ADROID Conference Series: Engineering and Technology
Volume 2, Issue 3, pp.229-239, 2026

Clothing Classification using Vision Transformer, Efficient-NET and CNN

B. Suvarna, Kilaru Chaitanya, Appikatla Jaswanth
Department of Computer Science and Engineering, Vignan University
suvarna1720@gmail.com, kilaruchaitanya9640@gmail.com, appikatlajaswanthsai@gmail.com

Abstract. Clothing classification is an important task in computer vision. It has key uses in fashion retail, inventory management, and recommendation systems. This paper presents a deep learning framework for clothing classification. It compares Vision Transformer (ViT), EfficientNet, and Convolutional Neural Networks (CNN). The Vision Transformer captures global relationships across image patches. EfficientNet offers a good balance between model accuracy and computational efficiency through compound scaling. CNN models extract local features, such as textures and patterns, that help distinguish visually similar clothing categories. Research was conducted to evaluate the proposed approach on a large labelled clothing dataset. This dataset includes various lighting conditions, backgrounds, and poses. The experimental results show that EfficientNet achieves the highest classification accuracy of 94.09%. It outperforms both the CNN and ViT models in accuracy and efficiency. These findings underscore the effectiveness of modern deep learning models for reliable and scalable clothing classification in real-world situations.

Keywords: Clothing classification, deep learning, Efficient-Net, CNN, Vision Transformer, Kaggle dataset, fashion recommendation, image classification.

1. INTRODUCTION

In recent years, rapid advancements in automation, machine learning, and computer vision have greatly changed the fashion and retail industries. One key application is clothing classification, which helps with automated tagging, inventory management, and personalized recommendations on e-commerce platforms. Clothing classification involves identifying and sorting clothing items into specific categories like shirts, pants, dresses, and jackets. As the demand for smart and scalable systems grows, so does the need for accurate and efficient classification models. Deep learning techniques have largely taken the place of traditional image classification methods because they can learn complex visual patterns more effectively. Convolutional Neural Networks (CNNs) are often used for image classification tasks since they can capture important features like edges, textures, and shapes. However, CNNs struggle to capture long-range dependencies and global context in images. To overcome these limitations, Vision Transformers (ViTs) have emerged as a strong alternative. ViTs view images as sequences of patches and use self-attention mechanisms to understand global relationships between different parts of an image. This allows for a better grasp of complex visual structures compared to methods based on CNNs. EfficientNet is another innovative architecture that aims to achieve high accuracy while reducing computational costs. It uses a compound scaling method that balances network depth, width, and resolution, leading to better performance and efficiency. Driven by the advantages of these models, this study aims to evaluate and compare CNN, Vision Transformer, and EfficientNet architectures for clothing classification tasks.

A. Objective of The Study

This project is primarily aimed at developing an excellent clothing classification system that utilizes efficient net as well as vision transformer and CNN algorithm. Our proposed system will set up the different kinds of clothes that are connected with dresses and hats and pants and shirts and shoes and different kinds of classification. The proposed project uses the efficient net and vision transformers as well as CNN algorithms to achieve the highest accuracy in classifying clothes that can boost the speed of detection in other applications like the online shopping system.

B. Problem Statement

The clothing classification system is a critical requirement as it allows the e-commerce platform to have automated tagging systems and recommendation systems and content search systems. The archaic system of classifying images involves tedious tagging and rudimentary machine learning algorithms, which is incapable of handling real-life complicated scenarios to generate relevant and correct outcomes. The advanced deep learning algorithms such as Efficient Net and Vision Transformer and CNN are used to solve the problem of automated classification of clothing items in the project. The system develops efficient fashion application solution that expands to meet various requirements of the fashion business.

