The concept of a deepfake detection system of biometric image modifications based on neural networks

1st Valery Dudykevych
Department of Cybersecurity
Lviv Polytechnic National University
Lviv, Ukraine
ORCID: 0000-0001-8827-9920

2nd Halyna Mykytyn
Department of Cybersecurity
Lviv Polytechnic National University
Lviv, Ukraine
ORCID: 0000-0003-4275-8285

3rd Khrystyna Ruda
Department of Cybersecurity
Lviv Polytechnic National University
Lviv, Ukraine
ORCID: 0000-0001-8644-411X

***Abstract* — The concept of a system for detecting deepfake modifications in biometric images is proposed. The concept is deployed based on the functioning of a convolutional neural network and the algorithm of the biometric image classifier according to the structure "sensitivity-Youden's index-optimal threshold-specificity".**

Keywords — biometric image, deepfake modification, information neural network technology, convolutional neural network, classification, decision support system, concept.

I. Introduction

The relevance of the problem of detecting deepfake modifications of biometric images is growing not only in Ukraine but also throughout the world in modern conditions of hybrid warfare. Modern impersonation systems – deepfakes [1] are used for propaganda and misleading the public, by generating believable appeals by officials that carry the narratives needed by the hostile party. However, the scope of use of these systems is not limited to propaganda[2], they can also be used to establish unauthorized access to confidential information. This poses a significant danger today not only for the person whose privacy was violated but also for the state in general, which is reflected in the Cybersecurity Strategy of Ukraine [3]. Detecting this kind of threat is a significant challenge, as more and more realistic forgery technologies are regularly emerging, which requires innovative approaches[4] to detect deepfake modification of biometric images.

In the space of tasks of digital transformation of society according to the strategy of intellectualization, the task of ensuring the confidentiality of the biometric image in various subject areas of the infrastructure of society is urgent[5]. In this direction, scientists are considering the application of deepfake detection technologies for biometric image modifications. For example, in [6] the authors determine the originality of digital images using the fractal nature of digital signals. In this case, the verification of the originality of digital biometric images takes place concerning the identification of functional characteristics of the digital equipment of their video recording. At the same time, there is a comparison of records stored on media or in the memory of digital video recording equipment in electronic form. The record is original if the identification features of the experimental and researched records match.

In the method described in [7], the detection of deepfake modifications occurs using the localization of a person's face and the correct detection of the area around it for analysis. It has been found that when a face is replaced, its modified biometric image will consist of three zones: the background original image, the replaced face, and a transition zone that will smooth the border between the first two. The system built based on this method implements the search for artifacts that appear in the area of overlapping faces and, based on their detection or non-detection, concludes whether the original or modified image was presented to it.

Another method, which is considered in [8], is based on the fact that the current deepfake algorithm cannot generate an image with a resolution higher than a certain threshold value adopted by the relevant criteria. Hence, these limited-resolution images must be further transformed using affine transforms to match the faces, which are to be replaced in the original video. Such transformations leave certain characteristic artifacts in the resulting modified deepfake video files, which can be effectively registered by a neural network model built according to the appropriate parameters.

In article [9] the degree of difference of unfiltered consecutive frames from each other is analyzed with the help of dispersion, and classification of the dispersion indicators is carried out, not biometric images per se. The authors of the article also justify the ineffectiveness of using a convolutional neural network by the fact that such a network, having a convolutional filter in its structure, ignores the effects of blurring, brightness, contrast, and noise and, accordingly, loses part of the data necessary for training the classifier.

The considered methods become the basis for building the concept of a deepfake detection system of biometric image modifications based on a convolutional neural network.

The purpose of the article is to develop the concept of a deepfake detection system for biometric image modifications based on information neural network technologies, which will ensure the confidentiality of the biometric image in the space of the main cyber security tasks.

