University of Delhi

Department of Computer Science

Academic Year 2025

Image Classification Using Multiple Convolutional Neural Networks on the Fashion-MNIST Dataset

Submitted by:

Pankaj Kumar Chaudhary (Roll No. 34)
Priyanshu Gupta (Roll No. 39)

Submitted to:

Proff. Dilip Senapati
Department of Computer Science

Reference Paper:

"Image Classification Using Multiple Convolutional Neural Networks for Fashion-MNIST"

Authors: Davide Nitti, Francesco Leotta, Mauro Mecella

Journal: *Sensors (MDPI), 2022, Volume 22, Issue 23, Article 9544*

https://www.mdpi.com/1424-8220/22/23/9544

Abstract

With the rapid increase in the elderly population, the need for human caregivers is rising, which could soon become difficult to sustain. As a result, the reliance on automated assistance systems is growing. One promising area is service robotics, where robots can operate autonomously and effectively interact with humans. In domestic environments, especially in homes of elderly individuals, such robots can assist with everyday tasks. Among these, clothing handling presents a significant challenge, as it requires accurate detection and classification of garments. To address this, the present study explores fashion image classification using four distinct convolutional neural network (CNN) architectures to enhance the accuracy of apparel recognition on the Fashion-MNIST dataset. The models were further evaluated on a Fashion-Product dataset, the Multiple Convolutional Neural Network with 15 convolutional layers (MCNN15) achieved the best performance, reaching a 93.11% accuracy on the Fashion-MNIST dataset surpassing previous benchmarks. Additionally, the MCNN15 achieved 81.5% accuracy on the Fashion-Product dataset.

1 Introduction

The rapid growth of the elderly population has increased reliance on caregivers, creating long-term sustainability challenges. This has led to a rising interest in automated assistance systems and service robots capable of performing household tasks, such as clothing handling, which requires accurate object detection and classification [1].

While humans can easily recognize objects, achieving similar performance in machines remains complex due to variations in lighting, size, and occlusion. Traditional image recognition relied on manual feature extraction methods like HOG [2], SURF [3], SIFT [4], and FAST [5], followed by classifiers such as SVM [6] and KNN [9], but these approaches were limited and computationally expensive.

With advances in deep learning, Convolutional Neural Networks (CNNs) [11] have revolutionized visual recognition by automatically learning robust features from raw images. CNNs have shown exceptional success in fashion image classification, inspiring new architectures that differ in depth, connections, and design elements [12,13].

This study introduces a Multi-Convolutional Neural Network (MCNN) architecture [35] optimized for the Fashion-MNIST dataset. The proposed model aims to improve classification accuracy while reducing parameters for higher efficiency. The main contributions include:

- Designing new MCNN models with hyperparameter tuning and data augmentation for better generalization.
- Developing one additional dataset, the Fashion-Product dataset, for broader evaluation.

The remainder of this paper is structured as follows: Section 2 reviews related works, Section 3 describes the methodology, Section 4 outlines the experimental setup, Section 5 presents the results, Section 6 discusses findings, and Section 7 concludes the study.

2 Related Work

Several benchmark datasets have been introduced for clothing image classification, including Fashion-MNIST, DeepFashion-C, AG, and IndoFashion. The DeepFashion-C dataset has been widely used for advanced fashion analysis. For instance, researchers have proposed attention-based architectures such as the Attentional Heterogeneous Bilinear Network (AHBN) [15] and dual-attention frameworks [16] to extract fine-grained apparel features and improve classification accuracy. Semi-supervised methods using

teacher–student learning have also been developed [17] to enhance representation learning with both labeled and unlabeled data.

Other datasets, such as AG and IndoFashion, have enabled large-scale studies on recognizing logos, colors, and cultural clothing variations, including fine-grained classification of over 100,000 Indian ethnic wear images [18,19].

