Image Classification on Fashion Dataset Using Inception V3

Maryamah Maryamah[1], Najma Attaqiya Alya[1], Muhammad Hanif Sudibyo[1], Ergidya Liviani[1], Razin Isyraq Thirafi[1]
,
[1]*Faculty of Advanced Technology and Multidiscipline, Universitas Airlangga, Surabaya, Indonesia*

***Abstract*— Fashion is a form of self-expression that allows us to be able to manifest our personality and identity more confidently. One of the effects of Covid 19 is the economics of the industry, especially the fashion category as the largest category in the e-commerce industry. However, A large number of categories required high labor costs, possibly misclassifications that have nearly the same clothing model and affected many subjective or objective factors. The other problem is sellers uploading pictures of products on the platform for sale and the consequent manual tagging involved. In this paper, we proposed image classification on the fashion dataset using inception V3. The methodology consists of scrapping data from the official websites of five famous fashion brands, data preprocessing, and classification with the Inception V3 method. The accuracy and F1-Score values obtained using Inception V3 are 92.86% and 92.85%. The proposed method is the highest result of the comparison method and can differentiate between knitted with a scarf that is difficult to classify when compared to other comparison methods.**

***Keywords*—classification, fashion image dataset, inception V3**

I. INTRODUCTION

Fashion is a form of self-expression that allows us to be able to manifest our personality and identity more confidently [1]. In 2020, the fashion industry was put to the major test by a global pandemic [2]. The Covid-19 crisis had a major impact on decreased sales in one of the luxury fashion industries report [3]. The Covid-19 pandemic has affected the fashion industry around the world. Many factories have taken policies to close manufacturing factories and cancellation of events in fashion demonstrations to reduce the spread of viruses. The “State of Fashion” report by McKinsey & Company shows that the economic profits of companies in the fashion industry fell by 90 percent in 2020 meanwhile the previous year increased by 4% [2].

With the reduction in Covid-19 cases and the termination of social distancing rules around the world, the fashion market is resurfacing. This policy affects the industrial economy where the fashion category is the largest in the e-commerce industry [4]. Supported by the government that allowed outdoor activities, many people paid more attention to their fashion. Consumer perception of a product understanding is influenced by online photos, which are often used to attract attention and influence consumer purchasing decisions in recent years. In the B2C e-commerce platform, different clothing brands can display different types of clothing styles. However, A large number of categories in each fashion brand allows shop owners to be misclassifications about the placement of items that have nearly the same clothing model. This process required high labor cost and affected many subjective or objective factors [5]. The other problem is sellers uploading pictures of products on the platform for sale and the consequent manual tagging involved.

In this paper, we proposed image classification on the fashion dataset using Inception V3. The methodology of this paper consists of scrapping data from the official websites of five famous fashion brands, data preprocessing, and classification with the Inception V3 method. 16 categories of fashion datasets are used because actual data categories from the fashion dataset are more numerous, and this total category is more than the dataset used in previous studies. The architecture of classification with the Inception V3 method consists of some distinct layer types: convolution, pooling, concat, dropout, fully connected, and dense. There are three-layer types that affect image representation: convolution, batch normalization, and max pooling.

II. RELATED WORKS

Inception V3 is a Convolutional Neural Network (CNN) architecture developed from the previous version of GoogleNet (Inception) and Inception V2. Inception V3 can be used for image classification. Fashion classification with an image dataset is also included in the image classification problem. Several studies related to fashion classification have previously been carried out using Convolutional Neural Network (CNN) Inception V3 and Transfer Learning. The dataset used was composed of 80.000 clothing images of 13 categories. The evaluation of Inception V3 and Transfer Learning obtain average accuracy of 70 %, with the highest classes of articles rate of 90%, and other categories have low accuracy because of the similarity of these classes [6]. Similar research conducted

Corresponding author:
Maryamah,
Faculty of Advanced Technology and Multidiscipline
Universitas Airlangga
Email: maryamah@ftmm.unair.ac.id

using a framework of Convolutional Neural Network (CNN) Inception V3 and Transfer Learning shows the methods can handle large datasets efficiently and accurately [7]. Hierarchical Convolutional Neural Networks (H-CNN) have also been used for the classification of fashion datasets (Fashion MNIST dataset) [8]. The H-CNN model result can improve the accuracy and loss gets decreased with the comparison with base methods without a hierarchical structure. H-CNN can get better performance in fashion classification.

