International Journal of Innovative Research and Reviews, 7(1) 22-26

International Journal of Innovative Research and Reviews
ISSN: 2636-8919
Website: www.injirr.com
doi:
Research paper, Short communication, Review, Technical paper

RESEARCH ARTICLE

Video Classification Results with Artificial Intelligence and Machine Learning

Elif AKARSU[1,*], Tevhit KARACALI[1]

[1] *Department of Electrical-Electronics Engineering, Faculty of Engineering, Atatürk University, Erzurum, Türkiye*

** Corresponding author E-mail: elif.akarsu@atauni.edu.tr*

ARTICLE INFO

Received : 05.22.2023
Accepted : 07.07.2023
Published : 07.15.2023

Keywords:
Artificial Intelligence
Classification
Video Processing
Machine Learning

ABSTRACT

The study is related to the classification of the videos of the UCF101 dataset obtained from kaggle with the help of artificial intelligence and machine learning. The ucf 101 dataset has six classes and 155 videos in each class, each of which has approximately 150 picture frames. and with 3 different preprocessing algorithms, features were obtained from each picture frame, and 3 different accuracies were obtained by sending them to the LSTM classifier and the obtained results were compared with each other. In the classification process, cross validation was used to confirm the accuracy obtained.

Contents

1. Introduction

Video classification is a broad topic. We can also call it video comprehension or video identification. It also includes many visual and auditory factors in its content. this creates a wide range of classifications [1]. Video classification is a very difficult event in terms of content. However, in terms of content, the attributes to be obtained depending on sound, text, image, movement and gesture processing vary [2]. If movement classification is done, what the movement or gesture is and the duration of the movement or gesture are important parameters.

Vision-based human gesture recognition involves predicting a gesture such as saluting, sign language gestures, or clapping using a series of video frames. One of the attractive features of gesture recognition is that it makes it possible for people to communicate with computers and devices without the need for external input equipment such as a mouse or remote control. It has many applications, from video to motion recognition, control of consumer electronics and mechanical systems, robot learning to computer games. For example, online prediction of multiple actions for videos from multiple cameras can be important for robot learning.

Compared to image classification, modeling human motion recognition using videos is difficult due to the large amount of false ground truth data for video datasets, the variety of motion that actors in a video can perform, datasets with a large class imbalance, and the large number of datasets. datasets. data needed to train a robust classifier from scratch. Deep learning techniques such as SlowFast two-way convolution networks [3].

Video is a digital sequence of multiple images. each frame of the video is a picture frame. and by extracting the features

Cite this article Akarsu E, Karacalı T. Video Classification Results with Artificial Intelligence and Machine Learning. *International Journal of Innovative Research and Reviews (INJIRR)* (2023) 7(1) 22-26
Link to this article: http://www.injirr.com/article/view/194

of these pictures, the attribute matrix of that video is created. this matrix can be determined in desired amounts from certain regions of each frame. but the most important factor here is to get enough number of attributes. Matrices are created with more than one feature extraction algorithm. These algorithms differ according to the number of layers and their depth.

Deep learning models, specifically convolutional neural networks (CNNs), are well known for understanding images. A number of CNN architectures are proposed and developed in the scientific literature for image analysis. Among these, the most popular architectures are LeNet-5 [4], AlexNet [5], VGGNet [6], GoogleNet [7], ResNet [8], and DenseNet [9]. These algorithms are feature extraction algorithms. and they allow us to obtain matrices containing all the information of the video. These matrices are obtained from each frame and form a whole.

The simplest form of these algorithms used today is Alexnet. considering both the number of layers and the amount of convolution layers, it is simpler than other pre-training algorithms. The Vgg19 algorithm includes 47 layers and 19 deep layers. more extensive data and higher-level features are obtained. However, the most comprehensive algorithm used in this study is the Resnet18 algorithm. It contains 71 layers and 18 deep layers. at this point the number of deep layers is less than Vggnet. Each of them has many different characteristics.

The important thing is which of these features is high. The effect of this situation will also be investigated. In this study the effect of three different pre-training stages on video classification was investigated. and it is aimed to determine the algorithm that gives the most accurate results. The classification of all these pre-training algorithms is done with the LSTM classifier.

