**Machine Learning Research**
2021; 6(1): 1-7
http://www.sciencepublishinggroup.com/j/mlr
doi: 10.11648/j.mlr.20210601.11
ISSN: 2637-5672 (Print); ISSN: 2637-5680 (Online)

# Designing a Voice-controlled Wheelchair for Persian-speaking Users Using Deep Learning Networks with a Small Dataset

**Mohammad Amiri, Manizheh Ranjbar, Mostafa Azami Gharetappeh**

Department of Computer Engineering, Technical and Vocational University (TVU), Tehran, Iran

**Email address:**
m-amiri@tvu.ac.ir (M. Amiri), manij.ranjbar@gmail.com (M. Ranjbar)

**To cite this article:**
Mohammad Amiri, Manizheh Ranjbar, Mostafa Azami Gharetappeh. Designing a Voice-controlled Wheelchair for Persian-speaking Users Using Deep Learning Networks with a Small Dataset. *Machine Learning Research.* Vol. 6, No. 1, 2021, pp. 1-7.
doi: 10.11648/j.mlr.20210601.11

**Received**: July 31, 2021; **Accepted**: August 12, 2021; **Published**: November 5, 2021

**Abstract:** With the advancement of technology, the demand for improving the quality of life of the elderly and disabled has increased and their hope to overcome their problem is realized by using advanced technologies in the field of rehabilitation. Many existing electrical and electronic devices can be turned into more controllable and more functional devices using artificial intelligence. In every society, some spinal disabled people lack physical and motor abilities such as moving their limbs and they cannot use the normal wheelchair and need a wheelchair with voice control. The main challenge of this project is to identify the voice patterns of disabled people. Audio classification is one of the challenges in the field of pattern recognition. In this paper, a method of classifying ambient sounds based on the sound spectrogram, using deep neural networks is presented to classify Persian speakers sound for building a voice-controlled intelligent wheelchair. To do this, we used Inception-V3 as a convolutional neural network which is pretrained by the ImageNet dataset. In the next step, we trained the network with images that are generated using spectrogram images of the ambient sound of about 50 Persian speakers. The experimental results achieved a mean accuracy of 83.33%. In this plan, there is the ability to control the wheelchair by a third party (such as spouse, children or parents) by installing an application on their mobile phones. This wheelchair will be able to execute five commands such as stop, left, right, front and back.

**Keywords:** Voice Recognition, Deep Learning, Convolutional Neural Network, Spectrogram, Inception-V3

## 1. Introduction

In modern society, mobility is more of a major public health challenge, especially for the elderly and disabled living alone. Therefore, the demand for assistive systems such as intelligent wheelchairs, especially for patients who are unable to walk normally due to injury, has increased significantly in the healthcare industry [1, 2]. People with mobility problems are more depressed or anxious than normal people [3]. Therefore, restoring their mobility will increase their mental and physical health.

The power of modern computer systems has created new opportunities for researchers in the field of human-machine interaction to use advanced technology to improve the quality of life for disabled people.

Today, the use of wheelchairs as the main support for mobility for the elderly as well as the disabled has become quite common [1]. Previous models of wheelchairs, although providing a way to move people with disabilities, are usually only suitable for patients with lower extremity mobility, and are not able to reduce their dependence on their caregivers.

The modern smart wheelchair is a step towards an independent lifestyle for people with physical disabilities, which has a population of about 650 million [4]. One of the characteristics of smart wheelchairs is their capability to use by people with cerebral palsy and higher disability problem. These people cannot use a normal wheelchair with manual control like joysticks [1, 3]. They can only move their eyes eighter speak [5]. in this article we used and intelligent algorithms to detect voice command patterns for Persian speakers to detect five commands including stop, left, right, back and forward to use as the heart of the smart wheelchair.

Nowadays, due to the widespread access to various and inexpensive sensors [6, 7], research in the field of identification and classification of audio data has accelerated.

Systems based on sound sensors in the fields of medicine, surveillance, security, etc., such as multimedia [8], Bioacoustics monitoring [9], Identification of intruders in wildlife areas [10], Audio monitoring [11], Monitoring and identification of animal species [12], Automatic diagnosis of heart disease [13], acoustic analysis of crying signal in infants [14], automatic diagnosis of lung disease [15], security monitoring in unstructured environments [6] have been used.

