Turkish Journal of Science & Technology
16(1), 103-112, 2021

Classification of 40 Different Human Movements with CNN Architectures and Comparison of Their Performance

Muhammed YILDIRIM[1*], Ahmet ÇINAR
[1] Computer Engineering Department, Engineering Faculty, Firat University, Elazığ, Turkey
[2] Computer Engineering Department, Engineering Faculty, Firat University, Elazığ, Turkey
[*1] 171129205@firat.edu.tr, [2] acinar@firat.edu.tr

(Geliş/Received: 27/01/2021; Kabul/Accepted: 09/02/2021)

Abstract: Detection of human movements has become one of the current issues with the developing technology. Recognition of human movements is used in many areas such as security systems, human computer interaction, human robot interaction. Due to the increase in data stored in databases, deep learning methods have recently become one of the most frequently used methods. At this study, it is aimed to classify human movements by using Convolutional Neural Network (CNN) architectures. Images are classified with InceptionV3, Googlenet and Alexnet architectures using a data set with 40 different motion classes. The highest accuracy rate with 76.15% was obtained in InceptionV3 architecture. Increasing the amount of data in CNN networks is a parameter that closely concerns the network uptime. Since 40 different motion classes are used in this study, the results obtained in the related architectures are obtained in different times.

Key words: Deep Learning, Human movements, Alexnet, Googlenet, Inceptionv3.

40 Farklı İnsan Hareketinin CNN Mimarileriyle Sınıflandırılması ve Başarım Oranlarının Karşılaştırılması

Öz: İnsan hareketlerinin tespit edilmesi gelişen teknolojiyle birlikte güncel konulardan biri haline gelmiştir. İnsan hareketlerinin tanınması güvenlik sistemleri, insan bilgisayar etkileşimi, insan robot etkileşimi gibi birçok alanda kullanılmaktadır. Veri tabanlarında saklanan verilerin artmasından dolayı son zamanlarda derin öğrenme yöntemleri en sık kullanılan yöntemlerden biri haline gelmiştir. Bu çalışmada Convolutional Neural Network (CNN) mimarileri kullanılarak insan hareketlerinin sınıflandırılması amaçlanmıştır. 40 farklı hareket sınıfına sahip olan bir veri seti kullanılarak InceptionV3 Googlenet ve Alexnet mimarileriyle görüntüler sınıflandırılmıştır. En yüksek doğruluk oranı %76.15 ile InceptionV3 mimarisinde elde edilmiştir. CNN ağlarında veri miktarının artması ağın çalışma süresini yakından ilgilendiren bir parametredir. Bu çalışmada 40 farklı hareket sınıfı kullanıldığından dolayı ilgili mimarilerde alınan sonuçlar farklı sürelerde elde edilmiştir.

Anahtar kelimeler: Derin Öğrenme, İnsan Hareketleri, Alexnet, Googlenet, Inceptionv3.

1. Introduction

After Alexnet architecture won the large-scale visual recognition (ILSVCR) competition in 2012, the Convolutional Neural Network (CNN) started a new era in image processing. Thanks to these deep learning-based neural networks, the classification process can be successfully performed today. In this study, Alexnet, who won the large-scale visual recognition competition in 2012, and the Googlenet and inceptionV3 architectures, which won the same competition in 2014, were used [1]. Classification of human movements has become an important issue with the developing technology. Recognition and classification of human movements are widely used in many areas such as security systems, human computer interaction, smart homes, object tracking, environment supported life, rehabilitation services in elderly care homes [2]. This situation makes the issue of recognition of human movements more and more important every day.

The purpose of the describe human movements is to try to recognize a human activity automatically using relevant images. In recent years, human movement recognition systems have been a major subject of study. It is very difficult to classify human movements with high accuracy, especially in studies involving more than one movement class [3]. The source of this difficulty lies in the similarity of most human movements. The fact that human movements are similar to each other complicates the result obtained in large-class data sets. The

[*] Corresponding author: 171129205i@firat.edu.tr. ORCID Number of authors: [1]0000-0003-1866-4721, [2] 0000-0001-5528-2226

performance rate may be higher in data sets with fewer classes. Since there are 40 different classes in the data set used in this study, the performance rate of the study decreases.

Many methods are used to classify human movements. Support Vector Machines (SVM) [4], Principal Component Analysis (PCA) [5], Histograms of Oriented Gradients(HOG) [6] are some of these methods used. However, methods based on deep learning have been used widely in the field of image classification in recent years [7]. In this study, it is aimed to classify 40 different human movements using the Stanford 40 Action data set taken from Stanford University. Alexnet, Googlenet and InceptionV3 models, which are CNN models, are used for the classification of this data. With these models, accuracy and loss curves are obtained separately. In addition, the performance rates of the models were calculated from the confusion matrix obtained and these models were compared with each other.

In the first part of the study, the classification of human movements is briefly mentioned. In the second chapter, the data set and used models are explained. In the third part, experimental results are evaluated. In the last section, the results are evaluated and future studies are mentioned.