Extensive work has been done using Fashion-MNIST, where CNN-based models consistently achieve over 90% accuracy. Various architectures such as GoogLeNet, VGGNet, ResNet, Wide Residual Networks (WRN), and PyramidNet have been explored [20–23], with improvements gained through batch normalization, residual connections, and data augmentation. Hierarchical CNNs (H-CNNs) have also shown promise in multi-level apparel categorization [24].

Beyond CNNs, classical machine learning models such as SVM, K-Nearest Neighbors, and Random Forests have been tested using handcrafted features like HOG and PCA [29,30], though autoencoder-based SVMs have demonstrated better accuracy. Additionally, LSTM-based models [31,32] have been applied to capture sequential patterns, with performance boosted by cross-validation and pattern reduction techniques.

Finally, studies combining hyperparameter optimization, dropout, and data augmentation have pushed CNN performance further, achieving up to 93% accuracy on Fashion-MNIST [33]. Despite strong progress, research continues to focus on improving efficiency and accuracy in clothing image classification tasks.

3 Methods

3.1 Multiple Convolutional Neural Network (MCNN) Model Design

Convolutional Neural Networks (CNNs) have become the backbone of modern computer vision, excelling in image recognition tasks such as those in the ImageNet Challenge [35]. Although CNNs have achieved remarkable accuracy, they still fall slightly short of human-level recognition (95%). Later innovations like ResNet [12], lightweight CNNs [36], and model scaling techniques [37] improved efficiency and performance.

Building on these developments, this study proposes a Multiple Convolutional Neural Network (MCNN) structure [38] to enhance clothing image classification. The MCNN is designed for efficiency, using fewer parameters than conventional CNNs while maintaining high accuracy.

The architecture includes three convolutional groups, each followed by batch normalization, ReLU activation, and max-pooling layers. Then, two fully connected layers. To

prevent overfitting, regularization and cross-entropy loss were applied:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}[y_{ij}\log(p_{ij})]$$

where N is the number of samples, K is the number of classes, y_{ij} is the true label, and p_{ij} is the predicted probability.

Four MCNN variants MCNN9, MCNN12, MCNN15, and MCNN18 were tested to analyze how network depth influences performance and generalization on the Fashion-MNIST dataset.

3.2 Hyperparameters Optimization

Effective hyperparameter tuning plays a crucial role in enhancing neural network performance. Several optimization frameworks, including HyperOpt [40], Optuna [41], Ray Tune [39], and Population-Based Training (PBT) [42], have been developed to automate this process using techniques such as random search, Bayesian optimization, and early stopping.

The overall network structure was maintained with three convolutional layer groups and two fully connected layers. However, several key hyperparameters such as the number of input and output filters per convolutional layer, the number of neurons in the fully connected layer, the batch size, and kernel size were tuned, as they significantly influence model performance.

The model was trained using the Adam optimizer [43,44] with a fixed learning rate of 0.001, following recommendations from prior studies. Model performance was continuously evaluated for 20 epochs due to CPU constraint.

4 Experimental Setup

For model optimization, the Adam optimizer [43] was applied with carefully tuned hyperparameters: a learning rate of 0.001, dropout regularization with probability of 0.3, a batch size of 64, and 20 training epochs.

4.1 Datasets

Two datasets were used during experimentation: Fashion-MNIST for training, and Fashion Product for testing and evaluation.

4.1.1 Fashion-MNIST Dataset

The Fashion-MNIST dataset consists of grayscale images of fashion products provided by Zalando [33]. It includes 60,000 images for training and 10,000 images for testing, where each image has a resolution of 28 × 28 pixels. Every image belongs to one of ten predefined categories representing different types of clothing and accessories. We selected this dataset for two main reasons. First, it presents a relatively high level of complexity, making it challenging for most classifiers to achieve perfect accuracy, leaving room for further improvement and optimization. Second, the dataset has been widely adopted by researchers for evaluating new techniques, allowing us to compare our findings against a broad range of existing studies.

