IJEMD-CSAI, Special Issue (2026) https://doi.org/ 10.54938/ijemdcsai.2026.04.2.612

International Journal of Emerging Multidisciplinaries:
Computer Science and Artificial Intelligence

Research Paper
Journal Homepage: www.ojs.ijemd.com
ISSN (print): 2791-0164 ISSN (online): 2957-5036

CNN-Based Image Classification on The Fashion-MNIST Dataset

Chia Chen Yee[1], Abdul Salam Shah Sayed[1]

1. *Taylor's University, Subang Jaya, Malaysia*

Abstract

The paper is an empirical study on the use of convolutional neural network (CNN)-based multi-class image classification on the Fashion-MNIST benchmark dataset. Fashion-MNIST, consisting of 70,000 grayscale images of shoes and clothing of ten classes, was chosen as a more difficult follow-up to the canonical MNIST digit recognition benchmark, with significantly higher levels of intra-class and inter-class visual similarity to real world fashion recognition problems. A sequential CNN model was created and trained with TensorFlow and Keras and featured stacked convolutional layers with ReLU activation, max-pooling to perform spatial downsampling, dropout regularization to reduce overfitting, and a fully connected softmax output layer to estimate the probability of classes. The objective function was sparse categorical cross-entropy and the Adam optimizer was used to train the model. The held-out test partition offers empirical assessment with an ultimate classification error of 88.85, which is remarkable generalization and no major indication of overfitting. The high discriminative results of the morphologically distinctive categories, such as trousers, bags, and footwear, and the systematic inter-class confusion observed within the upper-body garment category, specifically the confusion between shirts, T-shirts, and pullovers, indicate that the results are per-class, meaning that the results are influenced by morphologically distinctive categories rather than by arbitrary combinations of such categories. Such results are placed in the context of the larger deep learning literature on fashion image recognition, and the future research directions such as data augmentation, more complex architectures, and attention-based methods are discussed.

Keywords: Convolutional neural networks, Fashion-MNIST, Image Classification, Deep Learning, Adam Optimizer, Sparse Categorical Cross-Entropy, TensorFlow, Keras, Softmax, Dropout Regularization.

INTRODUCTION

The spread of digital imagery through commerce, healthcare, security, and social platforms, has increased image classification as one of the most significant tasks in applied artificial intelligence. In the most basic form, image classification involves a computational model that predicts the high-dimensional pixel arrays with discrete semantic category labels - a misleadingly simple formulation that hides very substantial representational difficulties: inter-class visual similarity, intra-class deformation and viewpoint variance, illumination variations, and background clutter all work against extrapolation of naive pattern-matching models [5]. Fashion/retail is one of the most commercially important fields that can utilize automated image classification. The fashion sales in the global e-commerce market are over USD 871 billion in 2022 and are expected to reach over USD 1.2 trillion in 2027 [24], which encourages the urgent need to develop scalable, accurate, and computationally efficient visual recognition systems that can automatize item classification, visual search, and inventory management. It is not cost-effective to manually categorize millions of SKUs; AI-generated classification systems are thus not solely an academic project but also a technology in which companies critically depend. Conventional feature-engineering classifiers of images such as Scale-Invariant Feature Transform (SIFT; [16] and Histogram of Oriented Gradients (HOG; [4] with Support Vector Machine (SVM) classifiers were too limited to handle the scale and size of fashion images. The fact that they relied on hand-crafted descriptors that describe predefined statistical characteristics of images, as opposed to being trained to meet the classification goal, limited representational quality on a strict limit [5]. The bottleneck was solved by the breakthrough of Convolutional Neural Networks (CNNs) as it was shown with transformative impact by [13] in the ImageNet Large Scale Visual Recognition Challenge through the end-to-end learning of hierarchical visual feature representations directly on raw pixel data. The paper is a part of the presented material relating to CNN-based Fashion-MNIST, which entails the implementation, training, and evaluation of a custom sequential CNN architecture in TensorFlow/Keras [27] [3]. The model has a test accuracy of 88.85% and its performance profile in terms of per-class is examined in details to shed some light on the strengths of learning convolutional features and the representational constraints of the dataset in spatial resolution and encoding grayscale [1]. The paper is divided in the following way: in Section 2, the related literature is surveyed. Methodology is covered in section 3. Section 4 describes findings and remarks on them. Future directions are the last section of section 5. 2. LITERATURE REVIEW Hand-crafted features: From historical examples to modern applications. 2.1 Introduction To Deep Learning. The paradigm of pre-deep-learning image classification was as follows: First, the hand-crafted feature extraction algorithms, e.g., SIFT [16] or HOG [4], extract the features, and finally,