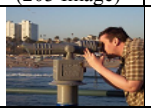
2. Materials and Methods

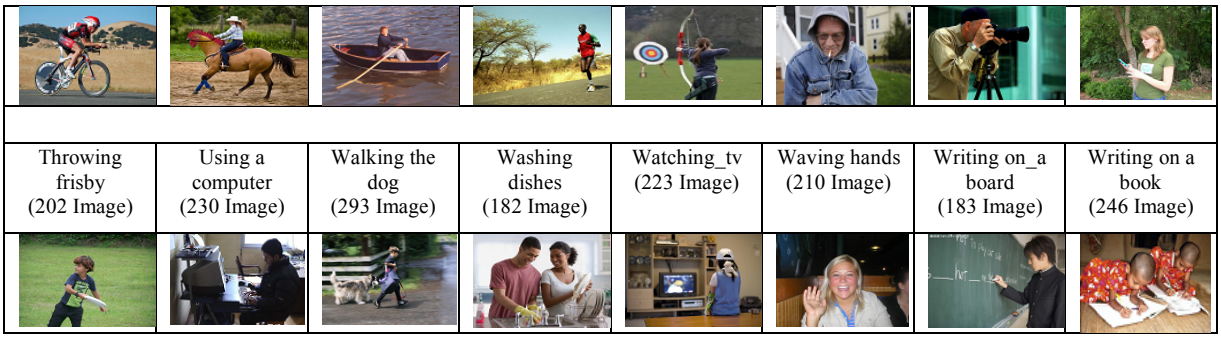
Classification of human movements has become an important issue with the developing technology. In this study, it is aimed to classify 40 different human movements. CNN-based models are used to classify human movements. The test results obtained with Alexnet [8], Googlenet [9], Inceptionv3 [10] models were compared with each other.

2.1. Data set

In this study, 40 Action datasets taken from the Stanford University web page are used [11]. The data set consists of 40 different human movement classes. The number of pictures in each class is different for these movements. The classes used and the number of data in these classes are given in table 1.

Table 1. Data counts and image examples in 40 action classes.

Applauding (285 Image)	Blowing bubbles (259 Image)	Brushing teeth (200 Image)	Cleaning the floor (212 Image)	Climbing (295 Image)	Cooking (288 Image)	Cutting trees (203 Image)	Cutting vegetables (189 Image)
							
Drinking (256 Image)	Feding a horse (286 Image)	Fishing (273 Image)	Fixing a bike (228 Image)	Fixing a car (251 Image)	Gardening (199 Image)	Holding an umbrella (292 Image)	Jumping (295 Image)
							
Looking through a microscope (191 Image)	Looking through a telescope (203 Image)	Phoning (259 Image)	Playing guitar (289 Image)	Playing violin (260 Image)	Pouring liquid (200 Image)	Pushing a cart (235 Image)	Reading (245 Image)
							
Riding_a_bike (293 Image)	Riding a horse (296 Image)	Rowing a boat (185 Image)	Running (251 Image)	Shooting an_arrow (214 Image)	Smoking (241 Image)	Taking photos (197 Image)	Texting message (193 Image)



Each class in the data set has an average of 180-300 images. The total number of images in 40 classes is 9532.

2.2. Models Used in the Paper

In this study, Alexnet, Googlenet and Inceptionv3, which are CNN models, are used.

2.2.1. Alexnet

The LeNet architecture, published by Yan LEcun in 1998, is seen as the first study in deep learning. However, the most important factor in the popularization of deep learning is the Alexnet architecture. Alexnet architecture won the large-scale visual recognition (ILSVRC) competition published in 2012. After this success of the Alexnet architecture, deep learning architectures have become popular. With this success of the Alexnet architecture, the object identification error value of computers decreased from 26.2% to 15.4% [12].

2.2.2. Googlenet

It is the winning model of ImageNet 2014 competition. The combination of starter modules in the Googlenet model has a complex structure. Since the error rate in Googlenet architecture is 5.7%, it has achieved successful results in data sets. Googlenet has a structure of 144 layers in total. The starting layers in Googlenet have the ability to filter in different sizes. With this feature, it differs from the previously suggested models [13]. This architecture has the distinction of being one of the first Cnn works in which layers have moved away from an ordered structure.

2.2.3. Inceptionv3

We can define Inceptionv3 model as a model consisting of three parts. This parts Convolutional block, improved start up module and classifier [14]. This model takes the input data in 299x299 size. Inceptionv3 model consists of 315 layers in total**.** In the Inceptionv3 model, Relu is used as the activation function, batch normalization for normalization, and max pooling and average pooling are used together for pooling.

2.3. Layers of Convolutional Neural Networks

2.3.1. Input Layer

This layer is the first layer of the developed CNN models. Images are first read from the input layer [15]. The entrance dimensions of the models used in the application are given in table 2.

