Comparison of Deep Learning Models in Carotid Artery Intima-Media Thickness Ultrasound Images: CAIMTUSNet

Araştırma Makalesi/Research Article

Serkan SAVAŞ[1], Nurettin TOPALOĞLU[2], Ömer KAZCI[3], Pınar NERCİS KOŞAR[4]

[1]Department of Computer Engineering, Çankırı Karatekin University, Çankırı, Turkey
[2]Department of Computer Engineering, Gazi University, Ankara, Turkey
[3]Department of Radiology, VM Medical Park Ankara Hospital, Ankara, Turkey
[4]Department of Radiology, Ankara Training and Research Hospital, Ankara, Turkey
serkansavas@karatekin.edu.tr, nurettin@gazi.edu.tr, omerkazci1990@gmail.com, p.kosar@hotmail.com
(Geliş/Received:24.10.2020; Kabul/Accepted:05.11.2021)
DOI: 10.17671/gazibtd.804617

Abstract- Deep learning is a machine learning technique that uses deep neural networks, which are multilayer neural networks that contain two or more hidden layers. In recent years, deep learning algorithms are also used to solve machine learning problems in medicine. Carotid artery disease is a type of cardiovascular disease that can result in a stroke. If a stroke is not diagnosed early, it is in the first place among the disabling diseases and the third place for the most common cause of death after cancer and heart disease. In this study, the classification performances of deep learning architectures in the biomedical field are compared, and Carotid Artery (CA) Intima-Media Thickness (IMT) Ultrasound (US) images were used. For an early diagnosis, AlexNet, ZFNet, VGGNet (16-19), which had successful results in the ImageNet competition, and authors' original CNNcc models were used for comparison. An image database of CA-IMT-US which contains 501 ultrasound images from 153 patients was used to test the models' classification performances. It is seen that AlexNet, ZFNet, VGG16, VGG19, and CNNcc models achieved rates of 91%, 89.1%, 93%, 90%, and 89.1% respectively. The CNNcc model was found to produce successful classification results on CAIMTUS images when different performance indicators are also taken into account. In addition, different performance indicators including confusion matrices were investigated and the results were announced. The results showed that deep architectures are promising in the biomedical field and can provide proper classification on biomedical images so; this can help clinics to diagnose the disease early.

Keywords- Deep learning, image processing, carotid artery, intima-media thickness, convolutional neural network, machine learning.

Karotis Arter Intima-Medya Kalınlığı Ultrason Görüntülerinde Derin Öğrenme Modellerinin Karşılaştırılması: CAIMTUSNet

Özet- Derin öğrenme, iki veya daha fazla gizli katman içeren çok katmanlı sinir ağları olan derin sinir ağlarını kullanan bir makine öğrenimi tekniğidir. Son yıllarda tıpta makine öğrenimi problemlerini çözmek için derin öğrenme algoritmaları da kullanılmaktadır. Karotis arter hastalığı, felçle sonuçlanabilen bir tür kardiyovasküler hastalıktır. İnme erken teşhis edilmezse, sakatlayıcı hastalıklar arasında ilk sırada, kanser ve kalp hastalıklarından sonra en sık ölüm nedeni olarak üçüncü sırada yer almaktadır. Bu çalışmada, derin öğrenme mimarilerinin biyomedikal alandaki sınıflandırma performansları karşılaştırılmış ve Karotis Arter (KA) Intima Media Thickness (IMT) Ultrason (US) görüntüleri kullanılmıştır. Erken teşhis için, ImageNet yarışmasında başarılı sonuçlar alan AlexNet, ZFNet, VGGNet (16-19) ve karşılaştırma için yazarların özgün CNNcc modelleri kullanılmıştır. 153 hastadan 501 US görüntüsünü içeren bir KA-IMT-US görüntü veritabanı, modellerin sınıflandırma performanslarını test etmek için kullanılmıştır. AlexNet, ZFNet, VGG16, VGG19 ve CNNcc modellerinin sırasıyla %91,%89.1, %93, %90 ve %89.1 oranlarına ulaştığı görülmüştür. CNNcc modelinin, farklı performans göstergeleri de hesaba katıldığında KAIMTUS görüntüleri üzerinde başarılı sınıflandırma sonuçları ürettiği bulunmuştur. Ayrıca çalışmada karışıklık matrislerini de içeren farklı performans göstergeleri incelenmiş ve sonuçlar açıklanmıştır. Sonuçlar, derin mimarilerin biyomedikal alanda ümit verici olduğunu ve biyomedikal görüntülerde uygun sınıflandırma sağlayabileceğini göstermiştir ki bu, kliniklerin hastalıkları erken teşhis etmesine yardımcı olabilir.

Anahtar Kelimeler- Derin öğrenme, görüntü işleme, Karotis arter, intima-media kalınlığı, evrişimli sinir ağları, makine öğrenmesi.

1. INTRODUCTION

The success achieved by the classification made by Krizhevsky and colleagues in the world's most important object recognition competition named ImageNet in 2012 via AlexNet's Deep Convolutional Artificial Neural Networks [1] has been the greatest impact of Deep Learning (DL) in the literature. After this success, DL began to use in different disciplines. Originally, DL was first introduced in 2006 in a method called Deep Belief Nets [2] by Hinton and colleagues.

In recent years, the techniques developed in DL research have influenced a wide range of information processing, both in traditional and new forms, in expanded contexts, including the most effective and important aspects of Machine Learning (ML) and Artificial Intelligence (AI). DL is a sub-field of ML and Deep Neural Networks (DNN) have become a common application field. In DL, the algorithm learns from the data itself and covers wider data set of solutions although ML solutions are covered by specified algorithms for each problem. DL is a promising approach to solving AI problems in ML.

DL is an ML technique that uses DNNs which are multilayer Neural Networks (NN) that contain two or more hidden layers [3]. DL is mainly based on learning from the representation of data. Representation for an image may be considered as a vector of density values per pixel, or features such as edge clusters, custom shapes, etc. Some of these features represent data better. As an advantage at this stage, DL methods use effective algorithms for hierarchical feature extraction that best represent data rather than manually extracted features [4]. DNNs consist of multi-hidden layers and they construct more interrelationships along with the complexity of the data. Every layer creates a relationship between itself and the layer before. So every input becomes more detailed and the network's accuracy becomes more. While constructing the DNNs, activation functions are used to generate an output due to the input value of the cell. Mainly, Sigmoid, TanH, and ReLU functions are used and it depends on different aspects of the data [5].

To run DL algorithms and solve problems, high-capacity machines, especially GPU, and large amounts of data are needed. Unlike usual ML algorithms that break down and solve problems individually, the problem is solved from start to end in DL. More importantly, the more data a DL algorithm feeds, the better the task will be accomplished. The time factor is also important and time limitless studies can produce better results when fed with large data. The major reasons for the increasing importance of DL today are the greatly increased processor capabilities, the vast increase in data used for training, and the latest developments in ML and signal/information processing researches. In addition to these, deep architectures such as AlexNet, ZFNet, ResNet, GoogLeNet, VGG16-19, Inception, etc., the introduction of DL platforms, and libraries such as Keras, Tensorflow, Caffe2, Pytorch, MatConvNet, etc., activation functions, data training, and factors such as data enhancement methods, and the development and use of effective optimizers by researchers have increased interest in deep architectures.

