DOI: 10.37943/22GEBT9085

Evan Yershov
Bachelor student, Faculty of Physics and Technology
yershov_evan@kaznu.edu.kz; orcid.org/0009-0006-2267-0365
Al-Farabi Kazakh National University, Kazakhstan

Madiyar Nurgaliyev
PhD, Faculty of Physics and Technology
madiyar.nurgaliyev@kaznu.edu.kz; orcid.org/0000-0002-6795-5384
Al-Farabi Kazakh National University, Kazakhstan

Gulbakhar Dosymbetova
PhD, Faculty of Physics and Technology
gulbakhar.dossymbetova@kaznu.edu.kz; orcid.org/0000-0002-3935-7213
Al-Farabi Kazakh National University, Kazakhstan

Batyrbek Zholamanov
PhD student, Faculty of Physics and Technology
zholamanov.batyrbek@kaznu.edu.kz; orcid.org/0000-0001-8206-7425
Al-Farabi Kazakh National University, Kazakhstan

Sayat Orynbassar
PhD student, Faculty of Physics and Technology
sayat.orynbassar@kaznu.edu.kz; orcid.org/0009-0001-9124-2560
Al-Farabi Kazakh National University, Kazakhstan

Tomiris Khumarbekkyzy
Master student, Faculty of Physics and Technology
khumarbekkyzy_t@kaznu.edu.kz; orcid.org/0009-0005-4945-6273
Al-Farabi Kazakh National University, Kazakhstan

CLASSIFICATION OF HUMAN EMOTIONS USING THERMOGRAMS AND NEURAL NETWORK

Abstract: As information systems and technologies continue to evolve, there remains a noticeable gap in the efficiency and practicality of data processing algorithms, especially in the field of emotion recognition. This study explores several neural network models designed to classify emotions based on thermal images (thermograms). The dataset used for training included 1,642 images, some of which were generated through augmentation, with all images captured while participants viewed emotionally charged videos. The goal was to recognize six basic emotions: joy, sadness, fear, disgust, anger, and surprise. To identify the most effective architecture, the performance of five models were compared: a standard convolutional neural network (CNN), Quadruplet Network, U-Net, Inception, and SqueezeNet. Each model was trained on the same dataset under consistent conditions. Classification accuracy and validation loss were the main evaluation metrics. In addition, data augmentation and early stopping were applied to improve generalization and prevent overfitting. Among the tested architectures, the Inception model achieved the highest test accuracy of 97.5%, while the Quadruplet Network achieved 96.85% accuracy with a lower validation loss of 0.571, indicating stronger generalization. These results suggest that both models are well-suited for real-time emotion recognition using thermal imaging. The findings highlight the potential of combining infrared data with modern neural architectures to advance emotion detection systems beyond traditional RGB-based methods.

Received: 02.12.2024 Accepted: 25.06.2025 Published: 30.06.2025

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

Keywords: neural network, convolution neural network, thermal imager, emotion recognition, inception, U-net, Quadruplet Network, Squeeze net.

Introduction

Technological development makes it possible to introduce more powerful and accurate computing devices. This process leads to the development of calculation methods and an increase in the range of technical capabilities. One of the new possibilities is classification, forecasting of numerical parameters based on input data, voice assistants based on language models and robots based on artificial intelligence. This has become possible due to the development of computing technologies and mathematical methods. One of these methods is neural networks [1]. Today, there are many different types of neural networks, the combination of which creates new architectures, which improves the learning process, accuracy and increases the range of tasks performed. There are such types of neural networks as FNN, RNN, LSTM, GRU and CNN. If we classify each type of neural network by the type of data processed, then FNN [2] works better with single data, neural networks with feedback (RNN, LSTM, GRU) [3] work with data arrays and time dependence, and CNN [4] with images. The development of methods and techniques of data processing is a product of the ordered and chaotic dissemination of information.

It is natural for a person to show their emotions with facial expressions, but if they hide them, their condition can be judged by their behavior, temperature, and pulse. When experiencing various emotions, a person's face changes its temperature. The greater the intensity of the emotions experienced, the more visible the temperature difference in different areas of the face is as a result of the work of the autonomic nervous system. As scientific works confirm, emotions have different natures: negative, positive, and neutral [5]. And the main ones are joy, surprise, sadness, disgust, fear, and anger [6]. The ability to accurately recognize human emotions will improve the human-machine interface and increase the quality of people's lives. To obtain this ability, it is necessary to implement an emotion recognition system using thermograms. CNN is best suited for this task.

The VGG architecture is characterized by its simplicity and depth, relying on small 3×3 convolutional filters to build deeper models with improved accuracy [7]. GoogleNet introduced the inception module, which combines multiple convolutional kernels (1×1, 3×3, 5×5) to better extract both local and global features [8]. ResNet advanced the field with residual connections, effectively mitigating the vanishing gradient problem and enabling the training of extremely deep networks [9]. DenseNet further improved gradient flow by establishing dense connections between layers, making the model more compact and promoting feature reuse [10].

Subsequent architectures focused on efficiency and scalability. EfficientNet introduced compound scaling of network depth, width, and resolution to balance performance and cost [11]. BiT applied transfer learning using large pre-trained models for strong domain adaptation [12]. The Vision Transformer (ViT) brought attention-based mechanisms from NLP to image processing by treating image patches as tokens [13].

Later developments such as Meta Pseudo-Labels [14], Swin Transformer [15], EfficientNetV2 [16], ConvNeXt [17], and Segment Anything Model (SAM) [18] expanded the field through semi-supervised learning, hierarchical attention structures, accelerated training, and universal segmentation capabilities.

In the context of thermal image classification, which involves low-resolution, low-texture, grayscale inputs, selecting an appropriate neural network architecture is critical. Inception networks, for example, are effective due to their multi-scale feature extraction capabilities. The Quadruplet Network, based on metric learning, improves inter-class separability by learn-

DOI: 10.37943/22GEBT9085

ing compact and discriminative embeddings – valuable when emotional states differ only subtly in thermal data. Other modern architectures like EfficientNet and ConvNeXt offer efficient training and generalization performance, which is especially important for small and specialized datasets such as thermograms.

The novelty of this research lies in the systematic evaluation of multiple state-of-the-art CNN architectures on a custom thermal image dataset specifically designed for facial emotion recognition. To our knowledge, this is one of the first studies to benchmark deep architectures such as Inception, EfficientNet, ConvNeXt, and the Quadruplet Network on thermal data for the task of classifying six basic human emotions.

When designing the architecture of a convolutional neural network, researchers can adjust the number of hidden layers, the type of convolution operations, activation functions, and loss functions. It is also possible to enhance the model through normalization techniques or advanced architectural modifications, such as using Siamese or metric-learning-based networks. By adjusting the nonlinear structure of the model, it can be adapted to the specific characteristics of thermal images.

The main objective of this work is to build a neural network to improve the classification of emotions according to one of the parameters: classification accuracy, learning speed and flexibility.

The Methodology section describes the neural network architectures used in this study, their specific configurations, and the mathematical formulations that define their operation. The Results and Discussion section presents and interprets the experimental findings obtained from training on thermal image data, including analysis of classification accuracy, learning efficiency, and model adaptability. Finally, the Conclusion section summarizes the main outcomes of the study and suggests possible directions for future research.

