International Journal of Artificial Intelligence,
ISSN 0974-0635; Int. J. Artif. Intell.
2026, Volume 24, Issue Number 2 [Published on 29/08/2026]

Architecture of a Real-Time Facial Emotion Recognition System for Intelligent Systems

Ayaguz Beksultanova[1], **Saule Kudubayeva**[2], **Kalybek Maulenov**[3], **Tatyana Radchenko**[4] and **Olga Ramp**[5]

[1]Akhmet Baitursynuly Kostanay Regional University,
Baitursynov Str., 47, 110000, Kostanay, Republic of Kazakhstan
a.beksult@hotmail.com

[2]Department of Artificial Intelligence Technologies, L.N. Gumilyov Eurasian National University
Satpayev Str., 2, 010008, Astana, Republic of Kazakhstan
s.kudub@outlook.com

[3]Department of Digital Technologies and Artificial Intelligence, Akhmet Baitursynuly Kostanay Regional University,
Baitursynov Str., 47, 110000, Kostanay, Republic of Kazakhstan
k.maulenov@hotmail.com

[4]Department of Physics, Mathematics and Digital Technologies, Akhmet Baitursynuly Kostanay Regional University,
Baitursynov Str., 47, 110000, Kostanay, Republic of Kazakhstan
t.radchenk@outlook.com

[5]Department of Physics, Mathematics and Digital Technologies, Akhmet Baitursynuly Kostanay Regional University,
Baitursynov Str., 47, 110000, Kostanay, Republic of Kazakhstan
o.ramp@hotmail.com

ABSTRACT

The study aimed to design and evaluate the architecture of a real-time facial emotion recognition system for intelligent applications integrated into the educational infrastructure of Akhmet Baitursynuly Kostanay Regional University (Republic of Kazakhstan). The paper analyzes modern machine learning approaches for computer vision, including convolutional neural networks and transformer architectures, to identify and compare baseline models for automated visual information analysis. Experimental evaluation was conducted using the publicly available Facial Expression Recognition 2013 (FER-2013) and Real-world Affective Faces Database (RAF-DB) datasets, which are widely used for benchmarking emotion recognition methods. The proposed architecture achieved higher emotion recognition accuracy than the baseline models while maintaining comparable computational costs and improving video processing efficiency. It also required fewer computational resources, particularly in terms of memory consumption and graphics processing unit load. The developed system demonstrated stable performance during real-time video stream processing and consistent recognition accuracy across different test datasets. In addition, the model maintained reliable emotion classification under variations in input images throughout the experiments. The practical significance of the research lies in the applicability of the proposed system to real-world video analysis tasks, including educational, research, and intelligent applications requiring automatic real-time recognition of users' emotional states under limited computational resources. The findings can support the development of intelligent systems for behavioral analysis and the investigation of human interaction with digital environments, providing an efficient and resource-conscious solution for real-time emotion recognition.

Keywords: Deep learning, computer vision, neural networks, convolutional models, algorithm optimisation.

Mathematics Subject Classification: 68T45, 68T07, 68T10

Journal of Economic Literature (JEL) Classification: C45, C63, O33

1. INTRODUCTION

The development of artificial intelligence and computer vision technologies during the period from 2015 to 2025 has significantly expanded the capabilities of human-machine interaction systems. At the same time, the tasks of automatic real-time emotion recognition remain highly relevant from the perspective of ensuring the stable performance of models and their implementation in applied intelligent systems. For educational institutions, including Akhmet Baitursynuly Kostanay Regional University, the use of systems for analysing students' emotional states during the learning process is of interest for supporting digital educational platforms. However, the deployment of such systems is accompanied by limitations in computational resources and variability in acquisition conditions, including lighting, viewing angle, video stream quality, and hardware configuration, which are characteristic of the university infrastructure. This necessitates the development of models designed for real-time video data processing using a standard computing environment while maintaining acceptable classification accuracy. An additional challenge arises from the discrepancy between real-world operating conditions and the laboratory scenarios on which most existing models are trained. Such inconsistency decreases the generalization capability of algorithms and demands their adaptation to practical operating conditions. Therefore, the studies aimed at increasing the classification accuracy and, at the same time, optimisation of the models' computational efficiency, including reduction of the latency of processing of the frame, the decrease of the memory consumption, increase of the speed of processing of the video stream, have become especially relevant.

Shaikemelova and Omarov (2025), and Uali et al. (2024) showed the prospects of deep learning methods for analysing human emotional states and solving tasks of automatic emotion classification. The study of Sapakova and Sarsembayev (2023) examined modern approaches to the analysis of emotional characteristics using deep learning technologies, which confirmed the significant potential of neural networks in processing complex behavioural data. In the study of Mutalova et al. (2024) methods of analysis of facial images and increasing the efficiency of computer vision systems used for visual object recognition are described.

Recent research has particularly focused on the robustness of algorithms to changes in the characteristics of input data and in real-world operating conditions. The analysis of the results obtained by Bekmurzayev et al. (2023) showed the importance of image pre-processing and increasing the robustness of models to changes in lighting, image quality and environmental conditions. The works of Gupta et al. (2023) and Wu et al. (2025) observed that the performance of emotion recognition systems can degrade significantly when deployed from controlled laboratory conditions to real-world application scenarios due to variations in lighting, camera angles, image quality, and other external factors. The authors stressed the importance of improving the generalisation ability of models and adapting them to practical operating conditions.

