UDC 004.8
IRSTI 28.23

https://doi.org/10.55452/1998-6688-2025-22-2-110-126

[1]**Yershov E.,**
Bachelor's student, ORCID ID: 0009-0006-2267-0365,
e-mail: yershov_evan@kaznu.edu.kz
[1]**Orynbassar S.,**
PhD student, ORCID ID: 0009-0001-9124-2560,
e-mail: sayat.orynbassar@kaznu.edu.kz
*[1]**Zholamanov B.,**
PhD student, ORCID ID: 0000-0001-8206-7425,
*e-mail: zholamanov.batyrbek@kaznu.kz
[1]**Nurgaliyev M.,**
PhD, ORCID ID: 0000-0002-6795-5384,
e-mail: madiyar.nurgaliyev@kaznu.edu.kz
[1]**Dosymbetova G.,**
PhD, ORCID ID: 0000-0002-3935-7213,
e-mail: gulbakhar.dossymbetova@kaznu.edu.kz
[1]**Khumarbekkyzy T.,**
Master's student, ORCID ID: 0009-0005-4945-6273,
e-mail: khumarbekkyzy_t@kaznu.edu.kz

[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan

EMOTION CLASSIFICATION USING CONVOLUTIONAL NEURAL NETWORKS WITH DIFFERENT ARCHITECTURES

Abstract

Thermal imaging offers a non-invasive and robust approach to emotion recognition by capturing facial temperature patterns that correlate with psychophysiological states. This study investigates the application of deep neural networks to classify six basic human emotions – happiness, sadness, fear, disgust, anger, and surprise – using facial thermograms. A balanced dataset was collected under controlled experimental conditions, and four de ep learning architectures were evaluated: Convolutional Neural Network (CNN), Fully Convolutional Network (FCN), EfficientNet, and MobileNet. The models were trained and tested on a curated set of preprocessed thermal facial images. Among the evaluated architectures, FCN achieved the highest classification accuracy of 90.04%. The results demonstrate that deep learning models, particularly FCNs, are well-suited for emotion recognition from thermal data, with potential applications in psychophysiological monitoring, healthcare, and real-time human-computer interaction systems.

Keywords: CNN, Efficient Net, Mobile Net, Fully Convolution Network, thermograms, neural networks.

Introduction

With the growing interest in developing intelligent systems capable of interpreting human behavior, object and image recognition technologies are rapidly advancing. These technologies, which fall under the field of computer vision, are widely applied in areas such as medicine, robotics, smart homes, autonomous driving, and human-computer interaction. One actively explored application is emotion recognition, which is being integrated into consumer electronics, hospital monitoring systems, smart cars, and personal devices [1, 2]. The effectiveness of such systems depends on both the quality of input data and the choice of recognition algorithms. While traditional approaches rely

on RGB images and handcrafted features, these are susceptible to visual obstructions such as facial expressions, lighting changes, and background variations.

To overcome these limitations, thermal imaging (thermography) has emerged as a promising alternative. Thermograms reflect the temperature distribution across the human face, which correlates with emotional states due to autonomic nervous system responses [6]. This modality is difficult to manipulate voluntarily and less affected by environmental noise, making it particularly valuable for affective computing. In recent years, deep learning, especially convolutional neural networks (CNNs), has proven effective in analyzing spatially structured data such as images. CNNs and their variants offer end-to-end learning capabilities and outperform traditional machine learning algorithms in most vision-related tasks [3–5]. Their application to thermal data has shown significant promise in emotion classification [6].

Earlier works have used conventional classifiers–such as SVM, k-NN, logistic regression, and naïve Bayes–combined with feature extraction methods like HOG, Fourier or wavelet transforms [7–14]. While useful, these methods depend heavily on manual feature engineering and often require dimensionality reduction techniques (e.g., PCA, LDA) or ensemble classifiers (e.g., Random Forests) [15–18]. However, they generally underperform compared to deep neural architectures on complex tasks involving spatial and contextual patterns.

This paper investigates the effectiveness of several deep neural network architectures–CNN, Fully Convolutional Network (FCN), EfficientNet, and MobileNet–for classifying six basic human emotions using thermal facial images. The aim is to evaluate the adaptability and classification performance of these models on a balanced and preprocessed dataset captured under controlled experimental conditions.

The main contributions of this work are as follows:

A comparative analysis of four deep learning architectures applied to thermal facial images for emotion recognition.

The construction of a balanced dataset using thermal imaging, capturing six fundamental emotional states under controlled experimental conditions.

The Research Methods section describes the experimental setup, dataset characteristics, preprocessing steps, and the neural network architectures employed in this study, including their configurations and underlying mathematical formulations. The Results and Discussion sections present and interpret the experimental findings obtained from training on thermal image data, with an emphasis on classification accuracy, learning dynamics, and model adaptability. Finally, the Conclusion section summarizes the key outcomes of the work and outlines potential directions for future research.

Materials and Methods

To develop a robust and high-quality neural network classification model, thermograms of people were collected while they were watching various videos that evoked six basic emotions. The videos evoked the following emotions: happiness, sadness, fear, disgust, anger, and surprise. Using a thermal camera, facial thermograms of the experiment participants were captured 15–20 times, depending on the manifestation of emotions while watching each video.

The experimental sessions were conducted in a controlled indoor environment at room temperature. Each participant sat comfortably in front of a monitor placed 0.7 meters away, which displayed audiovisual stimuli. The stimuli were carefully selected to elicit specific emotional responses. Each clip lasted 4 minutes, followed by a short break to allow emotional state normalization. The thermal camera (Fluke TiS20+ MAX; IR resolution: 120×90, 8–14 μm spectral range, 60 mK sensitivity) was mounted on a tripod 1 meter away, perpendicular to the face.

The dataset included ten healthy participants aged between 18 and 19 years. A total of 821 original thermal images were collected and labeled based on the elicited emotion. Each image was preprocessed by cropping around the face and resizing to 48×48 pixels to optimize training efficiency.

Data augmentation was applied using transformations such as rotation (±10°), horizontal flipping (50% chance), scaling (90–110%), and translation (±3 pixels), producing 821 additional samples.

Thus, the final dataset consisted of 1,642 thermographic facial images, evenly distributed across six emotion classes (274 per class, except anger with 272 images). The dataset was split into 80% training and 20% testing sets to ensure model generalizability. Figure 1 presents representative samples of facial thermograms for each emotion.

