

XG-ViT: Explainable and Generalizable Vision Transformer for Benchmark Image Classification

Sonia Bouzidi^{1,*†}, Imen Jdey^{1,†} and Fadoua Drira^{1,†}

¹ReGIM-Lab. REsearch Group in Intelligent Machines (LR11ES48), ENIS, Sfax, Tunisia

Abstract

In the ever-evolving fashion industry, sustainability has become a primary focus, driving companies to make significant changes to their business models. Corporate social responsibility is now crucial, emphasizing ethical practices and eco-friendly material sourcing throughout the supply chain. A key part of this collective effort involves the adoption of real-time image classification technologies. This paper introduces the XG-ViT method, an innovative approach that employs a customized Vision Transformer (ViT) for real-time clothing recognition, specifically utilizing the Fashion MNIST dataset. To ensure the ViT model's reliability, we use the Grad-CAM algorithm, which highlights the pixel areas that are most important during predictions in the final attention layer of our XG-ViT model.

Our experimental results demonstrate the state-of-the-art performance of the XG-ViT method on the Fashion MNIST benchmark for real-time image classification. Notable metrics include an impressive accuracy of 92.83%, precision of 92.87%, a loss of 21.13%, an F1 score of 92.63%, and a recall of 92.65%. These outcomes clearly validate the effectiveness of the XG-ViT method in meeting the demands of real-time image classification tasks.

Keywords

Real-time image classification, ViT, Grad-CAM, Fashion MNIST Benchmark, XG-ViT.

1. Introduction

In the current fashion industry landscape, a significant transformation in business strategies is occurring, driven largely by an increased focus on sustainability [1]. Companies are progressively shifting their operations to align with sustainable practices, emphasizing corporate social responsibility throughout their supply chains [2]. This strategic shift involves responsible sourcing, prioritizing eco-friendly materials, and adhering to ethical labor standards [3]. The push towards a sustainable future highlights the importance of integrating advanced technologies, such as real-time image classification, which can act as a catalyst for industry-wide change [4] [5]. Utilizing such technology enables businesses to enhance supply chain efficiency, minimize waste, and increase transparency [6].

Concurrently, the adoption of deep learning has emerged as a viable alternative to traditional machine learning for clothing recognition [7], offering a strategic approach to effectively categorize garments [8] [9]. Deep learning's superior accuracy helps reduce the likelihood of customer returns and dissatisfaction by supporting more informed purchasing decisions [10]. Although convolutional neural networks (CNNs) are widely used for clothing classification, their localized processing and high computational costs limit their practicality for real-time applications [11]. Vision Transformers (ViTs), which use self-attention mechanisms for a global understanding of context, provide a promising alternative, offering both computational efficiency and suitability for real-time tasks [12]. However, enhancing transparency and interpretability in ViT models remains crucial. Class Activation Mapping (CAM) techniques, particularly Grad-CAM, visually highlight the key areas influencing model predictions [13].

WAISS'2024: 1st Euro-Mediterranean Workshop on Artificial Intelligence and Smart Systems, October 15, 2024, Djerba, Tunisia (Co-located with the 17th International Conference on Verification and Evaluation of Computer and Communication Systems (VECoS'2024), October 15-18, 2024, Djerba, Tunisia)

*Corresponding author

†These authors contributed equally.

✉ sonia.bouzidi.doc@enetcom.usf.tn (S. Bouzidi); imen.jdey@fsegs.usf.tn (I. Jdey); Fadoua.drira@enis.tn (F. Drira)

🌐 https://www.researchgate.net/profile/Sonia-Bouzidi-3?ev=hdr_xprf (S. Bouzidi);

<https://www.researchgate.net/profile/Imen-Jdey> (I. Jdey); <https://www.researchgate.net/profile/Fadoua-Drira> (F. Drira)

🆔 0009-0004-7876-5211 (S. Bouzidi); 0000-0001-7937-941X (I. Jdey); 0000-0001-6706-4218 (F. Drira)



© 2025 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Despite the strong performance of ViTs, they are prone to overfitting, which can degrade accuracy and generalization, especially when trained on smaller datasets [14]. Overfitting leads to poor performance on unseen data, undermining the accuracy of clothing classification models.

To address these challenges, this research introduces the XG-ViT methodology, which combines the ViT model with the Grad-CAM algorithm. This integration allows for the visualization of specific pixel areas emphasized during predictions. Furthermore, the study employs k-fold cross-validation (K-FCV) to ensure a comprehensive representation of each class's characteristics and to mitigate potential overfitting issues. The paper is structured as follows: Section 2 explores the vision transformer model, the Grad-CAM algorithm, the overfitting problem, and the K-FCV technique. Section 3 provides a literature review on image classification using the Fashion-MNIST dataset. Section 4 discusses the dataset and the proposed XG-ViT method. Experimental results and discussions are presented in Section 5. Finally, Section 6 concludes the study and suggests future research directions.

