ICIC Express Letters
Volume **18**, Number **1**, January **2024**
 ISSN 1881-803X
pp. 1–8

PERFORMANCE ANALYSIS OF IMAGE DETECTION SYSTEM FOR SENSITIVE CONTENT WITH PERSONAL IDENTIFIABLE INFORMATION (PII) USING CONVOLUTIONAL NEURAL NETWORK

Alfian Ilarizky and Riri Fitri Sari*

Faculty of Electrical Engineering
University of Indonesia
Depok, Jawa Barat 16424, Indonesia
alfian.ilarizky@ui.ac.id; *Corresponding author: riri@eng.ui.ac.id

Received January 2023; accepted April 2023

Abstract. *Social media is one of the platforms most in demand by the public. However, social media also has risks that significantly impact its users. One of the risks of using social media is privacy concerns. Users tend to like sharing their personal information on social media platforms, especially information in the form of images. The mechanism for detecting images with sensitive content with Personal Identifiable Information (PII) can be used to address privacy issues on social networks. This research developed a content detection system to detect images with sensitive content with PII using VGG-16 based on the CNN architecture. In this study, VGG-16 was modified by changing the fully connected layer, which was trained using several transfer learning scenarios. The transfer learning scenario used in this study compares several freezing points of the VGG-16 architecture. The experimental results show that freezing after the fourth convolutional block gets a better accuracy value than the other scenarios. By applying modification and freezing of VGG-16 after the fourth convolution block, the Recall, Specificity, Precision, Accuracy and F1 Score values were obtained as 0.992, 0.967, 0.967, 0.979 and 0.979. This indicates that Modified VGG-16 frozen after the fourth convolutional block effectively detects PII-sensitive content images.*
Keywords: Social media, Detection, Image, Sensitive content, Personal identifiable information, CNN, VGG-16, Transfer learning

1. **Introduction.** Advances in technology platforms have changed business processes, government regulations, and relationships between individuals. One of the significant impacts of technological development is the ease of sharing information. One of the platforms for sharing information that is popularly used by people today is social media. Social media is a platform that can connect users with other parties, such as family, and friends [1]. Social media today is very easy to use. The increasing popularity of social media has attracted many people in various daily activities. This results in a large amount of data created by users [2]. Nowadays, people are very attached to social media [3]. With all the comfort and convenience it provides, social media has risks that have a significant impact on its users, one of which is privacy issues [4].

People tend to have a desire to express themselves and share personal information through social media [5,6]. Information stored on social media can be in the form of general information, as well as personal information and not worthy of being known by others. Privacy issues on social media are not just about what users share about themselves but also about what other people share about the privacy of another user [7].

One piece of information that is often shared by social media users is images containing Personal Identifiable Information (PII). The shared PII can contain direct information, as

DOI: 10.24507/icicel.18.01.1

well as indirect information [8]. PII is information that can be used to track or identify an individual [9]. In some cases, social networking platforms such as Instagram and Twitter do not implement checks on the personal information shared by their users, especially for image content. As a result, a lot of information can be used by unauthorized parties to support crimes. The more information that is shared, the more likely it is that someone can impersonate the user or encourage other crimes.

To overcome this problem, research is needed on content detection mechanisms that will be uploaded on social media. Sensitive content detection mechanisms are very important to minimize the risk of spreading personal data of social media users. In recent years, there have been many studies on sensitive content detection mechanisms on social media [10-13]. From the literature study that has been done, research on this problem always focuses on sensitive text detection mechanisms, but visual or image content detection has not been done.

Deep learning is one of the architectures commonly used in visual or image content detection [14-21]. One architecture that is often used for image detection is VGG-16. VGG-16 uses convolutional neural network architecture in performing object classification. This study uses VGG-16 architecture to detect images with sensitive content using PII. The dataset used in this study is a dataset obtained through the collection of image data with sensitive content with PII. By utilizing VGG-16, it is hoped that the detection process of images with sensitive content using PII can be carried out. Social media users can use the results of this detection before publishing their content. It is intended that social network users be aware of the importance of the confidentiality of personal information. These problems motivated us to use the VGG-16 model to detect images with sensitive content with PII. This paper proposes that VGG-16 is implemented in image detection system-sensitive content with PII. The main contributions of this paper are as follows:

- Creating an image detection system with sensitive content with PII by utilizing a privacy agent;
- Modification of the VGG-16 architecture on the fully connected layer by adding an average pooling layer and a dropout layer;
- Analysis of the use of transfer learning techniques by comparing freezing points on the third, fourth, and fifth convolution blocks and combining them with the modification of VGG-16.