Moreover, recent studies have addressed the problem of balancing the recognition performance and the computational efficiency of models. Kudubayeva et al. (2026) have shown that one of the main challenges in the development of intelligent systems for visual information analysis is the achievement of high recognition accuracy with acceptable computational costs. Thus, the literature shows the absence of universal architectures that can provide both high recognition accuracy and computational efficiency in real-time operation. This determines the relevance of developing optimised models for application under the computational resources' constraints of educational institutions. In addition, the importance of the trade-off between the classification performance and the hardware resource requirements is highlighted, which is particularly relevant to the implementation of such systems in large-scale educational settings.

The purpose of this research was to evaluate the performance of architectural solutions used in facial emotion recognition systems in the aspect of classification accuracy and computational efficiency when processing real-time video stream in the limitations of computational resource of educational infrastructure. To achieve this goal, the following objectives were set: to compare the performance of different types of neural networks, to analyse quality metrics using publicly available datasets, to evaluate the computational requirements of the developed models, to compare the architectures of Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) recurrent neural networks and Vision Transformers, to assess their suitability for real-time video stream processing, to evaluate the model performance using the FER-2013 and RAF-DB datasets, and to summarize the findings from the perspective of their practical applicability under conditions of limited computational infrastructure.

The scientific novelty of the study is a comparative assessment of convolutional neural network architectures, recurrent neural networks and Vision Transformers (ViT) in terms of the compromise between the accuracy of emotion recognition and computational efficiency during processing of real-time video stream. The main focus, contrary to most of the studies in the field, is to analyse the applicability of the models under the computational resource constraints of educational infrastructure in terms of classification accuracy, computational load and practical feasibility.

2. MATERIALS AND METHODS

The empirical research of the facial emotion recognition system was carried out in Akhmet Baitursynuly Kostanay Regional University. Source data. The FER-2013 and RAF-DB datasets, which are publicly available, were used. The FER-2013 dataset is composed of 35,887 facial images divided into seven emotional classes, while RAF-DB was employed to increase the sample variability and representativeness of the experimental data. The pre-processing procedure included resizing the images to a common size and normalising the pixel values.

Within the study, a comparative evaluation of the architectures of convolutional neural networks (CNNs), three-dimensional convolutional neural networks (3D-CNNs), Long Short-Term Memory (LSTM) recurrent neural networks, and Vision Transformers (ViT) was conducted. Model training and testing were performed under identical experimental conditions using the university's computing infrastructure and graphics processing units (GPUs).

Classification performance was evaluated using the Accuracy and F1-score metrics. Overall classification accuracy was calculated using Equation (1):

$$\mathrm{Accuracy} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}) \tag{1}$$

where TP – true positives; TN – true negatives; FP – false positives; and FN – false negatives.

Model performance was evaluated using the test subsets of the FER-2013 and RAF-DB datasets in accordance with the standard protocols for their application in facial emotion recognition tasks.

To evaluate computational efficiency, GPU utilisation, random access memory (RAM) consumption, and energy consumption were assessed. Measurements were performed during real-time processing of a Full HD (1920×1080) video stream. Data were collected using GPU monitoring tools (NVIDIA Management Library, NVML) and the psutil library. The obtained values are presented as average and peak measurements over the processing period of the test video stream.

The experimental validation of the system was conducted in university classrooms under real educational conditions. To evaluate model robustness, testing was performed at different head rotation angles (0°, ±15°, ±30°, and ±45°) under natural lighting conditions. The analysis included an evaluation of emotion recognition accuracy and video stream processing latency. The system was implemented in Python using machine learning and computer vision libraries. To improve computational efficiency, standard neural network architecture optimisation techniques aimed at reducing the computational complexity of the models were applied. The computational experiments were implemented locally, without data transfer to external servers, which confirms the autonomous execution of the computational experiments.

3. RESULTS

During the empirical study, a comparative evaluation of facial emotion recognition models was conducted using the publicly available FER-2013 and RAF-DB datasets. The obtained results showed that the baseline convolutional neural network (CNN) achieved an Accuracy of 66.7% on the FER-2013 dataset and 79.4% on the RAF-DB dataset, with an F1-score of 0.68. The use of the 3D-CNN architecture increased Accuracy to 68.9% and 80.8%, respectively, with an F1-score of 0.71. The Long Short-Term Memory (LSTM) recurrent neural network achieved an Accuracy of 65.4% on FER-2013 and 77.8% on RAF-DB, with an F1-score of 0.65. The highest performance among the baseline models was achieved by the Vision Transformer (ViT), which attained an Accuracy of 72.4% and 85.9%, respectively, with an F1-score of 0.74. In turn, the proposed architecture demonstrated the best overall performance, achieving an Accuracy of 74.1% on FER-2013 and 89.2% on RAF-DB, with an F1-score of 0.81.

The overall performance indicators of the evaluated models are presented in Table 1.

Table 1: Comparison of emotion recognition accuracy metrics.

Architecture	**Accuracy (FER-2013), %**	**Accuracy (RAF-DB), %**	**F1-score**
Baseline Convolutional Neural Network (CNN)	66.7	79	0.68
3D Convolutional Neural Network (3D-CNN)	68.9	80.8	0.71
Long Short-Term Memory (LSTM) Recurrent Neural Network	65.4	77.8	0.65
Vision Transformer (ViT)	72.4	85.9	0.74
Proposed Architecture	74.1	89.2	0.81

Note: the results were obtained during experimental testing using the constructed dataset.
Source: compiled by the authors.

