Zadeh, A.; Morency, L.-P. Words Can Shift: Dynamically Adjusting Word Representations Using Nonverbal Behaviors. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 7216–7223. [CrossRef]
21. Delbrouck, J.-B.; Tits, N.; Brousmiche, M.; Dupont, S. A Transformer-based joint-encoding for Emotion Recognition and Sentiment Analysis. *arXiv* **2020**, arXiv:2006.15955.
22. Emotions and Types of Emotional Responses, Verywell Mind. Available online: https://www.verywellmind.com/what-are-emotions-2795178 (accessed on 6 June 2023).
23. Torkamaan, H.; Ziegler, J. Exploring chatbot user interfaces for mood measurement: A study of validity and user experience. In Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers, Virtual Event, 12–17 September 2020; pp. 135–138.
24. Torkamaan, H.; Ziegler, J. Mobile mood tracking: An investigation of concise and adaptive measurement instruments. Proceedings of the ACM on Interactive, Mobile. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2020**, *4*, 155. [CrossRef]

25. Dahmane, M.; Alam, J.; St-Charles, P.-L.; Lalonde, M.; Heffner, K.; Foucher, S. A Multimodal Non-Intrusive Stress Monitoring from the Pleasure-Arousal Emotional Dimensions. *IEEE Trans. Affect. Comput.* **2020**, *13*, 1044–1056. [CrossRef]
26. Universal Emotions, Paul Ekman Group. Available online: https://www.paulekman.com/universal-emotions/ (accessed on 6 June 2023).
27. Niwlikar, A.B. Popular Theory of the 6 Basic Emotions by Paul Ekman, Careershodh, 25 February 2022. Available online: https://www.careershodh.com/popular-theory-of-the-6-basic-emotions-by-paul-ekman/ (accessed on 6 June 2023).
28. Lim, N. Cultural differences in emotion: Differences in emotional arousal level between the East and the West. *Integr. Med. Res.* **2016**, *5*, 105–109. [CrossRef] [PubMed]
29. Emotion-Detecting Tech Should Be Restricted by Law—AI Now, BBC News. Available online: https://www.bbc.com/news/technology-50761116 (accessed on 5 June 2023).
30. Nast, C. Job Screening Service Halts Facial Analysis of Applicants, Wired, 12 January 2021. Available online: https://www.wired.com/story/job-screening-service-halts-facial-analysis-applicants/ (accessed on 6 June 2023).
31. Cogito—AI for a Better Human Customer Service Experience, Digital Innovation and Transformation. Available online: https://digital.hbs.edu/platform-digit/submission/cogito-ai-for-a-better-human-customer-service-experience/ (accessed on 6 June 2023).
32. Vincent, J. AI 'Emotion Recognition' Can't Be Trusted, The Verge, 25 July 2019. Available online: https://www.theverge.com/2019/7/25/8929793/emotion-recognition-analysis-ai-machine-learning-facial-expression-review (accessed on 6 June 2022).
33. Tatan, V. Understanding CNN (Convolutional Neural Network), Medium, 23 December 2019. Available online: https://towardsdatascience.com/understanding-cnn-convolutional-neural-network-69fd626ee7d4 (accessed on 6 June 2023).
34. Dobilas, S. LSTM Recurrent Neural Networks—How to Teach a Network to Remember the Past, Medium, 5 March 2022. Available online: https://towardsdatascience.com/lstm-recurrent-neural-networks-how-to-teach-a-network-to-remember-the-past-55e54c2ff22e (accessed on 6 June 2023).
35. What is Supervised Learning? | IBM. Available online: https://www.ibm.com/cloud/learn/supervised-learning (accessed on 6 June 2023).
36. Text-Emotion-Detection. Available online: https://www.kaggle.com/dataset/f10c38f8f356a43b344ca82476b6b32b5d31b99af19276ba1f7846004c0851f2 (accessed on 5 June 2023).
37. RAVDESS Emotional Speech Audio. Available online: https://www.kaggle.com/uwrfkaggler/ravdess-emotional-speech-audio (accessed on 5 June 2023).
38. CK+48 5 Emotions. Available online: https://www.kaggle.com/gauravsharma99/ck48-5-emotions (accessed on 6 June 2023).
39. Kumar, S.; Bhuyan, M.K.; Chakraborty, B.K. Extraction of informative regions of a face for facial expression recognition. *IET Comput. Vis.* **2016**, *10*, 567–576. [CrossRef]
40. Facial Expression Dataset Image Folders (fer2013). Available online: https://www.kaggle.com/datasets/astraszab/facial-expression-dataset-image-folders-fer2013 (accessed on 25 October 2023).
41. Wang, X.; Wang, K.; Lian, S. A survey on face data augmentation for the training of deep neural networks. *Neural Comput. Appl.* **2020**, *32*, 15503–15531. [CrossRef]
42. Zhang, H.; Gou, R.; Shang, J.; Shen, F.; Wu, Y.; Dai, G. Pre-trained deep convolution neural network model with attention for speech emotion recognition. *Front. Physiol.* **2021**, *12*, 643202. [CrossRef]
43. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
44. CMU-MOSEI Dataset. Available online: http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/ (accessed on 24 October 2023).
45. Revina, I.; Emmanuel, W. A survey on human face expression recognition techniques. *J. King Saud Univ. Comput. Inf. Sci.* **2021**, *33*, 619–628. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.