Table 2. Input size of models

	Alexnet	Googlenet	InceptionV3
Input Size	227 227 3	224 224 3	299 299 3

2.3.2. Convolutional Layer

Convolution layer, which is considered as the main layer of CNN architecture, is the process of applying different filters such as 2x2, 3x3, 5x5 on the image. The filter operator is scrolled through the image. The main purpose of this process is to obtain feature maps [16]. The application of a 3x3 size filter to an 8x8 size image is shown in figure 1.

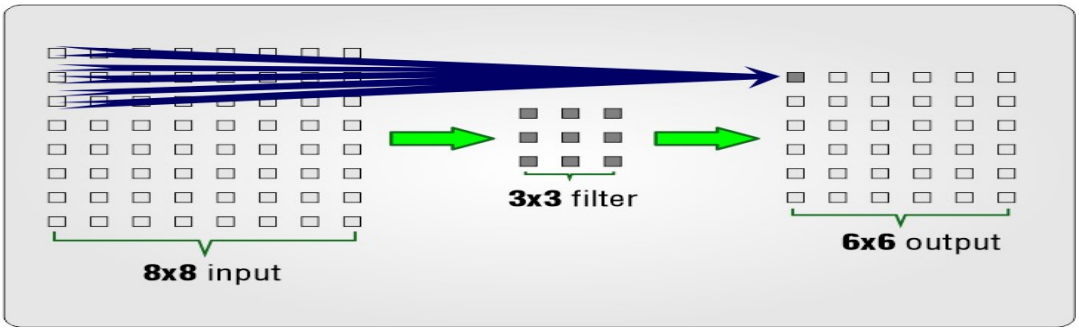


Figure 1. Applying a 3x3 filter to the 8x8 input image

Calculation of the output size is given in equation 1.

$o = ((i - k) + 2p)/s + 1$ (1)

i: Input Size, k: filter size, s: number of steps, p: padding process, o: output size

2.3.3. Activation Function (RELU)

With the activation process, a linear filter is obtained with a nonlinear function applied to each component of a feature map. Relu is an activation function commonly used in CNN networks. It is also used in sigmoid and tangent activation functions [17]. Relu equation 2, Sigmoid equation 3 and Tanh equation 4 are presented.

Relu= $f(x) = \begin{cases} 0, x < 0 \\ x, x \geq 0 \end{cases}$, $f(x)' = \begin{cases} 0, x < 0 \\ 1, x \geq 0 \end{cases}$ (2)

Sigmoid= $f(x) = \frac{1}{1+e^{-x}}$,$f'(x) = f(x)(1 - f(x))$ (3)

Tanh= $f(x) = \tanh(x) = \frac{2}{1+e^{-2x}}$ -1,$f'(x) = 1 - f(x)^2$ (4)

2.3.4. Normalization

This layer, where the normalization process is performed, mostly normalizes the values produced by the Fully Connected and convolution layers. After the normalization process, the training time of the network is shortened and the learning process of the network faster [18].

2.3.5. Dropout

One of the problems with CNN architectures is that the network is memorized. In CNN architectures, node dilution is performed to prevent memorization. In this process, some nodes of the network are removed randomly. The original structure of the network is shown in Figure 2.a, and its shape after the node dilution step is shown in Figure 2.b [19].

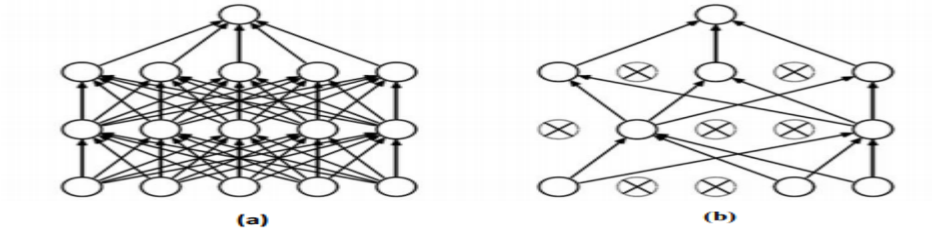


Figure 2. Dropout process

2.3.6. Pooling Layer

In the pooling layer, the data size is reduced and calculation costs are lowered. It is tried to prevent the network from memorizing. Since the data volume decreases in the pooled image, there is a shrinkage. The height and width of the image are reduced. As a result of this process, there is no change in the depth of the image [20]. In an image that has been pooled, the size of the new image is calculated by equation 5.

$$Y = \frac{W-F}{S} + 1 \quad (5)$$

Y=Size of the new image, W=Image_size, F=Filter_size, S=Number of steps.

2.3.7. Fully Connected

The classification stage comes after the feature extraction stage in CNN networks. Fully Connected takes feature maps as input and prepares them for classification. The matrix obtained from the feature maps in a multi-dimensional form is made unidimensional in this layer and given as input to the classifier [21].

2.3.8. Softmax and Classification

This layer comes after the Fully Connected layer and classification is done on this layer. The output value of this layer is equal to the number of objects to be classified [22].

