How May Deep Learning Testing Inform Model Generalizability? The Case of Image Classification

Giammaria Giordano*, Valeria Pontillo, Giusy Annunziata, Antonio Cimino, Filomena Ferrucci and Fabio Palomba

Software Engineering (SeSa) Lab — Department of Computer Science, University of Salerno (Italy)

Abstract

Artificial intelligence (AI) has become increasingly popular and is used in various fields, particularly image recognition. Several studies use images to train self-driving car models, security monitoring systems, recognize signals, etc. However, the approach taken to design and evaluate AI models can significantly affect the resulting performance of the models during operation. Hence, applying a rigorous approach to the design and evaluation of AI models may become crucial: this is the ultimate goal of the research field of *Software Engineering for Artificial Intelligence*. While current literature on image recognition proposed AI pipelines achieving good performance, it is still unclear how they would work in a real environment, where additional social and environmental factors come into play. In this paper, we propose a preliminary investigation into the role of input testing as a early indicator of the real-world performance of deep learning models in the context of image recognition. By taking the well-known Fashion-MNIST dataset into account, we first design a Convolutional Neural Network able to recognize images, in an effort of replicating the work done in previous studies and establishing a baseline. Then, we propose the use of input testing to simulate real-case conditions. Our preliminary results show that the devised CNN can lead to *precision*, *recall*, *F-Measure*, and *accuracy* close to 90%, hence confirming the results of previous experimentation in the field. Nonetheless, when input testing is applied, the performance of the model drastically drops (reaching ≈30%), possibly highlighting the need for revisiting image recognition models.

Keywords

Empirical Software Engineering, Software Engineering for Artificial Intelligence, Deep Learning.

1. Introduction

Software Engineering for Artificial Intelligence (SE4AI) refers to the use of software engineering principles to manage complex Artificial Intelligence (AI) models in order to rigorously test and ensure their scalability, interoperability, and maintenance over time [1]. Hence, the principles of SE4AI could be applied with the aim of developing effective, efficient, reliable, and sustainable AI models. In the last years, researchers and practitioners have been focusing on object recognition and image classification, developing a large amount of AI systems with good performance [2, 3]. The reason behind this choice is related to the availability of large datasets of images, *e.g.,* Fashion-MNIST or MNIST datasets, that can be applied in various studies spanning different fields, *e.g.,* from healthcare to self-driving cars [4, 5].

Unfortunately, all that glitters is not gold: despite the promising results obtained in previous studies, the applicability of these models in a real-world scenario still seems to be quite limited today due to external conditions, *e.g.,* environmental factors, which can render the systems unsuitable. As an example, Beede *et al.* [6] investigated the prediction correctness of deep learning models for diabetic eye disease with a strong performance in in-vitro experiments. Their results indicated poor performance due to socio-environmental factors that impacted the in-vivo experimentation. This study suggests that an improved assessment of these models would inform the design of effective solutions that may reach good performance when employed in production.

For this reason, this paper proposes a preliminary investigation into the *ecological validity* [7] of AI models proposed in the context of image recognition, namely we aim to understand how generalizable the experimental results previously presented would be in a real-case scenario. More specifically, starting from the Fashion-MNIST dataset, we first built a *Convolutional Neural Network* (CNN) using software engineering principles, hence conducting in-vitro experimentation in an effort of corroborating previous results and establishing a baseline. Then, we apply input testing [8], with the aim to understand to what extent the training set data fit the AI model, *i.e.,* altering the inputs of the model to simulate an in-vivo experimentation [9].

SATToSE'23: 15th Seminar Series on Advanced Techniques & Tools for Software Evolution, June 12–14, 2023, Fisciano, Italy

*Corresponding author.

giagiordano@unisa.it (G. Giordano); vpontillo@unisa.it (V. Pontillo); gannunziata@unisa.it (G. Annunziata); a.cimino10@studenti.unisa.it (A. Cimino); fferrucci@unisa.it (F. Ferrucci); fpalomba@unisa.it (F. Palomba)

0000-0003-2567-440X (G. Giordano); 0000-0001-6012-9947 (V. Pontillo); 0009-0002-0742-7261 (G. Annunziata); 0000-0002-0975-8972 (F. Ferrucci); 0000-0001-9337-5116 (F. Palomba)

On the one hand, our preliminary findings corroborate the image recognition performance reported in literature when considering the in-vitro experimentation: indeed, the engineered CNN developed reached levels of *precision*, *recall*, *F-Measure*, and *accuracy* close to 90%. On the other hand, we discover that the performance of the same model drastically drops when input testing is applied, hence suggesting that (1) the currently available models would not properly work in practice and (2) input testing may provide insights to machine learning engineers on the generalizability of the model in practice, hence possibly informing their design actions.