2. Material and Method

In this study, the UCF 101 dataset obtained from Kaggle was used. The dataset contains information on many human activities. And what we need to do is to perceive what these activities are and to categorize each activity. A 6-class video file is used. Figure 2 shows the classes of the 6-class video file. There are about 150 videos in each class and each picture contains 240*320 pictures.

These videos were classified using 3 different preprocessing algorithms. these algorithms are Alexnet, Vgg19 and Resnet18. 1000 features are taken in the fully connected layer of each algorithm. these attributes are taken from the frame of each video. AlexNet is a convolutional neural network that is 8 layers deep. The pre-trained network can classify images into 1000 object categories, such as keyboard, mouse, pencil, and many animals. As a result, the network has learned rich feature representations for a wide range of images. The network has an image input size of 227-by-227 [10]. Alexnet consists of 25 layers in total. 1000 attributes are taken from the last layer.

VGG-19 is a convolutional neural network that is 19 layers deep. The pretrained network can classify images into 1000 object categories, such as keyboard, mouse, pencil, and many animals. As a result, the network has learned rich feature representations for a wide range of images. The network has an image input size of 224-by-224 [10].

ResNet-18 is a convolutional neural network that is 18 layers deep. The pretrained network can classify images into 1000 object categories, such as keyboard, mouse, pencil, and many animals. As a result, the network has learned rich feature representations for a wide range of images. The network has an image input size of 224-by-224 [10]. After obtaining the feature extraction feature matrix, feature grouping and video classification processes for each method were performed in MATLAB 2019. The MATLAB code for the attributes to be sent at a certain rate before they are sent for classification is given in Figure 1.

```
numberofTraindata = numel(Traindata);
TraindataLengths = zeros(1,numberofTraindata);

for i = 1:numberofTraindata
    series = Traindata{i};
  TraindataLengths(i) = size(series,2);
end

figure
histogram(TraindataLengths)
title("Sequence Lengths")
xlabel("Sequence Length")
ylabel("Frequency")
```

Figure 1 The MATLAB code used in the study

The results obtained in each pre-training algorithm are divided into 5 parts and the training is done in 5 folds in order to prevent overfitting and to obtain more reliable results. After the features are extracted, they are subject to a classification process, which will be performed with the LSTM classifier. These attributes are sent to the time dependent LSTM classifier. The LSTM classifier is a time-dependent classification mechanism. The central role of an LSTM model belongs to a memory cell that maintains its state over time, known as the "cell state". The cell state is the horizontal line that goes over the top of the diagram below. It can be visualized as a conveyor belt where information flows unchanged. The video classification algorithm is also shown Figure 3.

Each video is composed of many frames juxtaposed and the attributes of each are obtained. at this point we will analyze each frame of the picture. In Figure 2, 6 different frame information belonging to 6 different video frames is presented as an example. Each class contains many different activities. they are quite unrelated to each other. This makes a significant contribution to the classification.

Figure 2 Videos of Six Classes

LSTM Classifier; recursive neural networks are frequently used for modeling sequential data such as text, audio, and video. Unlike classical neural networks, this model feeds back the inputs from the next layers. In summary, recursive neural networks examine the information in the input data by considering the value of the previous output. Theoretically, at time t, information from all previous steps can be retained, but in practice it becomes impossible to learn long-term requirements. This situation is known as gradient disappearance in the literature. In other words, as you add layers in very deep forward propagation networks, the network is untrainable. A long-short-term memory approach has been proposed to solve this problem. The long-term memory model consists of four layers. An ordinary LSTM unit consists of a cell, input, output, and forget layer. The cell remembers values at arbitrary time intervals, and the three layers regulate the flow of information into and out of the cell [11–13]

Deep learning is a sub-branch of machine learning and allows computers to learn from datasets of complex structures [14]. Deep learning, performed using artificial neural networks, learns patterns and relationships by processing data, similar to the functioning of the human brain. This technology is used in many different fields. For example, it can be used in image recognition, natural language processing, speech recognition, automated driving, game strategies, and more. For the machine learning field, deep learning is very important. Traditional machine learning models process datasets based on human-specified features. However, thanks to deep learning, computers identify features in data sets themselves and obtain more accurate results. Therefore, deep learning is considered a revolutionary development in the field of machine learning.