2. RELATED WORK

Recent developments in deep learning have significantly advanced the field of clothing image classification and fashion recommendation systems. Various models, including Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and hybrid architectures, have been widely explored.Bouzidi et al. [1] conducted a comprehensive survey on hybrid CNN–ViT models, highlighting their effectiveness in combining local and global feature extraction for fashion image classification. Nie and Zhao [2] proposed an intelligent clothing design system based on AIGC and virtual simulation, enabling automated and personalized fashion design.Wang et al. [3] introduced a semantic-driven progressive guidance learning approach for clothing-change person reidentification, improving recognition performance under varying clothing conditions. Basak et al. [4] developed a transfer learning framework for image classification, demonstrating improved accuracy and efficiency, especially with limited datasets.Kumar et al. [5] focused on e-commerce applications by classifying clothing reviews using deep neural networks to enhance recommendation systems. Cetin et al. [6] proposed a system for generating and recommending clothing combinations, improving user experience in online shopping platforms.Sun et al. [7] developed a computer vision-based model for detecting clothing insulation and estimating metabolic rates, contributing to smart environment applications. Demelash and Derb [8] applied deep learning techniques for cultural clothing classification, achieving high accuracy in recognizing traditional garments.Kumar et al. [9] explored fashion recommendation systems using visual analytics and deep learning techniques to enhance e-commerce platforms. Pitta et al. [10] utilized pre-trained CNN models for footwear classification, achieving robust performance with reduced computational cost.Rajani et al. [11]proposed a CNN-based ensemble learning approach to improve fashion product recommendation accuracy. Wang et al. [12] introduced a mobile-driven deep learning algorithm using multi-feature attributes for personalized clothing design.Hassan et al. [13] developed optimized neural networks for the Fashion-MNIST dataset using data augmentation and hyperparameter tuning, achieving high classification accuracy. Shin et al. [14] proposed a novel deep learning-based method for clothing image classification, demonstrating improved feature extraction and performance.Medina et al. [15] implemented a real-time clothing classification system integrated with IoT devices, enabling smart applications such as connected thermostats.

.The overall overview of the recent researches published in 2022-2026 reveals the latest achievements in the area of clothing classification with the help of deep learning models, especially the Vision Transformer (ViT), EfficientNet and CNN-based models. These papers underline the growing precision and effectiveness of ViT models, particularly ViT-based versions, in identifying complex clothing patterns and global image understanding, which is essential to such tasks as clothing recognition and fashion retrieval. The study also shows that there are major improvements in fine-grained classification problems with the combination of self-attention mechanisms, better feature extraction strategies, and large-scale and diverse datasets. Furthermore, the application of data augmentation, transfer learning, and hybrid models using CNNs and transformers have been demonstrated to improve the performance, in particular, when it comes to the changes in clothing styles, textures, and environmental conditions. Also, generative data and domain adaptation methods have been applied to enhance the capacity of the models to generalize to unseen fashion items and be robust when applied to real-world scenarios.

3. PROPOSED SYSTEM

Deep learning models are used to identify clothing objects to address such limitations in this system, and they include EfficientNet, Vision Transformer, and CNN. It possesses a complex attribute extraction and classification procedure, which enables the system to accommodate the large scale of fashion products. It is a cost effective, automatic and accurate method of sorting clothing. This begins with user registration and accessing where the user is given an entry into the system. The user can then make the dataset and feed in images. It possesses an efficient backend database where the users and datasets are stored.

Such modules of the system consist of such basic steps as data collection, data pre-processing, and data analysis. The following steps are needed to train the model. When the process of data processing is completed, the system begins training its model with the help of a range of machine learning algorithms to achieve maximum accuracy. Once the system is trained and new user details are provided to the system then the system makes predictions. The system provides an effective system of clothing classification since the user can receive predictions through the system.

4. METHODOLOGY

A. Dataset Overview

The dataset applied in this research is very important in the success of clothing classification activity. It is composed of a massive number of pictures depicting a wide range of clothes of various types in various categories including shirts, pants, dresses, t-shirts, jackets, etc. The pictures are systematized in labelled folders where each folder depicts a variety of clothing classes and makes it achievable to learn through supervision and to train the models efficiently. The dataset is further divided into training, validation and test so that the model can be trained, optimized and tested successfully. In order to deal with the variability in real-life situations, the dataset has a rich diversity of conditions, including varying lighting, backgrounds, poses and occlusions, which are useful in the development of solid models that can extrapolate to unseen data. Besides, the dataset is periodically augmented by methods like random cropping, flipping and color adjustments, to artificially scale the size and variety of the dataset, preventing the model to overfitting and to be effective on novel, unknown clothing items. Such a diversified and carefully selected dataset forms a good base to train state-of-the-art deep learning models, including Vision Transformers and CNN-based models, and plays a key role in reaching a high accuracy of clothing recognition tasks.

