Available online www.ejaet.com

European Journal of Advances in Engineering and Technology, 2022, 9(12):32-39

Research Article ISSN: 2394 - 658X

Design of an Internet Data Access Filtering System based on Convolutional Neural Networks (EfficientNet)

Christine Laure MANANGA*, Felix PAUNE and Léandre NNEME NNEME

Laboratory of Computer Science Engineering and Automation, University of Douala,
P. O. Box 1872 Douala, Cameroon
*Corresponding author: cmananga35@gmail.com

ABSTRACT

This paper consists in proposing a system for filtering access to data on the Internet, in a "general public" context, via the design of a plugin with the objective of improving the processing and control of dynamic information disseminated in web content. The aim of this approach is to determine which resources are relevant to a user. We have studied and proposed a hybrid method between browser plugin technology (content-based filtering) and convolutional neural network technology (efficient net). The working principle of our algorithm will be the detection and classification of a so-called illegal image. To present good results, we used Web Scraping technology to design our training database; the training of our model was based on EfficientNetb7, the Spyder Anaconda development environment and the Python programming language. During testing, the CNN results we obtained were very encouraging as they allowed us to assess the achievement of our objectives at 95%.

Key words: Algorithm, filtering, web content, plugin, neural networks

INTRODUCTION

With the explosive growth of ICTs today, around 80% of users find themselves overwhelmed by a flood of often-unsolicited information. The protection of minors on the Internet is a major social, political and economic issue. The recent governmental decisions to make parental control mandatory and free for Internet Service Providers (ISPs) attest to this. It is therefore important at this stage to have reliable measures to improve and strengthen these various controls in order to reduce the spread of illegal images on the Internet. Filtering on the Internet means controlling a flow of information and presenting only that which is likely to be of interest to the user (relevant information) [1]. Internet filtering (sometimes called Internet blocking) is not a recent activity. However, the concept covers such a wide range of practices, hardware, software and services. The primary purpose of Internet filtering is to prevent illegal content from reaching a personal computer or workstation. This can be done using a software or hardware product that monitors all Internet communications and determines whether to prevent the receipt and/or display of specifically targeted content. It can be adapted and configured differently to meet the needs of different categories of users (parents, children, teachers, students, etc.). It would be a mistake to assume that all types of filtering are the same or equally effective, that they have the same legal consequences, or that a given system could easily be used to target more than one type of content.

How can it be ensured that once a user is connected to the Internet, he or she cannot access illegal and unauthorised information? The main issue in this topic is the overall policy of dynamic information processing and content analysis.

In the literature to date, several equally effective algorithms have been used to filter information on the Internet, including memory-based algorithms [2], model-based algorithms (K-means, RecTree, FRAC) [3]-[5], recommendation systems [6] and deep learning [7]. Each of these techniques for filtering information and web content has strengths, weaknesses and risks in terms of overblocking (or excessive filtering) and underblocking (or insufficient filtering) [8].

We therefore propose a reliable and efficient algorithm based on convolutional neural networks (EfficientNet) [9]-[11], which will be integrated into a web browser (plugin). It should be able to analyse any web content and classify it as illegal or not (violence, pornography).

DESIGN, MATERIAL AND METHODS

The database was designed using Web Scraping technology. It consists of 3407 images divided into 6 main categories:

- Dressed people ;
- Images with pornographic titles ;
- Knives ;
- Tanks ;
- Firearms (rifles, pistols, shotguns, etc.).

The database creation scheme is shown above:

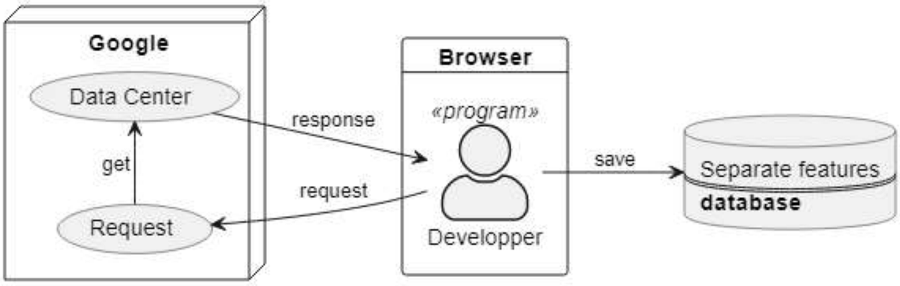

Fig. 1 Web scraping for data collection

The cleaning of our database was done manually. This phase consists of eliminating images of poor quality, which do not fit in with the researched theme. The images obtained are recorded in a database which will be subdivided into three, namely :