Many researchers have focused on DL architectures because of the success of CNN structures in object recognition [6]. Today, DL algorithms are also used in medicine to solving a wide range of ML problems. Segmentation of abdominal multi-organ, microscope, and biomedical images, metastatic breast cancer [7-10], detection of mitosis in breast cancer histology images [11], diagnosis of diabetic retinal fundus images [12], precision of pulmonary nodule detection [13], detection and diagnosis of seizure with encephalogram signals [14, 15], brain tumors classification [16], Alzheimer's disease detect and diagnosis, [17-20] and steganalysis on medical images [21] are some of these studies.

There are various reasons for the widespread use of DL architectures in the field of health in the last decade. The foremost of these; by replacing existing baseline assessments with highly accurate and reproducible measurements from related studies, image processing routines that can automatically analyze medical images have enormous potential value for improved diagnosis, treatment planning, and follow-up of individual patients [22]. However, in biomedical image analysis; while many problems prevent consistent results such as extensive research, creation of datasets, evaluation criteria of dependent and independent experts, DL methods prove that they partially overcome them [23]. In addition, while performing all these, it offers classification and segmentation solutions for diagnosis and treatments with state-of-art algorithms and techniques within DL [24].

In this study, the performance of DL architectures in the biomedical field is compared on Carotid artery (CA) Intima-Media Thickness (IMT) ultrasound (US) images. Since different imaging techniques and different types of images are used in the studies carried out in the field of medicine, it has not yet reached the stage of producing solutions with common models and algorithms. However, the architectures created in the ImageNet competition can classify from thousands of different image types. The deep aspects of deep architectures that are open to development in this area will be an important area for researchers soon.

Carotid Artery (jugular vein) is the first vein that separates from the large vein that emerges from the heart. The human brain is fed with clean blood from the CA passing through both sides of the neck. CA disease is an atherosclerotic disease, which can result in a stroke with a sudden blockage in the vessels [25]. Several factors accelerate a natural process of aging in all vessels of the body, causing some arteries to contract and block at an early age. Accumulation in the artery wall because of oil particles, cholesterol, and some different particles causes plaque formation. The volume of these plaques increases in time and creates a clot on the surface of the wall. Thus, the vessel is completely blocked. When the complete blockage

of the CA happens or small vessels of the brain are blocked by small lime particles or small clots detached from this atherosclerotic plaque, stroke occurs.

Stroke is the first in the list of disabling diseases if it is not diagnosed and treated early [26]. In the ranking of the diseases that cause the most common deaths, it comes third after cancer and heart diseases [27]. Approximately 16 million people have a stroke each year in the world [28]. The number of deaths due to Cerebrovascular Diseases (CVD) caused by stroke in Turkey has been identified as 64780 at the beginning of the 2000s [29]. Internal Carotid Artery stenosis is one of the most important causes of ischemic cerebral palsy [30].

Early detection and treatment of disease with such a high risk of death and/or disability with state-of-the-art methods and techniques is of vital importance. The last stage of AI studies applied in medicine today is DL techniques. For this reason, the performance of various DL models determined in the study on CAIMTUS images was compared.

The rest of the paper is organized as follows. In the second section of this study, the methodology and materials used in the study were described. The architectures and layers of the models, the database in which these models are used, and the image pre-processing operations performed on the images in this database are presented in this section. In the third section, the experimental results obtained were described. Each of the models' measurements was evaluated by accuracy and loss rates and the resulting ratios were interpreted. In addition, comparative rates and graphics of models were also indicated. To ensure the reliability of the results, confusion matrix values and ROC curve values are also examined in this section. The fourth section of the paper gives the main analysis and the discussion for the results. The last section explains the general results and contribution of the study to both clinic types of research and field studies.

2. METHODOLOGY AND MATERIALS

Deep learning architectures, which have become even more competitive with the worldwide competition ImageNet, are increasing every year with varying layer numbers and performance rates. The success of the AlexNet architecture in this competition has greatly increased the interest in DL. In this study, performances of AlexNet, ZFNet, VGG16-VGG19, and CNNcc (Convolutional Neural Network Carotid Classifier), models in classifying CA IMT US images were compared (CAIMTUSNet).

2.1. Models

AlexNet is the deep Convolutional Neural Network (CNN) for image classification winning the ILSVRC-2012 competition. The first five layers are convolutional, the last three are fully connected layers. These include “pooling” and “activation” layers. There are also input and output layers. The AlexNet architecture was designed to classify 1000 objects and the error rate for object identification was reduced from 26.2% to 15.3% [1]. The model architecture has a softmax layer that is employed to put the samples to the proper class. Stochastic gradient descent with momentum is used for minimizing the cost function in AlexNet architecture [31].

After AlexNet won the ImageNet competition, ZFNet [32] inspired by this architecture, became the winner of the ImageNet competition in 2013. With this architecture, the error rate in object recognition has been reduced to 11.2%. The difference from AlexNet; filter size is 7x7 and the number of steps is two. Here, a smaller filter size in the first convolution layer is intended to help preserve many original pixel information at the input size. In addition, "Cross-Entropy," "Stochastic Gradient Descent,” and “ReLU” algorithms are used in this architecture and it consists of seven layers.

There are two different types of VGGNet architecture: 16 and 19 layers; VGG16, VGG19. VGG16 architecture consists of 13 convolutions 3 fully connected layers used for better results in ImageNet 2014 competition [33]. There are a total of 41 layers with Max Pooling, Full-Connected Layer, ReLU Layer, Drop Out Layer, and Softmax Layer. The image format, which is placed in the input layer, is 224x224x3. The last layer is the classification layer [34]. The VGGNet architecture uses a 3x3 filter on all layers and uses the Convolution-ReLU layers one above the other before the Pooling layer. As in other deep architectures, VGG architecture decreases the height and width dimensions of the matrices from the input layer to the exit, while the depth value increases. It achieved a top-5 error rate of 7.3% in 2014.

In addition to these models, a DL model prepared for classification on biomedical images was used in the study. While the model uses 3x3 filters in Convolution layers, the first ten activation functions are selected as ReLU. The last activation function is the Softmax activation function. The filter size of the Pooling layers was determined as 2x2 and the Drop Out ratio was determined as 0.5. There are three Fully Connected layers and the first of these layers gives 256 outputs. The second one gives 128 outputs and the last one produces 2 outputs, which is basically, class number of images. The diagram of the Carotid Classifier Deep Learning Model (CNNcc) with CNN [5] created by the authors is shown in Figure 1.