In addition, by using the same dataset, the Convolutional Neural Network (CNN) model can increase the accuracy value by 2% [9]. CNN with batch normalization and dropout layers was proposed to reduce overfitting also outperforms accuracy compared with conventional CNN [10]. Other researchers using the same dataset and model also have proven that the results of the CNN model produce accurately classified the fashion dataset with DenseNet121 is the most powerful model compare with MobileNet and ResNet 50 [11]. There are several architectures that continue to develop to improve CNN performance, starting from LeNet, AlexNet, ZFNet, GoogLeNet, VGG to the latest ResNet. The differences in each architecture is based on the size of the input image, the number of layers, the choice of kernel size, and others [12]. Many studies related to fashion classification prove that CNN can predict the fashion image dataset correctly. However, the actual data categories from the fashion dataset are more numerous than the existing datasets. This can cause fashion image classification to have different results and challenges. In this paper, we used 16 categories which is more than the dataset used in previous studies.

III. Methodology

A. *Materials*

The image dataset used in this paper is obtained from scrapping data from five official websites of famous brands namely Gucci, LV, Dior, Prada, and Uniqlo. Imageeye software was used for scrapping the image dataset. There are 6,304 total image data then divided into 16 categories of fashion label. The image data used is 400 x 400 colored pixels (RGB). The amount of each data can be seen in Table 1.

Table 1. Research Data

Label	Amount of Data
Bags	423
Belts	263
Boots	392
Dresses	220
Hats	255
Jackets & Coats	396
Jewelry	310
Knitwear	665
Loafers	403
Pants	628
Sandals	531
Scarves	250
Skirt	268
Sneakers	357
Sunglasses	266
T-Shirts	677

B. *Methodology*

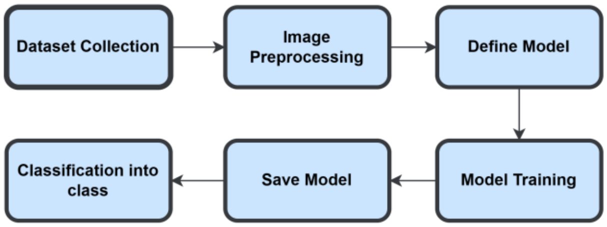


Figure. 1 The methodology process

The methodology of this paper shows in Figure 1 that consists of collecting a dataset with scrapping from five famous brands of fashion websites. The image processing began with splitting the data into ratios of 80% as training data and 20% as testing data. The next step after spitting data is data augmentation. Data augmentation is a process of modifying or manipulating an image and the original image in standard form will be changed in shape and position. In this paper, the augmentation image used was rotated, sheared, zoomed, and translated vertically and horizontally. Figure 2 is an example of data augmentation in this research. Our research makes one image dataset and augments the image into 32 different images to expand the dataset. Data augmentation is used to make machines learn and recognize different images and use them to reproduce data.

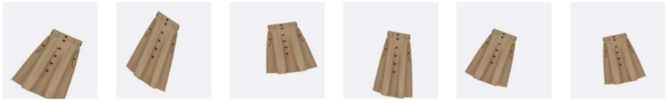


Figure. 2 Example augmentation process

Corresponding author:
Maryamah,
Faculty of Advanced Technology and Multidiscipline
Universitas Airlangga
Email: maryamah@ftmm.unair.ac.id

