

Lip reading using deep learning in Turkish language

Hadi Pourmoussa^{1,2}, Üstün Özen¹

¹Department of Management Information Systems, Faculty of Economics and Administrative Science, Atatürk University, Erzurum, Türkiye

²Department of Computer Engineering, Faculty of Engineering, Atatürk University, Erzurum, Türkiye

Article Info

Article history:

Received Oct 21, 2023

Revised Jan 27, 2024

Accepted Feb 10, 2024

Keywords:

Convolutional neural networks
Dataset
Deep learning
Lip-reading
Turkish language

ABSTRACT

Computer vision is one of the most important areas of artificial intelligence and lip reading is one of the most important areas of computer vision. Lip-reading, which is more important in noisy environments or where there is no sound flow, is one of the working areas that can help the hearing-impaired people. There is no dataset in Turkish for lip reading, which there are different datasets at alphabet, word, and sentence level in different languages. The dataset of this study was created by the author and video data were collected from 72 people for 71 words. Audio streams were removed from the collected videos and a dataset was created using only images. Due to the small size of the dataset, the data was replicated with the Camtasia application. After the model of the research was designed and trained, the model was tested on adjectives, nouns, and verbs dataset and success rates of 71.8%, 71.88%, and 79.69% were obtained, respectively.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Hadi Pourmoussa

Department of Management Information Systems, Faculty of Economics and Administrative Science

Atatürk University

Erzurum, Türkiye

Email: hadi.pourmoussa14@ogr.atauni.edu.tr

1. INTRODUCTION

The lip-reading process, which tries to understand what the speaker is saying from lip movements [1], is one of the most important field of human action recognition [2]. Lip reading which is performed using only visual information, is an impressive skill in noisy environments or when there is no audio streaming [3]-[5]. Because, the main purpose in lip reading studies is to detect and understand expressions that are spoken only with images without audio flow [6]-[8].

In recent years, lip-reading studies have started to increase with the widespread of deep learning [6]. The lip-reading process is applied at the letter, number, syllable, word, and sentence level [6], [9]. The lip reading is process is applied at the level of letter, number, syllable, word, and sentence [6], but in most studies it has been applied at the level of alphabet, word, and sentence [9].

The most important step in lip-reading studies is to acquire the mouth images of the speaker with some predefined point coordinates. Then, the movements of these points are analyzed by some classification methods, such as k-nearest neighbor classifier, hidden Markov models, and artificial neural networks. to understand what the speaker is saying [10]. Automatic lip reading is used in different fields. However, there are many challenges to be overcome in this area in Turkish language.

Guttural sounds (K, G, Ğ). Having multiple dialects (kardeş (brother) is pronounced as gardaş). Making some words and verbs meaningless by shortening (gidiyorum (I am going) verb changes to gidiyom). Absence of lip movement in some words (iyi (good), 40, and 2). Different datasets in different languages were created for the lip-reading area. Some of the important datasets created are presented in Table 1.

Table 1. Some of the important lip-reading datasets

Dataset	Language	Year	Number of dataset	Type	Number of class
TULIPSI [11]	English	1995	12	Number	4
M2VTS [12]	English	1997	37	Number	10
AVLetters [13]	English	1998	10	Alphabet	26
WAPUSK20 [14]	English	1999	20	Sentence	52
IBMViaVoice [15]	English	2000	290	Sentence	10500
AV@CAR [16]			208	Alphabet	26
	English	2004	20	Number	10
			20	Sentence	250
CUAVE [17]	English	2004	36	Number	10
UWB-05-HSAVC [18]	Czech	2005	100	Sentence	200
AVLetters2 [19]	English	2008	5	Alphabet	29
CENSREC-1-AV [20]	Japanese	2010	42	Number	10
NDUTAVSC [21]				Number	10
	German	2010	66	Word	6907
				Sentence	6907
OULUVS2 [22]				Number	10
	English	2010	53	Expression	10
				Sentence	540
LTS5 [23]	French	2012	20	Number	10
MIRACL-VC [24]	English	2012	15	Word	10
				Expression	10
HAVRUS [25]	Russian	2016	20	Sentence	1530
AV Digits [26]	English	2018	53	Number	10
			39	Expression	10
LRW-1000 [27]	English	2018	2000'den fazla	Word	1000
AVSD [28]	Arabic	2019	22	Expression	10
NSTDB [4]	Chinese	2020	349	Word	

2. RELATED WORK

In recent years, lip-reading studies have been carried out using deep learning techniques and datasets in different languages have been created and studies are increasing. Most of the studies have been carried out in English and different datasets have been created in English, unlike other languages, until today. However, today, datasets for other languages have begun to be created, even if they are small. In Turkish, a dataset consisting of 70 people was created for 20 numbers by Pourmousa and Özen [29], and it was trained and tested with convolutional neural network. In the study, 56.25% successes were achieved and it was reported that the absence of lip movements in some Turkish numbers and the similarity of lip movements in some numbers significantly affect the success.

Atila and Sabaz [6] created two new Turkish datasets, one with 111 words and the other with 113 sentences. In this study, pre-trained models and the bidirectional long short-term memory (Bi-LSTM) method were used to perform the classification. GoogleNet, ResNet-101, ResNet-50, ResNet-18, Nasnet-Large, Xception, DarkNet53, DarkNet19, AlexNet, Squeezenet, and DenseNet201 were used as a pre-trained model and success of models were investigated. As a result of the study, when Resnet-18 and Bi-LSTM were used together, the highest success was achieved with 84.5% at word level and 88.55% at sentence level.

Nambeesan *et al.* [30] developed a lip-reading system on the MIRACL-VC1 dataset by using the long short-term memory method. In this study, lip regions were extracted from all frames of a word video and recorded sequentially in a single photograph. As a result of the study, 85.5% accuracy was obtained. Sarhan *et al.* [31] developed a hybrid lip-reading model which is based deep convolutional neural network for lip reading from videos. The proposed model consists of preprocessing, encoder, and decoder stages to perform lip-reading. As a result of the study, 92% success was achieved in the unseen speaker and 99% in the overlapped speakers.

