Analysis of Machine Learning Algorithms for Violence Detection in Audio *

Bruno Veloso[1 [0000−0003−3243−1633]], Dalila Durães[1 [0000−0002−8313−7023]], and Paulo Novais[1 [0000−0002−3549−0754]]

ALGORITMI Centre, University of Minho, Braga, Portugal
a78352@alunos.uminho.pt, dalila.duraes@algoritmi.uminho.pt, pjon@di.uminho.pt

Abstract. Violence has always been part of humanity, however, there are different types of violence, with physical violence being the most recurrent in our daily lives. This type of violence increasingly affects many people's lives, so it is essential to try to combat violence. In recent years, human action recognition has been extensively studied, but mainly in video, an important computer vision area. Audio appears as a factor capable of circumventing these problems. Audio sensors can be omnidirectional, requiring less processing power and hardware and software performance when compared to the video. The audio can represent emotions. It is not affected by lighting or temperature problems, nor does it need to be at a favourable angle to capture the intended information. That said, audio is seen as the best way to recognize violence, applied with *Machine Learning/Deep Learning/Transfer Learning techniques*. In this paper we test a Convolutional Neural Network (CNN), a ResNet50, VGG16 and VGG19, in order to classify audios. Later we see that CNN obtains the best results, with a 92.44% accuracy in the test set. ResNet50 was the worst model used, obtaining an 86.34% accuracy. For the VGG models, both show a good potential but did not get better results than CNN.

Keywords: Audio Violence Detection, Deep Learning, Transfer Learning, Audio Action Recognition.

* Supported by organization ALGORITMI Centre.

1 Introduction

In 2020, 66408 cases were reported to the Portuguese Association for Victim Support (APAV- Associação Portuguesa de Apoio à Vítima), of which 31% correspond to "crimes and other forms of violence". Of these 31%, 94% represent acts of violence against people. [1]. The detection and recognition of violence have been areas of research interest, mainly in surveillance. The main objective of detecting and recognizing violence is to carry it out automatically and in real time, in order to be able to provide assistance to victims in a timely manner [2].

Violence has always been part of humanity, and it can be expressed in different ways. The fact that there are different ways of practicing violence, means that it has to be reduced in society. One of the types that is more present in society is domestic violence. Domestic violence is then recognized as a serious public health problem, which can not only cause physical harm to the victim, as well as mental harm to the victim [3].

When talking about violence detection, it is mainly associated with detecting violence through the video. However, capturing video requires great capacity and performance of hardware and software. Another method that can be used to violence detection is with the use of audio, as it can be identified and/or classified through *machine learning (ML)* [4]. Audio can be easily captured by microphones. These sensors are very powerful and can capture human behavior and emotions. Thus, a good representation of audio is critical to complement and prove the video classification [5].

Audio plays a critical role in understanding the environment around us, containing information that visual data cannot represent. Hence, its analysis is important, since, when analyzing the content of audio, it is possible to interpret the medium in which it was captured or its present situation. After analysing the importance of audio for understanding and environment, it is interesting that this can be used to build systems capable of automatically detecting and recognizing violence [6].

In this way, we find that automatic audio classification is a growing research area, with results that allow its application in real cases. The study and classification of audio can be very important for the resolution of several issues, namely in the detection of violence, where one can recognize whether the environment, in which a particular person is inserted, is in any way prone to violence or not. For that, we propose the use of *deep learning(DL)/transfer learning(TL)* to classify the audios in terms of violence or not.

The paper is organized as follows: next section presents the Literature Review with explanations of recognition og Human Actions, audio vs video, audio representation, and dataset. Section III described the dataset where it was explained the preprocessing, while Section IV, Experiments, describe the models of deep learning used and training details of the experiments. Section 5, present the results and the discussion. Finally, section VI concludes this work with some future directions.

2 Literature Review

This literature review begins with the explanations of recognition of Human Actions and the difference between action predictions and action recognition. Then some difference between audio and video is described. Following it analyses the audio representation as well as the different methods. Finally, it presents some existing public datasets.

2.1 Recognition of Human Actions

In the last decade, the analysis of human movement and recognition of actions have been extensively studied by researchers [2–7].

All human actions are done with some purpose. For example, to complete a physical exercise, the person interacts and responds with the environment using legs, arms, hands, etc. One of the biggest goals of artificial intelligence is to build a machine capable of understanding actions and human interactions.