- Training database consisting of 2303 images;
- Test database of 742 images;
- Evaluation database consisting of 362 images;

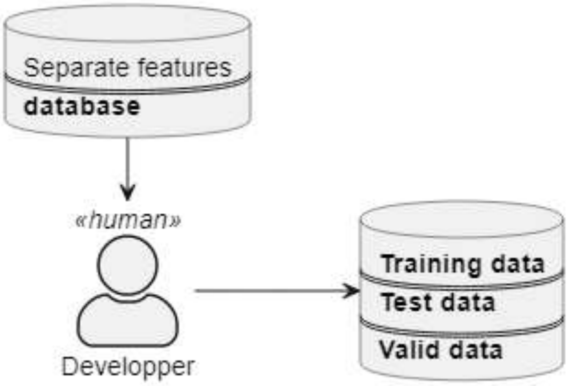

Fig. 2 Data cleaning of the database

The Google Colab development environment

The Google Colab environment is the tool that enabled us to design the learning model.

Hugginface

Deployment interface for open source artificial intelligence models. It is an API capable of being indexed by all programming platforms.

Web browser

It is the testing and development support for our extension.

Development methodology

EfficientNet

Based on the observation that better accuracy and efficiency can be achieved by imposing a balance between all dimensions of the network, the EfficientNet [12]-[14] has been proposed to improve the performance of CNNs by scaling in three dimensions, i.e. width, depth and resolution using a set of fixed scaling coefficients that meet certain specific constraints.

Transfer learning

For neural networks, it is essential to collect enough data for the training process [15]-[16]. Transfer learning allows the reuse of existing parameters, i.e. convolution weights of a model trained on large datasets to train new models with relatively fewer labelled images. We used pre-trained weights from ImageNet.

Learning the model

Our model was designed in two main phases:

- Data augmentation

This technique will allow us to increase the quality of our learning database and thus increase the performance of the learning model [17]. We have used:

- Changing the angle of view of the image ;
- Changing the Zoom of the image ;
- Changing the image dimensions (height and depth) ;

- Training the model

The training of our model is based on EfficientNetb7. Through object programming, we will instantiate it to have a Model object. The latter will be optimised by making its layers trainable except for the last 10 layers.