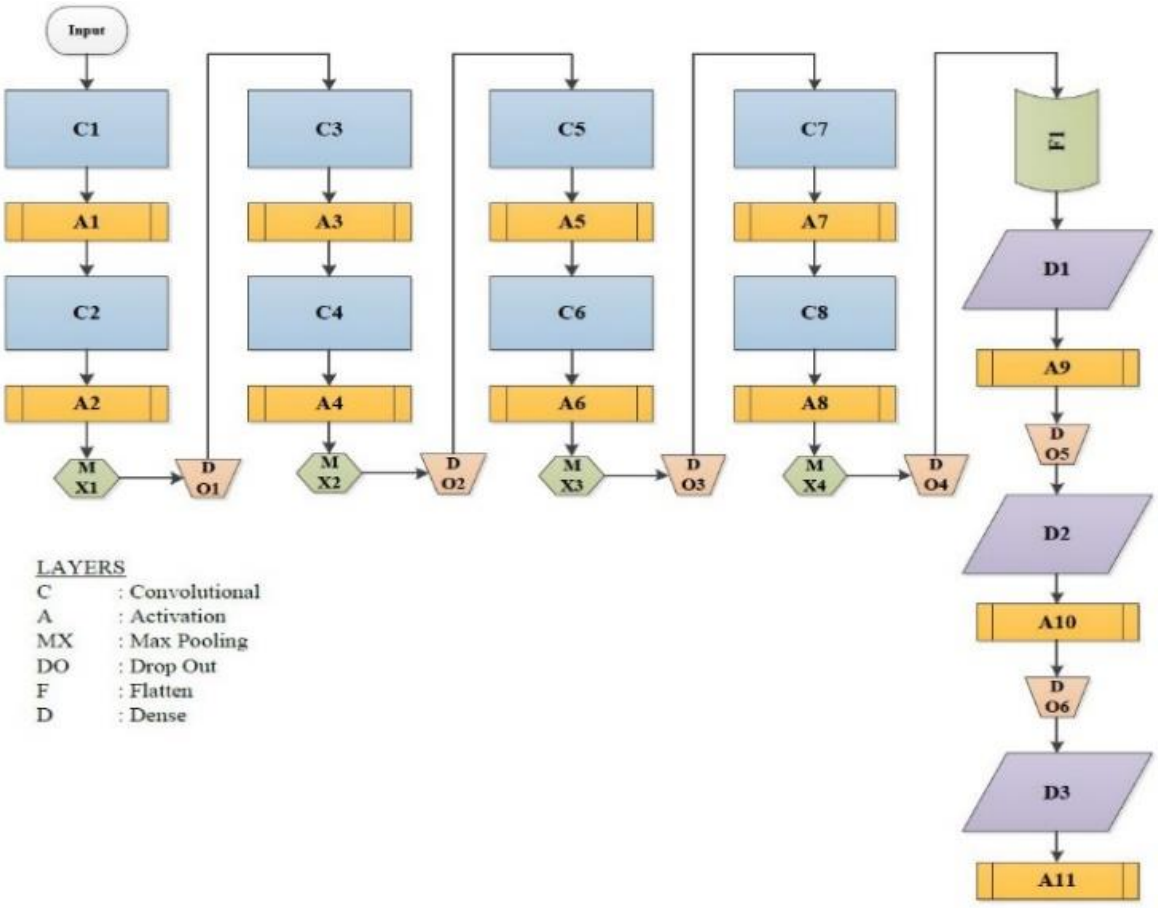


Figure 1. The layer structure of the CNNcc model

2.2. Database

For the study, 501 CA IMT US images were obtained from 153 patients of the Radiology Clinic of Ankara Training and Research Hospital between June 2018 - January 2019. 08/05/2018 dated and 2018-217 numbered Gazi University Ethics Commission's Ethics Approval Certificate was used to take the images. For US imaging, Toshiba Aplio 400 US device was used. Specialized two radiology department doctors of the hospital defined the images for the classification. The classes of images are defined as "IMT: 1" and "IMT: 0". "IMT: 1" images are the patients with signs of narrowing of the vein and "IMT: 0" images are the patients, which have veins without narrowing. There are 203 images in the "IMT: 1" class and 298 images in the "IMT: 0" class in the CA IMT US images database. Image samples from both classes are shown in Figure 2.

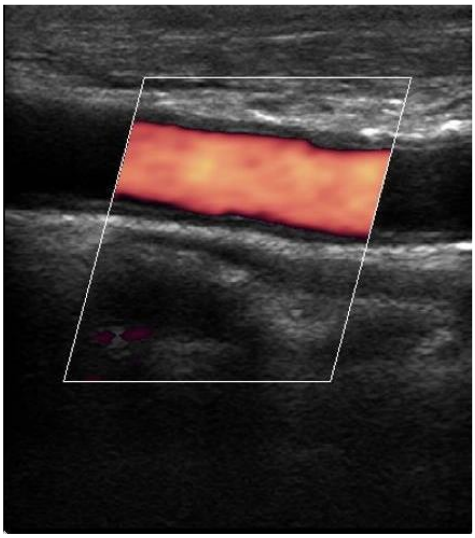
Usually, the IMT is manually measured by the specialist on the ultra-sound. It is possible to reduce the subjectivity and variability of manual approaches and detecting the IMT throughout the artery length using image segmentation algorithms. The vascular region (a) and intima-media region (b) images measured on US images in IMT measurements are shown in Figure 3 [35].

2.3. Image Pre-processing

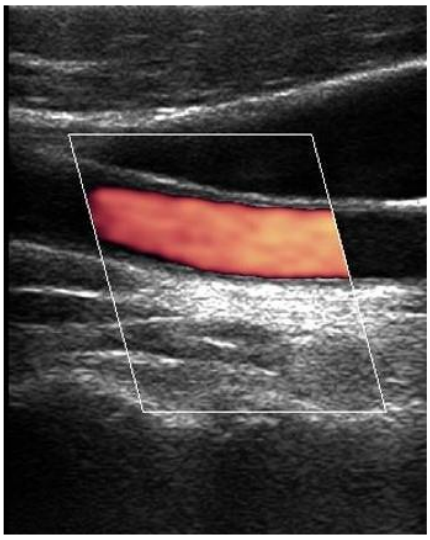
An image pre-processing operation was applied to CA IMT US images after the classification and labeling of images. The image resolution is fixed at 128x128 and the image channel is determined as a single channel since the grayscale image format is used. CA US IMT images taken from the image folder have been resized to 128x128 resolution. These resized images were then converted to grayscale image format and saved in the second folder which is determined during the identification phase. An example input image (a) and its state after the pre-processing stage (b) are shown in Figure 4, respectively.

The image sequence obtained from the last step of the image pre-processing stage is divided into two as training and test data. While 80% of the images (400 images) in the dataset are used for training, the remaining 20% (101 images) are used for testing. During the selection of the training images, random selection was made by maintaining the proportions of the images labeled "1" and "0" in the total image in the dataset. Because, if the selected images are performed in sequential order, the models are likely to memorize the data. Similarly, if model training is performed by selecting only weighted images from a certain class, there is a possibility that the model will memorize this class and fail in other class images. Therefore, it was aimed to ensure the sensitivity of the model in the study to the learning on the images in the

training set and to increase the accuracy of the model in this way. In addition, the possibility of memorizing or over-fitting the data is prevented.

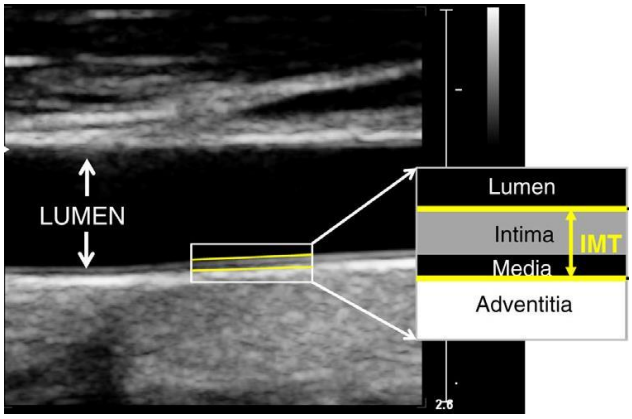


a)

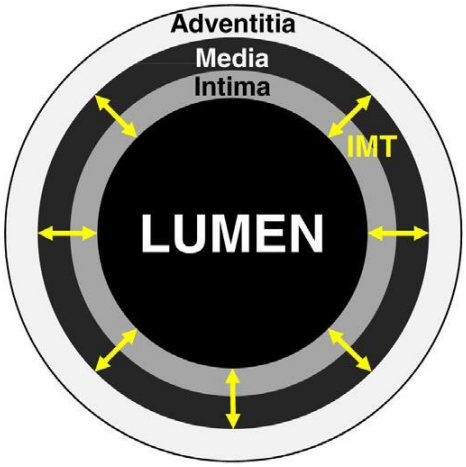


b)

