International Journal of Engineering Science Invention Research & Development; Vol. 12, Issue 8, February 2026
www.ijesird.com, E-ISSN: 2349-6185

FASHION-MNIST IMAGE CLASSIFICATION USING CONVOLUTIONAL NEURAL NETWORKS

Tanvi Gande[1], Chanchal pathade[2], Dinesh Barode[3],Paravin Dhole[4], Bharti Gawali[5]
[1-5]*Dr. Babasaheb Ambedkar Marathwada University, Chh. Sambhjinagar, Maharashtra, India, Pin-431001*

***Abstract*-The fast development of internet fashion retail stores like Meesho, Flipkart, and Amazon has generated a high need for"robotic" with "automated" image classification techniques. As the number of clothing products posted by sellers and seen by customers every day is in the millions, proper product categorization and search have become necessary tools to enhance user interaction and business activity. Conventional machine learning methods, like SVM and KNN, do not usually cope with variations of images on large scales and intricate visual patterns. To address these issues, this research paper explores the applicability of Convolutional Neural Networks (CNNs) to fashion image recognition. Based on common datasets like Fashion-MNIST, our study analyzes various classification methods and shows that CNNs are better when it comes to the ability to extract deep visual cues such as texture, shape, and pattern. The first CNN model was overfitted and a test accuracy of 0.91 was attained. The enhanced model (adding dropout layers to the architecture) performed better (at about 0.93 test accuracy) on generalization. The results affirm that the CNN-based model can greatly improve fashion item recognition and can be directly integrated with the e-commerce platforms to increase accuracy in product search, seller productivity, and customer satisfaction. The study will be useful in providing hands-on research in using deep learning in actual online retail systems.**

***Keywords*-Fashion-MNIST, CNN, Product Tagging, Visual Recognition, Fashion Technology.**

I. INTRODUCTION

The fashion business in the world is changing fast with the evolution of digital technologies and web-based markets. In modern scenarios, it is possible to use fashion products on websites like Meesho, Flipkart, Amazon, and other e-commerce platforms, where the fashion product is offered in convenient user interfaces and easy search systems, as well as on pages with personalized recommendations of products. With the growth of online shopping, several million online shopping users each day find it more convenient to find new fashions, compare prices, and receive a new garment through online shopping. The shift to online retail has also been advantageous to small and medium corporations, allowing local sellers and resellers to expand their reach thanks to the lack of the need to have brick-and-mortar stores. Some of these platforms, like Meesho in particular, have enabled individuals who operate businesses at home by bringing them simple product-listing proxies, affordable marketing, and fast delivery systems.

To enhance user interaction and beneficence of sellers, modern e-commerce applications use fast and smart search engines. These include content-based filtering, collaborative filtering, and deep learning-based image search that helps in the identification of suitable products in a matter of seconds. However, some previous machine-learning models, like traditional classification, support vector machines, and the use of k-Nearest Neighbors (KNN), have previously failed to deal with large datasets, changes in clothing styles, and more complex visual patterns. These constraints have driven the researchers to explore more superior techniques that can process large-scale image data with greater strength and precision [2] [5].

To address these issues, our study will focus on the use of the Convolutional Neural Networks (CNNs) to classify fashion images. CNNs are especially suitable in terms of the identification of patterns, textures,colors, and shapes of clothing images. In comparison to the conventional DNNs, CNNs use less preprocessing, are effective at processing large volumes of data, and are rated significantly higher in accuracy in image-related tasks. Using Octavius-created databases, specifically Fashion-MNIST, Fashion Products Dataset, and DeepFashion, CNN models are able to identify deep visual representations and classify clothing items with even higher accuracy [6] [7] [10].

This paper allows comparing different classification algorithms and proving that CNNs are superior to traditional models when it comes to fashion recognition. The applied methodology helps Internet resources, merchants, and companies to improve the search for products, the satisfaction of

customers, and the sales results in general. Our effort aims at making contributions to workable e-commerce applications in which proper and speedy categorization of fashion items is the key element in user interaction and corporate advancement [1] - [8].