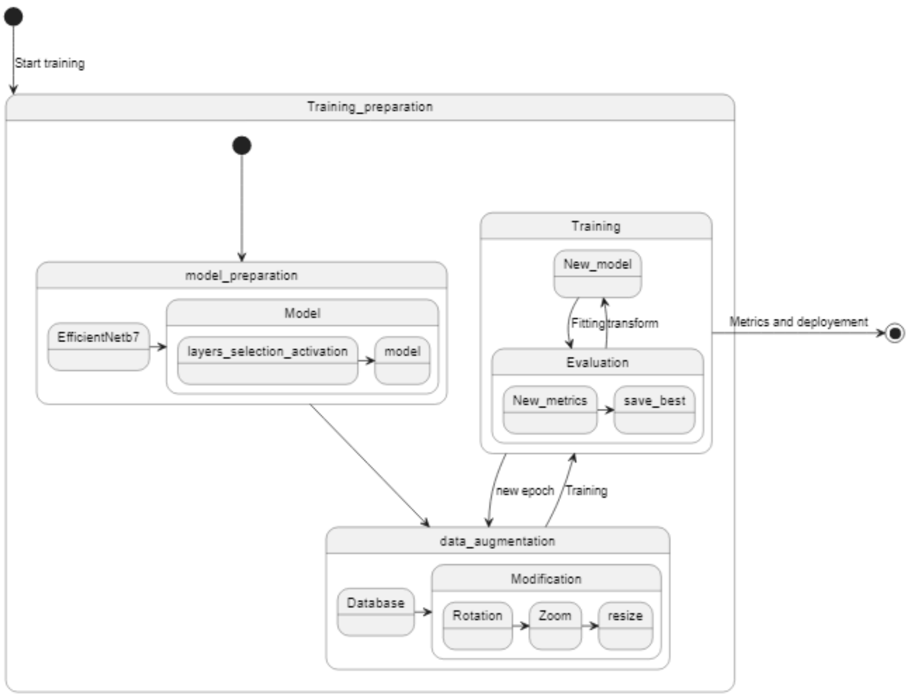

Fig. 3 Learning strategy of our model

RESULTS AND DISCUSSION

System structure

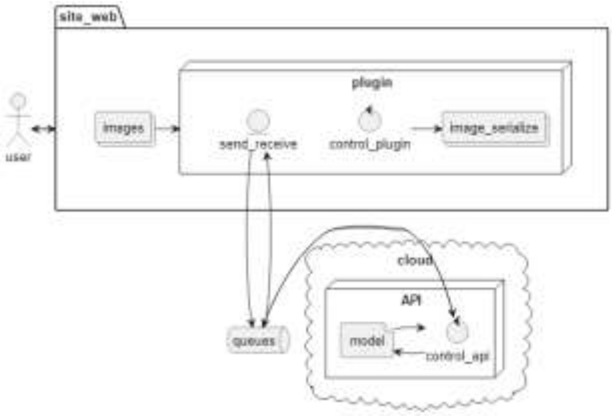

Fig. 4 Structure of the new data access filtering system

Calculation of performance metrics

The performance metrics allow us to have the quality of detection of the model with respect to our test database. They are represented by the figure below:

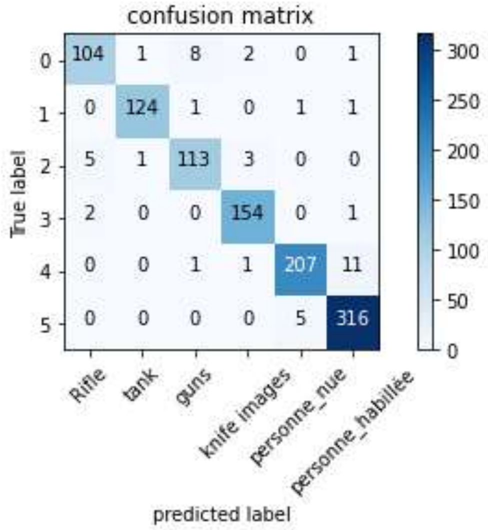

Fig. 5 Confusion matrix of the learning model

The data classification report by the Python Scikit-learn library is as follows:

	precision	recall	f1-score	support
Rifle	0.99	0.86	0.92	116
tank	0.88	0.98	0.93	123
guns	0.99	1.00	0.99	157
knife images	0.96	0.98	0.97	220
personne_nue	0.98	0.97	0.98	321
personne_habillée	0.98	0.99	0.99	127

Fig. 6 Classification ratio of the learning model

Results

- First experience :

The first experiment illustrates the influence of the number of filters needed to create adequate convolution maps. The experiment was performed on a base of 280 vectors.

Table -1 Recognition rates by different numbers of filters.

	First convolution layer	Second convolution layer	Recognition rate
Number of filters	**6**	**6**	97.5%
	6	**12**	**94.8%**
	12	**6**	**87.5%**

From the results obtained, we can say:

- In the feature extraction phase, the recognition rate is better if the number of filters in the first convolution layer is less than or equal to the number of filters in the second convolution layer.
- In the extreme case, if we have only 6 filters in the first layer and 6 or 12 filters in the second layer, the recognition rate reaches 95% and 97% for 280 vectors.

- **Second experiment**

In the second experiment, we increased the number of learning periods each time until it was higher than the set recognition rate.