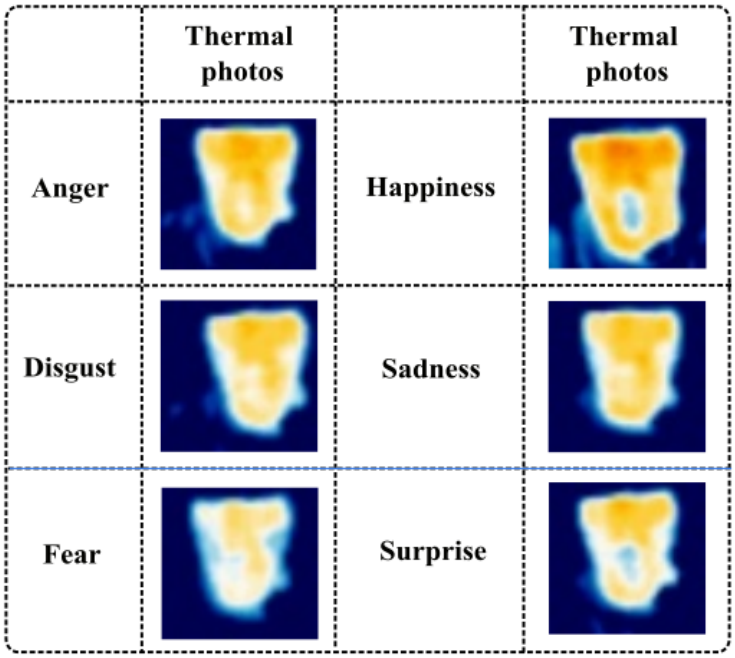

Figure 1 – Thermograms of a human face in different emotions

CNN

Unlike fully connected networks, CNNs have the property of local connectivity and parameter sharing, which reduce the complexity of the model and allow for efficient handling of spatial dependencies [19].

The basic operation in CNN is convolution. The input to a layer is a tensor X of size H×W×C. The convolution is performed with a kernel $W^{\{k\}}$ of size F×F×C. H is the height of the image and W is the width of the image, C is the number of channels, F is the kernel size, and k is the filter index. The filter or image is the image obtained at the output after convolution. The output of the convolution operation is calculated as (1):

$$y_{i,j}^{\{k\}} = \sum_{m=1}^{F} \sum_{n=1}^{F} \sum_{c=1}^{C} \left(X_{i+m,j+n,c} * W_{m,n,c}^{\{k\}} \right) + b^{\{k\}} \quad (1)$$

where $b^{\{k\}}$ is the bias and σ is the activation function (typically ReLU).

The ReLU activation function is defined as:

$$\sigma(z) = max(0,z) \quad (2)$$

The result of the convolution is an output tensor of size H'×W'×K, where K is the number of filters applied to the layer. After the convolution layer, a pooling layer is often applied, which

reduces the dimensionality of the data and thus reduces the number of parameters in the model, and also makes the network more robust to minor shifts in the image. The max pooling operation, which extracts the maximum value in each window of size p×p (3), reduces the dimensionality of the tensor by a factor of p along each spatial axis.

$$y_{i,j,k} = max_{m=1,\dots,p} max_{m=1,\dots,p} X_{i+m,j+n,k} \quad (3)$$

where X is the input tensor, $y_{i,j,k}$ is the subsampling result.

The Flatten layer transforms the 3D output tensor into a 1D vector. If the input tensor is of shape H×W×K, then the output is transformed into a vector of length H×W×K. After the convolution and pooling layers that extract spatial features, the CNN adds fully connected layers that provide classification (4).

$$y_i = \sigma(\sum_{i=1}^{N} w_{i,j} * x_i + b_i) \quad (4)$$

where x_i is the input vector, $w_{i,j}$is the weight coefficient for neuron j, b_j is the bias, and σ is the activation function (usually ReLU or softmax).

For a classification problem, the output layer is typically a layer with n neurons, where n is the number of classes, and a softmax activation function. Softmax transforms the output values into probabilities for each class (5).

$$Softmax(z_i) = \frac{e^{z_i}}{\sum_{j=0}^{n} e^{z_i}} \quad (5)$$

where $z_i$is the activation of the i-th neuron before softmax.

To train the network for the classification task, the categorical cross-entropy loss function (6) is used.

$$L = -\sum_{i=1}^{N} \sum_{j=1}^{C} y_{i,j} * log(\hat{y}_{i,j}) \quad (6)$$

where $y_{i,j}$ is the true class label and $\widehat{y_{i,j}}$ is the predicted probability.

Network optimization is performed using stochastic gradient descent (SGD) or its modifications such as Adam. At each optimization step, the parameters are updated using the gradient of the loss function for each parameter (7).

$$w \leftarrow w - \eta * \frac{dL}{dw} \quad (7)$$

where η is the learning rate, and $\frac{dL}{dw}$ is the gradient of the loss function with respect to weight w.

Thus, the CNN gradually adapts to the input data by changing the weights to minimize the prediction error and improve classification accuracy (Figure 2).

The model includes convolutional layers with ReLU activation, max-pooling layers to reduce spatial dimensions, a flattening operation, and fully connected layers for classification. The final softmax layer outputs the predicted probabilities for the six basic emotion classes.

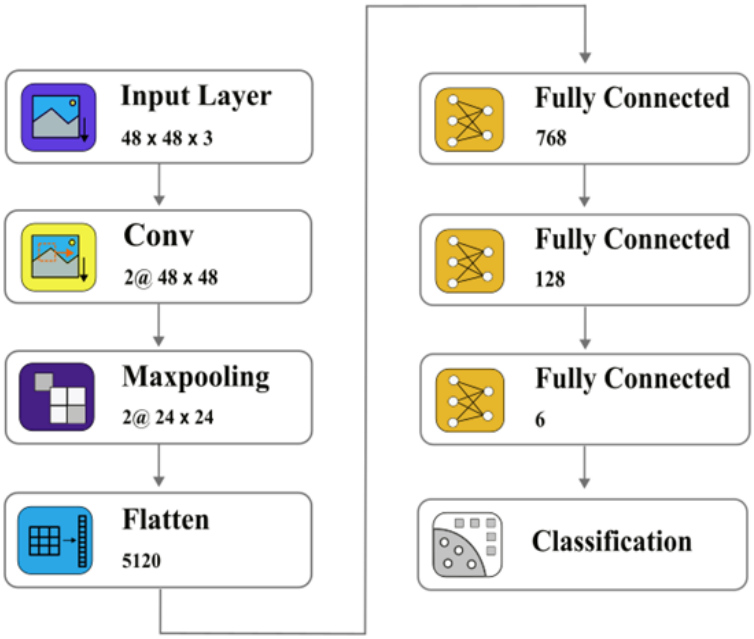

Figure 2 – Architecture of the Convolutional Neural Network (CNN) used in this study

Fully Convolution Network

Fully Convolutional Networks (FCNs) [20] are a neural network architecture in which all layers are convolutional or pooling, allowing the model to be applied to inputs of arbitrary size. Unlike conventional CNNs, which often terminate in fully connected layers for classification, FCNs consist exclusively of convolutional layers and are used primarily for tasks requiring dense pixel-level prediction, such as image segmentation.

In FCN, each layer is either a convolutional layer or a pooling layer. By eliminating fully connected layers, the network preserves spatial structure at each layer, which is important for tasks that require detailed information about the position and shape of objects. Pixel-wise probability maps are used for prediction, allowing each pixel to be classified, rather than just the entire image. At the level of convolutional and pooling layers, FCN is similar to CNN, but unlike CNNs, where convolutions and pooling can be used to extract features before fully connected layers, FCNs use progressive convolutions and pooling to gradually reduce spatial resolution, which helps to reveal hierarchical features.