Fig. 2. sample images in dataset

B. Data Preprocessing

Preprocessing data plays a very important role in the creation of a deep learning model, particularly in image classification. Preprocessing was used in this paper to make sure that the dataset was presented in an ideal form to be used in training and evaluation. The initial process was image resizing, in which all pictures were resized to a specific size (e.g., 224x224 pixels) to fit the input of the Vision Transformer (ViT) model.

Image normalization was then used to normalize pixel values. It consisted of mapping pixel intensities to between [0, 1] or normalizing them to have a mean of zero and a standard deviation of one, depending on a particular model need. Normalization facilitates quicker convergence of the model in training because it will make the distribution of features similar. The data augmentation techniques were applied to make the training data more diverse and more closely approximated to the model in general. These were random rotation, flipping, zoom and cropping. This sort of transformations is a simulation of real-world changes in the appearance of objects, e.g. clothes viewed at different angles or under different light conditions, and the resulting model which is more resistant to unobservable data. Even the brightness, contrast and saturation were tuned via color jittering which adds to the diversity of the dataset and prevents overfitting the model to a particular lighting setting.

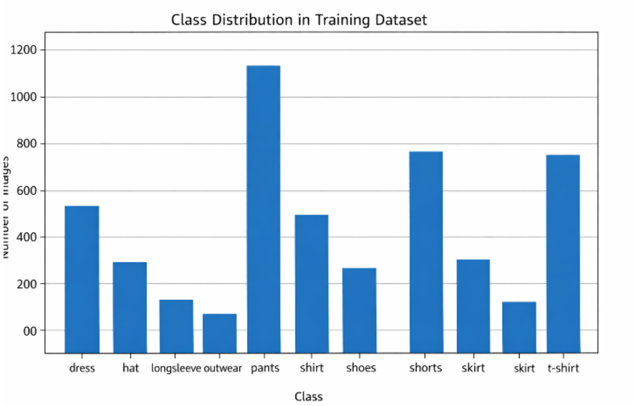

Fig. 3. Distribution of classes in training dataset

C. Model Training

The model training process is carried out based on a deep learning algorithm as a clothing classification which learns on the basis of the prepared dataset. The model is learned with the help of labelled images, where an image is assigned to a certain type of clothes (e.g., shirt, pants, dress).

C.1 Vision Transformer:

This dataset is fine-tuned on the Vision Transformer (ViT) model which uses a transformer-based architecture. In the course of training, the model works with each image and divides it into patches that are subsequently fed to the transformer layers to extract both local and global features. The training is done to optimize the parameters of the model by minimizing cross-entropy loss function which measures the difference between predicted class probabilities and real-life class labels.

Training of the model will be done in different epochs with mini batches to make sure that the model can be generalized effectively. At every epoch, the model changes its weights according to the calculated gradients using the backpropagation. The overfitting is avoided by regularization methods such as dropout and weight decay due to the complexity of the data that could be present. Accuracy and F1-score are the most important key performance measurements used to monitor the training process to allow tracking progress in classification as well as to ensure that the model does not overfit on the training material.

Also, the data augmentation methods are used during training to artificially increase the size of the dataset by performing random transformations such as flipping, rotation, and zoom. This enhances the strength and capability of the model to be extended to real-world changes in clothing appearance. When the model has achieved an acceptable degree of accuracy on the validation set, it is stopped and early stopping is employed to prevent unnecessary computations and overfitting.