We can better explain the usage area of deep learning with an example story. Let's say an e-commerce site wants to analyze the purchasing habits of users. A traditional machine learning model predicts users' purchasing habits based on certain characteristics such as gender, age, regional location. However, when deep learning is used, different features that affect users' purchasing habits are automatically determined. For example, the deep learning model makes more accurate predictions by analyzing what hours users shop, what products they search for, and how often other users purchase similar products.

In short, deep learning is a very important development in the field of machine learning and can be used in many different areas [15]. Thanks to this technology, computers can get more accurate results and process large data sets that humans cannot.

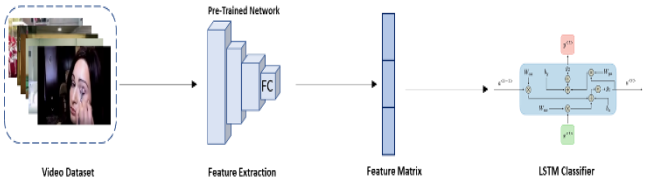

Figure 3 Video Classification Algorithm

In Figure 3, we see all the video processing stages. Here, with the help of the pre-training algorithm, the features of each frame are extracted with 3 different methods. Extracted attributes is sent to the LSTM classifier in the form of a matrix and classification is performed for 6 different classes. The classification process was done with 5 folds. The reason for this is to obtain more reliable results and to prevent the overfitting problem.

3. Results

There are approximately 150 videos in each of the 6 different classes. that is, a total of 900 videos were classified. and each video has about 120 frames. The classification process was based on recognizing motion in videos. When each class is examined, it is seen that they contain quite different movements from each other. This has a positive effect on stability. Classifications related to each pre-training algorithm were performed with 5 folds and 3 different accuracies were obtained.

Table 1 Accuracies of Three Different Feature Extraction Algorithms

Pre-Training Algorithm	Accuracy	Error Rate
Alexnet	%91.7	%8.3
Vgg19	% 93.8	% 6.2
Resnet18	%98.8	% 1.2

The accuracy here is highly affected by both the number of layers, the number of features and the deep layers. Table 1 shows the accuracy and error rates obtained with the three different algorithms. Accuracies, error rates and distribution of the features obtained for alexnet, resnet and vggnet were examined and a result was obtained. When the number of layers and deep networks are examined [16], Resnet18 architecture gives the most accurate result. followed by a sub-version of it, Vgg19, and the latest, simplest algorithm, Alexnet, with an accuracy of 91.7%. For Vgg19 and Resnet 18, the results are respectively% 93.8 and 98.8.

In Figure 4, the distribution of the features of 3 different pre-training algorithms is shown. It is desired to obtain 1000 features for each frame from the last layer of each feature extraction algorithm. How many frames there are in a video, 1000 attributes are taken for each frame. For a complete video file, the number 1000 is the number of rows in the matrix. represents the number of columns of each sequence length matrix seen in the histograms. frequency is how many times it is sent to be classified in this number. A certain amount of attributes of each algorithm is sent. Care was taken not to exceed 420. There is a significant decrease in the accuracies achieved when going above 420. Attribute submission amounts are between 260 and 420 for each. this is a situation that should be considered for more accurate

results [17]. Each histogram is 1*144 in size. these come together to form the video footage.