Methods and Materials

To develop a robust and high-quality neural network architecture for emotion recognition from thermal images, the study was structured into the following sequential stages: data collection, image preprocessing, image analysis, neural network construction, model training, and evaluation.

Thermal data were collected in a controlled indoor environment at room temperature. The participants included ten healthy volunteers aged between 18 and 19 years. Each participant was seated in front of a thermal imaging camera and a laptop that displayed a curated sequence of short video clips with audio. These audiovisual stimuli were specifically selected to elicit six basic human emotions: joy, sadness, fear, disgust, anger, and surprise.

During the viewing of these stimuli, changes in facial temperature were continuously recorded using a thermal camera. These thermal fluctuations reflect autonomic nervous system responses and are particularly observable in regions such as the forehead, cheeks, and areas surrounding the eyes. The resulting thermal data captured the dynamic physiological patterns associated with each emotional state. The characteristics of the thermal camera used for data collection are presented in Table 1.

Table 1. Device characteristics

Device name	Characteristics
Fluke TiS20+ MAX thermal imager	IR resolution: 120x90 Infrared spectral range: from 8 to 14 μm Temperature range: -20°C to 400°C Sensitivity: 60 mK Minimum focal length: 0.5 m

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

As the thermal images (TIs) were captured in the thermal imager's auto-calibration mode, the initial raw images exhibited a consistent appearance. To better visualize the subtle temperature differences on the human face under various emotional conditions, a specific temperature range was selected using the Fluke Connect software. This enhancement made it possible to highlight meaningful thermal features. After testing multiple temperature intervals, the range from 32°C to 35.5°C was found to reveal the most distinguishable features.

In the next stage, facial localization was achieved through temperature thresholding, isolating high-temperature areas associated with facial skin. The largest connected region within the selected temperature range was designated as the region of interest (ROI). A bounding box was drawn around the detected facial area, which was then cropped and resized to 48 × 48 pixels to balance computational efficiency with the retention of critical thermal details. The steps for processing thermographic images are shown in Figure 1.

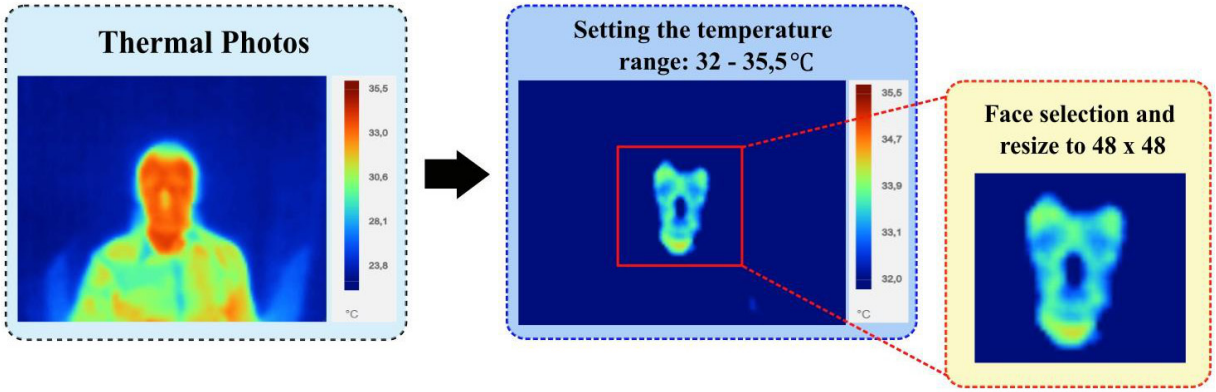

Figure 1. Representative thermal images of a person's face for each of the six basic emotions.

The final dataset comprised 1,642 thermal images, including 821 original thermograms and 821 augmented samples generated through standard image augmentation techniques. The methods used to increase the amount of data are presented in Table 2.

Table 2. Applied methods to increase the amount of data

Augmentation type	Transformation	Parameter range
Geometric	Rotation	± 10°
	Horizontal flip	50% probability
	Scaling	90% – 110%
	Translation	± 3 pixels (X, Y)

Each image was labeled based on the corresponding elicited emotion. The data were distributed as uniformly as possible among the six emotional classes, with 274 images assigned to each class, except for anger, which included 272 images. This near-balanced distribution ensured fairness in model training and minimized classification bias. The dataset was subsequently divided into 80% for training and 20% for testing, following standard machine learning practice.

Figure 2 presents representative examples of thermal facial images for each of the six basic emotions. The discernible variations in temperature distribution serve as the foundation for the neural network's ability to distinguish between emotional states based on thermographic input.

DOI: 10.37943/22GEBT9085

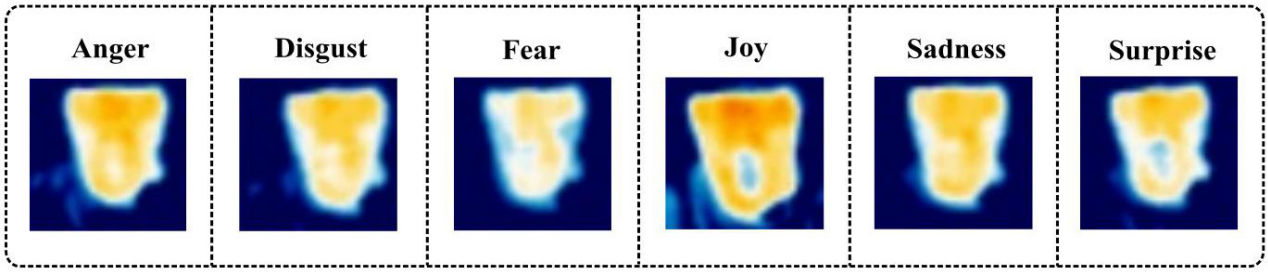

Figure 2. Representative thermal images of a person's face for each of the six basic emotions.

Although thermal datasets differ significantly from large-scale RGB datasets, their inherent properties—such as reduced noise, lower texture complexity, and consistent spatial patterns—make them well-suited for convolutional analysis. In this study, data augmentation techniques such as rotation, scaling, and horizontal flipping were employed to increase variability and improve the model's generalization. The selected architectures, including Inception and Quadruplet networks, were chosen for their proven performance in constrained-data scenarios, allowing efficient training while capturing essential thermal features.

These design choices enabled the model to effectively learn meaningful emotional patterns from thermographic input and demonstrate reliable classification performance without requiring massive datasets. This highlights the feasibility of emotion recognition from thermal images when supported by targeted architecture design and principled data handling.

CNN Architecture

Convolutional neural networks (CNNs) [4] are a powerful image processing tool that uses convolution, pooling, and fully connected layers to extract features. The basic idea is to apply a convolution operation to an input image using filters. The convolution is calculated as the sum of the products of the image pixel values and the corresponding filter weights:

$$y_{i,j} = \sum_{m=0}^{u-1} \sum_{n=0}^{u-1} \left(x_{i+m,j+n} * w_{m,n} \right) + b_i \quad (1)$$

where x is the input image, w is a filter of size $u \times u$ (commonly u=3 or u=5), b is the bias.