Ma *et al.* [32] developed a lip-reading system using convolutional neural network on lip reading in the wild (LRW) and LRW-1000 dataset. In this study, 88.5% accuracy was obtained on the LRW dataset and 46.6% on the LRW-1000 dataset. Elrefaei *et al.* [28] proposed a lip-reading method for Arabic language. In this study, 22 people were participated and recorded video by their smartphones for 10 words. As a result of the study, 79% accuracy was obtained. Noda *et al.* [5] used convolutional neural network method as feature extraction mechanism for visual speech recognition in their study titled lip reading using convolutional neural network. In this study, the convolutional neural network was trained using images of the speaker's mouth region. The proposed system was evaluated on an audio-visual speech dataset containing 300 Japanese words with six different speakers, and 58% success was achieved as a result of the study.

3. DATASET

In this section, the details of the dataset used are explained. First, the characteristics of the words used in the data set and how they were collected are explained, then how the dataset was created and the pre-processing done on the data set are explained. Finally, an example of the dataset that should be given as input to the system is presented.

3.1. Dataset properties

The dataset was created by the author. First of all, the most used adjectives, nouns and verbs in Turkish were found and some of them were chosen randomly [33], [34]. In the selection of these words, attention was paid to the selection of words with lip movements, and words with no lip movements such as "iyi" (good) were not chosen. In this framework, words are presented in Table 2. In this context, 71 words, including 19 adjectives, 33 nouns and 19 verbs, were selected to create the dataset.

Table 2. Turkish words dataset

Dataset	Word
Adjectives (19 Words)	büyük (big), küçük (small), başka (another), önemli (important), doğru (correct), güzel (beautiful), yüksek (high), kolay (easy), mümkün (possible), sürekli (continually), yavaş (slow), ekonomik (economic), beyaz (white), siyah (black), sosyal (social), uluslararası (international), tamam (ok), temiz (clean), lazım (required)
Nouns (33 Words)	parça (piece), hizmet (service), bilgisayar (computer), televizyon (television), masa (table), merkez (centre), ortam (environment), araba (car), sistem (system), alışveriş (shopping), abla (elder sister), fırsat (opportunity), piyasa (market), teknoloji (technology), fiyat (price), demokrasi (democracy), yumurta (egg), insan (human), zaman (time), dünya (world), hayvan (animal), baş (head), ülke (country), para (money), baba (father), devlet (government), bölüm (section), banka (bank), toplum (society), program (program), çalışma (work), cumhuriyet (republic), sigorta (insurance)
Verbs (19 Words)	bakmak (look), yapmak (do), beklemek (wait), bilinmek (be known), başlatmak (start), belirlemek (determine), bozmak (disrupt), kapatmak (close), kaybetmek (lose), vurmak (hit), toplamak (collect), fark etmek (notice), paylaşmak (share), satın almak (buy), düşünmek (think), uğraşmak (strive), evlenmek (marry), öldürmek (kill), geliştirmek (develop)

3.2. Dataset pre-processing

First, each word was extracted with the Camtasia application, and then new videos were created by turning them at different angles due to the small dataset. As seen in the Figure 1, five videos were created from each video by rotating it at three different angles, and thus, the dataset was reproduced. The video with the most frames was detected and the frame count (48) was recorded for further processing. The lip area was cut with the mediapipe library in each frame of the videos. There are 468 points in the face region in the Mediapipe library as shown in the Figure 2.

To create the final entry in the data set, the frames of the videos were first calculated and the highest frame was obtained. The maximum number of frames was obtained as 48 and the closest Square number was 49, so the designed photograph consisted of 7 rows and 7 columns. Then the dimensions of the lip areas were changed to 64×64 and then saved in a file as in Figure 3. Frames obtained from each video and recorded in a file were arranged sequentially on a photo, and a single photo was obtained for each video, as seen in Figure 4. Since each frame consists of 64×64 dimensions, the width and length of the single photograph designed consists of 7×64 length. Thus, the size of the photo was obtained as 512×512 and the frames were arranged in this photo.

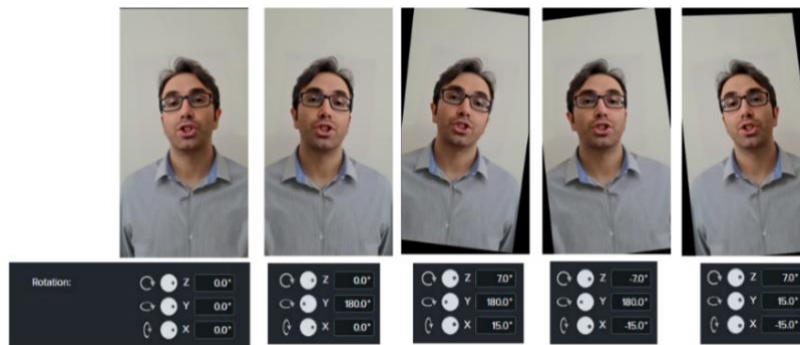


Figure 1. Five videos of word with rotation at different angles

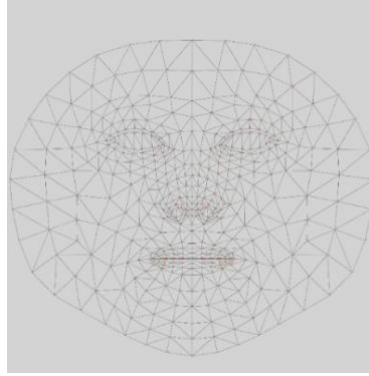


Figure 2. 468 points used for face in mediapipe library

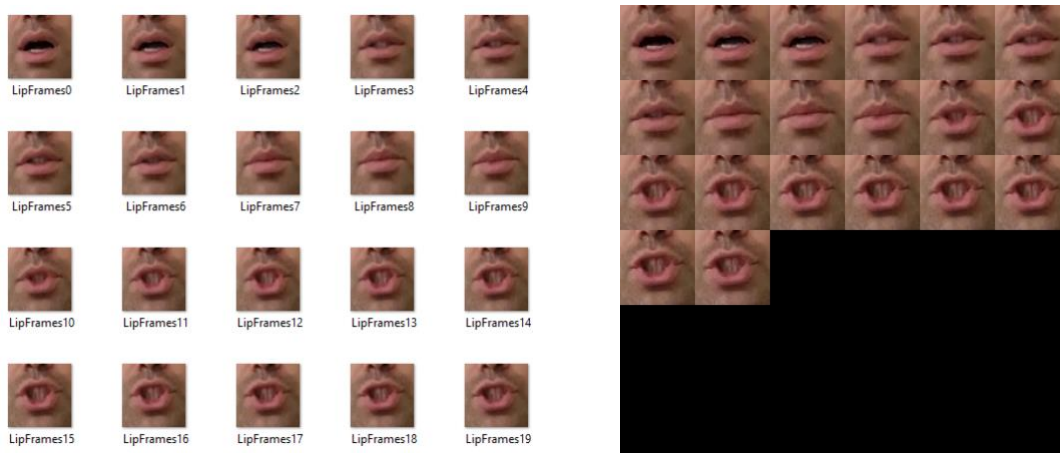


Figure 3. Recording of lip movement as frames Figure 4. Sequential recording of lip frames in one image

The dataset prepared as mentioned above is divided into training and validation datasets. The test dataset is completely different and consists of data not found in the training and validation dataset. The number of data in the each class was not equal due to the fact that some words were said incorrectly or there was an error in the video while saying it. Therefore, the training dataset contains at least 295 data for each class, 25 for validation and 20 for testing.

