Revista de Informática Teórica e Aplicada - RITA - ISSN 2175-2745
Vol. 33, Num. 1 (2026) 11-27

REVIEW ARTICLE

Synergizing CNNs and ViTs: A Survey on Hybrid Models for Benchmark Image Classification

Sinergizando CNNs e ViTs: Uma Revisão sobre Modelos Híbridos para Classificação de Imagens em Benchmarks

Sonia Bouzidi [1], Ghazala Hcini [1], Imen Jdey [1, 2], Fadoua Drira [1]

Abstract: The rapid expansion of online fashion retail has led to a massive increase in product images, making accurate clothing classification a crucial task. Misclassification not only disrupts customer satisfaction but also contributes to higher return rates and operational inefficiencies. Effective apparel classification enhances search results, improves personalized recommendations, and optimizes inventory management. In response to these challenges, deep learning models have gained attention for visual recognition tasks. This review focuses on the capabilities of convolutional neural networks (CNNs), vision transformers (ViTs), and hybrid models combining both, with an emphasis on the Fashion MNIST benchmark. CNNs are widely used for extracting local spatial features, while ViTs excel at capturing global dependencies through self-attention mechanisms. Recent works have proposed hybrid architectures to leverage the strengths of both models, combining fine-grained feature extraction with broader contextual understanding. This paper offers a comprehensive survey of these hybrid approaches, evaluating their performance in terms of accuracy, scalability, and adaptability. Additionally, we highlight the importance of explainability in fashion classification models and discuss how hybrid solutions enhance model interpretability. Comparative findings suggest that hybrid models outperform individual architectures, making them highly effective for real-world fashion applications demanding both transparency and robust performance.
Keywords: Image Classification — Convolutional Neural Networks (CNNs) — Vision Transformers (Vits) — Hybrid Deep Learning Models — Explainable AI (XAI) — Representation Learning

Resumo: A rápida expansão do varejo de moda online levou a um aumento massivo de imagens de produtos, tornando a classificação precisa de vestuário uma tarefa crucial. A má classificação não só prejudica a satisfação do cliente, como também contribui para maiores taxas de devolução e ineficiências operacionais. Uma classificação eficaz de artigos de vestuário melhora os resultados de busca, aprimora recomendações personalizadas e otimiza a gestão de estoques. Em resposta a esses desafios, os modelos de aprendizado profundo ganharam destaque em tarefas de reconhecimento visual. Esta revisão concentra-se nas capacidades das redes neurais convolucionais (CNNs), dos vision transformers (ViTs) e de modelos híbridos que combinam ambos, com ênfase no benchmark Fashion MNIST. As CNNs são amplamente utilizadas para extrair características espaciais locais, enquanto os ViTs destacam-se na captura de dependências globais por meio de mecanismos de autoatenção. Trabalhos recentes propuseram arquiteturas híbridas para aproveitar os pontos fortes de ambos os modelos, combinando a extração de características refinadas com uma compreensão contextual mais ampla. Este artigo oferece um levantamento abrangente dessas abordagens híbridas, avaliando seu desempenho em termos de precisão, escalabilidade e adaptabilidade. Adicionalmente, destacamos a importância da explicabilidade nos modelos de classificação de moda e discutimos como soluções híbridas melhoram a interpretabilidade dos modelos. Resultados comparativos sugerem que os modelos híbridos superam as arquiteturas individuais, tornando-os altamente eficazes para aplicações de moda no mundo real que exigem tanto transparência quanto desempenho robusto.
Palavras-Chave: Classificação de Imagens — Redes Neurais Convolucionais (CNNs) — Vision Transformers (ViTs) — Modelos Híbridos de Aprendizado Profundo — IA Explicável (XAI) — Aprendizado de Representação

[1] *ReGIM-Lab. Research Groups in Intelligent Machines (LR11ES48), University of Sfax, Tunisia.*
[2] *CITI lab., INSA de Lyon, Inria, France.*
*__Corresponding author__: imen.jdey@fsegs.usf.tn
DOI: http://dx.doi.org/10.22456/2175-2745.147347 • **Received:** 07/05/2025 • **Accepted:** 23/09/2025

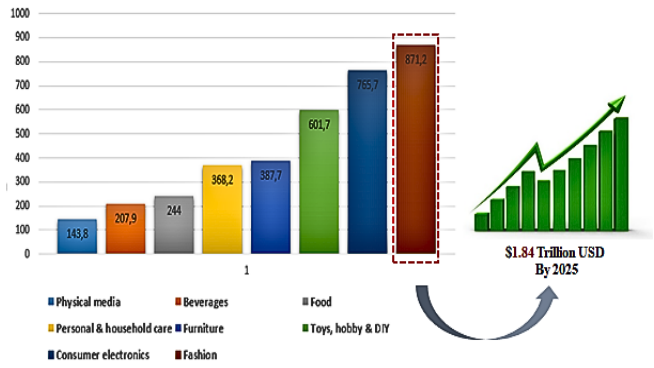

Figure 1. Estimated Annual Spending by Consumer Goods E-commerce Category in 2022 and Projected Online Fashion Sales through 2025.

1. Introduction

The growth of e-commerce has significantly changed how consumers buy and sell products and services online, utilizing electronic payment methods and digital communication between businesses and their customers. Over time, the momentum for e-commerce has increased considerably due to its ease of access, extensive variety of products, and international reach [1]. Technologies like machine learning (ML) and deep learning (DL) play a crucial role in this sector by allowing companies to make real-time decisions based on large data sets [2]. These advanced technologies tackle important issues such as personalized marketing, fraud detection, security measures, optimizing supply chains, and analyzing customer opinions [3].

By implementing deep learning techniques, e-commerce platforms can enhance user experience, streamline operations, and increase sales volumes contributing to their competitiveness in the online market [4]. Among various segments within e-commerce, fashion stands out as the most profitable one, dominating the market share as of 2022 [5]. A report from *Statista* [6] indicates that global online fashion sales reached approximately $871.2 billion, with an expected compound annual growth rate (CAGR) of 14.2% between 2017 and 2025 [6]. According to *UniformMarket* [7], the global apparel market is estimated at USD 1.84 trillion in 2025, with moderate growth ahead. These trends are illustrated in Figure 1, highlighting the significant growth and digital transformation within the global fashion sector. This financial significance highlights the vital importance of the fashion industry in today's rapidly evolving digital economy The swift expansion of online fashion retail brings along certain difficulties, especially when it comes to accurately distinguishing clothing items from a vast array of images on the internet. The sheer volume of product options can overwhelm shoppers, complicating their ability to understand the true features of the garments. This uncertainty often results in higher return rates, which not only diminishes customer satisfaction but also impacts retailers' profitability. Elevated returns have significant repercussions for online businesses, undermining long-term trust and loyalty in their brands.

Deep learning-driven apparel recognition provides a valuable approach to tackling these issues [8]. By using cutting-edge image classification technologies, deep learning models enhance the precision of clothing identification [9], contributing to a richer online shopping experience.These improvements may support customer satisfaction and indirectly reduce certain types of product returns [10], although size and fit remain the primary causes of returns. As the fashion industry continues to evolve within e-commerce, deep learning holds the potential to better align with consumer needs by refining apparel classification methods.

This research investigates the application of CNNs and ViTs as two different deep learning architectures for the classification of fashion images. Traditionally, CNNs are well-suited for image classification tasks because they can learn spatial hierarchies from images. ViTs employ self-attention mechanisms to model long-range dependencies and vary in input size. ViTs are, therefore very popular, owing to their ability to handle different input sizes and their power to learn long-range dependencies. While their advantages, such as flexibility and better contextual awareness come with some shortcomings: high computational cost, huge model sizes, scaling issues with big datasets, interpretability problems, sensitivity to adversarial attacks, and possible generalization shortcomings. Given this set of features, a comparative analysis between both architectures is therefore necessary in understanding their relative strengths and weaknesses.

The paper provides a comprehensive survey of deep learning models, specifically CNNs and ViTs, for garment classification in e-commerce [11]. The key contributions include:

- Comparative Analysis: A thorough evaluation of CNNs and ViTs, analyzing their strengths, weaknesses, and suitability for fashion classification.
- Hybridization Approaches: Discussion on the integration of CNNs and ViTs, exploring hybrid models to leverage both local feature extraction and long-range dependencies.
- Explainability Challenges: Examination of interpretability issues in CNNs and ViTs, emphasizing the need for more transparent AI solutions in e-commerce.

This paper is organized as follows. Section 2 provides background information on deep learning models used in image classification, focusing on the architectures, advantages, and limitations of CNNs and ViTs. This section highlights the core principles behind both approaches and sets the foundation for their comparative analysis. Section 3 presents an extensive literature review, exploring studies that have applied CNNs, ViTs, and hybrid models for fashion classification. It examines key contributions from recent research, identifying trends that motivate further investigation. Section 4 delves

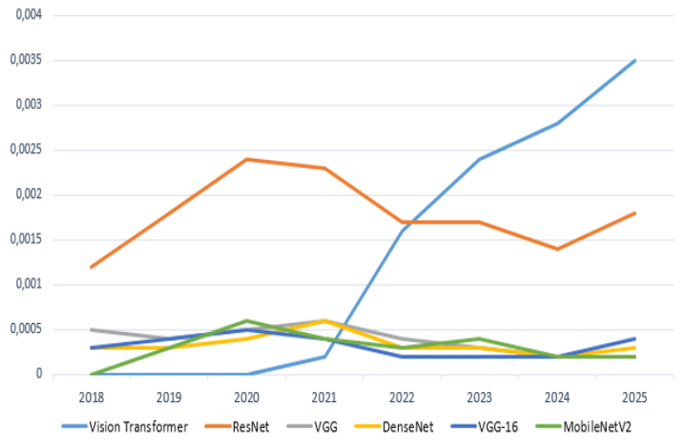

Figure 2. Number of Papers Using ViT and Different CNN Architectures in image Classification Tasks Between 2018 and 2025.

into a detailed discussion on model performance, hyperparameter tuning, and hybridization strategies, emphasizing the challenges and opportunities in integrating CNNs and ViTs. This section also addresses critical issues such as model interpretability, computational efficiency, and domain-specific challenges in e-commerce. Finally, Section 5 concludes the paper by summarizing key findings, discussing their implications, and outlining future research directions aimed at improving deep learning models for fashion classification in e-commerce.

