IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 March 2024, accepted 6 April 2024, date of publication 9 April 2024, date of current version 17 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3386644

RESEARCH ARTICLE

A New Approach for Effective Medical Deepfake Detection in Medical Images

MEHMET KARAKÖSE, (Senior Member, IEEE), HASAN YETİŞ, AND MERT ÇEÇEN
Department of Computer Engineering, Fırat University, 23119 Elâzığ, Turkey

Corresponding author: Hasan Yetiş (h.yetis@firat.edu.tr)

This work was supported in part by The Scientific and Technological Research Council of Turkey (TUBİTAK) under Grant 122E676, and in part by the FIRAT University Scientific Research Projects Unit (FUBAP) under Grant MF.24.07.

ABSTRACT In today's world, deepfake technology is being used to generate fake images, sounds, and videos from real images and sounds using deep learning and artificial intelligence techniques. It is possible to manipulate medical images with this technology. The manipulation of medical images can lead to incorrect diagnoses by medical professionals, disrupting the functioning of hospitals. As a result of these disruptions, hospitals may experience significant financial and life-threatening problems. In this study, it is aimed to obtain an effective deep learning-based method to detect manipulated medical images. Initially, two distinct datasets are created which contain Knee Osteoarthritis X-ray and lung CT scans. Data pre-processing and augmentation methods are applied for data standardization and variation. The instances in datasets are labeled as real or fake. The medical deepfake distinguish ability of YoloV3, YoloV5nu, YoloV5su, YoloV8n, YoloV8s, YoloV8m, YoloV8l, YoloV8x models tested on these datasets. In the analysis performed, all YOLO models showed almost full success in distinguishing Knee Osteoarthritis X-ray images. In lung CT scan images, although YoloV8 models generally achieved good performance, the YoloV5 models gave the best and worst results. While the best result was obtained from YoloV5su with a recall value of 0.997, the worst result was obtained from the YoloV5nu model with a recall value of 0.91. Furthermore, the best model (YoloV5su) works 60% faster than YoloV8x model, which has the second highest performance. The findings show that YoloV5su can be used for fast and accurate medical deepfake detection.

INDEX TERMS Medical deepfake image detection, deep learning, YOLO, convolutional neural networks.

I. INTRODUCTION

The concept of deepfake images was first developed in 2014 by training Generative Adversarial Networks with large amounts of data [1]. In the following years, with better training of neural networks, they became a part of life and began to be used in many different areas [2], [3]. In addition to being frequently used to create scenes in movies or fake videos and images of famous individuals, producing fake images and videos of government officials in the field of politics, it has also begun to be used in the medical field [4]. The fact that Computerized Tomography devices in medical centers generally use outdated software and rarely receive software updates allows security problems to occur by accessing the software of these machines [5]. Placing fake tumors in the images obtained from the machines may cause unnecessary treatment in the hospital, causing financial damage to the hospital, as well as causing a large financial deficit by deceiving insurance companies by using fake tumor images [6]. Nowadays, especially with the development of Machine learning and Artificial Intelligence technology, it becomes very easy to add fake tumors to medical images using Generative Adversarial Networks. In order to make a correct diagnosis of the patient's condition, medical data can be shared between different expert personnel and diagnosed with mutual decisions [7]. The use of deepfake technology in images used in this field can lead to various abusive situations such as misdiagnosis by experts. In this field, it is of great importance to detect manipulated images, as the use of images manipulated using deepfake technology carries the risk of causing fatal situations [8]. In the literature, different studies have been carried out to detect fake images created in

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang.

medical data using Deepfake technology. Using the Ocular Disease Recreation dataset, Yong Suk Kim and colleagues asked three ophthalmologists and two general practitioners to verify the manipulated images. In this experiment, they used 100 manipulated and real data. As a result of the evaluations, ophthalmologists received 0.72 points and general practitioners received 0.67 points. They also used the U-Net model when training the model, they mentioned in their article and achieved a very successful result with 0.913 lesions, which is approximately 20% of the group OPT score. Approximately 6,000 data were scanned in the research. They used Sparse CNN, a deep learning network, to detect manipulated images in approximately 1000 data, of which approximately 350 were original data, approximately 200 were manipulated data, and 400 were original and manipulation data, created using Cycle GAN [8]. With the development of technology, it is becoming very difficult to understand the difference between real and deepfake content. Convolutional Neural Networks (CNN) have a very important place in medical image processing studies as well as being important in the field of image processing when looking at other deep learning algorithms [9]. Convolutional Neural Networks have been highly preferred by researchers in detecting fake content produced with Deepfake technology because they require less pre-processing than other deep learning algorithms and can easily learn filters and features after training. There are different types of lung cancer. In one of these types, a round tissue mass called a solitary pulmonary nodule is formed. If the diameter of these masses is less than 8 millimeters, they are benign, but if their diameter is larger than 8 millimeters, they may indicate a malignant cancerous growth. In addition, the presence of many nodules larger than 8 millimeters in diameter indicates that the patient's risk of first cancer increases [10]. Saleh Albahli and colleagues conducted a study to detect Lung CT Scan-based deepfake examples. In the study, the authors used a customized EfficientNet-V2 model with B4 as the base network with extra dense layers to identify pristine images as fake. In addition, in order to strengthen the results of the used model in the classification process, they added an AM policy that helps the network to focus on the manipulated areas of the samples and increases the recognition potential. By performing extensive experiments on the dataset they used, they achieved an average accuracy score of 85.49% in lung CT Scan deepfake detection of their approach [11].

In this study, different versions of YOLO algorithms were applied to detect deepfake medical images. The used YOLO versions are YoloV3, YoloV5nu, YoloV5su, YoloV8n, YoloV8s, YoloV8m, YoloV8l, and YoloV8x. Two different datasets were used, including Osteoarthritis X-ray images and lung CT images compiled from the literature. By examining the model outputs, analyses were conducted regarding the use of YOLO in counterfeit detection and, if so, which version to use. The contributions of the study are listed below:

- It is shown that the YOLO object recognition algorithm provides good results for medical deepfake images.
- New approaches to deepfake detection have been introduced to obtain fast and reliable results.
- The high accuracy achieved is important to secure the patient and the hospital.
- The work carried out is suitable for real-time application
- In the study, the performances of YOLO versions are given comparatively, especially for deepfake detection in medical images.