II. Concept of a Detection System Based on Information Neural Network Technologies

To provide a holistic presentation of the problem of detecting deepfake modifications to ensure the confidentiality of biometric images and to solve this

2022 IEEE 3rd KhPI Week on Advanced Technology (KhPIWeek) | 979-8-3503-9920-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/KHPIWEEK57572.2022.9916378

979-8-3503-9920-2/22/$31.00 ©2022 IEEE

problem using information neural network technologies, the concept of a deepfake detection system based on convolutional neural networks is proposed (fig.1).

The development of informational neural network technologies for detecting deepfake modifications of a biometric image depends on the level of structured tasks and requires appropriate approaches in the context of ensuring the security of research objects as biometric images according to the confidentiality profile. Following the problematic situation in the context of deepfake detection of biometric image modification, the concept of a deepfake detection system for biometric image modification has the following structure: 1) the subject area of society's infrastructure – research objects (RO_1, RO_2, ..., RO_n); system analysis (principles of integrity, structure, and hierarchy); neural network-based approach to creating deepfake modifications recognition system; 2) information neural network technology (IT1) with the following structure: «object model – methodology – image classification accuracy»; 3) decision support system (IT2). Constructive algorithm IT1 «splitting the video into frames – detection – feature extraction – classification» implemented by a system based on neural networks in structure «indication – interpretation – identification – decision-making», in particular, the functionality of a convolutional neural network "input data – convolution – subsampling. The constructive algorithm IT2 is implemented using a data analysis system to identify the classifier's assessment and make a management decision to establish compliance concerning the detection of deepfake modification. The decision support system identifies the features of the studied biometric image, establishing correspondence with the selected classifier model. In case of discrepancy, the data analysis system decides for the user to build a new classifier model.

To ensure the unification of the methods of creating a system for detecting deepfake modifications based on neural networks in the area of ensuring the confidentiality of biometric images, the structure of the concept provides elements of standardization in the field of neural network technologies, biometric images, cyber security: ISO/IEC 30107-1:2016[10], ISO/IEC TR 24029[11], DSTU ISO/IEC 15408[12], C2PA Specification.

The tools for detecting deepfake modifications of a biometric image are information neural network technologies, the core of which is the system. The main stages of the neural network technology for detecting deepfake modifications of the human face are:

- *Splitting the video into frames*. Since the object of the study is an individual biometric image and not the video in general, for further processing it is necessary to extract images from the stream of frames;
- *Detection*. At this stage, the convolutional neural network detects the faces of people in the image and crops them for the next stage. Since at the next stage the image is studied based on biometric features, it is advisable to extract biometric images of people's faces from the general background;
- *Feature extraction*. At this stage, a specially designed convolutional neural network generates a matrix of features of each biometric image based on the calculated weights that were formed in the process of its training;
- *Classification*. The weights generated in the previous step together with an actual condition defined for each biometric image are used to train the classifier.

III. INFORMATION NEURAL NETWORK TECHNOLOGY

A. *Convolutional neural network.*

Taking into account that the object of research is a biometric image, the convolutional neural network architecture was chosen for its processing, since this architecture is designed to work with images [13]. In the proposed solution, convolutional neural networks are used for the detection of biometric images, as well as for the calculation of significant features of the biometric image. The absence of an output layer in the network, responsible for feature extraction, is compensated by the presence of a separate classifier, which uses the matrix of features of biometric images calculated by the network. The input data for the neural network are biometric images extracted from a video file previously split into frames. At the next stage, biometric images are processed by a convolutional neural network, the result of which is a matrix of image features.

B. *Classifier training.*

The approach to detecting modified images using a convolutional neural network is implemented in several stages. *The first stage* consists in splitting each of the video files into separate frames, which will be further processed with the help of a biometric image detector[14]. The detector recognizes the object, after which we get a biometric image from each recognized human face in the frame of the video recording, normalized to the specified dimensions with a marking to which class (real or modified) the investigated image belongs. *The second stage* uses the normalized biometric face images to be processed by a convolutional neural network to compute the feature vectors used to train the classifier. They are divided into training and test sets for actually training and testing the classifier.