II. Literature Review

The increasing scale of online fashion retail has created a strong demand for automated and accurate image-based product classification. Early studies relied on traditional machine learning methods such as Support Vector Machines and k-Nearest Neighbors, which depended on handcrafted visual features and showed limited robustness when applied to complex apparel images. These limitations motivated the transition toward deep learning–based approaches, particularly Convolutional Neural Networks (CNNs), which are capable of learning hierarchical visual representations directly from data.

A major milestone in fashion image classification research was the introduction of the Fashion-MNIST dataset by Xiao et al. [1]. Designed as a more challenging alternative to the MNIST benchmark, Fashion-MNIST contains grayscale images of clothing items with overlapping visual characteristics, making it well suited for evaluating CNN architectures. Since its release, the dataset has been widely adopted for benchmarking deep learning models in fashion recognition tasks. In parallel, large-scale datasets such as DeepFashion [9] and Clothing1M [10] have enabled the study of fashion classification under real-world conditions, including large intra-class variation and noisy labels.

CNNs, popularized through foundational work by LeCun et al. [2], have consistently demonstrated superior performance over traditional methods in visual recognition tasks. Their ability to automatically extract spatial features such as edges, textures, and shapes has made them particularly effective for apparel image analysis. However, as CNN architectures became deeper, overfitting emerged as a critical challenge, especially for datasets with visually similar classes. Srivastava et al. [3] addressed this issue through Dropout regularization, which has since become a standard technique for improving generalization in deep networks.

To further improve performance, researchers explored deeper architectures. Residual Networks (ResNet), introduced by He et al. [4], enabled the training of very deep CNNs by mitigating vanishing gradient problems. While highly effective, such architectures often involve high computational cost. To address efficiency concerns, lightweight models such as MobileNet [5] and MobileNetV2 [6] were proposed, significantly reducing parameter count and inference time while maintaining competitive accuracy. These models are particularly relevant for deployment in real-time and resource-constrained e-commerce environments.

More recent architectures have focused on optimizing the trade-off between accuracy and efficiency. EfficientNet, proposed by Tan and Le [7], introduced compound scaling to balance network depth, width, and resolution, achieving strong performance with fewer parameters. DenseNet, introduced by Huang et al. [8], improved feature reuse and gradient flow, reducing redundancy and overfitting in deep networks. Inception-based architectures [11] further enhanced performance by capturing multi-scale visual patterns, which is beneficial for handling variation in clothing shape and size.

Beyond standard CNN architectures, alternative strategies have been explored to enhance representation learning and interpretability. Capsule Networks proposed by Sabour et al. [12] aimed to preserve spatial relationships between features, although their computational complexity has limited widespread adoption. Zhou et al. [13] introduced Class Activation Mapping (CAM), enabling visualization of discriminative regions used by CNNs, which is particularly valuable for analyzing misclassification among visually similar apparel categories.

A. Research Gap and Motivation

Despite the extensive body of work, several gaps remain. Many high-performing models rely on deep or pretrained architectures that increase complexity and limit practical deployment. In addition, several studies emphasize overall accuracy without systematically analyzing overfitting behavior, class-

wise performance, or confusion among visually similar clothing categories. Furthermore, although regularization techniques such as Dropout are widely acknowledged, their impact is often not explicitly evaluated in compact CNN architectures trained from scratch.

B. Motivation of the Present Study

Motivated by these gaps, the present study proposes a compact CNN architecture trained from scratch on the Fashion-MNIST dataset. The study systematically evaluates the effect of Dropout on generalization and provides detailed class-wise performance analysis and confusion matrix interpretation. By balancing architectural simplicity, robustness, and interpretability, the proposed approach aims to offer a practical and effective solution for fashion image classification in real-world e-commerce applications

.