3. Application and Results

In the study, it is aimed to classify human movement images belonging to 40 different classes with CNN models. While 80% of the Stanford 40 actions dataset used is used for training the model, the remaining 20% is reserved for testing the model. The study is carried out in a Matlab environment in a computer with 8GB RAM memory of the 8th generation.

There are many parameters that measure the success rate of working in CNN architectures. Most of these parameters are obtained using a Confusion matrix [23]. An example of the Confusion matrix is given in table 3.

Table 3. Confusion Matrix

	A	B
A	TruePositive(TP)	FalsePositive(FP)
B	FalseNegative(FN)	TrueNegative(TN)

The success rates of the models are compared according to the accuracy matrix. Accuracy ratio is calculated by equation 6.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}. \quad (6)$$

Other parameter values used in the study are given in table 4.

Table 4. Used parameters and their values

Software	Model	MaxEpochs	Optimization	MiniBatchSize	LearningRate	ValidationFrequency
Matlab	AlexNet GoogleNet InceptionV3	5	Stochastic Gradient Descent (SGD)	16	10^{-5}	6

The accuracy and loss curves of the model trained using the Alexnet model are given in figure 3. 59.59% accuracy value is obtained with Alexnet. Confusion matrix created with the Alexnet architecture is as in figure 4. By looking at the confusion matrix, we can see which data is placed in which class.