4. CONVOLUTIONAL NEURAL NETWORKS

Convolutional neural network is a type of feed forward artificial neural network proposed by LeCun *et al.* [35] to extract features from data that has a grid pattern such as images and videos with convolutional structures [36], [37]. Convolutional neural network is a mathematical structure that consists of three types of layers: convolution, pooling, and fully connected layers. While the convolution and pooling layers are used for feature extraction from the input images, the fully connected layers use the extracted features for classification, which are the result stage [37], [38]. Multiple convolution layers and pooling are used to extract high-level features, and the output of each layer depends on the next layer [39]. Convolutional neural networks can use for any form of data as input, such as images [40], video [41], natural language [42], audio [43], and speech [44]. The visual geometry group (VGG) model, which is an important example of convolutional neural network architecture and proposed by Simonyan and Zisserman [45], is shown in the Figure 5.

As Figure 5, there are different layers in the network, and each layer can have several convolutional layers. A pooling layer can be applied after one or more convolution layers. Finally, there are one or more fully connected layers. The purpose of the convolution layer is to extract and learn the sensitive and high-level features of the input data (image, video or audio) using the local neurons in the previous layer [46], [47]. The convolution layer creates feature maps by applying kernels of different sizes and weights to the input queue. A completed convolution layer consists of multiple feature maps with different weight vectors so that multiple

features can be obtained from each location [35]. The pooling layer, which is usually added after a convolution layer, is an important step in convolution-based systems that reduce the dimensionality of feature maps [48]. The two main purposes of this layer are, firstly, to reduce the number of parameters or weights, thereby reducing the computational cost, and secondly, to control overfitting by reduce the spatial size of the data [49], [50]. Generally, two types of pooling methods are used, maximum pooling and average pooling (Figure 6).

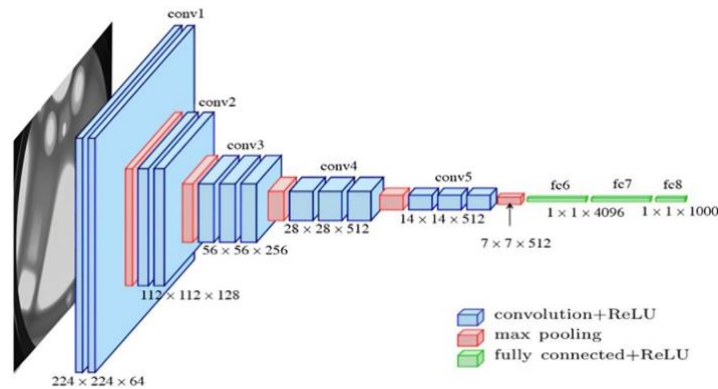


Figure 5. An important example of convolutional neural network architecture [45]

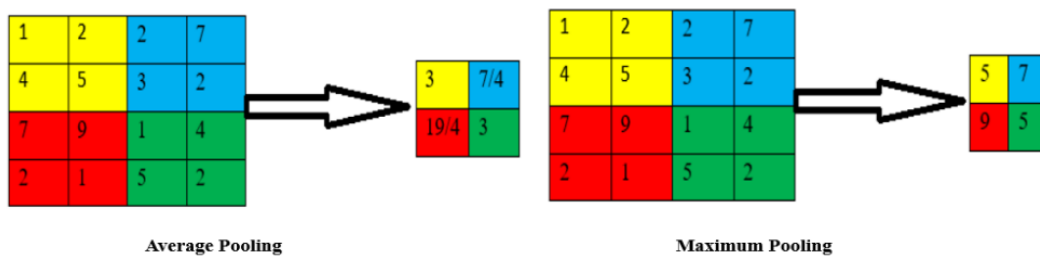


Figure 6. Pooling methods [51]

5. RESEARCH METHOD

In this study, convolutional neural networks with high success in image processing and classification were used. Because all images are divided into frames and recorded on a single image. Since there are not many features in the images, the images are given to the system in grayscale to increase the speed of the study. As seen in the model presented in the Figure 7, firstly the dataset was changed to 224x224x1 and sizes of all the images were reduced and changed to grayscale images to increase the speed of the system.

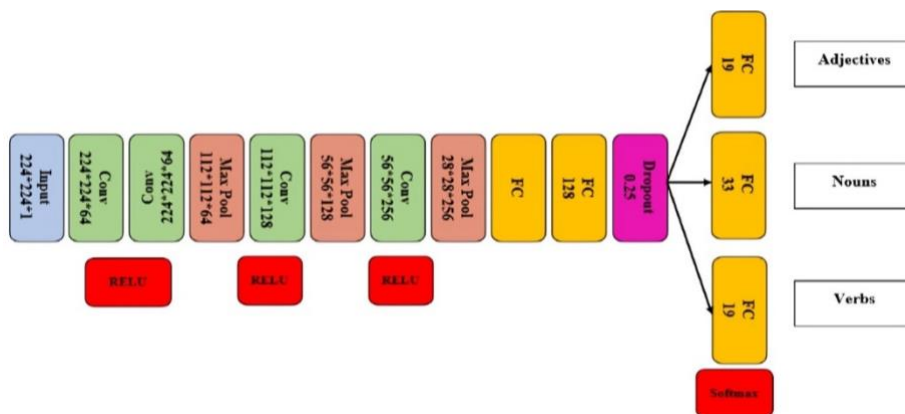


Figure 7. Proposed convolutional neural network model

6. RESULTS AND DISCUSSION

The model was run 3 times as there were different results. The model presented in Figure 7 are trained on the adjectives, nouns, and verbs dataset. For training, the model was run 50 epoch and the adam function was used as the optimizer. The training and validation accuracy and the training and validation loss for the adjectives dataset are shown in Figure 8. Training accuracy was 82.01% and validation accuracy was 81.26%.