In general, in the identification and classification of the system, due to the unsuitability of using raw audio data as a network input, it is necessary to first extract various features of raw audio data. Therefore, the sound recognition and classification process consist of three different steps including signal preprocessing, extraction of unique features, and finally the use of some classification tools to differentiate between classes. Various features can be extracted from audio data, the most common of which are spectroscopy, Mel spectroscopy, Mel frequency coefficient (MFCC), Stabilized hearing image (SAI) and Linear prediction coefficients (LPC) [16, 17]. In addition, various supervised machine learning algorithms including decision tree, random forest and nearest neighbor, support vector machine (SVM), hidden Markov models (HMM), Multilayer perceptron and deep learning networks have been used to develop voice recognition and classification systems [18, 19]. In recent years, due to the good results of using deep learning methods, especially convolutional neural networks (CNN), this method has been widely used in the field of sound identification and classification [20-22].

The purpose of this article is to design and build an intelligent wheelchair for Persian users using deep learning to classify audio spectrogram images. The simulation results show the capability of the proposed method.

The rest of this paper is as follows. In section 2, the overall architect of the device will be explained. The basic concepts and the proposed method for classifying ambient sounds are presented in Section 3. The data set used to evaluate the proposed method and test results are presented in Section 4. Finally, Section 5 sets out the conclusions and some future directions.

## 2. The General Architecture of the Proposed Voice-controlled Wheelchair

### *2.1. The General Overview*

In Figure 1, you can see an overview of the proposed design. The main components of the design are a wheelchair, two gearboxes, engine, driver, controller board, Wi-Fi module, microphone and a joystick. In the following, we will describe the performance of each section.

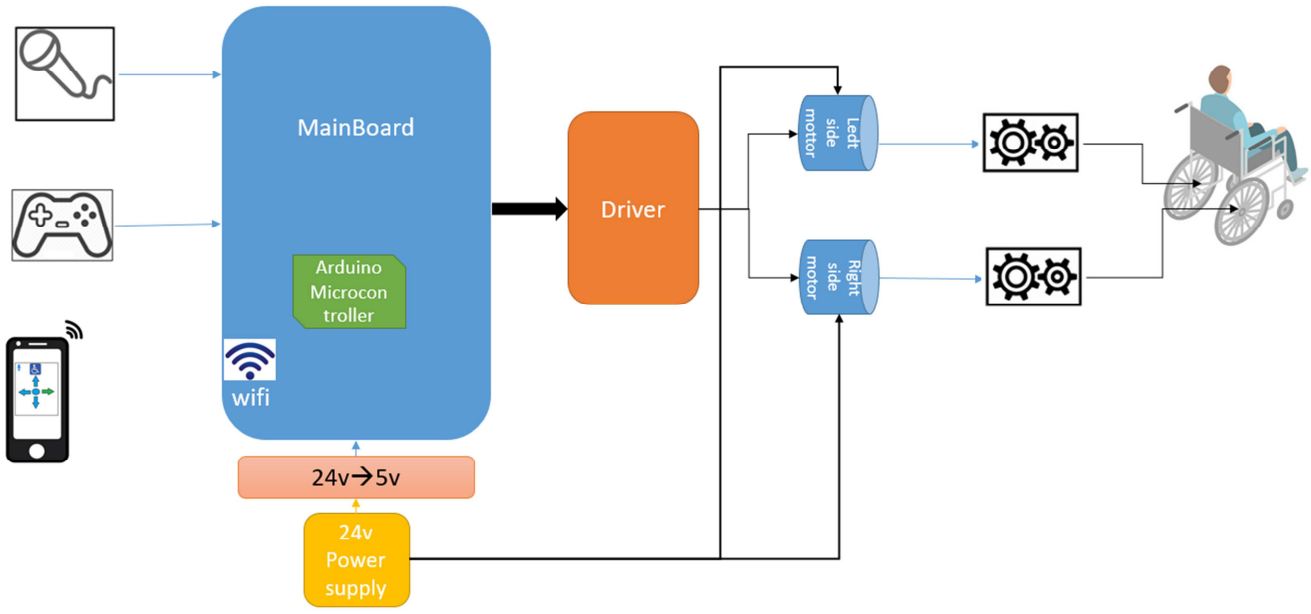


***Figure 1.*** *Overview of the proposed wheelchair design.*

### *2.2. Gearboxes*

To reduce the transmission speed from the engine to the wheels, two gearboxes with a ratio of 63 to 7 have been used. Since this wheelchair must be able to move an ordinary human weighing approximately 100 kg plus the weight of the wheelchair (about 10 kg), it requires a relatively high engine power. To reduce engine speed and thus reduce the speed of the wheelchair from the gearbox we used these two gearboxes.