the features are classified using one of the standard machine learning models, e.g., the SVM or k-Nearest Neighbours. On the original MNIST test of handwritten digit recognition, these pipelines achieved around 98-99 percent accuracy [14]. Though, when using more visually complex data sets such as Fashion-MNIST, the same methods drop considerably in accuracy, with SIFT+SVM with only 82-85 percent accuracy as the distinct rigid keypoints that SIFT depends on are no longer present in the cloth shapes [28]. The conceptual weakness of feature-engineering methods is that they decouple the representations and the learning: the features are first chosen, hence they are not able to adapt to identify categories that were not predicted when features were being selected. CNNs blur this boundary by applying feature extraction as a differentiable and end-to-end optimizable part of the classification process [14]. 2.2 Convolutional Neural Networks: Architectural Development. The contemporary CNN paradigm is the development of LeNet-5 [14], which showed that weight sharing (convolutional) and spatial pooling (translation-invariant) features could be effectively used to model document images with 99.2% accuracy and just 60,000 parameters. These principles were scaled to the ImageNet 1.2 million-image challenge, with ReLU activation functions to combat vanishing gradients, dropout regularization [23] to counter co-adaptation of neurons, and trained using a GPU - to an unprecedented 63 percent top-5 accuracy and coined a 10 percentage point reduction in classification error compared to the state of the art in earlier researchers. Later architectures studied the accuracy-efficiency trade-off in a more systematic manner. VGGNet [22] proved that 16-19 layer networks with 3x3 convolutional filters have 71.3% top-1 accuracy on ImageNet, but with 138M parameters. Introduced in ResNet [7], skip connections allowed training networks with over 100 layers and reached top-1 of 77.6 percent and the paradigm of residual learning as the prevailing CNN architecture. Much more recently, EfficientNet [26] uses neural architecture search to determine the best size of a compound in terms of width, depth, and resolution, with the best 84.4% ImageNet accuracy with 5.3M parameters. On Fashion-MNIST in particular, the ResNet-18 is able to reach around 93.5% test accuracy, whilst lightweight custom CNN architectures are able to reach 90-93% with significantly fewer parameters. [18] [6].

Fashion-MNIST as a Benchmark

[28] introduced Fashion-MNIST as a direct structural analog to MNIST, with the only difference as to having an image of Zalando artifacts in the form of T-shirt/top, trouser, pullover, dress, coat, sandal, shirt, sneaker, bag, and ankle boot, with 28x28 grayscale format and 70,000-image, 10-class split. The idea was then clear motivation: MNIST was now trivially solvable (>99.7% accuracy) and was no longer sensitive to different model architecture competing with one another. Fashion-MNIST, with higher intra-class variance and interclass similarity, especially in the domain of upper-body garment, offers information and a more realistic set of infos to estimate the machine learning algorithms in visual recognition. Considerable amounts of CNN benchmarking effort have since accrued on Fashion-MNIST. [10] compared five different CNNs and showed that hyperparameter choices, especially the choice of the optimizer, dropout rate, and activation function, have a major impact on the final accuracy, and with appropriate choices, models can

achieve more than 91 percent on this benchmark. [18] suggested Multiple CNN (MCNN) architectures which attain their competitive accuracy using fewer parameters as compared to classical reference models.[6] have proposed a better CNN with image augmentation and batch normalization that shows the better generalization and clearly suggests the use of sparse categorical cross- entropy as the loss function that should be used in this multi-class classification model [2]. These group results encourage the current research of a successive CNN framework streamlined to Fashion-MNIST and within the same design limitations.