The training and validation datasets are likely to be similar, but the test dataset consists of completely different data and is given to the system only while it is being tested. The success rate of the adjectives dataset was 75% on the test set (Figure 9). Elrefaei *et al.* [28] achieved 70.09% success on word dataset in Arabic. Garg *et al.* [7] achieved 56% success at word level on MIRACL-VC1 dataset in English. Chen *et al.* [4] achieved 73.19% success in Chinese with LipNet.

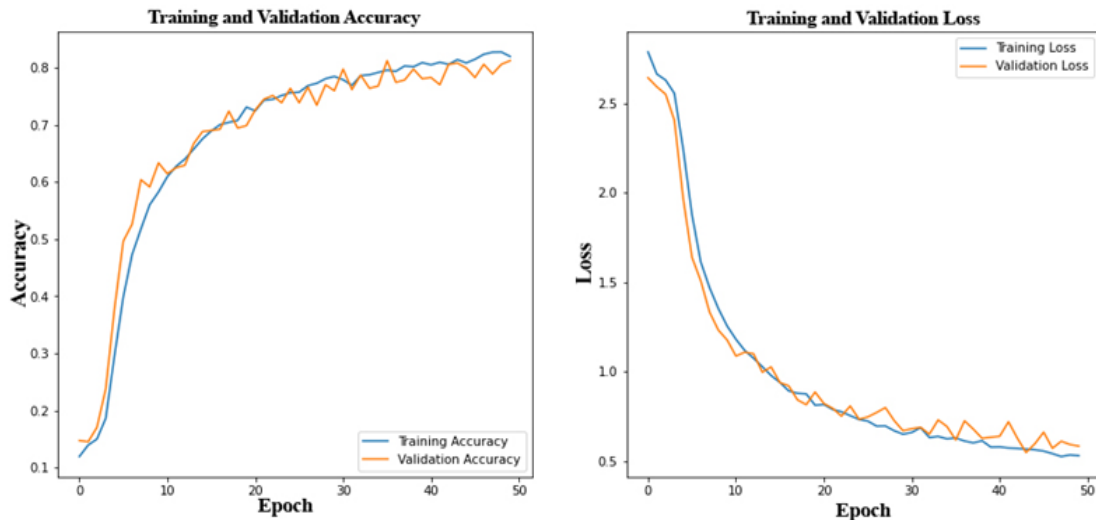


Figure 8. Training and validation accuracy and training and validation loss for adjectives dataset

```
[ ] X,y = test_generator.next()
score = model.evaluate(X, y)

4/4 [=====] - 0s 85ms/step - loss: 0.9325 - accuracy: 0.7500
```

Figure 9. Test accuracy for adjectives dataset

The large dataset in deep learning plays an important role in making the system more successful. The dataset of this study is very small and the data were mostly obtained by replacing other data. The confusion matrix of the adjectives dataset is presented in Figure 10. 100% success was achieved in most of the adjectives such as “Other”, “White”, “Correct”, “Economic”, “Easy”, “Small”, “Possible”, “Important”, “Black”, and “International” and the system predicted completely correctly. The system made an incorrect guess with 0% in "Slow" adjective, 20% in "continually" adjective, and 25% in "OK" adjective. In addition, in adjectives such as “Great”, “Beautiful”, “Necessary”, “Clean”, and “High”, the system can be said to have made almost correct predictions by showing 50% or more success.

The system showed more errors in words with similar lip movements (as seen in Figure 11). The lip movements of the adjectives "Beautiful" (in Figure 11(a)) and "Easy" (in Figure 11(b)) are similar. Therefore, the system showed the adjective "Easy" as a result in 33% of the " Beautiful " adjectives. Also, the lip movement of the "continually" is similar to the "Beautiful" and "High", and thus the system nearly failed with only 20% correct guesses. In order to eliminate these errors and to enable the system to learn better, the dataset and the number of epochs must be increased.

The model was run again with 50 epochs and adam optimization for the nouns dataset. Training and validation accuracy and training and validation loss for the nouns dataset are shown in Figure 12. Training accuracy was 78.67% and validation accuracy was 81.21%. The system showed lower performance than the adjective dataset in training and validation accuracy.