Rest of the paper is as follow. In Section II, the stages of the operations carried out in the studies carried out for the detection of fake images produced in medical data are explained in detail. Section III describes the methodology and workflow of the deep learning technology we used in this study. Section IV contains analyzes and explanations of the images of the experimental results we received as a result of training the model. In Section V, the results we reached in the study are explained.

II. MEDICAL DEEPFAKES

With the development of technology and the increasing use of deep learning methods, creating fake images has become easier, and image production has begun in different areas. One of these fields is the field of medicine. Different studies have been carried out in the literature on the detection of fake medical images created using deep learning methods and artificial neural networks. Fernandes and colleagues conducted a study on the use of Neural Ordinary Differential Equations (Neural ODEs) to estimate heart rate. Experiments conducted by the authors show that there is a significant difference between the heart rate of the original videos and the estimated heart rate of the depth fake videos, thus showing that real and fake videos can be distinguished by this method [12]. They conducted a study showing that the machine learning technology used by Google performed at a level equal to or superior to that of an ophthalmologist in diagnosing diabetic retinopathy using fundus images [13]. These studies can prevent such deceptions by detecting fake images that can deceive the insurance company in the medical data sent to insurance companies today. Although these images are not real diseases, they are used to obtain illegal health insurance [6]. Today, in different videos where deepfake technology is used, images of faces can be found in different locations and in areas with different environmental factors. In contrast to this situation, medical data can be obtained from patients using common machines under certain conditions to film certain areas. This makes medical images easier and more vulnerable to creating fake content. Systems that detect whether medical images are fake or not are likely to attract intense interest in the future. Convolutional Neural Networks [14], which have been available in the literature for a long time, are used to detect these images. Convolutional Neural Networks provide highly successful results in detecting objects, image processing, face recognition and analyzing videos [12]. Especially in biomedical images, the opportunity to access more images is more limited, as in other fields. For this reason, Generative

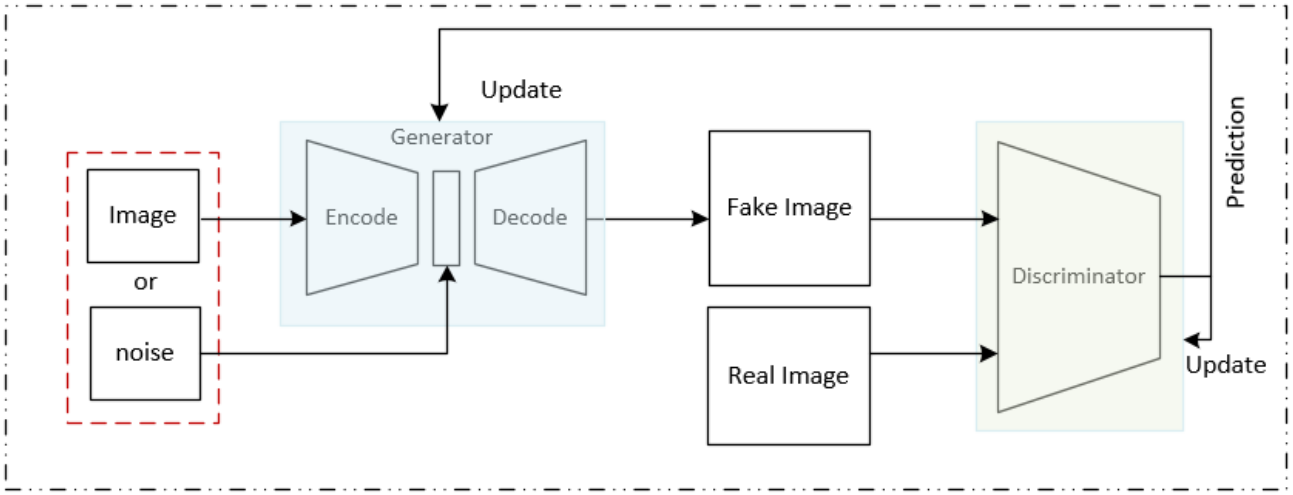

FIGURE 1. The generalized block diagram of noise-to-image and image-to-image GAN methods.

Adversarial Networks are used to increase the number of these images. Generative Adversarial Networks are used not only to reproduce images, but also to use and manipulate these images for malicious purposes [15].

A. CREATING DEEPFAKES

Today, with the advancement of artificial intelligence techniques, fake images have become a huge problem and have become a serious threat. Generative Adversarial Network is generally used to create these fake contents, which are the output of Deepfake technology, and to improve the image quality. The production of images created with Deepfake technology is suitable for use using Generative Adversarial Networks [16], [17]. Generative Adversarial Networks are very successful in making the created fake images more authentic, and with the development of these networks, the use of the created images in different areas has become widespread [14]. One of these fields was the field of medicine. In the field of medicine, they have carried out image generation in different studies using Generative Adversarial Networks. For example, Han and his colleagues conducted a study to obtain synthetic MR images using Generative Adversarial Network from brain MR images, and they stated that in the test performed on a specialist doctor, the doctor had difficulty distinguishing the images [18]. It is very difficult for researchers and students to access medical data due to confidentiality principles imposed by law. Generative Adversarial Networks are used in these areas and are also used to increase data sets. In addition, non-cancerous images can be shown as cancerous by manipulating real images using Generative Adversarial Networks. By performing the opposite process, factors that show symptoms of the disease in the images of sick individuals can be removed [5]. The GAN methods generally divided into noise-to-image and image-to-image methods. The general GAN architecture is given in Figure 1.

B. DEEPFAKE DETECTION

Generative Adversarial Networks are widely used in image synthesis. With the latest developments in Generative Adversarial Networks, we have entered a period where it is difficult for people to perceive whether the images are real or fake. Studies aimed at detecting this technology, which may pose great risks to humans, are also becoming widespread. Convolutional Neural Networks have also begun to be widely used in detecting fake images [19], [20]. Thanks to the stride parameter Convolution can be used in down-sampling images. With using them consecutive with pooling, CNN networks as feature extractors are obtained. The size of convoluted image is given in Eq. 1 and 2, where H is image height, W is image weight, pad is padding, K is convolution kernel, and S is stride [21].

$$H_{\text{out}} = \frac{H_{\text{in}} + (2 \times \text{pad}) - K_{\text{height}}}{S} \tag{1}$$

$$W_{\text{out}} = \frac{W_{\text{in}} + (2 \times \text{pad}) - K_{\text{width}}}{S} \tag{2}$$

Deepfake detection methods generally focus on the features of the image. Convolutional Neural Networks are generally used in the detections performed. Alexnet is a CNN architecture that performs substantially better than ImageNet in highly challenging object recognition [22], [23]. Convolutional Neural Networks has a structure that can automatically extract different features of the data. YOLO is an CNN based algorithm that can make predictions by learning the images it will detect at once and calculating bounding box coordinates and class probabilities [24], [25]. Although YOLO is widely used for object detection and classification, there are no study about using YOLO in deepfake detection.