To prevent the image from changing its size when convolution is applied, padding is often used, which adds extra zero pixels to the edges of the image. The convolution process also uses a stride s, which determines how many positions the filter moves across the image.After the convolution operation, the results are passed through a nonlinear activation function such as ReLU (Rectified Linear Unit), which has the form:

$$ReLU(x) = max(0, x) \quad (2)$$

This allows the model to consider only positive values, speeding up training and improving feature representation. Next, a pooling operation such as MaxPooling is applied, which reduces the dimensions of images, ensuring invariance to small changes in the position of objects:

$$y_{i,j} = max(x_{u*i,u*j}, x_{u*i+1,u*j}, x_{u*i,u*j+1}, x_{u*i+1,u*j+1}, \ldots, x_{u*i+u-1,u*j+u-1}) \quad (3)$$

After several convolutional and pooling layers, the image is converted into a one-dimensional vector, which is then fed to fully connected layers. For each neuron in these layers, a weighted sum of the input values is calculated:

$$y = W * x + b \quad (4)$$

where W is the weight matrix, x is the input vector, and b is the bias.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

At the output of the network, the Softmax activation function is applied for multi-class classification, which transforms logits into probabilities:

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

where z_i are the logits for each class, and the output of the Softmax function is the probability of each class.

To train the network, a loss function is used, such as cross-entropy for classification, which measures the discrepancy between the true labels and the predicted probabilities:

$$L = -\sum_{i=1}^{N} y_i * log(p_i) \quad (6)$$

where y_i is the true class label, and p_i is the predicted probability for class i.

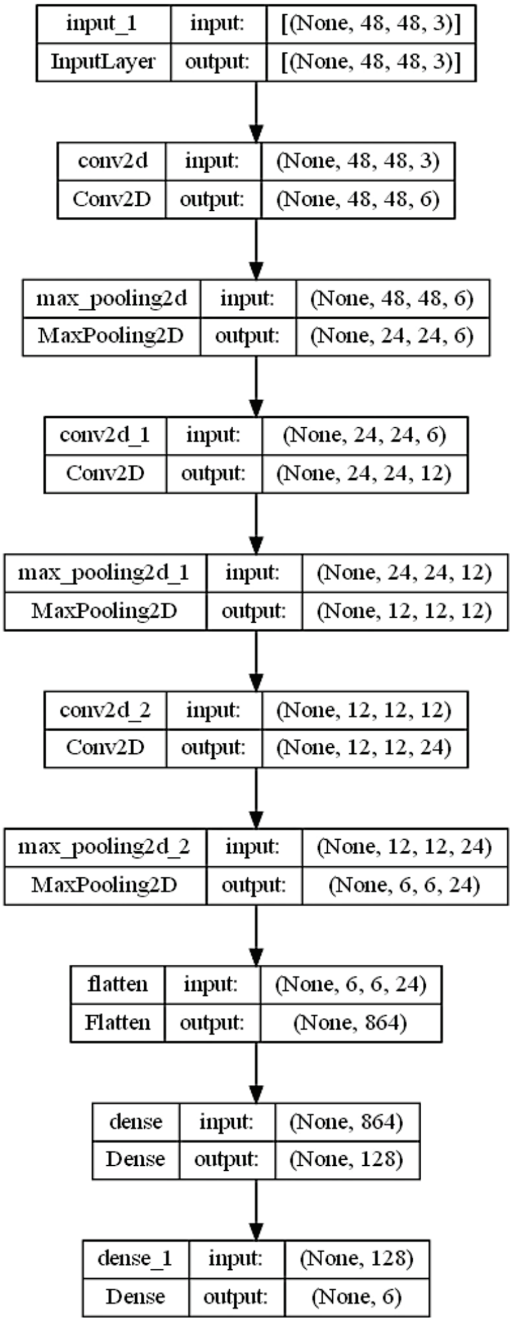

Figure 3. CNN architecture.

DOI: 10.37943/22GEBT9085

Thus, CNN efficiently extracts features, learns, and classifies images by minimizing the loss function using optimization techniques such as gradient descent. An example of CNN architecture is shown in Fig. 3.

Siamese CNN and Quadruplet Network Architecture

Siamese convolutional neural network (Siamese CNN) [19] is a neural network architecture that is used for the task of comparing two input images. These networks consist of two identical subnetworks that process two input images and compare their representations. This approach is used in tasks where it is important to determine the similarity degrees between objects, such as in face recognition or identity verification.

The basic idea is that both input images are passed through the same network architecture, resulting in similar or identical representations if the images are similar, or very different if the images are different.

The training of Siamese CNN consists of minimizing the contrast loss, which leads to an improvement in the network's ability to distinguish similarities and differences between pairs of images. Siamese CNN can be used to solve problems in the field of face recognition, signature verification, pattern matching, and other areas where it is necessary to compare objects by similarity.

Quadruplet Network is an extension of the Siamese Network concept designed for tasks related to image classification, verification, and ranking [20]. Unlike simpler architectures such as Siamese Network or Triplet Network, which work on pairs or triplets of images, Quadruplet Network processes four images in a single pass, which improves the network's ability to discriminate and generalize features.