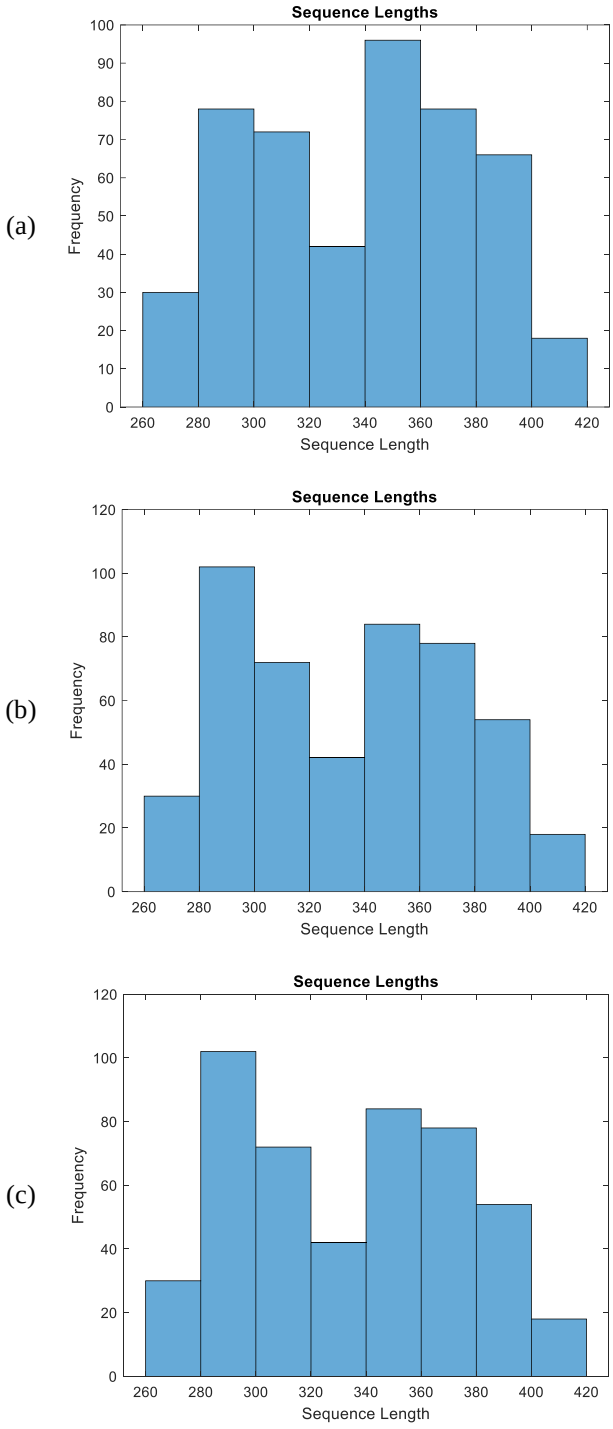

Figure 4 Graph representing the size of the attributes being sent to the network. Exclusion from the larger than 600. (a) Alexnet (b)Vgg19 (c) Resnet18

We see the feature distribution histograms of 3 different pre-training algorithms. For each, different amount of feature distributions were obtained at different points. The very high or very small amounts of these distributions greatly affect the accuracy. That's why these charts are very important.