2. Background

Research on architectures specifically designed for image processing has increased dramatically as a result of recent breakthroughs in DL. CNNs and ViTs, two of the most well-known models, have both attracted a lot of interest. These architectures work with DL frameworks that are supervised. Recently, for image classification tasks, ViT have become a competitive alternative to conventional CNNs. As this new approach continues to show promise, studies on ViTs will likely expand in the future as more scholars investigate its possibilities.

To determine which model is more influential in image classification studies, the number of research publications over the past few years is considered. This strategy makes it possible to compare different approaches according to how many articles have been published for each. The analysis was conducted manually, as detailed in Section 3. Transformers have become more common in recent years compared to various CNN architectures, such as ResNet, VGG, DenseNet, VGG16, and MobileNetV2, as illustrated in Figure 2.

We take inspiration from the rivalry between CNNs and my ViTs in image classification to outline the core ideas of each model.

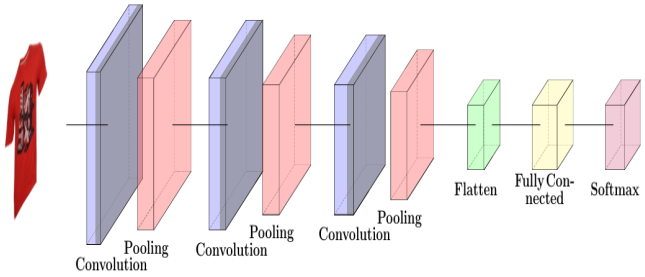

Figure 3. Standard CNN Architecture.

2.1 Convolutional Neural Networks

CNNs [12, 13] originated in the 1980s, but they gained significant traction in image classification following AlexNet's remarkable performance in the 2012 ImageNet Challenge. These networks [14] are designed with a hierarchical structure: convolutional layers play a crucial role by capturing essential features such as edges and textures, while pooling layers help reduce dimensionality to boost computational efficiency. Additionally, normalization techniques like Batch Normalization have been implemented to improve learning reliability and accelerate convergence [15].

The features procured from these processes are forwarded through fully connected layers that weave together spatial and structural information for accurate classification. Ultimately, as depicted in figure 3, the output layer employs a softmax function to assign probabilities to different classes, ensuring precise categorization of images [16].

CNNs are excellent at extracting local features for image identification by virtue of the fact that they capture complex local patterns. The usage of local connections and convolution operation of the CNNs contributes significantly in generalization due to less sensitivity towards noisy data. On the other hand, CNNs suffer some disadvantages such as difficulty of learning long-range dependencies due to their small receptive fields that affect tasks like object detection and image segmentation. In addition, the inductive bias of CNNs can lead to overfitting and thereby make generalization to the new data rather harder. In addition, they usually have a fixed size of input, thus restricting them to process images of any size.

2.2 Vision Transformers

An important development in the field of computer vision is the ViT [17, 18]. It was initially created for natural language processing (NLP) jobs in 2017 [19, 20], marking its beginnings. The name "vision transformer" was coined in 2020 as a result of its use in computer vision [21]. In terms of efficiency and performance, the ViT had surpassed CNNs by 2021, especially when it came to image classification [21]. The ViT's attention-based mechanism is one of its main advantages; it allows it to efficiently identify complex patterns in images [22] and offers a competitive alternative to conventional convolution-based designs [23]. These developments have made the ViT a promising method for classifying images,

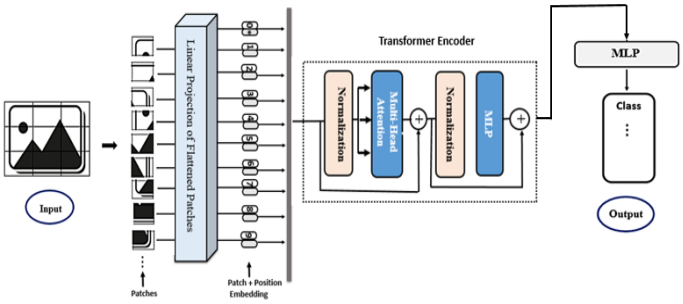

Figure 4. Standard ViT Architecture.

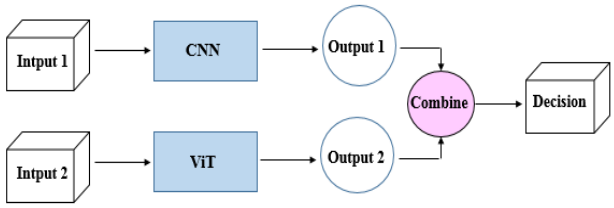

Figure 5. Parallel Hybridization.

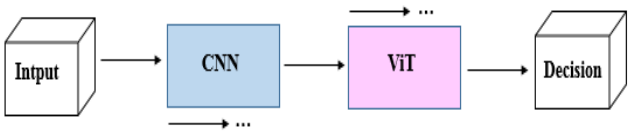

Figure 6. Sequential Hybridization.

opening up new possibilities for applying deep learning to visual problems [23]. To create a sequence of vectors, the ViT first divides an image into fixed-size patches, then linearly embeds each patch, adds positional embeddings, and then processes the resultant vectors through a standard Transformer encoder [18]. For classification purposes, the conventional approach includes appending a learnable "classification token" to the sequence [24], as shown in figure 4.

ViTs excel in modeling a global context through the use of self-attention to capture long-range dependencies and have their greatest successes in tasks such as image segmentation and object detection. The other benefit is scalability: ViTs may be trained to operate on images of nearly any size within a specified number of tokens. In comparison, ViTs face problems in capturing fine-graded local features that are needed in tasks like fine image recognition. Besides, their performance tends to decrease with noise since their performance is heavily dependent on global context; the noisy inputs tend to reduce the output quality. The main disadvantage of ViTs, though, comes from the very high computational costs, especially for larger images or more tokens, because of the quadratic complexity of the self-attention mechanism.

We shall, after a discussion of CNNs and ViTs' strengths and weaknesses, turn our focus to hybrid models based on them. Combining CNNs' skill of local feature extraction locally and with the capacity to extract wider context information from their training on larger corpora in its architecture, the hybrid model capitalizes on both networks' advantages. Thus, the integration of both architectures overcomes the limitations each individual architecture presents, leading to improved performance in tasks requiring the extraction of fine details on local contexts and a broader contextual understanding.

2.3 Forms of hybridization of ViT and CNN

Hybrid deep learning models that combine different architectures have demonstrated improved classification performance [25] by effectively leveraging their complementary strengths. Recently, hybrid architectures have emerged as a strong approach to get around the drawbacks of both CNNs and ViTs [26], including the former's long-range dependency capture, inductive bias, and strict input size requirements, and the latter's weak local feature extraction, noise sensitivity, and high memory costs. Through the combination of their individual benefits, picture classification tasks are performed more efficiently [27][28]. Several integration strategies, such as sequential, parallel, and hierarchical methods, are being investigated in order to maximize the complementarity within these methods.

2.3.1 Parallel Hybridization

CNN and ViT networks are used concurrently on the same input data in parallel hybridization [27]. This approach makes use of ViTs capacity to identify long-range correlations in the data and CNNs capacity for spatial representation (fig. 5). Parallel hybrid models combine these different representations in an effort to improve classification results in terms of accuracy and robustness.

2.3.2 Sequential Hybridization

CNNs and ViTs integrate sequentially, meaning that data is processed step-by-step by one architecture, which processes input in turn and passes the results to the subsequent architecture (fig. 6). A CNN often starts by extracting local features, which the ViT then analyzes to find long-range correlations. During sequential processing, it is crucial to ensure that the representations of these architectures align. This frequently calls for modifications or transformations, such as resizing, reshaping, or adapting the output features from one architecture to fit the input format required by the next architecture [29].

2.3.3 Hierarchical Hybridization

Hierarchical integration combines ViTs and CNNs in a tiered approach, utilizing their unique benefits at various stages of data processing (fig.7). While ViTs focus on understanding global context and long-range dependencies, CNNs are excellent at extracting local characteristics and collecting fine spatial details [30]. Ensuring alignment and consistency among these architectures' representations is essential in hierarchical integration. Smooth feature combination at various layers is made easier by techniques like cross-layer connections and hierarchical fusion.

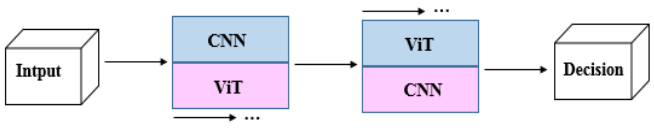

Figure 7. Hierarchical Hybridization.

Careful consideration of the dataset utilized is necessary for evaluating CNN, ViT, and their hybrid counterparts' performance in picture classification tasks. Because it significantly affects the models performance, choosing the dataset is an essential part of the evaluation process. Therefore, choosing the right dataset is essential for determining the accuracy and reliability of these models in image classification applications.