As technological advances increase, it is becoming possible to develop machines capable of understanding human actions. There are two cores topic, which are: action prediction, trying to predict the human action using data that do not correspond to the totality of the action; and action recognition, which tries to recognize/classify the action using data regarding the total execution of the action [7].

Action prediction focuses on the future state. Often, machines cannot wait for the execution of total action before acting, so it's important that they are able to predic whether we will be facing a risky action [7], using the data collected so far, that could put people'se lives at risk. For example, trying to predict that a robbery will occur, in order to prevent or contain it as soon as possible, as can be seen in figure 1b, the passenger, dressed in black, extends his hand around the driver, and not having the complete action, you can only try to predict what will happen next, in this case it was an assault on the driver.

Action recognition attemps to identify human action based on data that represent the action in its entirety. One of the biggest problems with recognizing human actions is the detection of violence. As the Figure 1a illustrates, the previously mentioned scene of the robbery is already complete, so we can try to recognize what the action is, which in this case this recognition would say that it was a robbery.

2.2 Audio vs Video

The signal produced by the sound of an audio contains information that visual data cannot represent [6].

Audio sensors (microphones) have very interesting particularities, to which video sensors (cameras) cannot compete. When detecting violence through audio, the need for bandwidth, storage, and computational resources are much lower wthen when compared to video [12]. This is due to the fact that audio is one-dimensional (time), unlike video which is three-dimensional (width x height

(a) Action recognition. (b) Action prediction.

Fig. 1: Classification of human actions, where a) represents an action recognition, and b) represents a prediction of an action. All rights belong to *DarrenLevyOfficial*, images taken from https://www.youtube.com/watch?v=BB5Y0j8RLE4.

x time), and this allows for a greater number of audio sensors (since its cheaper per unit) and also having a more complex signal processing due to requiring less computing resources. Cameras have a limited angular field of view, while the microphones are omnidirectional, this allowing a spherical field of view. Audio sensors do not have problems with lighting and temperatures, so the audio to be processed is not affected. A video, by itself alone, cannot represent information such as screams, explosions, words of abuse, or emotions [5]. Audio event acquisition is better because the audio wave length is longer and many surfaces allow reflections of acoustic waves, so obstacle in the way can be bypassed [12].

However, audio also has its problems, and sometimes these are difficult to avoid. Some of these problems are: there may be an overlap of several audios, for example a song that is in the background, which can affect the classification of the audio, when there is multipath propagation that results in an echo, and if the microphone is far from the audio we want to capture, it makes it harder to understand what kind of environment this one was captured [5].

The main problem, regarding any type of data capture, is due to privacy. The capture of images or audio raises very important ethical issues, which can lead to some debates about whether it is correct or not, but, as there is no other form of surveillance, this was put aside for the continuation of the study and it will be used audio.

2.3 Audio Representation

The audio can be represented, so that it can be interpreted by the human being. The main ideia is in taking the audio signal and converting it into a visual image. These images, generated from the audio, can then be used to extract features from them, either by hand, or fed directly to a DL/TL classifier, as there are classifiers that can learn and extract features [13].

There are some methods tha can be used to create this images (spectrograms), that represent the audio, and some are: *Short-Time Fourier Transform*, *Chromagram*, *Mel-Spectrogram* [12]. As we will only be using mel-spectrograms, this will be the only to be explained.

A Mel-spectrogram is a spectrogram whose y-axis has been applied a mel scale. For this to be obtained, some steps have to be fulfilled. These are [14]:

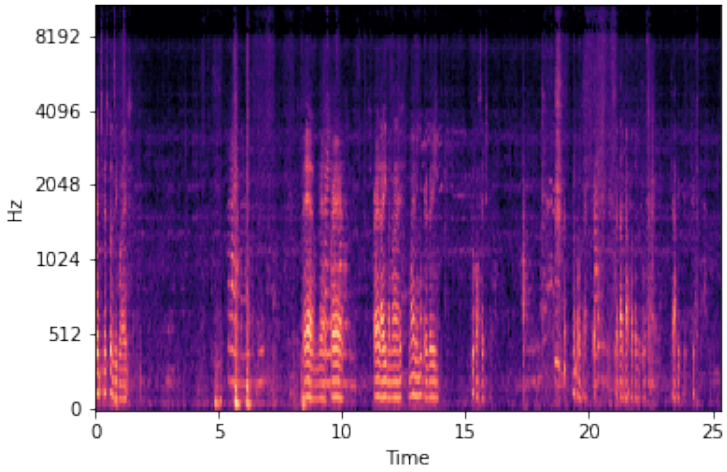

Fig. 2: Representation of a *mel-spectrogram*.