III. THE PROPOSED METHOD

Fake images produced with deepfake technology, which has become very difficult to detect, are almost no different from real images. However, deepfake technology has begun to be

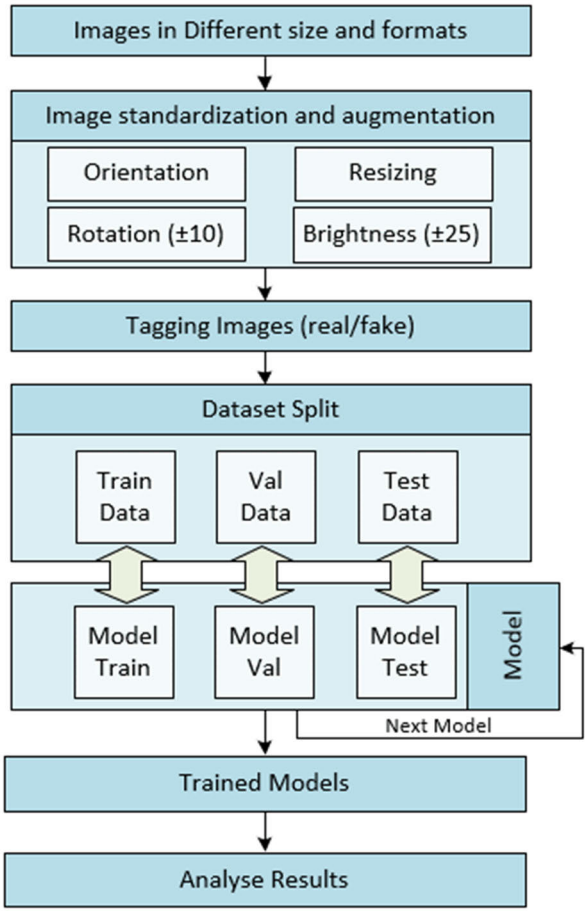

FIGURE 2. The block diagram of the proposed method.

used to produce fake images in different areas. One of these areas has been medical images.

In this study, it is aimed to obtain a fast and reliable model to detect medical fake images. First of all, because the lack of the medical datasets in literature we use several datasets in the literature for obtaining dataset. Two main datasets which consist of Osteoarthritis X-ray scans (dataset1) and lung CT scans (dataset2) are obtained. Image standardization and data augmentation techniques are applied to variate the datasets. The datasets split for using in train, validation and test stages. Model trainings are done using different YOLO versions. The used YOLO versions are YoloV3, YoloV5nu, YoloV5su, YoloV8n, YoloV8s, YoloV8m, YoloV8l, and YoloV8x. The model names end of the version name refers to network and parameter size of YOLO. For example, YoloV8n contains 3.2 million parameters, YoloV8x contains 68.2 million parameters. After training the models the results are analyzed. The performance comparisons are done. At the end, the model with the highest performance is obtained. The general block diagram of the proposed method is shown in Figure 2. In this section, the creation stages of the data set used in the study are explained, respectively. Then, the created model architecture and the model performance metrics are explained.

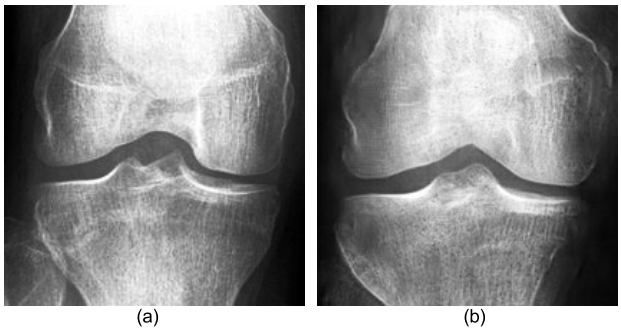

FIGURE 3. The sample images in dataset1. a) Real Image, b) Fake Image.

TABLE 1. The number of instances in dataset1 (osteoarthritis X-ray).

	Real	Fake	Total
Train	1500	1500	3000
Val	1500	1500	3000
Test	1500	1500	3000

A. DATASET

It is very difficult to access medical images because they are not often shared on the internet due to patients' personal rights. On the other hand, since studies on fraud detection generally focus on face replacement, the number of researchers working on producing deepfake medical datasets remains low. In this study, different data sets in the literature are used. The first dataset is the Osteoarthritis dataset. Prezja and his colleagues produced a total of 320,000 synthetic Osteoarthritis X-ray images with the GAN-based method they developed and made them available [26]. Since this data set only includes images produced by deepfake, real images are needed. A dataset for knee joint detection and knee KL (Kellgren and Lawrence) grading was prepared and made available by Chen [27]. In the study, original and deepfake X-ray images were obtained by taking data from these two distinct data sets. While the dataset containing real examples had different KL degrees, in the study, all classes were combined and labeled as ''real''. In this way, a data set that can be used to distinguish deepfakes. The number of dataset instances is given in Table 1. Sample real and fake X-ray images are given in the Figure 3.

Similarly, a dataset containing lung CT images was obtained with data collected from the literature. CT-GAN [6] is medical deepfake dataset viewed by radiologist. But there are limited fake images in the dataset. There are TB, TM, FB, FM classes in the deepfake datasets in the literature, and in the study, TB and TM are labeled as real, and FB and FM are labeled as fake images. Image examples of real and fake cases in this dataset are given in Figure 4. The lack of sufficient number of destroyed images in this dataset creates a disadvantage for our study. For this reason, data augmentation

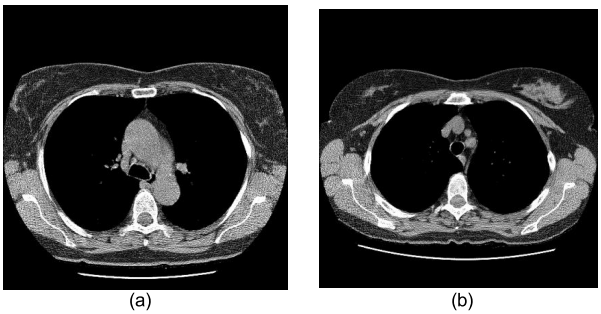

FIGURE 4. The sample images in dataset2. a) Real Image, b) Fake Image.