An analysis of the results presented in Table 1 shows that the proposed architecture demonstrates higher Accuracy and F1-score values than the models considered. On the RAF-DB dataset, the Accuracy reached 89.2%, exceeding the performance of the baseline CNN model. The results obtained (Table 2) confirm the effectiveness of the proposed approach for the task of facial emotion recognition.

Table 2: Technical performance characteristics of the system.

System parameter	**Baseline model**	**Proposed model**	**Percentage change**
Frame Rate (FPS)	32	48	+50%
Latency (ms)	31.5	18.2	-42.2%
VRAM Usage (GB)	4.2	2,.1	-50%

Source: compiled by the authors.

The analysis of the technical characteristics shown in Table 2 reveals that the proposed model enhances the system performance over the baseline architecture. The processing rate was increased to 48 FPS. The inference latency has been reduced to 18.2 ms. The video memory usage decreased to 2.1 GB. The results demonstrate the viability of the proposed model for real-time data processing tasks.

A confusion matrix was used to evaluate the classification performance for each class. The best classification accuracy was obtained for the “happy” and “neutral” classes. At the same time, the largest number of classification errors was observed in the recognition of the "fear" and "surprise" emotions. These results reflect differences in the recognition accuracy of individual emotional states.

To provide a clear analysis of the data representation structure, the image was visualised as a set of structural components. This approach makes it possible to represent the distribution of information within the original image in an interpretable form. An example of the visual representation of these components is presented in Figure 1.

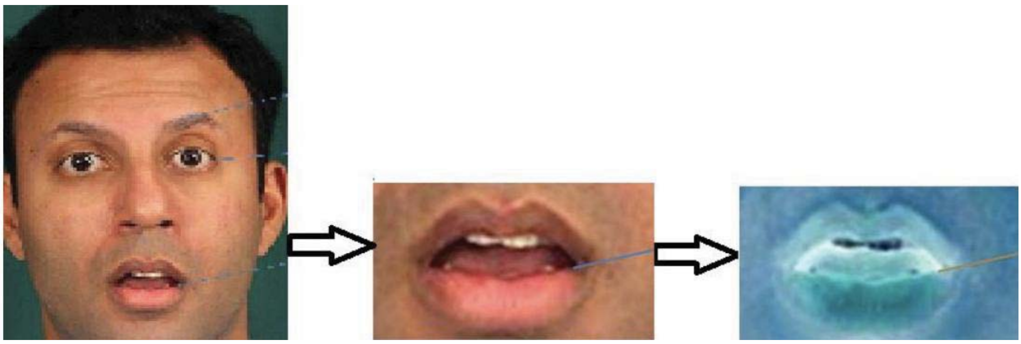

Figure 1. Representation of an image as a set of three-dimensional structures.
Source: compiled by the authors based on Kaur et al. (2026).

As shown in the figure, the original image is represented as a set of structural components that differ in the extent of their contribution to the final data representation. This visualisation makes it possible to illustrate the characteristics of the internal distribution of information within the image and to present the processing results in an interpretable form.

A comparative analysis of computational resource requirements is presented in Table 3. The proposed model demonstrates reduced computational resource utilisation compared with the architectures considered. The model required less random-access memory (RAM) and computational resources compared to baseline models, enabling more efficient data processing without sacrificing the performance of emotion recognition. The results obtained confirm the possibility of the application of the proposed architecture to real-time tasks and systems with limited computational resources.

Table 3: Analysis of system resource consumption during model inference.

architectural solution	gpu utilisation (%)	RAM usage (MB)	Peak power consumption (W)
CNN (ResNet-50)	78	1,450	120
3D-CNN	92	2,100	165
ViT	85	3,200	140
Proposed Architecture	45	850	85

Note: measurements were performed during the processing of a Full HD (1920×1080) video stream.
Source: compiled by the authors.

As demonstrated in Table 3 the proposed architecture is less computationally demanding than the models under consideration. In particular, use of the GPU, the consumption of random-access memory (RAM) and the peak energy consumption are reduced. This suggests a more efficient use of computing resources at inference time than the baseline architectures.

The experimental validation, carried out in university classrooms, evaluated the system stability when processing video streams. During the testing process, the average values of the important key performance indicators such as recognition accuracy and frame processing time were recorded. The obtained results show a stable operation of the system in terms of long-term operation and without significant deviations of the main performance indices.

Statistics of face detection effectiveness at different camera angles are shown in Table 4.

Table 4: Effectiveness of face detection according to camera angle.

Head Rotation Angle (Degrees)	**Detection Accuracy (%)**	**Latency (ms)**
0° (frontal)	99.8	3.8
±15°	98.5	4.1
±30°	94.2	4.5
±45°	87.6	5.2

Note: measurements were performed under the actual lighting conditions of university classrooms.
Source: compiled by the authors.

Analysis of the confusion matrix revealed individual systematic classification errors between emotional states with similar visual features. An increase was observed in the recognition accuracy of each individual class after making changes to the model.

The tests showed that the system provides stable face detection and emotion recognition on the test data. The model demonstrates stable performance during video stream processing without a significant decrease in classification quality. Analysis of the obtained results indicates the appropriate functioning of the system in various input data conditions in the performed experiments. Overall, stable values of the emotion recognition quality metrics were observed without pronounced fluctuations between data subsets. It was also shown that the system preserves its operational ability in processing successive video frames, demonstrating stable performance indicators in the test scenario. This indicates that the model can be applied to real-time video data analysis tasks.