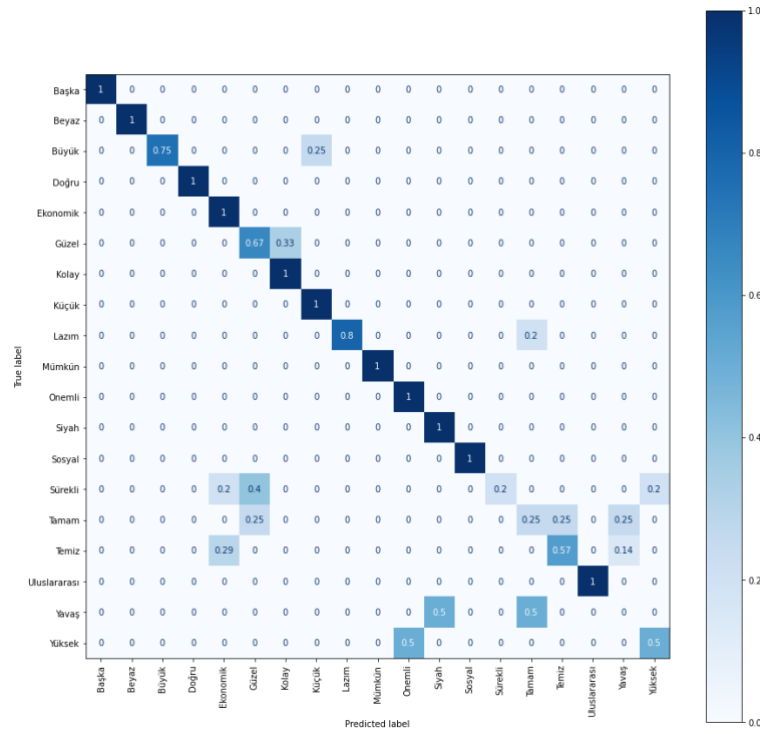


Figure 10. Confusion matrix for adjectives dataset

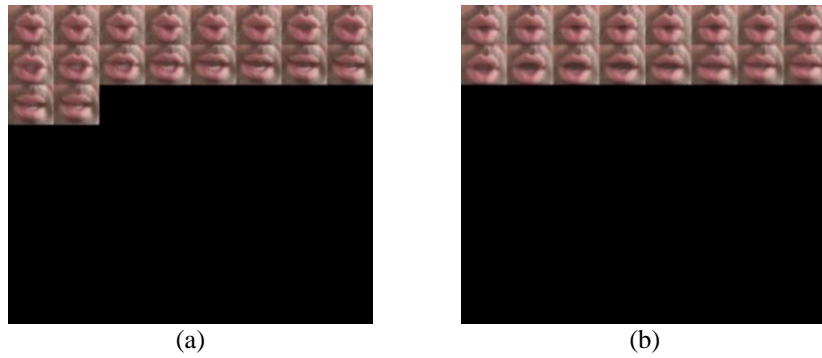


Figure 11. Lip movement of (a) “beautiful” and (b) “easy”

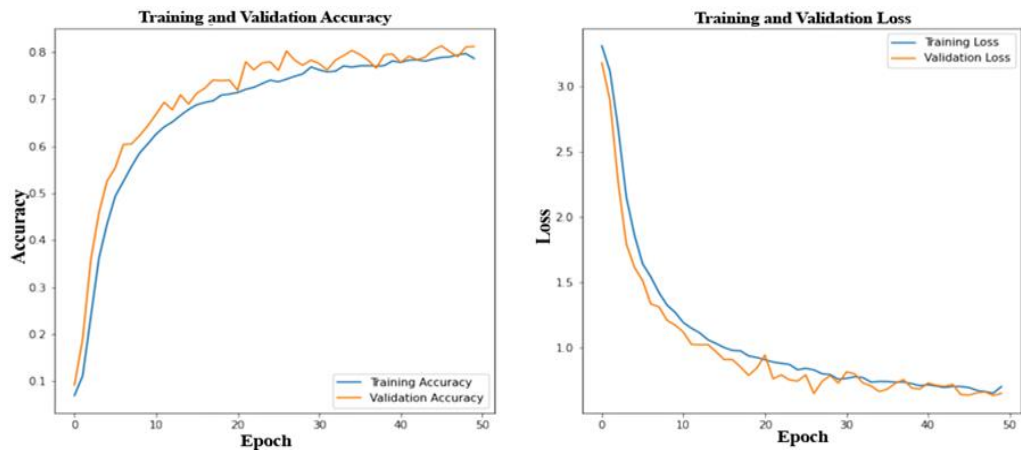


Figure 12. Training and validation accuracy and training and validation loss for nouns dataset

The success rate of the nouns dataset was 71.88% on the test set (Figure 13). As mentioned in the adjective dataset, the training and validation datasets are likely to be similar here, but the test dataset consists of completely different data and is given to the system only when testing. Faisal and Manzoor [52] achieved 62% success on word dataset in Urdu language.

```
X,y = test_generator.next()
score = model.evaluate(X, y)

2/2 [=====] - 0s 134ms/step - loss: 1.3726 - accuracy: 0.7188
```

Figure 13. Test accuracy for nouns dataset

The confusion matrix of the nouns dataset is presented in Figure 14. The system showed 100% success in 17 nouns. However, it failed completely with 0% on some words such as “Car”, “section”, “Technology”, and “Country” and partially failed with the nouns “Money” and “Part” under 50% guessing. In other words, 50% or more success was achieved.

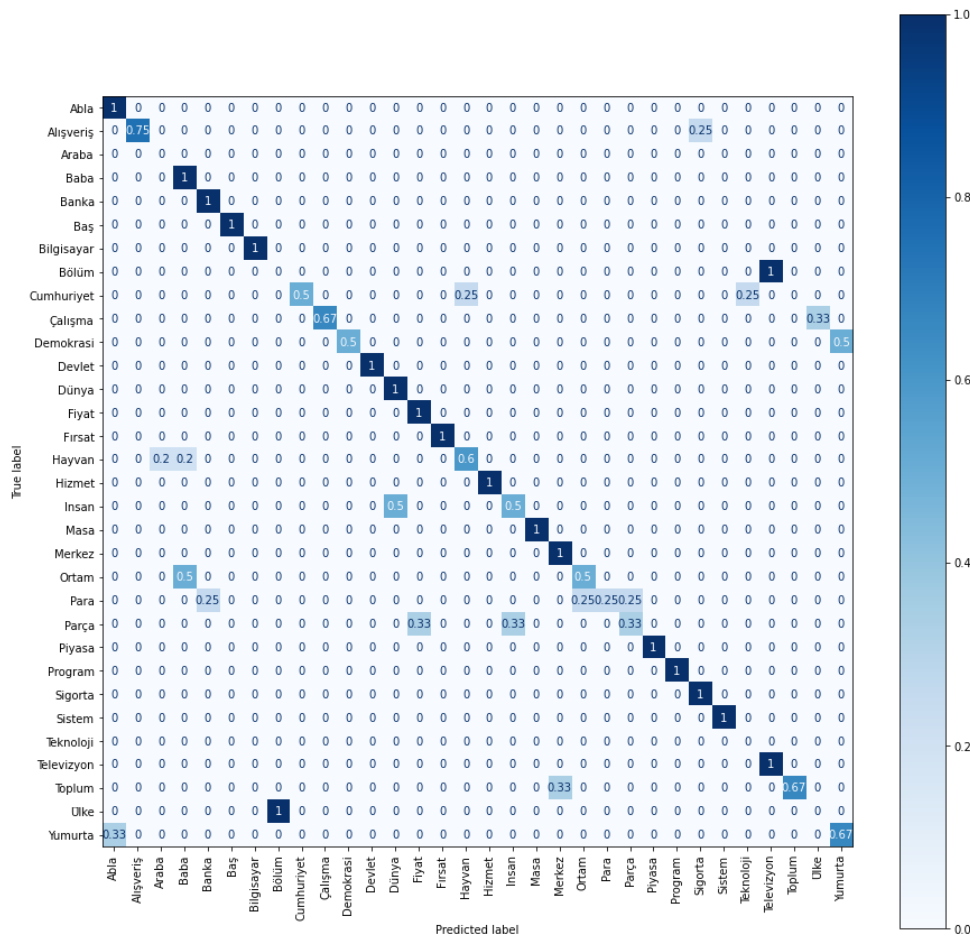


Figure 14. Confusion matrix for nouns dataset

The model was run again with 50 epochs and adam optimization for the verbs dataset. Training and validation accuracy and training and validation loss for the verbs dataset are shown in Figure 15. Training accuracy was 81.16% and validation accuracy was 78.74%. The system showed better performance than the adjective and nouns dataset in training and validation accuracy.

