DOI: https://doi.org/10.15276/aait.09.2026.01
UDC 681.3.07: 004.8

Improving the quality of mine detection by a mine-detecting drone by selecting an architecture for classifying objects in visible-light camera images

Oleg M. Galchonkov[1)]
ORCID: https://orcid.org/0000-0001-5468-7299; o.n.galchenkov@gmail.com. Scopus Author ID: 56081377900
Alexey M. Baranov[2)]
ORCID: https://orcid.org/0009-0002-5951-2636; oleksii.m.baranov@gmail.com
Illia O. Baskov[1)]
ORCID: https://orcid.org/0000-0002-3517-6773, illyabaskov@gmail.com
[1)] Odesa Polytechnic National University, 1, Shevchenko Ave, Odesa, 65044, Ukraine
[2)] Oracle Corporation, Oracle World Headquarters, Oracle Way 2300, Austin, USA TX 78741

ABSTRACT

The use of unmanned aerial vehicles (drones) is increasingly being used in humanitarian demining. Multiple sensors operating on different physical principles are used simultaneously to search for mines. This increases the probability of mine detection. To expedite the survey of the area, the initial processing of sensor signals is performed onboard the drone. Therefore, the requirement for sensor signal processing algorithms not only increases the probability of object detection but also improves the ratio of classification accuracy to the required computational effort. One of the sensors used is visible-light cameras. A large number of neural networks have been developed for classifying objects in images. However, a specific feature of their use for humanitarian demining is the lack of large datasets for training and testing. Therefore, the challenge arises of finding neural networks that offer a high ratio of classification accuracy to the required computational effort, while also being able to train on very small data sets. This paper compares convolutional neural networks and networks based on transformers. The networks were trained on a small dataset containing images of anti-tank mines, rocks, and various backgrounds. The study showed that convolutional neural networks train faster and are more resilient to image quality degradation. However, their potential is limited; increasing the number of layers does not significantly improve mine classification quality. Transformer-based neural networks offer greater flexibility and a wealth of options for architectural configuration, resulting in superior performance compared to convolutional networks. Although they are slower to train and potentially require an expanded training dataset, they are the preferred choice for implementing image processing on drones.

Keywords: Convolutional neural network; transformer; attention, weights; classification quality; tokenization

***For citation*: Galchonkov O. M., Baranov A. M., Baskov I. O. "Improving the quality of mine detection by a mine-detecting drone by selecting an architecture for classifying objects in visible-light camera images". *Applied Aspects of Information Technology*. 2026; Vol.9 No.1: 9–24. DOI: https://doi.org/10.15276/aait.09.2026.01**

1. INTRODUCTION

Autonomous unmanned aerial vehicles (quadcopters, UAVs) are increasingly used in a variety of fields to expedite surveys, conduct research in hard-to-reach places, and replace humans on dangerous missions [1]. To improve the quality of surveys, sensors of various physical nature are typically used, the signals of which are processed and integrated using neural networks. Some of the most popular sensors are optical cameras operating in the visible wavelength range. Very often, new applications do not have very large data sets that would allow for the effective training of a neural network for solving the problems of detecting and classifying objects in images. An additional difficulty in solving this problem is the limited computing power of the processor located onboard the UAV.

Therefore, the problem of choosing a neural network that ensures high classification quality with a small amount of computation, and which can be trained on a small data set, is extremely relevant for many new applications.

An important task of this type is the use of drones for humanitarian demining. Two approaches are possible. The first involves the drone flying at a high altitude along a predetermined trajectory and using algorithms such as YOLO [2] to detect objects. After an object is detected, its classification result is recorded by the YOLO algorithm itself, or its image is cropped and fed to an additional classifier. The second involves the drone flying at a low altitude and feeding camera images directly to the classifier. This approach is preferable, as it allows for the integration of camera signal processing in the visible, infrared, ultraviolet ranges with the results of signal processing from magnetic sensors, ground-penetrating radars, and other types

of sensors. This combined use of sensors of different physical nature significantly improves the accuracy of mine detection and classification. Therefore, the choice of a neural network for mine detection and classification in visible images for subsequent integration with the results of signal classification from other sensors is relevant.

2. RELATED WORKS

Neural networks used for mine detection and classification in images must ensure high classification quality after training on a small dataset with a low computational effort. By low computational effort, we mean that the neural network must be implemented on the UAV's computer and produce classification results in real time at a rate sufficient for the UAV to survey a given area of terrain in an acceptable time.

One approach to reducing computational effort is pruning [3]. This approach provides a significant reduction in computational effort while maintaining classification quality [4] or even improving it [5]. Closely related to this approach are low-rank factorization [6] and knowledge distillation [7]. A common drawback of this approach is training neural networks on relatively large datasets.

Another approach is the development of new neural networks specifically designed to meet all requirements. The most commonly used basis for this is the Visual Transformer (ViT) architecture, adapted for image processing [8]. It allows the volume of calculations to vary from a very small amount to billions of floating-point operations. The Transformer architecture was initially proposed for processing text data [9]. Unlike recurrent neural networks, the Transformer processes the entire sentence/the entire context at once. Its built-in attention mechanism focuses on relevant words. For each word, a query (Q – what is being searched for), a key (K – what the word represents), and a value (V – what information the word carries) are formed. The query of a given word is compared with the keys of other words; the greater the match, the more attention (weight) this word receives. The resulting attention of a word is defined as a weighted sum of the values of other words. When processing images, it is divided into patches, the equivalents of words in a sentence. Then the patches are transformed into flat vectors (Linear Projection), which are further used in calculations. While words are encoded by their place in a sentence/context, patches are encoded by their coordinates in the image. A special feature of Transformers is that they scale very well and require the maximum possible amount of data for training. Therefore, the effectiveness of video transformers for processing small images and small training data sets requires research for each specific dataset.

To ensure a high ratio of classification quality to computational effort, a number of special modifications of the Transformer architecture have been developed. In [10], a modification of the classical Transformer is proposed in which the regions into which the original image is divided (patches) are made overlapping. This allows for taking into account local image structures located at the patch boundaries. In [11], the flexibility in object placement in images is improved by introducing additional positional self-attention into the Transformer architecture. In [12], a multi-level cross-attention mechanism extracting low-level features is proposed. In [13], linear projections at the tokenization stage are replaced with convolutional projections. In [14], the self-attention mechanism, which has quadratic complexity, is replaced with an extrinsic attention mechanism with linear complexity. Thus, the Transformer architecture has great flexibility and allows for numerous modifications. However, the effectiveness of using certain modifications requires research and adaptation for specific target datasets. In [15], such a study was conducted for the CIFAR-10 dataset [16]. The following modifications of transformers were considered. Compact Convolutional Transformer (CCT) [17] – a convolutional neural network was added to the transformer at the input to reduce the patch size and improve the efficiency of training on small datasets. External Attention Transformer (EANet) [14] – additional blocks of a multilayer perceptron (MLP) was used to implement external attention. FNet [17] – the internal self-attention block was replaced with a simpler shuffling of tokens using a Fast Fourier transform. Swin Transformer (SwinTr) [19] – flexibility in processing objects with different scales is provided due to the calculation of self-attention not for the entire image, but only within shifted windows. In [15], it was shown that on the CIFAR-10 dataset, the CCT has the best ratio of classification quality and the amount of computations. However, for a set of images with mines, the efficiency of using CCT requires a separate study.

Unlike transformer-based architectures, which are sensitive to the size of the training dataset, simpler convolutional networks [20] are less sensitive to the size of the training data.

From the literature review, it follows that for mine detection and classification, taking into account the implementation on a computer installed

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

on a UAV, it is advisable to compare the efficiency of using convolutional networks such as CCN and ResNet-18 [20] with transformer-based networks – CCT and ViT variants with different dimensions.

3. RESEARCH AIM AND OBJECTIVES

The aim of this work is to improve the quality of mine detection by a mine-detecting drone by selecting an object classifier architecture in visible-light camera images.

To achieve this goal, the following objectives were set:

– parameter selection and comparison of CCN, ResNet-18, CCT, and ViT neural networks to ensure the best balance between classification quality and computational effort;

– analysis of the neural networks' resilience to input image quality degradation.