The following interrelated characteristics are involved in the algorithm of the classifier in the space of detection of modified biometric images: accuracy, sensitivity, specificity of the classifier, an optimal threshold value of the classifier, and the Youden index. The Youden index is related to the sensitivity and specificity of the classifier.

The result of the biometric image classification is a confusion matrix in space: correctly classified modified images, correctly classified unmodified images, falsely classified modified images, and falsely classified unmodified images.

The sensitivity of the classifier characterizes the proportion of true positive biometric images that are correctly identified [15]:

$$TPR = \frac{TP}{TP + FN},$$

where TPR – classifier sensitivity, TP – number of correctly classified positive test samples, FN – number of falsely classified negative test samples.

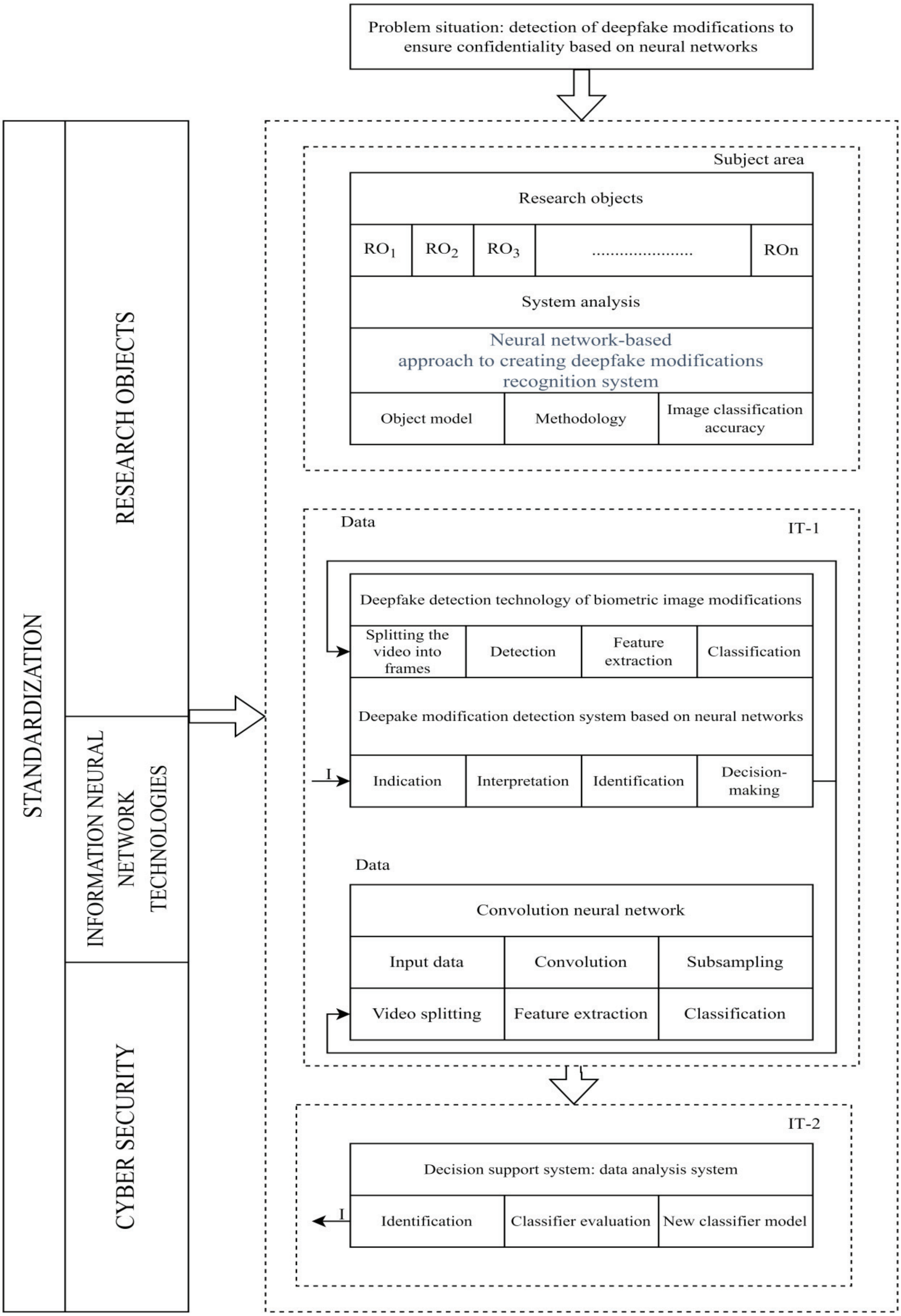
Problem situation: detection of deepfake modifications to ensure confidentiality based on neural networks
STANDARDIZATION
RESEARCH OBJECTS
INFORMATION NEURAL NETWORK TECHNOLOGIES
CYBER SECURITY
Subject area
Research objects
RO$_1$
RO$_2$
RO$_3$
....
ROn
System analysis
Neural network-based approach to creating deepfake modifications recognition system
Object model
Methodology
Image classification accuracy
Data
IT-1
Deepfake detection technology of biometric image modifications
Splitting the video into frames
Detection
Feature extraction
Classification
Deepake modification detection system based on neural networks
I
Indication
Interpretation
Identification
Decision-making
Data
Convolution neural network
Input data
Convolution
Subsampling
Video splitting
Feature extraction
Classification
IT-2
Decision support system: data analysis system
I
Identification
Classifier evaluation
New classifier model