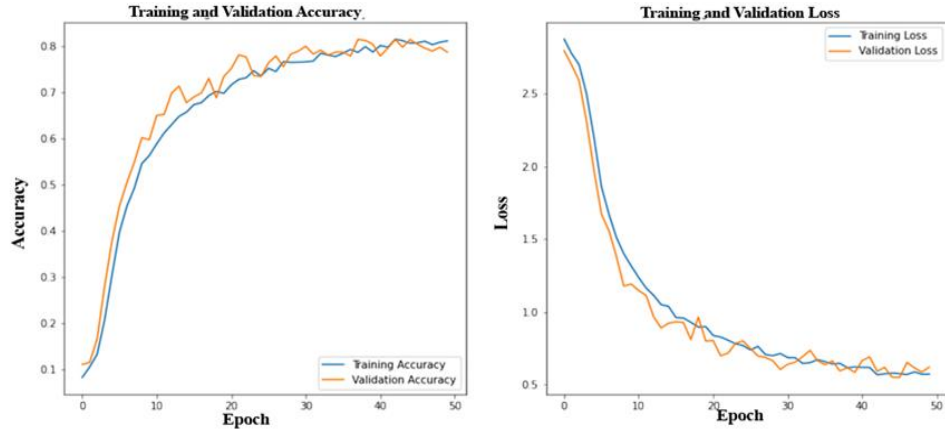


Figure 15. Training and validation accuracy and training and validation loss for verbs dataset

The success rate of the nouns dataset was 79.69% on the test set (Figure 16). As mentioned in the adjective dataset, the training and validation datasets are likely to be similar here, but the test dataset consists of completely different data and is given to the system only when testing. Ma *et al.* [32] achieved 88.5% success at word level with LRW-1000 dataset and Noda *et al.* [5] achieved 58% success in the word dataset in Japanese. The confusion matrix of the verbs dataset is presented in Figure 17. As seen in the complexity matrix, the model showed 100% success in 9 verbs and 50% or more success in all other verbs. The verbs dataset performed better than other datasets In the testing phase.

```
X,y = test_generator.next()
score = model.evaluate(X, y)

2/2 [=====] - 0s 145ms/step - loss: 0.7537 - accuracy: 0.7969
```

Figure 16. Test accuracy for verbs dataset

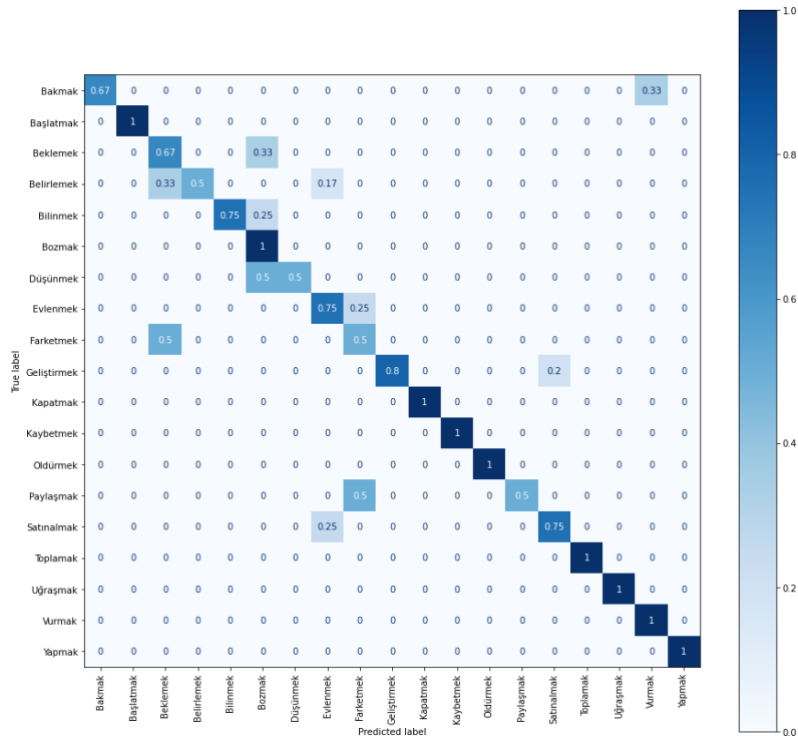


Figure 17. Confusion matrix for verbs dataset

There is lip movement in all verbs and lip movement of verbs are not similar to each other, and also the number of frames is high. So, the system learned better than other datasets. In addition to the developed convolutional neural network model, pre-trained models were also trained and tested on the datasets. In this study, pre-trained VGG and Inception-V3 models were trained on datasets and their success rates were examined. However, the VGG model showed only 10% success on the adjectives dataset and Inception-V3 model showed 64.06% success on the verbs dataset. Table 3 summarizes the results, the number of repetitions and success rates.

Table 3. Summary of the results

Model	Dataset	Epoch	Class	Optimizer	Success rate %
Proposed model	Adjectives	50	19	Adam	75
Proposed model	Nouns	50	33	Adam	71.88
Proposed model	Verbs	50	19	Adam	79.69
VGG	Adjectives	50	19	Adam	10
Inception-v3	Verbs	50	19	Adam	64.06

7. CONCLUSION

In this study, a visual lip-reading system was proposed for Turkish language. So, a convolutional neural network model was proposed and trained. Also, pre-trained models were used to increase success. The proposed model was trained and tested on the dataset of adjectives, nouns, and verbs and achieved 75%, 71.88%, and 79.69% success, respectively. In the adjectives dataset, the lip movements of adjectives such as “beautiful”, “easy”, and “continually” are similar to each other, and in the nouns dataset, the lip movements of words such as “money” and “piece” are similar to each other, and the lip movements of words such as “animal” and “car” are similar to each other. Therefore, they affected the success rate. In addition to these, one of the biggest problems of the study was that the data set was small and was not recorded in the required environment, and that most of the people did not look at the camera correctly. Most of the studies conducted at the word level in this area have achieved a success rate of less than 75%. Therefore, it can be said that the result is at a good level when the success achieved at the word level is compared with other studies. Most people did not send videos because they did not trust and thus the very small dataset was one of the major limitations of the study. In addition, the videos were not recorded in the required environment or the persons were not looking towards the camera. therefore, lip movements were not understandable by the system. For this reason, in future studies, it can be examined how much the success rate will be affected by keeping the camera fixed in one place and saying words by looking at the camera from different angles. In addition, Turkish sentences dataset can be collected and its success rate can be compared with numbers words. Also, other pre-trained models can be run on these datasets.