4. MATERIALS AND METHODS

To conduct the study, a special dataset was constructed, containing 1,826 images for training, 186 for validation, and 274 for testing. All images are color and have the dimensions (64,64,3).

All image groups are strictly balanced. Half contain images of antitank mines against various backgrounds. The other half contains only the background. The mine images were taken from the Landmines Detection Dataset (by Northumbria) [21]. In this dataset, images are stored in the YOLOv11 format with the coordinates of the object's center and its dimensions in the image, measuring 640 x 640 pixels. To create the dataset for the study, a 128 x 128 pixel region containing the mine image was cropped from the image, reduced to 64 x 64 pixels, and converted to the (64,64,3) format.

Examples of such images are shown in Fig. 1.

Background images were formed in a similar way from the same images, only the area that did not contain the mine was cut out (Fig. 2).

To increase the complexity of the classification task, 120 training background images, 24 validation background images, and 20 test background images were replaced with a background containing rocks that could be mistaken for mines [22]. Examples of such images are shown in Fig. 3. All images are labeled: 1 = contains a mine; 0 = does not contain a mine.

Fig. 1. **Examples of images with mines**
Source: **compiled by the authors**

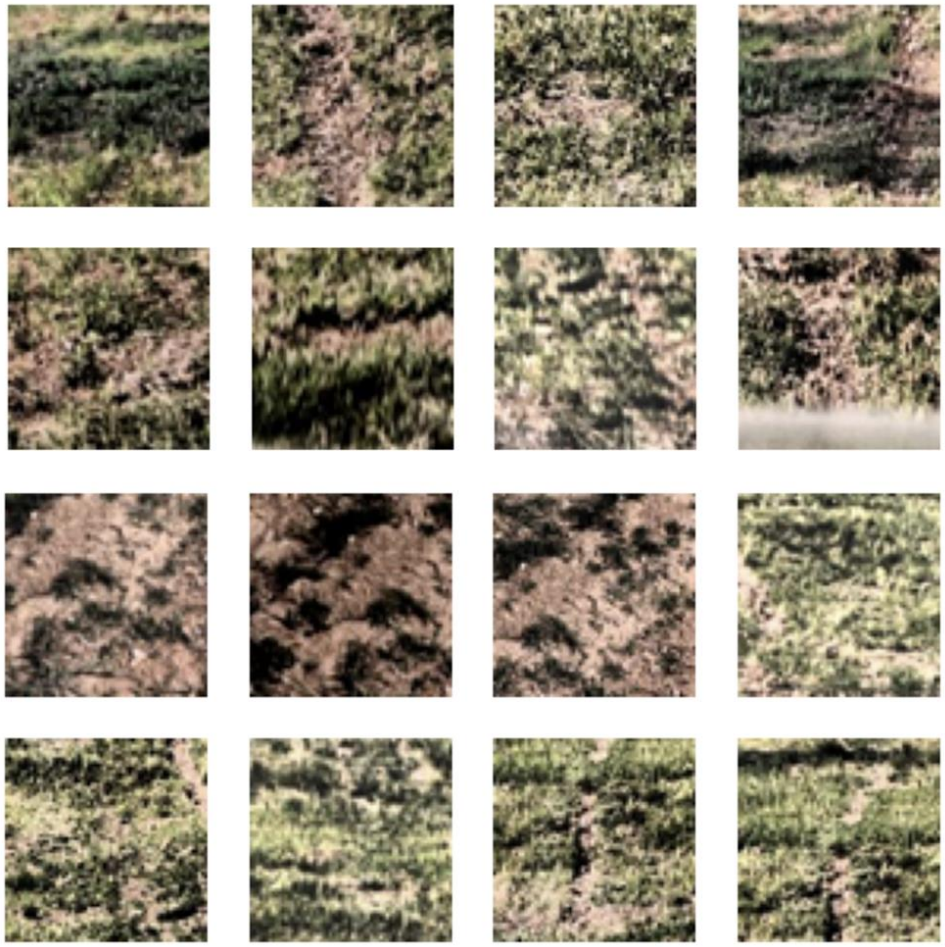

Fig. 2. **Examples of background images**
Source: **compiled by the authors**

The complete dataset is stored on GitHub [23]. The metric used is:

– accuracy, as the proportion of correctly recognized images out of the total number of images.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (1)$$

- completeness, the number of recognized mines from the total number of mines in the images,

$$Recall = \frac{TP}{TP + FN}, \quad (2)$$

where TP (True Positive) is correctly detected mines, TN (True Negative) is correctly classified absence of mines, FP (False Positive) – indications of mines for images where they are actually absent, FN (False Negative) is indications of the absence of mines for images where they are actually present.

The source code for all neural network programs used in the study is available on GitHub [23]. During training, the weight vector that provided maximum accuracy for the validation data was memorized. This vector was used to calculate the resulting characteristics for the test data.

5. RESEARCH RESULTS

5.1. A simple convolutional neural network

The network contains two convolutional layers with Relu activation function. After each of these, there's a MaxPool layer for dimensionality reduction. After that, there's a fully connected Dence layer with a Relu activation function of 64 dimensions, and a Dence output layer of 1 dimension with a sigmoid activation function:

```
# Block 1
x = layers.Conv2D(32, (3, 3), activation="relu")(x)
x = layers.MaxPool2D((2, 2))(x)

# Block 2
x = layers.Conv2D(64, (3, 3), activation="relu")(x)
x = layers.MaxPool2D((2, 2))(x)

# Fully Connected Layers
x = layers.Flatten()(x)
x = layers.Dense(64, activation="relu")(x)
x = layers.Dropout(0.5)(x)
outputs = layers.Dense(1, activation="sigmoid")(x)
```

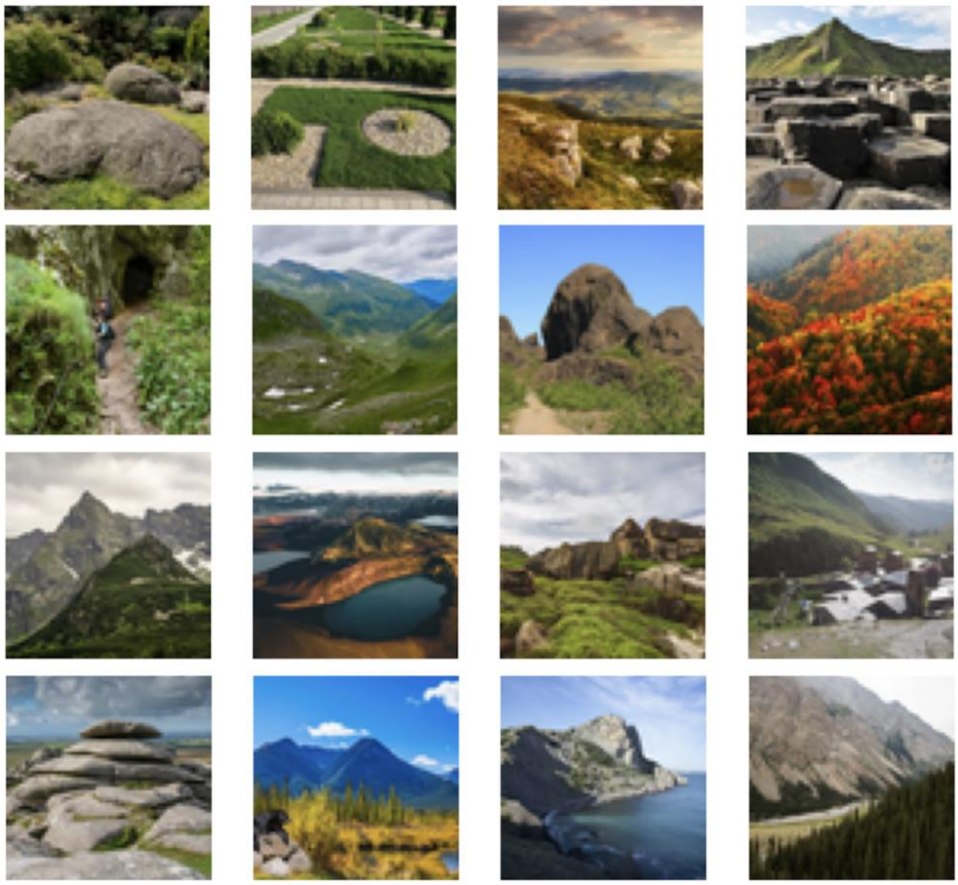

Fig. 3. **Examples of background images**
Source: **compiled by the authors**

The number of trainable weights in this neural network is 822,337.