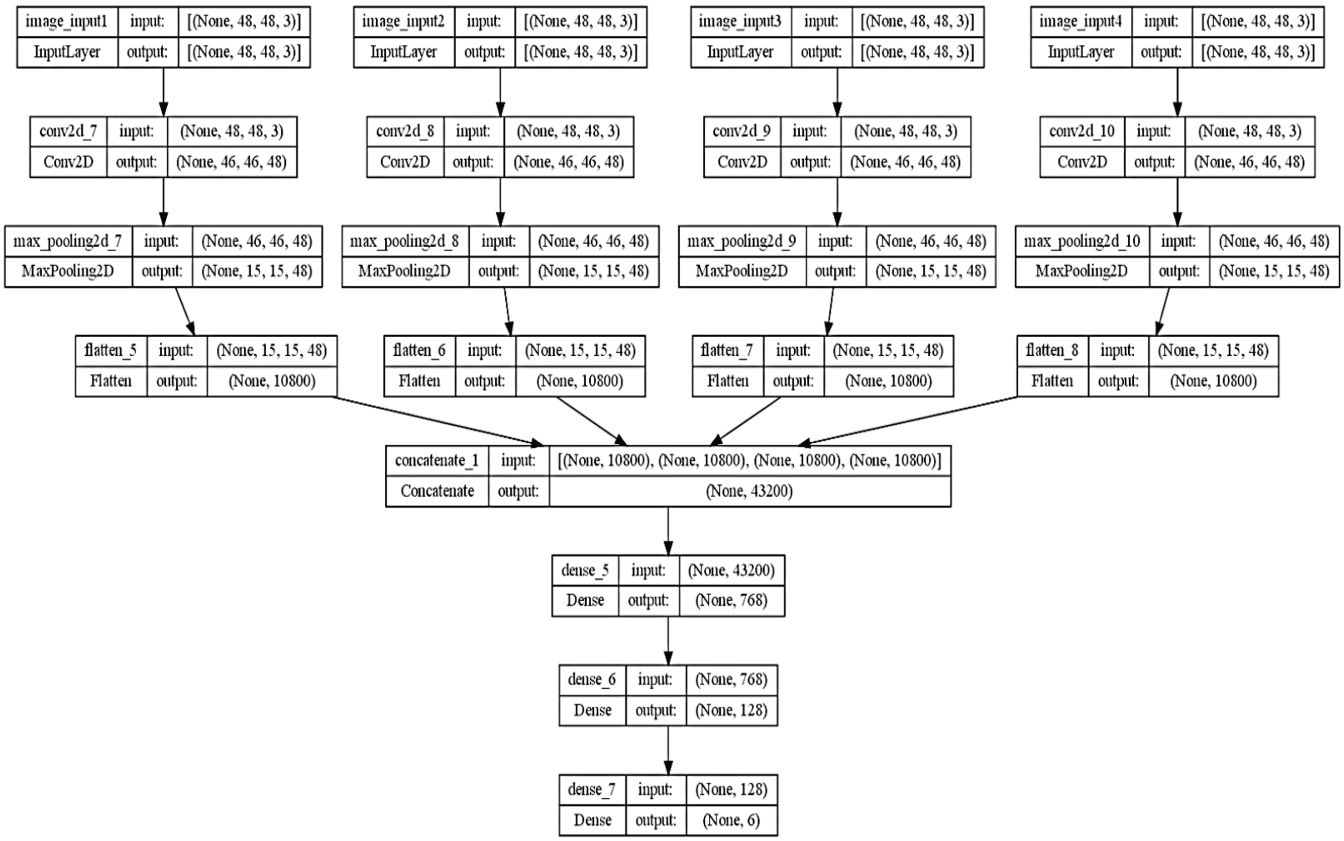

Figure 4. Quadruplet Network architecture.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

In this work, identical images were fed to the input of the Quadruplet Network. This was done in order to create interacting branches that minimize or maximize key properties of thermograms by changing weight coefficients. The considered architecture of the Quadruplet Network is shown in Fig. 4.

U-Net Architecture

U-Net is a neural network architecture specifically designed for image segmentation tasks, especially in the field of medical diagnostics [21]. It is used to highlight objects in images such as cells, tumors, and other anatomical structures.

The U-Net architecture consists of two main parts: an encoder and a decoder. The encoder is a series of convolutional and pooling layers that reduce the spatial dimensions of the image and extract key features. After that, the information passes through a series of deconvolutional layers that restore the feature map dimensions to the original image dimensions.

A special feature of U-Net is the use of so-called skip connections between the corresponding layers of the encoder and decoder. These connections allow high-level information about local features to be preserved, which improves segmentation accuracy, especially for small objects and objects with fuzzy boundaries. The encoder applies convolution and pooling operations, obtaining a smaller feature map. The decoder restores resolution using deconvolution.

The loss function for U-Net typically includes a cross-entropy or Dice coefficient. Cross-entropy measures the discrepancy between the true label and the prediction for each pixel (6). The Dice coefficient measures the similarity between the predicted and true segmentation:

$$Dice = \frac{2|A \cap B|}{|A| + |B|} \tag{7}$$

where A and B are sets of pixels corresponding to objects in the predicted and true segmentation.

The training process of U-Net is to minimize the loss function, which allows the network to segment images effectively. U-Net is widely used in medical segmentation for image analysis, such as MRI, CT, and in tasks where high accuracy of object detection in images is required. In this paper, U-Net is used for classification. An example of U-Net architecture is shown in Fig. 5.

DOI: 10.37943/22GEBT9085

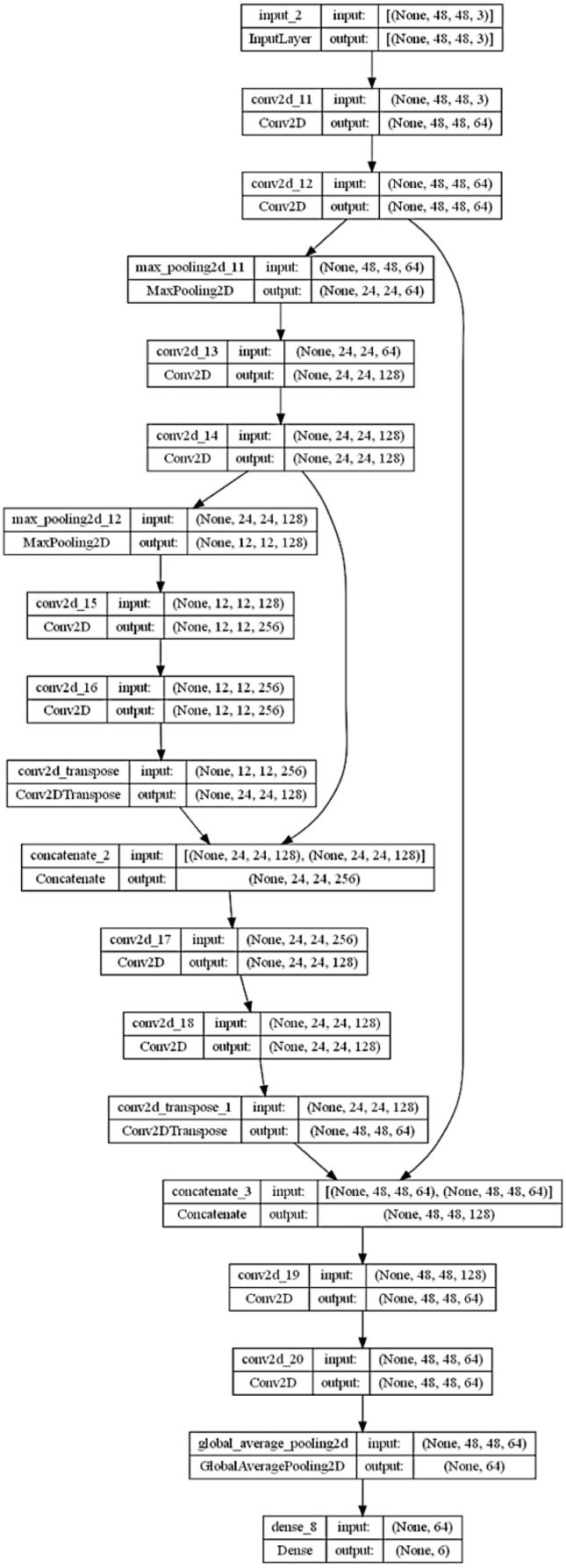

Figure 5. U-Net architecture.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

Inception Architecture

Inception is the result of years of research into improving the performance of convolutional neural networks for computer vision tasks such as image classification [22].

The basic idea behind Inception is to use different types of convolutions with different filter sizes in a single layer of the network. Instead of using only one type of filter (e.g. only 3x3 or 5x5), Inception uses multiple parallel convolution operations with filters of different sizes: 1x1, 3x3, and 5x5, as well as pooling operations with max pooling. The results are then concatenated, allowing the model to extract more diverse features at different scales in a single layer.

The Inception architecture has been improved over time, reaching its greatest success in Inception-v3. A single layer uses 1x1, 3x3, and 5x5 filters, and a 3x3 pooling operation with a stride of 2. This allows the network to extract features of different scales, improving its ability to handle complex images. Using 1x1 filters helps reduce the feature dimensionality and reduce computational costs. This also improves the quality of the network without adding significant computational costs. All outputs from different convolution operations are concatenated, allowing the model to leverage information from different filters and improve performance.

Operating on an input image x of size H×W, the output of a layer can be expressed as:

$$y = concat(conv_{1x1}(x), conv_{3x3}(x),\ conv_{5x5}(x),\ pool_{3x3}(x)) \tag{8}$$

where $conv_{1x1}(x), conv_{3x3}(x), conv_{5x5}(x), pool_{3x3}(x)$ are various convolution and subsampling operations with filters of different sizes, and $consat$ is the operation of combining all the results into one tensor.