After the convolution and pooling layer, dense prediction tasks such as segmentation require restoring the original image resolution. This is done by an upsampling operation, most commonly implemented by transposed convolution (also called inverse convolution or convolutional deconvolution) (8). The transposed convolution step increases the dimensionality of the output tensor, restoring the spatial resolution.

$$Y_{i,j,k} = \sum_{m=1}^{F} \sum_{n=1}^{F} \left(X_{i//s+m,j//s+n,k} * W_{m,n,k} \right) \quad (8)$$

where s is the step of the transposed convolution.

FCN also uses a skip connection mechanism to combine high-level and low-level features. This allows fine-grained spatial features from earlier layers to be combined with semantic features from deeper layers. As a result, FCN can reconstruct the resolution with higher accuracy and preserve spatial details important for segmentation. The final feature map Y is the sum of the early and late layer feature maps multiplied by weights and represents the feature map after reconstruction (9).

$$Y = \alpha Y_{low} + \beta Y_{high} \quad (9)$$

where Y_{low} is the feature map from the early layer, Y_{high} is the feature map from the deeper layer, and α and β are the weight coefficients that define the contribution of each layer.

For the semantic segmentation task, FCN generates a class map for each pixel as output, which allows each pixel of the input image to be classified. The prediction for a pixel can be expressed in terms of softmax (10).

$$p_{i,j,c} = \frac{exp(y_{i,j,c})}{\sum_{k=1}^{C} exp(y_{i,j,c})} \quad (10)$$

where $p_{i,j,c}$ is the probability of class c for pixel (i, j), and C is the total number of classes.

When solving image segmentation problems, the best results are achieved by using the categorical cross-entropy loss function at the pixel level (6). FCN is trained similarly to other neural networks using the gradient descent method, updating the weights to minimize the loss function. After training, the model can be applied to images of arbitrary size, which is an important advantage of FCN, since it does not depend on a fixed input image size. Thus, FCN allows for dense prediction at the pixel level, which makes it particularly suitable for semantic segmentation problems and other problems that require spatial understanding of data (Figure 3). By performing segmentation and adding classification accordingly, it is possible to improve the performance of the neural network.

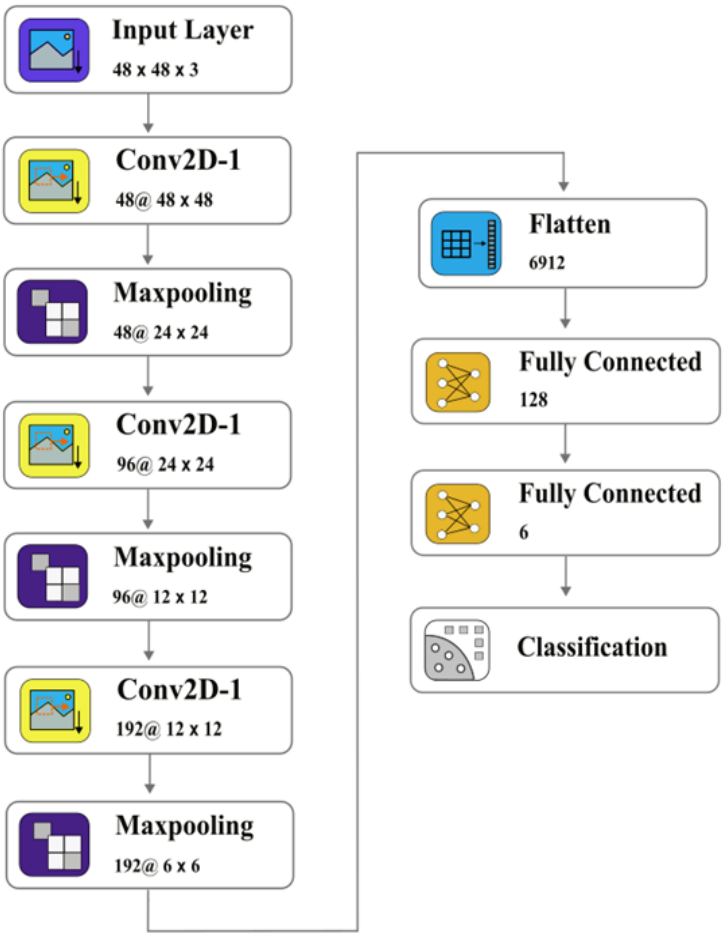

Figure 3 – Architecture of the Fully Convolutional Network (FCN) used in this study

The network consists entirely of convolutional and pooling layers and omits fully connected layers traditionally used in CNNs. For image-level classification, global average pooling is applied after the final convolutional block, followed by a dense softmax layer that outputs class probabilities.

This adaptation enables the FCN to perform robust classification while preserving spatial features and minimizing the number of parameters.

EfficientNet

EfficientNet models are a family of convolutional neural networks that offer an optimized method for scaling neural networks, allowing them to achieve high classification accuracy with minimal computational effort [21]. EfficientNet was proposed by the Google AI team and demonstrates a fundamentally new approach to scaling the depth, width, and resolution of a network, thereby achieving high performance on a wide range of tasks. One of the key ideas of EfficientNet is the simultaneous scaling of the width, depth, and resolution of the network (11). Traditional scaling approaches typically increase only one of these parameters, which can lead to inefficient use of resources. EfficientNet proposes a method called "joint scaling" (Compound Scaling).

$$d = \alpha^{\Phi},\ w = \beta^{\Phi},\ \ r = \gamma^{\Phi} \tag{11}$$

where d is the network depth (number of layers); w is the network width (number of channels in each layer); r is the image resolution; Φ is a scaling factor that controls the overall computational power of the model; α, β, γ are parameters that control the degree of increase in depth, width, and resolution, respectively.

The parameters α, β и γ are selected empirically to maintain a balance between them when changing computing resources.

EfficientNet-B0 is the base architecture of EfficientNet, on which other models (B1, B2, etc.) are scaled. The base model is built on an improved version of MobileNetV2, and is based on Depthwise Convolution and Pointwise Convolution blocks. An important feature of the architecture is also the use of swish activation (12):

$$swish(x) = x * \sigma(x) = \frac{x}{1+exp(-x)} \tag{12}$$

where $\sigma(x)$ is the sigmoid. The swish function helps improve the learning ability of the model by smoothing the activation.

An important component of EfficientNet is the MBConv (Mobile Inverted Bottleneck Convolution) block, which was borrowed from MobileNetV2 and adapted. MBConv includes a pointwise convolution (1x1) to increase the number of channels (13), a depthwise convolution (Depthwise Convolution) to process each channel separately (14), a pointwise convolution to reduce the number of channels back to the original (15).

$$x' = W^{(1)} * x \tag{13}$$

$$x'' = W^{(2)} * x, \tag{14}$$

$$y = W^{(3)} * x', \tag{15}$$

where $W^{(1)}$ are the point convolution weights; * is the convolution operation; $W^{(2)}$ are the depthwise convolution weights applied to each channel independently; $W^{(3)}$ are the weights of the last point convolution to restore the number of channels to the original value.