According to the specifications of this motor, which has a slip ring, its output torque can be calculated as:

$$P = T \times \omega \; or \; P = \frac{T \times 2\pi \times N}{60}$$

$$300 = \frac{T \times 2\pi \times 2500}{60} \longrightarrow T = 1.14 \ _{NM} \quad (1)$$

Therefore, we can control the output torque by changing the ratio of the gearboxes. The gearbox used in this project has a diagonal gear with a ratio of 63 to 7.

The reason for using diagonal gears is that in these gears, the ratio of engagement is more than one gear. The most popular ratio in industrial application is 1/6.

# 3. The Proposed Algorithm for Audio Classification for Persian Users

## *3.1. Inception-V3 Deep Network*

Inception-V3 [23] is a convolutional neural network architecture of the family of Inceptions. This architecture is an improved version of the GoogleNet suggested by Google in 2014. Because the core of the GoogleNet is the inception module, the GoogleNet is also named as inception network [24].

The inception module has some parallel layers, which usually includes a set of convolutional layers with three different sizes 1×1, 3×3 and 5×5 and also one max-pooling layer. In this way, information can be extracted at different levels. Therefore, spatial features can be extracted more efficiently. The network structure of the Inception is so that by increasing the depth of the network, it also reduces the number of network parameters, which reduces the computational complexity and increases the accuracy and efficiency of the network. In addition, it prevents overfitting. Hence, it is widely used in image classification tasks. In Figure 2 it is illustrated the structure of a typical inception module.

In the proposed inception-V3, to improve the performance of the Inception, the larger filters with the size of 5×5 and 7×7 that are computationally expensive has replaced with smaller size. For example, the 3×3 convolution is replaced by two 3×1 and 1×3. This replacement reduces the number of parameters and increases the speed of network training. In Figure 3 the structure of Inception-V3 is displayed.

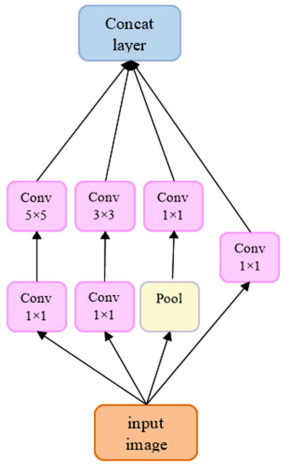


***Figure 2.*** *The overall structure of the Inception module [25]. One module includes some small convolution layers with a pooling layer.*

In Figure 3, the whole architecture of inception-V3 is described. The default input image size is 299×299 with three channels. The first layers include convolutional and pooling layers and the rest are inception modules and at the end, there are fully connected and SoftMax layers.

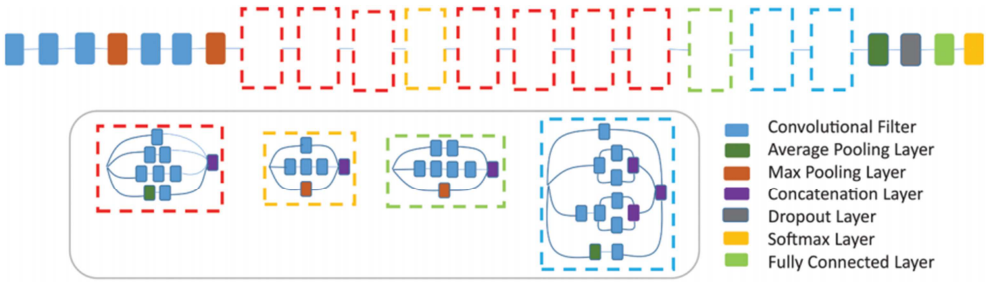


***Figure 3.*** *The Inception-V3 network [26]. It includes three main parts. Convolutional filter, Inception module and fully connected layers.*

## *3.2. Spectrograms*

A spectrogram or spectrometer is a method of visualizing the frequency spectrum of a sound wave. In simpler terms, display spectroscopy of audio data is based on a kind of real-time spectrometer [27-29]. Figure 4 shows an example of a spectrogram of an audio file. Spectrogram images can be used in conjunction with various machine learning classifiers, including deep learning methods. Studies by Liu et al. on deep learning models show that convolutional neural networks (CNN) have performed better than other models in images and video data [18]. Because a visual spectrum image is a signal of the frequency spectrum of a signal, spectrometer images are used in deep learning methods to extract and classify features. Audio signals are less frequent and create a different pattern in the display spectrum (see Figure 4).