Inception has significantly improved accuracy over traditional architectures such as AlexNet and VGG by using computational resources more efficiently. The model has become widely used for image classification, object recognition, and other computer vision applications.

One of the main advantages of Inception is the ability to extract different features from different levels of abstraction, which allows models to achieve high results with lower computational costs. However, the Inception architecture also has its limitations, such as complexity in implementation and the need for a large amount of training data. An example of the Inception architecture is shown in Fig. 6.

DOI: 10.37943/22GEBT9085

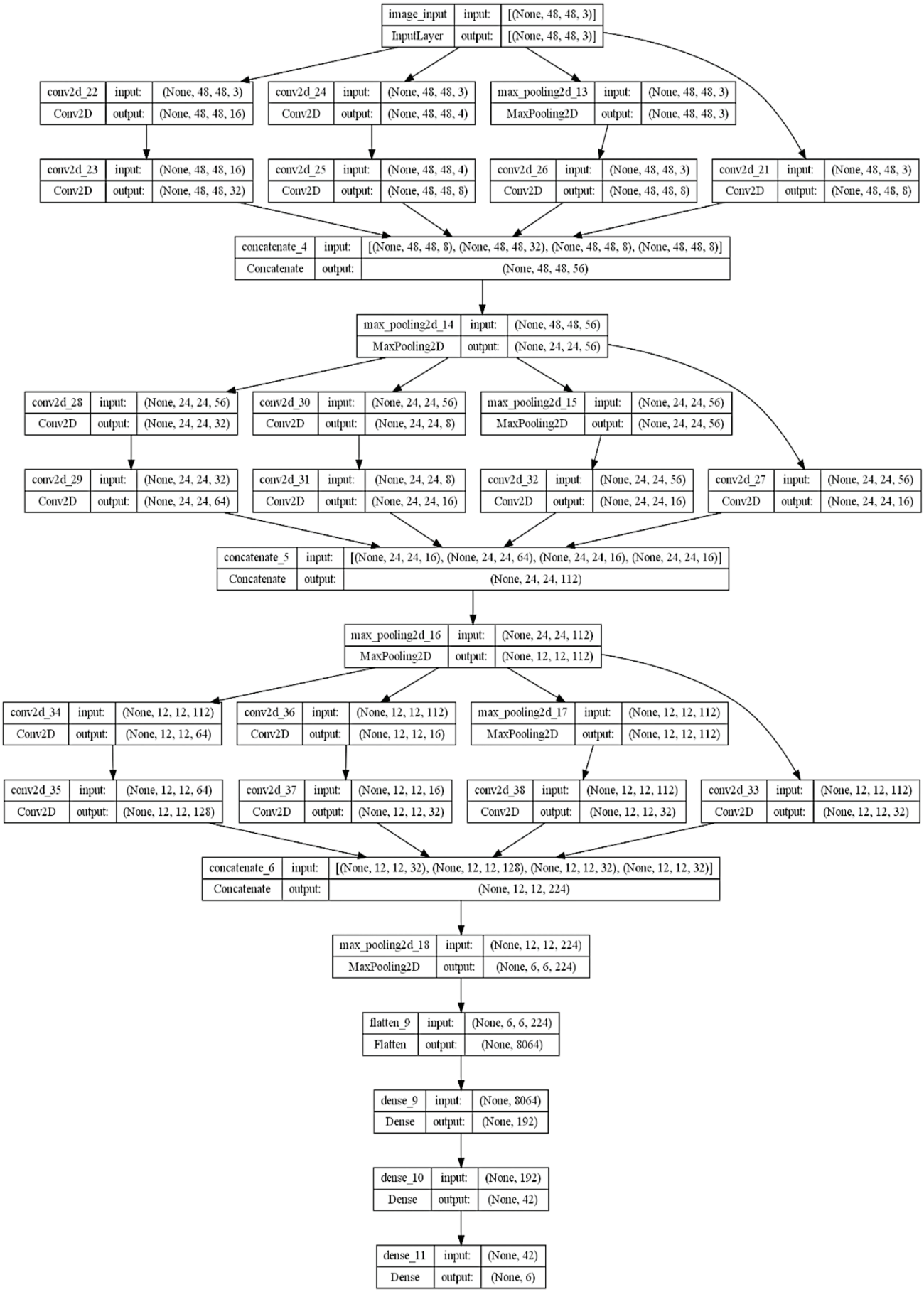

Figure 6. Inception architecture

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

SqueezyNet Architecture

SqueezeNet is a neural network architecture designed to reduce the number of parameters and computational cost while maintaining high accuracy [23]. One of the main differences of SqueezeNet is the use of a "squeeze" strategy, which significantly reduces the number of parameters, making the model more compact and suitable for use on devices with limited computing resources.

The basic idea behind SqueezeNet is to use so-called squeeze layers. These layers use 1x1 convolutions that reduce the dimensionality of the data before feeding it into more complex layers with larger filters, such as 3x3 convolutions. This allows the number of parameters in the network to be significantly reduced without losing the ability to extract important features from the data. Unlike standard models that use larger filters, SqueezeNet uses 1x1 convolutions as "squeeze" operations. This allows the number of channels and parameters to be significantly reduced, minimizing computational costs while still retaining the ability to extract information. SqueezeNet consists of so-called fire modules, which include two types of layers: squeeze layers (with 1x1 filters) and expander layers (with 3x3 filters). These modules allow for efficient information processing while keeping the number of parameters small. In each fire module, the data is first compressed using 1x1 convolutions, and then the data is passed through layers with 3x3 filters. This allows combining the efficiency of 1x1 convolutions with the power of larger filters for feature extraction.

For an input image x of size HxW, at the output of the fire module:

$$FireModule(conv_{1x1}(x), conv_{3x3}(conv_{1x1}(x))) \quad (9)$$

where $conv_{1x1}(x)$ is a 1x1 convolution operation for compression, and $conv_{3x3}(x)$ is a 3x3 convolution operation for expansion.

The advantage of this scheme is that the number of parameters in the network is significantly reduced, since 1x1 convolutions require less computation than standard convolutions of large dimensions, which makes the model suitable for use on mobile devices and in resource-constrained environments. By using 1x1 convolutions and fire modules, SqueezeNet significantly reduces the number of parameters (up to 50-60 times compared to traditional networks such as AlexNet), while maintaining high accuracy. Despite its smaller size, SqueezeNet demonstrates good performance in image classification tasks and is also suitable for running on devices with limited computing resources.