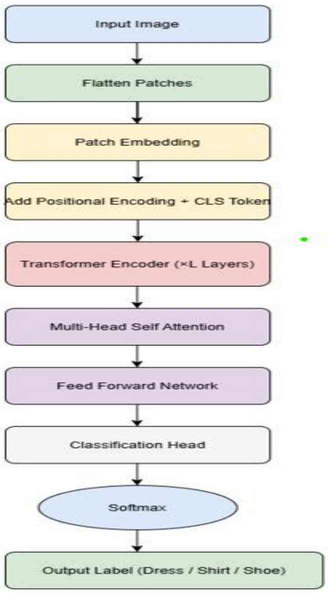

Fig. 4. basic model architecture of ViT model

Vision Transformer (ViT) is an image-classifying architecture, an alternative to traditional convolutional neural networks (CNNs), which is based on the transformer architecture. ViT is not meant to use convolutions, instead, it breaks an image into non-overlapping 16x16 pixel patches which are linearly encoded as vectors. They are regarded as a sequence of patches, which words in natural language transformers are viewed as. ViT model is constructed on the basis of multiple transformer layers which self-attend to learn image connections between all image patches so that the model is able to extract both local and global information.

The Vision Transformer (ViT) typically uses cross-entropy **loss** for classification tasks, which is a common loss function for training models in classification problems.
The loss function for Vision Transformer (ViT) is based on the cross-entropy loss, which is used for classification tasks. This loss function calculates the difference between the predicted class probabilities and the actual class labels, optimizing the model's performance over multiple training epochs. In ViT, the output from the model (after passing through the transformer layers and attention mechanisms) is a probability distribution across the predefined classes. The loss function aims to minimize the negative log likelihood of the correct class label, thus encouraging the model to make accurate predictions.

$$Loss = -\sum_{i=1}^{N} \sum_{j=1}^{C} y_{ij} log\left(p_{ij}\right) \quad (1)$$

Where:

- N is the number of images in the batch.
- C is the number of classes.
- y_{ij} is the binary indicator (0 or 1) of whether class j is the correct label for image i.
- p_{ij} is the predicted probability of class j for image i, which is the output of the softmax function applied to the logits produced by the model.

C.2 ResNet:

ResNet (Residual Network) revolution is among the deep learning models, which eliminate the difficulties in training extremely deep neural networks and suggests residual connections. These residual connections make the network attain residual functions that see it successfully develop networks with hundreds or even thousands of layers. The process of adding skip connections (or identity mappings) between the layers helps the model to bypass some of the layers and to overcome the vanishing gradient problem, which is one of the most serious problems of the traditional deep networks.

The deep structure of ResNet along with its residual connections allow it to be of much use in the complex image classification problems, including the task of clothing classification. ResNet can also be used to identify finer details and features of clothes in the image by being able to train deeper networks hence able to see fined-grained pattern and features even under difficult circumstances, such as different lighting, occlusions, and complex clothing textures. Global average pooling method guarantees that ResNet can work with other clothing types and hence it is a powerful model in terms of separating visually similar objects in clothing datasets. Mathematically, ResNet can be described as:

$$y = F(x, \{W_i\}) + x \quad (2)$$

Where:

- y is the output of the residual block.
- $F(x, \{W_i\})$ represents the residual function to be learned, where x is the input to the block and W_i represents the weights of the network.
- The term x is the identity mapping, which is added to the learned residual to form the final output.

C.3 EfficientNet:

EfficientNet is a series of convolutional neural network (CNN) models that Google researchers created to be highly performing with fewer parameters and a reduced cost of operation. The compound scaling method is the main concept of EfficientNet, and it empirically scaled the depth, width, and input resolution of the network to maximize its performance without adding too much computational complexity.

The models have demonstrated state-of-the-art performance on multiple benchmarks, including ImageNet classification tasks, while being more efficient in terms of computational cost and memory usage compared to traditional models like ResNet and Inception.

The following formula describes the scaling of EfficientNet:

$$New\ model = \alpha^{depth} \cdot \beta^{width} \cdot \gamma^{resolution} \quad (3)$$

EfficientNet has been found to be very efficient in different kinds of image classification like clothing recognition. It can enable the model to be highly accurate with a relatively low memory footprint and low cost of computation by employing the method of scaling a compound.

