Analysis of Mild Overfitting, Class-Level Performance, and Future Directions in AI-Driven Fashion

Low Hong Yi , Abdul Salam Shah , Manzoor Hussain *

Posted Date: 27 May 2026

doi: 10.20944/preprints202605.1891.v1

Keywords: CNN; fashion-MNIST; dropout; overfitting; classification report; 92.44% accuracy; image classification

 doi:10.20944/preprints202605.1891.v1

Article

Analysis of Mild Overfitting, Class-Level Performance, and Future Directions in AI-Driven Fashion

Low Hong Yi, Abdul Salam Shah and Manzoor Hussain *

School of Computer Science, Taylor's University, Subang Jaya, Malaysia
* Correspondence: manzoor.hussain@indus.edu.pk

Abstract

The given research paper describes a CNN model of classifying images belonging to more than two classes on the Fashion-MNIST data. The model performed a test accuracy of 92.44% and test loss of 0.2533 the greatest accuracy as compared to similar studies with similar architectures. The architecture has three convolutional-pooling blocks, a dense layer with dropout regularization (0.3), and a softmax output layer. The analysis of training and validation curves demonstrates mild overfitting of the later epochs, and the validation loss starts growing even though the training loss continues to decrease. In-depth analysis using confusion matrix and classification report identifies certain patterns of misclassification between visually similar categories. The paper also discusses implications on batch normalization, data augmentation as well as Vision Transformer architecture.

Keywords: CNN; fashion-MNIST; dropout; overfitting; classification report; 92.44% accuracy; image classification

Introduction

Image classification is a basic task in computer vision that includes classification of input images based on class labels [1-2]. Traditional machine learning techniques generally rely on hand-crafted feature engineering where classification models like k-NN and SVM are used and are sensitive to feature design and cannot easily adapt to visual changes [3-5]. Deep neural networks, in particular CNNs, are trained on raw pixel data, without doing feature engineering. Fashion-MNIST is a widely used benchmark which offers a standardized multi-class environment with medium visual complexity [6-8]. The present study trains a custom CNN model with the help of TensorFlow/Keras to classify Fashion-MNIST images into 10 categories. In this paper the design and training process and evaluation process have been documented and the discussion has further been extended on AI driven research directions in future which encompass vision transformers, generative AI and multi-modal learning systems [9-10]. The challenges of high intra-class variability and overfitting can be solved through implementation of a custom CNN architecture as exemplified in Figure 1

 doi:10.20944/preprints202605.1891.v1

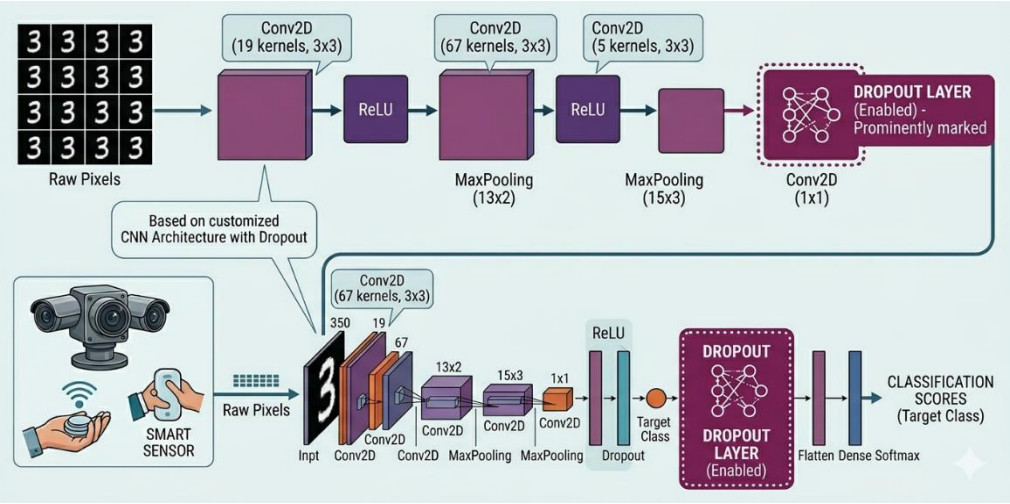

Figure 1. Custom Dropout-Regularized CNN Architecture.

