Karafan
Quarterly Scientific Journal of National University of Skill
Spring 2025, Vol. 22, No. 1, p. 183-206
Journal Homepage: https://karafan.nus.ac.ir/?lang=en
:10.48301/kssa.2025.490050.3053

Deep learning-based audio classification algorithm in a voice-controlled wheelchair for Persian-speaking users

Mohammad Amiri
[1] Assistant Professor, Department of Computer Engineering, National University of Skills (NUS), Tehran, Iran.

ARTICLE INFO

Article Type:
Original Research

Received: 11.30.2024
Revised: 02.27.2025
Accepted: 04.26.2025

Keyword:
Voice Recognition
Audio Classification
Deep Learning
Convolutional Neural Networks
Spectrogram
Voice-Controlled Devices
Inception-V3

***Corresponding Author:**
Mohammad Amiri
Email: *amiri.mgh@gmail.com*

ABSTRACT

In every society, some spinal disabled people lack physical and motor abilities such as moving their limbs; they cannot use a normal wheelchair and need a wheelchair with voice control. Audio classification is one of the challenges in the field of pattern recognition. Traditional methods for classifying voice commands primarily include simple algorithms and manual annotation techniques, which often have limited efficiency due to their inability to recognize complex patterns and the high variability of human speech. Convolutional neural networks (CNNs) have been widely used in audio recognition and classification since they often provide positive results. In this paper, a method for classifying ambient sounds based on the sound spectrogram, using deep neural networks, is presented to classify Persian speakers' sounds for building a voice-controlled intelligent wheelchair. To implement this, Inception-V3 was used as a convolutional neural network, pretrained by the InceptionV3 dataset. In the next step, the network was trained with images that were generated using spectrogram images of the ambient sound of about 50 Persian speakers. Due to the lack of a Persian speakers' dataset, the present research dataset was created with 50 participants including 35 males and 15 females, in the age range of 25 to 60 years old. The experimental results achieved a mean accuracy of 83.33%. Therefore, the wheelchair was able to execute five commands, such as stop, left, right, front, and back.

E-ISSN: 2538-4430
ISSN: 2382-9796

EXTENDED ABSTRACT

Introduction

Nowadays, due to the widespread access to various and inexpensive sensors, research in the field of identification and classification of audio data has accelerated. Systems based on sound sensors in the fields of medicine, surveillance, and security such as multimedia, bioacoustics monitoring, identification of intruders in wildlife areas, audio monitoring, monitoring and identification of animal species, automatic diagnosis of heart disease, acoustic analysis of crying signal in infants, automatic diagnosis of lung disease, security monitoring in unstructured environments 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 consists 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). 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. In recent years, due to the good results of using deep learning methods, particularly convolutional neural networks (CNN), this method has been widely used in the field of sound identification and classification.

The purpose of the present research was to design and build an intelligent wheelchair for Persian users using deep learning to classify audio spectrogram images. The simulation results showed the capability of the proposed method.

Methodology

Inception-V3 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. The inception module has some parallel layers, which usually include 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 such 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 1, the structure of a typical inception module is illustrated.

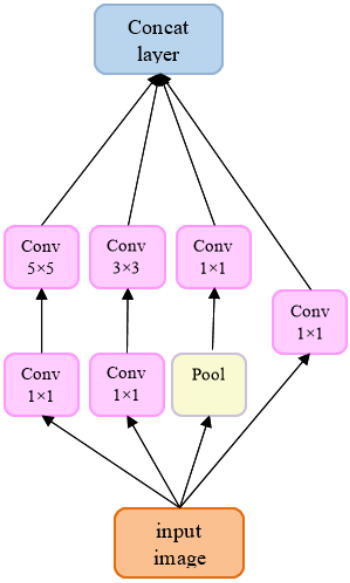


Figure 1. The overall structure of the Inception module [25]. One module includes some small convolution layers with a pooling layer.

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 were replaced with smaller sizes. For example, the 3×3 convolution was replaced by two 3×1 and 1×3 sizes. This replacement reduced the number of parameters and increased the speed of network training. In Figure 3, the structure of Inception-V3 is displayed.

In Figure 2, 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 module. At the end, there are fully connected and SoftMax layers.

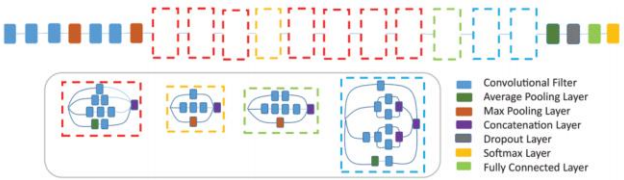


Figure 2. The Inception-V3 network [26]. It includes three main parts: convolutional filter, an Inception module and fully connected layers.

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) 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 2).

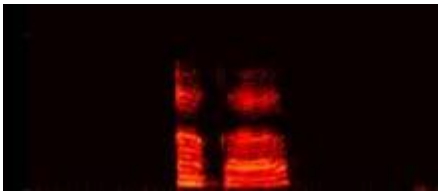


Figure 4. A spectrogram of an audio file.

This method involves the steps of filtering and converting the signal into an image. Input audio signals based on the STFT method are converted into audio images. The STFT indicates the frequency content of the input signal. It provides useful information about waveforms such as what frequencies and with what power are there in the waveform.

Results and discussion

The accuracy of the results obtained in the confusion matrix (Confusion matrix) for the two methods Incp-Svm and Incp-kSvm-rbf are 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 was obtained for command *stop* with the value of 100% in both methods and the worst accuracy was obtained for the *Right* command with 66.66% in incp-SVM method and 71.42% with Incp-kSvm-RBF method.

Table 1. 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 2. 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

Conclusion

In the present research, a voice-controlled wheelchair for Persian speakers was introduced using an artificial intelligence deep network. The efficiency of deep learning for image classification was proven. Therefore, the audio files were converted to images and applied one of the state-of-the-art deep networks, Inception-V3, to do the classification. Due to the lack of a database for Persian speakers, our database was created by recording the voices of 15 males and 35 females. Two algorithms, incp_SVM and Incp-kSVM-RBF, performed better than the others. The experimental results illustrated the efficiency of the proposed algorithm.

فصلنامه علمی دانشگاه ملی مهارت
بهار ۱۴۰۴، دوره ۲۲، شماره ۱، ۲۰۶-۱۸۳
آدرس نشریه: https://karafan.nus.ac.ir
:10.48301/kssa.2025.490050.3053

ارائه یک رویکرد طبقه بندی صوت برای ویلچرهوشمند فرمان پذیر صوتی برای کاربران فارسی زبان با استفاده از شبکه های یادگیری عمیق

محمد امیری

۱-استادیار، گروه مهندسی کامپیوتر، دانشگاه ملی مهارت، تهران، ایران.

اطلاعات مقاله

نوع مقاله: مقاله پژوهشی

دریافت مقاله: ۱۴۰۳/۰۹/۱۰
بازنگری مقاله: ۱۴۰۳/۱۲/۰۹
پذیرش مقاله: ۱۴۰۴/۰۲/۰۶

کلید واژگان:
تشخیص صدا
طبقه بندی صوت
یادگیری عمیق
شبکه های عصبی کانولوشنی
طیف نگار

***نویسنده مسئول:** محمد امیری
پست الکترونیکی:
amiri.mgh@gmail.com

چکیده

در هر جامعه ای، برخی از معلولان نخاعی فاقد توانایی های جسمی و حرکتی برای حرکت دادن اندام های خود هستند و نمی توانند از ویلچر معمولی استفاده کنند و به ویلچر با کنترل صوتی نیاز دارند. الگوریتم‌های طبقه‌بندی صوتی مبتنی بر یادگیری عمیق به‌عنوان یک جزء حیاتی در توسعه ویلچرهای فرمان پذیر صوتی محسوب می شوند.

طبقه‌بندی صوت یکی از چالش‌های حوزه شناسایی الگو می‌باشد. به دلیل نتایج مثبت حاصله، شبکه های عصبی کانولوشن به طور گسترده ای در زمینه تشخیص و طبقه بندی صدا مورد استفاده قرار گرفته اند. در این مقاله، روشی برای طبقه‌بندی صداهای محیطی بر اساس طیف‌نگار صوتی، با استفاده از شبکه‌های عصبی عمیق، برای طبقه‌بندی صداهای فارسی زبانان برای ساخت ویلچر فرمان پذیر صوتی ارائه شده است. برای پیاده سازی، از Inception-V3 به عنوان یک شبکه عصبی کانولوشن استفاده شده است که توسط مجموعه داده InceptionV3 از قبل آموزش داده شده است. در مرحله بعد با تصاویری که با استفاده از تصاویر طیف نگاری صدای محیط حدود ۵۰ فارسی زبان تولید شده بود، شبکه را آموزش دادیم. در فقدان مجموعه داده فارسی زبانان، مجموعه داده خود را با ۵۰ نفر شامل ۳۵ مرد و ۱۵ زن در محدوده سنی ۲۵ تا ۶۰ سال ایجاد کردیم. نتایج تجربی به میانگین دقت ۸۳,۳۳ درصد دست یافت. بنابراین ویلچر قادر به اجرای پنج دستور توقف، چپ، راست، جلو و عقب خواهد بود.