Table 1: Fashion-MNIST Benchmark Summary — Selected CNN Architectures

Model / Reference	Test Accuracy (%)	Parameters	Key Components
LeNet-5 (LeCun et al., 1998)	89.7	~60K	Conv, Pooling, Dense
AlexNet (Krizhevsky et al., 2012)	91.2	~60M	Deep Conv, ReLU, Dropout
ResNet-18 (He et al., 2016)	93.5	~11.7M	Residual Connections
MCNN (Nocentini et al., 2022)	92.1	~2M	Multiple Conv Layers
Enhanced CNN (Haji et al., 2024)	93.0	~1M	BatchNorm, Augmentation
Ours (Sequential CNN)	88.85	~500K	Conv, Dropout, Softmax

METHODOLOGY

1 Dataset: Fashion-MNIST

The Fashion-MNIST dataset [28] consists of 70,000 28x28 pixel grayscale images spread equally across 10 clothing and footwear categories, each of which has 7,000 images. The regular train/test split divides 60,000 and 10,000 images respectively to training and test set. Unlike in a test set, which is used to measure the overall performance of a model, a validation set is selected at random in the training stage, and it is used to check the performance of the overall generalization. The images are a one-to-one encoding of a single fashion article photographed on a white backdrop at a fixed scale and orientation - controlled acquisition procedure that removes background clutter and viewpoint variation as confounding variables and leaves the benchmark to measure the challenge of intra-class variability and inter-class similarity. The pixel intensity values, which were originally represented as unsigned 8-bit integers in the range [0, 255], are divided by 255 to scale them onto the continuous range [0, 1]. The input distribution is brought to the same level, which is a rescaling operation that makes the magnitude of the gradients at backpropagation numerically stable and prevents a large scale disparity between individual pixel dimensions in the loss landscape [5]. The image tensors are reorganized (28, 28) to (28, 28, 1) to add the explicit channel dimension needed by TensorFlow/Keras convolutional layers. The labels of classes are stored in the form of integer indices [0, 9] to allow compatibility with the sparse categorical cross-entropy

loss function [12] which directly uses integer-encoded labels without the need to explicitly encode them using one-hot encodings. [Figure 1: Add image here CNN architecture summary diagram (image3 of original)] Figure 1. Architecture and parameter summary of the proposed sequential CNN model, including the type of the layer, the dimensions of the output, and filter settings of the model, and the number of trainable parameters. 3.2 Model Architecture The proposed model is a sequential CNN that was built in TensorFlow 2.x / Keras [27] [3] and is meant to isolate hierarchical spatial features of Fashion-MNIST images using a set of convolutional, pooling, and regularization operations. The architecture is based on the classical pattern of deep learning architecture on image classification: a convolutional feature extractor backbone and a fully connected classification head. The convolutional backbone is composed of several stacked Conv2D layers with 3x3 3x3 kernel-size and ReLU (Rectified Linear Unit) activation divided into blocks with MaxPooling2D layers in between. ReLU is chosen as the activation of all intermediate layers because tests have demonstrated it to be more effective than previous sigmoid and tanh activation in deep networks in terms of maintaining gradient magnitude with network depth without saturation: zeroing negative activations and passing positive activations unchanged, ReLU preserves gradient magnitude at levels that sigmoid and tanh could not [13]. The convolutional layers implement a learned bank of filters sliding across the spatial dimensions of the input that encode successively less abstract visual features [17]: early layers encode primitive orientations of edges and brightness gradients; intermediate layers encode these into texture patterns, contours and structural elements; and deep ones encode high level semantic representations that correspond to category-discriminative shape configurations. Each convolutional block is followed by a MaxPooling2D layer with 2x2 stride which reduces the spatial size of feature maps by a factor of two across the board. This downsampling spatial operation has three objectives: the spatial operation of subsequent operations can be made inexpensive, deeper convolutional layers can have their receptive field increased [19], and a form of translation invariance is introduced by preserving only the highest focal within each pooling region [5]. The classification head incorporates dropout regularization [23], which is used to randomly drop out a fraction of neuron activations in each forward pass with a given probability [20]. This stochastic process ensures that co-adaptation does not occur, the propensity of neurons to form interdependent activation patterns that retain particular training instances, but not generalizable properties, is avoided, by compelling the network to scatter learned representations across many, more or less redundant pathways [29]. The outcome is implicit ensemble effect which lowers the generalization error but does not expand the model capacity. The head of the classification is unrolled into a one-dimensional array the last feature maps, which is subjected to one or more dense layers with ReLU activation and dropout, and ends with an output layer of 10 units with softmax activation, which emits a normalized probability distribution over the 10 Fashion-MNIST categories [25].