Structure of the paper. Section 2 overviews the background and the state of the art by pointing out the main differences between our work and the literature. Section 3 overviews the research questions driving our study and the research method, while Section 4 discusses our preliminary results. Finally, Section 5 summarizes the highlights of this work and outlines our future work.

2. Background and Related Work

This section describes the background and the related work that are the foundations of our proposed approach.

2.1. Background

Most of the research conducted on image recognition relied on the so-called Fashion-MNIST dataset.[1] This is the reason why the research presented in the remainder of this paper focuses on understanding the performance of a deep learning solution on this dataset. In particular, Fashion-MNIST is a clothes dataset based on the assortment on Zalando's website proposed by Xiao et al. [10]. It is considered a benchmark dataset containing images with the following characteristics: (1) all instances are normalized in a dimension of 28x28 pixels; (2) the images are preprocessed and converted into a gray scale; and (3) each pixel is composed of a value ranging from 0 to 255 based on the color intensity. The dataset contains over 70,000 examples of t-shirts, dresses, and so on, split into two sets, the training that contains 60,000 images and the test with 10,000 instances. In addition, the dataset is divided into 10 classes, one for each clothes category, *e.g.,* t-shirts, trousers, and pullovers. Figure 1 shows some images from the dataset.

2.2. Related Work

In the context of object detection and image classification [11, 12], Fashion-MNIST dataset appears in the top

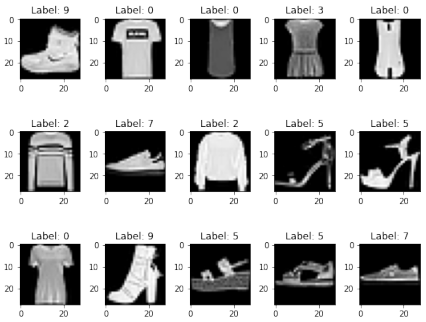

Figure 1: Example images from the Fashion-MNIST dataset.

10 most datasets used for several purposes, *e.g.,* investigation privacy [13] issues. Xiao et al. [10], the authors of Fashion-MNIST dataset, compared several classifiers, *e.g., Decision Tree* and *Extra Tree Classifier*, and they achieved performance on average around 80% in terms of *accuracy*. In 2021, Leithardt [14] performed a comparison between different classification methods *e.g., Support Vector Classification*, *Linear Support Vector Classification*, and various *Convolutional Neural Network* approaches. The best classification model was *CNN-dropout-3*, with accuracy above 99%, while the worst model was *Gaussian Naive Bayes* with accuracy around 51%. Saquib and Zahra [15] proposed an improvement of the Adam algorithm [16] named *Mean-ADAM* meant to reduce the oscillation of the weights—which is usually considered the general problem that makes the accuracy fall—and outperforms all other adaptive gradient methods until final training. A g_2^t stochastic optimization algorithm is used, by which the variance of the weights might increase during the optimization, therefore, there is a progressive reduction of the external weight that will improve the accuracy, generalizability, and data set invariance. Their results showed good *accuracy* (reaching ≈90%) for several neural networks, *e.g., ResNet*, *VGGNet*, and *Inception V1*.

Greeshma et al. [17] presented the classification of Fashion-MNIST dataset using a *Multiclass Support Vector Machine* (SVM); their results showed an *accuracy* above 86%. Similarly, Xhaferra et al. [12] used deep learning models in e-commerce to solve problems related to clothing recognition. The authors developed a neural network with an *accuracy* of 93.1%. Bhatnagar et al. [18] proposed three different convolutional neural network architectures using batch normalization and residual skip connections, reaching 90% *accuracy*. Finally, Kayed et al. [19] built upon previous studies and improved CNN's performance by leveraging a *LeNet-5* architecture. In this way, the authors achieved 98% accuracy.

While experimenting with multiple shallow and deep learning solutions, most of the studies discussed above

[1]The Fashion-MNIST dataset: https://github.com/zalandoresearch/fashion-mnist

reported the models based on Convolutional Neural Network (CNN) as the best solutions fitting the problem of image recognition. This aspect informed the design of our experiment, which indeed investigates the in-vitro and in-vivo performance of a CNN model.