2. Background

In this section, we will delve into the key elements of our novel approach: the Vision Transformer, the Grad-CAM algorithm, the fundamental issue of overfitting, and the K-fold cross-validation method. We will also emphasize how each of these components is pertinent to our contribution.

2.1. Vision Transformer

ViT (Vision Transformer) [15] is a recent breakthrough in the area of computer vision. The history of the ViT dates back to 2017, when it was initially designed for natural language processing (NLP) [16]. However, in 2020, its application expanded to computer vision tasks, marking the advent of the "vision transformer" [17]. In 2021, the ViT surpassed the Convolutional Neural Networks (CNNs) in terms of performance and efficiency, especially in image classification [5]. The ViT stands out for its ability to capture complex patterns in images thanks to its attention-based approach, thus offering an effective alternative to traditional architectures based on convolutions. These advances have positioned ViT as a promising method for image classification, opening up new perspectives in the field of deep learning applied to vision.

2.2. Gradient-weighted Class Activation Mapping (Grad-CAM)

Grad-CAM is widely used in image classification tasks to provide a visual interpretation of how deep neural networks make decisions. It achieves this by highlighting significant regions in an input image that influence its final classification [18]. This method produces a class activation map by capturing gradients associated with the target class in the feature maps of the final convolutional layer [19]. By identifying critical regions, Grad-CAM offers insights into the attention mechanisms of the network, thereby enhancing interpretability and transparency in deep learning models [19]. For an illustrative example, please refer to Figure 2.

2.3. Overfitting Problem and K-fold Cross Validation

Overfitting presents a common issue in attention models, where an excessive grasp of training data details impedes effective generalization [20]. Despite achieving high accuracy on training data, overfit models struggle with new data due to interpreting random variations as essential concepts. Addressing this concern, the widely adopted K-fold Cross Validation (K-FCV) technique in machine learning, particularly beneficial with limited datasets, partitions the dataset into k subsets [21]. The model undergoes iterative training k times, using k-1 folds for training and one fold for validation in each iteration, providing a comprehensive evaluation and minimizing the risk of overfitting to specific observations [21].

3. Literature review

Methods for identifying photos have advanced significantly recently, especially when the Fashion MNIST dataset is used. As Table 3 summarizes, the goal of these investigations has been to increase classification accuracy using different deep learning architectures. We summarize significant contributions and place our work within the framework of this developing topic below.

A convolutional neural network method for picture categorization was presented by Kadam et al. [22], and it was tested on the Fashion-MNIST dataset. Their method attained an accuracy of 93.5% on the Fashion-MNIST dataset by experimenting with multiple architectures and fine-tuning several hyperparameters, including activation functions, optimizers, and dropout rates. This showed that their method is effective in categorizing more complex picture.

A Multiple Convolutional Neural Network (MCNN15) with 15 convolutional layers was proposed by Nocentini et al. [23] to improve the accuracy of apparel image classification on the Fashion-MNIST dataset. Their strategy concentrated on solving problems related to garment manipulation in the context of service robotics for senior care. They assessed several neural network models and using the Fashion-MNIST dataset, they were able to obtain a classification accuracy of 94.04%.

Mukherjee et al. [24] introduced a new deep learning framework called OCFormer (One-Class Transformer Network for Image Classification), leveraging ViT. Their approach achieved an accuracy of 92.71%.

Chhabra et al. [25] introduced PatchRot, a self-supervised technique designed specifically for vision transformers. By rotating images and image patches and training the network to predict rotation angles, PatchRot effectively learns both global and local features, achieving an accuracy of 92.6%.

Chhabra et al. [26] introduced PatchSwap, a regularization technique that involves swapping patches between two pictures to create new inputs for transformer regularization. Its straightforward approach facilitates easy extension to semi-supervised environments with minimal effort, achieving an accuracy of 92.6%.

Sun et al. [27] introduced MADPL-net (Multi-layer Attention Dictionary Pair Learning Network), an integrated model that combines convolutional neural network learning schemes, deep encoder learning, and attention dictionary pair learning (ADicL) into a cohesive framework. Their approach achieved an accuracy of 91.24%.

A ViT model, optimized with transformer blocks and self-attention mechanisms, was presented by Abd Alaziz et al. [28] for the classification of fashion images. The model's efficacy across many CNN architectures was demonstrated by its 95.25% accuracy, 95.20% precision, 95.25% recall, and 95.20% F1-score on the Fashion-MNIST dataset.

Li et al. [29] introduced MLPEPS (Multi-Layered PEPS), a tensor network model designed for image classification. MLPEPS employs PEPS to extract features layer by layer from pictures, leveraging these features in Hilbert space to capture pixel correlations while preserving structural information. Their approach achieved a classification accuracy of 90.44%.