1. Divide the audio into fixed-size windows, with a smaller hop size between windows than division size.
2. For each window, apply *Fast Fourier Transform* to move from the time domain to the frequency domain.
3. Take the frequency spectrum, originated in the previous step, and apply *mel scale*.
4. For each window, decompose the magnitude of the signal and its components, corresponding to the frequencies from the mel scale.

The Mel-spectrogram, over time (x-axis), shows successive frequencies (y-axis), as well as the different amplitudes (colors, measured in dB) for each instant, as can be seen in the Figure 2.

2.4 Public Datasets

To find a dataset with audios involving violence or not, is really hard. But there are some datasets that are worth mentioned.

The Real Life Violence Situations dataset (RLVS) has 2000 videos. 1000 videos are classified as violence videos and the other 1000 are classified as non-violence videos. These videos were extracted from *YouTube*. The violent videos are extracted from many environments like prisons, streets, schools, etc., and the non violent videos represent different human actions like sports, eating, walking, etc. This dataset includes a wide variety of race, gender and age. Some of the videos, that the dataset contains, have no sound, and resolution can go from 480p to 720p (indoor or outdoor environments) [9].

NTU CCTV-Fights dataset has 1000 videos, some without sound, taken from *YouTube*. The actions in this dataset, goes from pushing, to kicking, fighting, pulling, among others.The dataset is divided into CCTV, which are videos captured by surveillance cameras, and NON-CCTV, which are videos capture by dash-cams, cell phone cameras, drones and helicopters. The CCTV group consists of 280 files, which include different types of fights, ranging from 5 seconds to 12 minutes, 8.54 hours in total. The NON-CCTV group has 720 videos, ranging from 3 seconds to 7 minutes, giving a total of 9.13 hours of videos [10].

XD-Violence dataset, contains 4754 videos, not all of them contain audio, these videos are divided into two categories, violence with 2405 videos and 2349 without violence, giving a total of 217 hours in videos. Videos marked with violence can be further distinguished between six different types of violence, which are: abuse, car accident, explosion, fight, riot and shooting. Each video of violence can also contain between 1 and 3 labels inclusive, and the order of the labels corresponds to the importance they have for the different events during the video. The videos in this dataset are clips from movies, cartoons, video games, news, sports, music, fitness, live scenes (captured by surveillance cameras, people recording with cell phone, etc.), etc [11].

3 Dataset

The dataset is required in order to evaluate the models implemented, but no dataset was found that contained the specific restrictions. So, a group of researchers created their own dataset, making all the recording of all the scences of violence, and non violence, inside a car with people who are not actors and during the pandemic. The dataset consist in videos, all of them have sound, and they represent 20 different scenarios. From that 20 different scenarios, 12 of them have violence included and the other 8 don't. Each scenario was recorded with 16 different pairs of actors. Some scenes contain the use of objects aswell.

3.1 Preprocessing the Dataset

The dataset had 795 violence videos and 494 non-violence videos. The next step was to go through each of these files and convert them to mp3 files, once that it is only necessary the audio of them. For this, a python script was created that traverses the folders and, with the use of a python library called *moviepy*[1], converts mp4 and MOV files to mp3. Due to the existence of very large non-violent audio files, it was decided that, for all those that had more than 40 seconds, they would be divided in half, using a python library called *pydub*[2], thus creating a new entry in the dataset. As for the audios of violence, these were generally longer than non-violence, and had the problem that, for the most part, the first 10 to 25 seconds did not contain violence. So, the solution was