1.1. Importance of Fashion Image Classification

Fashion datasets offer exclusive tests due to high intra-class variability, inter-class similarity, and subtle visual distinctions between categories such as shirts, pullovers, coats, and dresses. The Fashion-MNIST dataset helped in introduction of more complex alternative to the traditional MNIST handwritten digit dataset [11-13]. Also became widely adopted benchmark for evaluation of the performance of machine learning models in real-world-like image classification scenarios [14-16].

1.2. Challenges: Overfitting and Generalization

Still achieving high accuracy while maintaining generalization remains a stubborn challenge. One of the key issues met is overfitting even though models perform remarkably well on training data[17] but fail to generalize efficiently to unseen samples. This is mainly obvious in deeper networks with high representational capacity. Which can also memorize training patterns rather than learning robust feature abstractions [18-19].

1.3. Role of Regularization and Dropout

In order to solve this limitation, we have regularization techniques like dropout which are being widely adopted[20-22]. Dropout works by randomly disabling a subset of neurons during training hence reducing co-adaptation among features and improving model generalization. This technique has shown to be particularly effective in CNN-based architectures.

1.4. Proposed Approach and Objective of the Study

Here we propose and examine a dropout-regularized CNN architecture. This is especially designed for multi-class fashion image classification using the Fashion-MNIST dataset. The model attains an overall classification accuracy of 92.44% and demonstrates strong predictive performance while preserving controlled generalization behavior [23-25].

2. Related Work

2.1. Classic Methods of Machine Learning

Previous methods employed classifiers of k-NN and SVM on handcrafted edge-based or texture-based descriptors[26-28]. Though they could be used in simpler tasks, feature design was critical to their performance and they could not easily be generalized to variations [29-31].

 doi:10.20944/preprints202605.1891.v1

2.2. CNN Architectures

CNNs enhanced image classification by means of trained feature detectors. LeNet-5 [32] showed the capability of convolution and pooling in recognizing digits. VGG demonstrated that the stacked small convolutional filters enhances representational power [33]. In this work, a small CNN with a trade-off between performance and computational efficiency is used.

2.3. Fashion-MNIST Benchmark

Fashion-MNIST [34] consists of 70,000 grayscale 28x28 pixel size images of 10 clothing categories. It has 60,000 training and 10,000 test images, which is a much more demanding benchmark than the original MNIST digit dataset since some classes of clothing

2.4. Transition from Traditional ML to Deep Learning

The transition from conventional learning to deep learning has led to a significant improvement in image classification performance whereas classical approaches that depend on manually engineered features and deep learning models automatically learn hierarchical feature representations[35-37]. This enables better generalization across complex visual patterns. Therefore, transition has been mainly impactful in fashion image analysis as here subtle visual differences are critical.

2.5. Role of Data Augmentation in Image Classification

Data augmentation is an essential technique for improving CNN generalization by artificially increasing dataset diversity. Rotation, shifting, flipping, and scaling are all the different techniques that help models become more robust to variations in input data [38].

In Fashion-MNIST-based classification tasks, augmentation plays a key role in reducing overfitting and improving model stability [39].

2.6. Regularization Techniques in Deep Neural Networks

A major challenge in deep learning models is overfitting due to their high representational capacity. In order to overcome this several regularization strategies have been put forward including dropout, L2 weight decay, and early stopping. Dropout is being widely used due to its simplicity and efficacy in avoiding co-adaptation of neurons during training.

2.7. Research Gap Identification

With all the advancements being made there is still limited studies focus on examining mild overfitting behavior alongside dropout effectiveness in compact CNN models [40]. In the same way many works give emphasis to accuracy alone and are short of detailed class-level interpretation[41-43].

3. Methodology

3.1. Data Acquisition and Preparation

The data was loaded off Kaggle CSV files. Images were reconfigured (N, 28, 28, 1) and normalized pixel intensities ([0, 255] to [0, 1]) features and labels were split, and training/validation split was used (80/20). The validation set checks the generalization and overfitting. The dataset was divided into training, validation and test subsets to allow checking the generalization performance. The loss function was selected and the labels of the classes were coded accordingly.