2.4 Dataset

Technologies such as garment recognition, retrieval engines, and automated product recommendation systems are becoming more and more important for fashion-related firms as e-commerce keeps growing. However, because clothing has a wide range of qualities and classification is complex, precisely classifying clothing can be difficult. Multi-class clothing classification problems get more complex when attempting to distinguish between similar classes, since these challenges are magnified. As a result, there's a rising movement to use algorithms made to deal with massive amounts of data related to fashion. Fashion industry image identification is done using a variety of datasets (Table 1), including popular ones like "Fashion MNIST," "DeepFashion," "Watch and Buy," "Fashion IQ," "FGVCx Fashion," "iMaterialist," "ModaNet," "Chictopia10K," and "DeepFashion3D." [31][32] [33] [34] [35] [36] [37] [38] [39].

We use the number of research publications over time as a key parameter to identify the leading dataset and to facilitate comparisons in image classification tasks. The findings, as illustrated in Figure 8, indicate that "Fashion-MNIST" is the most widely used dataset, accounting for around 52.9% of research endeavors. A common benchmark used to assess the performance of computer vision and image classification systems is Fashion-MNIST. A key component of computer vision is image classification [40], which is widely used in the fashion industry to generate comprehensive product descriptions and optimize product tagging procedures. This automatic method is invaluable, particularly when handling large collections, which is a frequent occurrence on many e-commerce sites [41].

Zalando [42] assembled the Fashion MNIST, which is made up of 70,000 grayscale photos of different fashion items in ten categories. As shown in table 2, the dataset consists of a test set of 10,000 photos and a training set of 60,000 images. A label corresponding to one of the ten clothing classes is attached with each image, which is a 28x28 pixel representation [43].

Fashion MNIST is evaluated using several metrics in im-

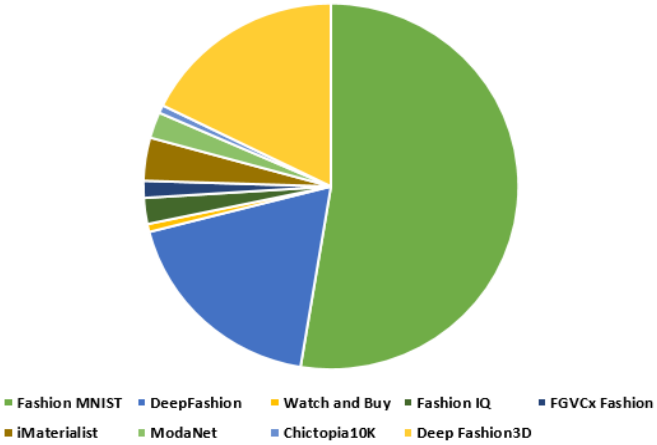

Figure 8. Several Papers That Use Various Fashion Industry Databases for Image Classification Task Between 2020 and 2025.

Table 1. Different Image Databases for Fashion Industry.

Dataset	Number of images	Dataset Link
Fashion MNIST	70 000	⟨https://github.com/zalandoresearch/fashion-mnist⟩
Fashion IQ	77 683	⟨https://github.com/XiaoxiaoGuo/fashion-iq⟩
DeepFashion	800 000	⟨https://www.kaggle.com/datasets/vishalbsadanand/deepfashion-1⟩
Watch and Buy	1,042,178	⟨https://tianchi.aliyun.com/competition/entrance/531893/information⟩
DeepFashion3D	2000	⟨https://drive.google.com/Folder⟩
iMaterialist	50 000	⟨www.kaggle.com/imaterialist-fashion⟩
FGVCx Fashion	55 000	⟨https://sites.google.com/view/fgvc7/home⟩
ModaNet	55 000	⟨https://github.com/modanet⟩
Chictopia10K	17,706	⟨https://files.is.tue.mpg.de/classner/gp/⟩

Table 2. Data distribution of Fashion MNIST [11][44].

Label	Description	Training Samples	Test Samples
0	Top	6000	1000
1	Trouser	6000	1000
2	Pullover	6000	1000
3	Dress	6000	1000
4	Coat	6000	1000
5	Sandal	6000	1000
6	Shirt	6000	1000
7	Sneaker	6000	1000
8	Bag	6000	1000
9	Boot	6000	1000

age classification tasks. A range of measures is typically applied to assess model performance with this dataset.

2.5 Evaluation Metrics

To evaluate the performance of classification models such as CNN and ViT, we used standard evaluation metrics widely adopted in the field, including accuracy, precision, recall, F1 score, and specificity. These metrics provide complementary perspectives on the effectiveness and robustness of the classification process.

3. Literature review on Fashion MNIST Classification

In this section, we provide an overview of research studies focusing on CNN-based methods, ViT, and their hybridization applied to Fashion MNIST dataset. The review process followed the well-established guidelines proposed by Keele et al. [45], which support systematic and transparent literature reviews.

To gather relevant articles, we performed a manual search targeting scientific publications from 2020 to 2025 using six well-known digital libraries: *IEEE Xplore*, *SpringerLink*, *ACM Digital Library*, *Wiley Online Library*, *Google Scholar*, and *ScienceDirect*. Our goal was to compare the evolution of standalone CNN and ViT models as well as recent hybrid approaches combining both.

Three tailored search queries were defined to retrieve relevant research papers:

- **C1:** ”Deep Learning”, ”Vision Transformer (ViT)”, and ”Fashion MNIST classification”
- **C2:** ”Deep Learning”, ”Convolutional Neural Networks (CNN)”, and ”Fashion MNIST classification”
- **C3:** ”Deep Learning”, ”Hybridization between Vision Transformer (ViT) and Convolutional Neural Networks (CNN)”, and ”Fashion MNIST classification”

After collecting the initial results, we eliminated duplicate entries and applied inclusion and exclusion criteria to ensure the scientific relevance and quality of selected studies.

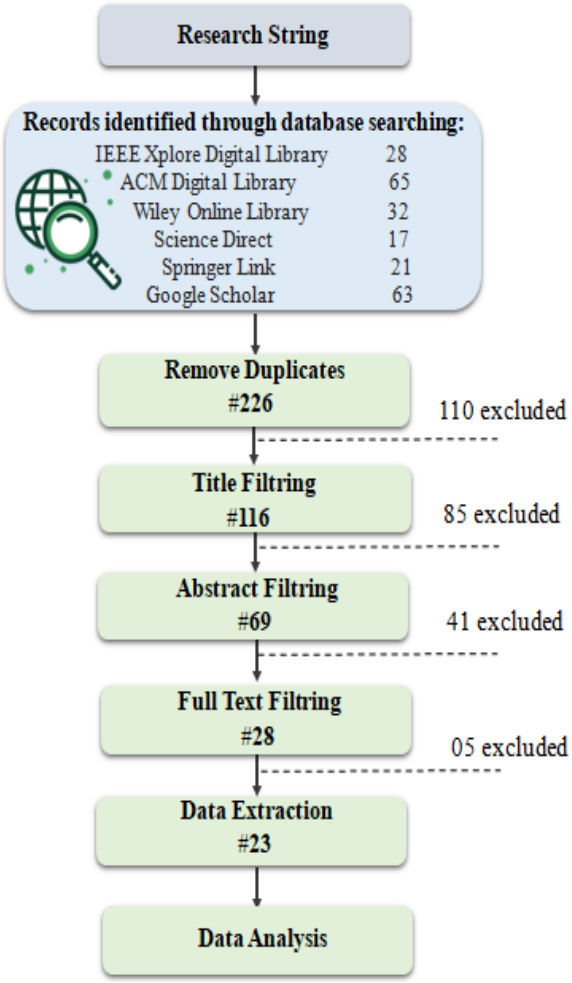

Figure 9. Systematic search and selection process.

- **Inclusion criteria:** peer-reviewed journal or conference papers, written in English, addressing CNN, ViT, or their hybrid use in Fashion MNIST classification, published between 2020 and 2025.
- **Exclusion criteria:** papers not published between 2020 and 2025, non-peer-reviewed articles, workshop papers, studies lacking experimental results, or those not involving the Fashion MNIST dataset.

After applying the inclusion and exclusion criteria, we identified 23 papers focusing on improving Fashion-MNIST classification using CNN and ViT techniques, as illustrated in Figure 9. Figure 10 highlights the significant evolution of CNN and ViT methods over the past few years. More recently, hybrid architectures combining both approaches have demonstrated increasing potential and are gaining momentum in image classification tasks.

3.1 CNNs Methods

As seen below and described in table 3, there has been a notable surge in CNN-based techniques published recently to tackle the challenge of image classification using the Fashion

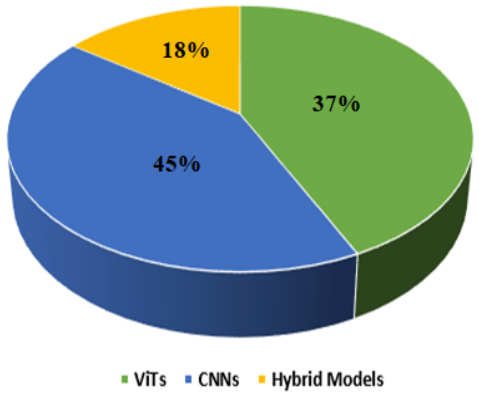

Figure 10. Number of Papers Using CNNs, ViTs and Their hybridization for Fashion MNIST Classification.