Figure 2. Classified CA US images. a) IMT:1 b) IMT:0



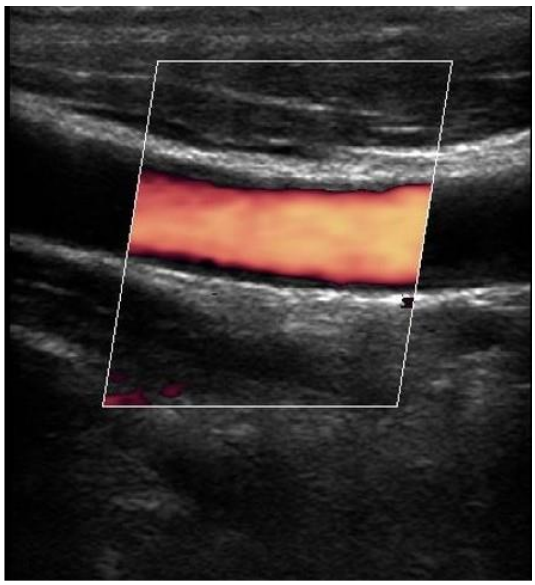
a)



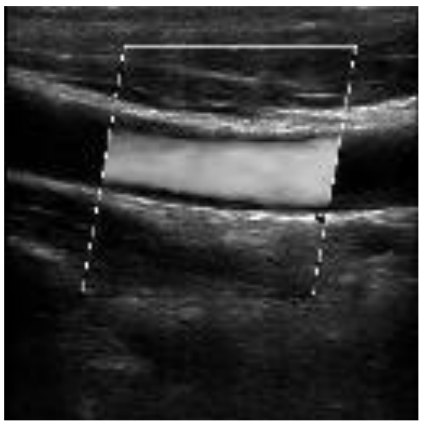
b)

Figure 3. a) Vascular region and b) Carotid artery wall layers [35]

As a result of the increasing prevalence of biomedical image processing studies in both field studies and clinical studies, many types of research and competitions have started to be carried out around the world. As such, the range of biomedical image processing has been gradually expanded by using different ways, methods, and contents in each study. Therefore, the researchers conducted a standardization study to make these studies easier to evaluate. In general, this study found a checklist and recommends that researchers stay within the limits of this checklist. Among the main headings of this checklist; title, abstract, keywords, introduction, method (organization, task, dataset, evaluation), results, and discussion. Under each title, there are sub-titles for competitions or research [36]. In this study, research that largely overlaps with the standards specified here (except for the competition organization titles) was carried out, and it was kept within the framework of standardization in medical image processing research.



a)



b)

Figure 4. Image pre-processing stage in the model. a) input image b) image after pre-processing

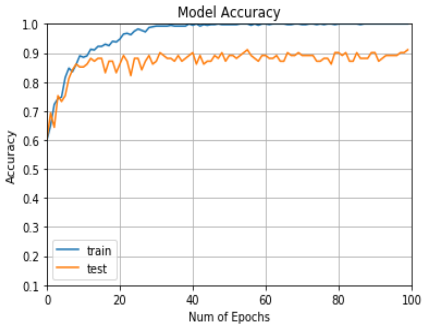
3. EXPERIMENTAL RESULTS

The models' training and testing processes were carried out by the Keras library of Python with the TensorFlow backend. The Scientific Python Development Environment (SPYDER) was used to implement models on images. The CNN models have been trained many times to access the best performances of hyper-parameters in the process of classification. The list of hyper-parameters used in models are like:

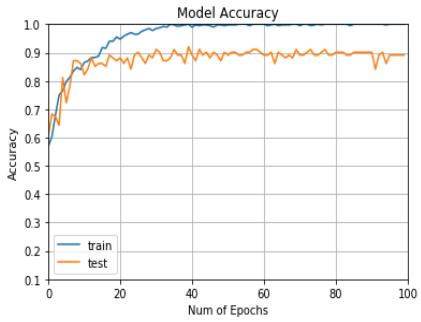
- Image-width:128 px
- Image-height: 128 px
- Image-channels: 1
- Batch size: 1
- Number of classes: 2
- Number of epoch: 100
- Loss: Binary crossentropy
- Optimizer: RMSprop
- Learning rate: 0,0001

The train and test accuracy graphs of the DL models used in the study are shown in Figure 5 and the classification accuracy results of the models are shown in Table 1.

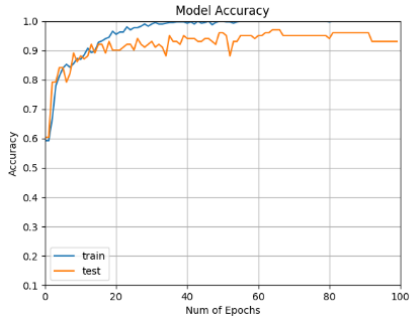
When the result graphs of the models used in the classification process on CA IMT US images are examined, AlexNet, ZFNet, VGG16, VGG19, and CNNcc models observed; 91%, 89.1%, 93%, 90%, and 89.1% accuracy rates respectively. Although accuracy is a determining factor for the model, the loss rate is also taken into consideration by researchers while evaluating the accuracy rates of the model created during the use of DL algorithms. The loss function is an important indicator of the discrepancy between the estimated value and the actual classification. As the loss value decreases, the robustness of the model increases [37]. Based on this information, the loss values of the DL models compared in the study were also measured in this study. The resulting loss graphs are shown in Figure 6.



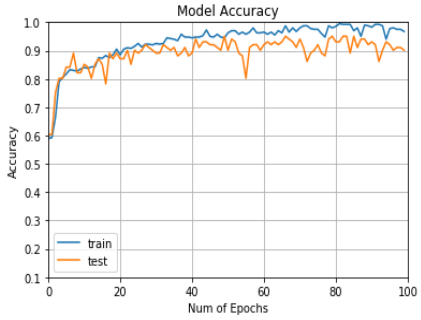
a)



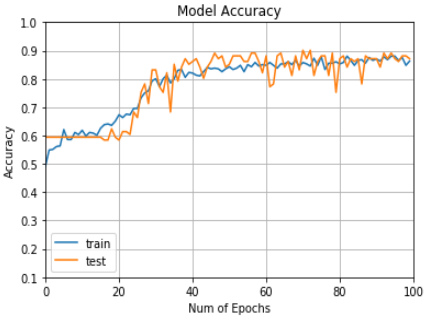
b)



c)



d)

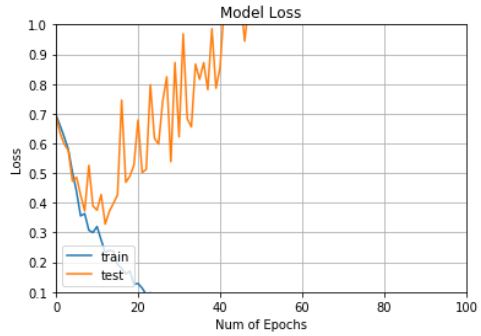


e)

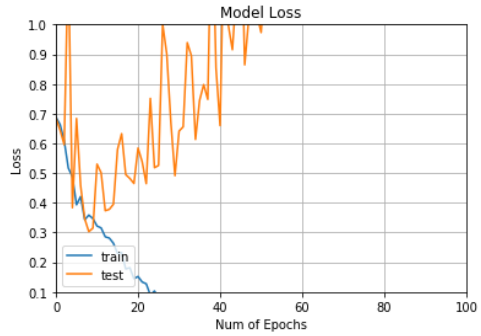
Figure 5. Accuracy graphics a) AlexNet b) ZFNet c) VGG16 d) VGG19 e) CNNcc