Table 1. Comparative Analysis of Existing Fashion Image Classification Techniques

Sr. No.	Authors	Year	Technique / Model	Dataset Used	Key Outcome
1	Xiao et al.	2017	CNN Benchmark Dataset	Fashion-MNIST	Established a challenging benchmark for fashion image classification
2	LeCun et al.	2015	CNN (Foundational)	MNIST / Vision Tasks	Demonstrated effectiveness of CNNs for visual recognition
3	Srivastava et al.	2014	CNN + Dropout	Image Benchmarks	Reduced overfitting and improved generalization
4	He et al.	2016	ResNet-50	ImageNet	Enabled training of very deep CNNs
5	Howard et al.	2017	MobileNet	ImageNet	Introduced lightweight CNNs for efficient inference
6	Sandler et al.	2018	MobileNetV2	ImageNet	Improved efficiency using inverted residuals
7	Tan & Le	2019	EfficientNet	ImageNet	Optimized CNN scaling with fewer parameters
8	Liu et al.	2016	CNN-Based Fashion Recognition	DeepFashion	Large-scale fashion dataset with rich annotations
9	Xiao et al.	2015	Noise-Robust CNN	Clothing1M	Addressed noisy labels in real-world fashion data
10	Dosovitskiy et al.	2021	Vision Transformer (ViT)	ImageNet	Introduced transformer-based image classification
11	Liu et al.	2021	Swin Transformer	ImageNet	Hierarchical transformer with shifted windows
12	Liu et al.	2022	ConvNeXt	ImageNet	Modern CNN redesigned using transformer principles
13	Khan et al.	2022	Vision Transformers Survey	Vision Datasets	Comprehensive review of transformer-based vision models

III. METHODOLOGY

The methodology used in this research has four major elements, which include dataset preparation, data preprocessing, model architecture development, and model training and evaluation. All of the phases have been carefully determined to help create a powerful Convolutional Neural Network (CNN) that can effectively recognize images in the Fashion-MNIST data set.

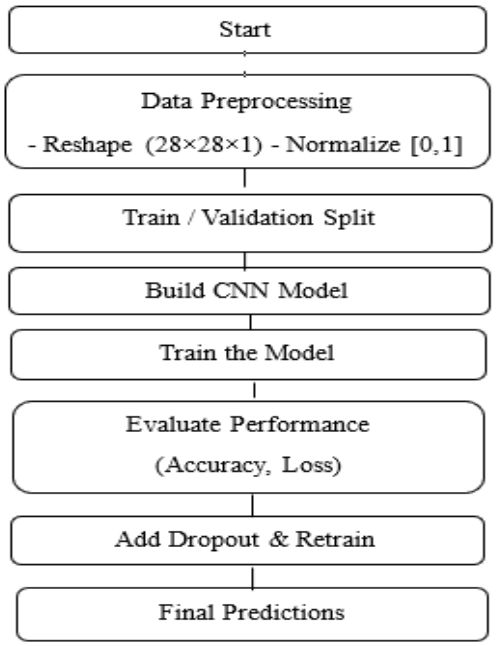

Figure 1. Overall Methodology Workflow for Fashion Image Classification Using CNN

A. *Dataset Preparation*

The Fashion-MNIST dataset consists of 70,000 grayscale images belonging to 10 distinct fashion categories. The dataset is divided into 60,000 training samples and 10,000 testing samples. Each image has a spatial resolution of 28 × 28 pixels and is represented as a 784-dimensional feature vector corresponding to grayscale intensity values ranging from 0 to 255. The dataset is intrinsically balanced, with each class contributing approximately 10 percent of the total samples, which helps mitigate class imbalance issues. In the CSV format used in this study, the first column represents the class label, while the remaining 784 columns correspond to pixel intensity values [1].

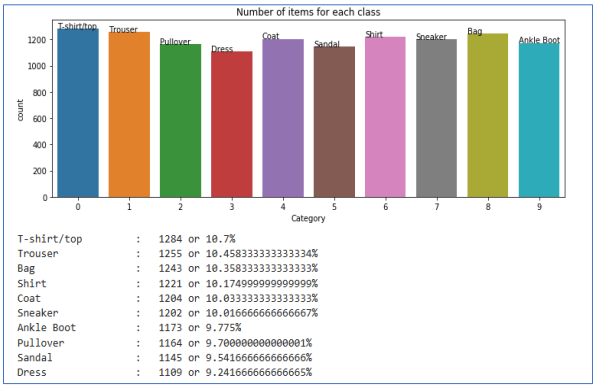

Figure 2. Class distribution of the Fashion-MNIST dataset showing the number of samples in each category.