شاپای الکترونیکی: ۲۵۳۸-۴۴۳۰
شاپای چاپی: ۲۳۸۲-۹۷۹۶

مقدمه

در جامعه مدرن، تحرک بیشتر یک چالش اساسی در بخش بهداشت عمومی است، به‌ویژه برای افراد مسن و معلول که تنها زندگی می‌کنند. لذا، تقاضا برای سیستم‌های کمکی مانند ویلچرهای هوشمند، به‌ویژه برای بیمارانی که به دلیل آسیب‌دیدگی قادر به راه رفتن عادی نیستند، در صنعت مراقبت‌های بهداشتی به طور قابل توجهی افزایش یافته است [۲;۱]. افراد دارای مشکلات حرکتی نسبت به افراد عادی بیشتر افسرده یا مضطرب هستند [۳]. بنابراین بازیابی تحرک آنها باعث افزایش رفاه روانی و بهداشتی شود.

قدرت سیستم‌های رایانه‌ای مدرن فرصت‌های جدیدی را برای محققان تعامل انسان و ماشین ایجاد کرده است تا با استفاده از فناوری‌های پیشرفته، کیفیت زندگی افراد مسن و معلول را بهبود ببخشند. امروزه استفاده از ویلچر به‌عنوان پشتوانه اصلی تحرک برای افراد مسن و همچنین معلولین کاملاً رایج شده است [۱]. مدل‌های قبلی صندلی‌های چرخدار اگرچه راهی برای جابجایی معلولان فراهم می‌کند، اما معمولاً فقط برای بیماران دارای ناتوانی حرکتی در اندام تحتانی آن‌ها مناسب است ازطرفی قادر به کاهش وابستگی آن‌ها به سرپرستانشان نیز نیست.

ویلچر هوشمند مدرن گامی است به سمت سبک زندگی مستقل برای افراد معلول جسمی که درحال‌حاضر جمعیت آن‌ها به حدود ۶۵۰ میلیون نفر می‌رسد [۴]. یکی از ویژگی‌های کلیدی ویلچر هوشمند، قابلیت استفاده و تعامل آسان با دستگاه برای کاربرانی است که دارای نقص شدید ناشی از فلج مغزی، اختلال حرکتی و ... هستند [۳;۱] و نمی‌توانند ویلچر را با استفاده از یک جوی استیک[1]استاندارد کنترل کنند اما قادر به استفاده از مهارت-هایی مانند حرکت چشم، صورت، دست، زبان یا صدا هستند [۵].

امروزه، با توجه به در دسترسی گسترده به حسگرهای مختلف و ارزان قیمت [۷;۶]، تحقیقات در زمینه شناسایی و طبقه‌بندی داده‌های صوتی شتاب گرفته است. سیستم‌های مبتنی‌بر سنسورهای صوتی در زمینه‌های پزشکی، نظارتی، امنیتی و چندرسانه‌ای [۹;۸]، نظارت بر زیست آکوستیک [۱۰]، شناسایی مزاحمان در مناطق حیات‌وحش [۱۱]، نظارت صوتی [۱۲]، نظارت و شناسایی گونه‌های جانوری [۱۳]، تشخیص خودکار بیماری قلبی [۱۴]، آنالیز صوتی سیگنال گریه در نوزادان [۱۵]، تشخیص خودکار بیماری ریوی [۱۶]، نظارت بر امنیت در محیط‌های بدون ساختار [۶] مورد استفاده قرار گرفته است.

یکی از چالشهای ویلچرهای فرمان پذیر صوتی، تجزیه و تحلیل فرمانهای صوتی دریافتی از کاربر می باشد که به دلیل اینکه کاربر در محیط بیرون از ویلچر استفاده می کند، صدای دریافتی همراه با نویز و صداهای ناخواسته محیط است.

به‌طورکلی در یک سیستم شناسایی و طبقه‌بندی، به دلیل مناسب نبودن استفاده از داده‌های صوتی خام به‌عنوان ورودی شبکه لازم است ابتدا ویژگی‌های مختلفی از داده‌های صوتی خام استخراج گردد. بنابراین، فرآیند تشخیص و طبقه‌بندی صدا، متشکل از سه مرحله مختلف شامل پیش‌پردازش[2]سیگنال‌ها، استخراج ویژگی‌های[3]

[1] Joystick
[2] Pre-processing
[3] Feature extraction

منحصربه‌فرد و درنهایت استفاده از برخی از ابزارهای طبقه‌بندی برای تمایز بین کلاس‌ها می‌باشد. ویژگی‌های مختلفی را می‌توان از داده‌های صوتی استخراج نمود که از رایج‌ترین این ویژگی‌ها می‌توان از طیف‌سنجی (اسپکتروگرام)[1]، طیف سنجی مل، ضریب کپسترال فرکانس مل، تصویر شنوایی تثبیت شده و ضرایب پیش‌بینی خطی نام برد. علاوه‌براین، از الگوریتم‌های مختلف یادگیری ماشین با ناظر ازجمله درخت تصمیم، جنگل تصادفی و نزدیکترین همسایه، ماشین بردار پشتیبان[2]، مدل‌های مخفی مارکوف، پرسپترون چند لایه و یادگیری عمیق[3] برای توسعه سیستم‌های تشخیص و طبقه‌بندی صدا استفاده شده است. نتایج بسیار مثبت بکارگیری روش‌های یادگیری عمیق به‌ویژه شبکه‌های عصبی کانولوشن[4] در سال‌های اخیر موجب استفاده گسترده این روش در زمینه شناسایی و طبقه‌بندی صدا شده است.

یادگیری عمیق به طور فزاینده‌ای در طبقه بندی گفتار مورد استفاده قرار گرفته است. شبکه های عمیق به دلیل توانایی آنها در یادگیری الگوهای پیچیده، می‌تواند دقت بهتری نسبت به رویکردهای سنتی به دست آورد [۱۸;۱۷]. رویکردهای سنتی معمولاً طبقه‌بندی صوت را به دو مرحله تقسیم می‌کنند: استخراج ویژگی و طبقه‌بندی [۱۹].

مدل‌های یادگیری عمیق به شدت به کیفیت داده‌هایی که بر روی آن‌ها آموزش دیده‌اند، وابسته هستند. اگر داده‌ها پر از نویز باشد، نتایج حاصله می‌تواند به طور قابل توجهی تحت تأثیر قرار گیرد [۲۰].

بنابراین، مدل‌های یادگیری عمیق برای طبقه‌بندی صوت برای حل وظایف پیچیده گسترش یافته‌اند. با این حال، برخی محدودیت‌ها هم در استفاده از این شبکه ها مانند هزینه محاسباتی بالا، عدم قابلیت تفسیر، بیش‌برازش و مجموعه‌های داده پیش‌بینی‌نشده وجود دارد. [۲۲;۲۱;۱۹]

زمان و همکاران [۲۳] از یک شبکه کانولوشنی با تصاویر طیف‌نگاری اسپکتروگرام برای طبقه‌بندی گفتار مضر در دستگاه‌های سمعک استفاده کردند. پس از تبدیل سیگنال صوتی به طیف‌نگاری، گفتارها به شش گروه شامل صدای سالم سالم و پنج نوع نویز مختلف طبقه‌بندی شد. نتایج تحقیقات آنها نشان داد که مدل آن‌ها می‌تواند گفتار مضر را با دقت ۹۹٪ به درستی طبقه‌بندی کند.

بالسترون و همکاران [۲۴] یک شبکه کانولوشنی به نام Deep4SNet را برای شناسایی صدای تقلبی با استفاده از هیستوگرام سیگنالهای صوتی پیشنهاد کردند. آن‌ها همچنین همان داده‌های گفتار را با استفاده از یک مدل یادگیری ماشین با ویژگی‌های دست‌ساز آموزش دادند. در یک مجموعه داده سفارشی، مدل آن‌ها به دقت ۹۸,۵٪ دست یافت.