MNIST dataset. Garment recognition algorithms were developed by Leithardt et al. [46] in response to the expanding needs of the online fashion business. Using the Fashion-MNIST dataset, their study used CNN models for fashion item classification. They reported using a novel CNN model called cnn-dropout-3 to obtain a considerable accuracy gain, from 89.7% to 99.1%. The present study highlights the efficacy of CNN models in accurately classifying fashion items, which could lead to improved sales tactics within the online fashion sector. Using the Fashion MNIST dataset as a case study, Khanday et al. [19] investigated the impact of filter size hyperparameters on CNNs for picture categorization. Three filter sizes (3×3, 5×5, and 7×7) were examined in the study, and there were always the same number of filters (32). On the Fashion MNIST dataset, it was shown that lower filter sizes—especially the 3×3 filter—achieved higher classification accuracy, reaching 92.68%. Additionally, they computed the loss values for various filter sizes; the findings indicated that a 3×3 filter had a 20% loss, a 5 x 5 filter had a 20.2% loss, and a 7×7 filter had a 24.3% loss. The study however point out that because of the increased processing demands, training times were longer for lower filter sizes. The significance of filter size selection in enhancing CNN performance for image classification tasks is shown by these results, especially when utilizing datasets like Fashion MNIST.

Using the Fashion-MNIST dataset, Nocentini et al. [47] investigated neural network models to enhance clothing image categorization with an emphasis on clothes alteration for elder care. Four neural network models were assessed, one of which, the Multiple Convolutional Neural Network with 15 layers (MCNN15), outperformed earlier benchmarks with a 94.04% classification accuracy on Fashion MNIST. This paper highlights how cutting-edge neural networks can be used to solve real-world problems in eldercare service robotics.

A unique training approach for convolutional layers in CNNs that doesn't rely on supervision or backpropagation was presented by Erkoç et al. [48]. By automatically calculating the ideal number of filters during filter extraction, this method enables feature extraction from unlabeled data and does away with the requirement for weight initialization procedures. The technique lessens the need for labeled data and hyperparameter tuning, simplifying CNN training. Remarkably, the approach achieved high test accuracies of 99.19%, 99.39%, 95.03%, and 90.11%, respectively, on datasets like MNIST, EMNIST-Digits, Kuzushiji-MNIST, and Fashion-MNIST, without using backpropagation or preprocessed data.

To help visually challenged people recognize clothing items, Swain et al. [49] suggested a computer vision solution utilizing CNN and the LeNet model. They trained and validated using the Fashion-MNIST dataset. With an accuracy of 93.7%, the CNN model performed poorly compared to the LeNet model's 98.4% accuracy. These results demonstrate how well deep learning techniques work when it comes to helping visually impaired people with apparel classification tasks.

A novel method for clothing categorization utilizing a CNN architecture known as FFENet (Frequency-Spatial Feature Enhancement Network) was presented by Yu et al. [50]. To extract spatial information from various frequency bands inside the image, their solution combines a Discrete Cosine Transform (DCT) frequency domain enhancement module. After combining these spatial feature maps to create a comprehensive feature map, a clothing feature extraction module processes the feature map further. To achieve the final apparel categorization results, the approach consists of two fully linked layers, global average pooling, and a 1×1 convolution. On the Fashion MNIST dataset, the approach achieved an impressive accuracy of 94.62%, while on the Clothing-8 dataset, it demonstrated promising performance.

A lightweight CNN model called LMFRNet was proposed by Wan et al. [51] and intended for situations with limited resources. A multi-feature block design is introduced by LMFRNet with the goal of minimizing computational load and model complexity without sacrificing performance. Notably, LMFRNet shows notable efficiency in image analysis tasks, with an accuracy of 94.6% on the CIFAR-10 dataset. The robustness and adaptability of the model were further confirmed by the authors by evaluating its performance on a variety of datasets, such as CIFAR-100, MNIST, and Fashion-MNIST. LMFRNet's accuracy of 94.11% on the Fashion-MNIST dataset is noteworthy as it highlights the network's suitability for real-world scenarios with constrained computational resources.

MADPL-net, a revolutionary end-to-end model that combines CNN learning, deep encoder learning, and attention dictionary pair learning (ADicL) into a single framework, was first presented by Sun et al. [52]. By updating dictionary learning and classification simultaneously, MADPL-net seeks to improve overall classification performance while addressing the drawbacks of earlier techniques. The model's classification capacity is further enhanced by the ability to choose image-attentive atoms, made possible by the integration of ADicL. The experimental findings show that MADPL-net outperforms other approaches with an astounding accuracy of

91.24%.

Shin et al. [53] have suggested an improved CNN architecture for apparel picture classification, which combines dropout layers strategically to reduce overfitting with revised convolutional and pooling layers activated by ReLU. They presented a brand-new dynamic learning rate approximation technique that automatically modifies as training progresses, cutting down on training time and enhancing convergence. An evaluation using the Fashion-MNIST dataset showed a noteworthy 93% classification accuracy, which is 4.6% better than the CNN model used as a baseline. This novel method holds potential for improving problems related to apparel picture recognition.

METLEK et al. [54] presented CNNTuner, a unique model created especially for clothing picture classification, in their work. They conducted their evaluations using the Fashion MNIST dataset. CNNTuner utilizes a 15-layer architecture in which the Keras Tuner tool is integrated to achieve hyperparameter adjustment. To guarantee ideal compatibility across all layers, crucial parameters like window sizes, activation functions, and the number of filters in convolutional layers are automatically chosen. The researchers remarkable performance metrics—which all exceeded 0.99 or 99.18%—include F1 scores, recall, precision, specificity, and accuracy. These results were obtained after much experimentation. With potential applications in clothes search and product recommendation, this automated system is a substantial development in clothing image classification.

Haji et al. [55] developed an enhanced CNN model aimed at addressing challenges in fashion item classification, such as variations in styles, textures, and patterns. To boost the model's performance, they utilized a range of image augmentation techniques, including rotation, width and height shifts, zooming, and flipping, which contributed to improving the model's robustness and generalization ability. Furthermore, the incorporation of a Batch Normalization layer helped stabilize the training process and accelerate convergence. By combining these augmentation and normalization methods, the model demonstrated improved adaptability to diverse data. Training the model on an augmented version of the Fashion MNIST dataset resulted in a significant improvement in test accuracy, achieving 91.97%, compared to the baseline CNN's accuracy of 88.5%. These results underscore the effectiveness of the proposed enhancements in boosting the performance of CNNs for fashion classification tasks.

Venkataravanappa et al. [56] proposed an approach that leverages two pre-trained Convolutional Neural Network (CNN) models, namely VGG19 and ResNet50, to classify images in the Fashion MNIST dataset. The method involved customizing the fully connected layers of both models to accommodate the dataset's 10 clothing categories. Training was conducted using the Adam optimizer with a learning rate of 1e-5, over 30 epochs and with a batch size of 64. The findings indicated that VGG19 achieved a training accuracy of 94.58% and a testing accuracy of 91.92%, demonstrating superior performance on unseen data. On the other hand, ResNet50 recorded a higher training accuracy of 99.41% but a slightly lower testing accuracy of 90.25%, hinting at possible overfitting. Additionally, ResNet50 exhibited faster inference times, with a latency of 3.522 seconds compared to 5.439 seconds for VGG19. These outcomes highlight the balance between training accuracy, generalization capability, and inference speed in selecting suitable CNN models for image classification.

Table 3. Existing CNN-based Approaches Evaluated on the Fashion MNIST Dataset

References	Methods	Performances
[46]	CNN-dropout-3	Accuracy: 99.1%
[19]	CNN	-Accuracy: 92.68% -Loss: Filter size 3 × 3 = 20% -Loss: Filter size 5 × 5 = 20.2% -Loss: Filter size 7 × 7 = 24.3%
[47]	MCNN15	Accuracy: 94.04%
[48]	Unsupervised Filter Learning	Accuracy: 90.11%
[49]	-CNN -LeNet	-Accuracy: 93.7% -Accuracy: 98.4%
[50]	FFENet	Accuracy: 94.62%
[51]	LMFRNet	Accuracy: 94.11%
[52]	MADPL-net	Accuracy: 91.24%
[53]	CNN	Accuracy: 93%
[54]	CNNTuner	-Accuracy: 99.18% -Precision: 99% -F1 score: 99% -Recall: 99% -Specificity: 99% -F2 score: 99%
[55]	Enhanced CNN	Accuracy:91.97%
[56]	-VGG19 -ResNet50	-Accuracy: 91.92% -Accuracy: 90.25%

3.2 ViTs Methods

A significant increase in ViT-based techniques has been published recently to address the problem of image classification with the Fashion MNIST dataset, as can be observed below and detailed in table 4.

PatchSwap is a regularization technique established by Chhabra et al. [57]. It generates fresh inputs for transformer regularization by swapping patches between two images. PatchSwap outperforms current state-of-the-art techniques, according to extensive testing. Its straightforward implementation makes it possible to quickly and simply apply it to semi-supervised learning environments. Surprisingly, PatchSwap produced an accuracy of 92.6%, demonstrating how much it may improve model performance.

Abd Alaziz et al. [58] enhanced fashion image categorization using their novel ViT architecture, which uses multilayer perceptron layers and self-attention. They used two

sets of fashion image datasets to test the model's performance against several CNN architectures (VGG16, DenseNet-121, MobileNet, and ResNet50) after adjusting the model's hyperparameters. According to the study, ViT outperformed other models in terms of accuracy, precision, recall, F1-score, and AUC. Specifically, on the Fashion-MNIST dataset, ViT showed its effectiveness in fashion image categorization tasks with a 95.25% accuracy rate. This study highlights the potential of ViT to improve classification methods.