Selecting a Vision Transformer for real-time garment recognition is extremely appropriate, given the growing popularity of ViTs in picture classification. ViTs are useful for real-time scenario detection of complex features and garment designs since they are good at seeing global patterns in photos. Moreover, ViTs can perform better than CNN-based models in terms of accuracy and computational efficiency, which is important for tasks demanding quick and accurate classification.

4. Methodology

This section offers a comprehensive explanation of our methodology, which includes the Fashion MNIST benchmark and details our approach involving the integration of ViT with the Grad-CAM method and k-FCV technique.

4.1. Fashion Mnist dataset

Several datasets are used for image recognition in the fashion sector, including Fashion MNIST [30], DeepFashion [31], Fashion IQ [32], FGVCx Fashion [33], iMaterialist [34], and ModaNet [35]. The quantity of research articles over the years is the primary criterion for determining which database is most prevalent. These data sets can be compared in terms of picture classification tasks thanks to this method. Adhering to the guidelines of Keele et al. [36], the approach employed for this analysis comprises manual searches conducted on digital resources, including IEEEExplore Digital Library, SpringerLink, Digital Library, ACM Digital Library, Wiley Online Library, and Science Direct. Based on the information gathered and displayed in Figure 1, "Fashion-MNIST" is the most often used dataset, accounting for nearly 59.9% of research projects.

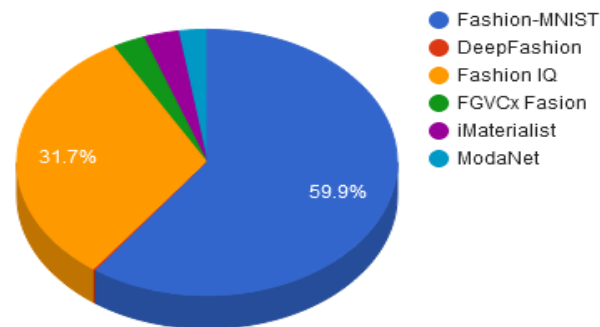





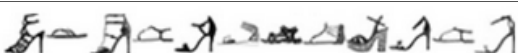






Figure 1: Different database of the fashion industry between 2020 and 2023 in the image classification task [8].

The Fashion MNIST dataset, created by Zalando, contains 70,000 grayscale images showcasing different fashion items in 10 distinct categories [8]. This collection is divided into a training set of 60,000 and a test set of 10,000 pictures, as shown in Table 1. Each image has dimensions of 28x28 pixels and is labeled according to its respective clothing category.

Table 1

Data distribution of Fashion Mnist.

Label	Description	Samples		Examples
		Training	Test	
0	Top	6.000	1.000	
1	Trouser	6.000	1.000	
2	Pullover	6.000	1.000	
3	Dress	6.000	1.000	
4	Coat	6.000	1.000	
5	Sandal	6.000	1.000	
6	Shirt	6.000	1.000	
7	Sneaker	6.000	1.000	
8	Bag	6.000	1.000	
9	Boot	6.000	1.000	

4.2. XG-ViT method and Experimental Setup

In this study, we presented our proposed XG-ViT architecture, and we detailed the implementation process by describing each step of the model’s design.

4.2.1. XG-ViT method:

Building upon the Fashion MNIST benchmark, this study elaborates on our methodology, focusing on key elements. We utilize the ViT approach for classification, aiming to enhance interpretability and pinpoint crucial image features using the Grad-CAM algorithm. Additionally, to address overfitting and ensure representative evaluation across diverse classes, we employ K-FCV.

Our XG-ViT architecture depicted in Figure 2 processes input images of size 72×72 . Initially, pictures are divided into patches, with XG-ViT accommodating variable patch counts. These patches are treated akin to word embeddings in natural language processing, undergoing transformer-based processing. The resulting patch embeddings are then linearly projected into a consistent model dimension D .

Linear Transformation of Patches to Vectors: Patches are transformed into vectors through a learnable linear projection:

$$X_{\text{linear}} = [X_p^1 E; X_p^2 E; \dots; X_p^N E] + E_{\text{pos_linear}}, \quad (1)$$

Here, X_p^i represents each patch, multiplied by a learned projection matrix E , with $E_{\text{pos_linear}}$ incorporating positional information.

Adding Position Tokens: Position tokens are integrated with class tokens and transformed patch vectors:

$$Z_0 = [X_{\text{class}}; X_{\text{linear}}] + E_{\text{pos}}, \quad (2)$$

where Z_0 combines class and linear outputs with positional information.

Encoder Layer with Grad-CAM Integration: The encoder processes Z_0 through multiple blocks, each containing Multi-Head Self Attention (MHSA) and Multi-Layer Perceptron (MLP) components.