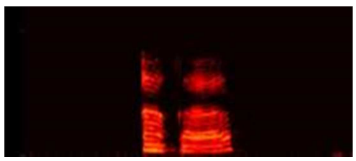

***Figure 4.*** *A spectrogram of an audio file.*

This method involves the steps of filtering and converting the signal into an image. Input audio signals based on the STFT[1] method are converted into audio images. The STFT

1 Short time Fourier transform

Indicates the frequency content of the input signal. It provides useful information about waveforms like what frequencies and with what power are there in the waveform.

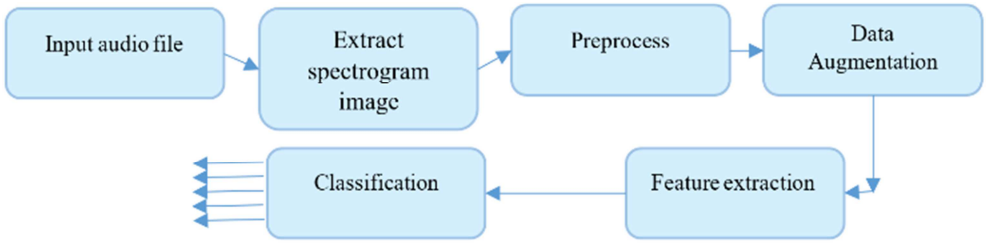


***Figure 5.*** *The pipeline of the proposed technique. The technique has five main modules: Extract spectrogram, preprocessing, data augmentation, feature extraction and image classification.*