The learning curves for accuracy, error, and ROC curves for the training and validation data are shown in Fig. 4, Fig. 5, and Fig. 6, respectively. Analysis of the learning curves shows that the network trained in approximately 40 epochs. The area under the ROC curve (AUC) is 0.99, indicating that the neural network distinguishes well between classes (presence of a mine in the image/absence of a mine in the image). The confusion matrix for the test data is shown in Fig. 7.

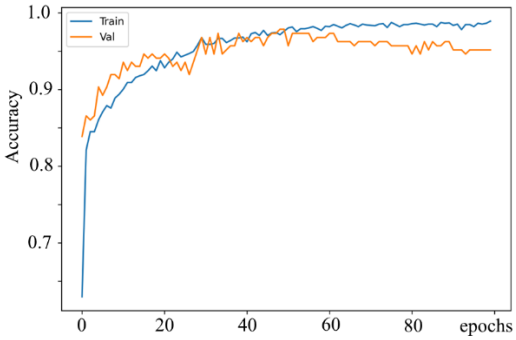

Fig. 4. **Learning curves for accuracy CNN**
Source: **compiled by the authors**

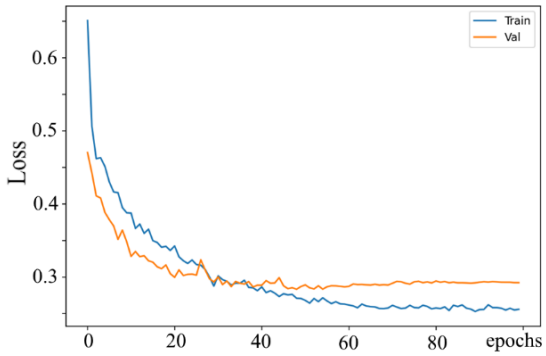

Fig. 5. **Learning curves for CNN Loss**
Source: **compiled by the authors**

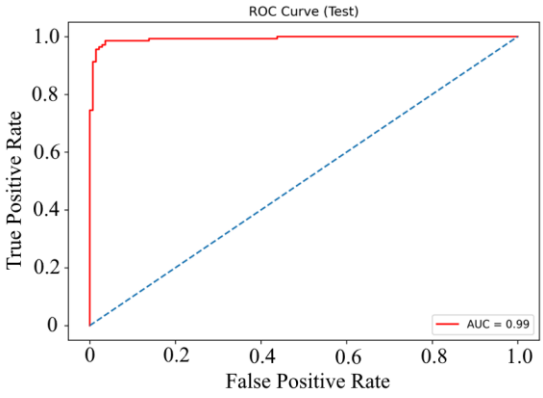

Fig. 6. **ROC curve for CNN**
Source: **compiled by the authors**

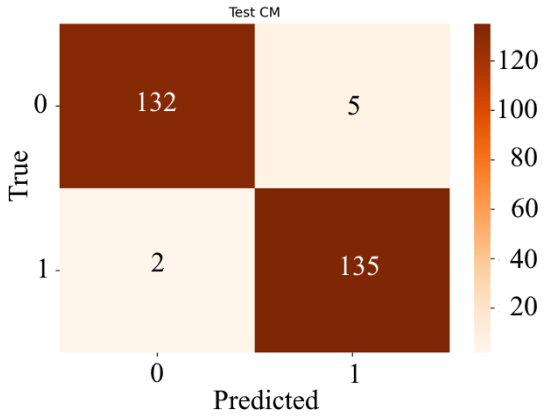

Fig. 7. **Confusion matrix on test data for CNN**
Source: **compiled by the authors**

Confusion matrix indicates that out of a total of 137 mine images, two were misclassified, meaning two mines were missed. There were five false positives.

The model's confidence distribution is shown in Fig. 8.

The confidence distribution shows that the classes are fairly well separated, with virtually no guesswork, though a small amount does occur. It's also worth noting that the edges of the distribution are slightly blurred, suggesting a predisposition to errors when the input images degrade. The images where errors were made are shown in Fig. 9.

From Fig. 9 it is clear that one omission of the mine was made on an image with a sparse distribution of illumination - there were glare from the mine, the second omission - the mine merged strongly with the background.

5.2. Convolutional neural network ResNet-18

The input of the convolutional network is an input convolutional layer, which receives the input image. Next, there are 8 blocks, each containing two convolutional layers. The activation function is Relu. The output of the second convolutional network is summed with the input of the first convolutional network (skip connections).

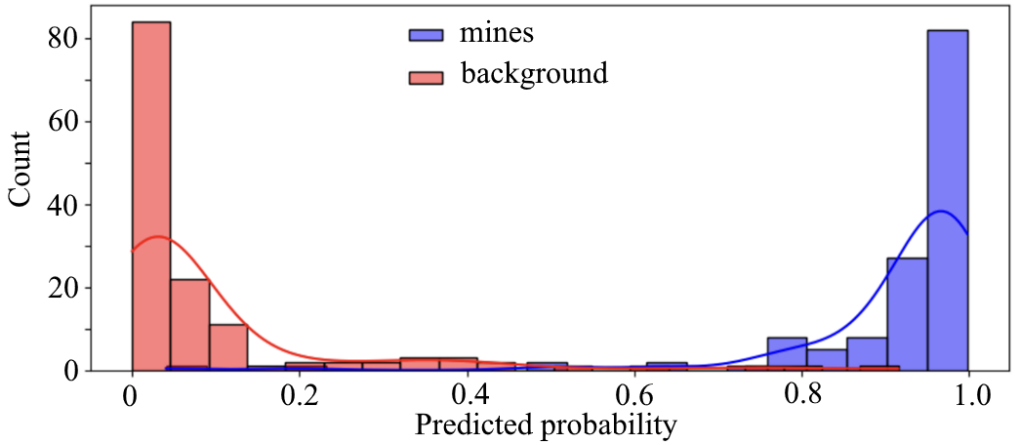

Fig. 8. **Confusion distribution for CNN**
Source: **compiled by the authors**

Fig. 9. **Images where CNN made errors (T – True, P – Predicted)**
Source: **compiled by the authors**

```
def residual_block(x, filters, kernel_size=3, stride=1,
downsample=False):
    shortcut = x
    if downsample:
        shortcut = layers.Conv2D(filters, 1, strides=stride,
padding="same")(shortcut)
        shortcut = layers.BatchNormalization()(shortcut)

    x = layers.Conv2D(filters, kernel_size, strides=stride,
padding="same", use_bias=False)(x)
    x = layers.BatchNormalization()(x)
    x = layers.Activation("relu")(x)

    x = layers.Conv2D(filters, kernel_size, strides=1,
padding="same", use_bias=False)(x)
    x = layers.BatchNormalization()(x)

    x = layers.Add()([x, shortcut])
    x = layers.Activation("relu")(x)
    return x
```

The output is a fully connected Dence layer with a sigmoid activation function. The ResNet-18 neural network contains a total of 18 layers with trainable weights.

The number of trainable weights in this neural network is 11,177,921. The learning curves for accuracy, error, and ROC curves for the training and validation data are shown in Fig. 10, Fig. 11, and Fig. 12, respectively.