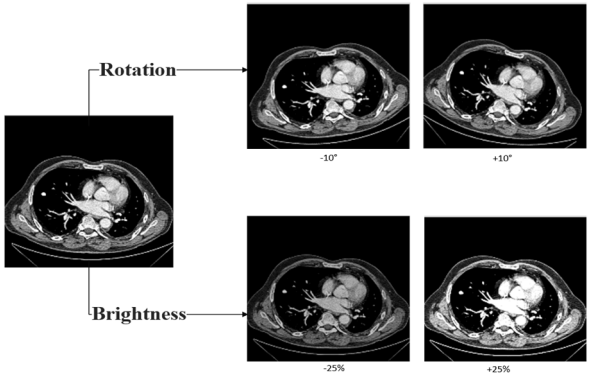

FIGURE 5. Studies carried out for data augmentation.

TABLE 2. The number of instances in dataset2 (lung CT scan).

	Real	Fake	Total
Train	6504	531	7035
Val	175	75	250
Test	175	75	250

techniques were used in the study, especially to increase the number of fake lung CT images.

In order to increase the number of images in the created data set and to better perceive the images, 10-degree rotation, 25 percent contrast tolerance adjustment and automatic orientation were applied to the images. Example images for data augmentation by setting brightness and rotating image are given in Figure 3. The number of samples in the data set obtained as a result of data augmentation techniques is given in Table 2.

B. DATA PRE-PROCESSING

Data collected from different sources may vary in size and color depth. Although the used algorithm works independently of the size, the data in the Osteoarthritis data set (dataset1) has been reduced to 210 × 210 in order to speed up the processes and to create a data standard. The reason for this is that the deepfake images used in the literature are of this size. On the other hand, Lung CT images can be found in DICOM (Digital Imaging and Communications in Medicine) format. The DICOM file format was developed by NEMA (National Electrical Manufacturers Association) to distribute and display MRI, CT scans and ultrasound images [28]. These medical images, stored in this format with the dcm extension, were first converted to JPEG format. The resulting images were labeled with the bounding box method using Roboflow. In order to maximize the learning efficiency of the learning model developed in the study, image resizing was performed to obtain the lung CT scan data (dataset2) to 416 × 416 size. Thanks to the image resizing process, the learning time of the deep learning model has been shortened.

C. EVALUATION METRICS

As with many object detection algorithms, precision, recall and mAP metrics are used to calculate the accuracy of the YOLO algorithm. First, Intersection over Union (IoU) is calculated between the detected image frame and the ground truth. A trust value (threshold) is determined. Frames above the confidence value (threshold) are considered valid, frames below are ignored. The classification performance is calculated based on the remaining frames. At this stage, confusion matrix is used. In the confusion matrix; True Positive: Predicted and actual class are positive; True Negative: Negative of the predicted and actual class; False Positive: Predicted class is positive, actual class is negative; False Negative: It is the number of cases where the predicted class is negative and the real class is positive. Recall value is the ratio of True positives to all true positives (TP/(TP+FN)). Prediction is calculated by the ratio of true positives to all positive predictions (TP/(TP+FP)). mAP value is mean average prediction and is the average of AP values of the classes. Here, mAP50 is the mAP value formed according to the configuration matrix when the confidence value is selected as 0.5. mAP50-95 is the average of mAP values occurring in the confidence value range of 0.5, 0.55, 0.65 …0.95, respectively.

The train values show the values obtained as a result of the measurement of the training set values, and the val values show the measurement results of the validation set values. The box_loss value focuses on measuring the loss rate values that will occur when labeling with a bounding box. The decrease in the values in the graph shows that the model has improved in terms of generalization and the data set is better labeled. The cls_los value indicates the measure of the loss rate resulting from the process of classifying images. A decrease in the value on the graph indicates a better classification.

IV. EXPERIMENTAL RESULTS

The proposed method is run in the Python environment. A computer with 64 GB RAM, RTX 4090 graphics card, Ryzen 9 7900X processor and Windows 11 was used. Training was carried out on the graphics card with CUDA. YoloV3, YoloV5nu, YoloV5su, YoloV8n, YoloV8s, YoloV8m, YoloV8l, YoloV8x versions were used in the training. The weights of the trained models are updated and stored. Stored weights help predict the class of data when new data arrives.

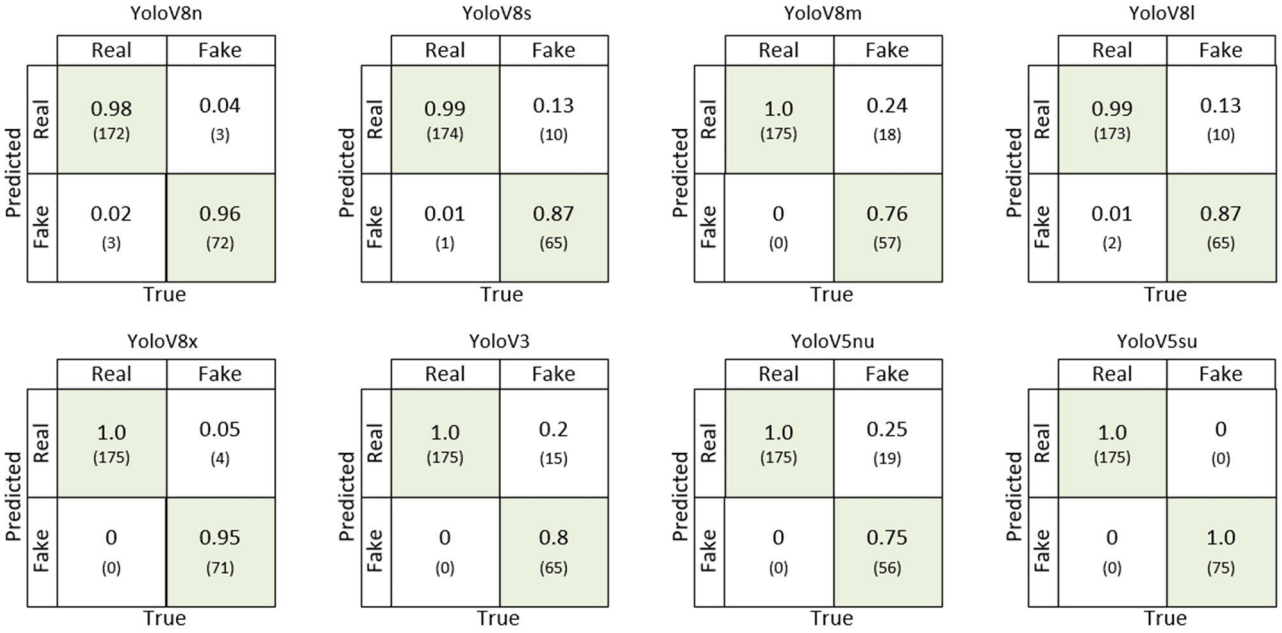

FIGURE 6. Confusion matrices for different YOLO versions on dataset2.