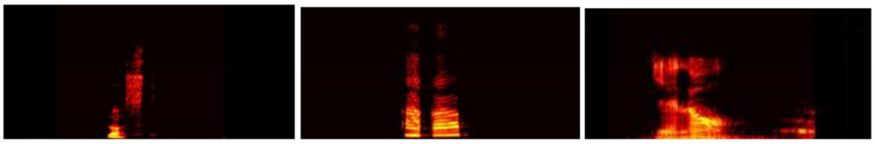

***Figure 6.*** *Example of some spectrogram images with Hamming=512, Noverlap=256 and Nfft=1024.*

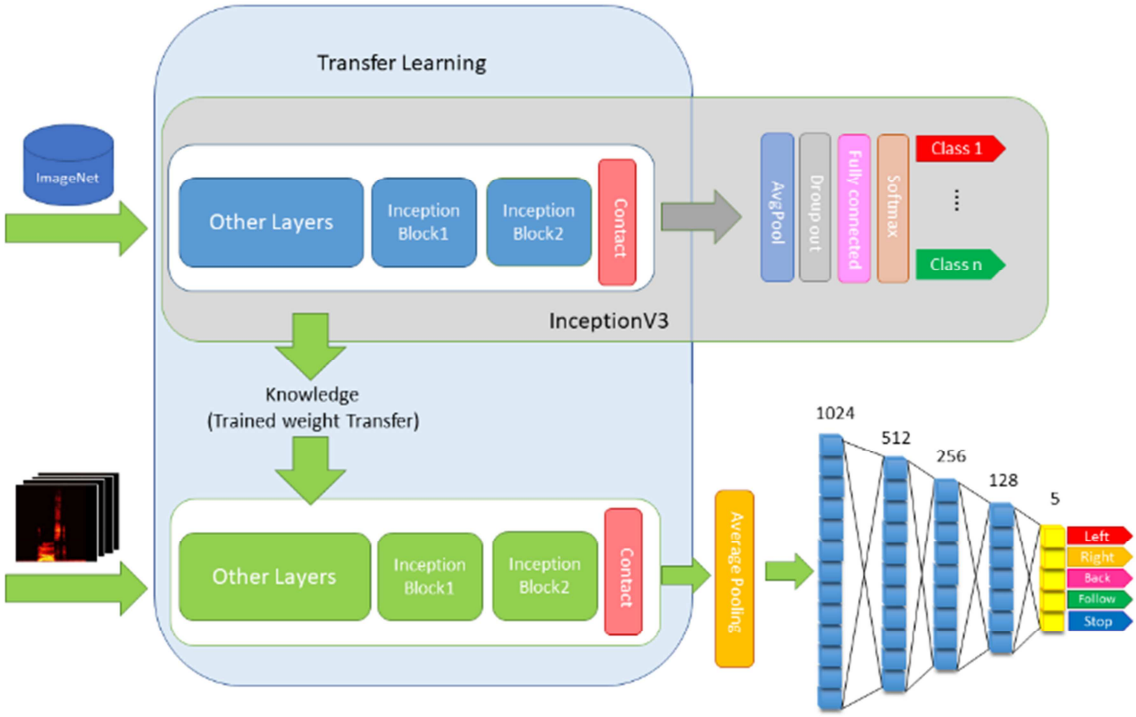


***Figure 7.*** *Usage of transfer learning to apply inception-V3 for sound classification. The last two modules and five fully connected layers trained using spectrogram images.*

### *3.3. Architectural Detail*

Our aim in this paper is to build a deep learning model for voice classification and recognition based on a small data set that was achieved using the proposed method. In general, a classification model consists of three main steps including preprocessing, feature extraction and classification. The inputs of the network are spectrogram images of audio data. Because convolutional neural networks require large data sets for training, we used both the data augmentation and transfer learning methods. For transfer learning, we used a pretrained inception-V3 network which has trained with the ImageNet dataset. This allows the network to learn rich feature representations for a wide range of images. The general structure of the proposed method is given in Figure 5.

As shown in Figure 6, the input data are audio files containing the sound of the words left, right, forward, backward and stop. In the first step, all audio files converted into spectrogram images using the MATLAB spectrogram function. In this formula, we set the Hamming parameter as 512, Noverlap=256 And Nfft=1024. Figure 6 shows an example of a spectrogram image.

For convolutional network models, the input images are often resized to maintain compatibility with the network architecture.

To apply incepception-V3 for this application, we added an average pooling and five full connection layers to the end of inception-V3. Layers one to four have 1024, 512, 256 and 128 nodes respectively. In these layers, we used ReLu as an activation function. The last layer has five nodes (related to the number of classes) and we used a SoftMax function as an activation function in this layer. In addition, we did fine-tune on the last two inception modules of the network. To have a good comparison, we also used the inception-V3 as feature extraction and SVM, Kernel SVM with different functions (e.g. poly, RBF[2] and sigmoid) as a classifier (see Figure 7).

# 4. Experimental Results

To evaluate the proposed method, all tests on a laptop equipped with a 2.30 GHz processor intel Core i5 and 4 GB RAM done. The programming language was MATLAB and Python.