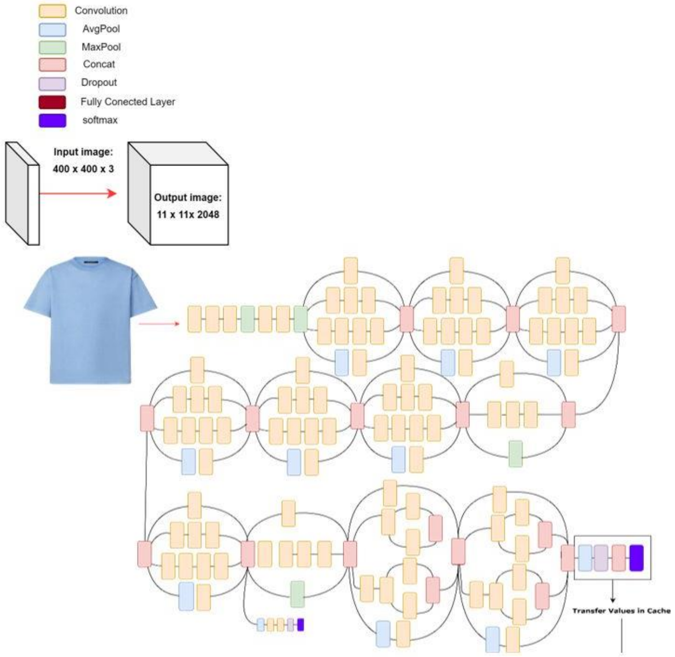


Figure. 3 Inception V3 model architecture

The next process is to build the Inception V3 model from the training image dataset. The model consists of 42 layers and has a lower error rate compared with Inception V1 and V2. From the architecture in Figure 3, all of those layers that are embedded in the model, there are a few distinct layer types: convolution, pooling, concat, dropout, fully connected, and dense. There are three-layer types that affect image representation: convolution, batch normalization, and max pooling. Figure 4 is a sample image that would be shown in the visualization of the sample image in training data to be converted into an image tensor.

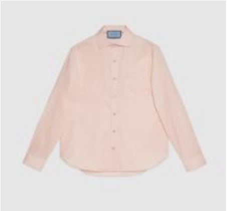
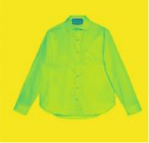
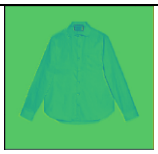
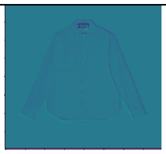
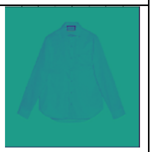


Figure. 4 Sample image in train data

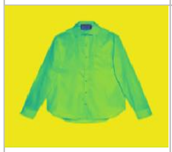
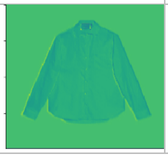
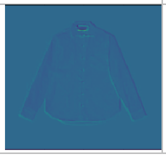
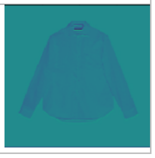
In convolution, the model will extract features that are embedded in images by doing matrix computation on each images dimension, resulting in a new matrix. That new matrix then gets fed into the next layer (pooling) where this layer works by reducing unnecessary information contained in the input data. From the first layer or in convolution, the primary feature that describes the image is discovered. Table 2 shows in the first channel, detail of the image is extracted example the wrinkle in the fabric of the cloth. At the third channel, the edge of the cloth is found and be used to describe how the outer cloth shape is classified. At the 35th layer, the depth of the image is found where the thickness of the fabric is more precise. At the 13th layer, the overall surface of the cloth is found to get an overview of the outer cloth.

Table 2. Convolution layer image result

Channel 1	Channel 3	Channel 35	Channel 13
			

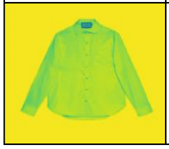
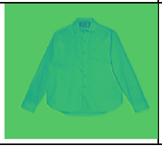
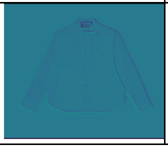
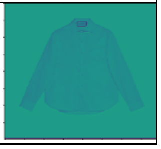
The pooling layer is used for reducing data and the deep learning model does not waste extra resources for training the same data. Table 3 shows from the first channel of the image until the last channel that the clarity of the image is degraded from the original. Reduced dimension from 400x400 pixels to 200x200 pixels used without loss of important features of the image. Because of the wide architecture of the model, unlike CNN which goes deep, this paper splits the structure and joined the layer again. The contact layer is used to combine the output of the split layers into one big matrix.