2.3. Limitations of the State of the Art

By analyzing the state of the art, we highlighted a number of challenges for the *Software Engineering for Artificial Intelligence* (SE4AI) research community. The interested reader might have a full overview of the current challenges in the field through the systematic literature review conducted by Giordano et al. [13].

First, we observed that previous work assessed the proposed approaches only through in-vitro experimentation, hence investigating the performance of machine and deep learning models in terms of performance indicators computed when running them against datasets using validation strategies such as percentage split or cross-fold validation. On the contrary, to the best of our knowledge, there is no study that attempted to provide indications of the ecological validity of the models.

In addition, most studies only experimented with the *accuracy* metric [20], namely the total amount of correct predictions made by a model. However, the use of accuracy can cause multiple biases. In the first place, the accuracy does not consider the distribution of the training and test sets. Suppose the training data is significantly different from the test data. In that case, the accuracy metric can be biased due to the learning effect where the model memorizes the training data instead of learning the true underlying data model. In the second place, although accuracy is one of the most analyzed metrics for understanding the effectiveness of an AI model, it is not an appropriate measure for unbalanced datasets since it does not distinguish between the numbers of correctly classified examples of different classes, leading to erroneous conclusions [21]. For this reason, it may be more appropriate to consider other evaluation metrics such as *F-Measure* and *recall* to assess the AI models.

In this work, we aim at addressing the two limitations above. We indeed devised a baseline CNN model to classify images that we first assessed through multiple performance indicators. Afterward, we experimented with input testing to investigate the potential ecological validity of the model in a real-case scenario.

3. Research Method

The ultimate *goal* of this study was to apply input testing methods to verify the behavior of a deep neural network model built in the context of image recognition, with the *purpose* of analyzing how the model would potentially work in a real-world scenario. The *perspective* is of both researchers and practitioners; the former are interested in assessing the current state of the art, hence understanding how software engineering practices can assist the development of AI solutions. The latter are interested in evaluating the capabilities of AI models in a real-context scenario. Based on the previous considerations, we ask:

> **RQ$_1$.** *What is the performance of an engineered Convolution Neural Network when applied for the task of image recognition?*

> **RQ$_2$.** *To what extent the application of input testing methods impact the performance of an engineered Convolution Neural Network when applied for the task of image recognition?*

Figure 2 shows the research method applied to answer our research questions. Specifically, to address **RQ$_1$**, we developed a *Convolutional Neural Network* (CNN) and applied it on the Fashion-MNIST dataset. We trained the algorithm applying the data augmentation, *i.e.,* a technique that allows us to increase the data available by modifying the initial images with filters to change the color palette to permit us to increase the original dataset size from 60,000 entries to 300,000. Finally, we divided the dataset by 85% for the training set and 15% for the test set. To understand the performance of our model, we evaluated the approach with a number of state-of-the-art metrics, *i.e., precision*, *recall*, *F-Measure*, and *accuracy* [22].

Once we had established a baseline, we proceeded with **RQ$_2$**, where we focused on the potential behavior of the CNN in a real-world context. Specifically, we applied input testing methods [8] to analyze the training data used to train the model, with the aim to identify potential issues in the training set data. Hence, we created customized instances of the Fashion-MNIST dataset by introducing different noises on the test set data to simulate a real-world scenario, e.g, rain or fog. For our preliminary evaluation, we applied a cut filtering on the Fashion-MNIST dataset to simulate the scenario in which images are not perfectly aligned to the center. Figure 3 shows an example of the application of this filter on Fashion-MNIST dataset: for each clothes category, the cut was made vertically in the center of the image, so the garment is not fully visible. The application of this filter is useful for simulating low visibility conditions, *e.g.,* traffic signs that are not fully visible in the context of self-driving cars. We then re-assessed the approach in terms of *precision*, *recall*, *F-Measure*, and *accuracy* [22].