Rodriguez et al. [59] introduced a novel approach to enhance privacy in picture classification by merging two distinct image transformation techniques: one that uses ViT embeddings, and the other that uses convolutional autoencoder (CAE) latent representations. These methods aim to develop robust classification models that can enhance defenses against reconstruction attempts and obscure critical visual information. Performance evaluations were conducted on a range of datasets, including Chest X-ray, Fashion MNIST, and CIFAR-10. The ViT-based method only achieved 87.32% classification accuracy on Fashion MNIST, whereas the CAE-based method achieved 91.43%. These results demonstrate how well both strategies support privacy safeguards while preserving model utility.

To improve ViTs for image identification, Shah et al. [60] introduced a unique technique called SWEKP-based ViT (Stochastic Weighted Composition of Contrastive Embeddings & Divergent Knowledge Dispersion for Heterogeneous Patch Encoding). Through the addition of three new embeddings—Spatially Gated, Fourier Token Mixing, and Multilayer Perceptron Mixture—to conventional linear projections and positional embeddings, this technique enhances the patch encoding module. Additionally, they implemented a Divergent Knowledge Dispersion (DKD) method to improve the spread of information inside the transformer network. They tested their model using four benchmark datasets, and on the Fashion-MNIST dataset, they obtained an accuracy of 93.57%, demonstrating notable gains above traditional ViT models.

In order to enable matrix multiplication, Xu et al. [61] presented an inventive technique called Optronic ViT (OPViT), which combines lens groups with spatial light modulators (SLMs). This represents the introduction of Transformer technology in optical neural networks. By maximizing the advantages of photonic computing, this innovative method significantly lowers computational costs. Reconfigurability, scalability, and low spatial complexity are demonstrated by OPViT. Notably, OPViT delivers remarkable classification results by utilizing two or three layers of optical convolution in conjunction with a single layer of Transformer. By achieving a test accuracy of 88.93% on Fashion-MNIST.

In order to overcome the difficulties presented by noisy images in practical situations, Lin et al. [62] introduced RViT (Robust Vision Transformer), a revolutionary approach for image categorization. By concentrating on generating robust representations that can tolerate noise and uncertainty frequently present in real-world settings, this model seeks to improve the accuracy and robustness of ViTs. When tested on the Fashion-MNIST dataset, the model's accuracy of 94.87% indicates a notable improvement over previous models.

Table 4. Existing ViTs-based Approaches Evaluated on the Fashion MNIST Dataset

References	Methods	Performances
[57]	PatchSwap	Accuracy: 92.6%
[58]	ViT	-Accuracy: 95.25% -Precision: 95.20% -Recall: 95.25% -F1-score: 95.20%
[59]	ViT	Accuracy: 87.32%
[60]	SWEKP-ViT	Accuracy: 93.57%
[61]	OPViT	Accuracy: 88.93%
[62]	RViT	Accuracy: 94.87%

3.3 CNN-ViT Models

As can be seen below and presented in figure 11, there have been an increasing number of proposed algorithms in recent years that hybridize ViT and CNN for image classification tasks utilizing the Fashion MNIST dataset.

The TSD architecture, which combines transformers and CNNs and makes use of novel components like convolutional parameter sharing multi-head attention (CPSA) and local feed-forward network (LFFN) for parallel integration, was first presented by Shao et al. [63]. Both local and distant picture elements are successfully captured by this design. The experimental results show that TSD maintains model efficiency and performs exceptionally well even when trained on small datasets from scratch. TSD obtained a top-1 accuracy of 96.56% on Fashion MNIST with lower computational and parameter costs than both CNN-based and transformer-based models. Because the model does not require positional token encoding, it may more easily modify input resolutions, which is important for a variety of vision applications.

Convolutional Attention Residual Network (CAReNet) is a novel hybrid design that combines attention processes and CNNs in a hierarchical fashion for image classification. It was first presented by Cools et al. [64]. Designed from the ground up to rectify architectural discrepancies discovered in pre-trained models, CAReNet starts with a convolutional block to extract basic features. It mixes parallel blocks with residual units and bottleneck structures to enhance feature extraction and information flow. Additionally, it employs element-wise addition to combine features and max pooling to fine-tune spatial hierarchy. This modular technique iterates three times with varying channel sizes to capture multi-scale information. The architecture also incorporates a CareNet block with bottleneck structures and attention mechanisms modeled after windows and grids. Consolidation and propagation through a second residual unit are subsequently accomplished using feature averaging. CAReNet shows competitive performance across benchmark datasets and demonstrates its effectiveness in many photo classification applications, as evidenced by its

impressive 95% top-1 validation accuracy on Fashion-MNIST.

In order to improve local and global visual feature extraction, Meng et al. [65] presented MixMobileNet, a unique technique that blends CNNs with ViTs. On the Fashion MNIST dataset, this architectural hybrid obtained a high accuracy of 95.37%. With an LFAE (Local-Feature Aggregation Encoder) for extracting local features and a GFAE (Global-Feature Aggregation Encoder) for capturing global features, MixMobileNet presents MMb (MixMobile) blocks. While the LFAE uses adaptive PConv (Partial-Conv) convolutions to capture local features at different sizes within an image, the GFAE improves computation by decreasing channel dimensions and implementing feature attention computations per channel. In a variety of datasets, MixMobileNet exhibits adaptability and efficacy in image classification.

HSViT (Horizontally Scalable Vision Transformer) is a revolutionary approach presented by Xu et al. [66] that aims to address important computer vision difficulties. HSViT tackles problems like limited pre-training on big datasets and mobile computing resources by hierarchically combining CNNs with ViTs. This method takes use of the inductive bias of CNNs inside the ViT framework by using image-level embeddings. In addition to reducing layers and parameters, the horizontally scalable design allows for cooperative training across several nodes. Results from experiments, especially on Fashion-MNIST, show that HSViT can reach up to 10% higher top-1 accuracy than other approaches, suggesting that it can be effectively used for inference and learning on a variety of platforms.

In order to improve clothing classification problems, Yuan et al. [67] suggested a hybrid technique called RST-Net (Residual Spatial Transformer). This method specifically addresses the difficulties caused by differences in shooting distance and angles, which might affect the size and shape of clothing. The suggested approach efficiently processes photos of apparel by combining the advantages of Transformers and CNN. Important spatial information and hierarchical patterns are captured by using ResNet as the basic model to extract local features from the photos. ResNet can handle images of varied sizes and shapes because these local features are processed hierarchically at different levels of the network. A Transformer Encoder, which processes the features at several sizes, is developed in order to capture the global correlations between the features. . This combination improves classification accuracy by allowing RST-Net to fully take into account both local and global information. A subset of the Fashion MNIST dataset, a common dataset for classifying images of apparel, was used to validate the approach. The findings demonstrated the superiority of the suggested method, with RST-Net exceeding both the original ResNet and Transformer models with an accuracy of 93.1%, with improvements of 2.8% and 3.3%, respectively.

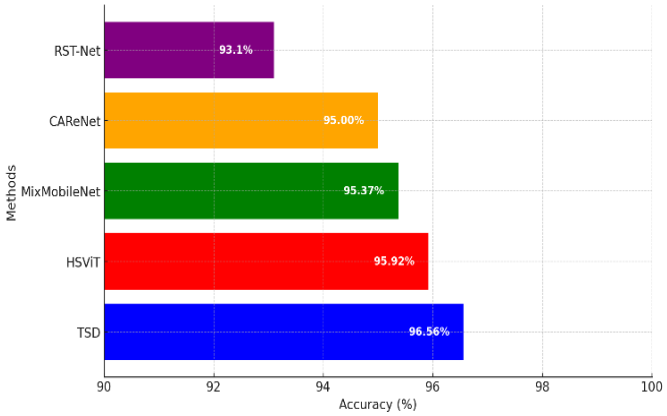

Figure 11. Performance of CNN–ViT Hybrid Models on the Fashion MNIST Dataset

4. Discussion

This section examines several important topics, such as the importance of architectural decisions by contrasting the benefits and limitations of CNN and ViT models. We look at fusion techniques that maximize the performance of both CNN and ViT models by utilizing the advantages of both architectures and the function of hyperparameter adjustment. Additionally, we examine the practical uses of these models in the fashion industry, such as clothing classification, recommendation systems, virtual try-on technologies, and trend analysis. Lastly, we discuss existing limitations and challenges, such as data scarcity, domain adaptation, and model interpretability, offering insights into possible future research directions and areas for improvement.

4.1 Performance of of Each Methods

The model's performance in image classification using the Fashion MNIST dataset is greatly influenced by the architectures used, such as CNNs and ViTs. Different techniques have been used to illustrate the unique benefits and drawbacks of each architecture.

Notable accuracy levels of 99.1% and 99.18% have been attained by CNN-based methods such as CNN-dropout-3 and CNNTuner, respectively. They are quite successful in image classification tasks because of their remarkable capacity to collect localized visual features. Other conventional CNN methods, on the other hand, have produced somewhat lower accuracy rates, ranging between 93.68% and 94.04%. However, ViTs also perform admirably when it comes to classification. Accuracy values of 93.57% and 95.25% have been reported for methods like SWEKP-ViT and ViT. By using attention mechanisms that identify global image patterns, ViTs may be able to increase accuracy and generalization.

When it comes to image classification tasks using the Fashion MNIST dataset, CNNs and ViTs each have unique advantages. CNNs are renowned for their dependability, flexibility, and proven performance, but ViTs use attention processes to offer innovative solutions. A possible development is combin-

ing the two architectures to create hybrid models, in which ViTs improve the model's comprehension of more general contextual patterns while CNNs are better at extracting local characteristics.