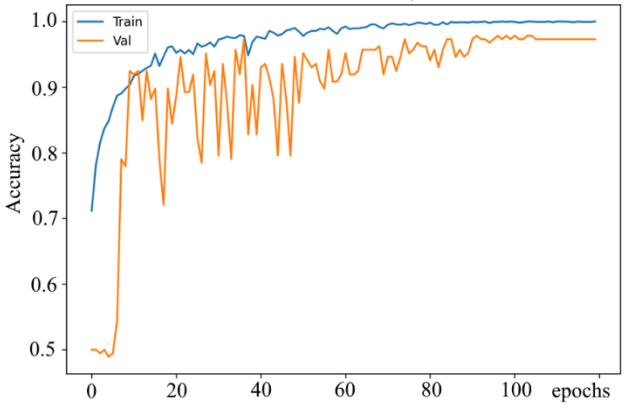

Fig. 10. **Learning curves for accuracy ResNet-18**
Source: **compiled by the authors**

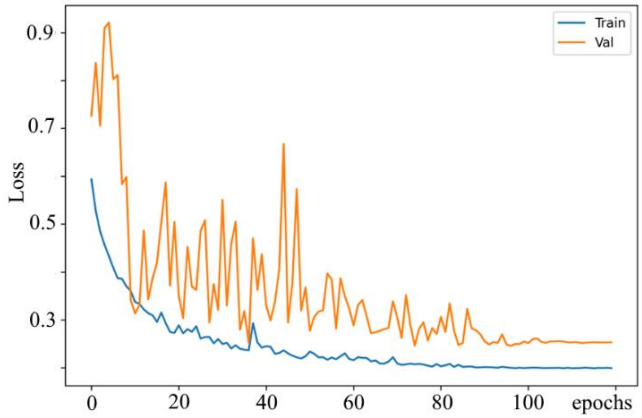

Fig. 11. **Learning curves for ResNet-18 Loss**
Source: **compiled by the authors**

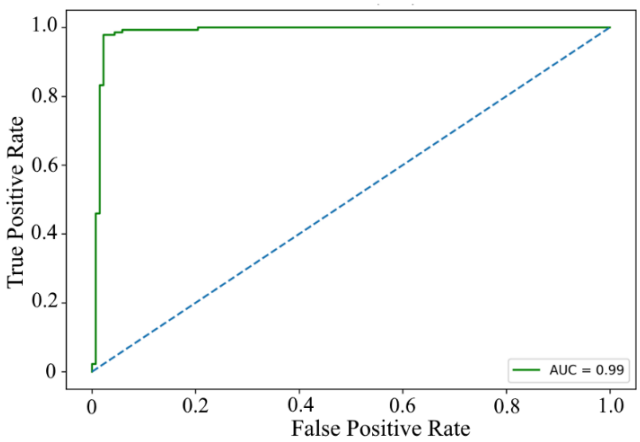

Fig. 12. **ROC curve for ResNet-18**
Source: **compiled by the authors**

Analysis of the learning curves shows that the network trained in approximately 100 epochs. The area under the ROC curve (AUC) is 0.99, indicating that the neural network distinguishes well between classes (presence of a mine in the image/absence of a mine in the image).

The confusion matrices for the test data are shown in Fig. 13.

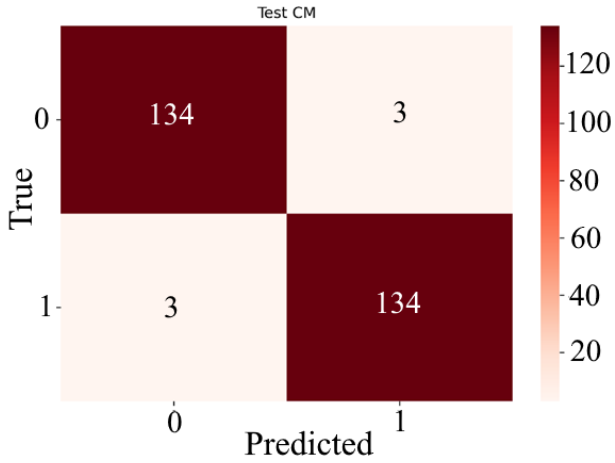

Fig. 13. **Confusion matrix ResNet-18 on test data**
Source: **compiled by the authors**

The confusion matrix shows that out of a total of 137 mine images, three were misclassified. There were three false positives. Despite ResNet-18 having more than an order of magnitude more weighting coefficients and an accuracy of 97.81 % (6 errors) than the simple CNN, three mines were missed. The CNN missed only two mines.

The model's confidence distribution is shown in Fig. 14.

The confidence distribution shows that the classes are fairly well separated, with no guesswork. It's also worth noting that the edges of the distribution are significantly less blurry than those of the CNN.

Images where errors were made are shown in Fig. 15.

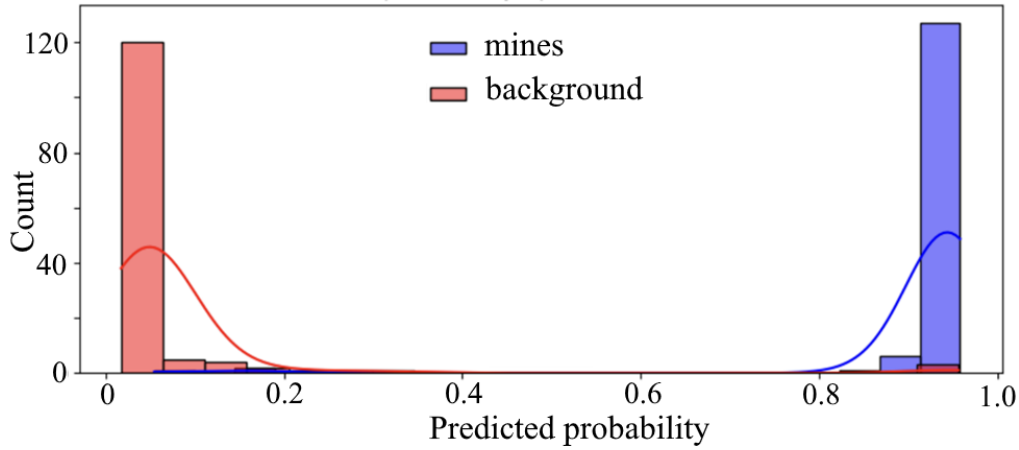

Fig.14. **Confidence distribution for ResNet-18**
Source: **compiled by the authors**

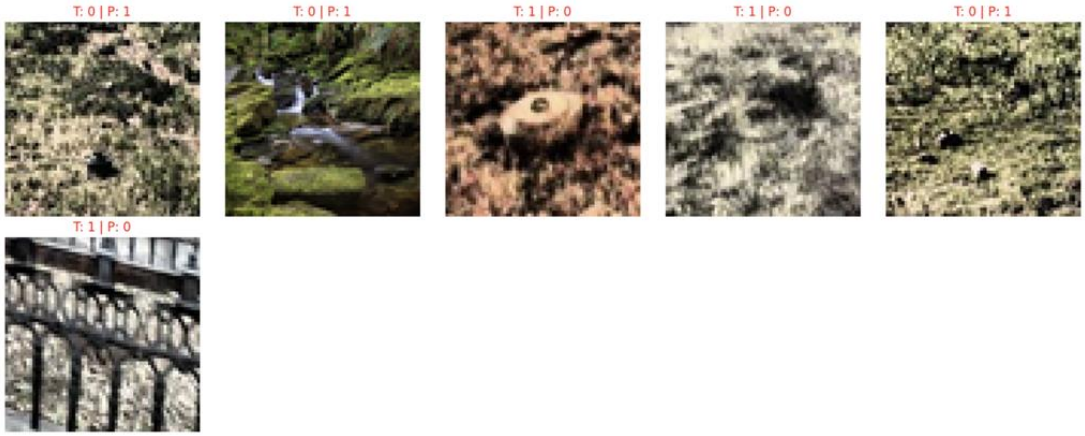

Fig. 15. **Images where ResNet-18 made errors (T – True, P – Predicted)**
Source: **compiled by the authors**

Since the classic video transform requires a large amount of training data, and this study is conducted on a very small data set, several computationally intensive variants of ViT were investigated. Given the small image size of 64x64 pixels, small patch sizes of 8x8 and 4x4 were used. The results are presented in Table 1.

Table 1 show that ViT-Tiny has the best accuracy-to-computation ratio, but it missed 6 mines on the test data. ViT-Lite has the second-best accuracy-to-computation ratio, missing 4 mines. The remaining two variants, ViT-Robust and ViT-Deep, provide higher accuracy and fewer missed mines, but also require significantly more computation. Therefore, based on the combined parameters, ViT-Lite provides the best result expected from classic ViT on the dataset under study.