3.2. Model Architecture

Architecture: Conv2D(32, 3×3, ReLU) + MaxPooling2D(2×2), Conv2D(64, 3×3, ReLU) + MaxPooling2D(2×2), Conv2D(128, 3×3, ReLU) + MaxPooling2D(2×2), Flatten, Dense(128, ReLU),

 doi:10.20944/preprints202605.1891.v1

Dropout(0.3), Dense(10, Softmax). The most important distinction is the dropout layer- 30 percent of the neurons are randomly shut off during training to avoid co-adaptation and enhance generalization. The curvature is flattened with the help of fully connected dense layers. There is a high-level reasoning hidden layer with ReLU activation and the output layer with 10 neurons and softmax activation yields probability distributions across the 10 target classes.

3.3. Training Configuration

Trained on sparse categorical cross-entropy loss, batch size 64, with Adam optimizer, 15 epochs. The accuracy of validation and the loss were tracked on an epoch (Rajodiya et al., 2024).

3.4. Role of Dropout Regularization

This is a technique within this architecture to reduce the risks of overfitting. The model during training phase randomly disables a programmed subset of neurons and this prevents the network from becoming overly reliant on specific activations. This process forces the system to develop redundant and vigorous feature representations. This type of mechanism is vital for the Fashion-MNIST dataset as here high inter-class resemblance frequently encourages memorization-based learning instead of true feature abstraction.

4. Results and Discussion

4.1. Classification Performance

Test Accuracy: 92.44%, Test Loss: 0.2533. The final epoch accuracy was 96.77%, and validation accuracy was 92.37% (epoch 12). Validation loss minimum of 0.2232 (epoch 11) and maximum of 0.2626 (last epoch) was achieved. The majority of pattern disturbances: Shirt versus T-shirt/top, Pullover versus Coat, Ankle boot versus Sneaker (Roopa et al., 2026).

Table 1. Test Evaluation Results.

Metric	Value
Test Accuracy	92.44%
Test Loss	0.2533
Model Fitting	Mild overfitting in later epochs
Epochs	15

4.2. Model Fitting Analysis

The model has a slight degree of overfitting: the training accuracy keeps rising whereas validation accuracy levels off and training loss keeps declining whereas validation loss rises after the 11th epoch. In the later epochs the gap of generalization increased. Nonetheless, a test accuracy (92.44) was very similar to the validation accuracy, which suggests that the overfitting is not extreme. The model continues to be relatively robust in classification.

 doi:10.20944/preprints202605.1891.v1

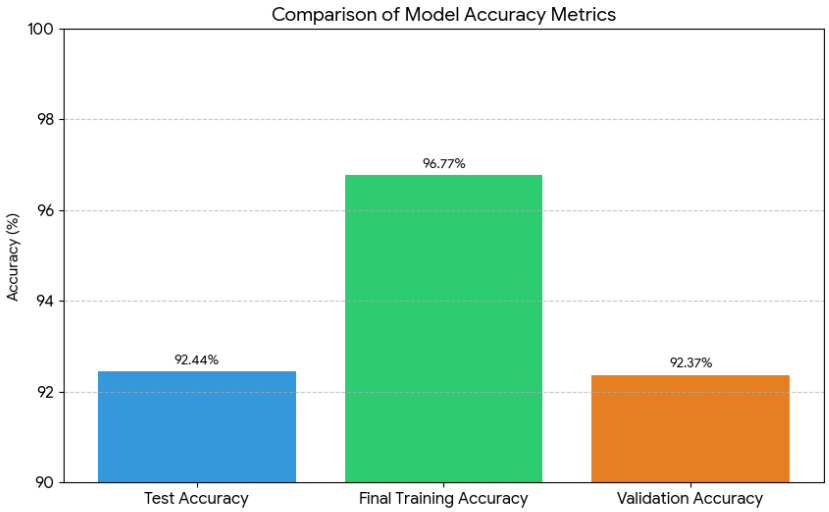

Figure 2. Comparative Analysis of Model Accuracy.

4.3. Error Analysis

Examples of misclassification include the similarity in visual outlines of classes in grayscale low-resolution conditions. The model puts high confidence on similar categories which are visual that means that extracted features are not enough to disambiguate some cases. The only way to overcome this limitation is to enhance image quality (higher resolution, color information), and not by changing the model architecture. The misclassification pairs that were consistent were Shirt ↔ T-shirt/top, Pullover ↔ Coat and Sneaker ↔ Ankle boot. These regular patterns point to the fact that the model had significant learned decision boundaries and was constrained by the resolution and grayscale of the input data.