In order to provide a general assessment of model behaviour on different subsets of test data, a further analysis of the results was conducted. The analysis showed that the performance of classification could be influenced by the complexity of the input images, but in all the experiments performed, the model maintained its performance under different conditions in general. The system performance was also estimated during the processing of video streams with different numbers of objects in the frame. It was shown that an increase in the number of faces resulted in an increase of processing time and a decrease of frame processing rate and the system was working in the mode of video data processing. Further investigation of the classification results for different emotional states was carried out separately. Non-uniform recognition performance for different classes was revealed, i.e., some emotions were recognised with higher accuracy than others, which is a characteristic feature of the multiclass classification task. The general analysis of the model robustness shows that the system performs stable under changes in characteristics of input data within the tests conducted, without any substantial degradation in quality.

Also, the robustness of the model against partial facial occlusions was studied. Recognition performance was found to degrade with the decrease in availability of informative image regions, in line with the general dependence of pattern recognition tasks on the completeness of the input data. Furthermore, an evaluation of computational resource utilization during system operation under a standard computing environment was also performed. The results demonstrate that the model can be run without specialized hardware and with a manageable computational resource load. The experimental results show that the proposed system achieves a stable trade-off between recognition accuracy and computational efficiency and can therefore be considered as adequate for video data analysis applications under near real-time conditions.

A comparison of the performance of the proposed model with respect to the baseline architectures was also performed. The results indicate that the proposed model performs better in terms of the main performance metrics than the alternative approaches used in the experiment. Both the Accuracy and F1-score steadily improved over the various test datasets. Analysis of the error distribution revealed that the classification performance could be different depending on the complexity of the classes to be recognised. This is in accordance with the general features of the multiclass emotion classification task. Also, it was found that the model maintains stable performance with changes in the parameters of the input video stream within the experimental testing. No significant deterioration in recognition performance was observed. Overall, the comparison results show that the proposed model provides better recognition performance than the base architectures while maintaining stable system operation.

Moreover, it should be noted that the comparison of the architectural approaches allows identifying a general trend of improving the recognition performance with the increasing model complexity. In particular, the transition from conventional convolutional models to architectures with enhanced spatio-temporal processing leads to a continuous increase in classification accuracy (Lavrentieva et al., 2019). Meanwhile, the proposed model has the most balanced recognition performance and computational complexity, making it more suitable for practical application. The analysis of results in Table 1 indicates that the models exhibit consistent differences across the different datasets. The baseline architectures are less accurate than state-of-the-art approaches that employ more sophisticated feature extraction mechanisms (Manghisi et al., 2020). Meanwhile, the proposed model still keeps its advantage on both test datasets, which demonstrates better generalisation ability. The results of the computational efficiency analysis confirm that the proposed architecture provides a significant reduction in the computational load in comparison to the baseline models. This is reflected in reduced memory consumption and lower graphics resource demands, while maintaining a comparable recognition performance. Thus, the model shows a more efficient use of computational resources. The analysis of the confusion matrix shows that the majority of the classification errors are due to the overlap between visually similar emotional classes. This is a characteristic of the emotion recognition task and was found over all the models considered. The proposed architecture shows a decrease in the number of such errors compared to the baseline approaches, which implies a better separation of the feature space. The results obtained in general confirm the efficiency of the proposed approach for the task of recognizing emotional states from facial images. The proposed model achieves better accuracy and robustness compared to the baseline architectures, while also being more efficient in terms of computational resources. This makes it a feasible solution to problems of visual data analysis in applied intelligent systems. The obtained results confirm that the differences between the architectures are reflected not only in the absolute values of the performance metrics, but also in their stability over different types of data. Simpler models are more sensitive to changes in the structure of the input data (Seidaliyeva and Smailov, 2025). More complex architectures have a more stable performance across the test datasets. The evaluation of the generalisation capability of the models should be specially considered. The proposed architecture achieves high performance on different datasets without significant loss of efficacy, indirectly showing stronger feature extraction and less dependence on the characteristics of the training dataset. Moreover, it is worth noting that the consistency among the Accuracy and F1-score metrics was noticed for all the considered cases, confirming the lack of a bias towards individual classes. This is common in multiclass classification problems where unbalanced predictions may affect the distribution of errors across classes and the overall performance measures of the model. The reduction in computational load achieved by the proposed model suggests a more efficient architectural organisation of the computational process (Bezrukova et al., 2022). This is reflected in lower resource utilization with similar performance, thus making the model more appropriate to be integrated into computationally resource-limited environments. The experimental study results reveal that the proposed architecture offers a more balanced trade-off between recognition performance and computational efficiency. This makes it possible to consider it as a balanced solution for automatic analysis of emotional states in video streams. Also, the nature of the classification errors is largely dictated by the visual similarity between individual emotion classes. This is typical for emotion recognition tasks and it sets an upper limit on the performance that can be achieved for all the considered models.

Furthermore, it is worth mentioning that the obtained results are in good agreement between the different parts of the model performance evaluation. In particular, a correspondence between classification accuracy and F1-score values was found, which indicates balanced predictions and the lack of a pronounced bias towards individual classes. This consistency is an important indicator of correct model behaviour in multiclass classification tasks, where the distribution of errors across classes may significantly affect the overall interpretation of the results. The comparative analysis of architectures shows that the performance improvement in recognition is due to more effective feature extraction and better model generalisation capability. Also worth noting are the differences between the models not only in the final metric values but also in the pattern of the distribution of errors which reflects the differences in their underlying operating principles. Simple architectures are more sensitive to variability in the input data, whereas more complex models provide more stable performance across different subsets. Special attention is deserved by the interpretation of the results regarding practical applicability. The proposed model allows to obtain a better trade-off between recognition performance and computational cost and makes it suitable for real-time systems.