TABLE 3. The hyper-parameters used in models.

Parameter	Value
epochs	50
batch	16
imgsz	416
lr0	0.01
lrf (lr1)	0.01
iou	0.0005
momentum	0.937
warmup_epochs	3
warmup_momentum	0.8
warmup_bias_lr	0.1

As a result of the studies carried out in this section, the model creation process and performance tests were carried out. The hyper-parameters are used are given in Table 3.

First, training was carried out on dataset1. Approximately 100% success rate was achieved in 50 epochs in each YOLO version used. Although the fake Osteoarthritis X-Ray images in dataset1 seem difficult to distinguish, the fakes are easily detected with YOLO algorithms. On the other hand, it is more difficult to distinguish the CT lung scan images in dataset2. The confusion matrices that emerged when classifying medical images in dataset2 are given in Figure 6. The data included here is data that has not been used in training before and consists of 175 real and 75 fake images in total. In general, when the confusion matrices of the models are examined, the fake image is predicted to be real in very few cases for all models. In other words, predictions of images that are actually real are more accurate. The prediction range for actually fake images varies between 0.75 and 1. The disadvantage of models is that they are more likely to perceive fake images as real. However, it can be said that the YoloV8n, and YoloV8x and YoloV5su models give successful results. As can be seen, the YoloV5su model classified all images correctly. It can be said that the YoloV8m and YoloV5nu models are the least successful models. The average time taken for each epoch during the training phase, the total time taken to complete the training, Map50 value, Map50-95 value and Recall value are given in Table 4. Although it can be seen in this table that models with more parameters generally take more time to train, it can be seen that the YoloV5su model works faster than the YoloV5nu model, which has fewer parameters. The highest recall value belongs to YoloV5su with 0.997, followed by YoloV8x with 0.994. When examined in terms of running times, it is seen that the YoloV5su model, which is more successful in the classification, runs approximately 60% faster than the YoloV8x model, which performs closest to it. YoloV5nu shows one of the worst performances with 0.973 MAp50, 0.899 Map50-95, 0.91 recall value. From the results, it is interpreted that the YoloV5 architecture is better in detecting deepfakes, but the number of parameters is not sufficient in the YoloV5nu model.

The precision-recall curve, F1-confidence curve and recall-confidence curve graphs of the best model and the worst model are given comparatively in Figure 7. Giving only the best and worst cases allows performance metrics to be examined through these graphs. The graphs given belong to the same training and were carried out for 50 epochs, as in the results given. Sample training metrics for YoloV5su are given in Figure 8 for 50 epochs.

TABLE 4. Elapsed times and performance metrics of YOLO models.

	Avg. Epoch Time (sec)	Elapsed Time (hours)	MaP50 (all classes)	MaP50-95 (all classes)	Recall
YoloV8n	21	0,340	0.993	0.944	0.965
YoloV8s	21	0,334	0.987	0.945	0.972
YoloV8m	28	0,475	0.994	0.939	0.982
YoloV8l	36	0,555	0.993	0.942	0.963
YoloV8x	**50**	**0,741**	**0.995**	**0.952**	**0.994**
YoloV3	51	0,771	0.981	0.92	0.959
YoloV5nu	22	0,380	0.973	0.899	0.91
Yolov5su	**20**	**0.319**	**0.995**	**0.931**	**0.997**