B. *Data Preprocessing*

Prior to model training, several preprocessing steps were applied to standardize the input data and improve learning efficiency. The original 784-dimensional pixel vectors were reshaped into 28 × 28 × 1 tensors to make them compatible with convolutional layers. Pixel intensity values were normalized to the [0, 1] range by dividing each value by 255, which helps stabilize gradient updates and accelerates convergence during training.

The training dataset was further split into 48,000 samples for training and 12,000 samples for validation to enable performance monitoring and prevent overfitting. To support multi-class classification, the categorical class labels were converted into one-hot encoded vectors. These preprocessing steps ensured consistent data representation and improved the effectiveness of feature learning in the CNN model.

C. *CNN Model Architecture*

Convolutional Neural Networks (CNNs) were selected for this study due to their proven effectiveness in image classification tasks, particularly their ability to automatically learn spatial features such as edges, textures, and shapes directly from raw pixel data [2]. Unlike traditional machine learning methods, CNNs require minimal manual feature engineering and are well suited for handling high-dimensional image inputs.

The proposed CNN architecture consists of three convolutional blocks, each designed to progressively extract higher-level feature representations. The first convolutional layer employs 32 filters of size 3 × 3 with ReLU activation, followed by a 2 × 2 max-pooling layer to reduce spatial dimensions. This layer primarily captures low-level features such as edges and simple textures. The second convolutional layer increases the number of filters to 64, enabling the network to learn more complex patterns and local structures. The third convolutional layer further expands the feature depth to 128 filters, allowing the model to extract higher-level semantic features from the input images. Stacking multiple convolutional layers in this manner enables

hierarchical feature learning, where low-level features are progressively combined into more abstract representations [2].

Following the convolutional blocks, the extracted feature maps are flattened and passed to a fully connected dense layer with 128 neurons using ReLU activation. This layer integrates the learned features and prepares them for final classification. The output layer consists of 10 neurons with Softmax activation, corresponding to the ten fashion categories in the Fashion-MNIST dataset, enabling multi-class probability estimation.

To mitigate overfitting and improve generalization, Dropout layers were incorporated after selected convolutional blocks and after the dense layer. Dropout randomly deactivates a fraction of neurons during training, preventing co-adaptation of features and encouraging the network to learn more robust representations [3]. The inclusion of Dropout is particularly important for datasets such as Fashion-MNIST, where certain classes exhibit strong visual similarity.

While deeper architectures such as ResNet and EfficientNet achieve high performance through increased depth and complex scaling strategies [4], [7], the present study adopts a compact CNN architecture trained from scratch. This design choice balances classification accuracy and computational efficiency, making the model more suitable for practical deployment in real-world e-commerce systems.

D. Model Training

The CNN was made to undergo 50 epochs based on the Adamoptimizer and categorical cross-entropy as the loss function. Classification accuracy was used as the primary evaluation metric. Two settings were studied, i.e., the model without dropout with about 99.5% classical training accuracy and 91.5% classical validation accuracy, which corresponds to a high degree of overfitting. A model using Dropout had training accuracy and validation accuracy of about 92 and 93 percent [3], respectively; the Dropout dramatically minimized overfitting and maximized generalization.

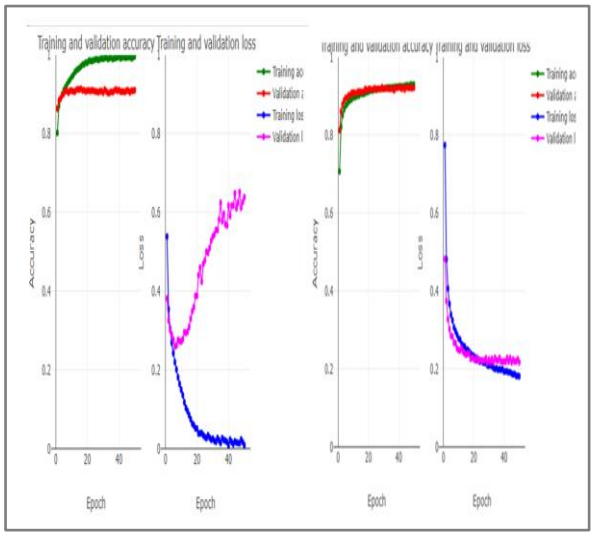

Figure 3. Training and validation accuracy of the CNN model without dropout, showing overfitting.