The proposed approach is more versatile and potentially applicable to a variety of practical tasks involving video data analysis due to the reduction in resource requirements while keeping a high level of accuracy. It should also be underlined that the results obtained in the experiments show the robustness of the model to changes in the input data within the test datasets used. This also applies to maintaining stable values of the key performance indicators when crossing different data sets and test scenarios. This robustness indicates good generalisation ability and the ability of the model to maintain its effectiveness when dealing with data that differ in structure and visual features. From an error analysis point of view, it can be observed that most of the errors are caused by the inherent complexity of the emotion recognition task, due to a high visual similarity among individual classes. This limits the maximum achievable classification accuracy and is a common property of all the models considered. However, the proposed architecture shows a more effective feature space separation, which is manifested in the decrease in the number of typical classification errors compared to the baseline approaches. Finally, the results support the conclusion that the proposed system provides a good trade-off between accuracy, robustness and computational efficiency, making it suitable for automatic emotion analysis in video stream processing conditions.

4. DISCUSSION

The obtained results showed that the proposed architecture had a more balanced relation between the performance in emotion classification and the computational efficiency than the baseline models. This suggests the possible use of the model for facial image analysis tasks with limited computational resources. The results were compared with the findings of Wang et al. (2026) and Paul et al. (2025) and the results are consistent with the patterns established before in computer vision. Specifically, it has been shown that Vision Transformer models have a higher classification accuracy than the conventional convolutional neural networks, as they model the global dependencies within the images more efficiently. Simultaneously, these studies also reported an increase in the computational complexity of transformer architectures, which is a characteristic limitation of this class of models. Similar findings were reported by Talaat (2023), who also noted the increased computational resource requirements of transformer architectures. The results of the current study showed that the adapted architecture can improve the classification performance while keeping the computational complexity at an acceptable level. Therefore, modified transformer-based approaches are a promising solution for facial emotion recognition tasks (Milloshi and Lako, 2024; Jaiprakash et al., 2020). Thus, the obtained results are consistent with the existing literature and extend it with regard to the analysis of the trade-off between model performance and computational efficiency in multiclass emotion classification tasks.

The study by Akula et al. (2026) noted that variability in visual conditions, including changes in brightness and the presence of shadows, may adversely affect the performance of computer vision systems. This is associated with the sensitivity of models to changes in the distribution of pixel intensities, which leads to reduced stability of feature extraction (Mazhibayeva et al., 2026). The findings of the present study are generally consistent with these observations, as they revealed differences in classification performance across emotional categories and variable visual examples in the FER-2013 and RAF-DB datasets. Aly (2025) emphasised the importance of identifying locally informative facial regions, including the eyes, eyebrows, and mouth, to improve emotion classification performance. The results obtained are consistent with these conclusions regarding the dependence of model performance on the effectiveness of extracting meaningful facial features. In particular, architectures with a greater capacity for identifying informative features achieved higher Accuracy and F1-score values (Smailov et al., 2025; Tavolzhanskyi et al., 2025). Thus, the findings of the study confirm the importance of image pre-processing and key feature extraction mechanisms for improving emotion recognition performance, which is consistent with previously published studies and extends them in the context of a comparative analysis of different neural network architectures.

The studies by Dwijayanti et al. (2022) noted that the effectiveness of emotion recognition systems largely depends on the computational resources required for model training and inference. The employment of high-end graphics processing units is considered as one of the main requirements for the functioning of modern deep learning algorithms and, therefore, restricts their implementation in case of limited computational infrastructure (Bilan et al., 2017; Ginters and Revathy, 2021; Ginters et al., 2018). The findings of the present work demonstrate that the model architecture differences have a large influence on computational complexity with similar classification performance. A comparison of the considered architectures demonstrates that it is possible to reduce computational costs by architectural optimisation without significant degradation of Accuracy and F1-score. We observed in particular reductions in GPU utilization, memory consumption and inference latency while maintaining high accuracy. These results are consistent with previous studies and confirm the importance of the optimization of neural network architectures for the efficient computer vision systems.

Abdelkader et al. (2026) have shown that some basic emotional states are characterized by partially overlapping facial features, which makes their automatic discrimination more difficult. Such overlap is most prominent in key facial regions like the eyes, eyebrows and mouth, which are used by most computer vision models for feature extraction. Gong et al. (2025) arrived at similar conclusions, demonstrating that classification accuracy can decrease when emotional expressions are weakly pronounced and visually similar patterns exist across different classes. The results of the present study corroborate these observations and contribute to the understanding of how visual feature similarity affects classification errors in facial emotion recognition tasks. It was found that the biggest difficulties are in the recognition of emotions of similar facial features, because their feature spaces partially overlap (Efremov, 2025; 2026; Jaiprakash et al., 2019). Hence, the results validate the limited distinguishability of specific emotional categories as an inherent property of multiclass classification tasks for facial images, which needs to be taken into account when analysing the classification errors of models.

Machine learning models trained on imbalanced datasets can be biased towards dominant classes, leading to poor classification performance for underrepresented categories (Duggento et al., 2019; Raj et al., 2025), as observed in the study of Fakhar et al. (2022). The results of the present study are in agreement with these observations on the effect of class distribution on the overall classification performance. In particular, the F1-score indicates the balance between precision and recall and allows to evaluate the model performance in conditions of heterogeneous data distribution (Salmanov, 2025; Salmanov and Mahmudov, 2012). The comparison with the previously published studies showed that the class imbalance is still one of the factors that affect the performance of emotion recognition systems, especially in the multiclass facial image classification tasks. The interpretation of the results was therefore carried out taking into account the distribution of data among the emotional categories.