---

[1] Spectrogram
[2] Support vector machine (SVM)
[3] Deep learning
[4] Convolutional neural network (CNN)

وِربچویچ و همکاران [۲۵] یک شبکه کانولوشنی با طیف‌نگاری برای طبقه‌بندی احساسات پیشنهاد کردند. آن‌ها ابتدا سیگنال‌های گفتار را به تصاویر طیف‌نگاری تبدیل کردند و تکنیک‌های مختلف افزایش داده، مانند کاهش نمونه و نویز، را به کار بردند تا اندازه داده‌های آموزشی را افزایش دهند.

سی و همکاران [۲۶] پیشنهاد کردند که از یک شبکه عصبی کانولوشنی برای دسته‌بندی صوتی در مجموعه‌های داده کم‌منابع استفاده شود، مانند مجموعه‌هایی که داده‌های آموزشی کمی دارند و مستعد بیش‌برازش هستند. آنها آزمایش‌هایی بر روی مجموعه‌های داده صوتی مختلف انجام دادند و بهبودهای قابل توجهی در دقت دسته‌بندی تا ۵٫۰٪ در شرایط کم‌منبع نسبت به مدل‌های پایه به دست آوردند.

فام و همکاران [۲۷] یک شبکه عصبی کانولوشنی ترکیبی برای دسته‌بندی صحنه‌های صوتی پیشنهاد کردند. آن‌ها سه طیف‌نگاشت را در یک تصویر ترکیب کردند که برای آموزش ترکیبی آن‌ها استفاده شد.

جنا و همکاران [۲۸] یک معماری عمیق چندوجهی برای دسته‌بندی ژانر موسیقی پیشنهاد کردند. آن‌ها به‌طور خاص از دو نوع ورودی گفتاری استفاده کردند؛ طیف‌نگاشت و موجک‌ها.

اسکارپینیتی و همکاران [۲۹] یک شبکه بازگشتی عمیق[1] با استفاده از حافظه طولانی کوتاه مدت[2] تکراری برای دسته‌بندی ضبط‌های صوتی در محل ساخت و ساز پیشنهاد کردند. ورودی شبکه از ویژگی‌های طیفی متعددی مانند طیف‌نگاشت مقیاس مل[3]، کروم و کنتراست طیفی تشکیل شده است. این مدل به دقت کلی تا ۹۷٪ در مجموعه تست دست یافت و از سایر مدل‌ها پیشی گرفت.

یو و همکاران [۳۰] یک مدل Bi-RNN با مکانیزم توجه ارائه کردند. آن‌ها همچنین دو مدل مبتنی بر توجه متفاوت، سری و موازی، را پیاده‌سازی کردند تا عملکرد آن‌ها را با استفاده از طیف‌نگاشت‌های اسپکتروگرام مقایسه کنند. نتایج نشان داد که مدل توجه موازی مؤثرتر است و نتایج بهتری نسبت به مدل توجه سری به دست می‌آورد.

سریواستاوا و همکاران [۳۱] استفاده از CNN-GRU و CNN-LSTM را برای طبقه‌بندی سیگنال‌های صوتی پیشنهاد کردند. آنها از طیف‌نگاشت مقیاس مل برای نمایش داده‌های صوتی استفاده کردند. نتایج نشان داد که دقت استفاده از CNN-GRU برابر ۸۵٫۷ درصد و دقت استفاده از معماری CNN-LSTM برابر ۸۷٫۵ درصد بوده است.

نیگرو و همکاران [۳۲] اثربخشی یک روش تأکید بر زمان- فرکانس- انرژی را نسبت به طیف‌نگاری مل برای طبقه‌بندی داده های صوتی با استفاده از CNN-RNN ارزیابی کردند. ارزیابی‌ها نشان داد که روش آن‌ها تعداد پارامترهای آموزشی را به نصف کاهش می‌دهد در حالی که دقت بهتری نسبت به استفاده از طیف‌نگاری مل حفظ می‌کند.

---

[1] Deep recurrunt neural network (DRNN)

[2] Long short- term memory (LSTM)

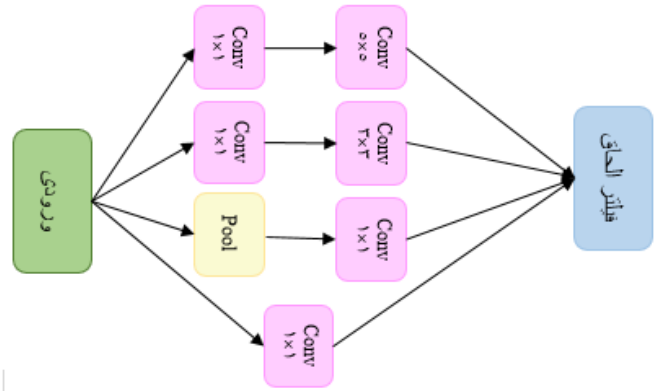
[3] Mel-Frequency Cepstral Coefficients (MFCC)

الگوریتم پیشنهادی برای طبقه بندی صوت برای کاربران فارسی

معماری InceptionV3

InceptionV3 [۳۳] یک معماری شبکه عصبی کانولوشن از خانواده Inception است. این معماری نسخه بهبود یافته معماری GoogleNet است که توسط گوگل در سال ۲۰۱۴ پیشنهاد شده است. از آنجا که هسته اصلی شبکه GoogleNet ماژول Inception است، شبکه GoogleNet نیز شبکه Inception نامیده می شود [۳۴].

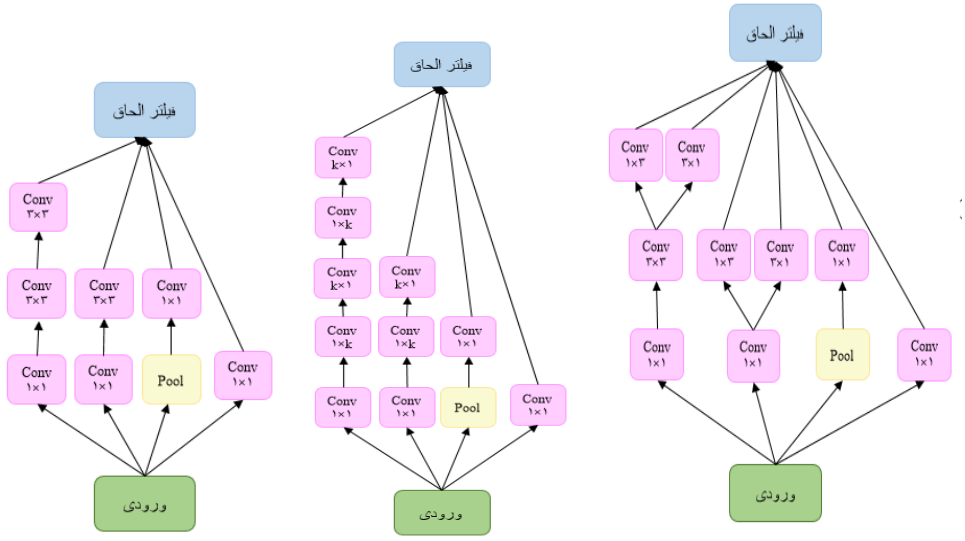
ماژول Inception دارای لایه‌هایی است که به صورت موازی قرار گرفته‌اند که معمولاً شامل مجموعه‌ای از لایه‌های کانولوشن با سه اندازه مختلف و یک لایه Max-pooling است. به این ترتیب می‌توان اطلاعات را در سطوح مختلف استخراج نمود. بنابراین ویژگی‌های فضایی می‌توانند به طور موثرتری استخراج شوند. ساختار شبکه Inception به‌گونه‌ایست که علاوه برافزایش عمق شبکه، تعداد پارامترهای شبکه را نیز کاهش می‌دهد که موجب کاهش پیچیدگی محاسباتی و افزایش دقت و کارایی شبکه می‌گردد و از برازش بیش از حد نیز جلوگیری می‌کند. از این رو، به طور گسترده‌ای در کارهای طبقه‌بندی تصویر استفاده می‌شود. شکل ۱ ساختار ماژول Inception را نشان می‌دهد.