4.2 Performance of hybrid CNNs-ViTs Architectures

These Architectures leverage the strengths of CNNs for local feature extraction and the ViTs attention mechanisms for understanding broader context. For example, By fusing the global learning potential of ViTs with the local feature extraction power of CNNs, hybridization between the two is becoming a more and more successful classification technique. The benefits of this integration are demonstrated by the TSD model's 96.56% accuracy, as shown in figure 11. Similarly, HSViT achieved a Top-1 accuracy of 95.92%, while CAReNet and MixMobileNet achieved accuracies of 95% and 95.37%, highlighting the efficacy of these hybrid techniques. These performance results confirm that hybrid CNN–ViT models outperform standalone CNN (Figure 3) and ViT (Figure 4) architectures on the Fashion MNIST dataset. By combining the local feature extraction strengths of CNNs with the global context modeling capabilities of ViTs, these hybrid models achieve higher accuracy and greater effectiveness in image classification tasks. While all results are based on evaluations using the same dataset, differences in training protocols and evaluation settings across studies may exist. These hybrid models may be essential to enhancing visual AI performance as research progresses, opening up new possibilities. Nevertheless, this combination presents several difficulties in spite of their efficacy. The additional complexity and processing needs that result from combining two different architectures are one of the main problems. Because ViTs and CNNs are both intrinsically complicated, combining them greatly increases the amount of resources needed. This additional complexity necessitates more potent technology and lengthens the training period, which can be especially difficult in settings with limited resources. Combining multiple architectures may also make it more difficult to interpret the models and take longer to create because it takes more time and effort to ensure correct integration and optimization.

4.3 Hyperparameters Selection

In deep learning, parameters and hyperparameters serve different purposes in shaping and training models. Parameters, such as weights and biases in neural networks, are values that the model learns and adjusts during the training process. These values are updated to reduce the loss function and enhance the model's performance on the specific task.

On the other hand, hyperparameters are preset parameters that control the model's structure and training procedure. These consist of variables like learning rate, batch size, number of layers, and architecture-specific parameters like CNN filter size or ViT patch size. Hyperparameters, in contrast to parameters, are chosen prior to training and are not learned during training. They have a significant impact on the model's generalization capacity, accuracy, and learning efficiency, making them essential to its overall performance. A model's effectiveness can only be maximized by properly adjusting its hyperparameters.

4.3.1 CNN Hyperparameters:

Hyperparameters play a major role in shaping CNN architecture and fine-tuning performance in image classification applications. The model's capacity to learn and generalize is impacted by these hyperparameters, which also affect other features of the model. Here are a few examples of how altering hyperparameters could enhance model performance:

- Number of layers: A neural network model's complexity and performance can be impacted by the number of hidden layers. In general, more complex connections and patterns in the data may be found by a network with more hidden layers.
- Activation functions: are crucial for altering the input that each neuron receives. For example, the Rectified Linear Unit (ReLU) and the softmax function are commonly employed in deep learning due to their effectiveness and ability to manage the vanishing gradient problem.
- Dropout Rate: prevents overfitting by asymmetrically deactivating connections during training, which promotes more generalization.
- Optimizers: Model weights and biases are adjusted using estimators like Adam, Adagrad, RMSprop, and Stochastic Gradient Descent (SGD) according to the gradient of the loss function. To enhance model performance, each optimizer has a unique collection of hyperparameters that can be adjusted.
- Learning rate: determines the step size at which the model updates its weights throughout the training phase. Choosing the appropriate learning rate is essential since a low rate may lead to slow convergence, while a high rate may cause the model to converge too quickly or overshoot the optimal solution.
- Epochs Number: describes how many times the model runs over the whole training set. While overfitting occurs when a model performs well on training data but poorly on new, unseen data, increasing the number of epochs can improve the model's performance.
- Batch Size: The quantity of data samples the model processes simultaneously during training is referred to as the batch size. It has an impact on learning speed and memory requirements. Model stability and training efficiency can be balanced by selecting the appropriate batch size.
- Kernel Size: affects the convolutional layers' receptive field, which in turn affects the model's ability to extract features from the input data. While smaller kernels

Table 5. Hyperparameters Values of CNN Architectures in Various Studies

Hyperparameters	Values	References
Number of Epochs	12, 50, 20, 100	[46], [48], [54], [47]
Batch size	123, 8, 128	[46], [50], [53]
Activation functions	SoftMax, ReLU and SoftMax, ReLU	[46], [19], [47] [48][49] [53] [54], [51]
Number of layers	2, 13, 5, 3, 15	[46] [49], [19], [47][48], [51], [54]
Optimizers	Adadelta, Adam, SGD, Adam and SGD	[46] [48], [19] [47] [49] [54], [50], [51]
Learning rate	0.001, 0.01, 0.1	[19] [47][49], [50], [51] [53]
Dropout Rate	0.5	[48]

focus on finer details, larger kernels can capture greater patterns. The dataset's characteristics and pertinent attributes should guide the choice of kernel size.

Optimizing CNN architectures performance and efficiency requires fine-tuning their hyperparameters. The optimal hyperparameter choices have a big influence on a model's capacity to learn from data, accelerate its convergence, and adapt to new data. A model's behavior can be tailored to a particular job or dataset by changing these settings, which will increase the model's accuracy and robustness. Using the Fashion MNIST dataset, table 5 offers a thorough examination of the hyperparameters employed in different CNN architectures. By using this technique, practitioners and academics can develop models with more consideration and maximize performance to get the best possible outcomes.

4.3.2 ViT Hyperparameters:

ViTs rely on various hyperparameters to define their architecture and optimize performance in image classification tasks. These hyperparameters influence the model's learning dynamics and its capacity to generalize well to unseen data. Below are some key hyperparameters that can be modified to improve the performance of ViT models:

- Attention head size: The attention mechanism enables the model to concentrate on important regions of the image. The size of the attention heads affects how the model distributes attention across different parts, which can influence its overall performance.
- Number of layers: This determines the depth of the model and its ability to grasp complex relationships in the image data. Increasing the number of layers allows the model to capture more intricate patterns.
- Patch size: In ViTs, images are divided into patches of a fixed size. The size of these patches affects the granularity of the visual information, thereby impacting the model's accuracy and learning process.
- Number of attention heads: This determines the quantity of parallel attention layers utilized. Each head assesses the input data independently. This hyperparameter is crucial for how the model identifies and interprets various features.
- Dropout rate: Is employed to prevent overfitting by randomly dropping units during training, which promotes better generalization.
- Embedding size: The dimension of the embedding space influences how the model encodes the input data. A larger embedding size can offer more information, but it also adds to the model's complexity.
- Activation function: Functions such as ReLU or GELU are utilized to process information within each layer. The selection of the activation function can have a significant effect on the model's performance.
- Optimizer: Optimizers such as Adam, Adagrad, and RMSprop are responsible for adjusting weights during the training phase. The choice of optimizer and its specific parameters can greatly influence the model's performance and efficiency.

Optimizing performance and efficiency in ViT designs requires fine-tuning hyperparameters. These hyperparameters can be changed to affect a model's learning efficiency, rate of convergence, and capacity to generalize to new data. Adapting the behavior of the model to certain tasks or datasets might result in increased robustness and accuracy.

The hyperparameters employed in several ViT architectures using the Fashion MNIST dataset are listed in table 6. When building models, this overview helps researchers and practitioners make well-informed decisions that will improve model performance and lead to better results.

4.4 Fashion Sector Particularities

The integration of sophisticated image classification and object detection techniques in the fashion sector has transformed the shopping experience for consumers and provided brands with valuable insights. Personalized recommendation systems utilize these technologies to deliver tailored outfit suggestions based on individual customer preferences and styles, improving the shopping experience and increasing sales through targeted recommendations.

Virtual try-on solutions allow customers to see how clothing and accessories appear on them in real time, employing object detection and classification to create an interactive online shopping environment. This innovation enhances consumer confidence in their purchases and may result in reduced return rates. Furthermore, by analyzing large datasets of fashion photos, trend analysis technologies employ image recognition

Table 6. Hyperparameters Values of ViTs in various studies

Hyperparameters	Values	References
Epochs	15, 10, 300, 30, 60	[68], [26], [57], [58], [60]
Encoder block	12 and 24, 6	[68], [26] [57]
Batch size	128, 256, 32	[26] [57],[58], [60]
Embedding size	768 and 102, 256	[68], [26]
Number of attention heads	4	[57]
Weight decay	$3 * 10^2$, 0.03	[26], [57]
Patch size	64, 4 and 8, 4	[68], [26], [57]
Learning rate	$1e^4$, $5 * 10^4$, $5 * 10^4$, 0.1, 0.001	[68], [57], [58], [60]
Activation function	GELU, Softmax	[68], [68] [58]
Optimizer	Adam	[68] [26] [58] [60]
Dropout rate	0.1	[57]

and classification to help fashion firms monitor and predict new trends. By highlighting popular styles, colors, and patterns, these insights support strategic decision-making about product design and marketing.

Research in this domain highlights the advantages of these technologies, such as highly accurate methods for categorizing fashion items, which facilitate inventory management and targeted advertising. Object detection algorithms can effectively classify fashion products in images, supporting trend forecasting and brand positioning.

The fashion industry has become a critical focus for image classification tasks in computer vision, surpassing other sectors in both importance and complexity. This is partly attributed to the growing size of datasets, with each new dataset containing an increasing variety of classes. These classes include various subcategories related to size, gender, color, brand, and other features. Moreover, the ever-changing nature of fashion trends means that the classes within these datasets frequently evolve.