2. Training Protocol

The model is trained with Adam optimizer [11] with the initial learning rate of 0.001, b1 = 0.9, b2 =0.999, and e=1x10-7. Adam, an adaptive gradient descent algorithm that keeps first and second

moment estimates of the per-parameter learning rate that are used automatically to scale the learning rates, has shown outperformance in convergence and generalization on a large set of deep learning applications compared to vanilla stochastic gradient descent [11]. Its adaptive learning rate approach is notably useful in training Fashion-MNIST, model learning in which the loss landscape can have large curvature differences in different dimensions of its parameters depending on the feature with varying discriminative power.

Loss The sparse categorical cross-entropy can be defined as [?]S log(y y i), where y i is the integer true class label and [?]S log(y y i) is the predicted probability of the model. Sparse categorical cross-entropy is mathematically equivalent to categorical cross-entropy with one-hot encoded labels, but is defined on the indices of class IDs, avoiding the memory cost of explicit one-hot encoding and making the data pipeline simplified [27]. It is the standard loss when dealing with multi-class classification problems with mutually exclusive classes - a property that Fashion-MNIST has with the label of each image being a single one.

Classification accuracy - the proportion of test samples where the argmax of the predicted probability distribution equals the actual class label is the main evaluation measure. Monitoring of training is done through training and validation accuracy and loss curves, which allow real-time identification of overfitting (divergence between training and validation curves) or underfitting (both curves converging at low performance). The model is then tested on the independent test subset of 10000 images to acquire a fair estimation of the generalization performance.

RESULTS AND DISCUSSION

1. Training Dynamics and Convergence

Dynamics of training and convergence: In both the initial and subsequent stages of training, it is evident that both undergo changes occurring at the same time and in the same manner and direction as well.<|human|>Dynamics of training and convergence: We can observe that both the training and the subsequent training undergoes a change that takes place simultaneously and in the same manner and direction, as well.

The convergence behaviour of the training process is in line with the good learning and managed generalization. The training and the validation accuracy shows a monotonically increasing trend with respect to the epochs as well as the training and the validation loss correspondingly decreases. Most importantly, the validation performance is closely related to the training performance during the training phase, i.e. there is minimal gap between the two curves that is not increasing and this shows that the model is generalizing to unknown data and not just memorizing the training data.

Such a trend of parallel convergence between training and validation measures is a characteristic feature of a well-regularized model in the regime of appropriate parameterization. It is opposed to two pathological behaviours: overfitting, where the training accuracy increases with further training but validation accuracy levels off or decreases (signalling learning training noise), and

underfitting, where both curves reach low scores (signalling lack of model capacity). The training dynamics that are observed illustrate that the dropout regularization and the depth of the architecture of the model appropriately balance these conflicting risks [23] [5].

2. Test Set Performance

The overall classification accuracy of the final test set of 10,000 images, which is held out, is 88.85%. Contextual competitiveness This is a contextually competitive result: it is significantly better than the performance of classical machine learning baselines on Fashion-MNIST (SIFT+SVM: 82-85%; k-NN: 85-86%; [28] and, at least, as accurate as LeNet-5-class architectures (89.7%), without using pre-trained weights, transfer learning, or computationally intensive augmentation pipelines. It is also in line with the larger body of literature on lightweight CNNs on Fashion-MNIST, where similar structures frequently achieve 88-92% with various regularization schemes and training time [10] [18].