The work of Subramanian et al. (2022) on real-time emotion recognition systems as a way to adapt digital services to users' emotional states supports the consistency of the findings of the present study with previously published evidence. The authors showed that the practical value of such solutions depends not only on the accuracy of classification but also on the ability of these solutions to work reliably under conditions of limited computational resources. Pabba and Kumar (2022) reported similar findings in their study on the use of facial expression recognition technologies for monitoring student engagement. Stable model performance in real-world operating conditions was the focus of that study, which is consistent with the findings in the present study.

The obtained results confirmed a general pattern characteristic of computer vision tasks: the quality of emotion classification is largely determined by the ability of the model to extract robust and informative facial features. The differences between the examined architectures were reflected in how effectively they could represent spatial dependencies and the level of generalisation achieved on the test data. The more sophisticated architectures such as Vision Transformers and the proposed model demonstrated superior performance over the baseline convolutional and recurrent approaches. This means that the application of global feature analysis mechanisms leads to the more accurate interpretation of visual information for multiclass emotion classification tasks (Ilyina, A., 2024; Zaki, et al., 2021; Zheng et al., 2024). At the same time, constraints related to the similarity of visual characteristics of some emotional categories were revealed, which caused partial overlap of feature spaces and increased the probability of classification errors. This characteristic is an intrinsic property of facial emotion recognition tasks and should be considered when interpreting the model performance (Ginters, 2020; Mohanraj et al., 2019).

In addition, it was shown that model performance was affected not only by architectural properties but also by the quality of input data, considering variations in lighting, viewing angles and the intensity of emotional expressions. It was observed that deterioration in visual conditions reduced the distinguishability of individual classes, particularly in cases involving only subtly expressed facial movements (Babak and Kulyk, 2023; Ilyina, 2025; 2026; Tan et al., 2025). It was also found that optimisation of the computational structure of the model reduced processing latency and increased the video stream processing rate without any substantial loss of classification accuracy, which is particularly important for real-time applications.

Thus, the study confirmed the importance of neural network architecture selection as one of the key factors influencing the quality of emotion classification and demonstrated the need for further improvement of methods for distinguishing closely related emotion classes, as well as enhancing model robustness to variability in the input data.

5. CONCLUSIONS

As a result of the study, an architecture for a facial emotion recognition system was developed and experimentally validated on the basis of a comparative analysis of deep learning models. Experimental testing using the FER-2013 and RAF-DB datasets demonstrated the superiority of the proposed solution over the baseline architectures. The comparison of the performance of CNN, 3D-CNN, LSTM and Vision Transformer models revealed differences in their ability to extract features, generalise data and computational efficiency. The results show that more sophisticated architectures, such as Vision Transformers and the proposed model, obtain higher Accuracy and F1-score values, while maintaining similar computational costs. These results confirm the remarkable impact of architectural features on the trade-off between classification accuracy and computational complexity.

The proposed model leverages the advantages of convolutional architectures designed for efficient spatial feature extraction and transformer-based approaches that allow modelling of global dependencies in images. This resulted in building a more informative feature representation than that of recurrent models. The experiments revealed that the LSTM architecture achieved lower classification performance than the other studied methods, which limits its usefulness for facial emotion recognition tasks on static facial images. The proposed architecture achieves better performance than the other models with Accuracy of 89.2% and F1-score of 0.81 on FER-2013 and RAF-DB datasets. The obtained results confirm the effectiveness of the proposed approach for the task of multiclass facial emotion classification in the context of limited computational resources. The study is limited by lower classification accuracy for visually similar emotional states due to partial overlap in feature spaces of individual classes. Future work should focus more on the further development of deep learning architectures and model training methods to improve the distinguishability of emotional categories and further enhance recognition performance.

6. REFERENCES

Abdelkader, B., Rakia, J., Gilles, B., 2026, CNN-DET: a hybrid deep learning architecture for emotion recognition. *Expert Syst. Appl.* **312**, 131377.

Akula, A., Budha, G., Bingi, G., Chanda, U., Borra, A.R., Yadav, D.B., Saravanan, M., 2026, Emotion recognition from facial expressions using CNNs. *Int. J. Eng. Ext. Technol. Res.* **8(1)**, 120-125.

Aly, M., 2025, Revolutionizing online education: Advanced facial expression recognition for real-time student progress tracking via deep learning model. *Multimed. Tools Appl.* **84(13)**, 12575-12614.

Babak, V.P., Kulyk, M.M., 2023, Possibilities and perspectives of the consumers-regulators application in systems of frequency and power automatic regulation. *Tech. Electrodyn.* **2023**(4), 72–80. DOI: 10.15470/techned2023.04.072.

Bekmurzayev, B., Yeremin, D., Kaliyeva, R., Alishin, T., 2023, Application of artificial neural networks in video surveillance systems. *Bull. Kazakh Acad. Transp. Commun.* **127(4)**, 309-321.

Bezrukova, N., Huk, L., Chmil, H., Verbivska, L., Komchatnykh, O., Kozlovskyi, Y., 2022, Digitalization as a Trend of Modern Development of the World Economy. *WSEAS Trans. Environ. Dev.* **18**, 120–129. DOI: 10.37394/232015.2022.18.13.