Table 3. Pooling layer image result

Channel 1	Channel 3	Channel 35	Channel 13
			

In this paper, the inception model has a total of 311 layers for the training model with the scrapping dataset. In part of the training model, dense and batch normalization layers were added to fit the model for the classification task. The final total layer for this model is 316 layers. At the batch normalization layer, normalization at each pixel on the image and centering of the data is performed and the position of the image is more center of the canvas. There is nothing major changed from the original image from this positioning of the image in the canvas. This process is used to reduce the likelihood of overfitting at model training. In batch normalization, the distinct colored pixels merged into one color to minimize needless information.

Table 4. Batch normalization layer image result

Channel 1	Channel 3	Channel 35	Channel 13
			

Flatten layer was a process of converting the calculated matrix into one row of data to be processed by a fully connected layer using 256 neurons to classify data. Classified data gets provided into the batch normalization layer where it gets normalized, and then goes into the dropout layer where this layer will deactivate some neurons and the model learns more

Corresponding author:
Maryamah,
Faculty of Advanced Technology and Multidiscipline
Universitas Airlangga
Email: maryamah@ftmm.unair.ac.id

efficiently than reducing overfit data training. The data will be classified into one of 16 labels using Softmax activation for the activation function to return a confidence percentage of the label or categories dataset.

Figure. 5 Inception V3 layer visualization

The layer depths of the model can be seen in Figure 5, the input for the first convolution layer is much bigger than the resulting convoluted matrix, and the dimension of the data decreases continuously for all the layers. This model consists of symmetrical and asymmetrical building blocks which include convolution, pooling, concatenations, dropouts, and fully connected layers. The loss metrics are calculated using Softmax and the training model was built.

The evaluation of the model in training results is used to understand the performance of the model. The training model is used to predict the categories of fashion in the testing dataset. In this paper, the evaluation metrics used accuracy, precision, recall, and F1-score metrics to evaluate the performance of the classification results. Comparison accuracy and loss of training and testing or validation dataset used to compare the performance.

IV. Result and Discussion

There are several analyzes carried out in this paper. First, the results of the epoch of training and testing or validation data using metrics accuracy and loss are in Figure 6. The performance of the model from accuracy metrics is great at training data but have lower result in testing data. The result was raised gradually until it stagnated result then the system did the early stopping of the model. The accuracy metric of the model scored 97,96% for the training data, while test data scored 92,86%. On loss metrics, training data keep decreasing gradually for more epochs. The testing data has stagnated results same as accuracy metrics then the system did the early stopping of the model. For the loss metric, the training data resulted in 0,066 scores while at test data have 0,305 scores. This concludes the model result is overfitted to training data. The strategy to decrease the loss metric is to improve training data and the model will learn the data more accurately.

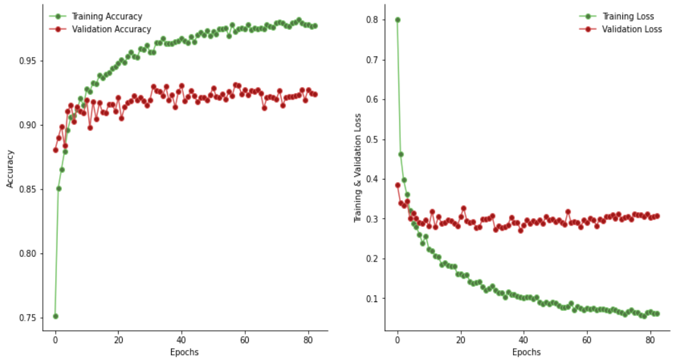


Figure. 6 Result epoch vs training and validation in accuracy/loss

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + True\ Negative + False\ Negative} \quad (1)$$