The remaining structure of this paper is as follows. Transfer learning concepts and VGG-16 are discussed in Section 2. Section 3 describes the proposed system for performing content detection on social media that utilizes convolutional neural networks and transfer learning. The experimental results and explanations are presented in Section 4. The conclusions of this study are presented in Section 5.

2. **Background.**

2.1. **Transfer learning.**

Transfer learning provides increased learning in new tasks through the transfer of knowledge from related tasks that are ready to be learned [24]. In the transfer learning process, the learning procedure does not start from the beginning but the learning outcomes that have previously been used to do other tasks [25]. The transfer learning process involves two things, including using previously collected knowledge, and neglect is a must to do learning from the beginning [15]. This makes the process required to run the model shorter and faster [26]. Transfer learning is used by freezing a model layer that has been previously trained using another dataset. As a result, the layer does not need to be retrained using the new dataset. Subsequently, the unfrozen model layers are retrained using the new dataset. The consequence of this is that the number of trainable parameters is small compared to the total model parameters.

2.2. **VGG-16.** VGG-16 is a CNN architecture created by Bagaskara and Suryanegara [22]. VGG-16 comprises 13 convolutional layers and three fully connected layers [23]. The convolutional layer in the VGG architecture is divided into five blocks. Each block performs a convolutional function with a size of 3×3 pixels. Each block is closed with a max pooling function. Then the output of the fifth block is fully connected using flatten, twice the dense 4096 functions, and closed with dense 1000. The final layer of VGG-16 uses a dense 1000 layer because the developer of this architecture utilizes the ILSVRC dataset. The ILSVRC dataset has 1000 classes. The activation function is used to close the fully connected layer using softmax. The VGG-16 architecture uses an input size of 224×224 pixels [24].

3. **Proposed Technique.** The research methodology used in this study uses the Design Science Research Methodology (DSRM) [25]. In general, this methodology has 6 steps that need to be carried out, namely problem identification and motivation, determining the goals of the solution, design, and development, demonstration, evaluation, and communication. The first step and the second step have been carried out in Section 1. The next step is to do the design and development. Based on the problems identified in Section 1, artifacts are designed and developed to overcome these problems. The system design proposed in this study is shown in Figure 1. The system uses a privacy agent that runs behind social media. This privacy agent will process each uploaded image and classify it into two, namely images that contain sensitive PII content and images that do not contain sensitive content. If the image contains sensitive content, the user will be notified that the uploaded image contains sensitive PII content and will be destroyed. The image will be displayed on social media if it does not contain sensitive content.

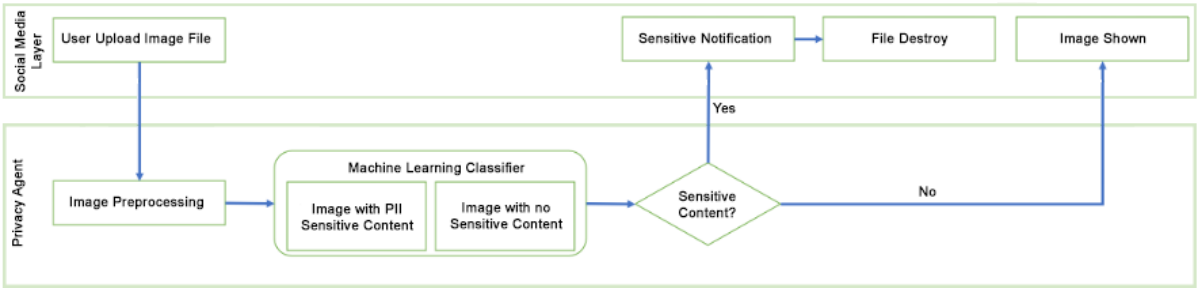

FIGURE 1. System design block diagram

The design of this system adapts the PII confidentiality protection guidelines issued by NIST [26] with the following adjustments:

- The system provides awareness and education to social media users by providing notifications that users upload images with sensitive content on social media.
- The system destroys images detected as having sensitive content. These images are not stored in the system to minimize the risk of social media providers to users' PII.