The general pattern between parameters and characteristics is evident from Table 1.

Increased ViT accuracy is achieved by:

– increasing the projection dimension (the length of the vector into which the patch is transformed);

– increasing the number of heads;

– increasing the number of Transformer layers;

– decreasing the patch dimension.

However, it's important to note that Transformer parameters don't change arbitrarily, but rather in specific increments. For example, the dimensions of the vectors into which patches are transformed must be evenly divisible by the number of heads used.

Fig. 15 shows that ResNet-18 missed mines in the same two images as the CNN. Furthermore, ResNet-18 missed a mine in the image where the mine was located behind a fence. Two of the three false positives were made in the same images as the CNN.

5.3. Classic vit (vision transformer)

ViT-Lite's learning curves for accuracy, error, and ROC curves for the training and validation data are shown in Fig. 16, Fig. 17, and Fig. 18, respectively.

The confusion matrix for the test data is shown in Fig. 19.

Table 1. **Parameters of the classical ViT variants**

Parameter	ViT-Tiny (Basic)	ViT-Lite	ViT-Robust (Reinforced)	ViT-Deep
Projection Dim	64	128	192	128
Heads	4	4	6	8
Transformer Layers	2	4	6	12
Patch Size	8×8	8×8	8×8	4×4
Label Smoothing	0.05	0.1	0.15	0.1
Advantage	Very fast, does not overfit on small data	Balanced in speed and accuracy	Better resistance to noise and fog	Maximum extraction of small features
Risk	May not capture complex mine textures	Standard option	Requires more memory and training time	High risk of overfitting
Trainable params	328705	563201	1856705	3667521
Accuracy	0.9635	0.9416	0.9526	0.9708
Recall	0.9562	0.9708	0.9854	0.9854
FN	6	4	2	2
Accuracy/ Trainable params	$2.93*10^{-6}$	$1.67*10^{-6}$	$0.51*10^{-6}$	$0.26*10^{-6}$

Source: **compiled by the authors**

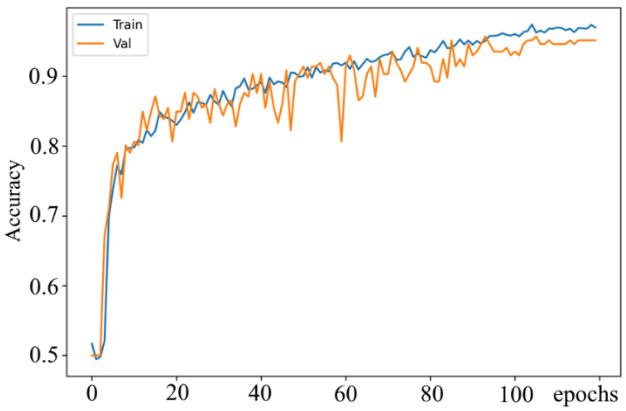

Fig. 16. **Learning curves for accuracy ViT-Lite**
Source: **compiled by the authors**

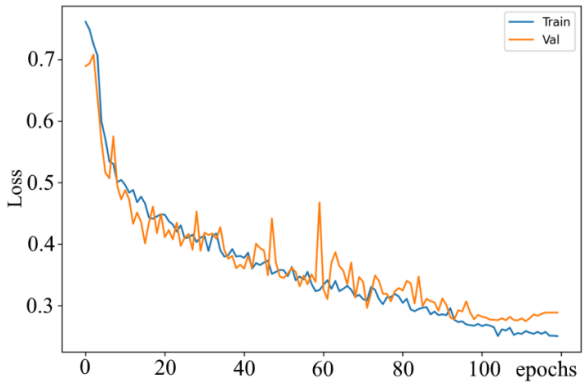

Fig. 17. Learning curves for ViT-Lite Loss
Source: **compiled by the authors**

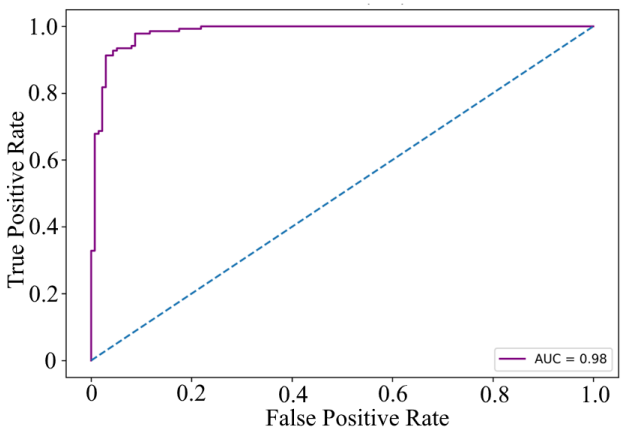

Fig.18. **ROC curve for ViT-Lite**
Source: **compiled by the authors**

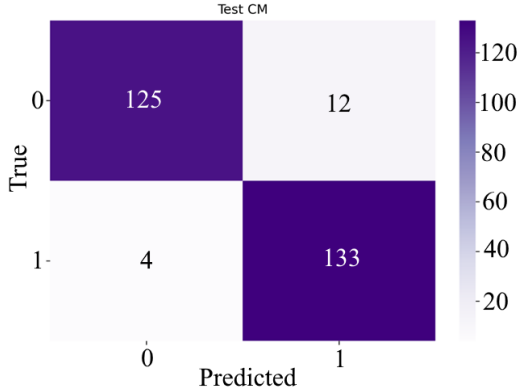

Fig. 19. **Confusion matrix ViT-Lite on test data**
Source: **compiled by the authors**

Analysis of the learning curves shows that the network trained in approximately 110 epochs. The area under the ROC curve (AUC) is 0.98, indicating that the neural network distinguishes well between classes (presence of a mine in the image/absence of a mine in the image).

The confusion matrix indicates that out of a total of 137 mine images, four were misclassified. There were 12 false positives. The model confidence distribution is shown in Fig. 20.

The confidence distribution shows that the classes are fairly well separated, but a small amount of guesswork still occurs. It's also worth noting that the edges of the distribution are slightly blurred, suggesting a predisposition to errors when the input images degrade.

The 15 images where errors were made are shown in Fig. 18.

5.4. Compact convolutional transformer (CCT)

To ensure higher classification accuracy while reducing computational overhead, the following changes were made to the CCT architecture compared to the classic ViT [17]:

1. Patch embedding has been replaced with a convolutional tokenizer. The tokenizer uses two convolutional layers with overlapping filters, a GELU activation function, and BatchNormalization. This allows for more efficient extraction of low-level features and takes into account local relationships between pixels.

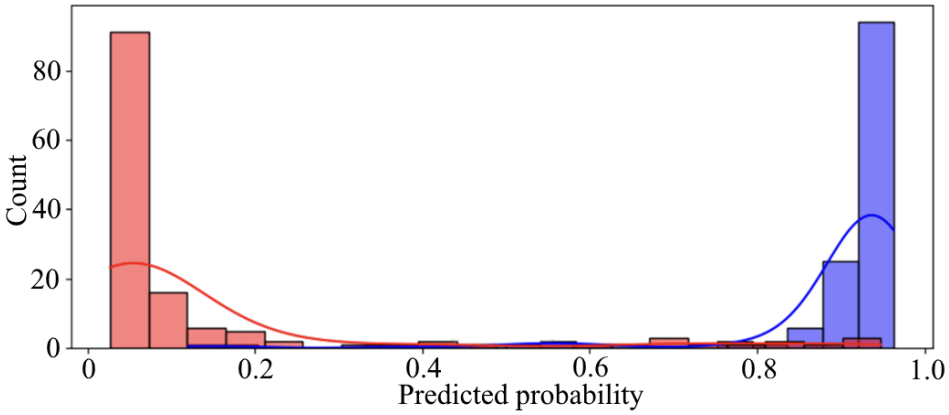

Fig. 20. **Confidence distribution for ViT-Lite**
Source: **compiled by the authors**

Fig. 21. **Images where ViT-Lite made errors (T – True, P – Predicted)**
Source: **compiled by the authors**

2. Global Average Pooling has been replaced with Sequence Pooling. In ViT, all tokens are equally weighted. CCT, using attention, learns to determine which parts of the image (tokens) are most important for classification (e.g., the location of a mine) and assigns them greater weight.