A comparison of other methods using two evaluation metrics was also conducted to determine the performance of the best method used for the classification of fashion image datasets. The evaluation metrics used are accuracy and F1-score. The accuracy metrics (1) is calculated based on the true positive class (correct classified positive class) plus true negative (correct classified negative class) divided by the sum of true positive, true negative, false positive (incorrectly classified of positive class), and false negative (incorrectly classified of negative class). The second metrics using F1-score that calculate from precision and recall value. The second metric uses an F1-score (2) calculated from two multiplied by multiplying precision and recall divided by precision plus recall. Precision (3) is calculated based on true positives divided by the addition of true positives and false positives. Recall (4) is calculated based on true positives divided by the addition of true positives and false negatives.

$$F1\ Score = 2\ \times \frac{Precision \times Recall}{Precision + Recall} \quad (2)$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (3)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (4)$$

In Table 5, the accuracy metric in this experiment is 92.86% using Inception V3. This result is the highest accuracy compared with other methods. The performance metric used for categorical accuracy is the adaptation of accuracy metrics for multinomial data that counts the percentage of correctly predicted labels. The other evaluation using F1-Score metrics from the Inception V3 model reach 92,85% with also the highest results. The results both in accuracy and F1-Score of the proposed method did not show a major difference. It can be concluded that the results of the proposed method are higher than other comparison methods: Naïve Bayes, Support Vector Machine, and XGBoost. Figure 7 shows both evaluation in F1-score and accuracy with the lowest result in the Naïve Bayes model. The SVM model gets the highest F1-Score of all three machine learning models and XGBoost is the second rank of the performance result.

Corresponding author:
Maryamah,
Faculty of Advanced Technology and Multidiscipline
Universitas Airlangga
Email: maryamah@ftmm.unair.ac.id

Table 5. Model performance

Comparison Methods	F1-Score	Accuracy
Naïve Bayes	53,08%	55,05%
SVM	87,47%	87,45%
XGBoost	86,38%	86,43%
Inception V3	92,85%	92,86%