Bilan, Y., Zos-Kior, M., Nitsenko, V., Sinelnikau, U., Ilin, V., 2017, Social component in sustainable management of land resources. *J. Secur. Sustain. Issues* **7**(2), 107–120. DOI: 10.9770/jssi.2017.7.2(9).

Duggento, A., Scimeca, M., Urbano, N., Bonanno, E., Aiello, M., Cavaliere, C., Cascella, G.L., Cascella, D., Conte, G., Guerrisi, M., Toschi, N., 2019, A random initialization deep neural network for discriminating malignant breast cancer lesions. *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., EMBS*, 912–915. Berlin, Institute of Electrical and Electronics Engineers. DOI: 10.1109/EMBC.2019.8856740.

Dwijayanti, S., Iqbal, M., Suprapto, B.Y., 2022, Real-time implementation of face recognition and emotion recognition in a humanoid robot using a convolutional neural network. *IEEE Access* **10**, 89876-89886.

Efremov, A., 2025, The Fear Primacy Hypothesis in the Structure of Emotional States: A Systematic Literature Review. *Psychol. Rep.* DOI: 10.1177/00332941241313106.

Efremov, A., 2026, Integration of AI and Neuroscience: New Methods in Studying the Causes of Mental Disorders. *Int. J. Psychiatry Med.* DOI: 10.1177/00912174261428872.

Fakhar, S., Baber, J., Bazai, S.U., Marjan, S., Jasinski, M., Jasinska, E., Hussain, S., 2022, Smart classroom monitoring using novel real-time facial expression recognition system. *Appl. Sci.* **12(23)**, 12134.

Ginters, E., 2020, New trends towards digital technology sustainability assessment. *Proc. World Conf. Smart Trends Syst. Secur. Sustain., WS4 2020*, 184–189. Virtual, London, Institute of Electrical and Electronics Engineers. DOI: 10.1109/WorldS450073.2020.9210408.

Ginters, E., Mezitis, M., Aizstrauta, D., 2018, Sustainability simulation and assessment of bicycle network design and maintenance environment. *2018 Int. Conf. Intell. Innov. Comput. Appl., ICONIC 2018*, article 8601225. Plaine Magnien, Institute of Electrical and Electronics Engineers. DOI: 10.1109/ICONIC.2018.8601225.

Ginters, E., Revathy, J.C., 2021, Hidden and latent factors' influence on digital technology sustainability development. *Mathematics* **9**(21), 2801. DOI: 10.3390/math9212801.

Gong, Q., Liu, X., Ma, Y., 2025, Real-time facial expression recognition based on image processing in virtual reality. *Int. J. Comput. Intell. Syst.* **18(1)**, 8.

Gupta, S., Kumar, P., Tekchandani, R.K., 2023, Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. *Multimed. Tools Appl.* **82(8)**, 11365-11394.

Ilyina, A., 2024, Crafting state policies to encourage innovations in human capital development. *Multidisciplinary Reviews* **7**(Special Issue), e2024spe038. DOI: 10.31893/multirev.2024spe038.

Ilyina, A., 2025, State Regulation and Innovation Infrastructure in the National System. *Revista de Cercetare si Interventie Sociala* **89**, 100–127. DOI: 10.33788/rcis.89.6.

Ilyina, A., 2026, Human capital in the circular economy: how to turn waste into innovation. *Heritage and Sustainable Development* **8**(1), 549–560. DOI: 10.37868/hsd.v8i1.1689.

Jaiprakash, H., Singh, A., Biswas, A., Mohanraj, J., Ghosh, S., 2019, E-PBL: An innovation to promote active learning and decrease cognitive overload among medical students. *Indian J. Public Health Res. Dev.* **10**(4), 1469–1473. DOI: 10.5958/0976-5506.2019.00922.7.

Jaiprakash, H., Singh, H.M.P., Jaiprakash, M., Joshi, V., Achanna, S., Jegasothy, R., 2020, Standardization of simulated patients: Assessing their efficiency and evaluating their perception. *Int. J. Adv. Sci. Technol.* **29**(7 Special Issue), 678–685.

Kaur, N., Singh, M., Singh, R., 2026, *Deep learning-based facial emotion recognition system*. in Smart Technologies and Intelligent Computing, Boca Raton: CRC Press.

Kudubayeva, S., Seksenbayev, Y., Yerimbetova, A., Daiyrbayeva, E., Sakenov, B., Telman, D., Turdalyuly, M., 2026, Unified visual synchrony: a framework for face-gesture coherence in multimodal human-AI interaction. *Big Data Cogn. Comput.* **10(3)**, 88.

Lavrentieva, O.O., Rybalko, L.M., Tsys, O.O., Uchitel, A.D., 2019, Theoretical and methodical aspects of the organization of students' independent study activities together with the use of ICT and tools. *CEUR Workshop Proc.* **2433**, 102–125.

Manghisi, V.M., Fiorentino, M., Boccaccio, A., Gattullo, M., Cascella, G.L., Toschi, N., Pietroiusti, A., Uva, A.E., 2020, A body tracking-based low-cost solution for monitoring workers' hygiene best practices during pandemics. *Sensors* **20**(21), 1–17. DOI: 10.3390/s20216149.

Mazhibayeva, G., Kenesbayev, S.M., Idrissov, S.N., Salgozha, I.T., Akhmetova, Z.A., 2026, Use of Innovative Methods in the Creation of Testing and Assessment Materials for Future Computer Science Teachers. *Higher Learning Research Communications* **16**(1). DOI: 10.18870/hlrc.v16i1.1623.