3. The MLP block expansion factor was increased from 2 to 3. This increases the number of weights in each layer processing features extracted by the tokenizer.

4. Warmup was introduced—a gradual increase in the learning rate followed by cosine decay. This allows the model to better learn on small datasets.

The learning curves for accuracy, error, and ROC curves for the training and validation data are shown in Fig. 22, Fig. 23, and Fig. 24, respectively.

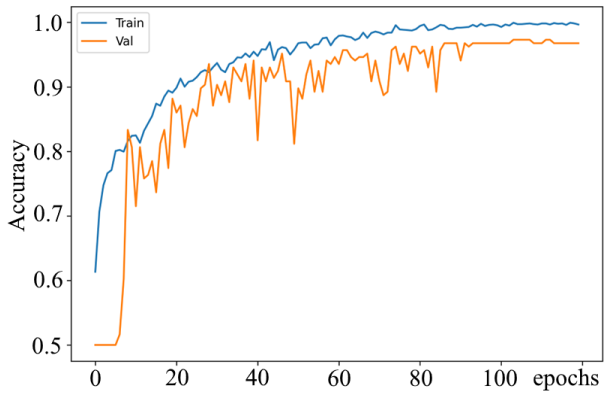

Fig. 22. **Learning curves for accuracy CCT**
Source: **compiled by the authors**

Analysis of the learning curves shows that the network trained in approximately 90 epochs. The area under the ROC curve (AUC) is 0.98, indicating that the neural network distinguishes between classes (presence of a mine in the image/absence of a mine in the image) well. The confusion matrix for the test data is shown in Fig. 25.

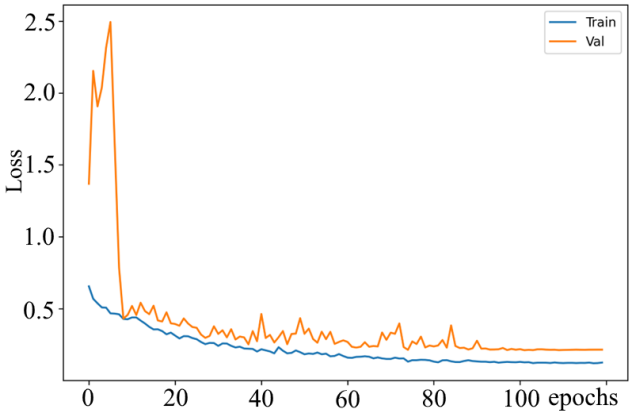

Fig. 23. **Learning curves for CCT Loss**
Source: **compiled by the authors**

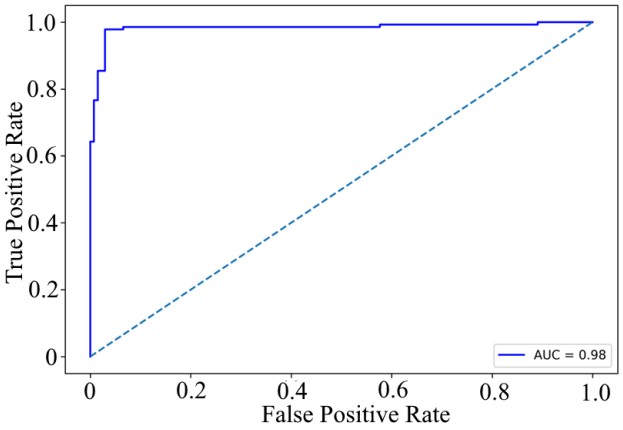

Fig. 24. **ROC curve for CCT**
Source: **compiled by the authors**

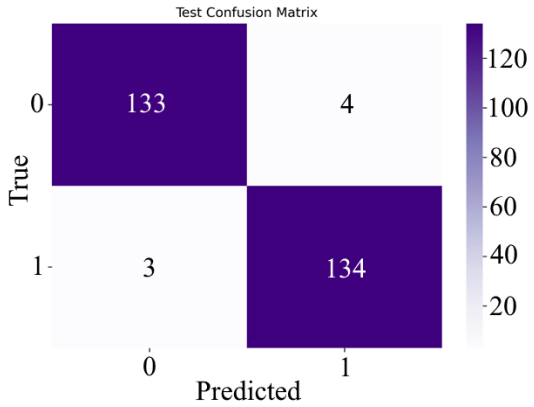

Fig. 25. **Confusion matrix CCT on test data**
Source: **compiled by the authors**

The confusion matrix indicates that out of a total of 137 images, three were misclassified. There were four false positives. The model's confidence distribution is shown in Fig. 26.

The confidence distribution shows that the classes are fairly well separated, but a small amount of guesswork remains. It's also worth noting that the edges of the distribution for CCT are significantly less blurred than for ViT-Lite.

The seven images where errors were made are shown in Fig. 27.

Two of the three mine-missing errors were made on the same images that produced errors in other neural networks. To illustrate the operation of the CCT, Fig. 28 shows an example of an input image with a mine and the attention map for this image in the first CCT layer. Fig. 29 shows the same image, but with stones instead of mines.

Lighter colors on the attention map indicate locations in the image where a mine is most likely to be located. The lighter the color, the greater the probability. Accordingly, these locations are weighted more heavily than others.

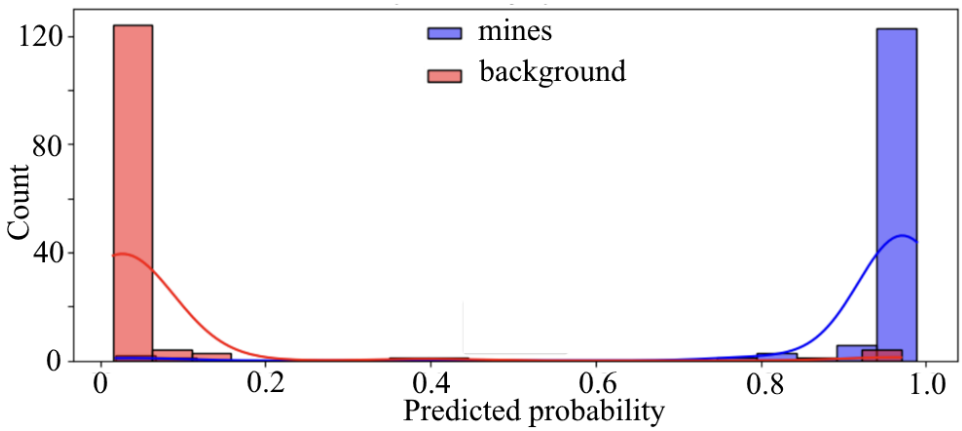

Fig. 26. **Confidence distribution for CCT**
Source: **compiled by the authors**

Fig. 27. **Images where CCT made errors (T – True, P – Predicted)**
Source: **compiled by the authors**

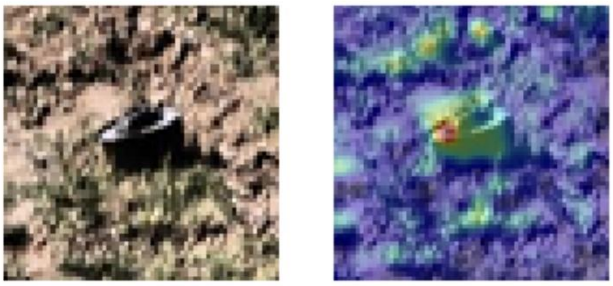

Fig. 28. **Input image with a mine (left) and the attention map in the first layer of CCT (right)**
Source: **compiled by the authors**

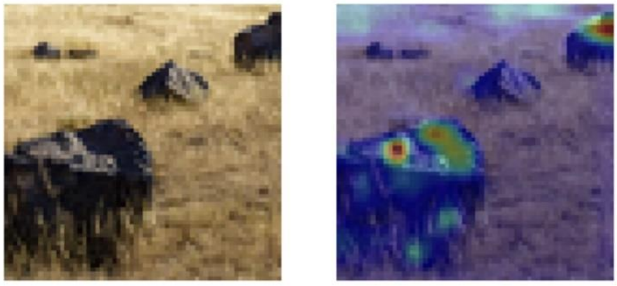

Fig. 29. **Input image with stones (left) and attention map in the first layer of CCT (right)**
Source: **compiled by the authors**

5.5. Analysis of the stability of neural networks to deterioration in the quality of input images

To test the neural networks' resilience to input image quality degradation, the following conditions were simulated for the test images:

– 15 % noise;

– cloudy day (40 % brightness reduction);

– fog (light haze, white content reduced to 20%).