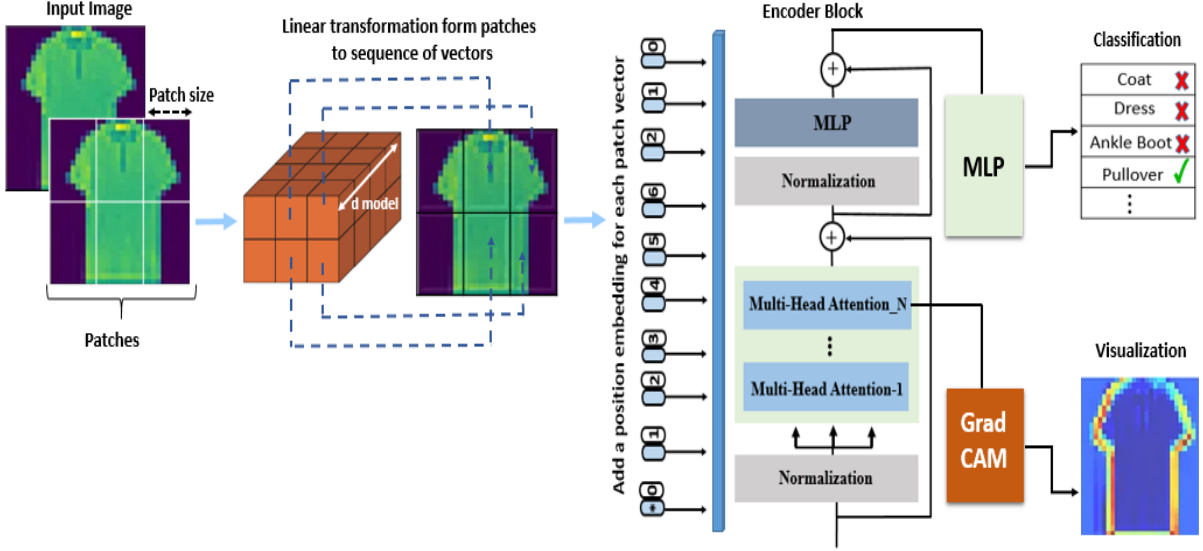


Figure 2: XG-ViT architecture.

MHSA involves scaled dot-product attention:

$$\text{Attention}(Q, K, V) = \text{Softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V, \quad (3)$$

Followed by:

$$\text{MHSA}(Q, K, V) = \text{concat}(\text{Atten}_1, \dots, \text{Atten}_h) W^O, \quad (4)$$

where W^Q, W^K, W^V are learned weights, h denotes attention heads, and W^O is the output matrix.

Classification Layer: The final encoder layer output, enhanced by Grad-CAM, feeds into a classifier:

$$y = \text{Layer Normalization}(Z_L^0). \quad (5)$$

Here, y represents model output, ensuring both accuracy and visual insights from Grad-CAM.

We integrated the k-FCV technique into our methodology, dividing Fashion MNIST into k subsets for comprehensive validation:

In our approach, the Fashion MNIST dataset is divided into k equal subsets. The K-fold cross-validation iterates k times, using one subset for testing and the rest for training per iteration. Performance metrics from all folds aggregate to assess model generalization, ensuring robust evaluation across diverse dataset subsets.

4.2.2. Experimental Setup

Both ViT and XG-ViT implementations were conducted using Python 3.x and TensorFlow on Google Colab. A Tesla K80 GPU was utilized to accelerate model training and results generation. Hyperparameters used in our experiments are detailed in Table 2. The input pictures were resized to 72×72 pixels, with a patch size of 6 for the input sequence. A batch size of 256 was chosen to optimize accuracy. The initial learning rate was set to 0.001 after extensive evaluation. The AdamW optimizer with a momentum of 0.9 was selected based on its superior performance in comparative experiments. Our model architecture includes 5 attention heads, 8 transformer layers, and was trained for 25 epochs. A weight decay of 0.0001 was applied to regularize the model. We employed 5-FCV to ensure robust evaluation and enhance generalization of the model across different subsets of the dataset.

Table 2

The range values of hyperparameters.

Hyperparameter	Definition	value
Image_size	Refers to the dimensions of an image, encompassing its height and width, measured in pixels.	72 × 72
Patch_size	Refers to the dimensions of a small segment within an image. A smaller patch implies a localized area with a compact size, where external influences are more pronounced and extend towards the inner parts.	6
Batch_size	Pertains to the number of data samples processed together during each iteration of model training	256
Learning rate	Governs the extent to which the model adjusts its weights in response to the computed error during each iteration.	0.001
Optimizer choice	TIt adjusts the model weights according to the gradients calculated during backpropagation to minimize the loss function.	AdamW
Data augmentation	Involves increasing the size and diversity of a training dataset by applying various transformations to existing data.	Auto-Augment
weight decay	Is a regularization technique penalizes large weights in the model, helping to prevent overfitting.	0.0001
Heads	The term 'heads' indicates the number of parallel self-attention heads present in the transformer layers.	5
Layer	The quantity of transformer layers defines the depth of the model.	8
N-epochs	is an important hyperparameter that determines how long the model trains and how many times it updates its parameters.	25
N-folds	Represents a partition or a subset created during the division of the dataset.	5

5. Results and Discussion

This section presents the outcomes of our methodology, highlighting the efficacy of our ViT-based approach, integration of K-FCV, and Grad-CAM algorithm. We discuss the achieved results in detail and analyze the strengths and challenges encountered during our experiments.