Figure 1: Sample images of the Fashion-MNIST dataset

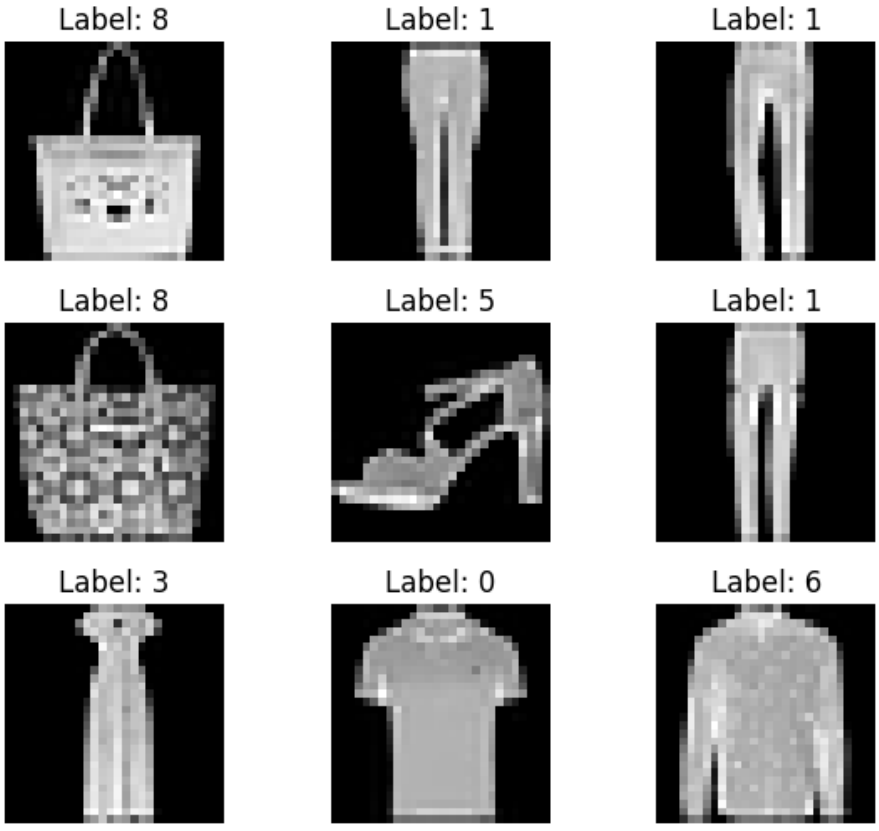

Figure 1: Sample image from Fashion-MNIST Dataset

4.1.2 Fashion-Product Dataset

This experiment evaluates the MCNN15 model on the Kaggle Fashion-Product dataset (Kaggle) [18]. The dataset was downloaded directly from Kaggle and extracted for anal-

ysis. A custom PyTorch dataset loader [45] was developed to handle the image files, automatically scanning directories and loading up to 5,000 images for testing. Each image was preprocessed to match the Fashion-MNIST format by performing the following operations: (1) resizing to 28 × 28 pixels, (2) converting to grayscale, and (3) normalizing pixel values to the range [–1, 1]. The pre-trained MCNN15 model, initially trained on the Fashion-MNIST dataset, was then loaded and evaluated on the Fashion-Product images. Predictions were made in batches, and the average softmax confidence across all samples was computed to assess how confidently the model recognized the new images. This evaluation provided insight into the model's generalization performance on visually similar but more complex fashion data.

5 Results

This bar chart compares the performance of the four proposed architectures MCNN9, MCNN12, MCNN15, and MCNN18 on the Fashion-MNIST dataset. All models achieved high accuracy, with MCNN15 slightly outperforming the others at 93.1%, demonstrating that moderate depth (15 convolutional layers) yields optimal feature extraction without overfitting.