Table 1. Accuracy and loss rates

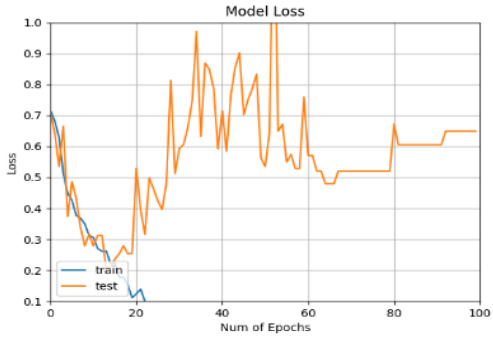
Model	*Test Accuracy*	*Loss*	*Train Accuracy*
AlexNet	%91	1.32	**%100**
ZFNet	%89,1	1.33	**%100**
VGG16	%93	0.65	**%100**
VGG19	**%90**	1.46	%94,5
CNNcc	%89,1	**0.29**	%89,4



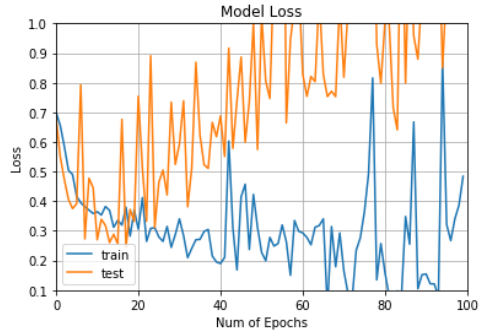
a)



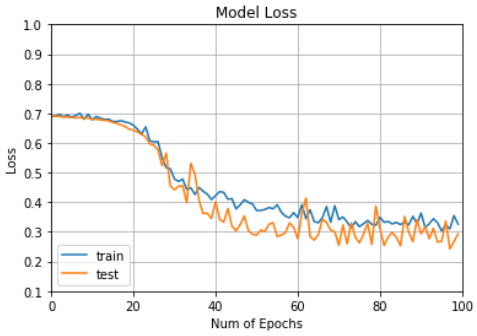
b)



c)



d)



e)

Figure 6. Loss graphics of a) AlexNet b) ZFNet c) VGG16 d) VGG19 e) CNNcc

When the inconsistency rates of the models on the labeled data in the database and the graphs are examined, the loss values of AlexNet, ZFNet, VGG16, VGG19, and CNNcc models are as follows; 1.33; 1.32; 0.65; 1.46, and 0.29 respectively. While the accuracy and loss parameters of the models are evaluated together, the last parameter to be added to this evaluation is the train accuracy rate. The accuracy of the models in the images reserved for training on the labeled data provides important clues about the model. When these ratios are examined, it is seen that AlexNet, ZFNet, and VGG16 models reach 100% accuracy. This ratio is undesirable during training because the model over-fitted the images in the data set and memorized the images. The reliability of the results has decreased due to the fall of these models to over-fitting. Weights obtained after training are not likely to produce the same results when used in different biomedical images. In VGG19 and CNNcc models, the situation is different. The VGG19 model has a 90% accuracy rate close to the CNNcc (89.1%) model. The accuracy rate obtained during the training was acceptable at 94.5% and no overfitting was performed on the data. On the other hand, the loss parameter, which is the other determinant parameter, has a value of 1.46 which is well above the CNNcc model.

To perform a more in-depth analysis of the results of the models on the images, the results of the confusion matrix were calculated separately for each model. The confusion matrix rows and columns are square matrices. Rows represent actual classes and columns represent predicted classes. The confusion matrix contains classification performance information of the model and sample distribution information of classes [38]. The confusion matrix results for AlexNet, ZFNet, VGG16, VGG19, and CNNcc models are shown in Table 2. Also as a result of the confusion matrices of the models; Accuracy, Error Rate, Recall, False Positive Rate, True Negative Rate, Precision, Prevalence, and F1-Score values were calculated and the results are shown in Table 3.

Table 2. Confusion matrices values of models

Number of Images:101		**Predicted**									
		Negative					Positive				
		AlexNet	ZFNet	VGG16	VGG19	CNNcc	AlexNet	ZFNet	VGG16	VGG19	CNNcc
Actual	Negative	54	55	56	59	57	7	6	5	2	3
	Positive	2	5	2	8	8	38	35	38	32	33

Table 3. Calculation of values of the confusion matrices and results

Value	*Calculation*	*AlexNet*	*ZFNET*	*VGG16*	*VGG19*	*CNNcc*
$Accuracy\ (Acc)$	$=\frac{TP+TN}{TP+TN+FP+FN}$	$\frac{92}{101}=0{,}91$	$\frac{90}{101}=0{,}891$	$\frac{94}{101}=0{,}93$	$\frac{91}{101}=0{,}90$	$\frac{90}{101}=0{,}891$
$Error\ Rate\ (ER)$	$=\frac{FP+FN}{TP+TN+FP+FN}$	$\frac{9}{101}=0{,}09$	$\frac{11}{101}=0{,}108$	$\frac{7}{101}=0{,}07$	$\frac{10}{101}=0{,}10$	$\frac{11}{101}=0{,}108$
$Recall$	$=\frac{TP}{TP+FN}$	$\frac{38}{40}=0{,}95$	$\frac{35}{40}=0{,}88$	$\frac{38}{40}=0{,}95$	$\frac{32}{40}=0{,}80$	$\frac{33}{41}=0{,}804$
$False\ Positive\ Rate\ (FPR)$	$=\frac{FP}{FP+TN}$	$\frac{7}{61}=0{,}11$	$\frac{6}{61}=0{,}09$	$\frac{5}{61}=0{,}08$	$\frac{2}{61}=0{,}03$	$\frac{3}{60}=0{,}05$
$True\ Negative\ Rate\ (TNR)$	$=\frac{TN}{FP+TN}$	$\frac{54}{61}=0{,}885$	$\frac{55}{61}=0{,}90$	$\frac{56}{61}=0{,}918$	$\frac{59}{61}=0{,}967$	$\frac{57}{60}=0{,}95$
$Precision$	$=\frac{TP}{TP+FP}$	$\frac{38}{45}=0{,}84$	$\frac{35}{41}=0{,}85$	$\frac{38}{43}=0{,}88$	$\frac{32}{34}=0{,}94$	$\frac{33}{36}=0{,}916$
$Prevalence$	$=\frac{TP+FN}{TP+TN+FP+FN}$	$\frac{40}{101}=0{,}40$	$\frac{40}{101}=0{,}40$	$\frac{40}{101}=0{,}40$	$\frac{40}{101}=0{,}40$	$\frac{41}{101}=0{,}405$
$F1\ Score$	$=\frac{2TP}{2TP+FP+FN}$	$\frac{76}{85}=0{,}894$	$\frac{70}{81}=0{,}864$	$\frac{76}{83}=0{,}916$	$\frac{64}{74}=0{,}865$	$\frac{66}{77}=0{,}857$

In the confusion matrix, accuracy calculation gives the ratio of correct predictions to total predictions. Error rate gives the ratio of wrong predictions to total predictions. Recall (or Sensitivity) indicates the correct prediction of positive values. FPR is the ratio of the wrong prediction of negative actual values. TNR indicates the correct prediction of negative values. Precision refers to the percentage of results, which are relevant. Prevalence is the frequency of positive values and F1-score is the harmonic mean of precision and recall. The estimated performance results of the models for each class on the labeled data are shown in Table 4.