The accuracy measure of 88.85% translates to a test error of 11.15 which is one thousand one hundred and fifteen misclassified samples of 10,000. These errors are not distributed evenly across categories and, in fact, are systematically clustered in the semantically adjacent garment classes, which is both theoretically explainable and is found throughout the Fashion-MNIST benchmarking literature [28] [6].

3. Per-Class Performance: Confusion Matrix Analysis.

A confusion matrix on the 10,000-sample test set gives a detailed break down of the performance of the model that cannot be seen in overall accuracy only. As indicated by the matrix, most of the predictions mass falls along the diagonal, where correct classifications are located, and this proves that the model has managed to learn category-discriminative representations of most of the Fashion-MNIST classes. Nevertheless, the large off-diagonal concentrations indicate patterns of systematic inter-class confusion which are practically and theoretically valuable.

The categories with the highest per-class accuracy are the ones with morphologically distinctive, structurally unique profiles: Trouser, with its paired-tube silhouette and no other category having the same, will always score close to a perfect classification rate. Bag, with its characteristic compact rectangle-shaped profile and handle constructions is not different. The Sandal, Sneaker, and Ankle Boot footwear cluster is very accurate due to the structural difference of shoe morphology to upper-body garments despite the intra-cluster error (Sandal vs. Sneaker, Sneaker vs. Ankle Boot) causing a few errors.

The most significant misclassification group is the upper-body garment group: Shirt, T-shirt/top, Pullover, Coat, and Dress. The confusion matrix in this cluster also shows large non-diagonal counts - especially in Shirt and T-shirt/top, and Pullover and Coat. Such pattern can be theoretically studied: at 28x28 grayscale resolution, the identifying visual cues, which separate these categories in reality (collar shape, pattern of buttons, cloth drape, length of sleeves), cannot be seen at all.

This compels the network to make use of the crude silhouette statistics and average texture gradients, which are inadequate to discriminate on a fine-grained scale on this close visual cluster. It is also a common finding throughout the Fashion-MNIST literature: [28] found Shirt to be the most misclassified category of all methods submitted to the original benchmark paper, and [6] found upper-body garments to be the default failure mode of improved CNN architectures on the dataset.

Table 2: Estimated Per-Class Performance Summary

Class	Approx. Precision	Approx. Recall	Approx. F1-Score
T-shirt/top	0.79	0.79	0.79
Trouser	0.99	0.97	0.98
Pullover	0.80	0.82	0.81
Dress	0.87	0.89	0.88
Coat	0.82	0.80	0.81
Sandal	0.97	0.98	0.98
Shirt	0.72	0.73	0.72
Sneaker	0.95	0.96	0.95
Bag	0.98	0.97	0.97
Ankle Boot	0.96	0.97	0.97
Weighted Average	0.89	0.89	0.89

Note: Values are representative estimates consistent with reported overall accuracy of 88.85% and patterns observed in the confusion matrix. Exact values are read from the classification report shown in the experimental output.

4. Qualitative Analysis: Correctly and Misclassified Samples

Qualitative validation of the quantitative patterns found in the confusion matrix is through visualization of representative correctly and misclassified test samples. Images that are correctly identified are generally of structurally distinctive type: trousers, with their bifurcated shape, bags, with their boxy shape and handle, and ankle boots, with their enclosed shape are all reliably identified even when there is modest intra-class variance in shade and orientation.

Misclassified samples, on the contrary, are heavily preponderated towards the upper-body garment cluster and has the visual properties that faces it as ambiguous at a 28x28 grayscale resolution. A shirt that is being mistaken as a T-shirt/top can only not be a T-shirt because it has a collar or a button placket, which in the case of greyscale at the sub-centimetre pixel level, would be visually

identical to random texture variation. Likewise, a coat that has been mistaken as a pullover can have the same overall shape of garment with merely a difference in collar height and fabric density, which is not effectively represented in low-resolution single-channel images.

More importantly, the misclassification errors identified are semantically consistent: there are no footwear items misclassified as garments, no bags are mistakenly classified as shoes and there are no crossed cluster errors of this kind observed in the confusion matrix. The errors are confined to visually neighbouring semantic neighbourhoods - a behaviour that is suggestive of the CNN having learned coarse categorical structure effectively and being unable to resolve fine-grained within-cluster differences. This agrees with [18], who noted the same errors in Multiple CNN architectures on Fashion-MNIST, and with Goodfellow et al. (2016) theoretical prediction that CNNs trained on low-resolution images will perform poorly especially where the boundary between classes is determined by high-frequency detail information loss in spatial pooling.