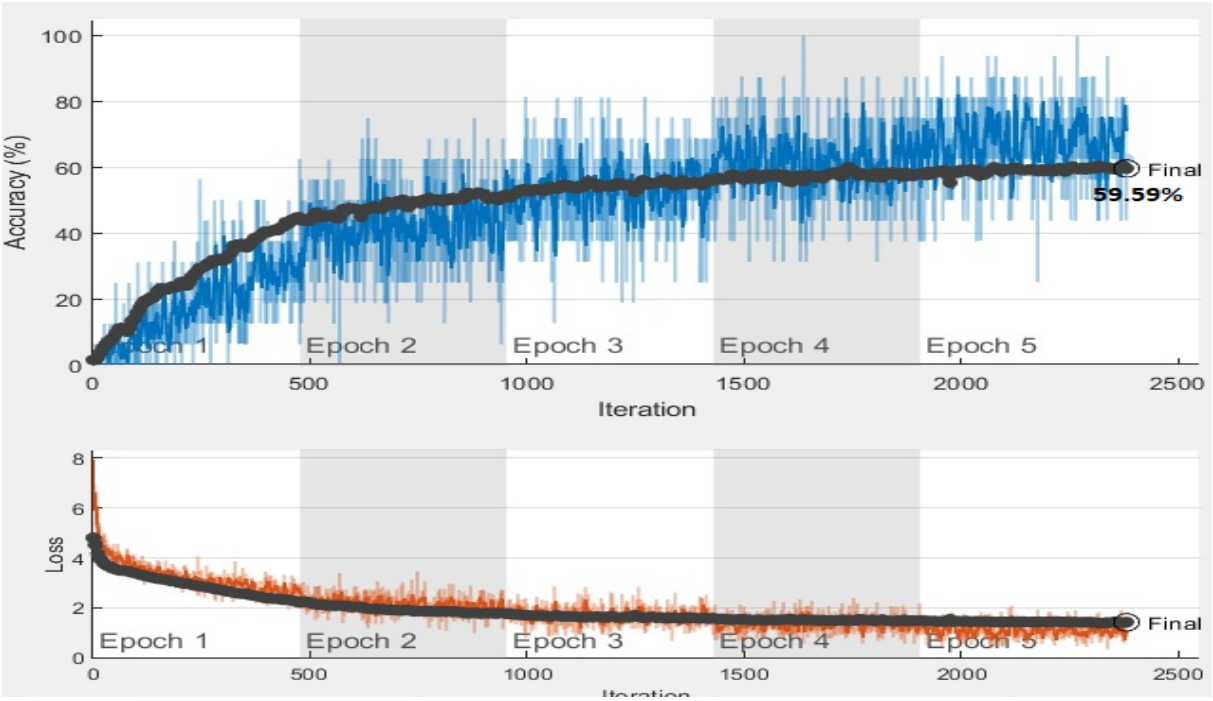


Figure 3. Alexnet's accuracy and loss curves

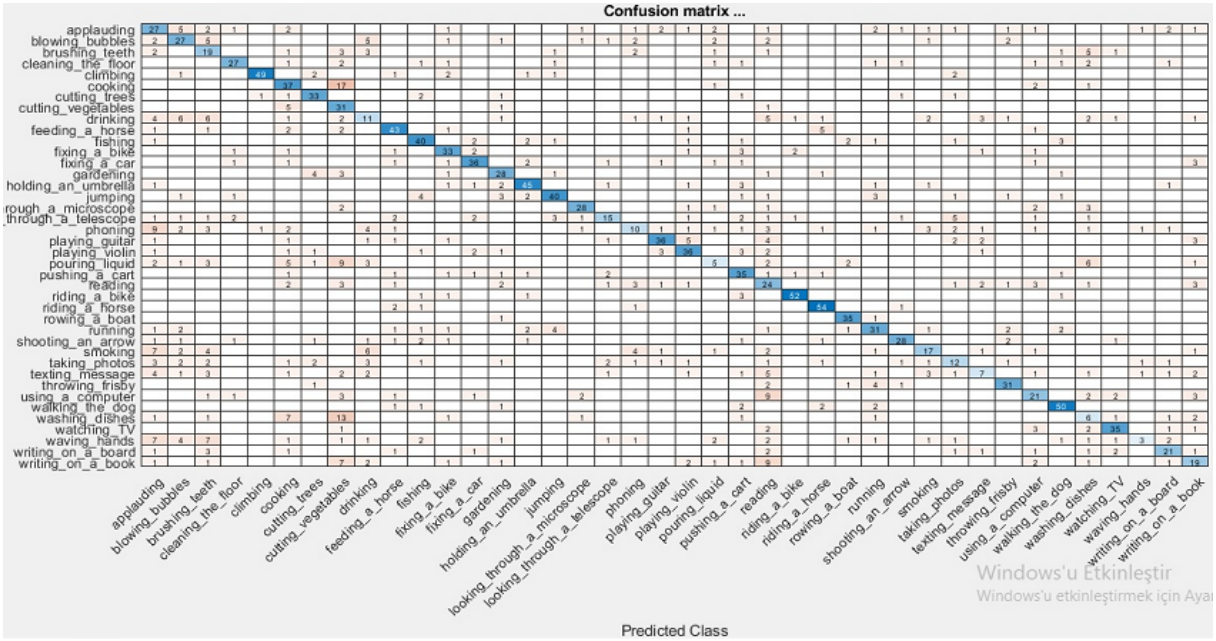


Figure 4. Alexnet's confusion matris.

The accuracy and loss curve acquired with the Googlenet model is as in figure 5. 71.02% accuracy is achieved with Googlenet. The confusion matrix obtained with Googlenet is as in figure 6.

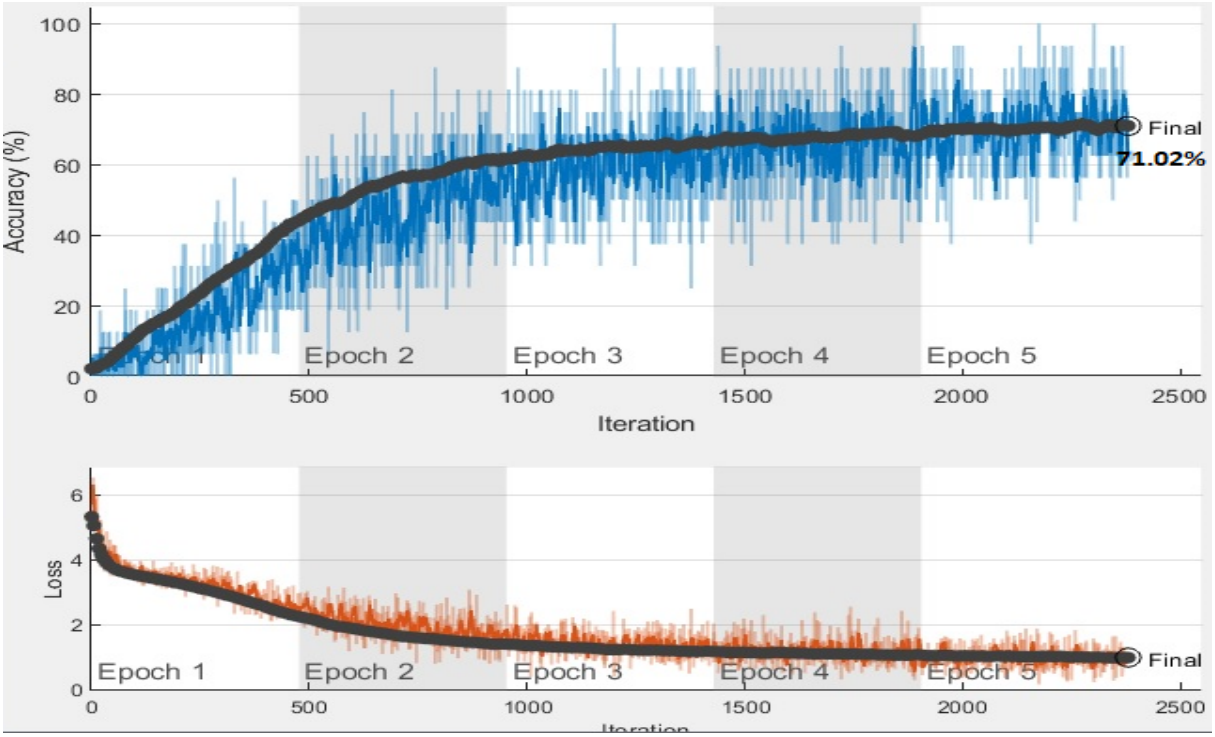


Figure 5. Googlenet's accuracy and loss curves.