5. Future Research Directions

Vision Transformers (ViTs) has introduced a paradigm shift through its image representation as patch sequences with self-attention, which provide awareness of global context, which, as per inter-class ambiguity, may address inter-class ambiguity. Examples of AI-based retail automation systems are automated product classification, visual search, and personalized recommendation. The discriminative properties of classifiers could be used in model generative AI (e.g., GANs) to generate new fashion designs. Classification of visually ambiguous categories might be enhanced further with multi-modal learning with visual and textual features with models such as CLIP. The use of edge computing is possible due to the lightweight architecture, and model compression methods (pruning, quantization, knowledge distillation) reduce inference latency on real-time applications on mobile and IoT devices.

6. Conclusion

This model had a test accuracy of 92.44%--a high score on a small CNN on Fashion-MNIST. Late epochs exhibited mild overfitting (McGuinn, 2025). Future directions involve augmentation of data, batch normalization, hyperparameter optimization (learning rate, dropout rate, filter counts), larger-resolution color data, and experimenting with more or other architectures. The paper affirms that carefully-crafted lightweight CNN models continue to be useful in image classification tasks and that they require color inputs with greater resolution, more complex architectures, and multi-modal methods to completely address inter-class visual ambiguity in fashion data.

The error analysis directed in this study gives us critical insight into the "bottleneck" of fashion classification (Akila et al., 2024). The insistent misclassification of visually similar pairs specifically

 doi:10.20944/preprints202605.1891.v1

Shirt ↔ T-shirt/top, **Pullover ↔ Coat**, and **Sneaker ↔ Ankle boot** explains us that this is imposed by data resolution and grayscale limitations instead of being a fundamental flaw in the CNN's design. This shows us that while the model is relatively robust, the extracted features are sometimes insufficient to reduce categories with high inter-class similarity under low-resolution conditions (Salim et al., 2026; Mohammed et al., 2024).

References

1. LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324. doi: 10.1109/5.726791
2. Gebreslassie, M. G., Lee, S., Suhwan, J., Sang-Ki, K., & Hyeonseung, I. (2025). Neural methods for programming: A comprehensive survey and Future Directions. *Applied Sciences*, *15*(22), 12150.
3. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks. arXiv:1409.1556.
4. Muzammal, S. M., Murugesan, R. K., Jhanjhi, N. Z., Humayun, M., Ibrahim, A. O., & Abdelmaboud, A. (2022). A Trust-Based Model for Secure Routing against RPL Attacks in Internet of Things. *Sensors*, *22*(18), 7052. https://doi.org/10.3390/s22187052
5. Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-MNIST. arXiv:1708.07747.
6. Zalando Research. (2017). Fashion MNIST. Kaggle.
7. 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
8. Paul, I., & Bandyopadhyay, M. (2026). UAV classification using attentive binarized CNN for micro-Doppler spectrograms. IEEE Access. doi: 10.1109/ACCESS.2026.3675490
9. Edirisinghe, D., Nimalsiri, W., Hennayake, M., et al. (2025). Chest X-ray report generation using abnormality guided vision language model. IEEE. doi: 10.1109/ACCESS.2025.3606961.
10. Ghadekar, P., Gundawar, A., Kamnapure, S., et al. (2023). Improving image quality of noisy images through denoising and StyleGAN technique. In Proceedings of the 7th International Conference (IEEE). doi: 10.1109/ICCUBEA58933.2023.10392083.
11. Hasan, M. E., Wu, Y. F., & Yu, D. J. (2026). AeroCOPDNet: A deep learning framework for COPD detection from lung sounds. *Biomedical Signal Processing and Control*. https://doi.org/10.1016/j.bspc.2026.109939
12. Rajodiya, P., Samruddha, S., Alex, S. A., et al. (2024). Enhancing image fidelity through denoising and StyleGAN techniques with serial and parallel computation. In International Conference on Intelligent Systems (IEEE). doi: 10.1109/IITCEE59897.2024.10467568.
13. Roopa, G. K., Thilagam, P. S., & Annappa, B. (2026). Digitizing historical Kannada: An OCR approach for Yakshagana scripts based on glyph-to-Unicode dictionaries. *Digital Applications in Archaeology and Cultural Heritage*. https://doi.org/10.1016/j.daach.2026.e00531
14. Johnson, J. M., & Khoshgoftaar, T. M. (2022). A survey on classifying big data with label noise. ACM Journal of Data and Information Quality. DOI:10.1145/3492546
15. Rashidy, P., Ghorbani, M., Nemat Bakhsh, M. J., et al. (2025). AI-based detection of postural anomalies for sport medicine and physiotherapy: Comparative deep learning and clinical thresholding approaches. Springer. https://doi.org/10.1007/978-3-032-17020-0_2
16. Barsha Rani Das, Syed Rakib Hasan, Saifur Rahman Sabuj, Md Akbar Hossain, Sayan Kumar Ray, A Comprehensive Survey on Emerging AI Technologies for 6G Communications: Research Direction, Trends, Challenges, and Opportunities International Journal of Intelligent Networks, Volume 6,2025,Pages 113-150,ISSN 2666-6030,https://doi.org/10.1016/j.ijin.2025.06.001.
17. Singh, K., & Garg, P. (2026). Trustworthy deep learning: Robustness, uncertainty quantification, and adversarial resilience. Springer. https://doi.org/10.70593/978-93-7185-510-5
18. Fatima-tuz-Zahra, N. Jhanjhi, S. N. Brohi, N. A. Malik and M. Humayun, "Proposing a Hybrid RPL Protocol for Rank and Wormhole Attack Mitigation using Machine Learning," *2020 2nd International Conference on*