شکل ۱. ساختار کلی ماژول Inception

در معماری پیشنهادی InceptionV3، برای بهبود کارایی معماری Inception، فاکتورسازی کانولوشن‌های بزرگ به کانولوشن‌های کوچکتر ارائه شد. درواقع در این معماری، ماژول Inception فیلترهای بزرگتر (مانند

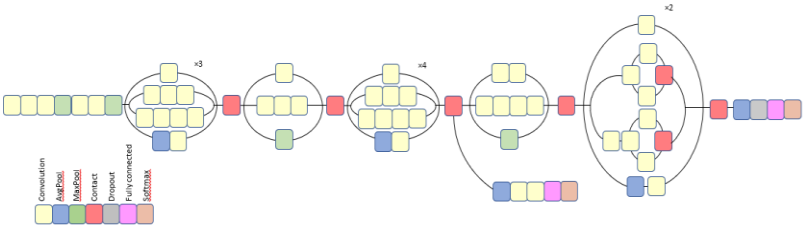
فیلترهایی با ابعاد ۵×۵، ۷×۷) را که از نظر محاسباتی گران هستند را با فیلترهای متوالی کوچکتر که دارای عملکرد یکسانی هستند، جایگزین می‌کند. به عنوان مثال، یک کانولوشن ۳×۳ به دو کانالوشن ۳×۱ و ۱×۳ جایگزین می-شود. این جایگزینی منجر به کاهش تعداد پارامترها و افزایش سرعت آموزش شبکه می‌شود. شکل ۲ ساختار ماژول‌های InceptionV3 را نشان می‌دهد. در شکل (۵-الف)، ماژول فاکتورسازی فیلترهای ۵×۵ به دو فیلتر ۳×۳، در شکل (۵-ب) ماژول فاکتورسازی برای کانولوشن‌هایی با ابعاد k×k و در شکل (۵-ج) یک بانک فیلتر گسترش یافته برای افزایش ابعاد نمایش‌ها ارائه شده است. بانک‌های فیلتر موجود در ماژول به منظور حذف گلوگاه[1] گسترش یافته اند (به‌جای عمیق‌تر، عریض‌تر شده‌اند). اگر ماژول به‌جای آن عمیق‌تر شود، منجربه کاهش بیش از حد ابعاد و در نتیجه از بین رفتن اطلاعات مفید می‌شود.



شکل ۲. ساختار ماژولهای InceptionV3 الف) سه لایه با فیلترهای مربعی به اندازه های ۳،۱ و ۳. ب) پنج لایه با فیلترهای مربعی اندازه های هم اندازه ۱. ج) سه لایه با اندازه های مربعی ۱ و ۳ و غیر مربعی ۳*۱

معماری شبکه InceptionV3 در شکل ۶ نشان داده شده است. اندازه ورودی این شبکه به‌صورت پیش‌فرض تصاویر با ابعاد ۲۹۹×۲۹۹ با سه کانال است. لایه‌های ابتدایی شامل لایه‌های کانولوشن و pooling و در ادامه ماژول‌های مختلف Inception (نوع 3،2،V3-1) هستند.

[1] Bottleneck



شکل ۳. معماری شبکه InceptionV3

انتقال یادگیری[1]

در هنگام کار بر روی مسائل نوظهور که مجموعه داده کافی وجود ندارد و یا کار بر روی مجموعه داده کوچکی از تصاویر، آموزش یک شبکه یادگیر عمیق با وزن‌های تصافی کارایی مناسبی نخواهد داشت. شبکه‌های یادگیر عمیق، باتوجه به تعداد زیاد پارامترها و لایه‌های موجود، به تعداد زیادی داده (میلیون‌ها داده) بعلاوه استفاده از روش‌های داده افزایی[2] برای آموزش شبکه نیاز دارند. اگر مجموعه داده آموزشی به اندازه‌ی کافی وجود نداشته باشد، شبکه در طول آموزش دچار کم برازش[3] می‌شود. برای بکارگیری شبکه‌های یادگیر عمیق روی یک مجموعه داده کوچک، از مفهوم انتقال یادگیری استفاده می‌شود. بر اساس تئوری یادگیری انتقال، وزن‌های شبکه‌های از قبل آموزش دیده، می‌توانند برای مجموعه داده‌های دیگر معنی‌دار باشند، لذا می‌توان از آن‌ها به عنوان استخراج-کننده ویژگی‌های تصاویر در حوزه‌های مختلف استفاده کرد.

با استفاده از روش انتقال یادگیری می‌توان سیستم‌های طبقه‌بندی کننده با عملکرد بالا برای طبقه‌بندی مجموعه داده‌های کوچک ایجاد نمود. در این زمینه می توان از مدل‌های از پیش آموزش دیده[4] روی مجموعه داده‌های بزرگ و یادگیری ویژگی‌های آن استفاده نمود. مدل‌هایی مانند مدل VGGNet، ResNet و Inception با مجموعه داده‌های بزرگی مانند InceptionV3 آموزش داده شده‌اند. بنابراین می‌توان، پارامترهای وزنی که توسط InceptionV3 از قبل آموزش دیده‌اند به‌عنوان وزن‌های اولیه شبکه، مقداردهی شوند. با این‌کار ویژگی‌های قبلاً آموخته شده به مدل ما انتقال داده می‌شود.

اسپکتروگرام

اسپکتروگرام یا طیف‌سنج یک روش برای تجسم طیف فرکانسی موج صدا است. به‌عبارت ساده‌تر، طیف‌سنجی نمایش از داده‌های صوتی مبتنی بر نوعی طیف‌سنج در زمان واقعی است [۳۷-۳۵]. شکل ۴ نمونه‌ای از اسپکتوگرام

[1] Transfer learning
[2] Augmentation data
[3] Underfitting
[4] Pre-train models

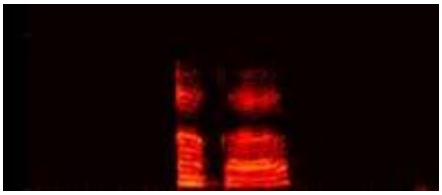
یک فایل صوتی را نشان می‌دهد. از تصاویر اسپکتروگرام می‌توان به همراه طبقه‌بندی کننده‌های مختلف یادگیری ماشین از جمله روش‌های مبتنی بر یادگیری عمیق استفاده کرد. بررسی‌های انجام شده توسط لیو و همکاران در مورد مدل‌های یادگیری عمیق، نشان می‌دهد که شبکه عصبی کانولوشن در تصاویر و داده‌های ویدئویی از سایر مدل‌ها بهتر عمل کرده است [۳۵]. از آنجایی که یک عکس از طیف، نمایشی تصویری از طیف فرکانسی یک سیگنال است، در روش‌های یادگیری عمیق برای انجام استخراج و طبقه‌بندی ویژگی‌ها از تصاویر طیف‌سنج استفاده می‌شود. سیگنال‌های صوتی کمتر مکرر بوده و الگوی متفاوتی را در طیف نمایش ایجاد می‌کنند.

تبدیل فوریه سریع[1] تبدیل فوریه زمان کوتاه[2](اسپکتروگرام)، تبدیل ویولت و چند تابع تبدیل دیگر می‌توانند برای تبدیل سیگنال صوتی یک بعدی به یک تصویر دوبعدی مورد استفاده قرار گیرند. تبدیل فوریه زمان کوتاه، نسخه بهبود یافته تبدیل فوریه است که مساله زمان را هم لحاظ می کند و بیان می‌کند که در هر پنجره زمانی مشخص، کدام مولفه های فرکانسی وجود دارند، این تبدیل اطلاعات مفیدی درباره شکل موج فراهم می‌کند. به‌عنوان مثال، چه فرکانسهایی و با چه میزان قدرتی در شکل موج وجود دارد. فرمول محاسبه این تبدیل به صورت زیر می‌باشد:

$$S(f,t) = \int_{-T}^{T} s(\theta) w(\theta - t) e^{-i2\pi f \theta} d\theta \quad (۱)$$

اسپکتروگرام از مربع اندازه تبدیل فوریه کوتاه به صورت زیر بدست می آید [۲۳]:

$$\text{Spectrogram} = |S(f,t)|^2 \quad (۲)$$



شکل ۴. تصویر اسپکتوگرام یک فایل صوتی

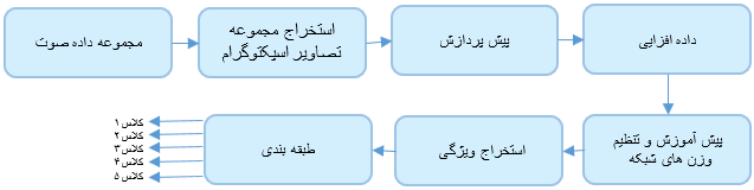
روش پیشنهادی

هدف ما در این مقاله ساخت یک مدل مبتنی‌بر یادگیری عمیق برای طبقه‌بندی و تشخیص صوت بر اساس یک مجموعه داده کوچک که با استفاده از روش پیشنهادی محقق گردید. به‌طورکلی یک مدل طبقه‌بندی کننده از سه مرحله اصلی شامل پیش‌پردازش، استخراج ویژگی و طبقه‌بندی داده تشکیل شده است. پس از بررسی تکنیک‌های فوق، ما معماری شبکه عمیق InceptionV3 را به‌عنوان استخراج کننده‌های ویژگی و طبقه‌بندی کننده تصاویر اسپکتوگرام در روش پیشنهادی در نظر گرفتیم. ورودی‌های سیستم، تصاویر اسپکتوگرام از داده‌های