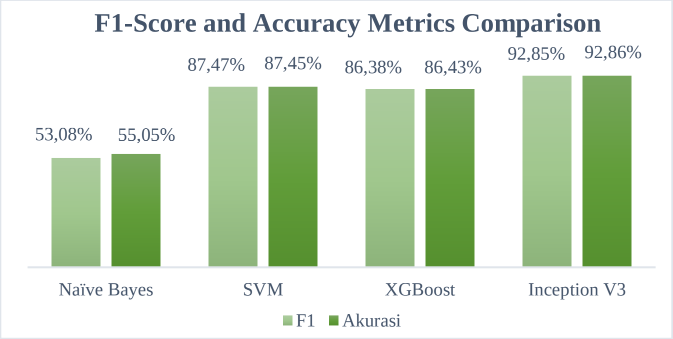


Figure. 7 Model performance comparison

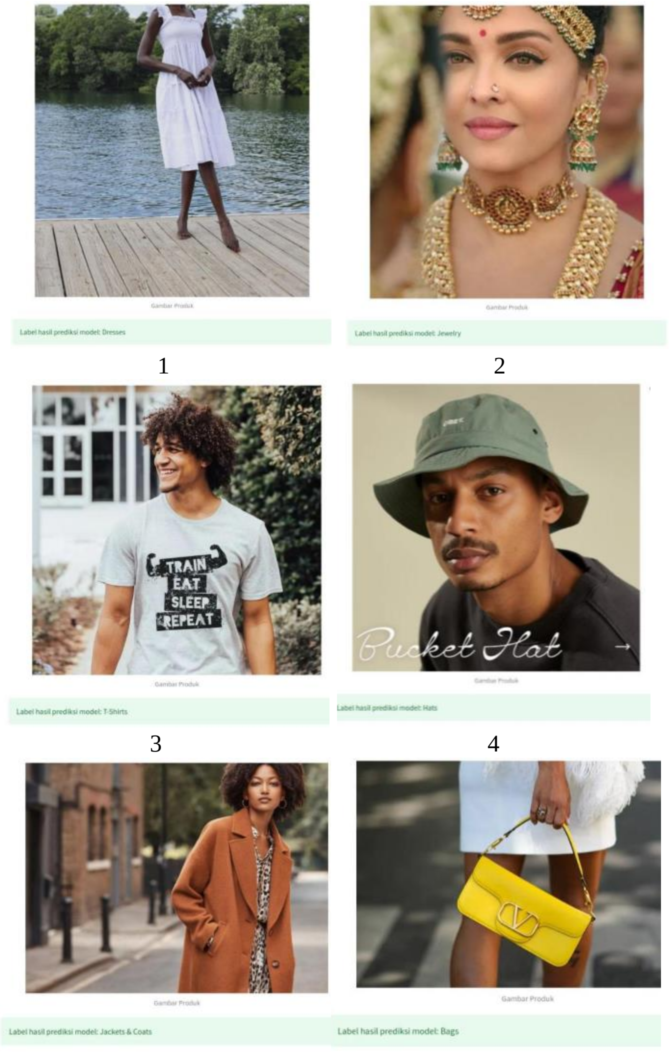
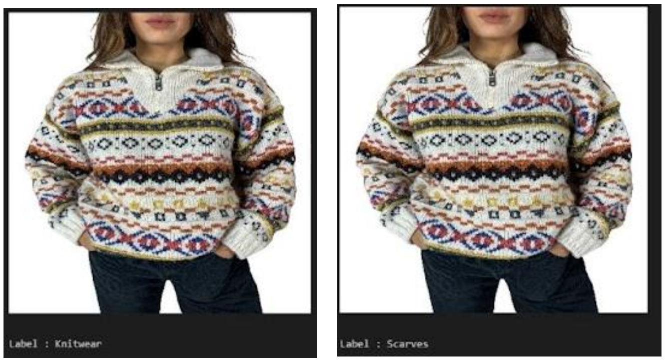

Gambar Produk
Label hasil prediksi model: Dresses
1
Gambar Produk
Label hasil prediksi model: Jewelry
2
TRAIN EAT SLEEP REPEAT
Gambar Produk
Label hasil prediksi model: T-Shirts
3
Bucket Hat
Gambar Produk
Label hasil prediksi model: Hats
4
Gambar Produk
Label hasil prediksi model: Jackets & Coats
5
Gambar Produk
Label hasil prediksi model: Bags
6

Figure. 8 Result of proposed method

Figure 8 shows some of the results of the proposed method with different categories: dresses, jewelry, t-shirts, hats, jackets & coats, and bags. The experiment shows the success of the system in classifying according to the right category. Figure 9 shows that SVM cannot distinguish between knitwear and jackets, but the proposed methods correctly predict the categories. This category has human noise greatly but the Inception V3 model can differentiate between knitwear. This concludes the model can classify data more correctly because the deep learning model can extract more information using the architecture layers.


Label : Knitwear

Label : Scarves

Figure. 9 prediction results of SVM and Inception V3 classification

The limitation of this research is model required high dimension images to obtain the details of the images (this paper using 400×400). This resolution is slower and less efficient in computer performance. The main point of image for classification also need to crop to get important features to avoid of mislead incorrect categories. Example the picture of a person wearing a belt identified as a t-shirt or jacket, because the proportion of clothes is bigger in the picture.

V. Conclusion

This paper proposes Inception V3 for classification in fashion image dataset of 16 categories in five global famous fashion brands. The methodology of image processing was augmentation image with rotated, sheared, zoomed, and translated vertically and horizontally of image dataset. Total 32 of image augmented was used from one image dataset. Based on the results of the analysis the Inception V3 model can classified fashion image dataset correctly with evaluation score 92.85% using F1-score and 92.86% in accuracy. In addition, classification using the Inception v3 model can distinguish tops from knitwear categories correctly compared to other comparison method: Naïve Bayes, Support Vector Machine, and XGBoost. The future works of this paper was improving more categories for image classification. This paper process fashion categories for teenager and adults from five of famous brand of fashion. Added a category of children clothing, more brand of fashion or culture of different country will be challenging task to classified.