5.1. Performance Evaluation Metrics

Our XG-ViT method's performance was evaluated using several key metrics to assess its effectiveness in image classification tasks. These metrics provide a comprehensive view of model performance across different aspects:

- Precision: Precision measures the accuracy of positive predictions. It is calculated as:

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

where TP (True Positive) and FP (False Positive) represent correct and incorrect positive predictions, respectively.

- Recall: Recall measures the proportion of correctly predicted positive instances out of all actual positive instances. It is defined as:

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

where FN (False Negative) represents instances incorrectly classified as negative.

- F1 Score: The F1 score is the harmonic mean of precision and recall, providing a balance between the two metrics:

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (8)$$

- Accuracy: Accuracy indicates the proportion of correctly classified instances among all instances:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

where TN (True Negative) represents correctly classified negative instances.

- Top 5 Accuracy: This metric measures the percentage of samples for which the correct label is among the top 5 predicted labels. It is computed as:

$$\text{Top 5 Accuracy} = \left(\frac{C}{N} \right) \times 100 \quad (10)$$

where C is the number of samples correctly classified among the top 5 predictions, and N is the total number of samples.

- Loss: Loss quantifies the difference between predicted values and actual labels in the dataset, providing a measure of model performance:

These metrics collectively assess the performance of our XG-ViT model, providing insights into its classification accuracy, predictive power, and robustness.

5.2. Experimental Results

Our experimental results demonstrate the performance of the XG-ViT model across various evaluation metrics. The model achieved a mean accuracy of 91.30%, precision of 90.64%, recall of 90.56%, training time of 31:56 min, test top-5 accuracy of 99.92%, loss of 22.86%, and F1 score of 90.51%. These metrics collectively indicate the model's robust performance in classifying Fashion MNIST dataset, as summarized in Table 4.

In analyzing the ViT results, we noted a significant disparity between training and testing accuracies. While the training accuracy reached 96.54%, the testing accuracy was 91.30%. This discrepancy suggests that ViT trained on the Fashion MNIST dataset struggles with generalization to new data, indicating potential overfitting. To validate our model's generalizability, we employed a 5-fold cross-validation method. This method effectively mitigates overfitting and enhances the robust utilization of the Fashion MNIST dataset. Through iterative evaluations with varying k values, we consistently found optimal results when $k \geq 5$. Figure 3 illustrates the progression of model evaluation across different k values up to 5. Moreover, this underscores the effectiveness of leveraging the Fashion MNIST dataset to its fullest extent, resulting in improved performance for image classification task.

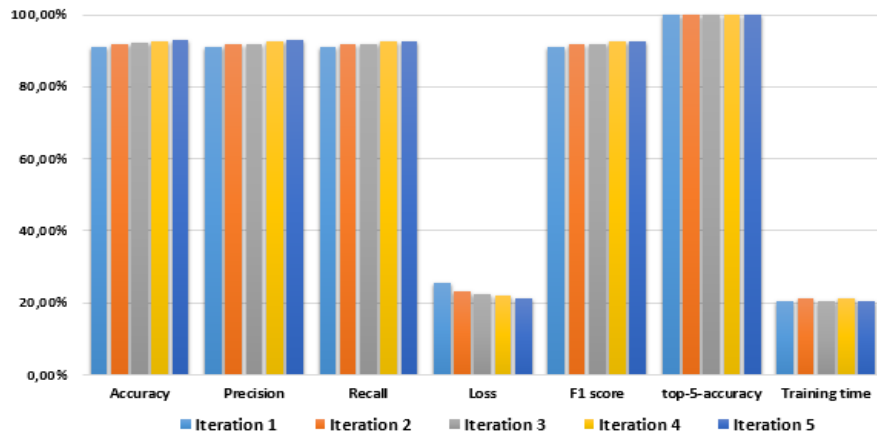


Figure 3: Performance evaluation of ViT model with 5-FCV.

As illustrated in Table 4, the application of k-FCV notably enhanced the performance of the ViT model. We observed significant improvements across key metrics: accuracy improved to 92.83%, precision to

92.87%, recall to 92.65%, training time reduced to 20 minutes and 39 seconds, top-5 accuracy reached 99.94%, and the F1 score increased to 92.63%. Moreover, the loss value decreased substantially to 21.13%. As the number of epochs increased, we observed improved performance. However, this also resulted in a significant increase in processing time. Therefore, it is essential to find a balance to enhance performance while reducing execution time.