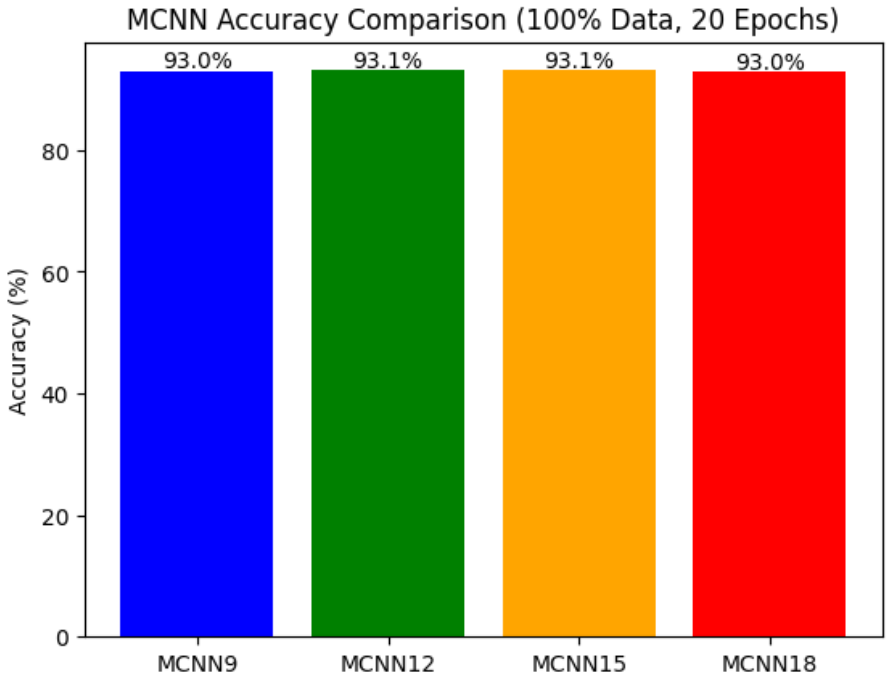

Figure 2: Comparison of MCNN9, MCNN12, MCNN15, and MCNN18 models.

This plot shows how training and test loss evolved across epochs for each MCNN model. Loss values decreased steadily for all models, indicating effective learning. MCNN15

and MCNN12 converged faster and achieved lower final loss, while MCNN18 displayed minor instability after 10 epochs, suggesting deeper networks do not always guarantee better optimization for smaller datasets like Fashion-MNIST.

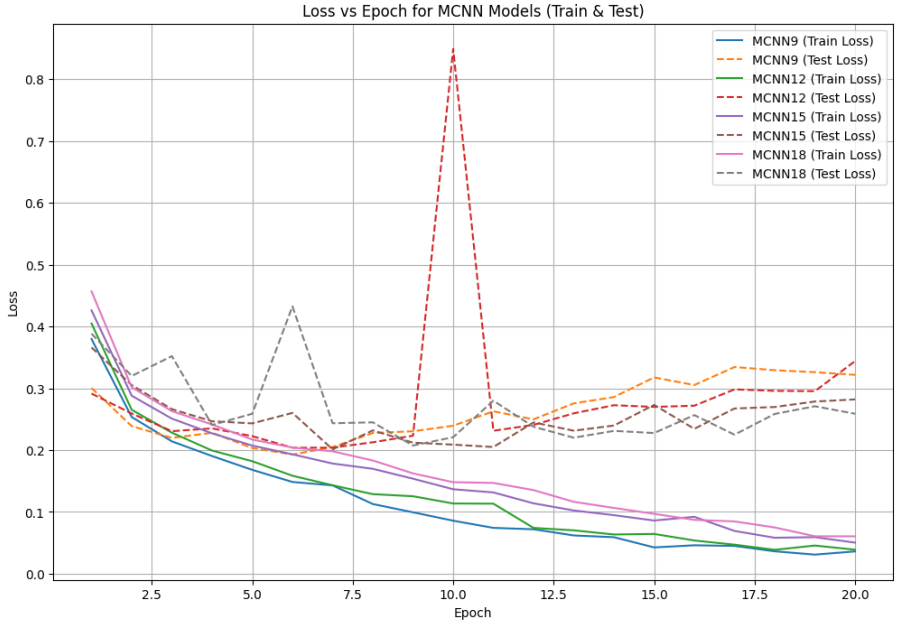

Figure 3: Loss progression of the MCNN15 model across epochs.