[1] Fast Fourier transform (FFT)

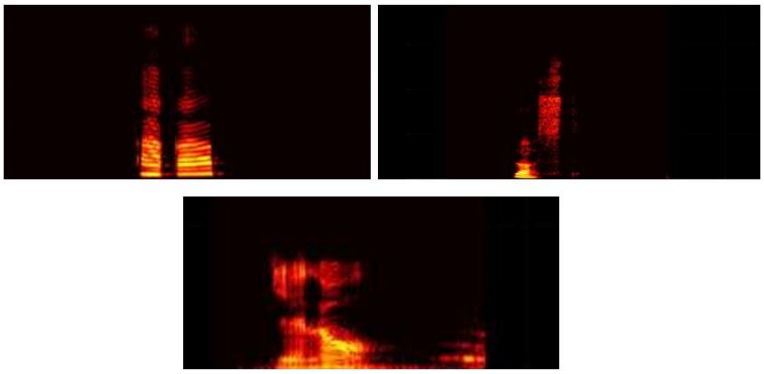
[2] Short-term Fourier transform (STFT)

صوتی است. ازآنجایی که شبکه‌های عصبی کانولوشن برای آموزش به مجموعه داده‌های زیادی نیاز دارند، از روش‌های داده افزایی استفاده نمودیم. در ادامه، برای بهبود تشخیص و عملکرد سیستم، با بکارگیری تئوری انتقال یادگیری، در مرحله اولیه، مدل از پیش آموزش دیده روی مجموعه داده تصاویر طبیعی InceptionV3 را در نظر گرفتیم. این کار باعث می‌شود شبکه بازنمایی ویژگی‌های غنی برای طیف گسترده‌ای از تصاویر را بیاموزد. انتقال یادگیری، ایده‌ای برای استفاده از وزن‌های شبکه‌های از قبل آموزش دیده، برای بهبود کارایی مدل یا انجام یک کار خاص مانند تصویربرداری پزشکی یا تشخیص رویدادهای صوتی در محیط‌های واقعی می‌باشد. شبکه InceptionV3 با استفاده از انتقال یادگیری، عملکرد طبقه‌بندی خوبی را در حل مسائل مختلف بدست آورده است. ما از مدل از پیش‌آموزش دیده InceptionV3 برای افزایش کارایی و عملکرد سیستم استفاده نمودیم. ساختار کلی متد پیشنهادی در شکل ۵ آورده شده است.



شکل ۵. ساختار کلی روش پیشنهادی

داده‌های ورودی، فایل‌های صوتی شامل صوت کلمات چپ، راست، جلو، عقب و ایست هستند. در ابتدا تمامی فایل‌های صوتی را به پسوند wav تبدیل، سپس داده‌های صوتی را به تصاویر اسپکتوگرام تبدیل نمودیم. برای انجام این کار از تابع اسپکتروگرام نرم افزار متلب استفاده شده است. پارامتر Hamming=512، Noverlap=256 و NFFT= 1024 درنظرگرفته شد. شکل ۶ نمونه تصاویر اسپکتوگرام از مجموعه داده‌های ورودی را نشان می‌دهد.



شکل ۶.سه نمونه از تصاویر اسپکتوگرام از مجموعه داده‌های ورودی

پیش‌پردازش فرآیندی بسیار رایج برای از بین بردن نویز یا صدای ناخواسته، تأکید بر جنبه‌هایی از تصویر که می‌توانند به وظیفه شناسایی کمک کنند یا حتی در مرحله آموزش یادگیری عمیق مفید باشند. برای مدل‌های شبکه کانولوشن، تصاویر ورودی اغلب برای حفظ سازگاری با معماری شبکه تغییر اندازه می‌یابند. بدین منظور ابعاد تصاویر به ۳×۲۹۹×۲۹۹ تغییر داده شد. بعد از انجام پیش‌پردازش روی تصاویر اسپکتوگرام بدست آمده، مجموعه داده‌های اولیه را ایجاد نمودیم. این مجموعه داده برای استخراج مقادیر توصیف‌کننده ویژگی هر صوت استفاده می‌شود. در ادامه با استفاده از تکنینک‌های سنتی داده افزایی مانند جابه‌جایی به چپ یا راست پیکسل‌های تصویر تعداد داده‌ها را افزایش دادیم. در مرحله استخراج ویژگی از مدل از قبل آموزش دیده InceptionV3 روی مجموعه تصاویر طبیعی InceptionV3 استفاده نمودیم.

بدین منظور از انجماد لایه‌ها (عدم اجازه تغییر آن‌ها)، حذف لایه‌ها، ایجاد لایه‌های جدید استفاده نمودیم. ابتدا به انتهای مدل یک لایه Average pooling و ۵ لایه تمام متصل[1] اضافه می‌نماییم. لایه تمام متصل اول تا چهارم با ابعاد ۱۰۲۴، ۵۱۲، ۲۵۶ و ۱۲۸ نرون و تابع فعال ساز Relu و لایه تمام متصل آخر با ابعاد ۵ و تابع فعال ساز Softmax می‌باشد. مقدار ۵ نشان‌دهنده تعداد کلاس‌های موجود در شبکه ما است. وزن شبکه برای لایه‌های اضافه شده به‌صورت تصادفی مقداردهی می‌شود. تمامی لایه‌های کانولوشن مدل از پیش آموزش داده شده را منجمد می‌نماییم. شبکه را برای batch_size=10، ۳۰ دورآموزش می‌دهیم. در این مرحله فقط مقادیر پارامترهای لایه‌های اضافه شده تغییر می‌کنند. در این مرحله لایه‌های بالایی به خوبی آموزش دیدند و در ادامه می‌توانیم پارامترهای لایه‌های کانولوشن شبکه InceptionV3 را تنظیم[2] نماییم. برای این کار ما تنها ۲ لایه Inception بالایی شبکه InceptionV3 را آموزش می‌دهیم بنابراین تمام لایه‌های پایینی شبکه را منجمد می‌کنیم و ۲ لایه بالای شبکه را از انجماد خارج می‌کنیم. مجدد شبکه را برای batch_size=10، ۳۰ دور آموزش می دهیم. در واقع شبکه آموزش‌دیده به‌عنوان استخراج کننده ویژگی عمل می‌کنند و دو لایه آخر (لایه‌های کاملاً متصل) طبقه‌بندی را انجام می‌دهند. این ساختار شبکه انتقال یادگیری می‌باشد. در ادامه برای بهبود شبکه، از Inception به عنوان استخراج کننده ویژگی و از طبقه‌بندی کننده‌های SVM و Kernel SVM با تابع کرنل-های مختلف Poly، RBF و Sigmoid برای طبقه‌بندی فایل‌های صوتی استفاده نمودیم.

پایگاه داده

همان‌طور گفته شد، هدف از این مقاله شناسایی و طبقه‌بندی صدا برای ساخت ویلچر هوشمند فرمان‌پذیر صوتی در یک محیط طبیعی است. لذا برای این کار، ما مجموعه داده صوتی متشکل از کلماتی برای هدایت و جابه‌جایی ویلچر هوشمند شامل: جلو، عقب، چپ، راست و ایست ایجاد نمودیم. این مجموعه داده شامل ۲۱۰ فایل

[1] Full connected
[2] Fine-tuning

صوتی می‌باشد که از ۵۰ نفر شامل ۱۵ زن و ۳۵ مرد جمع‌آوری شده است. توضیح مختصری از این مجموعه داده در جداول شماره ۱ و ۲ نشان داده شده است:

جدول ۱. مشخصات افرادی که نمونه فایل صوتی آن‌ها ضبط شده است.