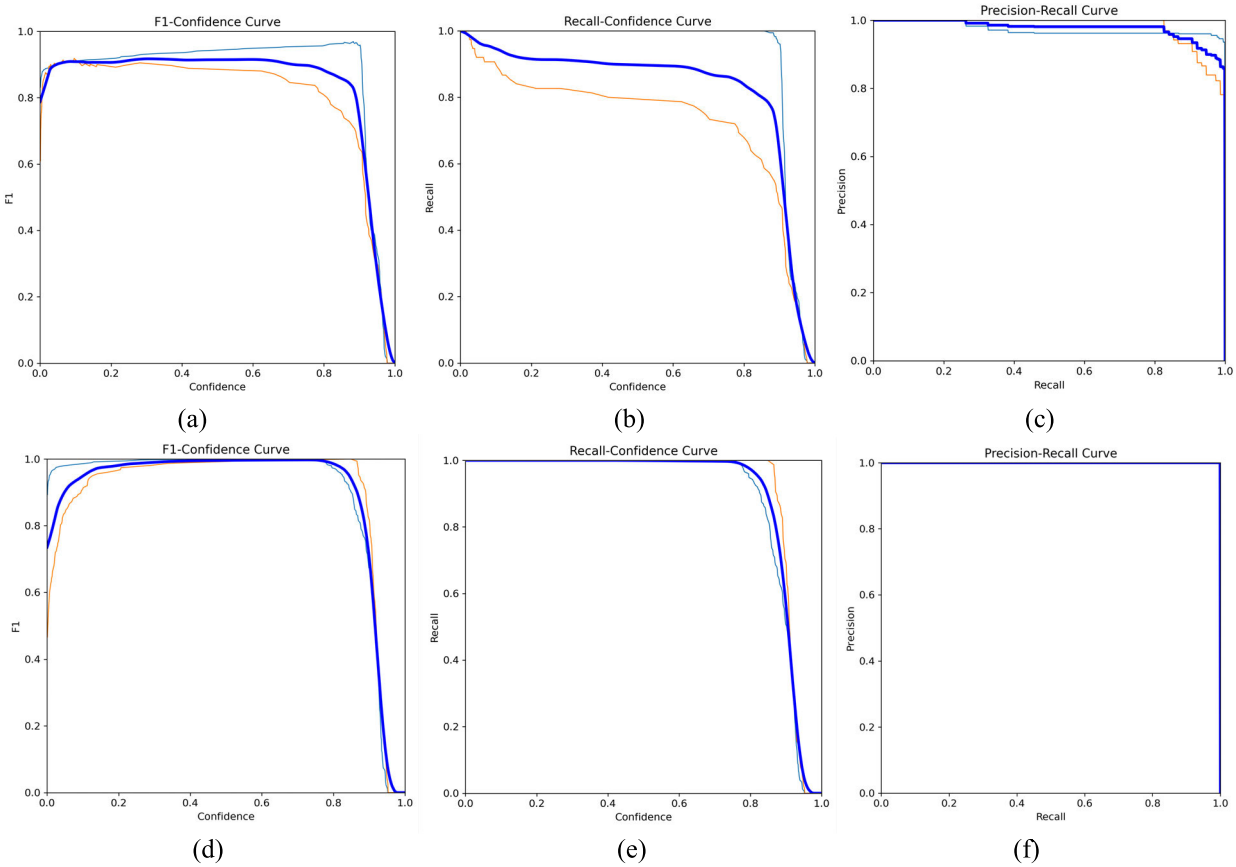

FIGURE 7. F1-confidence, Recall-confidence, precision-recall curves for worst (YoloV5nu) and best (YoloV5su) models. a) F1-confidence curve for worst model, b) Recall-confidence curve for worst model, c) Precision-recall curve for worst model, d) F1-confidence curve for best model, e) Recall-confidence curve for best model, f) Precision-recall curve for best model.

A. COMPARISON OF DEEP LEARNING STUDIES ON MEDICAL DEEPFAKE IMAGES

In the literature, different studies have been carried out using various data sets to detect medical deepfake images created using different deep learning methods and artificial neural networks. In this article, a table has been created showing the data used in studies in which medical images were detected using deep learning techniques and artificial neural networks, the images used in the methods applied to perform the detection, and the methods used in the detection processes. The explanation of the comparisons in Table 5 is detailed below.

In the first study in the table, Yong Suk Kim and his colleagues conducted an experiment using the Ocular Disease Recreation dataset. In the experiment, they used an image set consisting of 100 randomly manipulated images and real data. They used data that included four of the eight elements classified in the dataset they used: normal, glaucoma, diabetic retinopathy, and macular degeneration. As a result of studies evaluating the clinical effectiveness of manipulated data, oph-

TABLE 5. Comparison table with studies in the literature.

Reference	Images used in Medical Deepfake Detection	Methods used in Medical Deepfake Detection	Acc. / Recall
[8]	A dataset created from fake fundus and real fundus images manipulated with U-Net and Cycle GAN models was used.	Sparse CNN network, one of the deep learning networks, was used to detect the original fundus images and manipulated fundus images.	91%
[29]	They removed the parts with tumour regions in the images containing untampered or real CT scans from the LIDC-IDRI dataset and tampered CT scans from the CT-GAN dataset. They used this data set by obtaining a data set by performing enlargement, axis rotation and shifting operations on the obtained images.	In the study, they carried out comprehensive evaluations using DenseNet, ResNet, VGG and RC combinatorial architectures on two different image sets, which they created as a data set consisting of raw images and an evaluation set. In the tests carried out, a study was carried out on the use of CNN + RC, an ensemble-based architecture, to detect GAN-based tampering in medical images.	97.2%
[30]	The data set used by the authors in the study includes 3D Computed Tomography (CT) lung scans. He used a dataset containing medical deepfakes, in which some images were manipulated to inject fake cancer, and some images were tampered with to remove cancer.	The authors proposed an intelligent deep detection system for malicious tampering cancer to evaluate in distinguishing tampered and real data using machine learning algorithms and pre-trained DNN classifier.	93.19%
This Study	In the study, images created using a bi-conditional GAN (cGAN) network called CT-GAN, which was used to place pseudo nodules on nodule-free 3D CT lung images, were collected. These collected fake 3D CT lung images were combined with real nodule-containing CT lung images and a data set was obtained by applying various pre-processing steps and this data set was used.	Different versions of Yolo were trained to detect deepfake images. It has been observed that a method that can provide fast results such as YOLO can be used to detect medical deepfakes. The result of YOLOv5su of the best model has been reached.	99.7%

thalmologists received 0.72 points and general practitioners received 0.67 points. While only fundus images of the right eye were used among approximately 6,000 data in the study, the images were resized to 256 × 256. Sparse CNN structure, a deep learning network, was used to detect manipulated images in approximately 1000 data, of which approximately 350 were original data, approximately 200 were manipulated data, and 400 were original and manipulation data created using Cycle GAN. As a result of the experiments they conducted in the study, the authors found the average detection ability to be 91%. In the second study in the table, Budhiraja and colleagues used images containing untampered or real CT scans from the LIDC-IDRI dataset and tampered CT scans from the CT-GAN dataset. They extracted a 128 × 128 section limited to only one tumor region location. Data augmentation was achieved by enlarging the images, rotating and shifting them on the axis. They carried out the necessary development and testing procedures for the training and detected tampering on the original images and the tampered localized images. They achieved very successful results with the feature extraction-based classifiers they tested by performing comprehensive evaluations using DenseNet, ResNet, VGG and RC combinatorial architectures. For CNN+RC, an ensemble-based architecture, they observed a 90% relative increase in accuracy values in the images they detected using the raw images. They achieved a result accuracy of 97.2% with the combination of DenseNet201 and one-way RC. As a result of tamper detection classification for localized tumor regions and nodule locations in the second dataset they used in their evaluation, they achieved a 92.4% relative accuracy increase over DenseNet121-based classification for single localized images from the DenseNet121 and RC-based combination. In the third study in the table, the authors used a dataset containing medical deepfakes, in which some images of 3D Computed Tomography (CT) lung scans were manipulated and injected with fake cancer, and in some images the cancer was removed. They tried to improve the system detection performance by reducing the number of false alarms and increasing the detection rate. The aim of the DNN classifier-based method they propose is to reduce false alarm and error rates and increase detection rates. In their proposed model, the total detection accuracy rate was 93.19%, the total error rate was 6.70% and the percentage of total false alarms was 9.10%.