An example of the original image and the degraded images is shown in Fig. 30; the accuracy values for the test data are presented in Table 2.

Table 2 shows that convolutional networks exhibit higher robustness than Transformer-based networks. This can be explained by the larger number of trainable coefficients. However, it should be noted that operation under degraded conditions is not the primary mode, as the probability of missing mines increases significantly.

5.6. Resulting comparison of neural networks

The resulting parameters of the compared neural networks on the test data set are given in Table 3.

Table 3 shows that transformer-based neural networks have a higher accuracy-to-computation ratio. Given that ViT-Lite has lower accuracy and a higher number of missed mines, using CCT are the best option.

Comparison of the performance of CNN and ResNet-18 suggests that the potential of convolutional networks in this application is limited, since a radical increase in the number of trainable weight coefficients did not lead to a significant increase in accuracy, but only to an increase in confidence in the decisions made.

6. DISCUSSION OF THE RESULTS OF THE STUDY ON THE EFFECTIVENESS OF USING VARIOUS NEURAL NETWORKS FOR DETECTING MINES BY DRONES

The study showed that the potential of simple convolutional networks is limited. Significantly increasing the number of layers used does not significantly improve accuracy. This is because convolutional neural networks are designed to search for image details using small filters. This mechanism allows them to quickly learn on small datasets. This is clearly evident in the obtained learning curves. However, when the potential limit is approached, adding additional layers to ResNet-18, compared to CNNs that search for even larger details, does not significantly improve classification quality.

Neural networks built on Transformers differ from convolutional networks in that they process all image patches simultaneously and attempt to find connections between them, regardless of whether they are adjacent or on opposite sides of the image. This architecture results in such neural networks learning more slowly and requiring a large amount of training data. While CNNs study all image components without prioritizing any particular areas, Transformers generate attention maps, allowing them to identify areas of the most probable object locations.

Table 2. **Accuracy for comparable algorithms**

Conditions	CNN	ResNet-18	ViT-Lite	CCT
No deterioration	97,45 %	97,8 %	94,16 %	97,45 %
Noise	96,0 %	97,1 %	92,0 %	90,9 %
Nasty day	83,2 %	89,8 %	80,3 %	83,2 %
Fog	91,6 %	94,2 %	84,3 %	83,6 %

Source: **compiled by the authors**

a b c d

Fig. 30. **An example of a mine image, original image (a), with noise (b), cloudy day (c), fog (d)**
Source: **compiled by the authors**

Table 3. **Resulting parameters of neural networks**

Parameters	CNN	ResNet-18	ViT-Lite	CCT
Accuracy	97,45 %	97,8 %	94,16 %	97,45 %
Number of trainable weight coefficients	822337	11177921	563201	737858
The ratio of accuracy to the number of trainable weights	$1.19*10^{-6}$	$0.087*10^{-6}$	$1,67*10^{-6}$	$1.32*10^{-6}$
FN	2	3	4	3
Confidence in the decisions made	Reduced, fortune telling is possible	Increased	Reduced	Increased
Resistance to image degradation	Increased	Increased	Reduced	Reduced

Source: **compiled by the authors**

When processing small datasets, this leads to both CNNs and Transformers failing to provide a sufficiently high classification performance to computational effort ratio. However, while CNNs can generally only increase the number of layers used, Transformers offer significantly more flexibility in architectural modifications. This is evident in the CCT characteristics. The introduction of two convolutional layers in the tokenizer and division into overlapping patches ensures a high classification performance to computational effort ratio with high accuracy.

An analysis of images where the neural networks made errors shows that in some images, all networks made errors. This suggests that the dataset does not adequately represent all possible variations of mine images against different backgrounds. Expanding the dataset with similar images should improve mine recognition accuracy.

An analysis of the robustness of test images shows a higher dependence of Transformers on the dataset they were trained on. Convolutional networks are less sensitive to noise, illumination, and fog thanks to their feature-extracting architecture. By expanding the dataset to include images with degraded quality, we can expect an increase in the accuracy of all neural networks.

A limitation of the study is the small size of the dataset used. Nevertheless, this allowed us to identify the most promising architecture – the Compact Convolutional Transformer (CCT). By varying the nomenclature and number of additional architectural elements introduced into the Transformer, it is possible to significantly improve both the accuracy and the ratio of classification accuracy to the required computational effort. Combined with an increased number and variety of mine images against various backgrounds, this will enable the highest possible performance in the important field of humanitarian demining.

7. CONCLUSIONS

1. A comparison of convolutional neural networks and transformer-based networks revealed that the Compact Convolutional Transformer (CCT) architecture is the most promising architecture for use on mine-detection drones. It provides both high mine classification accuracy and a high accuracy-to-computation ratio. This is a result of the use of both the Transformer architecture itself and additional architectural components.

2. An analysis of the robustness to test image quality degradation demonstrated the superiority of convolutional neural networks. This necessitates increasing the number and range of images for training, which should improve the robustness of all the neural networks studied.

REFERENCES

1. Zhang, Z. & Zhu, L. “A review on unmanned aerial vehicle remote sensing: platforms, sensors, data processing methods, and applications. *Drones.* 2023; 7 (6): 398, https://scopus.com/pages/publications/85163140510. DOI: https://doi.org/10.3390/drones7060398.

2. Tian, Y., Ye, Q. & Doermann, D. “YOLOv12: Attention-Centric real-time object detectors”. *arXiv preprint*. 2025. DOI: https://doi.org/10.48550/arXiv.2502.12524.

3. Mary Shanthi Rani, M., Chitra, P., Lakshmanan, S., Kalpana Devi, M., Sangeetha, R. & Nithya, S. “DeepCompNet: A novel neural net model compression architecture”. *Hindawi. Computational Intelligence and Neuroscience*. 2022; 2022: 2213273, https://scopus.com/pages/publications/85139886348. DOI: https://doi.org/10.1155/2022/2213273.

4. Han, S., Mao, H. & Dally, W. J. “Deep compression: compressing deep neural networks with pruning, trained quantization and Huffman coding”. *arXiv*. 2015. DOI: https://doi.org/10.48550/arXiv.1510.00149.

5. Galchonkov, O., Nevrev, A., Glava, M. & Babych, M. “Exploring the efficiency of the combined application of connection pruning and source data preprocessing when training a multilayer perceptron”. *Eastern-European Journal of Enterprise Technologies. Information and controlling system*. 2020; 2 (9 (104)): 6–13, https://scopus.com/pages/publications/85085116742. DOI: https://doi.org/10.15587/1729-4061.2020.200819.

6. Wu, K., Guo, Y. & Zhang, C. “Compressing deep neural networks with sparse matrix factorization”. *IEEE Transactions on Neural Networks and Learning Systems*. 2019; 31 (10): 3828–3838, https://scopus.com/pages/publications/85074744315. DOI: https://doi.org/10.1109/TNNLS.2019.2946636.

7. Cheng, X., Rao, Z., Chen, Y. & Zhang, Q. “Explaining knowledge distillation by quantifying the knowledge”. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Seattle, WT, USA. 2020. p. 12925–12935, https://scopus.com/pages/publications/85091515938. DOI: https://doi.org/10.1109/CVPR42600.2020.01294.

8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. "An image is worth 16x16 words: transformers for image recognition at scale". *arXiv preprint.* 2021. DOI: https://doi.org/10.48550/arXiv.2010.11929.

9. Vaswani, A., Shazeer, N., Parmar, N., et al. "Attention is all you need". *31st Conference on Neural Information Processing Systems (NIPS 2017)*. Long Beach, CA, USA. 2017, https://scopus.com/pages/publications/ 85041738201. DOI: https://doi.org/10.48550/arXiv.1706.03762.