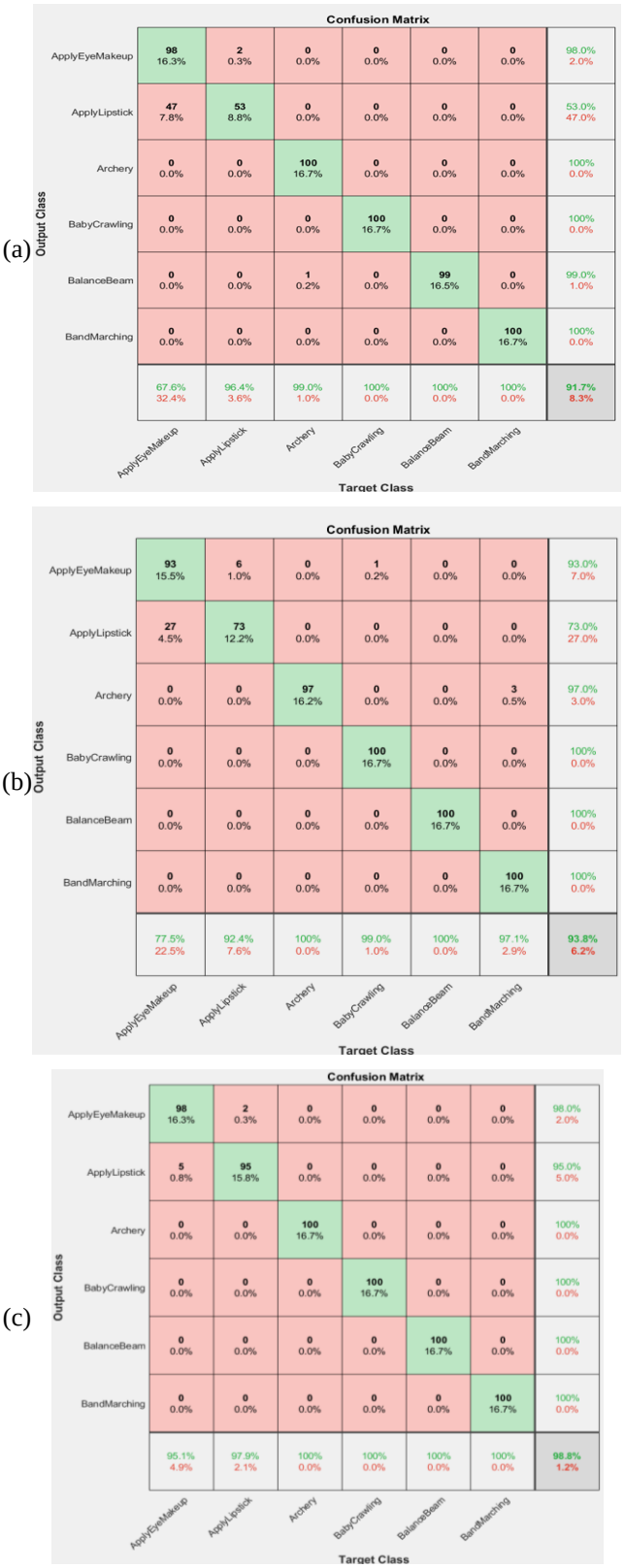

Figure 5 (a) Alexnet confusion matrix (b) Vgg19 confusion matrix (c) Resnet18 confusion matrix

Figure 5 shows the confusion matrices of the classification made. these matrices show the accuracies of 6-class video files. Although the result of each lagorithm is different, the results of 3 different pre-training algorithms overlap with each other.

The diagonal parts of the confusion matrix show correct predictions, while the other parts show incorrect predictions [18]. Therefore, the more data on the diagonals, the more accurate our prediction is. With 3 different pre-training algorithms, 3 different confusions were obtained. we have met the accuracy of each of them.

4. Discussion

Alexnet; “ImageNet” had succeeded in significantly increasing the classification accuracy. It consists of 5 “convolutional layers” and 3 “fully connected” layers. AlexNet uses ReLu (Rectified Linear Unit) as activation in non-linear parts [19]. Previous standard neural networks used tanh or sigmoid. In Vgg19, kernel sizes were not fixed in AlexNet, it started with the first 11 and continued as 5 and 3. VGG16 has fixed kernel dimensions. The idea behind this was that 11x11 and 5x5 kernels could be replicated with multiple 3x3 kernels. The total number of "convolutional" and "fully connected layers" of VGG16 is 16 [20]. It also has another version, VGG19. VGG16 and VGG19 are very similar, the only difference is that the number of layers is different. For Resnet 18, a technique called skip connection is used in this network. Skip Connections skips several layers and connects directly to the output. In this way, the problem of exploding / vanishing gradient is avoided. Exploding gradients, on the other hand, is the opposite of vanishing, which is the overgrowth of gradients. The results of 3 algorithms with all these advantages and disadvantages are the subject of this study. In this study, video classification was made with three different pre-processing algorithms and the results of these three different algorithms were compared. and as it can be seen from here, the most accurate results were obtained with the Resnet18 architecture, which contains 71 layers and has more accurate attribute information. followed by 47-layer Vgg19 and the simplest 25-layer Alexnet gave the most incorrect result. As can be seen from here, as the number of convolution layers increases, high-level features are obtained and this significantly affects the test accuracy obtained.

5. Conclusion

With the confusion matrices, we see the results from each pre-training algorithm. here is a video file with 6 classes and the main thing is to make motion classification. When examined from this point of view, it is seen that there are 6 classes that are very independent from each other. This contributes positively to the success of the classification. Also, cross validation has a great importance for our classification. If we examine this importance; Overfitting is prevented with K-fold and three different accuracies are obtained. it is also an important component to send the attributes to classification at a certain rate. Therefore, the graphs shows that the number of features in each series should not exceed 600.The classification of the six-class video dataset with three different pre-training algorithms is emphasized, and the results are compared. The most accurate test accuracy was found to be 98.8%. As a result Resnet18 gave the most successful results, followed by Vgg19 and Alexnet.