5. RESULTS AND DISCUSSION

The Vision Transformer (ViT) model's performance was assessed using a confusion matrix, which presents a comparison of predicted and actual labels across all dataset classes. This matrix provides insight into the model's proficiency in identifying specific categories, while also highlighting its shortcomings in the domain of clothing classification.The model gives strong performance in the t-shirt and long-sleeved categories because, with respect to the test cases, it captures the finest distinctions. The t-shirt class shows 47 accurate predictions while the long-sleeved class shows 70 accurate predictions which indicates that both categories were handled successfully. The shoes category achieves strong performance through its 68 accurate predictions.

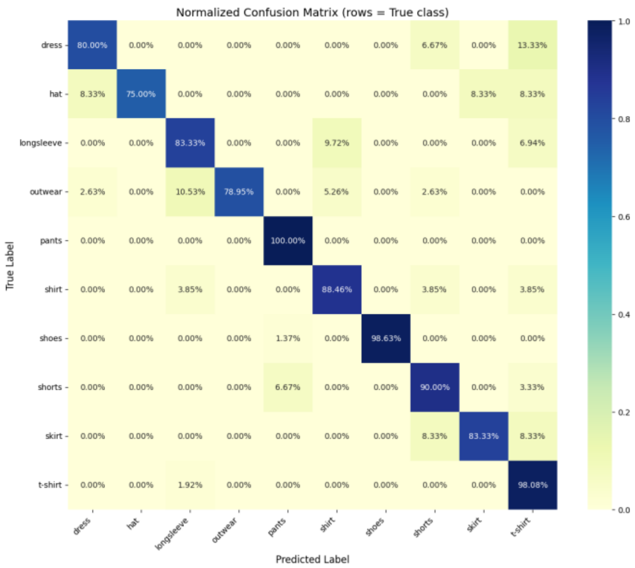

Fig. 5. confusion metrics of Vision Transformer

ViT model shows great performance when dealing with categories such as dress, hat, and long sleeve with a perfect precision and recall. Nevertheless, it has had problems with categories that are classified as Shirt, Pants, and Skirt, and has less precision and recall especially with Skirt. Further enhancements may be on the performance of these more challenging classes.

TABLE I. CLASSIFICATION REPORT FOR CLOTHING CLASSIFICATION MODEL

Class	Precision	Recall	F1-Score	Support
dress	1.00	1.00	1.00	15
hat	1.00	1.00	1.00	12
Long sleeve	1.00	0.97	0.99	72
outwear	0.94	0.89	0.92	38
pants	0.90	0.88	0.89	42
shirt	0.80	0.77	0.78	26
shoes	0.92	0.93	0.93	73
shorts	0.83	0.83	0.83	30
skirt	0.58	0.58	0.58	12
t-shirt	0.90	1.00	0.95	47

Macro Average:

- Precision: 0.89
- Recall: 0.89
- F1-Score: 0.89

Weighted Average:

- Precision: 0.91
- Recall: 0.91
- F1-Score: 0.91

This loss due to training reduces greatly and effective learning can be observed within a few epochs. The validation loss will vary in the beginning but will be steady and this will mean that the model will be generalizing after several epochs. The accuracy progressively climbs to a plateau which proves that the model is getting better and attains great performance at the culmination of the training.

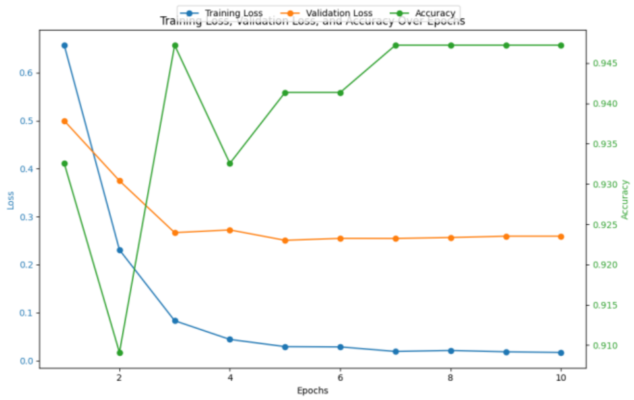

Fig. 6. result curve of ViT model training

The ViT model is also observed to be true in the distinction between the groups of long sleeve, outwear, and t-shirt with the highest degree of accuracy in the form of the values on the diagonal seen in the confusion matrix. There is however some mis-classification between dress and hat and the category of skirt shows a lot of confusion with other category. The model itself can be made better with a more advanced differentiation of similar categories that appear alike.