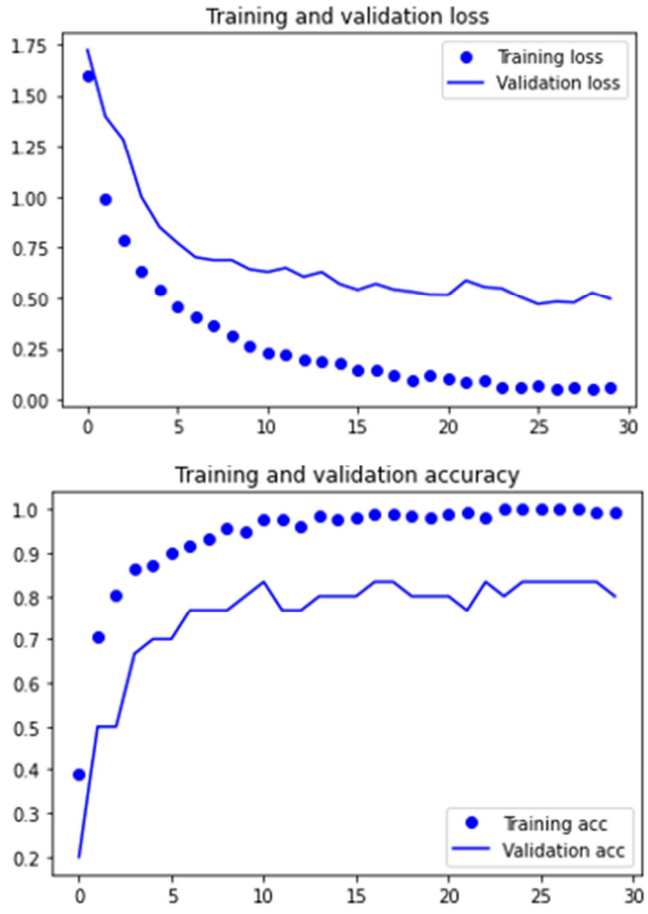


***Figure 8.*** *Convergence of the algorithm.*

### *4.1. Database*

As mentioned, the purpose of this article is to identify and classify sound to build a voice-controlled wheelchair in a natural environment. To do this, we created an audio dataset consisting of words to guide and move the smart wheelchair, including front, back, left, right and stop. This dataset includes 210 original samples recorded by 15 males and 35 females. A brief description of this dataset is described in Tables 1 and 2.

We used the data augmentation methods to generate more samples.

### *4.2. Changing the Spectrogram Function Variables*

The purpose of this experiment shows that the impact of Hamming, Noverlap and Nfft to extract spectrogram images in audio recognition and classification. In this test, we didn't do any fine-tuning and just trained the four last full connection layers. In Table 3 the result of these experiments is described. As it is explained, With Hamming=512, Noverlap=256 and Nfft=1024 the best result (80%) achieved by using the inception-V3 as feature extraction and SVM as a classifier. Using kSvm with the RBF function has the same result. If we change these parameters to 1024,512 and 2048 respectively the best result achieved in a combination of inception-V3 as feature extraction and SVM as a classifier.

As shown in Figure 8, The convergence of the algorithm is explained with iteration=30 and batch size of 10.

### *4.3. Activation Function*

In this experiment, we tried to understand which activation function has more efficiency. To do this, we did the comparison with two popular functions, ReLu and sigmoid, for all methods. The result is explained in Table 4. In this experiment, 10% of the data set is considered as test data. As it is described, using the ReLu activation function is better than Sigmoid in most cases. The best result is obtained by inception network as feature extraction and SVM as a classifier with the ReLu activation function.

### *4.4. Full Connection Layers with Different Iterations*

In this experiment, the performance of the methods Inception, Incp-Svm, Incp-KSvm-poly, Incp-kSvm-RBF and Incp-kSvm-sigmoid with different connection layers and iterations is evaluated. The results of these experiments are given in Table 5. In this experiment, the number of neurons for the two last fully connected layers is 1024 and 512, while the number of neurons when using four fully connected layers is adjusted to 1024, 512, 256 and 128 neurons respectively. According to the results in Table 5, it can be seen that Incp-kSvm-RBF and Incp-Svm with four full connection layers and using batch training in just one step, works better on our audio data set.

### *4.5. Precision, Recall and F1-score*

To have a better view of the efficiency of our algorithm, in addition to the accuracy rate, we also measured the precision, recall and F1_scores of the proposed algorithm. The results are shown in Table 6. These results prove the efficiency of our proposed method.

2 Radial basis function

***Table 1.*** *Details of people whose audio file samples have been recorded.*

| Age | Number of samples | Gender |
|---|---|---|
| 20-60 | 15 | Male |
| 25-60 | 35 | Female |

***Table 2.*** *Number of samples related to each voice command.*

| The word | Front | Back | Right | Left | Stop | Total |
|---|---|---|---|---|---|---|
| Before Augmentation | 41 | 42 | 45 | 42 | 40 | 210 |
| After Augmentation | 205 | 210 | 225 | 210 | 200 | 1050 |

***Table 3.*** *The accuracy rate with different values of the spectrogram function.*

| | inception | Incp-Svm | Incp-kSvm-poly | Incp-kSvm-rbf | Incp-kSvm-sigmoid |
|---|---|---|---|---|---|
| Hamming=1024, Noverlap=512, Nfft=2048 | 73.33 | 0.77 | 0.47 | 0.75 | 0.72 |
| Hamming=512, Noverlap=256, Nfft=1024 | 78.10 | 0.80 | 0.46 | 0.80 | 0.77 |

***Table 4.*** *The accuracy rate with different activation function.*

| | Inception | Incp-Svm | Incp-kSvm-poly | Incp-kSvm-rbf | Incp-kSvm-sigmoid |
|---|---|---|---|---|---|
| ReLu | 78.10 | 80 | 46 | 80 | 77 |
| sigmoid | 70.48 | 68 | 70 | 68 | 70 |