V. DISCUSSION

In this study, YOLO-based model training was carried out for effective and fast classification of medical fake images. With the trained models, a method with higher accuracy was obtained compared to the methods in the literature. Although studies in the literature do not provide information in terms of training times, it is known that YOLO-based algorithms provide an advantage in terms of speed compared to other CNN-based algorithms.

Two different data sets were used in the study. More successful results were obtained as a result of the training process on the first data set consisting of fake and real Osteoarthritis images. The data in the second dataset, which includes Lung CT images, is relatively small and was produced with the image-to-image technique using the CT-GAN method. It is seen that the method achieves relatively lower performance in the second data set. On the other hand, the first dataset

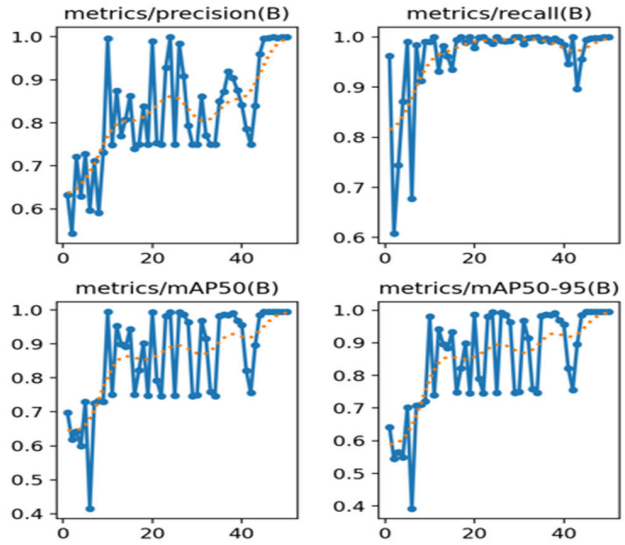

FIGURE 8. Training metrics for YoloV5su on dataset2.

obtains deepfake images from scratch, and the second dataset is obtained by adding and removing cancerous areas on existing images. It has been concluded that deepfake detection methods are more successful in detecting deepfake images produced with noise-to-image. Medical images are not limited to the data sets used, and the method must be trained with these data in order to work on different medical images. The rapid training process makes training different types of medical data practical.

Since the method performs rapid training and testing, it is possible to apply medical fake images in real time with the proposed method. The method works for known image formats such as jpg and png, but requires conversion from the dicom format where medical data is usually stored. By performing the conversion process with a middleware, this problem will not pose a problem for real world applications.

VI. CONCLUSION

Nowadays, deep learning methods and artificial neural networks are highly developed. With the development of Deepfake technology, which is used to produce fake images, audio and video using deep learning algorithms and artificial neural networks, the process of creating manipulated images in different fields has become widespread. One of these fields is the field of medicine. Accessing people's medical data and manipulating these data has also begun to be carried out, although this is not common. Manipulation of images obtained from devices that can capture images of patients, such as computerized tomography (CT) devices and MRI devices, causes a significant threat to people's private areas. Studies carried out to prevent these fake images from being used for malicious purposes, such as deceiving insurance companies and portraying non-patients as patients, have become increasingly important as fake manipulated images increase. Although the use of fake data produced by deepfake technology, which uses deep learning and artificial neural networks, has risky aspects, these situations can be prevented by detecting these images. While carrying out this study, studies created using Deepfake were investigated in detail and a model that can detect fake medical images was created.

First of all, Knee Osteoarthritis X-Ray and lung CT scan images were collected and 2 separate data sets were obtained. Data augmentation and pre-processing methods are applied for data standardization and variation. Then, these images were labeled as fake and real and a data set was created. Different YOLO models are trained on the datasets. YoloV3, YoloV5nu, YoloV5su, YoloV8n, YoloV8s, YoloV8m, YoloV8l, and YoloV8x models are used. In the analysis performed, all YOLO models showed almost full success in distinguishing Knee Osteoarthritis X-Ray images. In lung CT scan images, although YoloV8 models generally achieved good performance, the YoloV5 models gave the best and worst results. While the best result was obtained from YoloV5su with a recall value of 0.997, the worst result was obtained from the YoloV5nu model with a recall value of 0.91. Furthermore, the best model (YoloV5su) works 60% faster than YoloV8x model, which has the second highest performance. This situation shows that an effective approach has been proposed in detecting deepfake images when compared to studies in the literature. However, as deepfake techniques are constantly evolving and the model misperceives some images in the studies carried out, it is of great importance to constantly make efforts to update the model and recognize new deepfake methods. In future studies, it is aimed to expand the data set and obtain better results by integrating new technologies into our model. In this context, it is aimed to create different medical deepfake images and add them to the data set in order to enlarge the data set.

REFERENCES

[1] I. J. Goodfellow, "Generative adversarial nets," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, pp. 2672–2680, 2014.
[2] İ. İlhan and M. Karaköse, "Derin sahte videoların tespiti ve uygulamaları için bir karşılaştırma Çalışması," *Adıyaman Üniversitesi Mühendislik Bilimleri Dergisi*, vol. 8, no. 14, pp. 47–60, Jun. 2021.
[3] I. Ilhan, E. Bali, and M. Karaköse, "An improved DeepFake detection approach with NASNetLarge CNN," in *Proc. Int. Conf. Data Analytics Bus. Ind. (ICDABI)*, Oct. 2022, pp. 598–602.
[4] S. Solaiyappan and Y. Wen, "Machine learning based medical image deepfake detection: A comparative study," *Mach. Learn. Appl.*, vol. 8, Jun. 2022, Art. no. 100298.
[5] J. E. Dunn. (2018). *Imagine You're Having a CT Scan and Malware Alters the Radiation Levels—It's Doable the Register*. [Online]. Available: https://www.theregister.co.uk/2018/04/11/hacking_medical_devices/
[6] Y. Mirsky, T. Mahler, I. Shelef, and Y. Elovici, "CT-GAN: Malicious tampering of 3D medical imagery using deep learning," in *Proc. 28th USENIX Secur. Symp.*, Jan. 2019, pp. 461–478.
[7] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221–248, Jun. 2017.
[8] Y. S. Kim, H. J. Song, and J. H. Han, "A study on the development of deepfake-based deep learning algorithm for the detection of medical data manipulation," *Webology*, vol. 19, no. 1, pp. 4396–4409, Jan. 2022.
[9] A. G. Eker and N. Duru, "Deep learning applications in medical image processing," *Acta Infologica*, vol. 5, no. 2, pp. 459–474, 2021, doi: 10.26650/acin.927561.
[10] H. MacMahon, D. P. Naidich, J. M. Goo, K. S. Lee, A. N. Leung, J. R. Mayo, A. C. Mehta, U. Ohno, C. A. Powell, M. Prokop, G. D. Rubin, C. M. Schaefer-Prokop, W. D. Travis, P. E. V. Schil, and A. A. Bankier, "Guidelines for management of incidental pulmonary nodules detected on Ct images: from the Fleischner society," *Radiology*, vol. 284, no. 1, pp. 228–243, 2017.