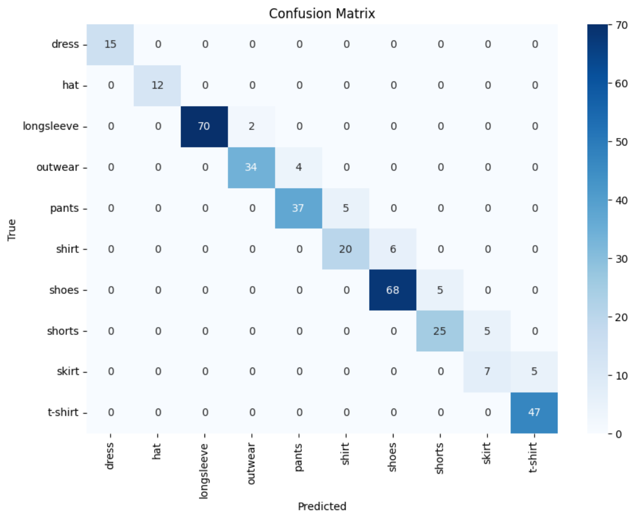

Fig. 7. Confusion matrix of CNN model for clothing classification

The test set confusion matrix of the EfficientNet model demonstrates the great accuracy, especially of such categories as t-shirt, shoes, and pants, and a low number of misclassifications. Nevertheless, dress and hat, and the likes of skirt and other categories are somewhat confused and this means there is difficulty in differentiating these items which are visually similar. Generally, the model attains the highest accuracy of 94.09 which is impressive upon the classification of clothing.

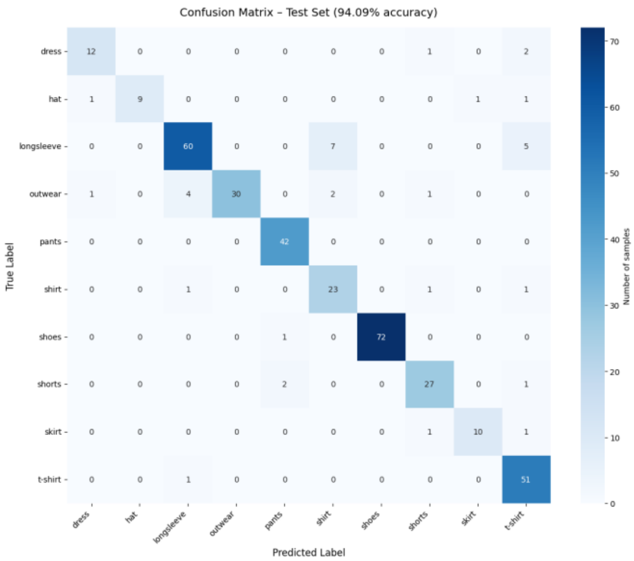

Fig. 8. Confusion matrix of EfficientNet model on the test set (94.09% accuracy)

Fig. 9 presents the precision, recall and F1-score of the EfficientNet model on the test dataset, per-class. The model has a high precision and recall on most of the categories with the highest performance of pants and shoes. The precision and recall scores of skirt and hat are a bit lower, which means that it is harder to correctly classify these categories. The F1-scores in general are indicative of balanced scores in the classes.

The model showed confusion of dress and hat where dress was mistaken as such more frequently. Equally, skirt also had a few misclassifications, which resulted in slightly reduced score in this category as observed in the precision, recall, and F1-scores. Nevertheless, overall performance of the model was high, which means that the difficulties are confined to several categories of visual similarity.