Following the implementation of k-FCV, we conducted an ablation experiment by integrating the Grad-CAM algorithm with ViT and k-FCV. Interestingly, our experiment revealed no significant variations in the results. This experiment confirmed that the inclusion of Grad-CAM did not substantially affect the performance of our method. Instead, its primary role was to enhance the interpretability of our model's decision-making process. The integration of Grad-CAM has proven instrumental in elucidating the critical regions within input pictures that influence the final classification of our EViT method. By leveraging Grad-CAM, we visually inspect these pivotal regions, providing a transparent portrayal of their impact on the model's decisions. This integration enhances transparency and instills confidence in the reliability of our model's predictions, as demonstrated in Figure 4.

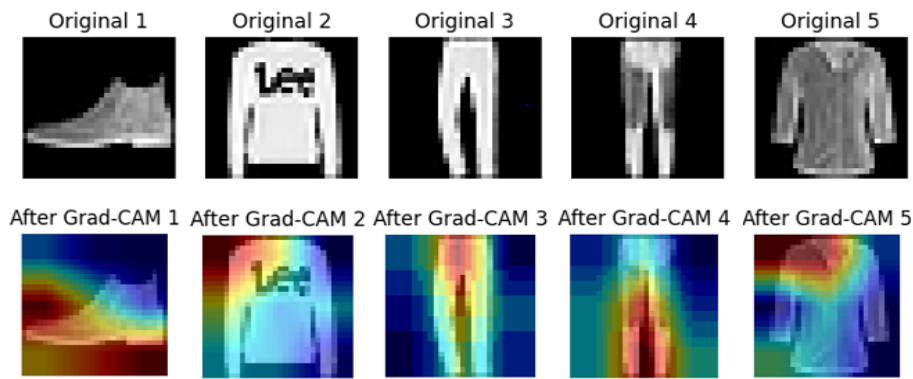


Figure 4: Correct attention regions of the model in the test images.

5.3. Discussion

Our study firmly establishes the superiority of the ViT model over CNNs. The self-attention mechanism in ViTs enhances its ability to capture semantic relationships across long distances, thereby significantly improving task performance. Importantly, existing research often lacks information on the execution times of CNNs, which is crucial for computationally efficient applications. CNNs, with their localized computation nature, may experience prolonged execution times when applied to complex architectures. In contrast, ViTs proficiency in handling global information suggests their potential to mitigate these challenges, opening up new possibilities for resource-intensive image processing applications.

In important aspects beyond accuracy, our suggested XG-ViT model performs better than not just our baseline ViT (Table 4) but also a number of approaches found in the literature (Table 3). Even though some techniques are more accurate than others, XG-ViT has better explainability thanks to Grad-CAM, which gives users a visual glimpse into the model's decision-making process and makes the results easier to understand and believe. Furthermore, in order to present a complete picture of its performance, XG-ViT makes use of a wide range of assessment metrics, including accuracy, precision, recall, F1 score, loss, training time, and test top-5 accuracy. This diversity guarantees that the model is assessed from several angles, such as its capacity to reduce false positives, identify pertinent examples, strike a balance between precision and recall, and monitor training efficiency and prediction mistakes. By lowering overfitting and enhancing generalizability, K-fold cross-validation enhances the model's robustness and guarantees dependability across various data subsets. Thus, XG-ViT shows itself to be a versatile and adaptive model that works well for real-world apparel detection tasks where robustness, explainability, and consistent performance are crucial.

Table 3

Existing approaches in literature performance.

References	Method	Accuracy (%)
[22]	CNN	93.56
[23]	MCNN15	94.04
[24]	OCFormer	92.71
[25]	PatchRot	92.6
[26]	PatchSwap	92.6
[27]	MADPL-net	91.24
[28]	ViT	95.25
[29]	MLPEPS	90.44

We experimented on two additional datasets, Deep Fashion [37], which has 68,000 images (58,000 for training and 10,000 for testing) across 50 categories, and FGVCx Fashion [17], which has 55,000 runway images (50,000 for training and 5,000 for testing) across 50 categories, in order to evaluate the generalizability of our method. These datasets are particularly difficult for classification tasks because they include high-resolution photographs with complicated backgrounds and substantial variation in stance, lighting, and apparel.

In comparison to the other two datasets, our results, which are summed up in Table 4, show a discernible performance gain on the Fashion MNIST dataset. There are several reasons for these variations in performance. First off, the Fashion MNIST dataset features pictures of distinct apparel items against consistent backdrops, which probably made our method’s classification task easier. Secondly, the Fashion MNIST dataset shows a lower training time while being larger (70,000 pictures total; 60,000 for training and 10,000 for testing). The reason for this is that the pictures were relatively simple (lower resolution, 28x28 pixels), which allowed for faster processing, and the extraction of visual information required a less complicated convolutional model. The performance might have also been affected by the fact that the class distribution in Fashion MNIST seems to be more balanced than in the other datasets. Finally, the presence of intricate elements like patterns or accessories in these photos, along with variations in annotation quality or selection biases in the Deep Fashion and FGVCx Fashion datasets, may have added to the longer training times and poorer performance. We can better understand why our method performed differently across datasets by looking at these characteristics, which increases our confidence in its capacity to generalize.