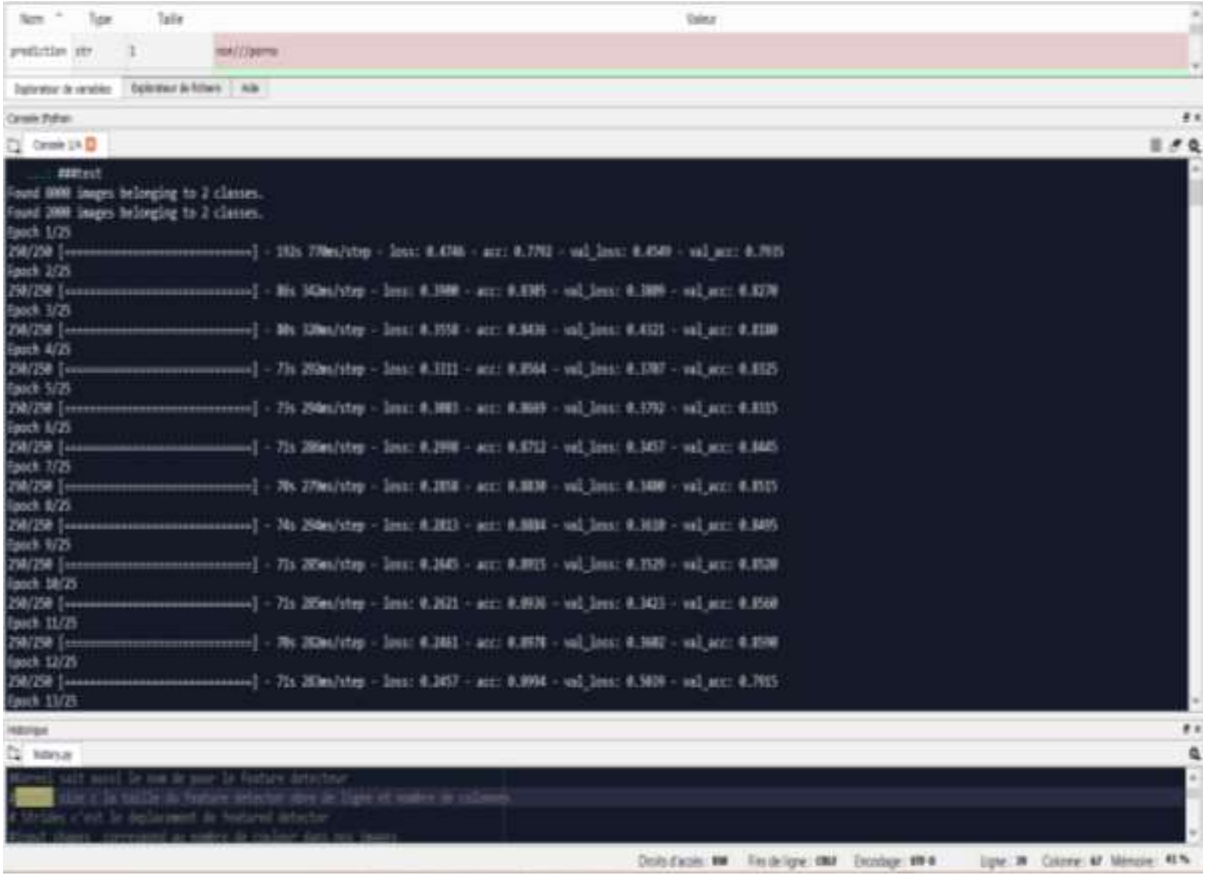

Fig. 7 Experimentation as a function of the number of epochs

Table -2 Recognition rates in relation to the number of periods

Number of eras	Number of images analysed	Recognition rate (acc)	Error rate (loss)	Running time
5	**250**	**86.69%**	**30.83%**	**73s 294ms/step**
10	**250**	**89.36%**	**26.21%**	**71s 285ms/step**
15	**250**	**91.29%**	**21.62%**	**76s 303ms/step**
20	**250**	**93.56%**	**16.52%**	**71s 283ms/step**
25	**250**	98.20%	7.61%	**76s 304ms/step**

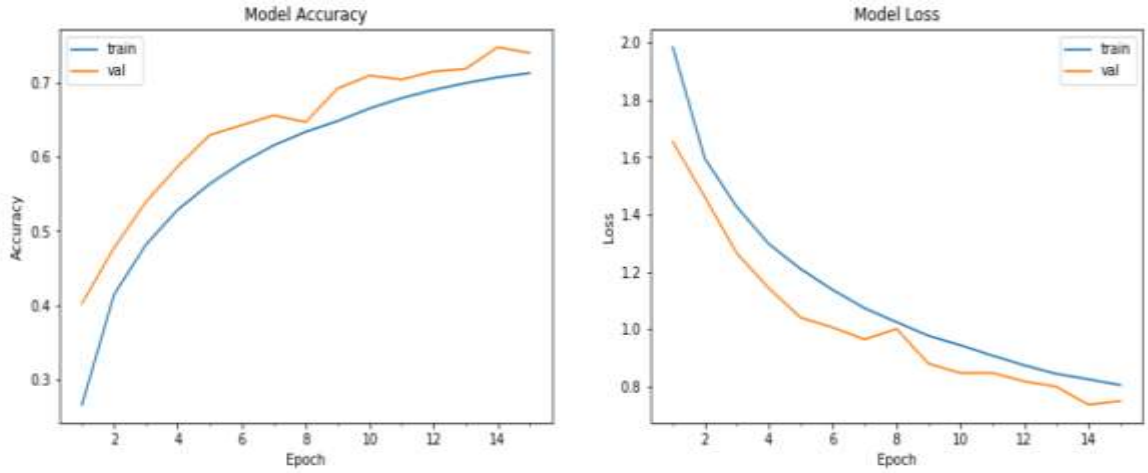

Fig. 8 Representation of the error and accuracy rate

The accuracy of learning and validation increases with the number of epochs, reflecting that with each epoch the model learns more information. If the accuracy is decreased then we will need more information to make our model learn and therefore we need to increase the number of epochs and vice versa. Similarly, the learning and validation error decreases with the number of epochs.