جنسیت	تعداد	سن
زن	۱۵	۲۰-۶۰
مرد	۳۵	۲۵-۶۰

جدول ۲. تعداد نمونه‌های مربوط به هریک از فرامین صوتی

کلمات	جلو	عقب	چپ	راست	ایست	تعداد کل
تعداد نمونه	۴۱	۴۲	۴۵	۴۲	۴۰	۲۱۰
تعداد نمونه بعد از داده افزایی	۲۰۵	۲۱۰	۲۲۵	۲۱۰	۲۰۰	۱۰۵۰

برای تولید نمونه‌های بیشتر از روش‌های داده‌افزایی استفاده شده است. روش‌های داده‌افزایی تکنیک‌هایی هستند که با افزودن نسخه‌هایی با تغییرات کم از داده‌های موجود یا داده‌های مصنوعی ایجاد شده از داده‌های موجود، به منظور افزایش تعداد داده‌ها استفاده می شود. بدین منظور با جابه‌جایی به چپ یا راست پیکسل‌های تصویر به میزان ۲۵ و ۵۰ پیکسل، تعداد داده‌ها را افزایش دادیم. در آزمایشات انجام شده، بطور تصادفی ۹۰٪ از داده‌ها به عنوان داده‌های آموزشی و ۱۰٪ به‌عنوان داده‌های تست درنظر گرفته شده است. همچنین برای پیش‌آموزش شبکه از مجموعه داده InceptionV3 استفاده شده است. InceptionV3، مجموعه‌ای متشکل از بیش از ۱۷ میلیون تصویر با وضوح بالا که روزبه‌روز بر تعداد آن افزوده می‌شود. این مجموعه داده حدود ۲۲۰۰۰ دسته را شامل می‌شود.

نتایج آزمایشات

در این بخش، آزمایشات انجام شده برای نشان دادن عملکرد روش پیشنهادی شرح داده شده است. چندین آزمایش برای ارزیابی کارآیی روش پیشنهادی روی مجموعه داده جمع‌آوری شده انجام شده است. آزمایش‌ها کارایی بالای روش پیشنهادی را نشان می‌دهد.

آزمایش ۱.

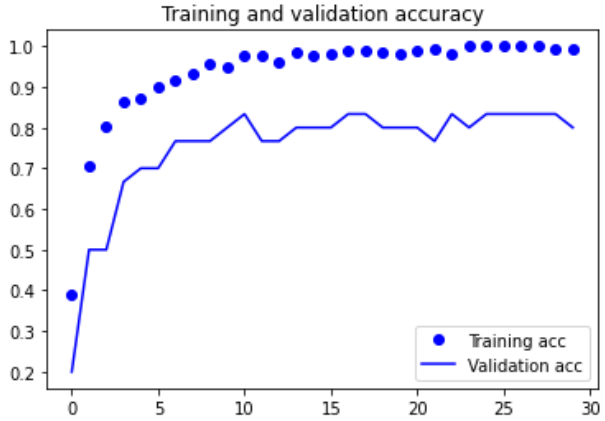
هدف از این آزمایش نشان دادن تاثیر مقادیر پارامترهای Hamming، Noverlap و NFFT برای استخراج تصاویر اسپکتوگرام در تشخیص و طبقه‌بندی صوت است. این آزمایش با افزودن ۴ لایه تماماً متصل دارای ۱۰۲۵، ۵۱۲، ۲۵۶ و ۱۲۸ نرون، برای تعداد تکرار ۲۵ دور برای تنظیم و تعداد تکرار ۲۵ دور برای آموزش انجام شده است. همچنین ۱۰٪ مجموعه داده، به عنوان داده تست درنظر گرفته شده است. نتایج آزمایش در جدول ۳ آورده شده است. مشاهده می‌شود، برای مقادیر Hamming = 512، Noverlap = 256 و NFFT = 1024 بهترین عملکرد را روش‌های Incp-Svm و Incp-kSvm-RBF با دقت ۸۰٪ و برای مقادیر Hamming = 1024، Noverlap = 512 و NFFT = 2048 بهترین عملکرد را Incp-Svm با درصد دقت برابر ۷۷٪ بدست آوردند. با توجه به اینکه برای مقادیر Hamming = 512، Noverlap = 256 و NFFT = 1024 اکثر روش‌ها سیستم دارای عملکرد بهتری می‌باشند، بنابراین در آزمایشات بعدی از این مقادیر استفاده شده است.

جدول ۳. بررسی حساسیت پارامترهای Hamming، Noverlap و NFFT

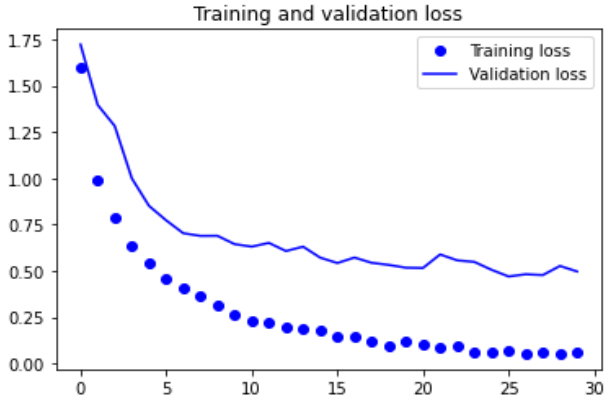
	Inception	Incp-Svm	Incp-KSvm-Poly	Incp-kSvm-RBF	Incp-kSvm-Sigmoid
Hamming = 1024 Noverlap = 512 NFFT = 2048	۷۳,۳۳	۷۷	۴۷	۷۵	۷۲
Hamming = 512 Noverlap = 256 NFFT = 1024	۷۸,۱۰	**۸۰**	۴۶	**۸۰**	۷۷

آزمایش۲.

هدف از این آزمایش نشان دادن همگرایی و میزان دقت الگوریتم پیشنهادی است. همانطور که در شکل ۷ و ۸ نشان داده شده است، برای تعداد تکرار برابر ۳۰ و تعداد دسته‌های برابر ۱۰ دقت سیستم برابر ۸۳,۳۳ می‌باشد.



شکل ۷. نمودار دقت روش پیشنهادی بر روی داده های آموزشی و اعتبار سنجی



شکل ۸. میزان خطای روش پیشنهادی بر روی داده های آموزشی و اعتبار سنجی

آزمایش۳.

در این آزمایش، کارایی روش‌های Inception، Incp-Svm، Incp-KSvm-Poly، Incp-kSvm-RBF و Incp-kSvm-Sigmoid برای دو تابع فعال‌ساز مختلف شامل Relu و Sigmoid مقایسه می‌کنیم. نتایج آزمایشات در جدول ۴ نشان داده شده است. این آزمایش با افزودن ۴ لایه تماماً متصل دارای ۱۰۲۵، ۵۱۲، ۲۵۶ و ۱۲۸ نرون، برای تعداد تکرار ۲۵ دور برای تنظیم و تعداد تکرار ۲۵ دور برای آموزش انجام شده است. در این آزمایش ۱۰٪ مجموعه داده به عنوان داده تست در نظر گرفته شده است. همان‌طور که از جدول شماره ۴ مشاهده می‌کنید، استفاده از تابع فعال‌ساز Relu بهتر از Sigmoid می‌باشد بهترین مقدار توسط دو روش Incp-Svm و Incp-kSvm-RBF با دقت ۸۰٪ بدست آمده است. فقط در روش Incp-KSvm-Poly تابع Sigmoid نسبت به Relu دارای عملکرد بهتری بوده است.

جدول ۴. مقایسه دو تابع فعال‌ساز Relu و Sigmoid

	Inception	Incp-Svm	Incp-KSvm-Poly	Incp-kSvm-RBF	Incp-kSvm-Sigmoid
Relu	۷۸,۱۰	۸۰	۴۶	۸۰	۷۷
Sigmoid	۷۰,۴۸	۶۸	۷۰	۶۸	۷۰

آزمایش۴.

در این آزمایش، عملکرد روش‌های Inception، Incp-Svm، Incp-KSvm-Poly، Incp-kSvm-RBF و Incp-kSvm-Sigmoid را به ازای مقادیر مختلف برای پارامترهای تعداد دور آموزش، تعداد دور برای مرحله تنظیم و تعداد لایه‌های تماماً متصل مورد ارزیابی قرار گرفت. همچنین تاثیر استفاده از آموزش دسته‌ای در هر دو مرحله آموزش و تنظیم یا تنها در مرحله آموزش نیز بررسی گردید. نتایج این آزمایشات در جدول ۵ آورده شده است. در آزمایشات انجام شده تعداد نرون‌ها برای ۲ لایه تماماً متصل برابر ۱۰۲۴ و ۵۱۲ می‌باشد. درحالیکه تعداد نرون‌ها در هنگام بکارگیری از ۴ لایه تماماً متصل برابر ۱۰۲۴، ۵۱۲، ۲۵۶ و ۱۲۸ نرون است. با توجه به نتایج بدست آمده در جدول مشاهده می‌شود که دو روش Incp-kSvm-RBF و Incp-Svm با تعداد لایه‌های تماماً متصل برابر ۴ و استفاده از آموزش دسته‌ای فقط در یک مرحله، بر روی مجموعه داده‌ی صوتی ما عملکرد بهتری دارد. همچنین با توجه به مقادیر حاصل، یادگیری واقعاً سریع است زیرا می‌بینیم که در چند دوره به مقادیر دقت بالا رسیده و بنابراین همگرایی سریع است.