Based on the design in Figure 1, it is necessary to design a system component, namely the machine learning component. The machine learning component is one of the critical components in constructing this detection system. Machine learning used in this study uses a modified VGG-16 architecture. Modification of VGG-16 is carried out on the fully connected layer by utilizing transfer learning techniques in each convolution neural network block. The proposed system architecture for use in a PII-sensitive content image detection system is shown in Figure 2.

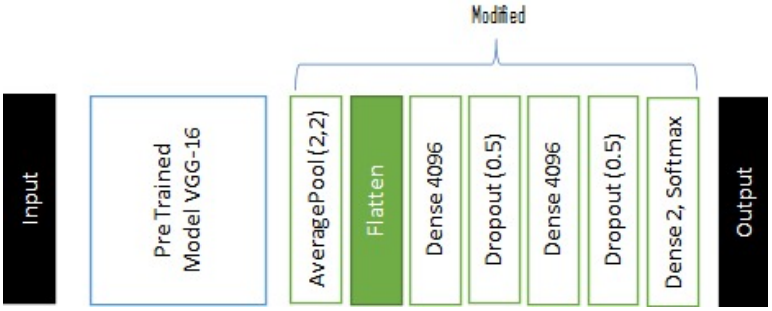

FIGURE 2. Modified VGG-16 architecture

4. Implementation and Evaluation.

4.1. **Dataset preparation.** The next step is to do the implementation. Carrying out the implementation required appropriate dataset. Dataset collection is done by doing image scrapping. The results of the image scrapping process obtained a total of 1200 images with details of 600 sensitive images with PII and 600 non-sensitive images. Each category's data is re-divided proportionally into training and validation data. The comparison of the amount of training data and validation data is 8 : 2. For using the model, the dataset needs to be augmented with data. The data augmentation process is carried out on the training and validation datasets. The parameters used to perform data augmentation on the dataset are shown in Table 1.

TABLE 1. Hyper-parameter data augmentation performed on the dataset

Dataset	Hyper-parameter	Value
Training dataset	Rescale, shear_range, zoom_range	1/255, 0.2, 0.2
Validation dataset	Rescale	1/255

4.2. **Model test design.** The model used in developing a sensitive content image detection system with PII is a modified VGG-16. This model was trained with epoch values and learning rates of 10 and 0.001. The model will be trained using a previously prepared dataset. The modified VGG-16 model proposed in this study will be simulated on a computer device with the specifications in Table 2.

TABLE 2. Specifications of hardware and software

Specification	Value
Operating system	Windows 11 Home Single Language
Processor	AMD Ryzen 5 4600G with Radeon Graphics 3.70 GHz
IDE	Anaconda dengan Jupyter Lab (Python 3.6.5)
Library machine learning	Scikit-learn, and Tensorflow Keras

The test is carried out by comparing the performance matrix between several transfer learning scenarios to the model made with the original VGG-16 model. The model needs to consider the dataset created to get good results. There are 3 transfer learning scenarios proposed in this study.

- The VGG-16 model is frozen in the 3rd block of the convolution layer. Subsequently, the 4th and 5th convolution blocks were rebuilt according to the original VGG-16 model. Furthermore, the fully connected layer is added to the modification section proposed in this study.

- The VGG-16 model is frozen in the 4th block of the convolution layer. Subsequently, the 5th convolution block was rebuilt according to the original VGG-16 model. Furthermore, the fully connected layer is added to the modification section proposed in this study.
- The VGG-16 model is frozen in the 5th block of the convolution layer. Furthermore, the fully connected layer is added to the modification section proposed in this study.

The training results on the VGG-16 modification are represented in accuracy and loss graphs. The results of this training can also show the values of Accuracy, Precision, Recall, Specificity, and the F1 Score of the model. The output of this training process is a file with the extension 'h5'. This file with the 'h5' extension will be used to construct the new model when implementing the system.

4.3. **Model training.** The modified VGG-16 model was trained on a dataset created in 10 epochs. This model is trained using the first, second, and third transfer learning scenarios. Results of this model training are represented on accuracy and loss graphs. The graph of accuracy and loss is shown in Figure 3. Performance evaluation of the first scenario VGG-16 modification is shown in Table 3. Table 3 shows that the performance of the modified VGG-16 with the second scenario portrays the Recall, Accuracy, and F1 Score values are better than the other two scenarios. This shows that modified VGG-16