These attributes present both challenges and opportunities for enhancing image classification tasks. Classification models must adjust to the fast-paced shifts in styles and trends within the fashion industry, contributing to improvements in accuracy and efficiency over time. By keeping pace with the continuously changing fashion landscape, image classification systems can achieve greater effectiveness in this field.

4.5 Lack of Explainability and Interpretability in CNN and ViT Models

Models based on DL, especially CNNs and ViTs, have greatly succeeded in image classification tasks in e-commerce production. High accuracy is achieved for tasks such as product recognition, recommendation systems, and visual search applications. A major hurdle, however, is present in their effectiveness small concern of explainability and interpretability, which might cast a shadow over their desirability in crucial decision-making situations.

CNNs, even with their hierarchical feature extraction, have a black-box challenge: the filters learned and activations in deep layers are usually not readily interpretable, thus hindering how we know what model is used in arriving at a certain classification decision.

4.6 Challenges in Explainability for E-Commerce Applications

E-commerce, where image classification is an appendix for functions of product categorization, fraud detection, and personalized recommendations, makes opacity in CNN and ViT models a compendium for huge concern. Customers and vendors alike often query about the reason for recommendations and classification decisions, especially when misclassifications or biases exist.

The opacity of the explainability of CNNs and ViTs is becoming a problem in their identification application in e-commerce. With the rollout of more advanced and complex deep learning models, the need for architectures to work with, interpret, and explain the critical aspects of the process becomes urgent for trustworthiness and usability. Future research should generate mixed methods that balance performance-based improvement with interpretability, to widen access and accountability of deep learning models in real-world applications.

4.7 Generalizability of CNN-ViT models

An important characteristic of deep learning models for image classification is their *generalizability*, which refers to the ability to maintain high performance when applied to datasets or domains different from the ones used during training. In image classification tasks, this property is particularly critical, as it reflects the robustness and adaptability of the model to real-world conditions and unseen data distributions. Evaluating a model on diverse benchmarks is therefore a common strategy to assess its generalization capacity beyond the source dataset.

In this context, hybrid CNN-ViT architectures have demonstrated remarkable generalization across multiple domains, including large-scale natural image benchmarks (ImageNet), general object recognition datasets (CIFAR-10/100), and specialized medical imaging datasets (HAM10000 for skin lesion classification). This versatility confirms the hybrid approach's ability to effectively capture both local and global features, making it suitable for a wide range of image classification applications.

In the medical imaging domain, Pacal al. [69] evaluated their CNN-ViT model on two dermatological benchmarks: ISIC 2019 and HAM10000. The model achieved an accuracy of 0.9254 on ISIC 2019 and 0.9501 on HAM10000, surpassing ten advanced CNN models and twenty state-of-the-art ViT models under identical training and evaluation conditions. Similarly, AbuAlkebash et al. [70] proposed a hybrid

YOLOv8-ViT framework tested on HAM10000, achieving balanced precision, recall, and F1-score of 93%, outperforming standalone ViT models. Beyond the medical field, Lu et al. [71] introduced SBCFormer, a lightweight CNN-ViT hybrid, which reached approximately 80% top-1 accuracy on the large-scale ImageNet-1K benchmark while maintaining real-time performance on single-board computers. For general object recognition, Gou and Ren [72] designed a multi-scale CNN-ViT model, achieving 98.65% accuracy on CIFAR-10 and 94.65% on CIFAR-100, confirming the approach's strong adaptability to different image classification scenarios.

These results collectively indicate that CNN-ViT hybrids are not only effective for Fashion MNIST but also excel across heterogeneous benchmarks, from everyday object recognition to highly specialized medical image classification. This confirms their robustness and adaptability, making them promising candidates for deployment in real-world, multi-domain classification tasks.

Conclusion

A thorough examination of deep learning approaches—particularly CNNs and ViTs—for Fashion MNIST image classification demonstrates their substantial contributions to the area. Both CNNs and ViTs exhibit excellent classification performance, with great efficiency and accuracy. Although large labeled datasets are available, challenges still persist. These include class imbalance, noisy or inconsistent labels, and variability in image quality—all of which can hinder model generalization and robustness in real-world applications. Other significant obstacles include restricted interpretability, sensitivity to hyperparameter adjustment, and high processing loads.

Our debate highlights these issues and the need for additional study to address them. Hybrid models, in particular, offer a promising route by combining CNNs' ability to capture fine-grained features with ViTs' ability to model long-range dependencies.

Moving forward, research efforts should focus on improving interpretability, developing more efficient hybrid architectures, leveraging semi-supervised and self-supervised learning techniques, and enhancing model security. These advancements hold significant potential for making deep learning models more adaptable, explainable, and effective in e-commerce applications and beyond.

Acknowledgements

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.

Author contributions

- Sonia Bouzidi: conducted the research and contributed to the writing of the manuscript.
- Ghazala Hcini: guided the manuscript drafting and ensured the coherence of the overall structure.
- Imen Jdey: coordinated the development of the work and supervised the methodological aspects.
- Fadoua Drira: supported the manuscript writing and carried out critical revisions.

References

[1] SHEIKH, A.-S. et al. A deep learning system for predicting size and fit in fashion e-commerce. In: *Proceedings of the 13th ACM conference on recommender systems*. [S.l.: s.n.], 2019. p. 110–118.

[2] RANE, N. L. et al. Artificial intelligence, machine learning, and deep learning for advanced business strategies: a review. *Partners Universal International Innovation Journal*, v. 2, n. 3, p. 147–171, 2024.

[3] WANG, S.; QIU, J. A deep neural network model for fashion collocation recommendation using side information in e-commerce. *Applied Soft Computing*, Elsevier, v. 110, p. 107753, 2021.

[4] JIN, L.; CHEN, L. Exploring the impact of computer applications on cross-border e-commerce performance. *IEEE Access*, IEEE, 2024.

[5] DASH. *E-commerce Statistics Report*. 2025. ⟨https://dash.app/blog/ecommerce-statistics⟩. Accessed: 2023-06-18.

[6] WIZISHOP, S. . *Starting an E-commerce Business: Key Figures and Trends*. 2025. ⟨https://www.statista.com/topics/871/online-shopping/#topicOverview⟩. Accessed: 2025-02-22.

[7] UNIFORMMARKET. *Global Apparel Industry Statistics*. 2025. ⟨https://www.uniformmarket.com/statistics/global-apparel-industry-statistics⟩. Accessed: 2025-08-17.

[8] DARGAN, S. et al. A survey of deep learning and its applications: a new paradigm to machine learning. *Archives of computational methods in engineering*, Springer, v. 27, p. 1071–1092, 2020.

[9] VIJAYARAJ, A. et al. Deep learning image classification for fashion design. *Wireless Communications and Mobile Computing*, Wiley Online Library, v. 2022, n. 1, p. 7549397, 2022.

[10] KIM, J.; FORSYTHE, S. Adoption of virtual try-on technology for online apparel shopping. *Journal of interactive marketing*, Elsevier, v. 22, n. 2, p. 45–59, 2008.

[11] BOUZIDI, S. et al. Convolutional neural networks and vision transformers for fashion mnist classification: A literature review. *arXiv preprint arXiv:2406.03478*, 2024.

[12] HCINI, G.; JDEY, I.; LTIFI, H. Hsv-net: a custom cnn for malaria detection with enhanced color representation. In: IEEE. *2023 International Conference on Cyberworlds (CW)*. [S.l.], 2023. p. 337–340.

[13] HCINI, G. et al. Hyperparameter optimization in customized convolutional neural network for blood cells classification. *J. Theor. Appl. Inf. Technol*, v. 99, p. 5425–5435, 2021.

[14] SLIMANI, N.; JDEY, I.; KHERALLAH, M. Performance comparison of machine learning methods based on cnn for satellite imagery classification. In: IEEE. *2023 9th International Conference on Control, Decision and Information Technologies (CoDIT)*. [S.l.], 2023. p. 185–189.

[15] CONG, S.; ZHOU, Y. A review of convolutional neural network architectures and their optimizations. *Artificial Intelligence Review*, Springer, v. 56, n. 3, p. 1905–1969, 2023.

[16] BRAHMI, W.; JDEY, I. Automatic tooth instance segmentation and identification from panoramic x-ray images using deep cnn. *Multimedia Tools and Applications*, Springer, v. 83, n. 18, p. 55565–55585, 2024.

[17] JAMIL, S.; PIRAN, M. J.; KWON, O.-J. A comprehensive survey of transformers for computer vision. *Drones*, MDPI, v. 7, n. 5, p. 287, 2023.

[18] BOUZIDI, S.; JDEY, I.; ALIMI, A. A vision transformer approach with l2 regularization for sustainable fashion classification. *Available at SSRN 4686032*.

[19] KHANDAY, O. M.; DADVANDIPOUR, S.; LONE, M. A. Effect of filter sizes on image classification in cnn: A case study on cfir10 and fashionmnist datasets. *Int J Artif Intell ISSN*, v. 2252, n. 8938, p. 8938, 2021.

[20] BOUZIDI, S.; JDEY, I.; DRIRA, F. Xg-vit: Explainable and generalizable vision transformer for benchmark image classification. In: *in International Conference on Verification and Evaluation of Computer and Communication Systems (VECoS)*. [S.l.: s.n.], 2025. p. 7–11.

[21] DOSOVITSKIY, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[22] WANG, Y. et al. Vision transformers for image classification: A comparative survey. *Technologies*, MDPI, v. 13, n. 1, p. 32, 2025.

[23] KHAN, S. et al. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, ACM New York, NY, v. 54, n. 10s, p. 1–41, 2022.