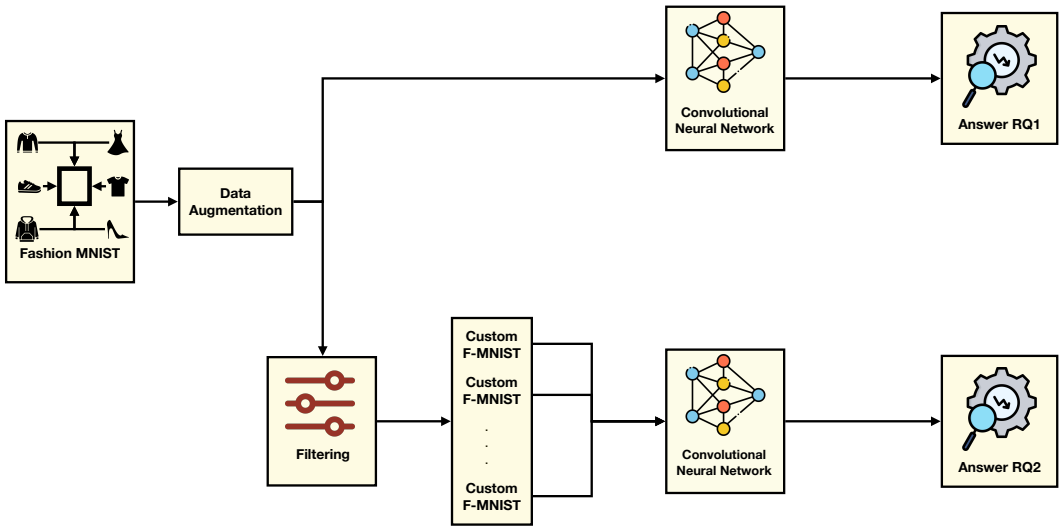

Figure 2: Overview of the research method.

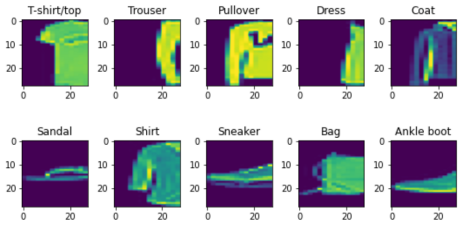

Figure 3: An example of the application of a cut filter.

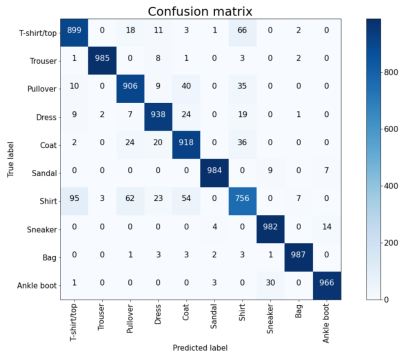

Figure 4: Confusion matrix after the data augmentation.

4. Preliminary Results

The following sections describe the preliminary results achieved to address the two research questions.

4.1. RQ_1 - Replicating Previous Experiments

Figure 4 shows the confusion matrix results after the data augmentation. The model achieved good prediction levels for all garments (from 89.9% to 98.7%) except the class `shirt`, in which the precision is around 76%. As the reader may observe, one in four elements is not classified correctly as a shirt but is misclassified as a `t-shirt` or `coat`—this probably happens because the three clothing classes are similar to each other.

ID	Epoch	Batch Size	Precision	Recall	F-Measure	Accuracy
0	30	128	0.92	0.92	0.92	0.92
1	50	128	0.93	0.93	0.93	0.93
2	60	128	0.93	0.93	0.93	0.93
3	70	128	0.93	0.93	0.93	0.93
4	30	256	0.93	0.93	0.93	0.93
5	50	256	0.93	0.93	0.93	0.93
6	60	256	0.93	0.93	0.93	0.93
7	70	256	0.93	0.93	0.93	0.93

Table 1
Results of the CNN model created to answer to $\mathbf{RQ_1}$.

Table 1 shows the results obtained by the CNN algorithm. We can observe that, in terms of *accuracy*, the model achieves good performance (above 90%), going

to confirm the results already shown in the literature [11, 12, 14]. Analyzing the other metrics, we can also see that the performance is always very positive (again above 90%) for each epoch and batch size considered. To conclude, our replication found results similar to those reported in previous experiments, hence confirming that a CNN approach can effectively recognize images when applied against the Fashion-MNIST dataset.

Key findings of RQ_1.

Our replication study corroborates previous findings in the field of image recognition through AI. The performance of the CNN model is over 90% in terms of *accuracy*, *F-Measure*, *Recall*, and *Precision*.