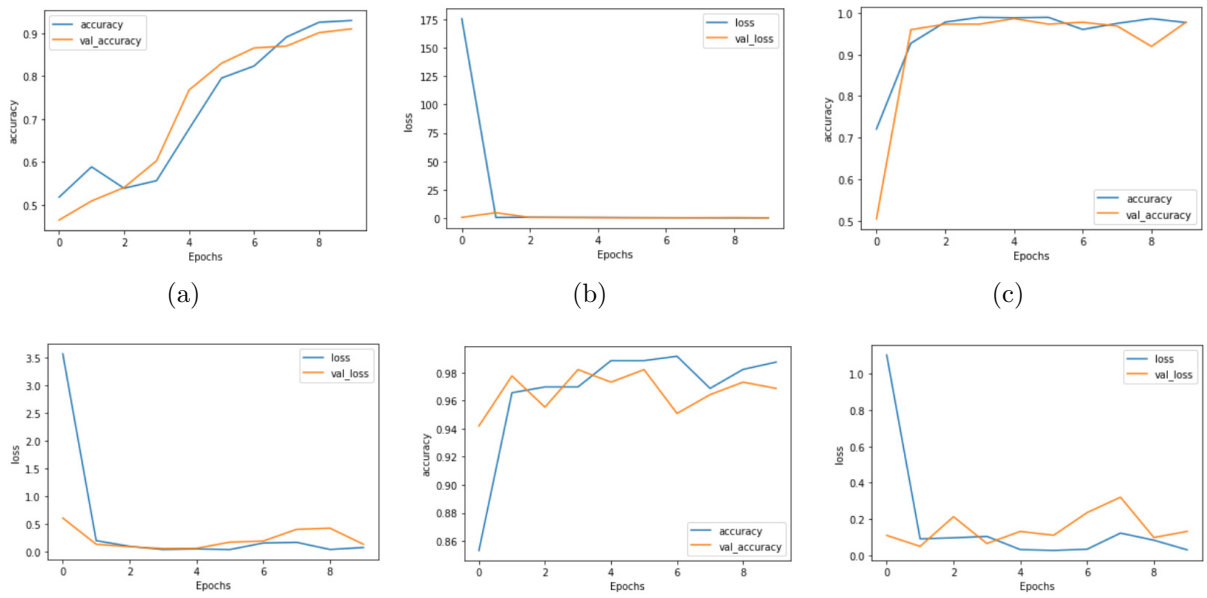

FIGURE 3. Results of this model training: (a) Accuracy of the first scenario VGG-16 modification; (b) loss of the first scenario VGG-16 modification; (c) accuracy of the second scenario VGG-16 modification; (d) loss of the second scenario VGG-16 modification; (e) accuracy of the third scenario VGG-16 modification; (f) loss of the third scenario VGG-16 modification

TABLE 3. Performance evaluation of the VGG-16 modification

Matrics	1st scenario	2nd scenario	3rd scenario
Recall	0.867	0.992	0.975
Specificity	0.95	0.967	0.967
Precision	0.945	0.967	0.967
Accuracy	0.908	0.979	0.971
F1 Score	0.904	0.979	0.971

with the frozen transfer learning scenario after the fourth convolution block shows the best results.

To provide a higher level of confidence in the use of this model, then the modified VGG-16 model with the transfer learning scenario frozen after the fourth convolution block will be compared to evaluate its performance against the original VGG-16 architectural model. The comparison model utilized two scenarios of the original VGG-16 architecture in this study. The first scenario used the original VGG-16 architecture trained on a sensitive image dataset with PII. In this scenario, the final layer is changed to Dense 2 layer. This is done to adjust the class of the dataset. The second scenario utilized the VGG-16 architecture trained with the ImageNet dataset and the sensitive image dataset with PII. The training in this scenario is carried out using transfer learning techniques by placing a freezing point before the final layer. The last layer is changed to Dense 2 layer. This is also done to adjust the output class based on the dataset. The performance comparison of the modified VGG-16 model with the second scenario with the two scenarios of the original VGG-16 is shown in Figure 4.