[1] Biblioteca *moviepy* https://github.com/Zulko/moviepy
[2] Biblioteca *pydub* https://github.com/jiaaro/pydub

to analyze audio by audio and see when violence started, and using the *pydub* library it was possible to split the audio into two audios, in which the first corresponds to non-violence and the second to violence. There were also some audios that did not contain any content, this could be because they accidentally recorded it or because they didn't know it was recording, and these were all removed, thus giving a total of 1175 non-violence and 755 violence audios. After all this process, the non-violence audios were analyzed again to see which ones could be removed in order to balance the dataset. The result was a dataset with 860 non-violence audios and 755 violence audios. The next step was to go to the RLVS dataset and see the violence audios that could be inserted to balance the dataset. We got 105 violent videos from the RLVS dataset, and these were converted to audio and added to the final dataset, giving us a total of 860 non-violence audios and 860 violence audios. The last thing to do was to loop through all the audios and create a mel spectrogram of each audio, using the python library called *librosa*[3], to be fed into the deep learning models. Finally, the dataset was then divided into 80% for training and 20% for testing. The training folder contains 1376 mel spectrograms, of which, 688 correspond to non violence mel spectrograms and 688 violence mel spectrograms. In the test folder there are a total of 344 mel spectrograms, and of these, 172 are non violence and 172 are violence.

4 Experiment

4.1 Deep Learning Models

In this subsection we present the four models that have been used, the reason is that they are some of the most used in the literature [12]. The models that we have applied are: Convolutional Neural Network (CNN), ResNet50, VGG16 e VGG19.

Convolutional Neural Network is a deep learning model that is focused on image classification. This neural network, with sufficient training, can learn features that are present in the images, and is able to capture spatial and temporal dependencies, in an image, by applying relevant filters. The architecture, of a CNN, was inspired by the connection of neurons in the human brain [15].

Residual Network (ResNet) was developed to resolve the vanishing gradient problem by skipping layers with identity functions. ResNet can have multiple network depths and the depth is followed by their name, for example the one we use is ResNet50 which means that it has depth 50 [8].

VGG, meaning Visual Geometry Group, is a deep CNN architecture with multiple layers. The deep refers to the number of layers they have. This is a transfer learning model, used for the classification of images, with 19.6 billion of parameters, and it was trained using the *ImageNet* dataset [16]. We used two implementations of VGG, which are VGG16 and VGG19, the first has 16

[3] Biblioteca *librosa* https://librosa.org/doc/latest/index.html

convolutional layers and the last one has 19 convolutional layers, and these convolutional layers are used to extract features from the images [16].

To prepare for the algorithms, the train part of the dataset was divided, randomly, into train and validation. So 20% of the train folder will be for validation, which means that we have 1102 for the training part and 274 images for the validation part. Each class has the same quantity of images.

4.2 Training Details

During the experiments, the CNN network was trained for 600 epochs, and ResNet50, VGG16 and VGG19 were trained for 35 epochs. As for CNN, the learning rate used was 0.01 and for ResNet50, VGG16 and VGG19, the learning rated applied was 0.001. All the models were trained on desktop with a NVIDIA GeForce GTX 1070 Ti GPU, 16Gb of RAM, and has an AMD Ryzen 5 2600 processor.

Each model has the same callbacks. *EarlyStopping* callback that is checking validation loss and has a patience of 10, *ModelCheckpoint* to save the weights and is checking validation loss aswell, and last *ReduceLROnPlateau* to reduce learning rate when a metric has stopped improving, checking validation loss with a patience of 20 and a factor of 0.1.

Table 1 shows all the training parameters applied.

Model	**Optimizer**	**Learning Rate**	**Epochs**
CNN	Adam	0.01	600
ResNet50	Adamax	0.001	35
VGG16	Adamax	0.001	35
VGG19	Adamax	0.001	35

Table 1: Training details for every trained model.

5 Results and Discussion

In Figure 3, for each model we present two graphs, the first one is for the accuracy curve, and the second one is to analyze the loss curve. In all the graphs, the orange line represents the train set, and blue line is the validation set. Analyzing the loss curve, can give us an idea of how the model is learning, and if the model is underfitting or overfitting.

As shown in Figure 3, the accuracy, for the train set and validation set, are close to each other in all models, except in the ResNet50 model. In ResNet50, validation set takes a much higher accuracy than the training set. As for the loss, it follows the same principle, where ResNet50 is the only model that have a much higher difference in training loss and validation loss. *EarlyStopping* callback, CNN only trained for 36 epochs.