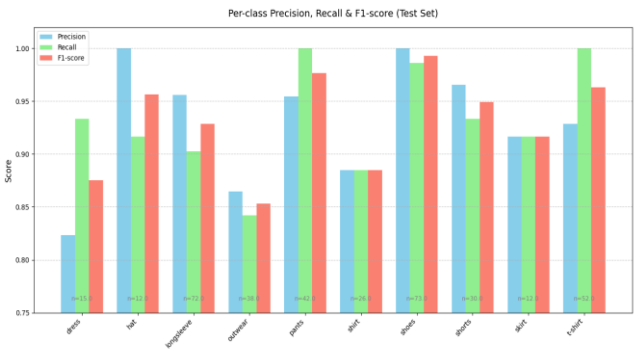

Fig. 9. Per-class Precision, Recall, and F1-score (Test Set)

This bar chart shows the quantity of the samples of the tests per the class in the data. The greatest number of samples is 72 in longsleeve and 79 in shoes whereas those categories as dress and hat have smaller samples, 15 and 12 respectively. This unequal sample sizes may possibly impact on the model performance especially on classes that are underrepresented.

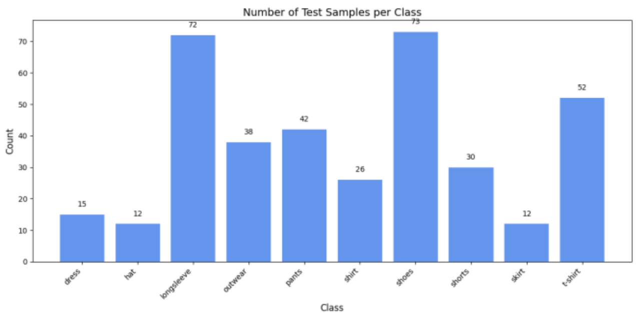

Fig. 10. Number of Test Samples per Class

The EfficientNet model had a high score on the test set with a total accuracy of 94.09 as indicated in the confusion matrix. The model was great in identifying t-shirt, shoes and pants in terms of recall and accuracy. Such groupings always predicted accurately with a small fraction of misclassification evident by the values of the diagonal in the confusion matrix. The model being able to categorically classify these classes is an indicator of its aptitude in addressing everyday apparel that has clear visual attributes.

The values of the per-class precision, recall, and F1-score analysis (Fig. 9) support the findings of the analysis of the confusion matrix. Particularly good results occurred in the categories such as pants and shoes with both high precision and recall with F1-score near to 1.0. The class of the skirt, on the other hand, scored lower which means that the model had some problems in differentiating between skirt and the other similar types. This implies that the model is robust in general, but more refinement can be required to classify under more complicated or problematic categories.

6. CONCLUSION

This paper suggested a clothing classification framework with the ViT and EfficientNet models. EfficientNet model proved to be the most successful with 94.09 percent accuracy on the test set. It was very precise and recallable on such categories like t-shirt, shoes, and pants, but was challenged to categorize more similar classes like skirt and hat. This could be enhanced with the use of a balanced dataset and data augmentation methods to boost the performance of the model particularly of the underrepresented classes. The research indicates the usefulness of EfficientNet in image classification of fashion-related tasks. All in all, the model is a sound approach to real-time clothing recognition in e-commerce and other fashion-related systems.

7. FUTURE ENHANCEMENT

To improve the clothing classification model in the future, more effort can be put on the performance of visually similar categories like skirt and hat through more sophisticated feature extraction methods like multi-scale attention or CNNs/ViT hybrids. Also, the method of data augmentation and synthetic data generation may be introduced to balance the dataset and minimise the effect of the class imbalances. Future studies should consider using bigger pre-trained models because they can enhance the accuracy of the model, particularly on the underrepresented classes. The other possible improvement would be to make the model efficient in real-time inference with mobile and embedded devices to enable efficient and scalable deployment. The model could be made more flexible to adapt to different clothing styles by refining it on domain specific datasets on fashion. Finally, explainable AI methods such as Grad-CAM might be used to offer transparency and interpretability to model predictions in fashion-related AI.