Table 4. Performance measures of models

Number of Images:101	*Precision*					*Recall*					*F1-score*				
	AlexNet	ZFNet	VGG16	VGG19	CNNcc	AlexNet	ZFNet	VGG16	VGG19	CNNcc	AlexNet	ZFNet	VGG16	VGG19	CNNcc
IMT:0	0.96	0.92	0.97	0.88	0.88	0.89	0.90	0.92	0.97	0.95	0.92	0.91	0.94	0.92	0.91
IMT:1	0.84	0.85	0.88	0.94	0.92	0.95	0.88	0.95	0.80	0.80	0.89	0.86	0.92	0.86	0.86
avg/total	0.92	0.89	0.93	0.90	0.89	0.91	0.89	0.93	0.90	0.89	0.91	0.89	0.93	0.90	0.89

The estimated results of the models on the labeled images according to the performance criteria shown in Table 3 are shown in Table 5.

When Table 5 is examined, the ZFNet and VGG16 models correctly predicted all of the eight randomly selected images as test data, and the AlexNet, VGG19, and CNNcc models made an error. Since the ZFNet and VGG16 models are overfitting the data, there is no reliability in this estimation process. The interesting thing here is that although the AlexNet model is again in an over-fitting situation, it made an error in the prediction process. The VGG19 and CNNcc were the other models that made an error.

Receiver Operating Characteristic (ROC) analysis is useful for clinical decision-making. It can provide easy and low-cost cut-off value indicator determination in a short time. Thus, the diagnosis process can be protected from different limitations like time, cost, equipment, and qualified personnel [39]. In different clinical tests, sensitivity and specificity graphics allow comparison of success. The ROC curves of the models are shown in Figure 7.

Table 5. Results of test and prediction

AlexNet								
Test values	0	0	1	0	1	0	0	0
Predicted Values	[1. 0.]	[1. 0.]	[0. 1.]	[0. 1.]	[0. 1.]	[1. 0.]	[1. 0.]	[1. 0.]
Result	√	√	√	X	√	√	√	√
ZFNet								
Test values	0	0	1	1	1	0	0	0
Predicted Values	[1. 0.]	[1. 0.]	[0. 1.]	[0. 1.]	[0. 1.]	[1. 0.]	[1. 0.]	[1. 0.]
Result	√	√	√	√	√	√	√	√
VGG16								
Test values	0	0	1	1	1	0	0	0
Predicted Values	[1. 0.]	[1. 0.]	[0. 1.]	[0. 1.]	[0. 1.]	[1. 0.]	[1. 0.]	[1. 0.]
Result	√	√	√	√	√	√	√	√
VGG19								
Test values	0	0	1	1	0	0	0	0
Predicted Values	[1. 0.]	[1. 0.]	[0. 1.]	[0. 1.]	[0. 1.]	[1. 0.]	[1. 0.]	[1. 0.]
Result	√	√	√	√	X	√	√	√
CNNcc								
Test values	0	1	0	0	1	0	0	0
Predicted Values	[0. 1.]	[0. 1.]	[1. 0.]	[1. 0.]	[0. 1.]	[1. 0.]	[1. 0.]	[1. 0.]
Result	X	√	√	√	√	√	√	√

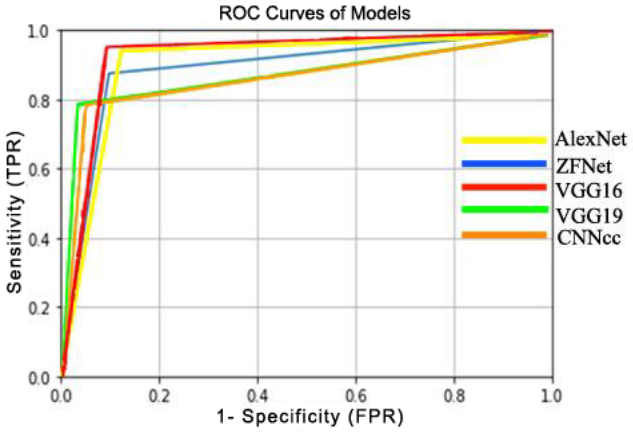


Figure 7. ROC Curve for models

4. DISCUSSION

Especially in recent years, the metrics in which the results are evaluated have become increasingly important to establish evaluation standards due to the increasing medical image processing studies. Among the evaluations used in the studies, there are various methods such as comparing accuracy and confusion matrix values calculated with mean and standard deviations [40], determination of standard parameters for all medical image processing studies [41], and characterization of tasks with the algorithm [42]. Based on this, a comprehensive evaluation process was carried out in this study. First, standard parameters were determined for all models using CNN architecture and compared in the study, and a comparison was made on a machine with the same hardware. In the next process, the reliability of the results was ensured by comparing the models over the accuracy, loss, confusion matrix, and ROC curves, and the focus was on metrics that became standardized in medical image processing evaluations. As a result of all these analyzes, the first issue to be discussed was the accuracy graph. When the accuracy rates on the test data were examined, the two models (VGG16: 93% and AlexNet: 91%) that reached the highest rate reached 100% accuracy on the training data and fell into the unreliable over-fitting situation for the test results. In addition, the Loss values of these models remain higher than the Loss value of the CNNcc model, an indication that the models mix between the labeled data during training. Many studies conducted in the field have clearly stated that a rate of 100% seen during training of models is not reliable, this rate is an indicator of over-fitting status, and that such a model may not produce similar results on different images.

In the next stage of the evaluations, the confusion matrix results were evaluated. These results allow for an in-depth analysis of different parameters of each image class. Among the models used in the comparison, AlexNet and VGG16, the two models with the highest Recall ratio used for the correct estimation of positive values, were the most successful models with 95%. However, similarly, the VGG19 (94%) and CNNcc (91.6%) models were successful in the targeted Precision value, where the false positive estimation was low in the correct estimation of positive values. VGG19 (96.7%) and CNNcc (95%) were the two most successful models in the TNR ratio used for the correct estimation of negative values. The model with the highest F1 score is the VGG16 model, and other models produced rates close to each other.

In medical image processing problems, each wrong guess can create problems that can seriously affect the results. Therefore, FPR and TNR ratios are of great importance. Among the models compared in the study, VGG19 and CNNcc, the models with the lowest FPR ratio and the highest TNR ratio, were an indication that the accuracy rate alone would not be sufficient for decision making in biomedical image processing. The ROC curve created in the study was also created to confirm these results.

In the light of all these evaluations, it was concluded that VGG19 and CNNcc models are suitable for application on biomedical images. Among these two models, CNNcc has been proposed as the most appropriate model since the inconsistency rate on estimating the labeled data is low and the accuracy rate approaches the VGG19 model. When all the results are examined, the CNNcc model does not fall into the overfitting state (accuracy: 89.4%), produces a

result close to the VGG19 model at the accuracy rate, and consistency capture with the lowest result of 0.29, which is the lowest parameter in the loss parameter, is considered as the most applicable model on biomedical images. The following determinations were made to use the models that were successful in ImageNet competition such as AlexNet, ZFNet, and VGG16 on biomedical images:

- Layer numbers of models should be arranged.
- Drop Out layers should be added to the models and loss parameters should be adjusted.
- Overfitting should be prevented by adding Drop Out layers.
- The number of images should be increased.
- Adjustments should be made on the hyper-parameters used in the models.

The results of the study showed that the performance of deep architectures in the biomedical field is promising. The classification results of the DL models used in the study were more pronounced than the ML algorithms. This study is especially important in terms of revealing certain DL models that can be applied to different imaging techniques in the biomedical field. Thus, researchers and clinicians will be able to use a common model instead of using separate solutions and algorithms for each problem.