E. Model Evaluation

The final evaluation on the 10,000-image test set resulted in an overall accuracy of 92.56%, with 9,256 images correctly classified and 744 misclassified. The class-wise analysis revealed notable variations in performance. Categories such as trouser,sandal,bag,sneaker, and ankle boot achieved the highest accuracy due to their distinct structural and visual characteristics, which made them easier for the CNN to recognize.

In contrast, classes with strong visual similarities,most notably Shirt, Pullover, and Coat,showed lower accuracy and were frequently misclassified. These categories share overlapping textures, shapes, and outlines, making it challenging for the model to differentiate between them, especially given the limited 28×28 grayscale resolution of the Fashion-MNIST images. This indicates that while the CNN successfully captures discriminative features for most fashion categories, performance declines when fine-grained visual distinctions are required [13].

To further understand model behavior, visual samples of correct and incorrect predictions were analyzed. The correctly classified images (displayed in green) demonstrate strong recognition of distinct apparel types, whereas the misclassified images (displayed in red) highlight common areas of confusion, particularly between visually similar classes. This visual evidence supports the quantitative findings and provides deeper insight

into the strengths and limitations of the CNN model.

Class	Precision	Recall	F1-Score	Support
T-shirt/top	0.88	0.87	0.88	1000
Trousers	0.99	0.99	0.99	1000
Pullover	0.93	0.83	0.88	1000
Dress	0.93	0.95	0.94	1000
Coat	0.87	0.91	0.89	1000
Sandal	0.99	0.98	0.99	1000
Shirt	0.76	0.80	0.78	1000
Sneaker	0.96	0.96	0.96	1000
Bag	0.99	0.99	0.99	1000
Ankle Boot	0.96	0.97	0.97	1000

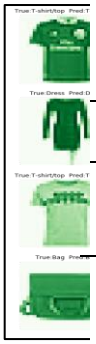

Figure 4. Correctly Classified Fashion-MNIST Images (Green)

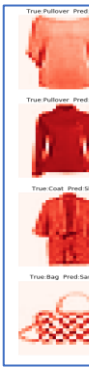

Figure 5. Misclassified Fashion-MNIST Images (Red)

IV. Results& Discussion

The improved Convolutional Neural Network (CNN) model with Dropout was evaluated on the Fashion-MNIST test dataset of 10,000 images. The model achieved a test accuracy of 0.9256 (93%), correctly classifying 9,256 images and misclassifying 744 images. Class-wise performance metrics indicated that categories such as Trouser, Sandal, Sneaker, Bag, and Ankle Boot achieved the highest precision and recall, all above 0.95. These classes have distinct shapes and textures, making them easier to identify. In contrast, Shirt, Pullover, and Coat showed lower performance, with F1-scores between 0.78 and 0.89, likely due to visual similarities and overlapping features. The confusion matrix further confirmed that most misclassifications occurred between Shirt (Class 6) and Pullover (Class 2). Overall, the results demonstrate that the CNN model effectively learns discriminative features and performs strongly on fashion image classification.

Table 2. Overall Test Performance of the Proposed CNN Model

Test Accuracy	92.56%
Correct Predictions	9,256
Incorrect Predictions	744

Table 3. Class-Wise Precision, Recall, and F1-Score for Fashion-MNIST Dataset

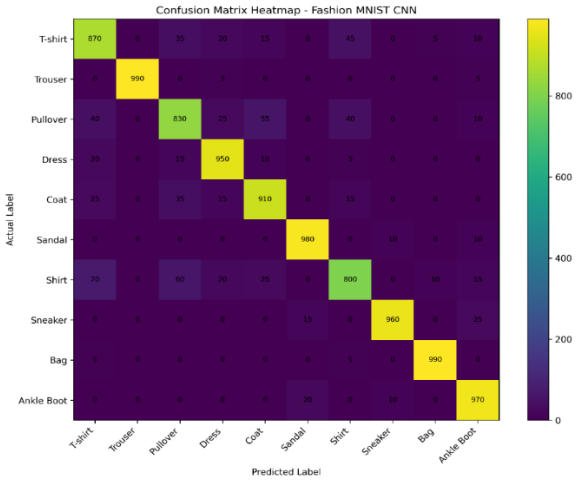

Figure 6. Confusion Matrix Heatmap of the Proposed CNN Model on the Fashion-MNIST Test Dataset