Table 4

Comparison of the Performance to ViT of different datasets.

Method	Accuracy (%)	Precision (%)	Recall (%)	Training time	Test Top 5 accuracy (%)	Loss (%)	F1 score (%)
ViT + Fashion MNIST	91.30	90.64	90.56	31min: 56sec	99.92	22.86	90.51
XG-ViT + Fashion MNIST	92.83	92.87	92.65	20min: 39sec	99.94	21.13	92.63
XG-ViT + Deep Fashion	90.12	87.32	88.15	50min: 29sec	95.02	43.02	89.71
XG-ViT + GVCx Fashion	88.23	81.13	80.35	34min: 22sec	94.13	38.22	81.45

6. Conclusion

In conclusion, this research highlights the effectiveness of employing the ViT model with the Grad-CAM method and K-FCV method for fashion MNIST classification, particularly in the realm of sustainable fashion. Our study demonstrates the capability of our approach to achieve high accuracy in image classification tasks. The Grad-CAM provides valuable insights by highlighting crucial regions in images that influence classification decisions, while K-FCV ensures robust model generalization and helps mitigate overfitting. Our experiments underscore the approach’s resilience to hyperparameter tuning, showcasing its effectiveness in handling diverse classes and intricate image features. Looking ahead, we plan to explore the applicability of our method across various domains to ensure its scalability and broader utility in future research endeavors. In particular, we aim to apply this methodology in

healthcare applications, especially for the skin skin disease recognition tasks, where explainability and reliable predictions are critically important.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

Acknowledgment

This work was supported by the Ministry of Higher Education and Scientific Research of Tunisia through the ReGIM-Lab. REsearch Groups in Intelligent Machines (LR11ES48). The authors gratefully acknowledge this support.

References

- [1] T. S. Thorisdottir, L. Johannsdottir, Corporate social responsibility influencing sustainability within the fashion industry. a systematic review, *Sustainability* 12 (2020) 9167.
- [2] R. Bartkutė, D. Streimikiene, T. Kačerauskas, Between fast and sustainable fashion: The attitude of young lithuanian designers to the circular economy, *Sustainability* 15 (2023) 9986.
- [3] F. James, A. Kurian, Sustainable packaging: A study on consumer perception on sustainable packaging options in e-commerce industry, *Natural Volatiles & Essential Oils* 8 (2021) 10547–10559.
- [4] X. Zhang, F. Guo, T. Chen, L. Pan, G. Beliakov, J. Wu, A brief survey of machine learning and deep learning techniques for e-commerce research, *Journal of Theoretical and Applied Electronic Commerce Research* 18 (2023) 2188–2216.
- [5] S. Bbouzidi, G. Hcini, I. Jdey, F. Drira, Convolutional neural networks and vision transformers for fashion mnist classification: A literature review, *arXiv preprint arXiv:2406.03478* (2024).
- [6] J. Wang, Retracted: Classification and identification of garment images based on deep learning, *Journal of Intelligent & Fuzzy Systems* 44 (2023) 4223–4232.
- [7] I. Jdey, A. Toumi, A. Khenchaf, M. Dhibi, M. Bouhlel, Fuzzy fusion system for radar target recognition, *International Journal of Computer Applications & Information Technology* 1 (2012) 136–142.
- [8] S. Bouzidi, I. Jdey, A. Alimi, A vision transformer approach with l2 regularization for sustainable fashion classification, Available at SSRN 4686032 (2024).
- [9] G. Hcini, I. Jdey, A. Heni, H. Ltifi, Hyperparameter optimization in customized convolutional neural network for blood cells classification, *J. Theor. Appl. Inf. Technol* 99 (2021) 5425–5435.
- [10] I. Jdey, Trusted smart irrigation system based on fuzzy iot and blockchain, in: *International Conference on Service-Oriented Computing*, Springer, 2022, pp. 154–165.
- [11] Y. He, L. Yang, L. Chen, Real-time fashion-guided clothing semantic parsing: A lightweight multi-scale inception neural network and benchmark., in: *AAAI Workshops*, volume 1, 2017.
- [12] B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jégou, M. Douze, Levit: a vision transformer in convnet’s clothing for faster inference, in: *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12259–12269.
- [13] O. Katar, O. Yildirim, An explainable vision transformer model based white blood cells classification and localization, *Diagnostics* 13 (2023) 2459.
- [14] X. Li, B.-B. Zhang, Fv-vit: Vision transformer for finger vein recognition, *IEEE Access* 11 (2023) 75451–75461.
- [15] O. N. Manzari, H. Ahmadabadi, H. Kashiani, S. B. Shokouhi, A. Ayatollahi, Medvit: a robust vision transformer for generalized medical image classification, *Computers in biology and medicine* 157 (2023) 106791.