MBConv also uses a dilation factor t, which determines how many times the number of channels is increased before applying depthwise convolution. The MBConv block also uses a skip connections mechanism, which allows the model to preserve useful features at each layer. EfficientNet offers a set of models from B0 to B7, which are scaled by the parameter Φ. As Φ, the depth, width, and resolution are increased according to pre-fitted values α, β, γ. This allows EfficientNet-B7 to achieve high

accuracy on a variety of tasks, outperforming standard CNNs with less computational cost. Training is done using the Adam or RMSprop algorithm with adaptive learning rate scaling, which helps in efficient training on large datasets. EfficientNet is optimized for images of arbitrary resolution and is suitable for a variety of computer vision tasks, including classification, segmentation, and object detection. Thanks to its scaling approach, EfficientNet-B7, for example, demonstrates better accuracy than many classical architectures such as ResNet and Inception, while requiring significantly less computational resources. EfficientNet offers a cost-effective and flexible solution for computer vision tasks, providing efficient architectural scaling and high performance with minimal computational costs (Figure 4).

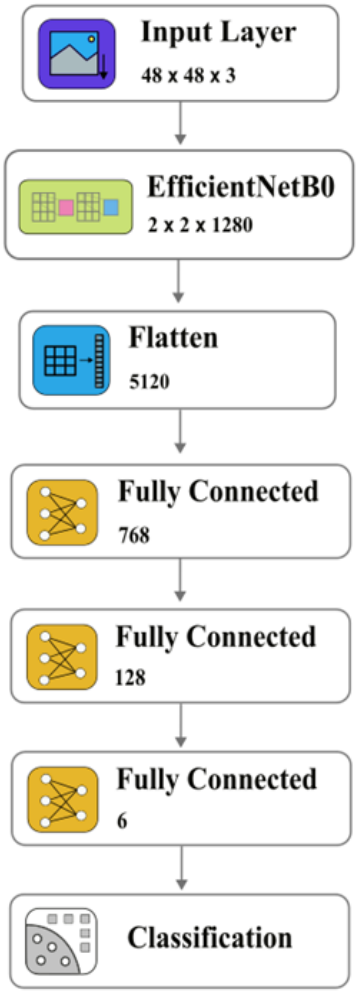

Figure 4 – Architecture of the EfficientNetB0-based neural network used in this study

The network uses a preconfigured EfficientNetB0 backbone to extract high-level spatial features from 48×48×3 input images. The output feature map (2×2×1280) is flattened and passed through three fully connected layers (768, 128, and 6 neurons). The final layer uses softmax activation to output the probabilities for six emotion classes. This architecture combines EfficientNet's optimized compound scaling with a custom classification head for emotion recognition.

MobileNet

MobileNet is a class of convolutional neural network architectures designed for computationally constrained devices such as mobile devices and embedded systems [22]. The basic idea of MobileNet is to use Depthwise Convolution and Pointwise Convolution to reduce the computational complexity and model size. MobileNet introduces the concept of Depthwise Separable Convolution, which

decomposes the standard convolution into a depthwise convolution and a pointwise convolution. Depthwise convolution is performed separately for each channel, which significantly reduces the number of operations. Pointwise convolution combines the results from each channel using a regular 1x1 convolution to mix the channels. This approach reduces the number of computations compared to traditional convolution while maintaining good classification accuracy.

To further tune the model complexity, MobileNet introduces a width multiplier and a resolution multiplier. The width multiplier is used to reduce the number of channels in each layer. If the width multiplier is less than 1, it reduces the number of filters, reducing the computational complexity and model size. The resolution multiplier reduces the resolution of the input image, which also reduces the number of computations. Like other CNN architectures, MobileNet uses the categorical cross-entropy function for the classification task. The use of depthwise sparse convolution reduces the computational requirements by 8–9 times compared to traditional convolutional networks of similar size. This makes MobileNet effective for tasks that require low latency, low power consumption, and compact models, such as tasks on mobile devices. MobileNet achieves a trade-off between accuracy and performance, especially with low values of the multipliers and r. MobileNet has been successfully applied to classification, detection, and segmentation tasks where high efficiency is required. Improved models such as MobileNetV2 and MobileNetV3 have also been developed based on MobileNet, introducing additional optimizations to improve classification quality while maintaining efficiency. MobileNet offers a simple and efficient architecture for mobile and embedded applications, achieving high accuracy with minimal computational overhead, making it one of the popular models for devices with limited capabilities (Figure 5).

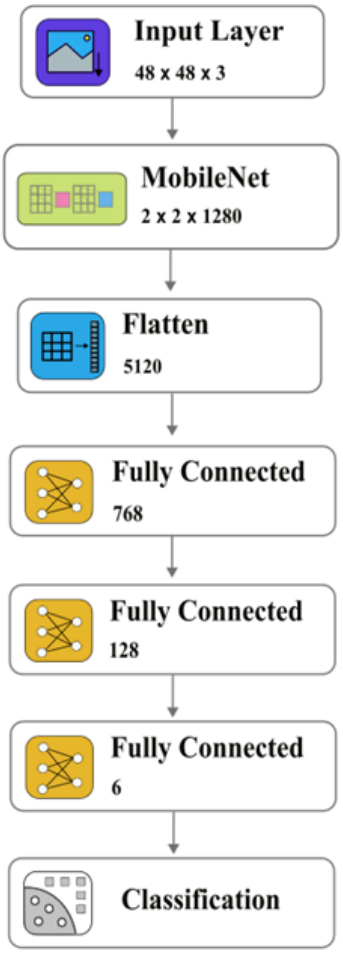

Figure 5 – Architecture of the MobileNet-based neural network used in this study

The model uses a MobileNet backbone for efficient feature extraction from low-resolution input images (48×48×3), producing a compact feature map of size 2×2×1280. This output is flattened and passed through three fully connected layers (768, 128, and 6 neurons). A final softmax layer outputs classification probabilities for the six basic emotions. The architecture is optimized for low computational cost, making it suitable for deployment on mobile or embedded systems.

Results

The number of training epochs for each neural network is defined as the number of epochs at which the error stops decreasing within 20 epochs. The batch size was 32. The learning rate was set to 0.0001 and the Adam optimizer was used for all models. The Adam optimizer was used with default parameters (β_1=0.9, β_2=0.999, ε=1e−8).

Table 1 – Final performance of neural networks after training

Neural network architecture	CNN	Fully Convolution Network	Efficient Net	MobileNet
loss	0.053	0.0502	0.2942	0.0636
accuracy	0.9908	0.9880	0.9114	0.9991
val loss	0.6715	0.513	1.8307	0.5003
val accuracy	0.8524	0.9004	0.31	0.8413

The Fully Convolution Network (FCN) has the best training results, and a relatively small error compared to CNN, Efficient Net, and MobileNet. Although FCNs are typically applied in segmentation tasks, in this study the FCN architecture was adapted for image-level classification by applying global average pooling followed by a softmax-activated dense layer. This allowed for aggregating spatial features into class-level predictions. FCN demonstrates the best balance between accuracy and loss function on the validation set. MobileNet achieves high accuracy on training, but its performance on validation is lower than that of FCN. EfficientNet requires improvement or adaptation, since its low performance on validation indicates a strong discrepancy between training and testing. CNN maintains stable results, which makes it suitable for problems where a trade-off between accuracy and model complexity is required. Such results emphasize the importance of choosing architecture depending on the specifics of the problem and available resources. Figures 6–9 present the dynamics of neural networks training.