The experimental results demonstrate that the incorporation of Dropout significantly improved the generalization capability of the CNN model used for Fashion-MNIST classification. The initial model, although capable of achieving a very high training accuracy of approximately 99%, exhibited clear signs of overfitting, as reflected by a considerably lower validation accuracy of around 91%. After introducing Dropout at multiple stages of the network, the model's validation accuracy increased to nearly 93%, indicating that regularization effectively reduced memorization of training data and enabled the model to learn more robust features. The class-wise evaluation further reveals that categories with distinct visual patternssuch as Trouser, Sandal, Sneaker, Bag, and Ankle Bootachieved near-perfect accuracy, while classes with subtle differences, particularly Shirt and Pullover, remained challenging. These

misclassifications are largely attributed to the limited spatial resolution of 28×28 grayscale images and the high similarity between certain clothing types. Nevertheless, the improved model demonstrates strong overall performance and aligns with findings from previous research that CNNs, when properly regularized, perform effectively on fashion image datasets. Future work could explore strategies such as data augmentation, deeper architectures, transfer learning, or attention mechanisms to address classification weaknesses and further enhance accuracy.

V. CONCLUSION

The findings of this research reinforce the effectiveness of convolutional neural networks as a reliable and robust framework for fashion image classification. Using the FashionMNIST dataset, the study demonstrated that a well-designed CNN is capable of learning meaningful visual representations even from low-resolution grayscale images. Although the baseline model achieved strong performance on the training set, its inability to generalize revealed considerable overfitting. The integration of dropout layers substantially improved generalization and stabilized learning, resulting in an enhanced test accuracy of nearly 93%. The class-wise analysis further highlighted the strengths and limitations of the model: while highly distinguishable categories were classified with near-perfect accuracy, classes with subtle structural similarities,particularly Shirt and Pullover,remained challenging. These observations align with existing literature, emphasizing the importance of robust feature extraction for fine-grained fashion recognition. Overall, the study provides strong empirical evidence that CNN-based approaches are highly suitable for real-world e-commerce applications, where accurate and automated product classification is essential. Future research may explore advanced architectures, transfer learning, or attention-based mechanisms to address misclassification challenges and further elevate model performance.

REFERENCES

[1] H. Xiao, K. Rasul, and R. Vollgraf, “Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms,” arXiv preprint arXiv:1708.07747, 2017. doi: 10.48550/arXiv.1708.07747

[2] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015. doi: 10.1038/nature14539

[3] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929–1958, 2014.

[4] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.doi: 10.1109/CVPR.2016.90

[5] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “MobileNets: Efficient convolutional neural networks for mobile vision applications,” arXiv preprint arXiv:1704.04861, 2017.

[6] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510–4520.doi: 10.1109/CVPR.2018.00474

[7] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6105–6114.doi: 10.1109/CVPR.2019.00627

[8] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, “DeepFashion: Powering robust clothes recognition and retrieval with rich annotations,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1096–1104. doi: 10.1109/CVPR.2016.118.

[9] T. Xiao, J. Zhang, K. Yang, Y. Peng, and Z. Zhang, “Learning from massive noisy labeled data for image classification,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2691–2699.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 874–884.doi: 10.1109/CVPR46437.2021.00020.

[11] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin Transformer: Hierarchical vision transformer using shifted windows,” in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10012–10022. doi: 10.1109/ICCV48922.2021.00986

[12] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A ConvNet for the 2020s,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11976–11986.doi: 10.1109/CVPR52688.2022.01167

[13] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, “Transformers in vision: A survey,” ACM Computing Surveys, vol. 54, no. 10s, pp. 1–41, 2022. doi: 10.1145/3505244