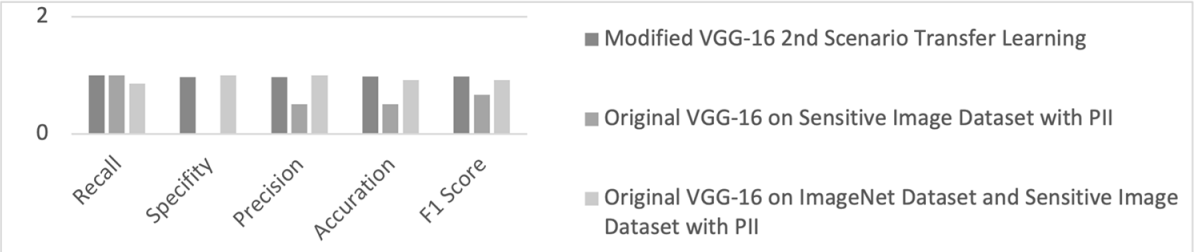

FIGURE 4. Comparison of modified VGG-16 2nd scenario transfer learning with original VGG-16

Based on Figure 4, the original VGG-16 model trained against a sensitive image dataset with PII showed poor performance. This is because the depth level of VGG-16 is not good enough if it is built using a small dataset. VGG-16 can offer improved performance when trained using large datasets such as ImageNet datasets. This can be proven by running the second scenario of the original VGG-16 architecture. However, suppose the two scenarios of the original VGG-16 architecture are compared with the modified VGG-16 model with a transfer learning technique by freezing after the fourth convolution block. In that case, it shows that the modified VGG-16 model with transfer learning technique by freezing after the fourth convolution block has better performance than the others. Based on that information, the model used in the implementation of the privacy agent in the image detection system with sensitive content with PII on social media is the modified VGG-16 model with a transfer learning technique by freezing after the fourth convolution block.

4.4. **Implementation of privacy agent on social media simulator.** Implementing the modified VGG-16 model on the privacy agent in the image detection system with sensitive content with PII on social media is carried out in a social media simulation. This simulation utilizes the python programming language in the Google Collab environment. In this implementation, the model training files on the dataset are used to build a new model in the system. The test scenarios for the implementation of this privacy agent include the following: testing by uploading images with sensitive content with PII on the system, and testing by uploading non-sensitive images on the system. The test results of this system show promising results in detecting images with sensitive content with PII. The system successfully detects images with PII-sensitive content and non-sensitive images.

5. **Conclusions.** An image detection system with sensitive content with PII on social media is necessary because many users are not aware of the importance of personal data confidentiality. The VGG-16 model modified by transfer learning technique by freezing after the fourth convolution block has good performance compared to the original model. The performance matrix of the proposed model in this study obtained Recall, Specificity, Precision, Accuracy, and F1 Score values of 0.992, 0.967, 0.967, 0.979, and 0.979, respectively. For future research, researchers suggest development using several machine learning architectures. In addition, researchers can also conduct research with implementation on real social media platforms.

Acknowledgment. This work is supported by the Ministry of Education, Culture, Research, and Technology (Kemendikbudristek) of Indonesia under PDD Grant with contract number NKB-1016/UN2.RST/HKP.05.00/2022.

REFERENCES

[1] A. K. Jain, S. R. Sahoo and J. Kaubiyal, Online social networks security and privacy: Comprehensive review and analysis, *Complex & Intelligent Systems*, vol.7, no.5, pp.2157-2177, DOI: 10.1007/s40747-021-00409-7, 2021.

[2] G. Beigi and H. Liu, Privacy in social media: Identification, mitigation and applications, *arXiv.org*, arXiv: 1808.02191, 2018.

[3] N. Karusala, A. Bhalla and N. Kumar, Privacy, patriarchy, and participation on social media, *Proc. of the 2019 on Designing Interactive Systems Conference*, San Diego, CA, USA, pp.511-526, DOI: 10.1145/3322276.3322355, 2019.

[4] G. S. O'Keeffe, K. Clarke-Pearson and Council on Communications and Media, The impact of social media on children, adolescents, and families, *Pediatrics*, vol.127, no.4, pp.800-804, DOI: 10.1542/peds.2011-0054, 2011.

[5] N. C. Krämer and J. Schäwel, Mastering the challenge of balancing self-disclosure and privacy in social media, *Current Opinion in Psychology*, vol.31, pp.67-71, DOI: 10.1016/j.copsyc.2019.08.003, 2020.

[6] L. Bioglio and R. G. Pensa, Analysis and classification of privacy-sensitive content in social media posts, *EPJ Data Sci.*, vol.11, no.1, 12, DOI: 10.1140/epjds/s13688-022-00324-y, 2022.

[7] J. M. Such and N. Criado, Multiparty privacy in social media, *Commun. ACM*, vol.61, no.8, pp.74-81, DOI: 10.1145/3208039, 2018.