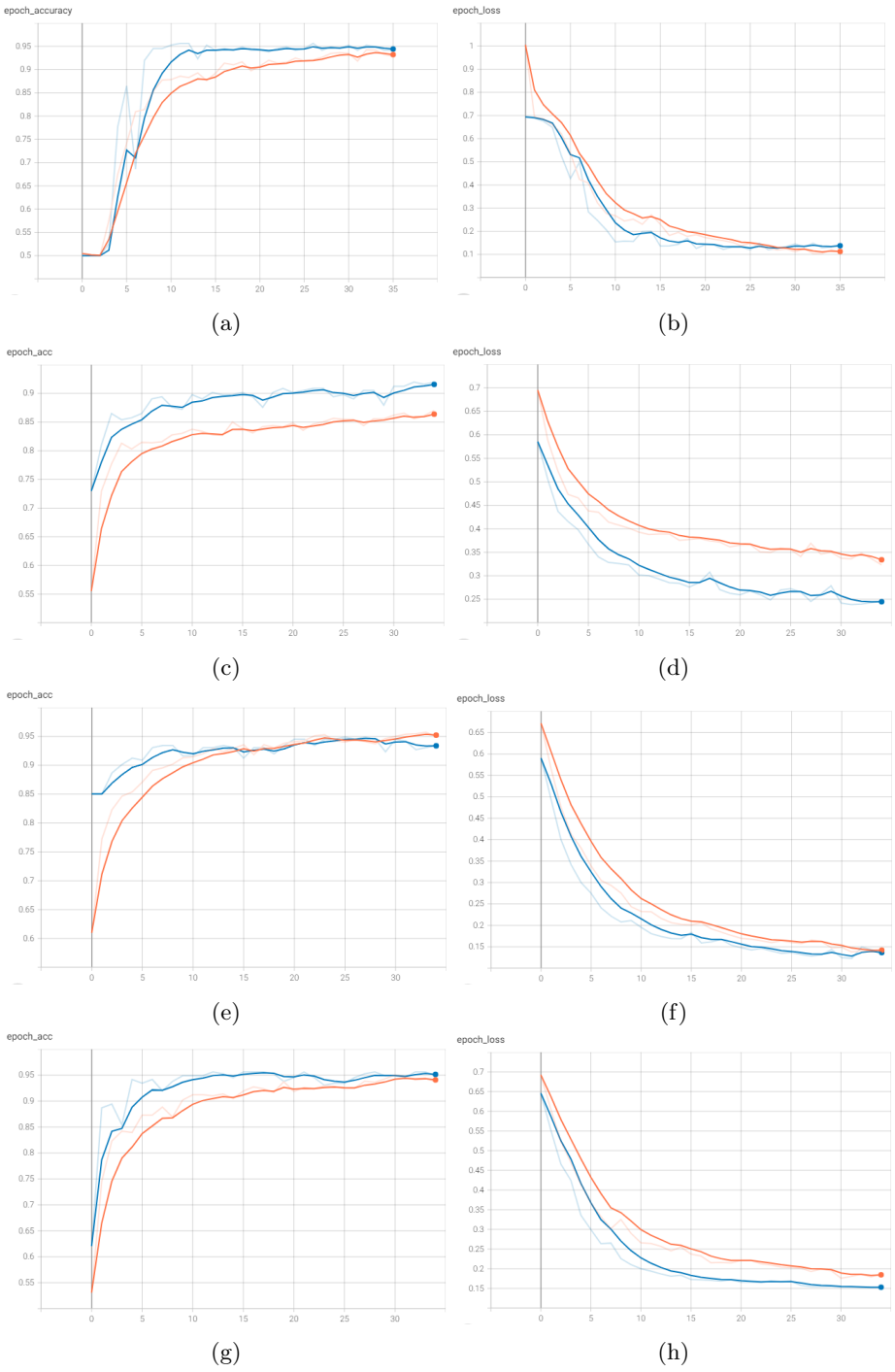

Fig. 3: Accuracy curve, on the left, and loss curve, on the right. Orange line is train set and blue line is validation set. a) and b) represent the use of CNN, c) and d) the use of ResNet, e) and f) model VGG16, and, g) and h) the VGG19.

The Figure 3 also shows that validation accuracy, in the CNN, almost did not improve during every epoch, and it also had a similar behaviour compared to VGG16. With that said, the validation loss, in CNN, did not decrease during every epoch.

Table 2 shows the best accuracy obtained for every model, in the train set, validation set and test set. The model that obtained best accuracy in the train set was VGG16, in validation set the best model was VGG9 and for the test set was the CNN.