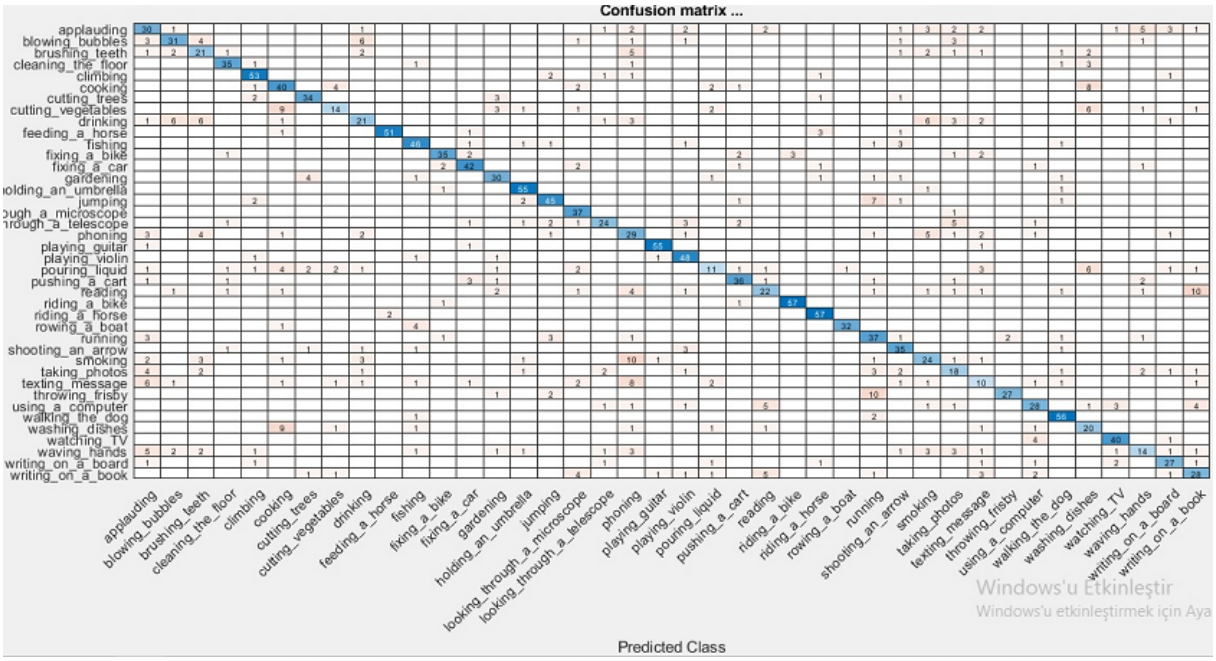


Figure 6. Googlenet's confusion matris.

The accuracy and loss curve obtained with the InceptionV3 model is as in figure 7. An accuracy of 76.15% is achieved with InceptionV3. This rate is the highest accuracy rate among the architectures used. The confusion matrix obtained with InceptionV3 is as in figure 8.