***Table 5.*** *Evaluation of network with different FC layers and iterations. FC=2,50 BN=1,10 means two fully connected layers with 50 neurons in each layer with batch number=10 and one iteration.*

| Iter1/iter2 | inception | Incp-Svm | Incp-kSvm-poly | Incp-kSvm-rbf | Incp-kSvm-sigmoid |
|---|---|---|---|---|---|
| FC=2,50 BN=1,10 | 74.29 | 76 | 49 | 73 | 68 |
| FC=2,15 BN=1,10 | 74.29 | 71 | 65 | 72 | 72 |
| FC=4,15 BN=1,10 | 79.05 | 77 | 59 | 78 | 73 |
| FC=4,15 BN=2,10,10 | 70.48 | 72 | 52 | 70 | 72 |
| FC=4, 25 BN=1,10 | 78.10 | 80 | 46 | 80 | 77 |

***Table 6.*** *Precision, recall and F1score of the proposed method.*

| class | Incp-kSvm-rbf | | | Incp-Svm | | |
|---|---|---|---|---|---|---|
| | precision | recall | f1-score | Precision | recall | f1-score |
| Back (عقب) | 100 | 71 | 83 | 100 | 71 | 83 |
| Forward (جلو) | 84 | 76 | 80 | 85 | 81 | 83 |
| Left (چپ) | 74 | 81 | 77 | 81 | 81 | 81 |
| Right (راست) | 79 | 71 | 75 | 78 | 67 | 72 |
| Stop (ایست) | 72 | 100 | 84 | 68 | 100 | 81 |
| Average | 82 | 80 | 80 | 82 | 80 | 80 |

### *4.6. Confusion Matrix*

Accuracy results obtained in the confusion matrix (Confusion matrix) for two methods Incp-Svm and Incp-kSvm-rbf has been shown in Tables 7 and 8. In both cases, it is clear that all classes are well known. According to these experiments, the best result is for command *stop* with the value of 100% in both methods and the worst accuracy was for the *Right* command with 66.66% in incp-SVM method and 71.42% with Incp-kSvm-RBF Method.

***Table 7.*** *Confusion matrix of incp_SVM algorithm.*

| Back | 71.42 | 9.52 | 0 | 14.28 | 4.76 |
|---|---|---|---|---|---|
| Forward | 0 | 80.95 | 9.52 | 0 | 9.52 |
| Left | 0 | 4.76 | 80.95 | 4.76 | 9.52 |
| Right | 0 | 4.76 | 4.76 | 66.66 | 23.8 |
| Stop | 0 | 0 | 0 | 0 | 100 |

***Table 8.*** *Confusion matrix of Incp-kSvm-rbf Algorithm.*

| Back | 71.42 | 9.52 | 0 | 14.28 | 4.76 |
|---|---|---|---|---|---|
| Forward | 0 | 76.19 | 14.28 | 0 | 9.52 |
| Left | 0 | 4.76 | 80.95 | 4.76 | 9.52 |
| Right | 0 | 0 | 14.2 | 71.42 | 14.2 |
| Stop | 0 | 0 | 0 | 0 | 100 |

# 5. Conclusion

In this paper, we introduced a voice-controlled wheelchair for Persian speakers using artificial intelligence deep network. The efficiency of deep learning for image classification has been approved. Therefore, we converted the audio files to images and applied one of the states of the art deep network, Inception-V3, to do the classification. Due to the lack of a database for Persian speakers, we created our database by recording the voice of 15 males and 35 females. Two

algorithms, *incp_SVM* and *Incp-kSVM-RBF* has better performance than others. The experimental results illustrate the efficiency of our algorithm.

# References


[1] Ghorbel, A., N. B. Amor, and M. Jallouli, A survey on different human-machine interactions used for controlling an electric wheelchair. Procedia Computer Science, 2019. 159: p. 398-407.

[2] Mazo, M., et al., Electronic control of a wheelchair guided by voice commands. Control Engineering Practice, 1995. 3 (5): p. 665-674.

[3] Tomari, M. R. M., Y. Kobayashi, and Y. Kuno, Development of Smart Wheelchair System for a User with Severe Motor Impairment. Procedia Engineering, 2012. 41: p. 538-546.

[4] Kumar, D., R. Malhotra, and S. R. Sharma, Design and Construction of a Smart Wheelchair. Procedia Computer Science, 2020. 172: p. 302-307.