Corresponding author:
Maryamah,
Faculty of Advanced Technology and Multidiscipline
Universitas Airlangga
Email: maryamah@ftmm.unair.ac.id

References

[1] M. Shajini, A. Ramanan, “A knowledge-sharing semi-supervised approach for fashion clothes classification and attribute prediction,” *The Visual Computer*, vol. 38, no. 11, pp. 3551–3561, 2022, doi: 10.1007/s00371-021-02178-3.

[2] K. Bilińska-Reformat, A. Dewalska-Opitek, “E-commerce as the predominant business model of fast fashion retailers in the era of global COVID 19 pandemics,” *Procedia Computer Science*, vol. 192, pp. 2479–2490, 2021.

[3] J. Xie, C. Youn, “How the Luxury Fashion Brands Adjust to Deal with the COVID-19,” *International Journal of Costume and Fashion.*, vol. 20, no. 2, pp. 50–60, 2020, doi: 10.7233/ijcf.2020.20.2.050.

[4] A. Vijayaraj, P. T. Vasanth Raj, R. Jebakumar, P. Gururama Senthilvel, N. Kumar, R. Suresh Kumar, and R. Dhanagopal, “Deep Learning Image Classification for Fashion Design,” *Wireless Communications and Mobile Computing*, pp. 1–13, 2022, doi: 10.1155/2022/7549397.

[5] Z Zhou, M Liu, W Deng, Y Wang, Z Zhu, "Clothing Image Classification with DenseNet201 Network and Optimized Regularized Random Vector Functional Link.", *Journal of Natural* Fibers, vol. 20, no. 1, 2023.

[6] M. Elleuch, A. Mezghani, M. Khemakhem, and M. Kherallah, “Clothing Classification Using Deep CNN Architecture Based on Transfer Learning,”, *19th International Conference on Hybrid Intelligent Systems*, 2019, pp. 240–248. doi: 10.1007/978-3-030-49336-3_24.

[7] F. Li, S. Kant, S. Araki, S. Bangera, and S. S. Shukla, “Neural Networks for Fashion Image Classification and Visual Search”, *Applied Machine Learning Term Paper*, 2020, doi: 10.48550/ARXIV.2005.08170.

[8] Y. Seo and K. Shin, “Hierarchical convolutional neural networks for fashion image classification,” *Expert systems with applications*, vol. 116, pp. 328–339, 2019, doi: 10.1016/j.eswa.2018.09.022.

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

[10] E Xhaferra, E Cina, and L Toti, "Classification of Standard FASHION MNIST Dataset Using Deep Learning Based CNN Algorithms.", *2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pp. 494-498, 2022

[11] Z Chen, M Lv, S Tian, and S Yin, “Fashion-MNIST Classification Based on CNN Image Recognition,” *Highlights in Science, Engineering and Technology*, vol. 34, pp 196–202, 2023.

[12] M. Swapna, Y. K. Sharma, dan B. M. G. Prasadh, “CNN architectures: Alex Net, Le Net, VGG, Google Net, Res Net,” *Int J Recent Technol Eng*, vol. 8, no. 6, hlm. 953–960, 2020.

Maryamah is a lecturer at Data Science Technology Study Program, Faculty of Advanced Technology and Multidiscipline Universitas Airlangga.

Najma Attaqiya Alya is a bachelor student at Data Science Technology Study Program, Faculty of Advanced Technology and Multidiscipline Universitas Airlangga.

Corresponding author:
Maryamah,
Faculty of Advanced Technology and Multidiscipline
Universitas Airlangga
Email: maryamah@ftmm.unair.ac.id

Muhammad Hanif Sudibyo is a bachelor student at Data Science Technology Study Program, Faculty of Advanced Technology and Multidiscipline Universitas Airlangga.

Ergidya Liviani is a bachelor student at Data Science Technology Study Program, Faculty of Advanced Technology and Multidiscipline Universitas Airlangga.

Razin Isyraq Thirafi is a bachelor student at Data Science Technology Study Program, Faculty of Advanced Technology and Multidiscipline Universitas Airlangga.