This results show that CNN had the best results so far, and ResNet50 may be to much complex for the problem to solve. As for the VGG, both VGG16 and VGG19 show good potential.

Model	**Train**	**Validation**	**Test**
CNN	92.83	94.16	92.44
ResNet50	87.02	91.97	86.34
VGG16	95.01	93.43	90.41
VGG19	93.74	94.89	90.41

Table 2: Best accuracy results in every model.

6 Conclusion

We have created a dataset with twenty different scenarios, twelve of them have violence included, and the other eight don't. Each scenario was recorded with sixteen different pairs of actors. Some scenes contain the use of objects as well.

Following previous work in the literature, we have used the Mel-spectrogram method for represented the audio signal. Next, we have made an extensible experience using four different deep learning models to classify violence based on the audio signal. These models were CNN, ResNet50, VGG16, and VGG19.

The results show that CNN had the best results so far, and ResNet50 may be to much complex for the problem to solve. As for the VGG, both VGG16 and VGG19 show promising prospects.

In future work is necessary to apply this models to public dataset and compare the audio violence recognition results.

Acknowledgments

This work is supported by: FCT Fundação para a Ciência e Tecnologia within the RD Units Project Scope: UIDB/00319/2020.

References

1. APAV, "Estatisticas_APAV_Relatorio_Anual_2020.Pdf.", apav.pt/apav_v3/images/pdf/Estatisticas_APAV_Relatorio_Anual_2020.pdf, 2021 . Accessed on 22-10-2021.
2. Durães, Dalila, et al. "Comparison of Transfer Learning Behaviour in Violence Detection with Different Public Datasets." EPIA Conference on Artificial Intelligence. Springer, Cham, 2021.
3. Souto, Helton, Rafael Mello, and Ana Furtado. "An acoustic scene classification approach involving domestic violence using machine learning." Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional. SBC, 2019.
4. Sharma, Shivangi, Introduction to Audio Classification, https://www.analyticsvidhya.com/blog/2021/06/introduction-to-audio-classification/, 2021. Accessed on 22-10-2021.
5. Crocco, Marco, et al. "Audio surveillance: A systematic review." ACM Computing Surveys (CSUR) 48.4 (2016): 1-46.
6. Souto, Helton, Rafael Mello, and Ana Furtado. "An acoustic scene classification approach involving domestic violence using machine learning." Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional. SBC, 2019.
7. Kong, Yu, and Yun Fu. "Human action recognition and prediction: A survey." International Journal of Computer Vision (2022): 1-36.
8. Boesch, Gaudenz, Deep Residual Networks (ResNet, ResNet50) – Guide in 2022, https://viso.ai/deep-learning/resnet-residual-neural-network/, 2021. Accessed on 30-01-2022.
9. M. Soliman, M. Kamal, M. Nashed, Y. Mostafa, B. Chawky, D. Khattab, " Violence Recognition from Videos using Deep Learning Techniques", Proc. 9th International Conference on Intelligent Computing and Information Systems (ICICIS'19), Cairo, pp. 79-84, 2019
10. Rapid-Rich Object Search Lab, NTU CCTV-Fights Dataset, https://rose1.ntu.edu.sg/dataset/cctvFights/. Accessed on 08-01-2022.
11. Wu, Peng, et al. "Not only look, but also listen: Learning multimodal violence detection under weak supervision." European Conference on Computer Vision. Springer, Cham, 2020.
12. Santos, Flávio, et al. "In-Car Violence Detection Based on the Audio Signal." International Conference on Intelligent Data Engineering and Automated Learning. Springer, Cham, 2021.
13. Nanni, Loris, et al. "Ensemble of convolutional neural networks to improve animal audio classification." EURASIP Journal on Audio, Speech, and Music Processing 2020.1 (2020): 1-14.
14. Gartzman, Dalya, Getting to Know the Mel Spectrogram, https://towardsdatascience.com/getting-to-know-the-mel-spectrogram-31bca3e2d9d0, 2019. Accessed on 29-01-2022.
15. Saha, Sumit, A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way, https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53, 2018. Accessed on 30-01-2022.
16. Tiwari, Ravi, Transfer Learning — Part — 4.0!! VGG-16 and VGG-19, https://becominghuman.ai/transfer-learning-part-4-0-vgg-16-and-vgg-19-d7f0045032de, 2021. Accessed on 30-01-2022.