4.2. RQ_2 - On the Impact of Input Testing

Figure 5 shows the confusion matrix results for our model evaluated in a real-world scenario. This confusion matrix is very different from the one obtained in the previous analysis, in fact, we can observe a decrease of the performance in all clothing classes, especially for `trouser` and `dress` classes where the precision is less than 10%.

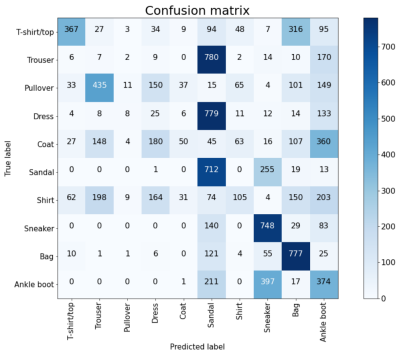

Figure 5: Confusion matrix of the model after applying the cut filter to the Fashion-MNIST dataset.

The only classes in which precision achieves good performance (above 70%) are `sandal`, `sneaker`, and `bag` classes. This could happen because, although the elements were cut in half, they still preserved elements that make them always distinguishable.

Finally, Table 2 reports the results for the CNN model after the application of the cut filtering. In this case, we can observe a severe decrease of all metrics, especially for the *F-Measure*, that reached no more than 27% against the previous 93%.

In addition, also the other metrics, *i.e., Precision*, *Recall*, and *Accuracy*, do not reach values above 40%. These results suggest that when the model cannot consider the entire image of a garment, then it may have a large loss of information, *e.g.*, on the shape, which leads to lower performance. While further analysis is required to understand how usable and generalizable deep learning models are in a real-world context, our findings suggest that (1) existing models would not properly work in conditions where the images are not perfectly passed as input; and (2) input testing seems to be a valid instrument to establish the performance of AI models, possibly informing machine learning engineers and data scientists on the need for taking further actions.

ID	Epoch	Batch Size	Precision	Recall	F1-Score	Accuracy
0	30	128	0.30	0.25	0.19	0.25
1	50	128	0.33	0.31	0.23	0.31
2	60	128	0.31	0.28	0.21	0.28
3	70	128	0.28	0.28	0.21	0.28
4	30	256	0.32	0.28	0.21	0.28
5	50	256	0.33	0.32	0.27	0.32
6	60	256	0.29	0.26	0.19	0.26
7	70	256	0.30	0.26	0.20	0.26

Table 2
Results of the CNN model applying the cut filter.

Key findings of RQ_2.

Our preliminary results indicated that the application of input testing methods lets the performance of the CNN decrease up to 60% with respect to what reported in literature. The overall performance ranged, indeed, between 19% to 33% in terms of *precision*, *recall*, *F-Measure*, and *accuracy*.

5. Conclusion

This paper provided a preliminary analysis of how existing deep learning solutions work when they are experimented in seemingly real conditions through the application of input testing methods. We first replicated the design of a previously defined CNN model in the context of image recognition, finding similar performance as those reported in the literature. Afterward, we re-assessed the performance of the model after the application of input testing methods, discovering a notable drop in terms of all performance indicators measured.

The reported results might open some discussion on the validation procedures to adopt when experimenting with AI solutions, possibly paving the way to new methodologies and standards to address the performance of those models. At the same time, our findings suggest that the research conducted in the field of image recognition through AI might be worth of re-visitation to properly understand the actual soundness of those techniques in practice.

Our future research agenda includes an extension of this work, in which we aim to assess the performance

of CNN-based models when considering a larger variety of input testing methods and in-vivo scenarios. Furthermore, we aim to experiment with additional use cases, like the models employed in the context of self-driving cars, security assessment, and others.

Acknowledgments

Fabio is partially funded by the Swiss National Science Foundation through SNF Projects No. PZ00P2_186090.

References

[1] E. Nascimento, A. Nguyen-Duc, I. Sundbø, T. Conte, Software engineering for artificial intelligence and machine learning software: A systematic literature review, arXiv preprint arXiv:2011.03751 (2020).

[2] E. R. G, S. M, A. R. G, S. D, T. Keerthi, R. S. R, Mnist handwritten digit recognition using machine learning, in: 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 2022, pp. 768–772. doi:10.1109/ICACITE53722.2022.9823806.

[3] Y. Rangoni, F. Shafait, J. van Beusekom, T. M. Breuel, Recognition driven page orientation detection, in: 2009 16th IEEE International Conference on Image Processing (ICIP), 2009, pp. 1989–1992. doi:10.1109/ICIP.2009.5413722.