The difficulty of obtaining images in a CAIMT study is the foremost limitation of the study. Because in this treatment method, image recording is not a frequently used application, and it is usually decided by the doctor's observation on the screen. When the literature is examined, studies on the CAIMT are generally segmentation studies [35, 43-46], but there are also some classification studies. ML algorithms were mostly used in classification studies. Among these algorithms, the accuracy of 71% [47] to 73% [48] was achieved in the studies carried out with the use of NN, while there were studies that reached accuracy rates between 73% [49] and 83% [50] with Support Vector Machines. When compared with the results obtained with different techniques before, it is seen that DL algorithms achieve more successful results in this regard. In this study, it has been demonstrated that more successful results are obtained with state-of-art methods and techniques.

5. CONCLUSION

In this study, the performance of the DL models on the CA IMT US images have been compared. The classification methods proposed by comparing their performance are important for early diagnosis of CVD and treatment of the disease.

AlexNet model, which has made significant improvements in the world with DL, and ZFNet, VGG16, VGG19 models that have succeeded in ImageNet competition and a CNNcc model prepared by the authors were tested. In addition, their classification performances were compared. 501 US images from 153 patients in Ankara Training and Research Hospital were used to test the models. As a result of the tests, while AlexNet, ZFNet, and VGG16 models achieved 91%, 89.1%, and 93% accuracy, respectively, the reliability of the results of these models decreased because they performed over-fitting to the data during the training. Changes in hyper-parameters on the layers of these models should be adapted to the data by adding drop-out layers and making various improvements. The VGG19 model achieved 90% performance and produced positive results. The loss parameter of the VGG19 model is 1.46. This value means inconsistency in the detection of labeled data for the model and some layer arrangements are required in the model by drop out method. Finally, the accuracy rate of the CNNcc model was measured as 89.1% and it was determined to be the most suitable model for classification on the images in the study together with the 0.29 loss parameter ratio and the confusion matrix performance.

This study is important to show that the performance of DL models on medical images can be more efficient than ML methods. Deep architectures successfully perform classification on different images. From this point of view, the formation of certain common DL models on biomedical images will be an important development in terms of the AI discipline. Although adequate models have not yet been produced and necessary steps have not yet been taken, this study is an example of the emergence of new models of DL that can demonstrate high performance on biomedical images.

For the CAIMTUSNet, a study to be carried out on ensemble models, which has become increasingly widespread in recent years, is planned as future work. Ensemble models often achieve successful results in studies such as segmentation [51, 52] with DL, detection [53], and feature selection [54, 55] with ML algorithms. For this reason, ensemble feature selection and ensemble model comparison studies can be performed in the CAIMTUSNet.

ACKNOWLEDGMENTS

The essence of this article was presented at the HORA-2019 International Congress on Human-Computer Interaction, Optimization and Robotic Applications [56] and it was finalized with additions and improvements made after the congress. This article is an extension of that presentation with new analysis and additions.

The authors would like to thank the Radiology Department of Ankara Training and Research Hospital for their kind cooperation and for providing all the ultrasound images used.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, 25: 1097-1105, 2012.

[2] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, 18(7): 1527-1554, 2006. doi: 10.1162/neco.2006.18.7.1527.

[3] K. Phil, **Matlab Deep Learning with Machine Learning, Neural Networks and Artificial Intelligence**. Seoul, Soul-t'ukpyolsi, Korea: Apress, 2017.

[4] H. A. Song and S.-Y. Lee, "Hierarchical representation using NMF," in **International conference on neural information processing**, 2013: Springer, 466-473.

[5] S. Savaş, N. Topaloğlu, Ö. Kazcı, and P. N. Koşar, "Classification of Carotid Artery Intima Media Thickness Ultrasound Images with Deep Learning," *Journal of Medical Systems*, 43(8): 273, 2019, doi: 10.1007/s10916-019-1406-2.

[6] O. Güler and İ. Yücedağ, "Hand Gesture Recognition from 2D Images by Using Convolutional Capsule Neural Networks," Arabian Journal for Science and Engineering, 2021/06/25 2021. doi: 10.1007/s13369-021-05867-2.

[7] P. J. Hu, F. Wu, J. L. Peng, Y. Y. Bao, F. Chen, and D. X. Kong, "Automatic abdominal multi-organ segmentation using deep convolutional neural network and time-implicit level sets," *Int. J. Comput. Assist. Radiol. Surg.*, 12(3): 399-411, 2017. doi: 10.1007/s11548-016-1501-5.

[8] O. Z. Kraus, J. L. Ba, and B. J. Frey, "Classifying and segmenting microscopy images with deep multiple instance learning," *Bioinformatics*, Article; Proceedings Paper, 32(12): 52-59, 2016. doi: 10.1093/bioinformatics/btw252.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in **International Conference on Medical image computing and computer-assisted intervention**, 2015: Springer, 234-241.

[10] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep learning for identifying metastatic breast cancer," *arXiv preprint arXiv:1606.05718*, 2016.

[11] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in **International conference on medical image computing and computer-assisted intervention**, 2013: Springer, 411-418.

[12] V. Gulshan *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *Jama*, 316(22): 2402-2410, 2016.

[13] Q. Dou, H. Chen, L. Q. Yu, J. Qin, and P. A. Heng, "Multilevel Contextual 3-D CNNs for False Positive Reduction in Pulmonary Nodule Detection," *IEEE Trans. Biomed. Eng.*, 64(7): 1558-1567, 2017. doi: 10.1109/tbme.2016.2613502.

[14] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, and H. Adeli, "Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals," *Computers in biology and medicine*, 100: 270-278, 2018.

[15] L. A. Yeola and M. P. Satone, "Deep neural network for the automated detection and diagnosis of seizure using EEG signals," *International Research Journal of Engineering and Technology (IRJET)*, 6(7): 381-385, 2019.

[16] K. Gürkahraman and R. Karakış, "Brain tumors classification with deep learning using data augmentation," *Journal of the Faculty of Engineering and Architecture of Gazi University*, 36(2): 997-1011, 2021, doi: 10.17341/gazimmfd.762056.

[17] S. Luo, X. Li, and J. Li, "Automatic Alzheimer's disease recognition from MRI data using deep learning method," *Journal of Applied Mathematics and Physics*, 5(9): 1892-1898, 2017.

[18] W. Lin *et al.*, "Convolutional Neural Networks-Based MRI Image Analysis for the Alzheimer's Disease Prediction From Mild Cognitive Impairment," *Frontiers in Neuroscience*, 12(777), 2018, doi: 10.3389/fnins.2018.00777.

[19] Y. Ding *et al.*, "A Deep Learning Model to Predict a Diagnosis of Alzheimer Disease by Using 18F-FDG PET of the Brain," *Radiology*, 290(2): 456-464, 2019. doi: 10.1148/radiol.2018180958.

[20] S. Savaş, "Detecting the Stages of Alzheimer's Disease with Pre-trained Deep Learning Architectures," *Arabian Journal for Science and Engineering*, 2021/09/20, 2021, doi: 10.1007/s13369-021-06131-3.

[21] R. Karakış and K. Gürkahraman, "Medikal Görüntülerde Derin Öğrenme ile Steganaliz," *Bilişim Teknolojileri Dergisi,* 14(2): 151-159, 2021.

[22] B. H. Menze *et al.*, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, 34(10): 1993-2024, 2015. doi: 10.1109/TMI.2014.2377694.

[23] A. E. Kavur *et al.*, "CHAOS Challenge - combined (CT-MR) healthy abdominal organ segme ntation," *Medical Image Analysis*, 69: 101950, 2021, doi: 10.1016/j.media.2020.101950.