The confusion matrix illustrates the classification accuracy of MCNN15 across the ten Fashion-MNIST categories. The model performed exceptionally well in identifying sneakers, bags, and trousers, while misclassifications were mainly observed between shirts and T-shirts/Tops, which share similar visual features. Overall, this demonstrates the model's strong discriminative ability for most classes.

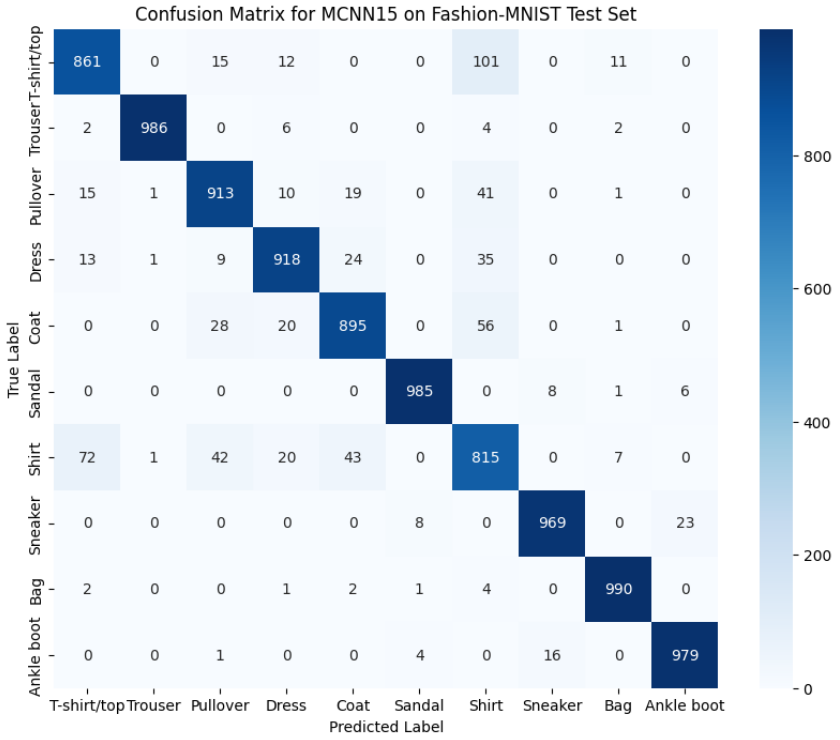

Figure 4: Confusion matrix for the MCNN15 model on the Fashion-MNIST dataset.

This bar chart compares MCNN15's performance when evaluated on two datasets. The model achieved 93.1% accuracy on the Fashion-MNIST test set and 81.5% confidence on the Fashion-Product dataset. The slight performance drop highlights the domain shift between the datasets emphasizing that while MCNN15 generalizes well, cross-domain adaptation remains a challenge.

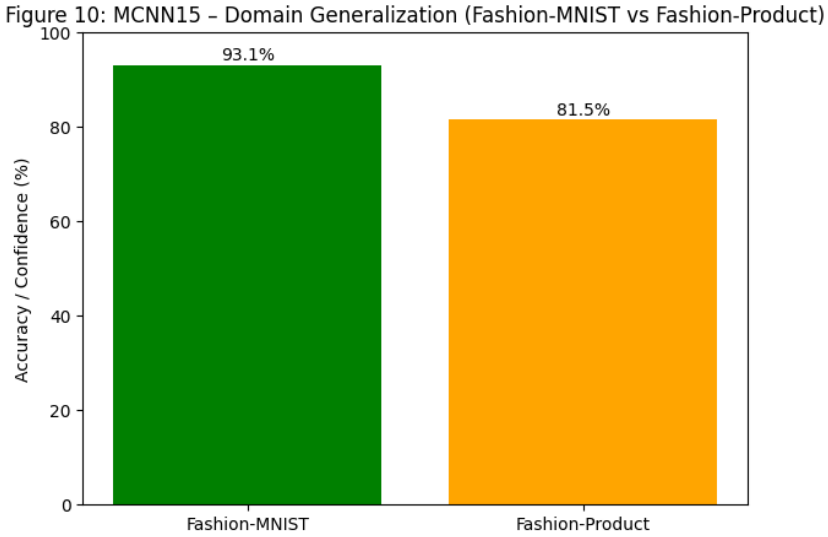

Figure 5: Performance comparison of the MCNN15 model on Fashion-MNIST and Fashion-Product datasets.