[11] S. Albahli and M. Nawaz, "MedNet: Medical deepfakes detection using an improved deep learning approach," *Multimedia Tools Appl.*, vol. 2023, pp. 1–19, Nov. 2023, doi: 10.1007/s11042-023-17562-5.
[12] S. Fernandes, S. Raj, E. Ortiz, I. Vintila, M. Salter, G. Urosevic, and S. Jha, "Predicting heart rate variations of deepfake videos using neural ODE," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1721–1729.
[13] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, p. 2402, Dec. 2016.
[14] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989.
[15] M. Sharafudeen and S. S. V. Chandra, "Medical deepfake detection using 3-dimensional neural learning," in *Lecture Notes in Computer Science (Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13739. Deutschland, Germany: Springer, 2023, pp. 169–180.
[16] K. Barrera, A. Merino, A. Molina, and J. Rodellar, "Automatic generation of artificial images of leukocytes and leukemic cells using generative adversarial networks (syntheticcellgan)," *Comput. Methods Programs Biomed.*, vol. 229, Feb. 2023, Art. no. 107314.
[17] K. Barrera, J. Rodellar, S. Alférez, and A. Merino, "Automatic normalized digital color staining in the recognition of abnormal blood cells using generative adversarial networks," *Comput. Methods Programs Biomed.*, vol. 240, Oct. 2023, Art. no. 107629.
[18] C. Han, H. Hayashi, L. Rundo, R. Araki, W. Shimoda, S. Muramatsu, Y. Furukawa, G. Mauri, and H. Nakayama, "GAN-based synthetic brain MR image generation," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 734–738.
[19] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.
[20] Y. Aslam and N. Santhi, "A review of deep learning approaches for image analysis," in *Proc. Int. Conf. Smart Syst. Inventive Technol. (ICSSIT)*, Nov. 2019, pp. 709–714, doi: 10.1109/ICSSIT46314.2019.8987922.
[21] J. Du, "Understanding of object detection based on CNN family and Yolo," *J. Phys., Conf. Ser.*, vol. 1004, Apr. 2018, Art. no. 012029.
[22] A. Krizhevsky, I. Sutskever, and E. H. Geoffrey, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1–9.
[23] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. F. Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[24] S. K. Rajput, J. C. Patni, S. S. Alshamrani, V. Chaudhari, A. Dumka, R. Singh, M. Rashid, A. Gehlot, and A. S. AlGhamdi, "Automatic vehicle identification and classification model using the YOLOv3 algorithm for a toll management system," *Sustainability*, vol. 14, no. 15, p. 9163, Jul. 2022.
[25] S. V. Viraktamath, M. Yavagal, and R. Byahatti, "Object detection and classification using YOLOv3," *Int. J. Eng. Res. Technol.*, vol. 10, no. 2, pp. 1–6, 2021.
[26] F. Prezja, J. Paloneva, I. Pölönen, E. Niinimäki, and S. Äyrämö, "DeepFake knee osteoarthritis X-rays from generative adversarial neural networks deceive medical experts and offer augmentation potential to automatic classification," *Sci. Rep.*, vol. 12, no. 1, p. 18573, Nov. 2022.
[27] P. Chen. (2018). *Knee Osteoarthritis Severity Grading Dataset*. Mendeley Data, V1. [Online]. Available: https://www.10.17632/56rmx5bjcr.1
[28] M. Santosa and N. P. Rochab, "A big data approach to explore medical imaging repositories based on DICOM," in *Proc. Centeris Int. Conf. Enterprise Inf. Syst./ProjMAN Int. Conf. Project Manag./HCist Int. Conf. Health Social Care Inf. Syst. Technol.*, 2022, pp. 1–9.
[29] R. Budhiraja, M. Kumar, M. K. Das, A. S. Bafila, and S. Singh, "MeDiFakeD: Medical deepfake detection using convolutional reservoir networks," in *Proc. IEEE Global Conf. Comput., Power Commun. Technol. (GlobConPT)*, India Habitat Centre, Lodhi Road, New Delhi, India, Sep. 2022, pp. 1–6.
[30] K. M. A. Alheeti, A. Alzahrani, N. Khoshnaw, and D. Al-Dosary, "Intelligent deep detection method for malicious tampering of cancer imagery," in *Proc. 7th Int. Conf. Data Sci. Mach. Learn. Appl. (CDMA)*, Mar. 2022, pp. 25–28.

MEHMET KARAKÖSE (Senior Member, IEEE) received the B.S. degree in electrical engineering and the M.S. and Ph.D. degrees in computer engineering from Fırat University, Elâzığ, Turkey, in 1998, 2001, and 2005, respectively.

From 1999 to 2005, he was a Research Assistant with the Department of Computer Engineering, Fırat University. He was an Assistant Professor and Associate Professor with the Department of Computer Engineering, Fırat University, from 2005 to 2014 and from 2014 to 2020, respectively. He is currently a Professor Doctor with the Department of Computer Engineering, Fırat University. His research interests include fuzzy systems, intelligent systems, quantum computing, simulation and modeling, fault diagnosis, computer vision, railway inspection systems, and photovoltaic systems.

HASAN YETİŞ received the B.S. degree in computer engineering from Inonu University, Malatya, Turkey, in 2014, and the M.S. and Ph.D. degrees in computer engineering from Fırat University, Elâzığ, Turkey, in 2017 and 2022, respectively.

From 2015 to 2022, he was a Research Assistant with the Department of Computer Engineering, Fırat University. He is currently an Assistant Professor with the Department of Computer Engineering, Fırat University. His research interests include quantum computing, artificial intelligence, optimization, and computer vision.

MERT ÇEÇEN was born in Elâzığ, Turkey, in 2000. He received the B.S. degree in computer engineering from Fırat University, Elâzığ, where he is currently pursuing the master's degree with the Department of Computer Engineering. His research interests include deep learning and artificial intelligence.