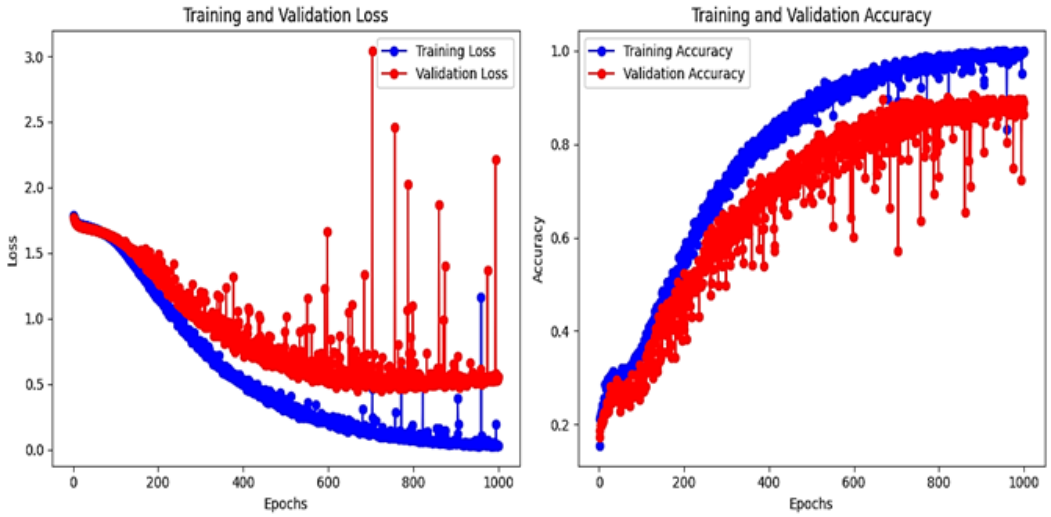

Figure 6 – Training and validation loss and accuracy curves
for the Fully Convolutional Network (FCN)

The FCN demonstrates a steady decrease in training loss and consistent improvement in validation accuracy. While the validation loss exhibits moderate oscillations throughout training, the model maintains high generalization performance, ultimately achieving the best validation accuracy among all tested architectures. This suggests that the FCN structure, adapted for image-level classification, effectively captures relevant thermal features despite dataset limitations.

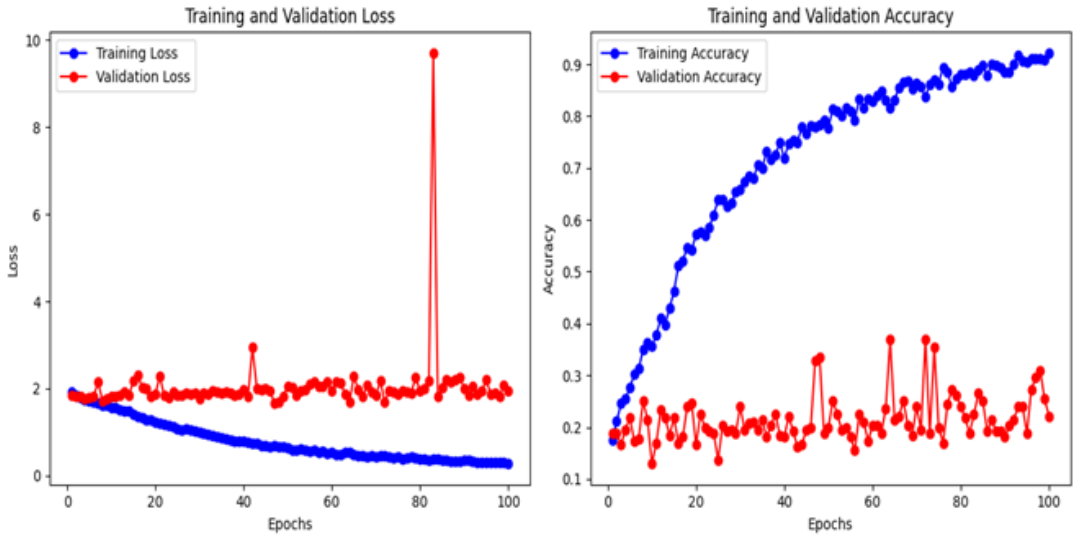

Figure 7 – Training and validation loss
and accuracy curves for EfficientNetB0

The model achieves progressively lower training loss and high training accuracy, indicating effective fitting on the training data. However, the validation loss remains high and unstable, while validation accuracy fluctuates with low values, showing a clear overfitting trend. The large gap between training and validation metrics suggests insufficient generalization, possibly due to the model's complexity relative to the dataset size.

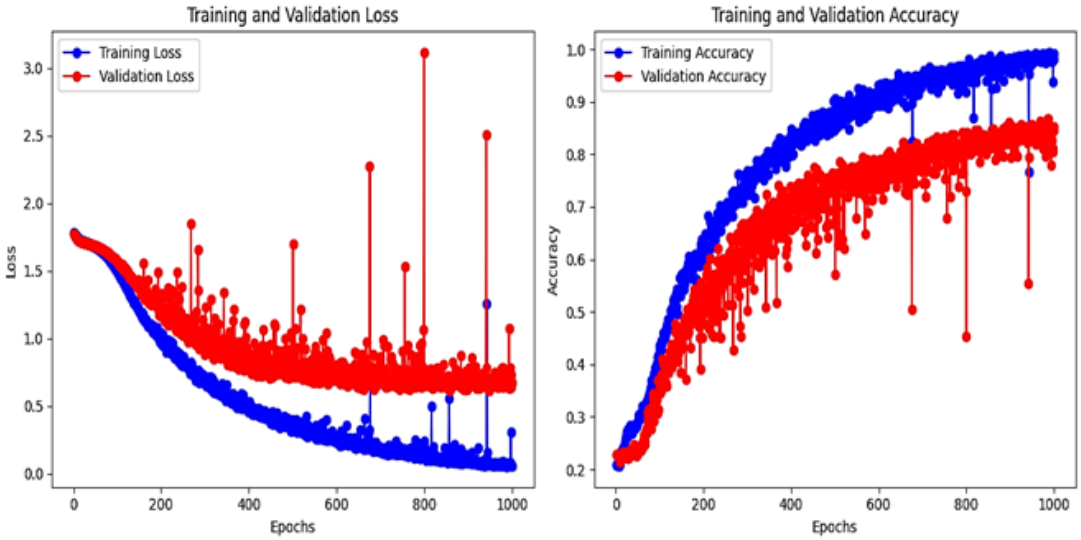

Figure 8 – Training and validation loss and accuracy curves
for the Convolutional Neural Network (CNN)

The CNN exhibits steady improvement in both training and validation accuracy. Training loss decreases consistently, while validation loss shows moderate oscillations, especially in the later epochs. This behavior suggests that the model generalizes reasonably well but may be sensitive to the learning rate or the absence of regularization techniques. Overall, CNN demonstrates robust performance with a good balance between accuracy and model complexity.