DOI: 10.37943/22GEBT9085

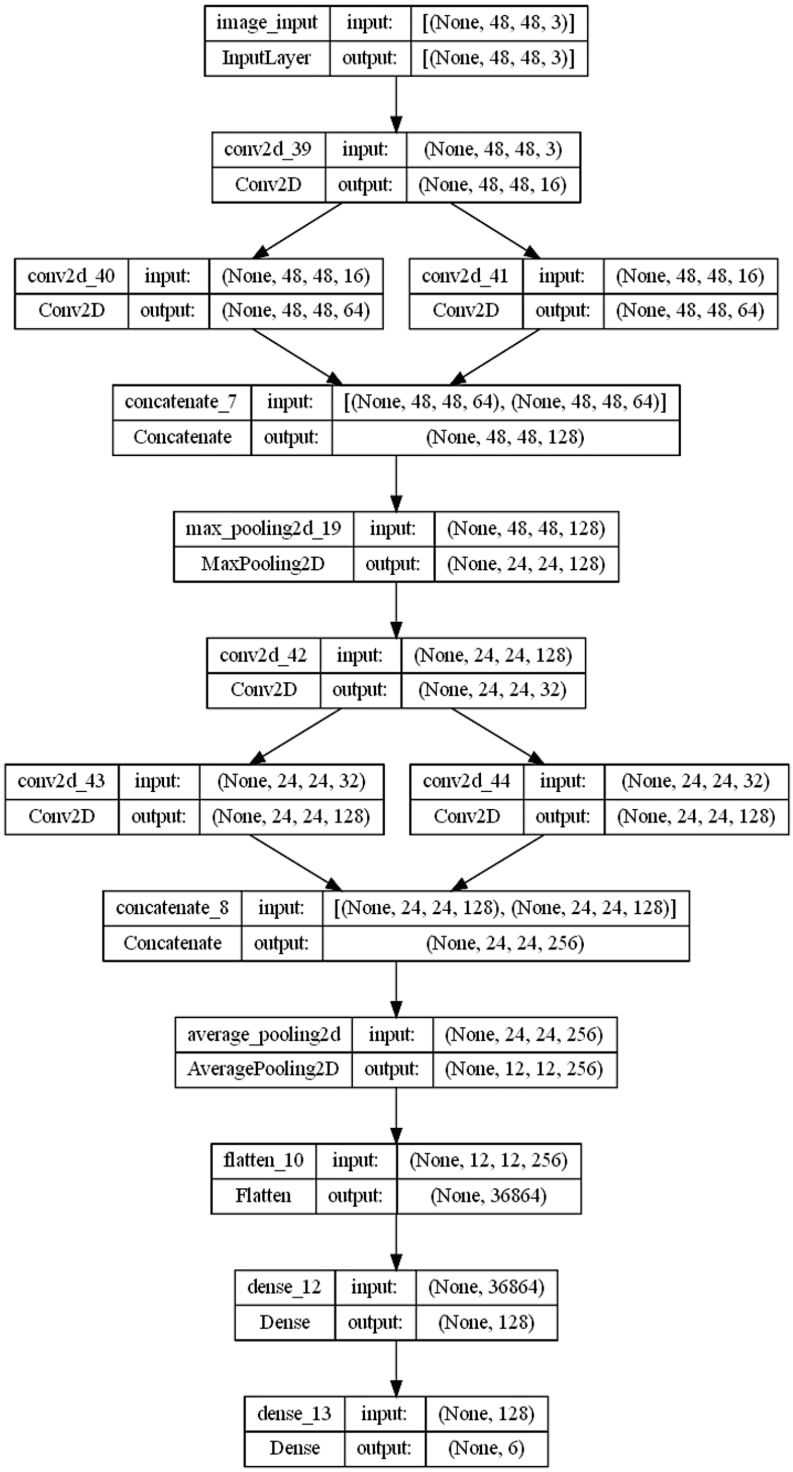

Figure 7. SquezyNet architecture.

SqueezeNet is ideal for applications where model compactness is important, such as mobile devices, robotics, and other systems with limited computing power. An example of the SqueezeNet architecture is shown in Fig. 7.

Results and Discussion

Key performance metrics for evaluating neural network models include classification accuracy and loss. To prevent overfitting during training, the EarlyStopping strategy was imple-

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

mented with a patience parameter of 10 epochs, terminating training when validation loss ceased to improve. Table 3 summarizes the training outcomes for each tested architecture, including the total number of epochs, training and validation losses, training and validation accuracies, and the final test accuracy achieved on the unseen data.

Table 3. Training and validation metrics for different neural network architectures

Name of architecture	CNN	Quadruplet Network	U-Net	Inception	SquezyNet
Number of epochs	133	200	200	188	178
Train loss	0.3574	0.0615	0.2488	0.0235	0.0615
Validation loss	0.6869	0.571	0.7957	0.6565	0.8029
Train accuracy	0.8653	0.9908	0.9197	0.9945	0.9815
Validation accuracy	0.7758	0.8791	0.823	0.8968	0.8496
Test accuracy	0.8474	0.9685	0.9004	0.975	0.9551

Figures 8–12 illustrate the training and validation dynamics of each model in terms of loss and accuracy progression over epochs.