REFERENCES




- [1] Y. Li, Y. Takashima, T. Takiguchi, and Y. Arika, “Lip reading using a dynamic feature of lip images and convolutional neural networks,” in *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, Jun. 2016, pp. 1–6, doi: 10.1109/ICIS.2016.7550888.
- [2] S. Agrawal, V. R. Omprakash, and Ranvijay, “Lip reading techniques: A survey,” in *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, 2016, pp. 753–757, doi: 10.1109/ICATccT.2016.7912100.
- [3] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Lip reading sentences in the wild,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 3444–3453, doi: 10.1109/CVPR.2017.367.
- [4] X. Chen, J. Du, and H. Zhang, “Lipreading with DenseNet and resBi-LSTM,” *Signal, Image and Video Processing*, vol. 14, no. 5, pp. 981–989, Jul. 2020, doi: 10.1007/s11760-019-01630-1.
- [5] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, “Lipreading using convolutional neural network,” in *Interspeech 2014*, Sep. 2014, pp. 1149–1153, doi: 10.21437/Interspeech.2014-293.
- [6] Ü. Atila and F. Sabaz, “Turkish lip-reading using Bi-LSTM and deep learning models,” *Engineering Science and Technology, an International Journal*, vol. 35, p. 101206, Nov. 2022, doi: 10.1016/j.jestech.2022.101206.
- [7] A. Garg, J. Noyola, and S. Bagadia, “Lip reading using CNN and LSTM,” *Technical Report, Stanford University*, 2016. [Online]. Available: https://cs231n.stanford.edu/reports/2016/pdfs/217_Report.pdf.
- [8] B. Martinez, P. Ma, S. Petridis, and M. Pantic, “Lipreading using temporal convolutional networks,” in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6319–6323, doi: 10.1109/ICASSP40776.2020.9053841.
- [9] T. Ozcan and A. Basturk, “Lip reading using convolutional neural networks with and without pre-trained models,” *Balkan Journal of Electrical and Computer Engineering*, vol. 7, no. 2, pp. 195–201, Apr. 2019, doi: 10.17694/bajeece.479891.
- [10] A. Yargic and M. Dogan, “A lip reading application on MS Kinect camera,” in *2013 IEEE INISTA*, Jun. 2013, pp. 1–5, doi: 10.1109/INISTA.2013.6577656.
- [11] J. R. Movellan, “Visual speech recognition with stochastic networks,” in *Advances in Neural Information Processing Systems*, 1994, pp. 851–858.
- [12] O. Vanegas, K. Tokuda, and T. Kitamura, “Location normalization of HMM-based lip-reading: experiments for the M2VTS database,” in *Proceedings 1999 International Conference on Image Processing*, 1999, pp. 343–347 vol.2, doi:

- 10.1109/ICIP.1999.822914.
- [13] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 198–213, 2002, doi: 10.1109/34.982900.
 - [14] A. Vorwerk, X. Wang, D. Kolossa, S. Zeiler, and R. Orglmeister, "WAPUSK20 - a database for robust audiovisual speech recognition," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010, pp. 3016–3019.
 - [15] C. Neri *et al.*, "Audio visual speech recognition," *Workshop Final Report*, pp. 1-84, 2000.
 - [16] A. Ortega *et al.*, "AV@CAR: a Spanish multichannel multimodal corpus for in-vehicle automatic audio-visual speech recognition," in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, 2004.
 - [17] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," in *IEEE International Conference on Acoustics Speech and Signal Processing*, 2002, pp. 2017–2020, doi: 10.1109/ICASSP.2002.5745028.
 - [18] C. Petr, Ž. Miloš, K. Zdeněk, K. Jakub, Z. Jan, and M. Luděk, "Design and recording of Czech speech corpus for audio-visual continuous speech recognition," in *Auditory-Visual Speech Processing Workshop 2005*, 2005, pp. 1–4.
 - [19] S. Cox, R. Harvey, Y. Lan, J. Newman, and B.-J. Theobald, "The challenge of multispeaker lip-reading," in *Audio Visual Speech Processing AVSP, Brisbane, 2008*, 2008, pp. 179–184.
 - [20] S. Tamura *et al.*, "CENSREC-1-AV: an audio-visual corpus for noisy bimodal speech recognition," *International Conference on Audio-Visual Speech Processing*, pp. 1-4, 2010.
 - [21] A. G. Chitu, K. Driegl, and L. J. M. Rothkrantz, "Automatic lip reading in the Dutch language using active appearance models on high speed recordings," in *Text, Speech and Dialogue*, Springer Berlin Heidelberg, 2010, pp. 259–266.
 - [22] I. Anina, Ziheng Zhou, Guoying Zhao, and M. Pietikainen, "OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, May 2015, pp. 1–5, doi: 10.1109/FG.2015.7163155.
 - [23] V. Estellers and J.-P. Thiran, "Multi-pose lipreading and audio-visual speech recognition," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, p. 51, Dec. 2012, doi: 10.1186/1687-6180-2012-51.
 - [24] A. Rekić, A. B. -Hamadou, and W. Mahdi, "A new visual speech recognition approach for RGB-D cameras," in *Lecture Notes in Computer Science*, Springer International Publishing, 2014, pp. 21–28.
 - [25] V. Verkhodanova, A. Ronzhin, I. Kipyatkova, D. Ivanko, A. Karpov, and M. Železný, "HAVRUS corpus: high-speed recordings of audio-visual Russian speech," in *Speech and Computer*, Springer International Publishing, 2016, pp. 338–345.
 - [26] S. Petridis, J. Shen, D. Cetin, and M. Pantic, "Visual-only recognition of normal, whispered and silent speech," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 6219–6223, doi: 10.1109/ICASSP.2018.8461596.
 - [27] S. Yang *et al.*, "LRW-1000: a naturally-distributed large-scale benchmark for lip reading in the wild," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 2019, pp. 1–8, doi: 10.1109/FG.2019.8756582.
 - [28] L. A. Elrefaie, T. Q. Alhassan, and S. S. Omar, "An Arabic visual dataset for visual speech recognition," *Procedia Computer Science*, vol. 163, pp. 400–409, 2019, doi: 10.1016/j.procs.2019.12.122.
 - [29] H. Pourmousa and Ü. Özen, "Lip reading using CNN for Turkish numbers," *Journal of Business in The Digital Age*, vol. 5, no. 2, pp. 155-160, Sep. 2022, doi: 10.46238/jobda.1100903.
 - [30] A. S. Nambesani, C. Payyappilly, E. J. C. J. P., and M. S. Alex, "Lip reading using facial feature extraction and deep learning," *International Journal of Innovative Science and Research Technology*, vol. 6, no. 7, pp. 92–96, 2021.
 - [31] A. M. Sarhan, N. M. Elshennawy, and D. M. Ibrahim, "HLR-Net: a hybrid lip-reading model based on deep convolutional neural networks," *Computers, Materials & Continua*, vol. 68, no. 2, pp. 1531–1549, 2021, doi: 10.32604/cmc.2021.016509.
 - [32] P. Ma, B. Martinez, S. Petridis, and M. Pantic, "Towards practical lipreading with distilled and efficient models," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 7608–7612, doi: 10.1109/ICASSP39728.2021.9415063.
 - [33] K. Sarigül, "En çok kullanılan 1000 türkçe kelime," *Türkçe Öğretimi*. Accessed: Oct. 21, 2021. [Online]. Available: <https://www.turkceogretimi.com/tavsiyeler/en-cok-kullanilan-1000-turkce-kelime>
 - [34] K. Sarigül, "Türkçede en çok kullanılan 200 fiil," *Türkçe Öğretimi*. Accessed: Oct. 21, 2021. [Online]. Available: <https://www.turkceogretimi.com/tavsiyeler/turkcede-en-cok-kullanilan-200-fiil>
 - [35] Y. LeCun *et al.*, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, Dec. 1989, doi: 10.1162/neco.1989.1.4.541.
 - [36] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: analysis, applications, and prospects," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022, doi: 10.1109/TNNLS.2021.3084827.
 - [37] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into Imaging*, vol. 9, no. 4, pp. 611–629, Aug. 2018, doi: 10.1007/s13244-018-0639-9.
 - [38] A. Kamilaris and F. X. P. -Boldú, "A review of the use of convolutional neural networks in agriculture," *Journal of Agricultural Science*, vol. 156, no. 3, pp. 312–322, 2018, doi: 10.1017/S0021859618000436.
 - [39] K. Ryczko, K. Mills, I. Luchak, C. Homenick, and I. Tamblin, "Convolutional neural networks for atomistic systems," *Computational Materials Science*, vol. 149, pp. 134–142, Jun. 2018, doi: 10.1016/j.commatsci.2018.03.005.
 - [40] T. Guo, J. Dong, H. Li, and Y. Gao, "Simple convolutional neural network on image classification," in *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*, Mar. 2017, pp. 721–724, doi: 10.1109/ICBDA.2017.8078730.
 - [41] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. F.-Fei, "Large-scale video classification with convolutional neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732, doi: 10.1109/CVPR.2014.223.
 - [42] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1746–1751, doi: 10.3115/v1/D14-1181.
 - [43] S. Hershey *et al.*, "CNN architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 131–135, doi: 10.1109/ICASSP.2017.7952132.
 - [44] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1533–1545, Oct. 2014, doi: 10.1109/TASLP.2014.2339736.
 - [45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for Large-Scale image recognition," *Computer Vision and Pattern Recognition*, Sep. 2014.




- [46] C. Affonso, A. L. D. Rossi, F. H. A. Vieira, and A. C. P. de L. F. de Carvalho, "Deep learning for biological image classification," *Expert Systems with Applications*, vol. 85, pp. 114–122, Nov. 2017, doi: 10.1016/j.eswa.2017.05.039.
- [47] F. Güven, "Using text representation and deep learning methods for Turkish text classification," *Master Thesis*, Çukurova University, Adana, Turkey, 2019.
- [48] H. Gholamalinezhad and H. Khosravi, "Pooling methods in deep neural networks, a review," *Computer Vision and Pattern Recognition*, Sep. 2020.
- [49] D. T. Tran, A. Iosifidis, and M. Gabbouj, "Improving efficiency in convolutional neural networks with multilinear filters," *Neural Networks*, vol. 105, pp. 328–339, Sep. 2018, doi: 10.1016/j.neunet.2018.05.017.
- [50] H. Wu and J. Zhao, "Deep convolutional neural network model based chemical process fault diagnosis," *Computers & Chemical Engineering*, vol. 115, pp. 185–197, Jul. 2018, doi: 10.1016/j.compchemeng.2018.04.009.
- [51] A. M. Karim, "A new framework by using deep learning techniques for data processing," *Ph.D. Thesis*, Department of Computer Engineering, Ankara Yıldırım Beyazıt University, Ankara, Turkey, 2018.
- [52] M. Faisal and S. Manzoor, "Deep learning for lip reading using audio-visual information for Urdu language," *Computer Vision and Pattern Recognition*, Feb. 2018.

BIOGRAPHIES OF AUTHORS



Hadi Pourmousa    holds a Doctor of Management Information Systems degree from Atatürk University, Türkiye in 2022. He also received his B.Sc. (Information Technology Engineering) from Tabriz University, Iran in 2011 and M.Sc. (MIS) from Atatürk University, Türkiye in 2017 and respectively. He is currently a PhD student in Computer Engineering. His research includes image processing, computer vision, machine learning, and deep learning. He can be contacted at email: pourmousahadi@gmail.com or hadi.pourmousa14@ogr.atauni.edu.tr.



Dr. Üstün Özen    is full Professor and Senior Lecturer in Management Information Systems Department at Atatürk University, Erzurum, Türkiye. He also serves as Department chair. His research interests focus on social media analysis, data analytics and health informatics. He has several journal and conference papers, and books. He can be contacted at email: uozen@atauni.edu.tr.