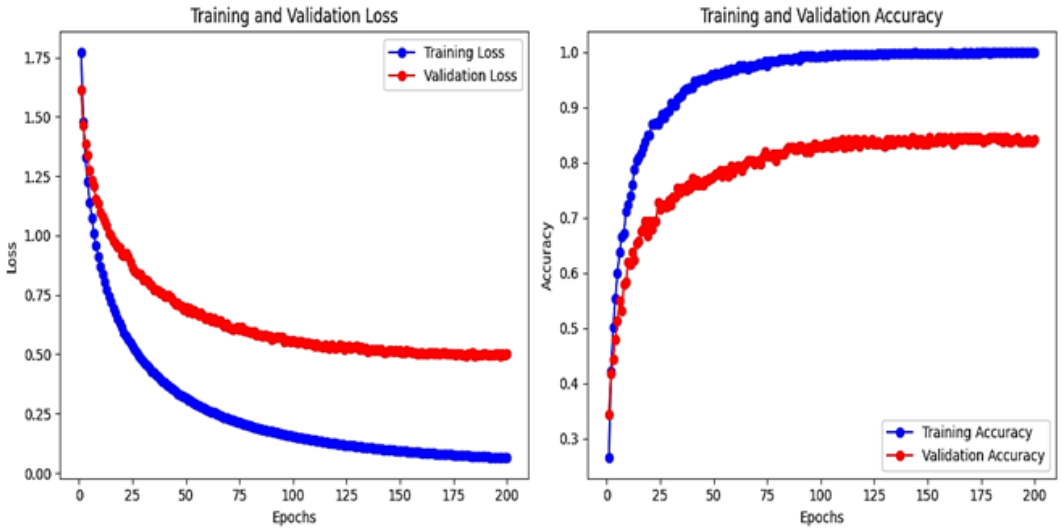

Figure 9 – Training and validation loss and accuracy curves for the MobileNet architecture

MobileNet demonstrates fast and stable convergence, with both training and validation loss decreasing smoothly. The validation accuracy increases steadily and reaches over 80%, although a performance gap between training and validation persists, indicating a degree of overfitting. Compared to other models, MobileNet shows the most stable learning dynamics with minimal oscillations, highlighting its efficiency and suitability for small-scale datasets and low-resource applications.

Discussion

The work performed demonstrates the relevance of using thermal images in emotion classification tasks utilizing neural network architectures. A distinctive feature of this study is the use of thermograms, which are more resilient to visual variations such as facial expressions, lighting conditions, and background clutter, compared to RGB images.

Among the evaluated models, the Fully Convolutional Network (FCN) demonstrated the highest classification accuracy, which is consistent with its architectural strength, namely, the ability to preserve spatial dependencies and integrate multi-scale feature representations. Its successful application underscores the adaptability of FCNs for tasks requiring detailed data processing, including segmentation. The CNN model also yielded stable and acceptable results, making it a viable option when a trade-off between computational cost and accuracy is needed. However, its performance was inferior to FCN in terms of spatial representation capacity. EfficientNet, despite its theoretical advantages in parameter efficiency and scalability, exhibited weak generalization on the validation set, suggesting a need for hyperparameter tuning or architecture adjustments tailored to the thermographic input modality. MobileNet showed high training accuracy but weaker performance on validation, indicating a tendency toward overfitting. These outcomes highlight the importance of aligning neural network architecture with the characteristics of the data and task-specific constraints. Additionally, the relatively small size of the dataset (1,642 images) may have

limited the learning potential of more complex models such as EfficientNet. Future work should therefore consider expanding the dataset and applying advanced data augmentation techniques to improve generalization and model robustness.

Furthermore, to better interpret model behavior and identify class-specific strengths and weaknesses, future research should incorporate complementary performance metrics such as precision, recall, and confusion matrices. These metrics would provide deeper insight into which emotions are more prone to misclassification and help guide targeted architectural or preprocessing improvements.

This study underscores the critical role of neural network architecture selection in thermal image-based emotion classification. The FCN architecture appears most promising under current conditions. However, ongoing research should aim to improve the performance of other architectures through thermogram-specific adaptations and extended evaluation using more comprehensive performance metrics.

Conclusion

The study examined the capabilities of various neural network architectures for classifying human emotions using thermograms. The results showed that the Fully Convolution Network (FCN) demonstrated the best balance between accuracy and robustness on the validation set, achieving a classification accuracy of 90.04%, making it a promising tool for analyzing thermograms. CNN also demonstrated stability and acceptable accuracy, which allows us to recommend it for tasks that require a balanced approach between computational complexity and accuracy. Despite the high potential of EfficientNet, its low performance at the validation stage indicates the need for improvement, including tuning hyperparameters and increasing the size of the training set. MobileNet demonstrated excellent results at the training stage, but its validation accuracy was lower, which may indicate a tendency to overfitting.

Thus, the work emphasizes the key role of architectural adaptation of neural networks for the specifics of thermogram processing tasks. The use of thermograms is a promising direction for the analysis of human emotions due to their resistance to external factors. For further research, it is advisable to focus on optimizing existing architectures and developing new approaches focused on the specific features of thermographic data, as well as increasing the volume of data for training and validation of models.

REFERENCES

1 Szeliski R. Computer vision: algorithms and applications. – Springer Nature. – 2022. https://doi.org/10.1007/978-3-030-34372-9

2 Huang Zi-Yu, et al. A study on computer vision for facial emotion recognition // Scientific reports. – 2023. – Vol. 13. – No. 1. – P. 8425. https://doi.org/10.1038/s41598-023-35446-4

3 Issa D., Demirci MF, Yazici A. Speech emotion recognition with deep convolutional neural networks // Biomedical Signal Processing and Control. – 2020. – Vol. 59. – P. 101894. https://doi.org/10.1016/j.bspc.2020.101894

4 Gautam, Chahak, and K.R. Seeja. Facial emotion recognition using Handcrafted features and CNN // Procedia Computer Science. – 2023. – Vol. 218. – P. 1295–1303. https://doi.org/10.1016/j.procs.2023.01.108

5 Prasad, Babu Rajendra, and B. Sai Chandana. Human face emotions recognition from thermal images using DenseNet // International journal of electrical and computer engineering systems. – 2023. – Vol. 14. – P. 155–167. https://doi.org/10.32985/ijeces.14.2.5

6 Assiri, Basem, and Mohammad Alamgir Hossain. Face emotion recognition based on infrared thermal imagery by applying machine learning and parallelism // Mathematical Biosciences and Engineering. – 2023. – Vol. 20. – P. 913–929. https://doi.org/10.3934/mbe.2023042