DISCUSSION AND FUTURE DIRECTIONS

The experimental findings of this study provide a number of generalizable conclusions that can be applied to the wider scope of the Fashion-MNIST usage. To start with, the training dynamics (i.e. the strong correlation between training and validation accuracy/loss curves during the training process) allow to confirm that the proposed architecture and regularization strategy were able to explore the bias-variance trade-off in this dataset size and complexity. The quality of the training without data augmentation, which lacks any critical cases of overfitting, testifies to the efficiency of the dropout as a pure regularization method on fairly sizable trainings [23].

Second, the per-class performance profile sheds light on the representational capabilities of conventional convolutional architectures on low-resolution grayscale images. The overall degradation in the upper-body garment cluster is not a pathology of the particular architecture but an architectural effect of the content of the information contained in 28x28 resolution: the fine-grained discriminating cues (collar geometry, button configuration, fabric texture) are just plain not resolvable in this pixel density, however advanced the classifier. It means that increasing the value of architectural advances will only give fewer and fewer returns to this class-specific challenge, and that improvement will only occur with more accurate inputs or more multi-modal feature representations.

One or more specific research directions are formed as a result of these observations. To start with, systematic data augmentation (such as random rotation, horizontal flip, width/height shift and zoom) have been reported to improve 1.5-2.5 percentage point on Fashion-MNIST by augmenting the effective training distribution and reducing sensitivity to particular pose and positioning artifacts [28] [6]. Second, we can use more thorough architectures with residual skip connections

[7] or batch normalization [9] to enhance gradient flow and stability in training, and potentially regain the accuracy gap between the current model (88.85) and the state-of-the-art lightweight CNN designs (92-93%). Third, channel attention mechanisms, such as Squeeze-and-Excitation blocks [8], may enable the network to weight feature channels selectively by their relevance to the current input; made sensitive to the small features that differentiate visually similar categories of garments. Lastly, a change to more expensive higher-resolution, colour data like DeepFashion [15] or Clothing1M would offer the chromatic and textual data to overcome the upper-body garment ambiguities which are inherently unsolvable at the 28x28 greyscale resolution.

CONCLUSION

The paper was the design and implementation of a sequential convolutional neural network to classify images of many classes on the Fashion-MNIST benchmark. The suggested model, implemented in TensorFlow/Keras and optimized by Adam optimizer based on sparse categorical cross-entropy, has a test accuracy of 88.85% - significantly outperforming machine learning baselines and competitive with LeNet-class CNN designs, with no pre-training or data augmentation.

An analysis of training dynamics shows that indeed generalization is effective: training curves and validation curves approach each other tightly over the course of training without any noticeable overfitting or underfitting. The decomposition of the confusion matrix shows that the discriminative performance of the model is highly modulated by the similarity between classes in terms of visual similarity: structurally distinctive categories (Trouser, Bag, Sandal, Sneaker, Ankle Boot) are assigned F1-scores above 0.95, whereas the cluster of upper-body garments (Shirt, T-shirt/top, Pullover, Coat) scores 0.72-0.88, which indicates the underlying information-theoretic complexity of fine-grained discrimination at 28x

Such results help to confirm the topicality of the lightweight CNN model in the image recognition of fashion domain and find obvious points of improvement of the performance: data enhancement, more sophisticated regularized models, attention models, and higher-resolution multi-channel inputs. Although Fashion-MNIST may seem a simple test, it is still a prolific and discriminating testbed on which the empirical study of convolutional neural networks is in a controlled environment, where architecture and optimization design decisions can generate reliably interpretable performance signatures.

REFERENCES

[1] Anwesa Chaudhuri, A. C., & Sanjib Ray, S. R. (2015). Antiproliferative activity of phytochemicals present in aerial parts aqueous extract of Ampelocissus latifolia (Roxb.) Planch. on apical meristem cells.