[8] J. Frankenfield, *Personally Identifiable Information (PII)*, https://www.investopedia.com/terms/p/personally-identifiable-information-pii.asp, Accessed on Feb. 25, 2022.

[9] R. Rana, R. N. Zaeem and K. S. Barber, US-centric vs. international personally identifiable information: A comparison using the UT CID identity ecosystem, *2018 International Carnahan Conference on Security Technology (ICCST)*, pp.1-5, DOI: 10.1109/CCST.2018.8585479, 2018.

[10] A. Barushka and P. Hajek, Spam detection on social networks using cost-sensitive feature selection and ensemble-based regularized deep neural networks, *Neural Computing and Applications*, vol.32, no.9, pp.4239-4257, DOI: 10.1007/s00521-019-04331-5, 2020.

[11] Y. Xu, X. Meng, Y. Li and X. Xu, Research on privacy disclosure detection method in social networks based on multi-dimensional deep learning, *Computers, Materials & Continua*, vol.62, no.1, pp.137-155, DOI: 10.32604/cmc.2020.05825, 2020.

[12] J. Alemany, V. Botti-Cebriá, E. del Val and A. García-Fornes, Detection and nudge-intervention on sensitive information in social networks, *Logic Journal of the IGPL*, DOI: 10.1093/jigpal/jzac004, 2022.

[13] X. Meng and Y. Xu, Research on sensitive content detection in social networks, *CCF Transactions on Networking*, vol.2, no.2, pp.126-135, DOI: 10.1007/s42045-019-00021-x, 2019.

[14] S. Noppitak and O. Surinta, Ensemble convolutional neural network architectures for land use classification in economic crops aerial images, *ICIC Express Letters*, vol.15, no.6, pp.531-543, 2021.

[15] B. Baheti, S. Gajre and S. Talbar, Detection of distracted driver using convolutional neural network, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Salt Lake City, UT, USA, pp.1145-1151, DOI: 10.1109/CVPRW.2018.00150, 2018.

[16] V. Kumar, A. Zarrad, R. Gupta and O. Cheikhrouhou, COV-DLS: Prediction of COVID-19 from X-rays using enhanced deep transfer learning techniques, *Journal of Healthcare Engineering*, vol.2022, pp.1-13, DOI: 10.1155/2022/6216273, 2022.

[17] J. Pardede, B. Sitohang, S. Akbar and M. L. Khodra, Implementation of transfer learning using VGG16 on fruit ripeness detection, *IJISA*, vol.13, no.2, pp.52-61, DOI: 10.5815/ijisa.2021.02.04, 2021.
[18] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*, MIT Press, 2016.
[19] A. Khan and Z. Ansari, Improved VGG-16 convolutional neural network based lung cancer classification and identification on computed tomography, *Journal of Network Communications and Emerging Technologies (JNCET)*, vol.11, no.2, 15, 2021.
[20] E. Kozegar, Cystoscopic image classification by an ensemble of VGG-nets, *IJNAA*, vol.12, no.1, DOI: 10.22075/ijnaa.2021.4876, 2021.
[21] J. Wilches, VGG fine-tuning for cooking state recognition, *State Recognition Symposium*, DOI: 10.32555/2019.dl.007, 2019.
[22] A. Bagaskara and M. Suryanegara, Evaluation of VGG-16 and VGG-19 deep learning architecture for classifying dementia people, *2021 4th International Conference of Computer and Informatics Engineering (IC2IE)*, Depok, Indonesia, pp.1-4, DOI: 10.1109/IC2IE53219.2021.9649132, 2021.
[23] M. F. Haque, H.-Y. Lim and D.-S. Kang, Object detection based on VGG with ResNet network, *2019 International Conference on Electronics, Information, and Communication (ICEIC)*, Auckland, New Zealand, pp.1-3, DOI: 10.23919/ELINFOCOM.2019.8706476, 2019.
[24] W. Rawat and Z. Wang, Deep convolutional neural networks for image classification: A comprehensive review, *Neural Computation*, vol.29, no.9, pp.2352-2449, DOI: 10.1162/neco_a_00990, 2017.
[25] W. Behutiye, P. Rodríguez and M. Oivo, Quality requirement documentation guidelines for agile software development, *IEEE Access*, vol.10, pp.70154-70173, DOI: 10.1109/ACCESS.2022.3187106, 2022.
[26] S. Radack, *Guide to Protecting Personally Identifiable Information*, https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=905656, Accessed in Apr., 2010.