7 Zuo W. et al. Gradient histogram estimation and preservation for texture enhanced image denoising // IEEE transactions on image processing. – 2014. – Vol. 23. – No. 6. – P. 2459–2472. https://doi.org/10.1109/TIP.2014.2316423

8 Bouchene, Mohammed Mehdi. Bayesian optimization of histogram of oriented gradients (HOG) parameters for facial recognition // The Journal of Supercomputing. – 2024. – Vol. 80. – No. 14. – P. 20118–20149. https://doi.org/10.1007/s11227-024-06259-7

9 Ma J. Based on the fourier transform and the wavelet transformation of the digital image processing // 2012 International Conference on Computer Science and Information Processing (CSIP). – IEEE, 2012. – P. 1232–1234. https://doi.org/10.1109/CSIP.2012.6309081

10 Lindeberg T. Scale invariant feature transform. – 2012. https://doi.org/10.4249/scholarpedia.10491

11 Thai LH, Hai TS, Thuy NT Image classification using support vector machine and artificial neural network // International Journal of Information Technology and Computer Science. – 2012. – Vol. 4. – No. 5. – P. 32–38. https://doi.org/10.5815/ijitcs.2012.05.05

12 Ma L., Crawford MM, Tian J. Local manifold learning-based k-nearest-neighbor for hyperspectral image classification // IEEE Transactions on Geoscience and Remote Sensing. – 2010. – Vol. 48. – No. 11. – P. 4099–4109. https://doi.org/10.1109/TGRS.2010.2055876

13 Dong Q., Zhu X., Gong S. Single-label multi-class image classification by deep logistic regression // Proceedings of the AAAI conference on artificial intelligence. – 2019. – Vol. 33. – No. 01. – P. 3486–3493. https://doi.org/10.1609/aaai.v33i01.33013486

14 Timofte R., Tuytelaars T., Van Gool L. Naive bayes image classification: beyond nearest neighbors // Asian Conference on Computer Vision. – Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. – P. 689–703. https://doi.org/10.1007/978-3-642-37331-2_52

15 Wang H. et al. Image recognition of plant diseases based on principal component analysis and neural networks // 2012 8th International Conference on Natural Computation. – IEEE, 2012. – P. 246–251. https://doi.org/10.1109/ICNC.2012.6234701

16 Tharwat A. et al. Linear discriminant analysis: A detailed tutorial // AI communications. – 2017. – Vol. 30. – No. 2. – P. 169–190. https://doi.org/10.3233/AIC-170729

17 Zhao, Shuping, et al. Linear discriminant analysis. Nature Reviews Methods Primers. – 2024. – Vol. 4. – No. 1. – P. 70. https://doi.org/10.1038/s43586-024-00346-y

18 Xu B., Ye Y., Nie L. An improved random forest classifier for image classification //2012 IEEE international conference on information and automation. – IEEE, 2012. – P. 795–800. https://doi.org/10.1109/ICInfA.2012.6246927

19 Badrulhisham NAS, Mangshor NNA Emotion recognition using convolutional neural network (CNN) // Journal of Physics: Conference Series. – IOP Publishing, 2021. – Vol. 1962. – No. 1. – P. 012040. http://dx.doi.org/10.1088/1742-6596/1962/1/012040

20 Maggiori E. et al. Fully convolutional neural networks for remote sensing image classification // 2016 IEEE international geoscience and remote sensing symposium (IGARSS). – IEEE, 2016. – P. 5071–5074. https://doi.org/10.1109/IGARSS.2016.7730322

21 Espejo-Garcia B. et al. Using EfficientNet and transfer learning for image-based diagnosis of nutrient deficiencies //Computers and Electronics in Agriculture. – 2022. – Vol. 196. – P. 106868. https://doi.org/10.1016/j.compag.2022.106868

22 Wang W. et al. A novel image classification approach via dense- MobileNet models //Mobile Information Systems. – 2020. – Vol. 2020. – No. 1. – P. 7602384. https://doi.org/10.1155/2020/7602384

REFERENCES

1 Szeliski R. Computer vision: algorithms and applications. Springer Nature (2022). https://doi.org/10.1007/978-3-030-34372-9

2 Huang, Zi-Yu, et al. A study on computer vision for facial emotion recognition. Scientific reports, 13 (1), 8425 (2023). https://doi.org/10.1038/s41598-023-35446-4

3 Issa D., Demirci MF, Yazici A. Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control, 59, 101894 (2020). https://doi.org/10.1016/j.bspc.2020.101894

4 Gautam, Chahak, and K.R. Seeja. Facial emotion recognition using Handcrafted features and CNN. Procedia Computer Science, 218, 1295–1303 (2023). https://doi.org/10.1016/j.procs.2023.01.108

5 Prasad, Babu Rajendra, and B. Sai Chandana. Human face emotions recognition from thermal images using DenseNet. International journal of electrical and computer engineering systems, 14 (2), 155–167 (2023). https://doi.org/10.32985/ijeces.14.2.5

6 Assiri, Basem, and Mohammad Alamgir Hossain. Face emotion recognition based on infrared thermal imagery by applying machine learning and parallelism. Mathematical Biosciences and Engineering, 20 (1), 913–929 (2023). https://doi.org/10.3934/mbe.2023042

7 Zuo W. et al. Gradient histogram estimation and preservation for texture enhanced image denoising. IEEE transactions on image processing, 23 (6), 2459–2472 (2014). https://doi.org/10.1109/TIP.2014.2316423

8 Bouchene, Mohammed Mehdi. Bayesian optimization of histogram of oriented gradients (HOG) parameters for facial recognition. The Journal of Supercomputing, 80 (14), 20118–20149 (2024). https://doi.org/10.1007/s11227-024-06259-7

9 Ma J. Based on the fourier transform and the wavelet transformation of the digital image processing. 2012 International Conference on Computer Science and Information Processing (CSIP) (IEEE, 2012), pp. 1232–1234. https://doi.org/10.1109/CSIP.2012.6309081

10 Lindeberg T. Scale invariant feature transform (2012). https://doi.org/10.4249/scholarpedia.10491

11 Thai LH, Hai TS, Thuy NT Image classification using support vector machine and artificial neural network. International Journal of Information Technology and Computer Science, 4 (5), 32–38 (2012). https://doi.org/10.5815/ijitcs.2012.05.05

12 Ma L., Crawford MM, Tian J. Local manifold learning-based k-nearest-neighbor for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 48 (11), 4099–4109 (2010). https://doi.org/10.1109/TGRS.2010.2055876

13 Dong Q., Zhu X., Gong S. Single-label multi-class image classification by deep logistic regression. Proceedings of the AAAI conference on artificial intelligence, 33 (01), 3486–3493 (2019). https://doi.org/10.1609/aaai.v33i01.33013486

14 Timofte R., Tuytelaars T., Van Gool L. Naive bayes image classification: beyond nearest neighbors. Asian Conference on Computer Vision (Berlin, Heidelberg: Springer Berlin Heidelberg, 2012), pp. 689–703. https://doi.org/10.1007/978-3-642-37331-2_52