[24] ALADHADH, S. et al. An effective skin cancer classification mechanism via medical vision transformer. *Sensors*, MDPI, v. 22, n. 11, p. 4008, 2022.

[25] BANSAL, A. et al. Enhancing fashion cloth image classification through hybrid cnn-svm modeling: A multi-class study. In: IEEE. *2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS)*. [S.l.], 2023. p. 484–489.

[26] LONG, H. Hybrid design of cnn and vision transformer: A review. In: *Proceedings of the 2024 7th International Conference on Computer Information Science and Artificial Intelligence*. [S.l.: s.n.], 2024. p. 121–127.

[27] PENG, Z. et al. Conformer: Local features coupling global representations for visual recognition. In: *Proceedings of the IEEE/CVF international conference on computer vision*. [S.l.: s.n.], 2021. p. 367–376.

[28] DAI, Z. et al. Coatnet: Marrying convolution and attention for all data sizes. *Advances in neural information processing systems*, v. 34, p. 3965–3977, 2021.

[29] YUNUSA, H. et al. Exploring the synergies of hybrid cnns and vits architectures for computer vision: A survey. *arXiv preprint arXiv:2402.02941*, 2024.

[30] HASSANI, A.; SHI, H. Dilated neighborhood attention transformer. *arXiv preprint arXiv:2209.15001*, 2022.

[31] XIN, J. et al. Convolutional neural network for fashion images classification (fashion-mnist). *Journal of Applied Technology and Innovation*, v. 7, n. 4, p. 11, 2023.

[32] AN, H. et al. Conceptual framework of hybrid style in fashion image datasets for machine learning. *Fashion and Textiles*, Springer, v. 10, n. 1, p. 18, 2023.

[33] RAO, J. et al. Watch and buy: A practical solution for real-time fashion product identification in live stream. In: *Proceedings of the 1st Workshop on Multimodal Product Identification in Livestreaming and WAB Challenge*. [S.l.: s.n.], 2021. p. 23–31.

[34] WU, H. et al. 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*. [S.l.: s.n.], 2021. p. 11307–11317.

[35] HU, S. X. et al. 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*. [S.l.: s.n.], 2022. p. 9068–9077.

[36] TZIKAS, T.-R. et al. Towards fashion image annotation: A clothing category recognition procedure. In: *Proceedings of the 11th Hellenic Conference on Artificial Intelligence (SETN 2020), Workshops*. [S.l.]: CEUR Workshop Proceedings, 2020. p. 42–47.

[37] BOUZIDI, S.; JDEY, I.; DRIRA, F. *Towards Explainable Skin Cancer Diagnosis: A Vision Transformer Approach with Grad-CAM Visualization*. 2025. 1-8 p.

[38] PANG, S. et al. An efficient style virtual try on network for clothing business industry. *arXiv preprint arXiv:2105.13183*, 2021.

[39] MENG, X.; CHEN, W.; YANG, B. Neat: Learning neural implicit surfaces with arbitrary topologies from multi-view images. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. [S.l.: s.n.], 2023. p. 248–258.

[40] BHATT, D. et al. Cnn variants for computer vision: History, architecture, application, challenges and future scope. *Electronics*, MDPI, v. 10, n. 20, p. 2470, 2021.

[41] RATHORE, B. Cloaked in code: Ai & machine learning advancements in fashion marketing. *Development*, v. 6, n. 2, 2017.

[42] ARAÚJO, M. A. P. de. *Zalando: An Equity Research Analysis of the E-Commerce Giant*. Dissertação (Mestrado) — Universidade NOVA de Lisboa (Portugal), 2023.

[43] XIAO, H.; RASUL, K.; VOLLGRAF, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[44] TALAVERA, G. L. Image classification using embedded spaces generated by siamese networks. 2019.

[45] KEELE, S. et al. *Guidelines for performing systematic literature reviews in software engineering*. [S.l.], 2007.

[46] LEITHARDT, V. Classifying garments from fashion-mnist dataset through cnns. *Advances in Science, Technology and Engineering Systems Journal*, v. 6, n. 1, p. 989–994, 2021.

[47] NOCENTINI, O. et al. Image classification using multiple convolutional neural networks on the fashion-mnist dataset. *Sensors*, MDPI, v. 22, n. 23, p. 9544, 2022.

[48] ERKOÇ, T.; ESKIL, M. T. A novel similarity based unsupervised technique for training convolutional filters. *IEEE Access*, IEEE, v. 11, p. 49393–49408, 2023.

[49] SWAIN, D. et al. An intelligent fashion object classification using cnn. *EAI Endorsed Trans. Ind. Networks Intell. Syst.*, v. 10, n. 4, p. e2, 2023.

[50] YU, F. et al. Ffenet: frequency-spatial feature enhancement network for clothing classification. *PeerJ Computer Science*, PeerJ Inc., v. 9, p. e1555, 2023.

[51] WAN, G.; YAO, L. Lmfrnet: A lightweight convolutional neural network model for image analysis. *Electronics*, MDPI, v. 13, n. 1, p. 129, 2023.

[52] SUN, Y. et al. Madpl-net: Multi-layer attention dictionary pair learning network for image classification. *Journal of Visual Communication and Image Representation*, Elsevier, v. 90, p. 103728, 2023.

[53] SHIN, S.-Y.; JO, G.; WANG, G. A novel method for fashion clothing image classification based on deep learning. *Journal of Information and Communication Technology*, v. 22, n. 1, p. 127–148, 2023.

[54] ÇETINER, H.; METLEK, S. Cnntuner: Image classification with a novel cnn model optimized hyperparameters. *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, Bitlis Eren University, v. 12, n. 3, p. 746–763, 2023.

[55] HAJI, L. M. et al. Enhanced convolutional neural network for fashion classification. *Engineering, Technology & Applied Science Research*, v. 14, n. 5, p. 16534–16538, 2024.

[56] VENKATARAVANAPPA, V. et al. Conquering fashion mnist with cnns using computer vision by pretrained models: Vgg19 and resnet50. In: AIP PUBLISHING. *AIP Conference Proceedings*. [S.l.], 2024. v. 3131, n. 1.

[57] CHHABRA, S.; VENKATESWARA, H.; LI, B. Patchswap: A regularization technique for vision transformers. In: *BMVC*. [S.l.: s.n.], 2022. p. 996.

[58] ALAZIZ, H. M. A. et al. Enhancing fashion classification with vision transformer (vit) and developing recommendation fashion systems using dinova2. *Electronics*, MDPI, v. 12, n. 20, p. 4263, 2023.

[59] RODRIGUEZ, D.; KRISHNAN, R. Learnable image transformations for privacy enhanced deep neural networks. In: IEEE. *2023 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*. [S.l.], 2023. p. 64–73.

[60] SHAH, S. M. A. H. et al. A hybrid neuro-fuzzy approach for heterogeneous patch encoding in vits using contrastive embeddings and deep knowledge dispersion. *IEEE Access*, IEEE, v. 11, p. 83171–83186, 2023.

[61] XU, C. et al. Transformer in optronic neural networks for image classification. *Optics & Laser Technology*, Elsevier, v. 165, p. 109627, 2023.

[62] LIN, Z. et al. Rvit: Robust fusion vision transformer with variational hierarchical denoising process for image classification. *Guidance, Navigation and Control*, World Scientific, v. 4, n. 03, p. 2441007, 2024.

[63] SHAO, R.; BI, X.-J. Transformers meet small datasets. *IEEE Access*, IEEE, v. 10, p. 118454–118464, 2022.

[64] COOLS, A.; MAHMOUDI, S. A.; BELARBI, M. A. Carenet: A novel architecture for low data regime mixing convolutions and attention. In: *Proceedings of the 20th International Conference on Content-Based Multimedia Indexing (CBMI)*. [S.l.]: IEEE, 2023. p. 1–6.

[65] MENG, Y. et al. Mixmobilenet: A mixed mobile network for edge vision applications. *Electronics*, MDPI, v. 13, n. 3, p. 519, 2024.

[66] XU, C. et al. *HSViT: Horizontally Scalable Vision Transformer*. 2024. ⟨https://arxiv.org/abs/2404.05196⟩. ArXiv preprint arXiv:2404.05196, accessed August 2025.

[67] YUAN, M.; XU, C. A novel approach for clothing classification: Integrating cnn and transformer in rst-net. In: IEEE. *2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*. [S.l.], 2024. p. 1652–1656.

[68] SIKDER, A. S.; ROLFE, S. The power of e-commerce in the global trade industry: A realistic approach to expedite virtual market place and online shopping from anywhere in the world.: E-commerce in the global trade industry. *International Journal of Imminent Science & Technology.*, v. 1, n. 1, p. 79–100, 2023.

[69] PACAL, I. et al. A novel cnn-vit-based deep learning model for early skin cancer diagnosis. *Biomedical Signal Processing and Control*, Elsevier, v. 104, p. 107627, 2025.

[70] ABUALKEBASH, H.; SALEH, R. A.; ERTUNC, H. M. Automated explainable deep learning framework for multiclass skin cancer detection and classification using hybrid yolov8 and vision transformer (vit). *Biomedical Signal Processing and Control*, Elsevier, v. 108, p. 107934, 2025.

[71] LU, X.; SUGANUMA, M.; OKATANI, T. Sbcformer: lightweight network capable of full-size imagenet classification at 1 fps on single board computers. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. [S.l.: s.n.], 2024. p. 1123–1133.

[72] GOU, Q.; REN, Y. Research on multi-scale cnn and transformer-based multi-level multi-classification method for images. *IEEE Access*, IEEE, 2024.