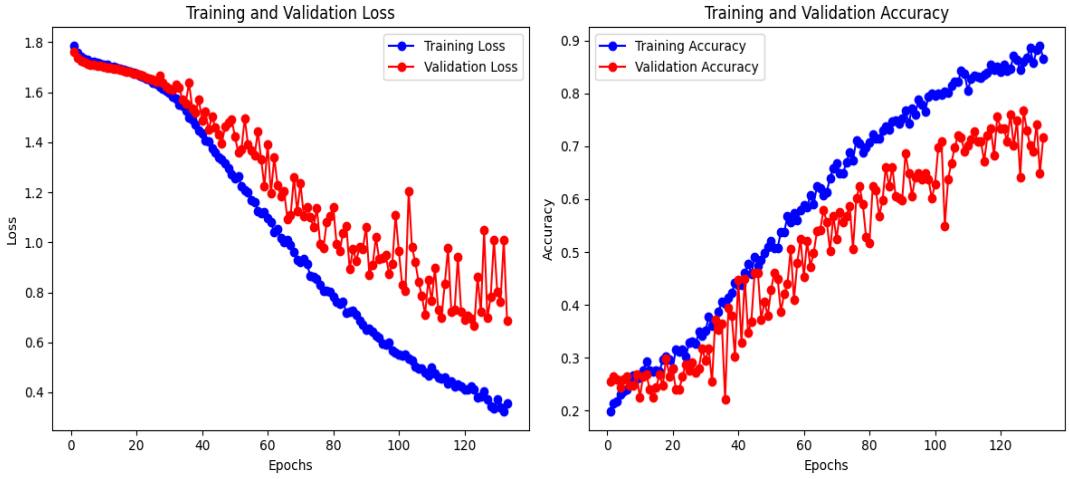

Figure 8. CNN

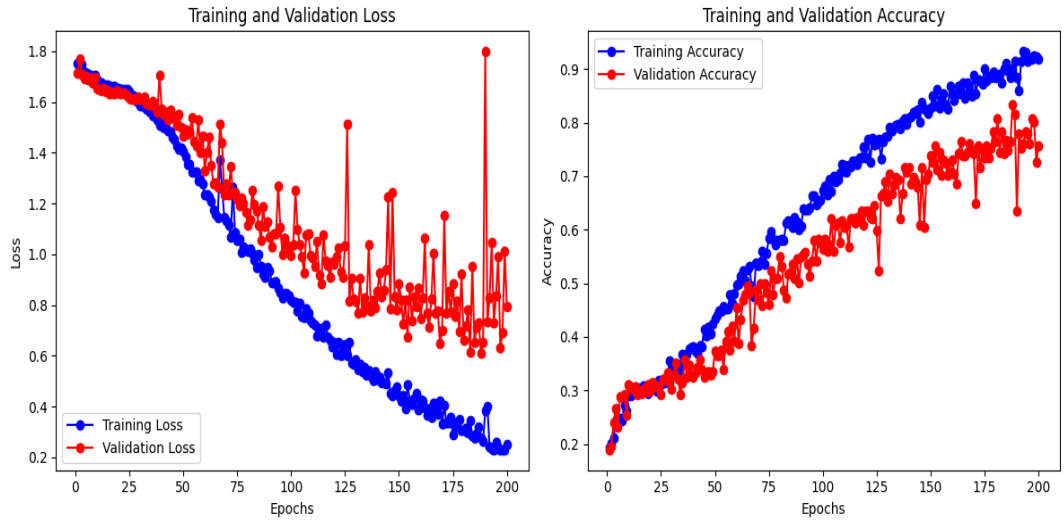

Figure 9. Quadruplet Network

DOI: 10.37943/22GEBT9085

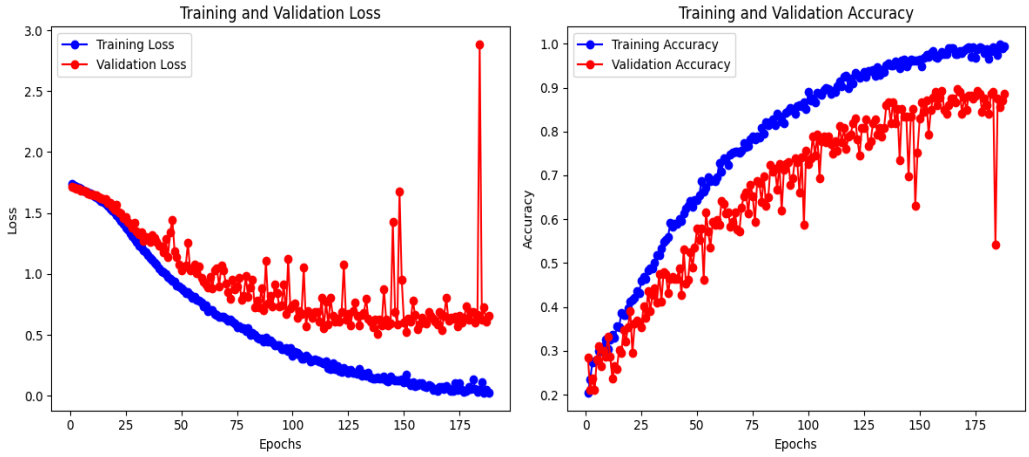

Figure 10. U-Net.

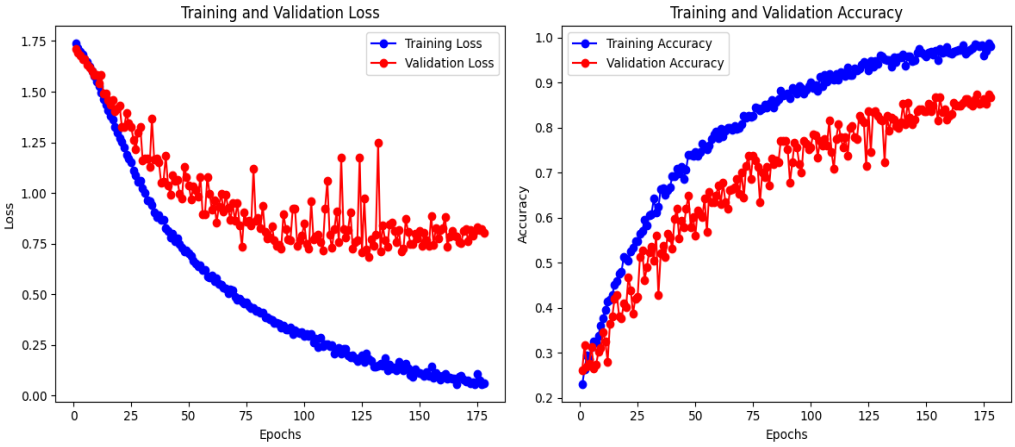

Figure 11. Inception.

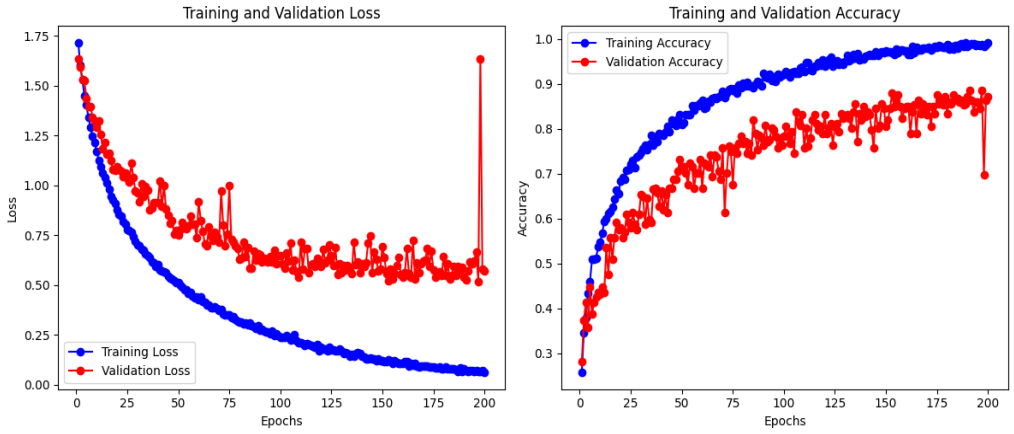

Figure 12. SquezyNet.

Most models exhibit a consistent upward trend in validation accuracy during training. However, fluctuations in validation loss suggest potential overfitting in deeper or more complex architectures, particularly under the constraints of a relatively small dataset.

Despite its limited size, the dataset was collected under controlled laboratory conditions using a Fluke thermal imager, ensuring consistency in resolution (48×48 pixels), environmen-

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

tal factors, and sensor calibration. To enhance model generalization, geometric data augmentation techniques were applied, including rotation, horizontal flipping, scaling, and translation. These transformations were carefully selected to preserve essential thermal gradients, which are critical for emotion classification.

Among the five evaluated architectures, Inception achieved the highest test accuracy of 97.5%, supported by a training accuracy of 99.45% and validation accuracy of 89.68%. In comparison, the Quadruplet Network reached a slightly lower test accuracy of 96.85%, but demonstrated more stable validation behavior, with a validation accuracy of 87.91% and a lower validation loss of 0.571.

These results highlight a trade-off between representational capacity and generalization: Inception is more effective at extracting high-level features, whereas Quadruplet Network maintains stronger validation performance under similar training conditions. Both models demonstrate strong potential for emotion recognition from thermal images. U-Net, originally developed for segmentation tasks, showed lower accuracy, possibly due to its emphasis on pixel-level reconstruction. SqueezeNet, while computationally efficient, exhibited limited performance when applied to low-resolution, textureless thermal data.

In this study, the primary evaluation was based on overall accuracy metrics. While this provided insight into model performance across architectures, further analysis may include class-wise metrics such as precision, recall, and F1-score, as well as confusion matrices, to better understand model behavior with respect to individual emotional categories.

The findings suggest that, with careful preprocessing and model selection, deep learning methods can be successfully applied to thermal image-based emotion recognition. These approaches may contribute to the development of non-contact, privacy-aware systems for affective computing, healthcare monitoring, smart environments, and human–machine interaction.