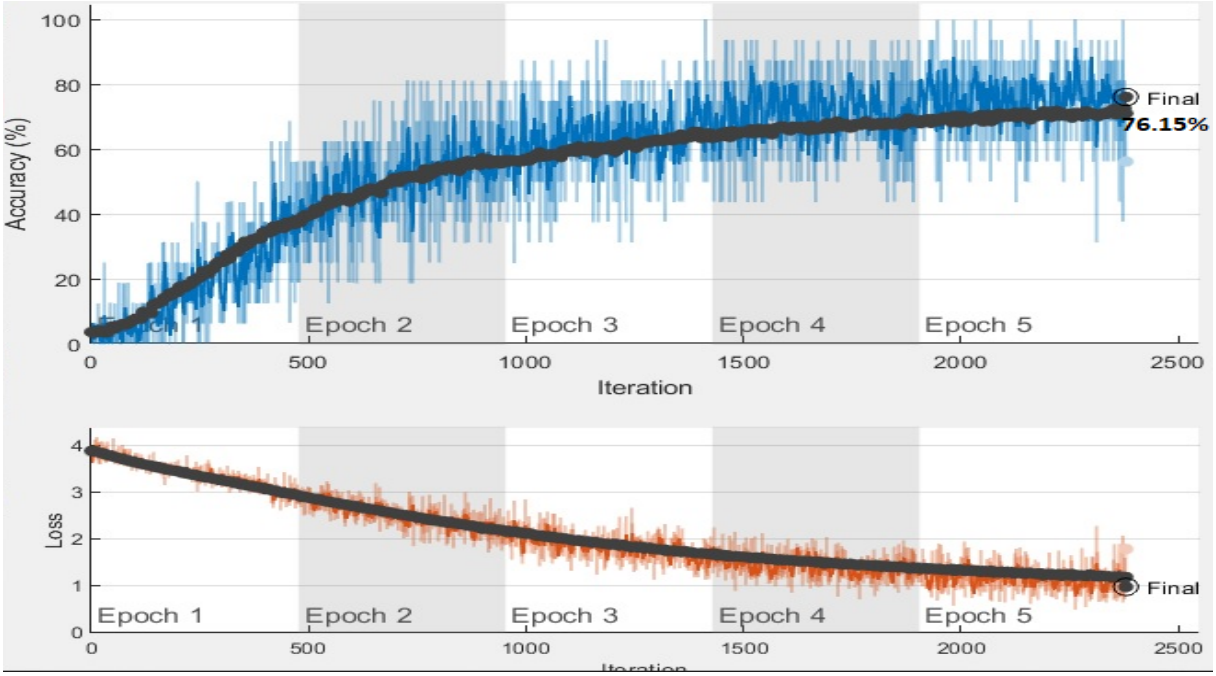


Figure 7. InceptionV3's accuracy and loss curves.

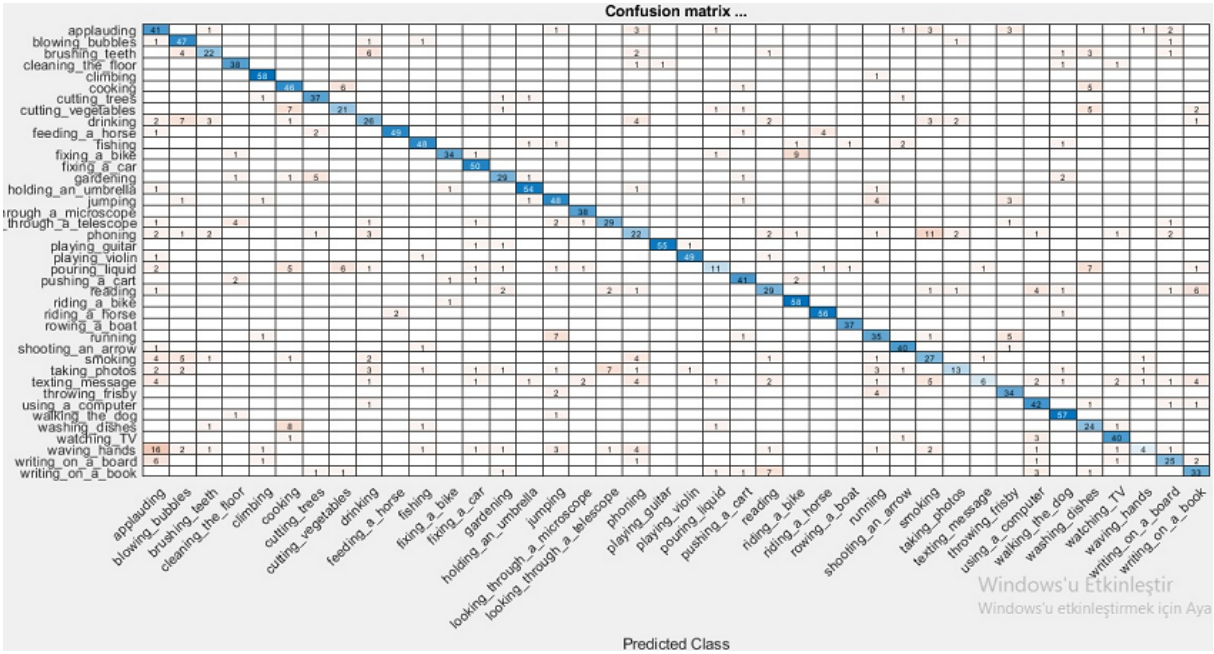


Figure 8. Inceptionv3's confusion matrix.

Classification of human movements is a difficult subject of study due to the large number of human movement class and the similarity of these movements.

4. Conclusions

With the advancing technology, the process of classifying human movements has become an important issue. With the development of technology in recent years, it is of great importance to use confidential data and make inferences from this data. After deep learning architectures became popular, human computer interaction has become one of the most current topics in recent times. There are many studies on the classification of human movements. Classification of human movements is used in many areas from health to safety, from sports to social life, from object tracking to the game industry. In this study, Alexnet, Googlenet and Inceptionv3 models from CNN models were used to classify images of human movements. The highest accuracy value is obtained in InceptionV3 model with 76.15%. Accuracy values of 59.59% are obtained in the Alexnet model and 71.02% in the Googlenet model. In future studies, a new architecture will be studied to improve performance in classifying human movements.