جدول ۵. بررسی حساسیت پارامترهای مختلف

	Inception	Incp-Svm	Incp-KSvm-Poly	Incp-kSvm-RBF	Incp-kSvm-Sigmoid
Iter=50-50	۷۴,۲۹	**۷۶**	۴۹	۷۳	۶۸
FC=2					
BN=10					
Iter=15-15	**۷۴,۲۹**	۷۱	۶۵	۷۲	۷۲
FC=2					
BN=10					
Iter=15-15	**۷۹,۰۵**	۷۷	۵۹	۷۸	۷۳
FC=4					
BN=10					
Iter=15-15	۷۰,۴۸	**۷۲**	۵۲	۷۰	**۷۲**
FC=4					
BN=10-10					

Iter=25-25	۷۸,۱۰	۸۰	۴۶	۸۰	۷۷
FC=4					
BN=10					

آزمایش۵.

در این آزمایش برای ارزیابی روش‌های Incp-Svm و Incp-kSvm-RBF، از معیارهای Precision، Recall و F1-score استفاده شده است. نتایج در جدول ۶ نشان داده شده است. مجموعه داده به مجموعه آموزش و مجموعه اعتبارسنجی[1](۹۰٪ به ۱۰٪) با تعداد برابر برای هر کلاس تقسیم شده است. با توجه به نتایج بدست آمده، شبکه نتایج بالایی را در تمام معیارها بدست می‌آورد و اثربخشی خود را در این حوزه خاص نشان می‌دهد.

جدول ۶. ارزیابی روش‌های Incp-Svm و Incp-kSvm-RBF، با استفاده از معیارهای Precision، Recall و F1-score

کلاس	Incp-kSvm-RBF			Incp-Svm		
	precision	recall	f1-score	Precision	recall	f1-score
عقب	۱۰۰	۷۱	۸۳	۱۰۰	۷۱	۸۳
جلو	۸٤	۷٦	۸۰	۸٥	۸۱	۸۳
چپ	۷٤	۸۱	۷۷	۸۱	۸۱	۸۱
راست	۷۹	۷۱	۷٥	۷۸	٦۷	۷۲
ایست	۷۲	۱۰۰	۸٤	٦۸	۱۰۰	۸۱
میانگین وزنی	۸۲	۸۰	۸۰	۸۲	۸۰	۸۰

آزمایش۶.

نتایج صحت بدست آمده در ماتریس سردرگمی[2]در جداول ۷ و ۸ برای دو روش Incp-Svm و Incp-kSvm-RBF نشان داده شده است. در هر دو جدول، مشخص است که همه کلاس‌ها با تقریب خوبی به‌درستی شناخته شده‌اند. دقت هر دو روش برابر ۸۰٪ می‌باشد اما با توجه به ماتریس سردرگمی، در روش Incp-Svm

[1] Validation
[2] Confusion Matrix

بهترین نتیجه برای برچسب "ایست" با دقت ۱۰۰٪ و بدترین نتیجه برای برچسب "راست" با دقت ۶۶,۶۶٪ حاصل شد. درحالی‌که در روش Incp-kSvm-RBF بهترین نتیجه برای برچسب "ایست" با دقت ۱۰۰٪ و بدترین نتیجه بدست آمده برای برچسب های "راست" و "عقب" برابر ۷۱,۴۲٪ می باشد.

جدول ۷. ماتریس سردرگمی حاصل از روش Incp-Svm بر حسب درصد

	عقب	جلو	چپ	راست	ایست
عقب	۷۱,۴۲	۹,۵۲	۰	۱۴,۲۸	۴,۷۶
جلو	۰	۸۰,۹۵	۹,۵۲	۰	۹,۵۲
چپ	۰	۴,۷۶	۸۰,۹۵	۴,۷۶	۹,۵۲
راست	۰	۴,۷۶	۴,۷۶	۶۶,۶۶	۲۳,۸۰
ایست	۰	۰	۰	۰	۱۰۰

جدول ۸. ماتریس سردرگمی حاصل از روش Incp-kSvm-RBF بر حسب درصد

	عقب	جلو	چپ	راست	ایست
عقب	**۷۱,۴۲%**	۹,۵۲%	۰%	۱۴,۲۸%	۴,۷۶%
جلو	۰	**۷۶,۱۹**	۱۴,۲۸	۰	۹,۵۲
چپ	۰	۴,۷۶	**۸۰,۹۵**	۴,۷۶	۹,۵۲
راست	۰	۰	۱۴,۲	**۷۱,۴۲**	۱۴,۲
ایست	۰	۰	۰	۰	**۱۰۰**

نتیجه گیری

در این مقاله، روشی برای طبقه‌بندی صداهای محیطی بر اساس طیف‌نگار صوتی (اسپکتروگرام)، به منظور استفاده در ویلچرهای فرمان پذیر صوتی با قابلیت دریافت پنج فرمان توقف، چپ، راست، جلو و عقب با استفاده از شبکه‌های عصبی عمیق، برای طبقه‌بندی صداهای فارسی زبانان ارائه شده است. برای پیاده سازی، از Inception-V3 به عنوان یک شبکه عصبی کانولوشن استفاده شده است که توسط مجموعه داده InceptionV3 از قبل آموزش داده شده است. نتایج بدست آمده از میزان صحت الگوریتم، کارایی قابل قبولی را نشان می دهد برای کارهای آینده می توان جهت بهبود عملکرد الگوریتم تا رسیدن به میزان صحت بهتر می توان کار کرد.

References

[1] Ghorbel, A., Amor, N. B., & Jallouli, M. (2019). A survey on different human-machine interactions used for controlling an electric wheelchair. *Procedia Computer Science*, *159*, 398-407. https://doi.org/10.1016/j.procs.2019.09.194

[2] Mazo, M., Rodríguez, F. J., Lázaro, J. L., Ureña, J., García, J. C., Santiso, E., & Revenga, P. A. (1995). Electronic control of a wheelchair guided by voice commands. *Control Engineering Practice*, *3*(5), 665-674. https://doi.org/10.1016/0967-0661(95)00042-S

[3] Tomari, M. R. M., Kobayashi, Y., & Kuno, Y. (2012). Development of Smart Wheelchair System for a User with Severe Motor Impairment. *Procedia Engineering*, *41*, 538-546. https://doi.org/10.1016/j.proeng.2012.07.209

[4] Kumar, D., Malhotra, R., & Sharma, S. R. (2020). Design and Construction of a Smart Wheelchair. *Procedia Computer Science*, *172*, 302-307. https://doi.org/10.1016/j.procs.2020.05.048

[5] Ruíz-Serrano, A., Posada-Gómez, R., Sibaja, A. M., Rodríguez, G. A., Gonzalez-Sanchez, B. E., & Sandoval-Gonzalez, O. O. (2013). Development of a Dual Control System Applied to a Smart Wheelchair, using Magnetic and Speech Control. *Procedia Technology*, *7*, 158-165. https://doi.org/10.1016/j.protcy.2013.04.020

[6] Scardapane, S., Scarpiniti, M., Bucciarelli, M., Colone, F., Mansueto, M. V., & Parisi, R. (2015). Microphone array based classification for security monitoring in unstructured environments. *AEU - International Journal of Electronics and Communications*, *69*(11), 1715-1723. https://doi.org/10.1016/j.aeue.2015.08.007

[7] Maccagno, A., Mastropietro, A., Mazziotta, U., Scarpiniti, M., Lee, Y.-C., & Uncini, A. (2021). A CNN Approach for Audio Classification in Construction Sites. In A. Esposito, M. Faundez-Zanuy, F. C. Morabito, & E. Pasero (Eds.), *Progresses in Artificial Intelligence and Neural Systems* (371-381). Springer Singapore. https://doi.org/10.1007/978-981-15-5093-5_33