[4] S. Du, H. Guo, A. Simpson, Self-driving car steering angle prediction based on image recognition, arXiv preprint arXiv:1912.05440 (2019).

[5] S. A. Fatima, A. Kumar, A. Pratap, S. S. Raoof, Object recognition and detection in remote sensing images: a comparative study, in: 2020 International Conference on Artificial Intelligence and Signal Processing (AISP), IEEE, 2020, pp. 1–5.

[6] E. Beede, E. Baylor, F. Hersch, A. Iurchenko, L. Wilcox, P. Ruamviboonsuk, L. M. Vardoulakis, A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 1–12. URL: https://doi.org/10.1145/3313831.3376718. doi:10.1145/3313831.3376718.

[7] M. A. Schmuckler, What is ecological validity? a dimensional analysis, Infancy 2 (2001) 419–436.

[8] V. Riccio, G. Jahangirova, A. Stocco, N. Humbatova, M. Weiss, P. Tonella, Testing machine learning based systems: a systematic mapping, Empirical Software Engineering 25 (2020) 5193–5254.

[9] A. H. M. Rubaiyat, Y. Qin, H. Alemzadeh, Experimental resilience assessment of an open-source driving agent, in: 2018 IEEE 23rd Pacific rim international symposium on dependable computing (PRDC), IEEE, 2018, pp. 54–63.

[10] H. Xiao, K. Rasul, R. Vollgraf, Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. arXiv:cs.LG/1708.07747.

[11] S. Bhatnagar, D. Ghosal, M. H. Kolekar, Classification of fashion article images using convolutional neural networks, in: 2017 Fourth International Conference on Image Information Processing (ICIIP), 2017, pp. 1–6. doi:10.1109/ICIIP.2017.8313740.

[12] E. Xhaferra, E. Cina, L. Toti, Classification of standard fashion mnist dataset using deep learning based cnn algorithms, in: 2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), 2022, pp. 494–498. doi:10.1109/ISMSIT56059.2022.9932737.

[13] G. Giordano, F. Palomba, F. Ferrucci, On the use of artificial intelligence to deal with privacy in iot systems: A systematic literature review, Journal of Systems and Software 193 (2022) 111475. URL: https://www.sciencedirect.com/science/article/pii/S0164121222001613. doi:https://doi.org/10.1016/j.jss.2022.111475.

[14] V. Leithardt, Classifying garments from fashion-mnist dataset through cnns, Advances in Science, Technology and Engineering Systems Journal 6 (2021) 989–994.

[15] N. Saqib, F. T. Zahra, An improved adaptive optimization technique for image classification, in: 2020 Joint 9th International Conference on Informatics, Electronics & Vision (ICIEV) and 2020 4th International Conference on Imaging, Vision & Pattern Recognition (icIVPR), 2020, pp. 1–6. doi:10.1109/ICIEVicIVPR48672.2020.9306620.

[16] Uxiliary, Emma, On the convergence of adam and beyond, ArXiv abs/1904.09237 (2018).

[17] K. Greeshma, K. Sreekumar, Fashion-mnist classification based on hog feature descriptor using svm, International Journal of Innovative Technology and Exploring Engineering 8 (2019) 960–962.

[18] S. Bhatnagar, D. Ghosal, M. H. Kolekar, Classification of fashion article images using convolutional neural networks, in: 2017 Fourth International Conference on Image Information Processing (ICIIP), IEEE, 2017, pp. 1–6.

[19] M. Kayed, A. Anter, H. Mohamed, Classification of garments from fashion mnist dataset using cnn lenet-5 architecture, in: 2020 international conference on innovative trends in communication and computer engineering (ITCE), IEEE, 2020, pp. 238–243.

[20] Y. Lu, X. Huang, K. Zhang, S. Maharjan, Y. Zhang,

Communication-efficient federated learning and permissioned blockchain for digital twin edge networks, IEEE Internet of Things Journal 8 (2021) 2276–2288. doi:10.1109/JIOT.2020.3015772.

[21] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, F. Herrera, A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42 (2012) 463–484. doi:10.1109/TSMCC.2011.2161285.

[22] R. Baeza-Yates, B. Ribeiro-Neto, et al., Modern information retrieval, volume 463, ACM press New York, 1999.