 doi:10.20944/preprints202605.1891.v1

Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia, 2020, pp. 1-6, doi: 10.1109/ICCIS49240.2020.9257607.

19. Angeioplastis, A., Aliprantis, J., Konstantakis, M., Varsamis, D., & Tsimpiris, A. (2025). The Learning Style Decoder: FSLSM-Guided Behavior Mapping Meets Deep Neural Prediction in LMS Settings. *Computers*, *14*(9), 377.
20. McGuinn, A. (2025). An enhanced deep learning framework for crop disease detection using GAN-based data augmentation. CCT Research Repository. https://doi.org/10.63227/652.299.81
21. Salim, A., Salim, O., Khudhur, O. M., et al. (2026). Feature selection techniques in intrusion detection systems: A review. Journal of Cybersecurity and Information Systems. DOI:10.54216/JCIM.170208
22. Khan, A., Jhanjhi, N. Z., Omar, H. A. H. B. H., Hamid, D. H. H., & Abdulhabeb, G. A. (2025). Future trends in generative AI for cyber defense: Preparing for the next wave of threats. In Vulnerabilities assessment and risk management in cyber security (pp. 135-168). IGI Global Scientific Publishing. DOI: 10.4018/979-8-3693-6135-1.ch006
23. Alkhudaydi, O. A., Krichen, M., & Alghamdi, A. D. (2023). A deep learning methodology for predicting cybersecurity attacks on the internet of things. Information, 14(10), 550. https://doi.org/10.3390/info14100550
24. Aldughayfiq, B., Ashfaq, F., Jhanjhi, N. Z., & Humayun, M. (2023). A Deep Learning Approach for Atrial Fibrillation Classification Using Multi-Feature Time Series Data from ECG and PPG. Diagnostics, 13(14), 2442. https://doi.org/10.3390/diagnostics13142442
25. Zhang, W., Belcheva, V., & Ermakova, T. (2025). Interpretable Deep Learning for Diabetic Retinopathy: A Comparative Study of CNN, ViT, and Hybrid Architectures. *Computers*, *14*(5), 187.
26. Jaffar, H. M., Mohaddes, A., Shiri, E., & Zare, F. (2026). Detection of Rumors in Social Media Using Cluster Graph Convolutional Networks and Big Data Analysis. DOI:10.13140/RG.2.2.24044.22404
27. Shahrabani, M. M. N., & Apanaviciene, R. (2025). Evaluation of smart building integration into a smart city by applying machine learning techniques. *Buildings*, *15*(12), 2031.
28. Muzafar, S., & Jhanjhi, N. Z. (2020). Success stories of ICT implementation in Saudi Arabia. In Employing Recent Technologies for Improved Digital Governance (pp. 151-163). IGI Global Scientific Publishing. DOI: 10.4018/978-1-7998-1851-9.ch008
29. Bansal, A., Sharma, R., Jain, A. K., et al. (2023). Enhancing fashion cloth image classification through hybrid CNN-SVM modeling: A multi-class study. IEEE Conference Proceedings. doi: 10.1109/ICSCSS57650.2023.10169791.
30. Vijayaraj, A., Vasanth Raj, P. T., et al. (2022). Deep learning image classification for fashion design. Wireless Communications and Mobile Computing. DOI:10.1155/2022/7549397
31. Saeed, S., Abdullah, A., Jhanjhi, N.Z. et al. New techniques for efficiently k-NN algorithm for brain tumor detection. *Multimed Tools Appl* **81**, 18595–18616 (2022). https://doi.org/10.1007/s11042-022-12271-x
32. Bai, Q., & Hu, X. (2024). Superposition-enhanced quantum neural network for multi-class image classification. Chinese Journal of Physics. https://doi.org/10.1016/j.cjph.2024.03.026
33. D. Akila, S. R. Raja, J. S, M. Revathi, F. Ashfaq and A. A. Khan, "Text Clustering on CCSI System using Canopy and K-Means Algorithm," 2024 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC), Windhoek, Namibia, 2024, pp. 1-6, doi: 10.1109/ETNCC63262.2024.10767494.
34. Anh, P. T. Q., Thuyet, D. Q., & Kobayashi, Y. (2022). Image classification of root-trimmed garlic using multi-label and multi-class classification with deep convolutional neural networks. Postharvest Biology and Technology. https://doi.org/10.1016/j.postharvbio.2022.111956
35. Experimental Investigation on Mechanical Characterization of Epoxy-E-Glass Fiber-Particulate Reinforced Hybrid Composites, Raffi Mohammed, Irfan Anjum Badruddin, Abdul Saddique Shaik, Sarfaraz Kamangar, and Abdul Azeem KhanACS Omega 2024 9 (23), 24761-24773DOI: 10.1021/acsomega.4c01365
36. Jhanjhi, N.Z. (2025). Investigating the Influence of Loss Functions on the Performance and Interpretability of Machine Learning Models. In: Pal, S., Rocha, Á. (eds) Proceedings of 4th International Conference on Mathematical Modeling and Computational Science. ICMMCS 2025. Lecture Notes in Networks and Systems, vol 1399. Springer, Cham. https://doi.org/10.1007/978-3-031-91005-0_43