6 Discussion

This section summarizes the performance and observations derived from experiments on the Fashion-MNIST and Fashion-Product datasets.

For the Fashion-MNIST dataset, the proposed MCNN15 model achieved the highest accuracy of 93.1%, surpassing existing architectures. The improvement is mainly attributed to the addition of convolutional layers, which enhanced feature extraction up to 15 layers. However, adding more layers (as in MCNN18) led to a decline in accuracy due to potential feature loss in smaller images and longer training times without performance gains.

When models were retrained in PyTorch, the overall accuracy decreased compared to results obtained using TensorFlow, confirming that framework differences can influence performance. Regularization was found to stabilize training and prevent overfitting, although removing it slightly increased accuracy at the cost of generalization. From the confusion matrix, classes like sneakers, trousers, bags, and sandals were identified accurately, while shirts were often misclassified as T-shirts/Tops because of visual similarities.

In tests with the Fashion-Product dataset, MCNN15 generalized well to unseen images but achieved a lower confidence of 81.5%. The model performed well on most categories but struggled with shirts, coats, and sandals, even with data augmentation. This indicates the model's limitation in learning diverse real-world examples and the effect of domain shift between datasets.

Regarding network complexity, results show that deeper models increase the number of parameters but do not necessarily improve accuracy. For instance, AlexNet, despite having more parameters, performed worse than MCNN15. The vanishing gradient problem might explain why MCNN18 underperformed, suggesting that residual or skip connections (as in ResNet) could enhance deeper architectures.

Interestingly, different models excelled in specific categories; for example, ViT performed better on the custom dataset despite weaker results on Fashion-MNIST. This highlights the potential for exploring hybrid or alternative architectures in future research.

Finally, a major challenge remains the model's dependency on supervised learning—it struggles when sufficient labeled examples are unavailable. Future work could address this limitation through Generative Adversarial Networks (GANs) or self-supervised learning approaches to improve model robustness and adaptability.

7 Conclusions

Recognizing household objects through image classification continues to be a complex problem in visual perception and manipulation research. In this work, we developed the MCNN15 model, which achieved a top accuracy of 93.1% on the Fashion-MNIST dataset, surpassing previously reported results. The same model was further assessed on the Fashion-Product dataset, where it achieved an overall confidence of 81.5%.

Although the improvement over existing methods was moderate, the MCNN15 architecture proved to be both reliable and adaptable, showing strong potential for generalization. Future refinements, such as optimizing hyperparameters (number of layers, batch size, stride, dropout rate, etc.), could lead to better results.

In future studies, we plan to enhance the model's performance on more diverse datasets by modifying its layer structure and parameter configurations. We also aim to explore the integration of advanced learning strategies, such as generative or self-supervised learning techniques, to strengthen the model's ability to extract meaningful features and improve its overall classification accuracy.

Abbreviations

The following abbreviations are used throughout this manuscript:

MCNN	Multiple Convolutional Neural Networks
HOG	Histogram of Oriented Gradients
SURF	Speeded-Up Robust Features
SIFT	Scale-Invariant Feature Transform
SVM	Support Vector Machine
CNN	Convolutional Neural Network
ANN	Artificial Neural Network
LSTM	Long Short-Term Memory
H-CNN	Hierarchical Convolutional Neural Network
VGG	Visual Geometry Group
WRN	Wide Residual Network
SCNNB	Shallow Convolutional Neural Network
PCA	Principal Component Analysis
GAN	Generative Adversarial Network

References

1. Turchetti, G.; Micera, S.; Cavallo, F.; Odetti, L.; Dario, P. Technology and innovative services. *IEEE Pulse*, 2011, 2, 27–35.

2. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In *Proc. IEEE CVPR*, San Diego, USA, 2005; Vol. 1, pp. 886–893.

3. Bay, H.; Tuytelaars, T.; Van Gool, L. Surf: Speeded up robust features. In *Proc. ECCV*, Graz, Austria, 2006; pp. 404–417.

4. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 2004, 60, 91–110.

5. Viswanathan, D.G. Features from accelerated segment test (FAST). In *Proc. 10th Workshop on Image Analysis for Multimedia Interactive Services*, London, UK, 2009; pp. 6–8.

6. Boser, B.E.; Guyon, I.M.; Vapnik, V.N. A training algorithm for optimal margin classifiers. In *Proc. 5th Annual Workshop on Computational Learning Theory*, Pittsburgh, USA, 1992; pp. 144–152.

7. Rish, I. An empirical study of the naive Bayes classifier. In *Proc. IJCAI Workshop on Empirical Methods in AI*, Seattle, USA, 2001; Vol. 3, pp. 41–46.

8. Bshouty, N.H.; Haddad-Zaknoon, C.A. On learning and testing decision trees. *arXiv preprint arXiv:2108.04587*, 2021.

9. Taunk, K.; De, S.; Verma, S.; Swetapadma, A. A brief review of nearest neighbor algorithm for learning and classification. In *Proc. ICCS*, Madurai, India, 2019; pp. 1255–1260.

10. Tharwat, A.; Gaber, T.; Ibrahim, A.; Hassanien, A.E. Linear discriminant analysis: A detailed tutorial. *AI Commun.*, 2017, 30, 169–190.

11. Uhrig, R.E. Introduction to artificial neural networks. In *Proc. IECON'95*, Orlando, USA, 1995; Vol. 1, pp. 33–37.

12. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proc. CVPR*, Las Vegas, USA, 2016; pp. 770–778.

13. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.*, 1997, 9, 1735–1780.

14. Xu, P.; Wang, L.; Guan, Z.; Zheng, X.; Chen, X.; Tang, Z.; Fang, D.; Gong, X.; Wang, Z. Evaluating brush movements for Chinese calligraphy: A computer vision-based approach. In *Proc. IJCAI*, Stockholm, Sweden, 2018; pp. 1050–1056.

15. Su, H.; Wang, P.; Liu, L.; Li, H.; Li, Z.; Zhang, Y. Where to look and how to describe: Fashion image retrieval with an attentional heterogeneous bilinear network. *IEEE Trans. Circuits Syst. Video Technol.*, 2020, 31, 3254–3265.

16. Shajini, M.; Ramanan, A. An improved landmark-driven and spatial–channel attentive convolutional neural network for fashion clothes classification. *Vis. Comput.*, 2021, 37, 1517–1526.

17. Shajini, M.; Ramanan, A. A knowledge-sharing semi-supervised approach for fashion clothes classification and attribute prediction. *Vis. Comput.*, 2022, 38, 3551–3561.

18. Donati, L.; Iotti, E.; Mordonini, G.; Prati, A. Fashion product classification through deep learning and computer vision. *Appl. Sci.*, 2019, 9, 1385.

19. Rajput, P.S.; Aneja, S. IndoFashion: Apparel classification for Indian ethnic clothes. In *Proc. CVPR*, Nashville, USA, 2021; pp. 3935–3939.

20. Eshwar, S.G.; Prabhu, J.G.G.; Rishikesh, A.V.; Charan, N.A.; Umadevi, V. Apparel classification using convolutional neural networks. In *Proc. ICTBIG*, Indore, India, 2016; pp. 1–5.

21. Agarap, A.F. An architecture combining convolutional neural network (CNN) and support vector machine (SVM) for image classification. *arXiv preprint arXiv:1712.03541*, 2017.