Once the plugin is installed in the browser, when the proposed filtering system detects an image as an illegal image, it automatically blocks the display of the image.

Implementation of the new filtering system

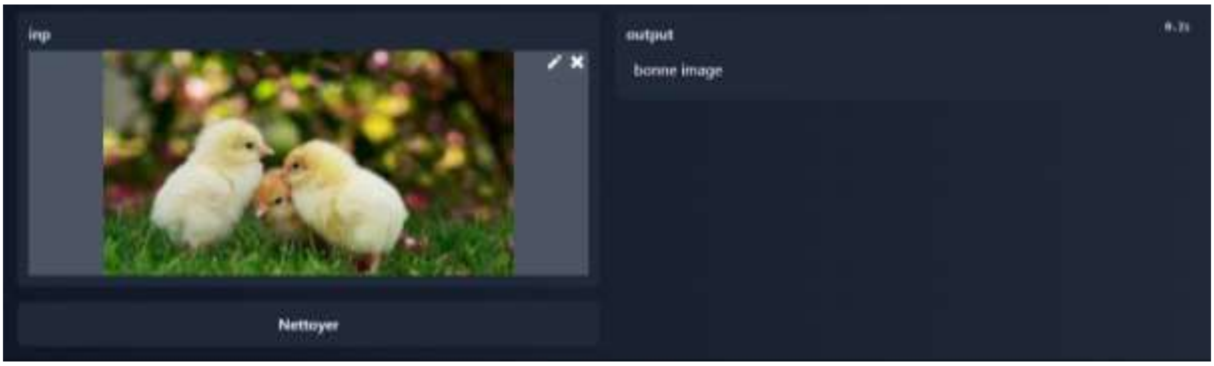

Fig. 9 Correct image detection results

Fig. 10 Illegal image detection results

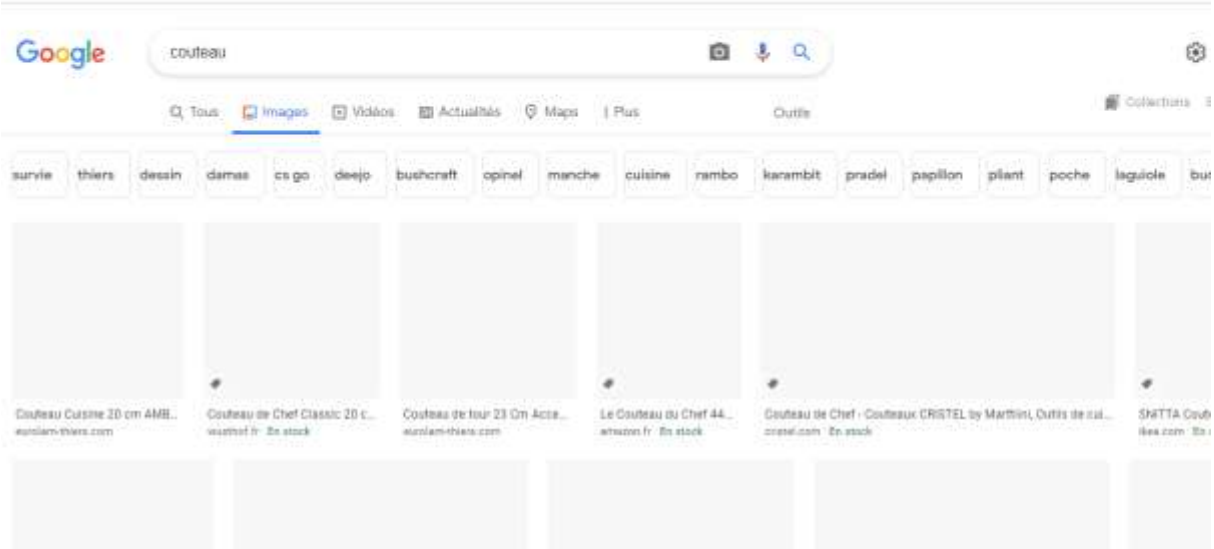

Fig. 11 Illegal image search results

CONCLUSION

In this paper, we have proposed the design of a web browser-based data access filtering system based on EfficientNet, a convolutional neural network technology. The main objective of this system is to control the spread of illegal images and improve the filtering of information on the Internet for better parental control and user protection. The added value of our approach is that it is mixed and located between the content-based filtering method and convolutional neural networks. The designed algorithm can be applied to the recognition and prediction of all categories of images; it will only be necessary to fill in the Training_set and the Test_set in order to supply the training base with images. As a result of the results obtained, our system presents a positive rate of illicit image recognition of 98% for a processing time of 76s.

REFERENCES

[1]. M. Belkheir, "Le filtrage de 1 Information sur Internet," DEA SIC synthesis note, ENSSIB, Lyon, 1995.
[2]. P. Lassalle. "Study of the scaling of segmentation and classification algorithms in remote sensing for the processing of massive volumes of data," PhD thesis, University of Toulouse, Nov.2015.
[3]. A. Belloui, "L'usage des concepts du web sémantique dans le filtrage d'information collaboratif", Master's thesis, National Institute of Computer Science, Algiers, 2008.
[4]. J. Leskovec, A. Rajaraman, J. Ullman, "Mining of Massive Datasets", 2nd Edition, 2014, pp 307-340.
[5]. F. Ricci, L. Rokach, B. Shapira and P. B. Kantor. "Recommender Systems Handbook," in proc. Springer editors, 2011.
[6]. E. NEGRE. "Bypassing web filtering using an SSH tunnel," internet: https://interstices.info/les-systemes-de-recommandation-categorisation/, Sept. 2018 [Jan. 20, 2022].