Conclusion

Each neural network architecture exhibits unique characteristics depending on the degree of data nonlinearity, the volume of training data, and the type of input. In this study, nonlinearity was defined as the variability of thermal facial features under different emotional states. Due to the limited dataset size (821 thermograms), geometric data augmentation techniques such as rotation, flipping, scaling, and translation were applied to increase data diversity. The models were trained to classify six basic emotions from low-resolution thermal images. Among the evaluated architectures, Inception and Quadruplet Network achieved the highest classification performance, indicating their suitability for extracting discriminative spatial features in thermographic data. CNN, U-Net, and SqueezeNet demonstrated lower but acceptable accuracy given the constraints of the data.

The results confirm the applicability of deep learning methods for thermal emotion classification. While the evaluation in this study was based on quantitative metrics, future work may include qualitative analysis approaches, such as visualization of feature importance and attention mechanisms (e.g., class activation maps or attention heatmaps), to explore how neural networks focus on informative thermal regions during classification. This could provide further insight into the decision-making process of the models and contribute to improved model transparency. The findings suggest that with appropriate data preprocessing and model selection, deep learning approaches can be applied to thermal imaging for contactless emotion recognition in applications such as healthcare, smart environments, and affective computing.

DOI: 10.37943/22GEBT9085

References

[1] Abiodun, O.I., Jantan, A., Omolara, A.E., Dada, K.V., Mohamed, N.A., & Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. Heliyon, 4(11). https://doi.org/10.1016/j.heliyon.2018.e00938

[2] Cao, W., Wang, X., Ming, Z., & Gao, J. (2018). A review on neural networks with random weights. *Neurocomputing*, *275*, 278-287. https://doi.org/10.1016/j.neucom.2017.08.040

[3] Tsantekidis, A., Passalis, N., & Tefas, A. (2022). Recurrent neural networks. In *Deep learning for robot perception and cognition* (pp. 101-115). Academic Press. https://doi.org/10.1016/B978-0-32-385787-1.00010-5

[4] Ketkar, N., Moolayil, J., Ketkar, N., & Moolayil, J. (2020). Deep Learning with Python: Learn Best Practices of Deep Learning Models with PyTorch. https://doi.org/10.1007/978-1-4842-5364-9

[5] Nematollahi, J., & Firoozabadi, M. (2017, November). Recognition of Positive, Negative and Neutral Emotions Using Brain Connectivity Patterns. In 2017 24th National and 2nd International Iranian Conference on Biomedical Engineering (ICBME) (pp. 330-333). IEEE. https://doi.org/10.1109/ICBME.2017.8430281

[6] Wang, Z., Ho, S. B., & Cambria, E. (2020). A review of emotion sensing: categorization models and algorithms. Multimedia Tools and Applications, 79, 35553-35582. https://doi.org/10.1007/s11042-019-08328-z

[7] Yaseliani, M., Hamadani, A.Z., Maghsoodi, A.I., & Mosavi, A. (2022). Pneumonia detection proposing a hybrid deep convolutional neural network based on two parallel visual geometry group architectures and machine learning classifiers. IEEE access, 10, 62110-62128. https://doi.org/10.1109/ACCESS.2022.3182498

[8] Anand, R., Shanthi, T., Nithish, M. S., & Lakshman, S. (2020). Face recognition and classification using GoogleNET architecture. In Soft Computing for Problem Solving: SocProS 2018, Volume 1 (pp. 261-269). Springer Singapore. https://doi.org/10.1007/978-981-15-0035-0_20

[9] Peng, S., Huang, H., Chen, W., Zhang, L., & Fang, W. (2020). More trainable inception-ResNet for face recognition. Neurocomputing, 411, 9-19. https://doi.org/10.1016/j.neucom.2020.05.022

[10] Li, B. (2022). Facial expression recognition by DenseNet-121. In Multi-Chaos, Fractal and Multi-Fractional Artificial Intelligence of Different Complex Systems (pp. 263-276). Academic Press. https://doi.org/10.1016/B978-0-323-90032-4.00019-5

[11] Li, R. (2023, June). Face detection and recognition technology based on EfficientNet and BNNeck. In International Conference on Mathematics, Modeling, and Computer Science (MMCS2022) (Vol. 12625, pp. 485-490). SPIE. https://doi.org/10.1117/12.2670429

[12] Harakannanavar, S.S., Prashanth, C.R., Raja, K.B., & Madiwalar, C.T. (2018, May). Face Recognition based on the fusion of Bit-Plane and Binary Image Compression Techniques. In 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (pp. 1889-1894). IEEE. https://doi.org/10.1109/RTEICT42901.2018.9012230

[13] Xue, F., Wang, Q., Tan, Z., Ma, Z., & Guo, G. (2022). Vision transformer with attentive pooling for robust facial expression recognition. IEEE Transactions on Affective Computing, 14(4), 3244-3256. https://doi.org/10.1109/TAFFC.2022.3226473

[14] Pham, H., Dai, Z., Xie, Q., & Le, Q. V. (2021). Meta pseudo labels. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11557-11568). https://openaccess.thecvf.com/content/CVPR2021/papers/Pham_Meta_Pseudo_Labels_CVPR_2021_paper.pdf

[15] Qin, L., Wang, M., Deng, C., Wang, K., Chen, X., Hu, J., & Deng, W. (2023). SwinFace: a multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation. IEEE Transactions on Circuits and Systems for Video Technology. https://doi.org/10.1109/TCSVT.2023.3304724

[16] Grd, P., Tomičić, I., & Barčić, E. (2024). Transfer Learning with EfficientNetV2S for Automatic Face Shape Classification. Journal of Universal Computer Science (JUCS), 30(2). https://doi.org/10.3897/jucs.104490

[17] Hoo, S.C., Ibrahim, H., & Suandi, S.A. (2022). ConvFaceNeXt: Lightweight networks for face recognition. Mathematics, 10(19), 3592. https://doi.org/10.3390/math10193592

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

[18] Osco, L.P., Wu, Q., De Lemos, E.L., Gonçalves, W.N., Ramos, A.P.M., Li, J., & Junior, J.M. (2023). The segment anything model (sam) for remote sensing applications: From zero to one shot. International Journal of Applied Earth Observation and Geoinformation, 124, 103540. https://doi.org/10.1016/j.jag.2023.103540

[19] Kumar, C.R., Saranya, N., Priyadharshini, M., & Gilchrist, D. (2023). Face recognition using CNN and siamese network. Measurement: Sensors, 27, 100800. https://doi.org/10.1016/j.measen.2023.100800

[20] Ren, G., Lu, X., & Li, Y. (2021). Joint face retrieval system based on a new quadruplet network in videos of multi-camera. IEEE Access, 9, 56709-56725. https://doi.org/10.1109/ACCESS.2021.3072055

[21] Chatterjee, S., & Chu, W. T. (2019, December). Thermal face recognition based on transformation by residual U-net and pixel shuffle upsampling. In International Conference on Multimedia Modeling (pp. 679-689). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-37731-1_55

[22] Kwon, D.H., & Yu, J.M. (2024). Real-time Multi-CNN-based Emotion Recognition System for Evaluating Museum Visitors' Satisfaction. *ACM Journal on Computing and Cultural Heritage*, *17*(1), 1-18. https://doi.org/10.1145/363112

[23] Sangamesh, H., Viswanatha, V.M., Petli, V., & Patil, N.B. (2023, February). A Novel Approach for Recognition of Face by Using Squeezenet Pre-Trained Network. In 2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS) (pp. 1-5). IEEE. https://doi.org/10.1109/ICICACS57338.2023.10100097