- [16] S. Jamil, M. Jalil Piran, O.-J. Kwon, A comprehensive survey of transformers for computer vision, *Drones* 7 (2023) 287.
- [17] E. Triantafillou, T. Zhu, V. Dumoulin, P. Lamblin, U. Evci, K. Xu, R. Goroshin, C. Gelada, K. Swersky, P.-A. Manzagol, et al., Meta-dataset: A dataset of datasets for learning to learn from few examples, *arXiv preprint arXiv:1903.03096* (2019).
- [18] H. Jiang, J. Xu, R. Shi, K. Yang, D. Zhang, M. Gao, H. Ma, W. Qian, A multi-label deep learning model with interpretable grad-cam for diabetic retinopathy classification, in: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, 2020, pp. 1560–1563.
- [19] S. Kumar, A. A. Abdelhamid, Z. Tarek, Visualizing the unseen: exploring grad-cam for interpreting convolutional image classifiers, *J. Full Length Artic* 4 (2023) 34–42.
- [20] J. Li, F. Gao, S. Lin, M. Guo, Y. Li, H. Liu, S. Qin, Q. Wen, Quantum k-fold cross-validation for nearest neighbor classification algorithm, *Physica A: Statistical Mechanics and its Applications* 611 (2023) 128435.
- [21] B. Anandan, M. Manikandan, Machine learning approach with various regression models for predicting the ultimate tensile strength of the friction stir welded aa 2050-t8 joints by the k-fold cross-validation method, *Materials Today Communications* 34 (2023) 105286.
- [22] S. S. Kadam, A. C. Adamuthe, A. B. Patil, Cnn model for image classification on mnist and fashion-mnist dataset, *Journal of scientific research* 64 (2020) 374–384.
- [23] O. Nocentini, J. Kim, M. Z. Bashir, F. Cavallo, Image classification using multiple convolutional neural networks on the fashion-mnist dataset, *Sensors* 22 (2022) 9544.
- [24] P. Mukherjee, C. K. Roy, S. K. Roy, Ocformer: One-class transformer network for image classification, *arXiv preprint arXiv:2204.11449* (2022).
- [25] S. Chhabra, P. B. Dutta, H. Venkateswara, B. Li, Patchrot: A self-supervised technique for training vision transformers, *arXiv preprint arXiv:2210.15722* (2022).
- [26] S. Chhabra, H. Venkateswara, B. Li, Patchswap: A regularization technique for vision transformers., in: *BMVC*, 2022, p. 996.
- [27] Y. Sun, G. Shi, W. Dong, X. Xie, Madpl-net: Multi-layer attention dictionary pair learning network for image classification, *Journal of Visual Communication and Image Representation* 90 (2023) 103728.
- [28] H. M. Abd Alaziz, H. Elmannai, H. Saleh, M. Hadjouni, A. M. Anter, A. Koura, M. Kayed, Enhancing fashion classification with vision transformer (vit) and developing recommendation fashion systems using dinova2, *Electronics* 12 (2023) 4263.
- [29] L. Li, H. Lai, Multi-layered projected entangled pair states for image classification, *Sustainability* 15 (2023) 5120.
- [30] J. Xin, T. J. Yi, V. P. Yi, P. J. Yu, Z. A. A. Salam, Convolutional neural network for fashion images classification (fashion-mnist), *Journal of Applied Technology and Innovation* 7 (2023) 11.
- [31] H. An, K. Y. Lee, Y. Choi, M. Park, Conceptual framework of hybrid style in fashion image datasets for machine learning, *Fashion and Textiles* 10 (2023) 18.
- [32] H. Wu, Y. Gao, X. Guo, Z. Al-Halah, S. Rennie, K. Grauman, R. Feris, Fashion iq: A new dataset towards retrieving images by natural language feedback, in: *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2021, pp. 11307–11317.
- [33] S. X. Hu, D. Li, J. Stühmer, M. Kim, T. M. Hospedales, Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9068–9077.
- [34] S. Guo, W. Huang, X. Zhang, P. Srikhanta, Y. Cui, Y. Li, H. Adam, M. R. Scott, S. Belongie, The imaterialist fashion attribute dataset, in: *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 0–0.
- [35] X. Wang, Towards color compatibility in fashion using machine learning, 2019.
- [36] S. Keele, et al., Guidelines for performing systematic literature reviews in software engineering, Technical Report, Technical report, ver. 2.3 ebse technical report. ebse, 2007.
- [37] Z. Liu, P. Luo, S. Qiu, X. Wang, X. Tang, Deepfashion: Powering robust clothes recognition and

retrieval with rich annotations, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1096–1104.