[7]. H. Michel, "Qu'est Ce Que Le Deep Learning (Apprentissage Profond)," Internet: https://ledatascientist.com/deep-learning-apprentissage-profond/, Jan. 13, 2019 [Sept. 20, 2020].
[8]. K. Djamila, "Using neural networks as a tool for datamining: Generating a behavioural model of a physical process from data," Master's thesis, University of Batna 2, 2012.
[9]. M. Zakaria, "Image classification with convolutional neural networks," Master's thesis, Abou Bakr Belkaid Tlemcen University, Algiers, 2017.
[10]. N. Srivastava, C. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from overfitting," Journal of Machine Learning Research, vol. 15, 2014, pp 1929-1958.
[11]. Mehdi Yedroudj, Marc Chaumont, and Frédéric Comby. How to Augment a Small Learning Set for Improving the Performances of a CNN-Based Steganalyzer? In Proceedings of Media Watermarking, Security, and Forensics, MWSF'2018, Part of IS&T International Symposium on Electronic Imaging, EI'2018, page 7, Burlingame, California, USA, 28 January - 2 February 2018.
[12]. A. G. HOWARD, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications " Internet : https://ledatascientist.com/deep-learning-apprentissage deep/ , Jan. 13, 2019 [Sept. 20, 2020].
[13]. Mehdi Yedroudj, Marc Chaumont, Frederic Comby, Ahmed Oulad Amara, and Patrick Bas. Pixels-off: Data-Augmentation Complementary Solution for DeepLearning Steganalysis. In Proceedings of the 2020 ACM Workshop on Information Hiding and Multimedia Security, IHMSec '20, page 39-48, Virtual Conference due to Covid (Formerly Denver, CO, USA), June 2020.

[14]. Han CAI, Ji LIN, Song HAN, "Efficient Method for Deeplearning", Computer Vision and Pattern Recognition, 2022, Pages 159-190.

[15]. Kamilaris ANDREAS, Prenafeta-BOLDU, Francesc XAVIER and Augusti FONTS, "Deeplearning in agriculture", Magazine article, 2018, pp 70-90.

[16]. Sara MEFTAH, Nasredine Semmar, Youssef Tamaazoust, Hassane Essafi, "Exploration Transfer learning for social network text analysis", Conference on Natural Language Processing - TALN-RECITAL 2019, 2019.

[17]. Tom KELDENICH, "Data Augmentation, rapidly improving your Deep Learning", Internet: https://inside-machinelearning.com/data-augmentation-ameliorer quickly-his-modele-of-deep-learning//, Feb. 22, 2021 [Jul. 21, 2022].