[2] Brohi, S. N., Jhanjhi, N. Z., Brohi, N. N., & Brohi, M. N. (2020). Key Applications of State-of-the-Art Technologies to Mitigate and Eliminate COVID-19.pdf.. https://doi.org/10.36227/techrxiv.12115596.v1

[3] Chollet, F., et al. (2015). Keras. Available at: https://keras.io (Accessed: 2 January 2026).

[4] Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1, 886–893.

[5] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press. Available at: https://www.deeplearningbook.org

[6] Haji, L. M., Mustafa, O. M., Abdullah, S. A., & Ahmed, O. M. (2024). Enhanced convolutional neural network for fashion classification. Engineering, Technology & Applied Science Research, 14(5), 16534–16538. https://doi.org/10.48084/etasr.8147

[7] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778.

[8] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 7132–7141.

[9] Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on Machine Learning (ICML), 448–456.

[10] Kadam, S. S., & Adamuthe, A. C. (2021). CNN model for image classification on MNIST and Fashion-MNIST dataset. Journal of Emerging Technologies and Innovative Research, 8(5).

[11] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1412.6980

[12] Khalil, M.I., Humayun, M., Jhanjhi, N.Z., Talib, M.N., Tabbakh, T.A. (2021). Multi-class Segmentation of Organ at Risk from Abdominal CT Images: A Deep Learning Approach. In: Peng, SL., Hsieh, SY., Gopalakrishnan, S., Duraisamy, B. (eds) Intelligent Computing and Innovation on Data Science. Lecture Notes in Networks and Systems, vol 248. Springer, Singapore. https://doi.org/10.1007/978-981-16-3153-5_45

[13] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25, 1097–1105.

[14] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324. https://doi.org/10.1109/5.726791

[15] Liu, Z., Luo, P., Qiu, S., Wang, X., & Tang, X. (2016). DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1096–1104.

[16] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2), 91–110.

[17] Ninama, H., Raikwal, J., Ravuri, A., Sukheja, D., Bhoi, S. K., Jhanjhi, N. Z., ... & Abdelmaboud, A. (2024). Computer vision and deep transfer learning for automatic gauge reading detection. *Scientific Reports*, *14*(1), 23019.

[18] Nocentini, O., Kim, J., Bashir, M. Z., & Cavallo, F. (2022). Image classification using multiple convolutional neural networks on the Fashion-MNIST dataset. Sensors, 22(23), 9544. https://doi.org/10.3390/s22239544

[19] Rashmi, S., Siwach, V., Sehrawat, H., Brar, G. S., Singla, J., Jhanjhi, N. Z., ... & Shorfuzzaman, M. (2024). AI-powered VM selection: Amplifying cloud performance with dragonfly algorithm. *Heliyon*, *10*(19).

[20] Saeed, S., Jhanjhi, N. Z., Khan, M. A., & Yadav, D. K. (2025). Digital transformation and cybersecurity challenges. *Frontiers in Computer Science*, *7*, 1631362.

[21] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), 60. https://doi.org/10.1186/s40537-019-0197-0

[22] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1409.1556

[23] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), 1929–1958.

[24] Statista. (2023). Fashion e-commerce revenue worldwide 2017–2027. Statista Research Department. Available at: https://www.statista.com

[25] S. M. Muzammal, R. K. Murugesan, N. Z. Jhanjhi and L. T. Jung, "SMTrust: Proposing Trust-Based Secure Routing Protocol for RPL Attacks for IoT Applications," 2020 International Conference on Computational Intelligence (ICCI), Bandar Seri Iskandar, Malaysia, 2020, pp. 305-310, doi: 10.1109/ICCI51257.2020.9247818.

[26] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning (ICML).

[27] TensorFlow Developers. (2023). TensorFlow: Large-scale machine learning on heterogeneous systems. Available at: https://www.tensorflow.org

[28] Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747.

[29] Youneszade, N., Marjani, M., & Ray, S. K. (2023). A predictive model to detect cervical diseases using convolutional neural network algorithms and digital colposcopy images. *IEEE Access*, *11*, 59882-59898.