 doi:10.20944/preprints202605.1891.v1

37. Clement, D., Agu, E., Suleiman, M. A., Obayemi, J., et al. (2022). Multi-class breast cancer histopathological image classification using multi-scale pooled image feature representation and one-versus-one support vector machines. Applied Sciences. https://doi.org/10.3390/app13010156
38. Chen, X., Deng, Y., Di, C., Li, H., Tang, G., & Cai, H. (2022). High-accuracy clothing and style classification via multi-feature fusion. Applied Sciences. https://doi.org/10.3390/app121910062
39. Kunwar, S. (2026). The garbage dataset (GD): A multi-class image benchmark for automated waste segregation. arXiv preprint. https://doi.org/10.48550/arXiv.2602.10500
40. Sindiramutty, S. R., Jhanjhi, N. Z., Ray, S. K., Jazri, H., Khan, N. A., & Gaur, L. (2024). Metaverse: Virtual Meditation. In Metaverse Applications for Intelligent Healthcare (pp. 93-158). IGI Global Scientific Publishing. DOI: 10.4018/978-1-6684-9823-1.ch003
41. Dutta, P., Dudhoria, S., Paul, T., Acharya, A., & Datta, D. Early Stage Mental Health Screening for Students using Machine Learning Techniques. In *Machine Learning based Approaches for Pedagogical Data Analysis* (pp. 167-186). CRC Press.
42. Abidi, S. M. H., Hassan, S. A., Raza, S. M., & Beliatis, M. J. (2026). Advances in Face Recognition: A Comprehensive Review of Feature Extraction and Dataset Evaluation. *Electronics*, *15*(2), 338.
43. Jiang, D., Shah, A., Yeung, S., Zhu, J., Singh, K., & Goldenberg, G. (2024). Multi-Label Classification for Fashion Data: Zero-Shot Classifiers via Few-Shot Learning on Large Language Models. In *KDIR* (pp. 250-257).