[5] Ruíz-Serrano, A., et al., Development of a Dual Control System Applied to a Smart Wheelchair, using Magnetic and Speech Control. Procedia Technology, 2013. 7: p. 158-165.

[6] Scardapane, S., et al., Microphone array based classification for security monitoring in unstructured environments. AEU - International Journal of Electronics and Communications, 2015. 69 (11): p. 1715-1723.

[7] Maccagno, A., et al., A CNN Approach for Audio Classification in Construction Sites, in Progresses in Artificial Intelligence and Neural Systems, A. Esposito, et al., Editors. 2021, Springer Singapore: Singapore. p. 371-381.

[8] Wold, E., et al., Content-based classification, search, and retrieval of audio. IEEE MultiMedia, 1996. 3 (3): p. 27-36.

[9] Weninger, F. and B. Schuller. Audio recognition in the wild: Static and dynamic classification on a real-world database of animal vocalizations. in 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2011.

[10] Ghiurcau, M. V., et al., Audio based solutions for detecting intruders in wild areas. Signal Process., 2012. 92 (3): p. 829–840.

[11] Rabaoui, A., et al., Using One-Class SVMs and Wavelets for Audio Surveillance. Trans. Info. For. Sec., 2008. 3 (4): p. 763–775.

[12] Xu, W., et al., A multi-view CNN-based acoustic classification system for automatic animal species identification. Ad Hoc Networks, 2020. 102: p. 102115.

[13] Deperlioglu, O., Heart sound classification with signal instant energy and stacked autoencoder network. Biomedical Signal Processing and Control, 2021. 64: p. 102211.

[14] Mahmoudian, S., et al., Acoustic Analysis of Crying Signal in Infants with Disabling Hearing Impairment. Journal of Voice, 2019. 33 (6): p. 946. e7-946. e13.

[15] Messner, E., et al., Multi-channel lung sound classification with convolutional recurrent neural networks. Computers in Biology and Medicine, 2020. 122: p. 103831.

[16] Alías, F., J. C. Socoró, and X. Sevillano, A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. Applied Sciences, 2016. 6 (5): p. 143.

[17] Lyon, R. F., et al., Sound retrieval and ranking using sparse auditory representations. Neural computation, 2010. 22 (9): p. 2390-2416.

[18] Thiruvengatanadhan, R., Speech Recognition using SVM. International Research Journal of Engineering and Technology (IRJET), 2018. 5 (9): p. 918-921.

[19] Alifani, F., T. W. Purboyo, and C. Setianingsih. Implementation of Voice Recognition in Disaster Victim Detection Using Hidden Markov Model (HMM) Method. in 2019 International Seminar on Intelligent Technology and Its Applications (ISITIA). 2019. IEEE.

[20] Totakura, V., B. R. Vuribindi, and E. M. Reddy. Improved Safety of Self-Driving Car using Voice Recognition through CNN. In IOP Conference Series: Materials Science and Engineering. 2021. IOP Publishing.

[21] Chandankhede, P. H., A. S. Titarmare, and S. Chauhvan. Voice Recognition Based Security System Using Convolutional Neural Network. In 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS). 2021. IEEE.

[22] Sharan, R. V., S. Berkovsky, and S. Liu. Voice command recognition using biologically inspired time-frequency representation and convolutional neural networks. in 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2020. IEEE.

[23] Szegedy, C., et al., Rethinking the Inception Architecture for Computer Vision. 2015.

[24] Dong, N., et al., Inception v3 based cervical cell classification combined with artificially extracted features. Applied Soft Computing, 2020. 93: p. 106311.

[25] Szegedy, C., et al. Going deeper with convolutions. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[26] Ding, Y., et al., A deep learning model to predict a diagnosis of Alzheimer disease by using 18F-FDG PET of the brain. Radiology, 2019. 290 (2): p. 456-464.

[27] Khamparia, A., et al., Sound Classification Using Convolutional Neural Network and Tensor Deep Stacking Network. IEEE Access, 2019. 7: p. 7717-7727.

[28] Altes, R., Detection, estimation, and classification with spectrograms. Journal of the Acoustical Society of America, 1980. 67: p. 1232-1246.

[29] Hussein, W., M. Hussein, and T. Becker, Spectrogram Enhancement by Edge Detection Approach Applied To Bioacoustics Calls Classification. International Journal of signal and image processing, 2012. 3: p. 1-20.