10. Yuan, L., Chen, Y., Wang, T., et al. "Tokensto-token vit: Training vision transformers from scratch on imagenet". *arXiv preprint.* 2021. DOI: https://doi.org/10.48550/arXiv.2101.11986.

11. d'Ascoli, S., Touvron, H., Leavitt, M., Morcos, A., Biroli, G. & Sagun, L. "Convit: Improving vision transformers with soft convolutional inductive biases". *arXiv preprint.* 2021. DOI: https://doi.org/10.48550/arXiv.2103.10697.

12. Yuan, K., Guo, S., Liu, Z., Zhou, A., Yu, F. & Wu, W. "Incorporating convolution designs into visual transformers". *arXiv preprint*. 2021. DOI: https://doi.org/10.48550/arXiv.2103.11816.

13. Wu, H., Xiao, B., Codella, N., et al. "CvT: Introducing convolutions to vision transformers". *arXiv preprint.* 2021. DOI: https://doi.org/10.48550/arXiv.2103.15808.

14. Guo, M.-H., Liu, Z.-N., Mu, T.-J., Hu, S.-M. "Beyond self-attention: External attention using two linear layers for visual tasks". *arXiv preprint.* 2021. DOI: https://doi.org/10.48550/arXiv.2105.02358.

15. Galchonkov, O., Babych, M., Zasidko, A. & Poberezhnyi, S. "Using a neural network in the second stage of the ensemble classifier to improve the quality of classification of objects in images". *Eastern-European Journal of Enterprise Technologies*. 2022; 3 (9 (117)): 15–21, https://scopus.com/pages/publications/85134657446. DOI: https://doi.org/10.15587/1729-4061.2022.258187.

16. Krizhevsky, A. "The CIFAR-10 dataset". – Available from: https://www.cs.toronto.edu/~kriz/cifar.html.

17. Hassani, A., Walton, S., Shah, N., Abuduweili, A., Li, J. & Shi, H. "Escaping the big data paradigm with compact transformers". *arXiv preprint.* 2021. DOI: https://doi.org/10.48550/arXiv.2104.05704.

18. Lee-Thorp, J., Ainslie, J., Eckstein, I. & Ontañón, S. "FNet: Mixing tokens with fourier transforms". *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* 2021. p. 4296–4313, https://scopus.com/pages/publications/ 85136450689.DOI: https://doi.org/10.18653/v1/2022.naacl-main.319.

19. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. & Guo, B. "Swin transformer: hierarchical vision transformer using shifted windows". *arXiv preprint.* 2021. DOI: https://doi.org/10.48550/arXiv.2103.14030.

20. Brownlee, J. "Deep learning for computer vision. Image classification, object detection and face recognition in Python". – Available from: https://machinelearningmastery.com/deep-learning-for-computer-vision. – [Accessed: Feb, 2025].

21. "Landmines Detection Dataset". *Northumbria*. – Available from: https://universe.roboflow.com/northumbria/landmines-detection-dataset. – [Accessed: Jan, 2026].

22. "Freepik". – Available from: https://freepik.com. – [Accessed: Jan, 2026].

23."Landmine_new_64_3". – Available from: https://github.com/oleksii-m-baranov/landmines-recognition. – [Accessed: Jan, 2026].

Conflicts of Interest: The authors declare that they have no conflict of interest regarding this study, including financial, personal, authorship or other, which could influence the research and its results presented in this article

Data Availability: The article has linked data in the data repository. Links are provided in the text of the article

Use of artificial intelligence tools: The authors used artificial intelligence technologies within the permissible framework to provide their own verified data, which is described in the research methodology section

Received 26.12.2025
Received after revision 12.02.2026
Accepted 19.02.2026

DOI: https://doi.org/10.15276/aait.09.2026.01
УДК 681.3.07: 004.8

Підвищення якості виявлення мін дроном-міношукачом за рахунок вибору архітектури класифікатора об'єктів на зображеннях з камери у видимому діапазоні

Галчонков Олег Миколайович[1)]
ORCID: https://orcid.org/0000-0001-5468-7299; o.n.galchenkov@gmail.com. Scopus Author ID: 56081377900
Баранов Олексій Миколайович[2)]
ORCID: https://orcid.org/0009-0002-5951-2636; oleksii.m.baranov@gmail.com
Баськов Ілля Олександрович[1)]
ORCID: https://orcid.org/0000-0002-3517-6773; illyabaskov@gmail.com
[1)] Національний університет «Одеська політехніка», пр. Шевченка, 1. Одеса, 65044, Україна
[2)] Oracle Corporation, Oracle World Headquarters, Oracle Way 2300. Austin, USA TX 78741

АНОТАЦІЯ

Використання безпілотних літальних апаратів (дронів) знаходить все ширше застосування при гуманітарному розмінуванні. Для пошуку мін застосовують одночасно кілька датчиків, що працюють за різними фізичними принципами. Це дозволяє підвищити можливість виявлення мін. Для прискорення обстеження території первинна обробка сигналів датчиків проводиться на борту дрону. Тому до алгоритмів обробки сигналів датчиків пред'являється вимога як забезпечувати підвищену ймовірність виявлення об'єктів, так і підвищене співвідношення точності класифікації та необхідного обсягу обчислень. Одним із використовуваних датчиків є камери у видимому діапазоні. На даний момент розроблено велику кількість нейронних мереж для класифікації об'єктів на зображеннях. Однак особливістю їх використання для гуманітарного завдання розмінування є відсутність великих наборів даних для навчання і тестування. Тому виникає завдання пошуку нейронних мереж, що відрізняються високим співвідношенням точності класифікації та необхідного обсягу обчислень і в той же час здатних навчатися на дуже невеликих обсягах даних. У роботі проведено порівняння згорткових нейронних мереж і мереж, побудованих на базі трансформерів. Навчання мереж проводилося на невеликому наборі даних, що містить зображення протитанкових мін, каміння та різне тло. Дослідження показало, що згорткові нейронні мережі швидше навчаються і стійкіші до погіршення якості зображень. Але їх потенціал обмежений, збільшення кількості шарів не призводить до суттєвого підвищення якості класифікації мін. Нейронні мережі на базі трансформерів мають більшу гнучкість і розмаїття можливостей для конфігурації архітектури, що забезпечує більш високі характеристики, порівняно зі згортковими мережами. Незважаючи на те, що вони повільніше навчаються і потенційно вимагають розширення використовуваного для навчання набору даних, вони є кращими при реалізації обробки зображень на бортовому обладнанні дронів.

Ключові слова: згорткова нейронна мережа; трансформер; увага; вагові коефіцієнти; якість класифікації; токенізація

ABOUT THE AUTHORS

Oleg M. Galchonkov - PhD, Assistent Professor, Information Systems Department. Odesa Polytechnic National University, 1, Shevchenko Ave, Odesa, 65044, Ukraine
ORCID: https://orcid.org/0000-0001-5468-7299; o.n.galchenkov@gmail.com. Scopus Author ID: 56081377900
***Research field*:** Artificial intelligence; evolution systems; machine learning

Галчонков Олег Миколайвич - кандидат технічних наук, доцент кафедри Інформаційних систем. Національний університет "Одеська політехніка", пр. Шевченка, 1. Одеса, 65044,Україна

Alexey M. Baranov - Software Engineer, Oracle Corporation, Oracle World Headquarters, Oracle Way 2300, Austin, USA TX 78741
ORCID: https://orcid.org/0009-0002-5951-2636; oleksii.m.baranov@gmail.com
***Research field*:** Artificial intelligence; evolution systems; machine learning

Баранов Олексій Миколайвич - програміст, Oracle Corporation, Oracle World Headquarters, Oracle Way 2300, Austin, USA TX 78741

Illia O. Baskov - Senior Lecturer, Information Systems Department. Odesa Polytechnic National University, 1, Shevchenko Ave. Odesa, 65044, Ukraine
ORCID: https://orcid.org/0000-0002-3517-6773; illyabaskov@gmail.com
***Research field*:** Artificial intelligence; evolution systems; machine learning

Баськов Ілля Олександрович - старший викладач кафедри Інформаційних систем. Національний університет"Одеська політехніка", пр. Шевченка, 1. Одеса, 65044,Україна