[8] Wold, E., Blum, T., Keislar, D., & Wheaten, J. (1996). Content-based classification, search, and retrieval of audio. *IEEE MultiMedia*, *3*(3), 27-36. https://doi.org/10.1109/93.556537

[9] Akoushideh, A., Tourani, A., Shahbahrami, A., & 4, M. M. (2021). Design and Implementation of Automatic License Plate Recognition System for Security Gates. *Karafan*, *18*(3), 237-252. https://doi.org/10.48301/kssa.2021.130288

[10] Weninger, F., & Schuller, B. (2011, 22-27 May 2011). Audio recognition in the wild: Static and dynamic classification on a real-world database of animal vocalizations. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), https://opus.bibliothek.uni-augsburg.de/opus4/frontdoor/deliver/index/docId/72683/file/72683.pdf

[11] Ghiurcau, M. V., Rusu, C., Bilcu, R. C., & Astola, J. (2012). Audio based solutions for detecting intruders in wild areas. *Signal Process.*, *92*(3), 829–840. https://doi.org/10.1016/j.sigpro.2011.10.001

[12] Rabaoui, A., Davy, M., Rossignol, S., & Ellouze, N. (2008). Using One-Class SVMs and Wavelets for Audio Surveillance. *Trans. Info. For. Sec.*, *3*(4), 763–775. https://doi.org/10.1109/tifs.2008.2008216

[13] Xu, W., Zhang, X., Yao, L., Xue, W., & Wei, B. (2020). A multi-view CNN-based acoustic classification system for automatic animal species identification. *Ad Hoc Networks*, *102*, 102115. https://doi.org/10.1016/j.adhoc.2020.102115

[14] Deperlioglu, O. (2021). Heart sound classification with signal instant energy and stacked autoencoder network. *Biomedical Signal Processing and Control*, *64*, 102211. https://doi.org/10.1016/j.bspc.2020.102211

[15] Mahmoudian, S., Aminrasouli, N., Ahmadi, Z. Z., Lenarz, T., & Farhadi, M. (2019). Acoustic Analysis of Crying Signal in Infants with Disabling Hearing Impairment. *Journal of Voice*, *33*(6), 946.e947-946.e913. https://doi.org/10.1016/j.jvoice.2018.05.016

[16] Messner, E., Fediuk, M., Swatek, P., Scheidl, S., Smolle-Jüttner, F.-M., Olschewski, H., & Pernkopf, F. (2020). Multi-channel lung sound classification with convolutional recurrent neural networks. *Computers in Biology and Medicine*, *122*, 103831. https://doi.org/10.1016/j.compbiomed.2020.103831

[17] Hoseini, F., Sepehrzadeh, H., & 2, A. T. (2024). MRI Segmentation Using Inception-based U-Net Architecture and Up Skip Connections. *Karafan*, *21*(1), 63-88. https://doi.org/10.48301/kssa.2023.394044.2530

[18] Benhari, M., & Hosseini, R. (2024). An Intelligent Ensemble Model of Uncertainty Management in Belief Network for the Classification of Microscopic Images to Detect Cervical Cancer. *Karafan*, *21*(1), 89-69. https://doi.org/10.48301/kssa.2023.404913.2625

[19] Tsalera, E., Papadakis, A., & Samarakou, M. (2021). Comparison of pre-trained CNNs for audio classification using transfer learning. *Journal of Sensor and Actuator Networks*, *10*(4), 72. https://doi.org/10.3390/jsan10040072

[20] Dong, X., Yin, B., Cong, Y., Du, Z., & Huang, X. (2020). Environment sound event classification with a two-stream convolutional neural network. *IEEE Access*, *8*, 125714-125721. https://doi.org/10.1109/ACCESS.2020.3007906

[21] Bahle, G., Fortes Rey, V., Bian, S., Bello, H., & Lukowicz, P. (2021). Using privacy respecting sound analysis to improve bluetooth based proximity detection for COVID-19 exposure tracing and social distancing. *Sensors*, *21*(16), 5604. https://doi.org/10.3390/s21165604

[22] Abeysinghe, A., Tohmuang, S., Davy, J. L., & Fard, M. (2023). Data augmentation on convolutional neural networks to classify mechanical noise. *Applied Acoustics*, *203*, 109209. https://doi.org/10.1016/j.apacoust.2023.109209

[23] Zaman, K., & Direkoğlu, C. (2020). Classification of Harmful Noise Signals for Hearing Aid Applications using Spectrogram Images and Convolutional Neural Networks. 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), https://doi.org/10.1109/ISMSIT50672.2020.9254451

[24] Ballesteros, D. M., Rodriguez-Ortega, Y., Renza, D., & Arce, G. (2021). Deep4SNet: deep learning for fake speech classification. *Expert Systems with Applications*, *184*, 115465. https://doi.org/10.1016/j.eswa.2021.115465

[25] Vrebčević, N., Mijić, I., & Petrinović, D. (2019). Emotion classification based on convolutional neural network using speech data. 2019 42nd international convention on information and communication technology, electronics and microelectronics (MIPRO), https://doi.org/10.21437/interspeech.2019-184110.23919/Eusipco47968.2020.928780210.1109/CICT56698.2022.999796110.23919/MIPRO.2019.8756867

[26] Si, S., Wang, J., Sun, H., Wu, J., Zhang, C., Qu, X., Cheng, N., Chen, L., & Xiao, J. (2021). Variational information bottleneck for effective low-resource audio classification. *arXiv preprint arXiv:2107.04803*.https://arxiv.org/pdf/2107.04803

[27] Pham, L. D., McLoughlin, I., Phan, H., & Palaniappan, R. (2019). A Robust Framework for Acoustic Scene Classification. INTERSPEECH, https://doi.org/10.21437/interspeech.2019-1841

[28] Jena, K. K., Bhoi, S. K., Mohapatra, S., & Bakshi, S. (2023). A hybrid deep learning approach for classification of music genres using wavelet and spectrogram analysis. *Neural Computing and Applications*, *35*(15), 11223-11248. https://doi.org/10.1007/s00521-023-08294-6

[29] Scarpiniti, M., Comminiello, D., Uncini, A., & Lee, Y.-C. (2021). Deep recurrent neural networks for audio classification in construction sites. 2020 28th European Signal Processing Conference (EUSIPCO), https://doi.org/10.21437/interspeech.2019-184110.23919/Eusipco47968.2020.9287802

[30] Yu, Y., Luo, S., Liu, S., Qiao, H., Liu, Y., & Feng, L. (2020). Deep attention based music genre classification. *Neurocomputing*, *372*, 84-91. https://doi.org/10.1016/j.neucom.2019.09.054

[31] Srivastava, N., Ruhil, S., & Kaushal, G. (2022). Music genre classification using convolutional recurrent neural networks. 2022 IEEE 6th Conference on Information and Communication Technology (CICT), https://doi.org/10.21437/interspeech.2019-184110.23919/Eusipco47968.2020.928780210.1109/CICT56698.2022.9997961

[32] Nigro, M., Rueda, A., & Krishnan, S. (2022). Acoustic Scene Classification Using Time–Frequency Energy Emphasis and Convolutional Recurrent Neural Networks. Artificial Intelligence and Evolutionary Computations in Engineering Systems: Computational Algorithm for AI Technology, Proceedings of ICAIECES 2020, https://link.springer.com/chapter/10.1007/978-981-16-2674-6_21

[33] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015). *Rethinking the Inception Architecture for Computer Vision* IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA. http://arxiv.org/abs/1512.00567

[34] Dong, N., Zhao, L., Wu, C. H., & Chang, J. F. (2020). Inception v3 based cervical cell classification combined with artificially extracted features. *Applied Soft Computing*, *93*, 106311. https://doi.org/10.1016/j.asoc.2020.106311

[35] Khamparia, A., Gupta, D., Nguyen, N. G., Khanna, A., Pandey, B., & Tiwari, P. (2019). Sound Classification Using Convolutional Neural Network and Tensor Deep Stacking Network. *IEEE Access*, *7*, 7717-7727. https://doi.org/10.1109/ACCESS.2018.2888882

[36] Altes, R. (1980). Detection, estimation, and classification with spectrograms. *Journal of the Acoustical Society of America*, *67*, 1232-1246. https://doi.org/10.1121/1.384165

[37] Hussein, W., Hussein, M., & Becker, T. (2012). Spectrogram Enhancement By Edge Detection Approach Applied To Bioacoustics Calls Classification. *International Journal of signal and image processing*, *3*, 1-20. https://doi.org/10.5121/sipij.2012.3201