Conflict of Interest

The authors declared no conflict of interest.

References

[1] Hochreiter S, Schmidhuber J. Long short-term memory. *Neural computation* (1997) **9**(8):1735–1780. doi:10.1162/neco.1997.9.8.1735.

[2] Roach MJ, Mason J, Xu L, Stentiford F, Heath M. Recent Trends In Video Analysis: A Taxonomy Of Video Classification Problems. *Proceedings of the 6th International Conference on Internet and Multimedia Systems and Applications (IASTED)* (2002).

[3] Christoph, F., Haoqi, F, Jitendra M, Kaiming H. SlowFast Networks for Video Recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019). p. 6202–6211.

[4] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based Learning Applied to Document Recognition. *Intelligent Signal Processing* (2001) **86**(11):2278–2324. doi:10.1109/5.726791.

[5] Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. *Advanced Neural Information Processing Systems* (2012) **2**:1097–1105. doi:10.1145/3065386.

[6] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: *Proceedings of the 3rd International Conference on Learning Representations* (2015). p. 1–14.

[7] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2015). p. 1–9.

[8] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2016). p. 770–778.

[9] Huang G, Liu Z, Van L, Weinberger KQ. Densely connected convolutional networks. In: *Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition* (2017). p. 2261–2269.

[10] Yang K, Qinami K, Fei-Fei L, Deng J, Russakovsky O. Towards fairer datasets. In: Hildebrandt M, Castillo C, Celis E, Ruggieri S, Taylor L, Zanfir-Fortuna G, editors. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM (2020). p. 547–558.

[11] Chollet F. *Python ile Derin Öğrenme [Deep Learning with Python]*. Ankara: Buzdağı Yayınevi (2019). 1–52.

[12] Ayyüce Kızrak M, Bolat B. Derin Öğrenme ile Kalabalık Analizi Üzerine Detaylı Bir Araştırma [A Comprehensive Survey of Deep Learning in Crowd Analysis]. *Bilişim Teknolojileri Dergisi* (2018) **11**(3):263–286. doi:10.17671/gazibtd.419205.

[13] Shervine, A, Afshine, A. *Recurrent Neural Networks cheatsheet*. Lecture Notes for the Course CS230:Deep Learning (2022).

[14] Irene A, Gianmarco B, Francesco L. Image and Video Forensics. *Journal of Imaging* (2021):7–242. doi:10.3390/jimaging7110242.

[15] Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN computer science* (2021) **2**(3):160. doi:10.1007/s42979-021-00592-x.

[16] Walther J. *Hierarchical Electrical Load Forecasting of Industrial Production Systems in the Manufacturing Industry based on Deep Learning*. PhD Thesis. Fachbereich Maschinenbau an der Technischen Universität Darmstadt. Darmstadt (2022). doi:10.26083/TUPRINTS-00021767.

[17] Dimitrova N, Agnihotri L, Wei G. Video classification based on HMM using text and faces. In: *10th European Signal Processing Conference* (2000). p. 1–4.

[18] Abhale AB, Manivannan S.S. Deep Learning Algorithmic Approach for Operational Anomaly Based Intrusion Detection System in Wireless Sensor Networks. *Pre-Print at Research Square* (2021):1–29. doi:10.21203/rs.3.rs-777010/v1.

[19] Özkara C, Ekim P. Real-Time Facial Emotion Recognition for Visualization Systems. In: *2022 Innovations in Intelligent Systems and Applications Conference* (2022). p. 1–5.

[20] Thilagaraj M, Arunkumar N, Petchinathan G. Classification of Breast Cancer Images by Implementing Improved DCNN with Artificial Fish School Model. *Computational Intelligence and Neuroscience* (2022)(Special Issue: Mental Illness Detection and Analysis on Social Media). doi:10.1155/2022/6785707.