Conflict of Interest

The authors declare no conflicts of interest regarding the current research.

References

[1] S.Bouzidi, G. Hcini, I. Jdey, and F. Drira, "Synergizing cnns and vits: A survey on hybrid models for fashion image classification," *Revista de Informatica´ Teorica´ e Aplicada*, vol. 33, no. 1, pp. 11–27, 2026.

[2] X. Nie and L. Zhao, "Intelligent clothing design system based on aigc and virtual simulation technology," *International Journal of Computer Information Systems and Industrial Management Applications*, vol. 18, pp. 18–18, 2026.

[3] S. Wang, M. Song, J. Shi, S. Shi, F. Yang, and L. Yang, "Clothing semantic-driven progressive guidance learning for clothing-change person re-identification," *Engineering Applications of Artificial Intelligence*, vol. 164, p. 113290, 2026.

[4] C. K. Basak, G. Sarkar, and P. K. Kundu, "Deep transfer learning framework for thermal tomographic image classification: a comparative study," *Multiscale and Multidisciplinary Modeling, Experiments and Design*, vol. 9, no. 1, p. 77, 2026.

[5] A. Kumar, B. Paul, R. Chaudhary, and K. Kumar, "From text to recommendations: Classifying e-commerce clothing reviews using deep neural networks," in *2025 International Conference on Artificial Intelligence and Machine Vision (AIMV)*. IEEE, 2025, pp. 1 – 6.

[6] E. Cetin, M. B. Ozbek,¨ S. Biner, C. Ulus, and M. F. Akay, "Development of a system for creating and recommending combination collections in the e-commerce clothing industry," *Computing and Artificial Intelligence*, pp. 1987–1987, 2025.

[7] Z. Sun, B. Song, S. Gao, S. Zhao, L. Yang, and Y. Zhai, “Dynamic detection models for clothing insulation and metabolic rate based on computer vision,” *Building and Environment*, vol. 283, p. 113372, 2025

[8] A. Demelash and E. Derb, “Habesha cultural cloth classification using deep learning,” *Scientific Reports*, vol. 15, no. 1, p. 14000, 2025.

[9] S. V. Kumar, P. Kumar, and S. Saravanan, “Enhancing e-commerce using fashion recommendation system,” in *2025 International Conference on Visual Analytics and Data Visualization (ICVADV)*. IEEE, 2025, pp. 864–868.

[10] A. B. Pitta, N. R. Pingala, N. V. M. C. Jaladhi, S. Bogolu, and B. Suvarna, “Footwear classification using pretrained cnn models with deep neural network,” in *2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI)*. IEEE, 2025, pp. 1729–1736.

[11] C. Rajani, S. Moturi, M. K. Rao, P. Saiganesh, S. K. Moula, and P. K. Khan, “Enhancing fashion product recommendations through cnn-based ensemble learning,” in *Congress on Control, Robotics, and Mechatronics*. Springer, 2025, pp. 193–207.

[12] Z. Wang, A. Aris, and P. Zhang, “Mobile-driven deep learning algorithm for personalized clothing design using multi-feature attributes.” *International Journal of Interactive Mobile Technologies*, vol. 19, no. 18, 2025.

[13] Y. Hassan, S. Abdelhamid, A. Fathi, P. Edward, and O. G. Elkhouly, “Optimized neural networks for fashion-mnist: A high-accuracy approach using augmentation and tuning,” in *2025 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*. IEEE, 2025, pp. 21–27.

[14] S.-Y. Shin, G. Jo, and G. Wang, “A novel method for fashion clothing image classification based on deep learning,” *Journal of Information and Communication Technology*, vol. 22, no. 1, pp. 127–148, 2023.

[15] A. Medina, J. I. Mendez,´ P. Ponce, T. Peffer, A. Meier, and A. Molina, “Using deep learning in real-time for clothing classification with connected thermostats,” *Energies*, vol. 15, no. 5, p. 1811, 2022.