References

[1] Ebrahim, M., Al-Ayyoub, M., & Alsmirat, M. A. (2019, June). Will Transfer Learning Enhance ImageNet Classification Accuracy Using ImageNet-Pretrained Models?. In 2019 10th International Conference on Information and Communication Systems (ICICS) (pp. 211-216). IEEE. DOI: 10.1109/IACS.2019.8809114

[2] AYDIN, İ., & AŞICI, B. (2020). İnsan Hareketlerinin Tanınması için Parçacık Sürü Optimizasyonu Tabanlı Topluluk Sınıflandırıcı Yöntemi. Fırat Üniversitesi Mühendislik Bilimleri Dergisi, 32(2), 381-390. https://doi.org/10.35234/fumbd.671403

[3] Murad, A., & Pyun, J. Y. (2017). Deep recurrent neural networks for human activity recognition. Sensors, 17(11), 2556. https://doi.org/10.3390/s17112556

[4] Batool, M., Jalal, A., & Kim, K. (2019, August). Sensors Technologies for Human Activity Analysis Based on SVM Optimized by PSO Algorithm. In 2019 International Conference on Applied and Engineering Mathematics (ICAEM) (pp. 145-150). IEEE. DOI: 10.1109/ICAEM.2019.8853770

[5] Altun, K., & Barshan, B. (2010, August). Human activity recognition using inertial/magnetic sensor units. In International workshop on human behavior understanding (pp. 38-51). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-14715-9_5

[6] Ke, S. R., Thuc, H. L. U., Lee, Y. J., Hwang, J. N., Yoo, J. H., & Choi, K. H. (2013). A review on video-based human activity recognition. Computers, 2(2), 88-131. https://doi.org/10.3390/computers2020088

[7] Yildirim, M., Çinar, A. (2019). Classification of white blood cells by deep learning methods for diagnosing disease. Revue d'Intelligence Artificielle, Vol. 33, No. 5, pp. 335-340. https://doi.org/10.18280/ria.330502
[8] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).
[9] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).
[10] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).
[11] Stanford University, http://vision.stanford.edu/Datasets/40actions.html
[12] Özkan, İ. N. İ. K., & Ülker, E. (2017). Derin öğrenme ve görüntü analizinde kullanılan derin öğrenme modelleri. Gaziosmanpaşa Bilimsel Araştırma Dergisi, 6(3), 85-104.
[13] YILDIZ, O. (2019). Derin öğrenme yöntemleriyle dermoskopi görüntülerinden melanom tespiti: Kapsamlı bir çalışma. Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi, 34(4), 2241-2260. https://doi.org/10.17341/gazimmfd.435217
[14] Lin, C., Li, L., Luo, W., Wang, K. C., & Guo, J. (2019). Transfer learning based traffic sign recognition using inception-v3 model. Periodica Polytechnica Transportation Engineering, 47(3), 242-250. https://doi.org/10.3311/PPtr.11480
[15] Çinar, A., Yildirim, M. (2020). Classification of malaria cell images with deep learning architectures. Ingénierie des Systèmes d'Information, Vol. 25, No. 1, pp. 35-39. https://doi.org/10.18280/isi.250105
[16] Yildirim, M., Cinar, A. (2020). A deep learning based hybrid approach for COVID-19 disease detections. Traitement du Signal, 37(3): 461-468. https://doi.org/10.18280/ts.370313
[17] Pei, J.Y., Shan, P. (2019). A micro-expression recognition algorithm for students in classroom learning based on convolutional neural network. Traitement du Signal, Vol. 36, No. 6, pp. 557-563. https://doi.org/10.18280/ts.360611
[18] Jiang, X., Chang, L., & Zhang, Y. D. (2020). Classification of Alzheimer's disease via eight-layer convolutional neural network with batch normalization and dropout techniques. Journal of Medical Imaging and Health Informatics, 10(5), 1040-1048. https://doi.org/10.1166/jmihi.2020.3001
[19] Öztürk, Ş., Yigit, E., & Özkaya, U. Fused Deep Features Based Classification Framework for Covid-19 Classification with Optimized MLP. Konya Mühendislik Bilimleri Dergisi, 8, 15-27. https://doi.org/10.36306/konjes.821782
[20] Çinar, A., Yıldırım, M. (2020). Detection of tumors on brain MRI images using the hybrid convolutional neural network architecture. Medical Hypotheses, 139: 109684. https://doi.org/10.1016/j.mehy.2020.109684
[21] Kanda, Y., Sasaki, K. S., Ohzawa, I., & Tamura, H. (2020). Deleting object selective units in a fully-connected layer of deep convolutional networks improves classification performance. arXiv preprint arXiv:2001.07811.
[22] Kadam, V. J., Jadhav, S. M., & Vijayakumar, K. (2019). Breast cancer diagnosis using feature ensemble learning based on stacked sparse autoencoders and softmax regression. Journal of medical systems, 43(8), 1-11. https://doi.org/10.1007/s10916-019-1397-z
[23] Zeng, G. (2020). On the confusion matrix in credit scoring and its analytical properties. Communications in Statistics-Theory and Methods, 49(9), 2080-2093.https://doi.org/10.1080/03610926.2019.1568485