Fig. 1. The cut-off threshold of informed classified images from uninformed classified ones

The investigated biometric image is independently determined by the probability of belonging to the corresponding class. Youden index [16] is used to calculate the optimal threshold value for image classification

$$J = max\ (TPR(t) + TNR(t) - 1),$$

where J – Youden index, t – cut-off threshold of informed classified images from uninformed classified ones, TNR – classifier specificity, which can be expressed as

$$TNR\ = \frac{TN}{TN+FP},$$

where TN – the number of correctly classified negative test samples, FP – the number of falsely classified positive test samples.

Figure 2 presents the threshold value for cutting off the informed classified images from the uninformed classified ones in the space of the relationship of the Youden index with the sensitivity of the classifier (TPR) and its specificity (TNR). Its maximum value determines the optimal threshold value of the classifier, which provides a criterion for balancing the characteristics of the sensitivity of the classifier and its specificity.

The accuracy of biometric image classification in the space of detecting deepfake modifications is determined, among other things, by the number of studied training images, which makes it possible to achieve a sensitivity value of the classifier greater than 0.85.

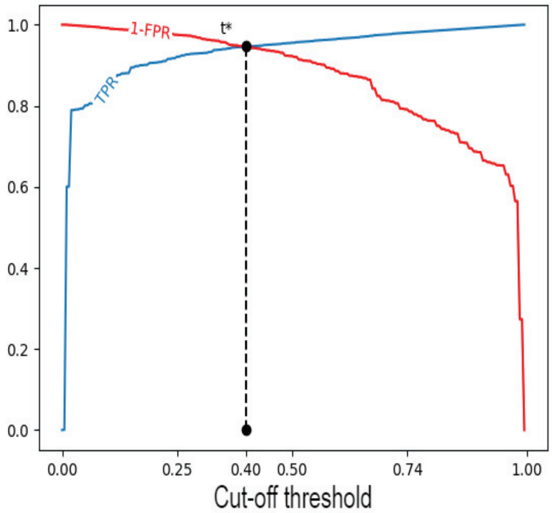

Fig. 2. The cut-off threshold of informed classified images from uninformed classified ones

Conclusions

The proposed concept of a system for detecting deepfake modifications of a biometric image based on neural networks allows us to outline the direction of research – ensuring the confidentiality of information in the space of cyber security of biometric images using: 1) the approach to detecting deepfake modifications "object model – methodology – accuracy of image classification"; 2) systems for detecting deepfake modifications "indication – interpretation – identification –decision-making" based on a convolutional neural network "video separation – feature processing – image classification"; 3) data analysis systems "identification – evaluation of the classifier – new model of the classifier"; 4) algorithmic and software implementation of data processing.