[24] X. Zhuang *et al.*, "Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge," *Medical image analysis*, 58: 101537, 2019.

[25] Internet: A. Civelek. "Karotis Arter Hastalığı." http://www.alicivelek.com/karotis-arter-hastaligi 1.1.2021.

[26] M.-G. Bousser, "Stroke prevention: an update," *Frontiers of medicine*, 6(1): 22-34, 2012.

[27] K. Strong, C. Mathers, and R. Bonita, "Preventing stroke: saving lives around the world," *Lancet Neurol.*, 6(2): 182-187, 2007. doi: 10.1016/s1474-4422(07)70031-5.

[28] A. Demirci Şahin, Y. Üstü, and D. Işık, "Management of Preventable Risk Factors of Cerebrovascular Disease," *Ankara Medical Journal*, 15(2): 2015.

[29] S. Bakanlığı, **Türkiye hastalık yükü çalışması 2004**, Hıfzıssıhha Mektebi Müdürlüğü, 2006.

[30] C. C. Phatouros *et al.*, "Carotid artery stent placement for atherosclerotic disease: Rationale, technique, and current status," *Radiology*, 217(1): 26-41, 2000. doi: 10.1148/radiology.217.1.r00oc2526.

[31] N. Daldal, Z. Cömert, and K. Polat, "Automatic determination of digital modulation types with different noises using Convolutional Neural Network based on time–frequency information," Applied Soft Computing, 86: 105834, 2020. doi: 10.1016/j.asoc.2019.105834.

[32] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in **European conference on computer vision**, 2014: Springer, 818-833.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[34] F. Doğan and I. Türkoğlu, "Comparison of Leaf Classification Performance of Deep Learning Algorithms," *Sakarya University Journal of Computer and Information Sciences*, 1: 10-21, 2018.

[35] R.-M. Menchón-Lara, J.-L. Sancho-Gómez, and A. Bueno-Crespo, "Early-stage atherosclerosis detection using deep learning over carotid ultrasound images," *Applied Soft Computing*, 49: 616-628, 2016, doi: 10.1016/j.asoc.2016.08.055.

[36] L. Maier-Hein *et al.*, "BIAS: Transparent reporting of biomedical image analysis challenges," *Medical image analysis*, 66: 101796, 2020.

[37] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss Functions for Image Restoration With Neural Networks," *IEEE Transactions on Computational Imaging*, 3(1): 47-57, 2017. doi: 10.1109/TCI.2016.2644865.

[38] D. Ballabio, F. Grisoni, and R. Todeschini, "Multivariate comparison of classification performance measures," *Chemometrics and Intelligent Laboratory Systems*, 174: 33-44, 2018.

[39] S. Kılıç, "ROC analysis in clinical decision making," *Psychiatry and Behavioral Sciences*, 3(3): 135, 2013.

[40] A. E. Kavur *et al.*, "Comparison of semi-automatic and deep learning-based automatic methods for liver segmentation in living liver transplant donors," *Diagnostic and Interventional Radiology*, 26(1): 11, 2020. doi: 10.5152/dir.2019.19025.

[41] L. Maier-Hein *et al.*, "Why rankings of biomedical image analysis competitions should be interpreted with care," *Nature communications*, 9(1): 1-13, 2018.

[42] M. Wiesenfarth *et al.*, "Methods and open-source toolkit for analyzing and visualizing challenge results," *Scientific Reports*, 11(1): 2369, 2021. doi: 10.1038/s41598-021-82017-6.

[43] R. Rocha, A. Campilho, J. Silva, E. Azevedo, and R. Santos, "Segmentation of the carotid intima-media region in B-mode ultrasound images," *Image and Vision Computing*, 28(4): 614-625. 2010. doi: 10.1016/j.imavis.2009.09.017.

[44] M. C. Bastida-Jumilla, R. M. Menchón-Lara, J. Morales-Sánchez, R. Verdú-Monedero, J. Larrey-Ruiz, and J. L. Sancho-Gómez, "Frequency-domain active contours solution to evaluate intima–media thickness of the common carotid artery," *Biomedical Signal Processing and Control*, vol. 16, pp. 68-79, 2015/02/01/ 2015, doi: 10.1016/j.bspc.2014.08.012.

[45] U. Kutbay, F. Hardalaç, M. Akbulut, Ü. Akaslan, and S. Serhatlıoğlu, "A Computer-Aided Diagnosis System for Measuring Carotid Artery Intima-Media Thickness (IMT) Using Quaternion Vectors," *J Med Syst*, 40(6): 149, 2016. doi: 10.1007/s10916-016-0507-4.

[46] N. Ikeda *et al.*, "Automated segmental-IMT measurement in thin/thick plaque with bulb presence in carotid ultrasound from multiple scanners: Stroke risk assessment," *Comput Methods Programs Biomed*, 141: 73-81, 2017. doi: 10.1016/j.cmpb.2017.01.009.

[47] E. Kyriacou *et al.*, "Ultrasound imaging in the analysis of carotid plaque morphology for the assessment of stroke," *Stud Health Technol Inform*, 113: 241-75, 2005.

[48] C. I. Christodoulou, C. S. Pattichis, M. Pantziaris, and A. Nicolaides, "Texture-based classification of atherosclerotic carotid plaques," *IEEE Trans Med Imaging*, 22(7): 902-912, 2003. doi: 10.1109/tmi.2003.815066.

[49] E. Kyriacou *et al.*, "Classification of atherosclerotic carotid plaques using morphological analysis on ultrasound images," *Applied Intelligence*, 30(1): 3-23, 2009, doi: 10.1007/s10489-007-0072-0.

[50] R. U. Acharya *et al.*, "Symptomatic vs. asymptomatic plaque classification in carotid ultrasound," *J Med Syst*, 36(3): 1861-1871, 2012, doi: 10.1007/s10916-010-9645-2.

[51] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, 18(2): 203-211, 2021, doi: 10.1038/s41592-020-01008-z.

[52] A. E. Kavur, L. I. Kuncheva, and M. A. Selver, "Basic Ensembles of Vanilla-Style Deep Learning Models Improve Liver Segmentation From CT Images," *arXiv preprint arXiv:2001.09647*, 2020.

[53] S. Buyrukoğlu, "Early Detection of Alzheimer's Disease Using Data Mining: Comparision of Ensemble Feature Selection Approaches," *Konya Mühendislik Bilimleri Dergisi,* 9(1): 50-61, 2021, doi: 10.36306/konjes.731624.

[54] G. Buyrukoğlu, S. Buyrukoğlu, and Z. Topalcengiz, "Comparing Regression Models with Count Data to Artificial Neural Network and Ensemble Models for Prediction of Generic Escherichia coli Population in Agricultural Ponds Based on Weather Station Measurements," *Microbial Risk Analysis*, 100171, 2021. doi: 10.1016/j.mran.2021.100171.

[55] S. Buyrukoğlu, "New hybrid data mining model for prediction of Salmonella presence in agricultural waters based on ensemble feature selection and machine learning algorithms," *Journal of Food Safety*, e12903, 2021. doi: 10.1111/jfs.12903.

[56] S. Savaş, N. Topaloğlu, Ö. Kazcı, and P. N. Koşar, "Performance Comparison of Carotid Artery Intima Media Thickness Classification by Deep Learning Methods," presented at the **International Congress on Human-Computer Interaction, Optimization, and Robotic Applications**, Urgup, Nevşehir, Turkey, 2019. doi: 10.36287/setsci.4.5.025.