Milloshi, R., Lako, A.Z., 2024, The Importance of Nursing Management for Patients with Internal Diseases in Albania. *Materia Socio-Medica* **36**(4), 294–299. DOI: 10.5455/msm.2024.36.294-299.

Mohanraj, J., Yuen, K.O., Eshvary, P., Sangwan, A., Thasya, A., Sahlan, N.S., Ashikin, N., Kaur, H., 2019, Exploring the Digital Divide between Pre-clinical Teachers and Students in an Integrated Medical Curriculum from a Malaysian Private University. *Universal J. Educ. Res.* **7**(12A), 1–9. DOI: 10.13189/ujer.2019.071901.

Mutalova, Zh.S., Shaushenova, A.G., Isakova, G.O., Nurpeisova, A.A., Ongarbayeva, M.B., Abdygalikova, G.A., 2024, A method for human recognition from facial images based on guided point movement. *Acad. Sci. J. Comput. Sci.* **3**, 118-137.

Pabba, C., Kumar, P., 2022, An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition. *Expert Syst.* **39(1)**, e12839.

Paul, J., Mitra, A., Kulhadiya, D., Pawar, T., Roy, A., Sil, J., 2025, A comprehensive approach to real-time attendance systems: integrating face recognition and emotion detection and with web technologies. *Procedia Comput. Sci.* **258**, 3436-3446.

Raj, L.V., Sasi, S., Rajeswari, P., Pushpa, B.R., Kulkarni, A.V., Biradar, S., 2025, Design of FBG-based optical biosensor for the detection of malaria. *J. Opt.* DOI: 10.1007/s12596-025-02780-x.

Salmanov, V., Mahmudov, N., 2012, Existence of the solution of a discrete optimal control problem for linear concentrated systems with lions special quality test. *2012 4th Int. Conf. Problems Cybernetics Informatics, PCI 2012 – Proc.*, 1–4. Baku, Azerbaijan. DOI: 10.1109/ICPCI.2012.6486464.

Salmanov, V.I., 2025, Optimal Control Problem and Integral Quality Criteria with a Special Gradient Term. *Eurasia Proc. Sci. Technol. Eng. Math.* **34**, 310–319. DOI: 10.55549/epstem.1754735.

Sapakova, B.S., Sarsembayev, A.A., 2023, Review of emotion classification methods based on audio data analysis using deep learning. *Int. J. Inf. Commun. Technol.* **4(1)**, 95-104.

Seidaliyeva, U., Smailov, N., 2025, Leveraging drone technology for enhanced safety and route planning in rock climbing and extreme sports training. *Retos* **63**, 598–609. DOI: 10.47197/retos.v63.110869.

Shaikemelova, A.M., Omarov, B.S., 2025, Emotion recognition in speech using parallel convolutional neural networks. *Bull. Almaty Univ. Power Eng. Telecommun.* **1**, 291.

Smailov, N., Kadyrova, R., Abdulina, K., Uralova, F., Kubanova, N., Sabibolda, A., 2025, Application of facial recognition technologies for enhancing control in information security systems. *Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska* **15**(3), 55–58. DOI: 10.35784/iapgos.7073.

Subramanian, B., Kim, J., Maray, M., Paul, A., 2022, Digital twin model: a real-time emotion recognition system for personalized healthcare. *IEEE Access* **10**, 81155-81165.

Talaat, F.M., 2023, Real-time facial emotion recognition system among children with autism based on deep learning and IoT. *Neural Comput. Appl.* **35**, 12717-12728.

Tan, Q.L., Hashim, S.B., Abdullah, N.L., Zaki, H.O., Zheng, Z., 2025, Bibliometric insights into the impact of values on consumer sustainable environmental behavior: Current trends and future directions. *Sustainable Futures* **9**, 100582. DOI: 10.1016/j.sftr.2025.100582.

Tavolzhanskyi, O.V., Shumeiko, O., Burda, O., Orobets, K.M., Struchaiev, M., 2025, Using Big Data in Criminal Investigations: Between Privacy and Efficiency. *Khazanah Hukum* **7**(3), 312–324. DOI: 10.15575/kh.v7i3.45201.

Uali, A.B., Naukenova, A.S., Korsun, O.N., Tulekbayeva, A.K., 2024, Investigation of the emotional state of operators of oil refining process units using the methodology of deep convolutional neural networks. *Bull. South Kaz. Sci.* **(3)**, 107-121.

Wang, H., Wang, C., Xu, H., Bao, C., Xu, X., 2026, Enhancing real-time facial emotion recognition in classrooms via Attention-ResNet optimization. *Vis. Comput.* **42(1)**, 46.

Wu, Y., Mi, Q., Gao, T., 2025, A comprehensive review of multimodal emotion recognition: techniques, challenges, and future directions. *Biomimetics* **10(7)**, 418.

Zaki, H.O., Kamarulzaman, Y., Mohtar, M., 2021, Cognition and Emotion: Exploration on Consumers Response to Advertisement and Brand. *Jurnal Pengurusan* **63**, 05. DOI: 10.17576/pengurusan-2021-63-05.

Zheng, Z., Hafizuddin-Syah, B.A.M., Zaki, H.O., Tan, Q.L., 2024, A scoping review of the impact of ageing on individual consumers' insurance purchase intentions. *Heliyon* **10**(18), e37501. DOI: 10.1016/j.heliyon.2024.e37501.