References

[1] Kietzmann J., Lee L. W., McCarthy I. P., and Kietzmann T. C., Deepfakes: Trick or treat? Business Horizons, vol. 63(2), pp. 135-146, 2020

[2] Yevseiev S., Ponomarenko V., Laptiev O., Opirskyy I., Milov O., Korol O., Milevskyi S. et. al.; Yevseiev S., Ponomarenko V., Laptiev O., Opirskyy I., Milov O. (Eds.) Synergy of building cybersecurity systems. Kharkiv: PC TECHNOLOGY CENTER, p.188, 2021.

[3] The Cybersecurity Strategy of Ukraine (2021-2025), [online] Available: https://www.rnbo.gov.ua/files/2021/STRATEGIYA%20KYBERBEZPEKI/proekt%20strategii_kyberbezpeki_Ukr.pdf

[4] Ruban I., Bolohova N., Martovytskyi V., Koptsev, O. Digital image authentication model. Advanced Information Systems, vol. 5(1), pp.113–117, 2021.

[5] Karpinski M, Khoma V., Dudvkevych V., Khoma Y. and Sabodashko D., "Autoencoder Neural Networks for Outlier Correction in ECG- Based Biometric Identification", *2018 IEEE 4th International Symposium on Wireless Systems within the International Conferences on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS-SWS)*, pp. 210-215, 2018.

[6] Rybalskyi O. V., Soloviev V. Y., "On the development of the theory, methods and means of conducting the examination of digital photo, video and sound recording materials, methods and means of conducting the examination of digital photo, video and sound recording materials". Modern Special Technique, vol. 3 (30), pp. 119–121, 2012, (in Russian). Рибальский О. В., Соловьев В. И., К развитию теории, методов и средств проведения єкспертизи материалов цифрових фото, видео и звукозаписи. Сучасна спеціальна техніка, №3 (30), С 119-121, 2012.

[7] Li L., Bao J., Zhang T., Yang H., Chen D., Wen F., & Guo B., Face X-ray for more general face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5001-5010, 2020.

[8] Li Y., Lyu S., Exposing deepfake videos by detecting face warping artifacts. In Proceedings of the IEEE 14 Conference on Computer Vision and Pattern Recognition Workshops, pp. 46-52, 2019.

[9] Lee G., Kim M., Deepfake Detection Using the Rate of Change between Frames Based on Computer Vision. Sensors vol. 21:7367, 2021.

[10] ISO/IEC 30107-1:2016, Information technology — Biometric presentation attack detection — Part 1: Framework

[11] SC42 WG3: Assessment of the robustness of neural networks - part 1: Overview.Tech. Rep. CD TR 24029-1, ISO/IEC JTC 1/SC 42 Artificial Intelligence (2019)

[12] DSTU ISO/IEC 15408-1:2017 Informatsiini tekhnolohii. Metody zakhystu. Kryterii otsinky. Chastyna 1. Vstup ta zahalna model [Information technology — Security techniques — Evaluation criteria for IT security — Part 1: Introduction and general model] (ISO/IEC 15408-1:2009, IDT).

[13] Albawi S., Mohammed T. A. and Al-Zawi S., "Understanding of a convolutional neural network," International Conference on Engineering and Technology (ICET), pp. 1-6, 2017.

[14] Svyrydov A, Kuchuk H., and Tsiapa, O., "Improving efficiency of image recognition process: Approach and case study," *2018 IEEE 9th International Conference on Dependable Systems, Services and Technologies (DESSERT)*, pp. 593-597, 2018.

[15] Yerushalmy S., Statistical problems in assessing methods of medical diagnosis, with special reference to x-ray techniques. Public Health Rep, 1947.

[16] Schisterman E. F., Perkins N. J., Liu A., Bondell H., Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. Epidemiology. vol. 16 (1), pp. 73–81, 2005