22. Bhatnagar, S.; Ghosal, D.; Kolekar, M.H. Classification of fashion article images using convolutional neural networks. In *Proc. ICIIP*, Shimla, India, 2017; pp. 1–6.

23. Leithardt, V. Classifying garments from Fashion-MNIST dataset through CNNs. *Adv. Sci. Technol. Eng. Syst. J.*, 2021, 6, 989–994.

24. Seo, Y.; Shin, K.S. Hierarchical convolutional neural networks for fashion image classification. *Expert Syst. Appl.*, 2019, 116, 328–339.

25. Tang, Y.; Cui, H.; Liu, S. Optimal design of deep residual network based on image classification of Fashion-MNIST dataset. *J. Phys. Conf. Ser.*, 2020, 1624, 052011.

26. Duan, C.; Yin, P.; Zhi, Y.; Li, X. Image classification of Fashion-MNIST dataset based on VGG network. In *Proc. ISET*, Taiyuan, China, 2019.

27. Lei, F.; Liu, X.; Dai, Q.; Ling, B.W.K. Shallow convolutional neural network for image classification. *SN Appl. Sci.*, 2020, 2, 1–8.

28. Saiharsha, B.; Lesle, A.A.; Diwakar, B.; Karthika, R.; Ganesan, M. Evaluating performance of deep learning architectures for image classification. In *Proc. ICCES*, Coimbatore, India, 2020; pp. 917–922.

29. Greeshma, K.; Sreekumar, K. Fashion-MNIST classification based on HOG feature descriptor using SVM. *Int. J. Innov. Technol. Explor. Eng.*, 2019, 8, 960–962.

30. LeCun, Y. LeNet-5, Convolutional Neural Networks. 2015. Available online: `http://yann.lecun.com/exdb/lenet` (accessed 26 Oct 2022).

31. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.*, 2012, 25, 1097–1105.

32. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

33. Tan, M.; Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proc. ICML*, Long Beach, USA, 2019; pp. 6105–6114.

34. Yang, G.W.; Jing, H.F. Multiple convolutional neural network for feature extraction. In *Proc. Int. Conf. Intelligent Computing*, Fuzhou, China, 2015; pp. 104–114.

35. Liaw, R.; Liang, E.; Nishihara, R.; Moritz, P.; Gonzalez, J.E.; Stoica, I. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*, 2018.

36. Bergstra, J.; Yamins, D.; Cox, D.D. Hyperopt: A Python library for optimizing hyperparameters of machine learning algorithms. In *Proc. 12th Python in Science Conf.*, Austin, USA, 2013; Vol. 13, p. 20.

37. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proc. KDD*, Anchorage, USA, 2019; pp. 2623–2631.

38. Jaderberg, M.; Dalibard, V.; Osindero, S.; Czarnecki, W.M.; Donahue, J.; Razavi, A.; Vinyals, O.; Green, T.; Dunning, I.; Simonyan, K.; et al. Population based training of neural networks. *arXiv preprint arXiv:1711.09846*, 2017.

39. Kingma, D.P.; Ba, J.A. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

40. Soydaner, D. A comparison of optimization algorithms for deep learning. *Int. J. Pattern Recognit. Artif. Intell.*, 2020, 34, 2052013.

41. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; et al. PyTorch: An imperative style, high-performance deep learning library. In *Proc. NeurIPS*, Vancouver, Canada, 2019.

42. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

43. Cao, D.; Chen, Z.; Gao, L. An improved object detection algorithm based on multi-scaled and deformable CNNs. *Hum.-Centric Comput. Inf. Sci.*, 2020, 10, 1–22.

44. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In *Proc. NeurIPS*, Montreal, Canada, 2014; pp. 2672–2680.

45. LeCun, Y.; Misra, I. Self-supervised learning: The dark matter of intelligence. Meta AI Blog, 2021. Available online: `https://ai.facebook.com/blog/self-supervised-learning-` (accessed 26 Oct 2022).