15 Wang H. et al. Image recognition of plant diseases based on principal component analysis and neural networks. 2012 8th International Conference on Natural Computation (IEEE, 2012), pp. 246–251. https://doi.org/10.1109/ICNC.2012.6234701

16 Tharwat A. et al. Linear discriminant analysis: A detailed tutorial. AI communications, 30 (2), 169–190 (2017). https://doi.org/10.3233/AIC-170729

17 Zhao, Shuping, et al. Linear discriminant analysis. Nature Reviews Methods Primers, 4 (1), 70 (2024). https://doi.org/10.1038/s43586-024-00346-y

18 Xu B., Ye Y., Nie L. An improved random forest classifier for image classification. 2012 IEEE international conference on information and automation (IEEE, 2012), pp. 795–800. https://doi.org/10.1109/ICInfA.2012.6246927

19 Badrulhisham NAS, Mangshor NNA Emotion recognition using convolutional neural network (CNN). Journal of Physics: Conference Series, IOP Publishing,1962 (1), 012040 (2021). http://dx.doi.org/10.1088/1742-6596/1962/1/012040

20 Maggiori E. et al. Fully convolutional neural networks for remote sensing image classification. 2016 IEEE international geoscience and remote sensing symposium (IGARSS) (IEEE, 2016), pp. 5071–5074. https://doi.org/10.1109/IGARSS.2016.7730322

21 Espejo-Garcia B. et al. Using EfficientNet and transfer learning for image-based diagnosis of nutrient deficiencies. Computers and Electronics in Agriculture, 196, 106868 (2022). https://doi.org/10.1016/j.compag.2022.106868

22 Wang W. et al. A novel image classification approach via dense- MobileNet models. Mobile Information Systems, 2020 (1), 7602384 (2020). https://doi.org/10.1155/2020/7602384

[1]Ершов Э.,
студент, ORCID ID: 0009-0006-2267-0365,
e-mail: yershov_evan@kaznu.edu.kz
[1]Орынбасар С.,
PhD студент, ORCID ID: 0009-0001-9124-2560,
e-mail: sayat.orynbassar@kaznu.edu.kz
[1]*Жоламанов Б.,
PhD студент, ORCID ID: 0000-0001-8206-7425,
*e-mail: zholamanov.batyrbek@kaznu.kz
[1]Нұрғалиев М.,
PhD, ORCID ID: 0000-0002-6795-5384,
e-mail: madiyar.nurgaliyev@kaznu.edu.kz
[1]Досымбетова Г.,
PhD, ORCID ID: 0000-0002-3935-7213,
e-mail: gulbakhar.dossymbetova@kaznu.edu.kz
[11]Хумарбекқызы Т.,
магистрант, ORCID ID: 0009-0005-4945-6273,
e-mail: khumarbekkyzy_t@kaznu.edu.kz

[1]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы қ., Қазақстан

ӘРТҮРЛІ АРХИТЕКТУРАЛАРЫ БАР КОНВОЛЮЦИЯЛЫҚ НЕЙРОНДЫҚ ЖЕЛІЛЕРДІ ҚОЛДАНУ АРҚЫЛЫ ЭМОЦИЯЛАРДЫ КЛАССИФИКАЦИЯЛАУ

Аңдатпа

Термография – адамның психофизиологиялық күйімен байланысты беттегі температуралық өзгерістерді тіркейтін, эмоцияларды тануға арналған инвазивті емес әрі сенімді әдіс. Бұл зерттеуде адамның бет термограммаларын пайдаланып, алты негізгі эмоцияны – қуаныш, мұң, қорқыныш, жиіркену, ашу және таңғалу – тану үшін терең нейрондық желілерді қолдану қарастырылады. Эксперименттік жағдайларда жиналған теңестірілген деректер жиынтығы негізінде төрт архитектура бағаланды: конволюциялық нейрондық желі (CNN), толық конволюциялық желі (FCN), EfficientNet және MobileNet. Модельдер алдын ала өңделген бет термограммаларында оқытылып, тексерілді. Бағаланған архитектуралардың ішінде FCN ең жоғары – 90,04% дәлдік көрсетті. Бұл зерттеу терең нейрондық желілердің, әсіресе FCN архитектурасының, термографиялық деректер негізінде эмоцияларды тану міндеттерінде тиімді екенін дәлелдейді және оны психофизиологиялық бақылау, денсаулық сақтау, сондай-ақ адам мен машина арасындағы өзара әрекеттесу жүйелерінде қолдануға болатынын көрсетеді.

Тірек сөздер: CNN, Efficient Net, Mobile Net, Fully Convolution Network, термограмма, нейрондық желі.

[1]Ершов Э.,
студент, ORCID ID: 0009-0006-2267-0365,
e-mail: yershov_evan@kaznu.edu.kz
[1]Орынбасар С.,
PhD студент, ORCID ID: 0009-0001-9124-2560,
e-mail: sayat.orynbassar@kaznu.edu.kz
[1]*Жоламанов Б.,
PhD студент, ORCID ID: 0000-0001-8206-7425,
*e-mail: zholamanov.batyrbek@kaznu.kz
[1]Нұрғалиев М.,
PhD, ORCID ID: 0000-0002-6795-5384,
e-mail: madiyar.nurgaliyev@kaznu.edu.kz
[1]Досымбетова Г.,
PhD, ORCID ID: 0000-0002-3935-7213,
e-mail: gulbakhar.dossymbetova@kaznu.edu.kz
[1]Хумарбекқызы Т.,
магистрант, ORCID ID: 0009-0005-4945-6273,
e-mail: khumarbekkyzy_t@kaznu.edu.kz

[1]Казахский национальный университет им. аль-Фараби, г. Алматы, Казахстан

КЛАССИФИКАЦИЯ ЭМОЦИЙ С ИСПОЛЬЗОВАНИЕМ СВЕРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ С РАЗЛИЧНЫМИ АРХИТЕКТУРАМИ

Аннотация

Тепловизионная съемка представляет собой неинвазивный и надежный подход к распознаванию эмоций, фиксируя температурные изменения на лице, связанные с психофизиологическим состоянием человека. В настоящем исследовании рассматривается применение глубоких нейронных сетей для классификации шести базовых эмоций – радость, грусть, страх, отвращение, гнев и удивление – по термограммам лица. Сбалансированный набор данных был собран в контролируемых экспериментальных условиях, и были оценены четыре архитектуры глубокого обучения: сверточная нейронная сеть (CNN), полностью сверточная сеть (FCN), EfficientNet и MobileNet. Модели обучались и тестировались на предварительно обработанных термографических изображениях лица. Среди исследуемых архитектур наивысшую точность – 90.04% – показала FCN. Результаты демонстрируют, что модели глубокого обучения, особенно FCN, хорошо подходят для задач распознавания эмоций по тепловизионным данным и могут быть использованы в психофизиологическом мониторинге, здравоохранении и системах взаимодействия человек – машина в реальном времени.

Ключевые слова: CNN, Efficient Net, Mobile Net, Fully Convolution Network, термограмма, нейронные сети.

Article submission date: 03.12.2024