Received: 2026-01-20
Accepted: 2026-03-11
Published Online: 2026-03-30
DOI: 10.70454/JRIST.020102

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926, P- 3117-535X

Neural Networks for Fashion Image Classification and Visual Search

Kamal Kishor Rajak*[1], Dr. Pharindra Kumar Sharma [2]

[1] M.Tech Student, Dept. of CSE, SRCEM, India.
[2] Associate Professor, Dept. of CSE, SRCEM, India.
[1]kamalkishorrajak905@gmail.com
[2]dr.pharindra@gmail.com
*Corresponding Author: kamalkishorrajak905@gmail.com

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

In modern internet commerce and digital retail, fashion classification and visual search are very important jobs. They make it possible to quickly recommend products, find them, and keep track of inventories. Even though deep learning has come a long way, current CNN (convolutional neural network) methods still have trouble with class disparities, overlapping categories, and picking up on fine-grained visual details in manner datasets. To solve these problems, this study suggests Fashion ViT-SA, a hybrid neural network that combines a Vision Transformer (ViT-Base, Patch16) core with a Spatial Attention Module. The model uses the transformer's capacity to encode global context while also using spatial attention to highlight local clothing features, which improves discriminative representation. We used the Deep Fashion multi-modal Dataset and did some preprocessing, such as filtering categories, encoding labels, separating the data into groups, and adding data both online and offline to make sure the class distributions were even. Fashion ViT-SA was used to extract features that were then used to sort items into seven fashion types and to search for items visually based on their content using an Annoy-inspired approximate nearest neighbour index. We trained the model with weighted cross-entropy loss, improved it with Adam-W, and then tested it for accuracy, precision, recall, F1-score, and loss. Experimental findings indicate that Fashion ViT-SA attains 83% accuracy, surpassing a baseline CNN model by 13%, and delivers solid, real-time retrieval performance for visually analogous products. The research underscores the promise of hybrid transformer-based architectures in fashion AI applications, merging classification precision with scalable visual search, thus propelling both scholarly inquiry and practical e-commerce innovations.

Keywords: Fashion Image Classification, Vision Transformer, Spatial Attention, Visual Search, Retrieval, Attention

Introduction

The fashion industry is growing fast, and e-commerce and digital platforms are even faster. This has led to a growing demand for intelligent systems that can efficiently and correctly classify, categorise, and retrieve fashion-related content. Consider online fashion marketplaces like Amazon,

Received: 2026-01-20
Accepted: 2026-03-11
Published Online: 2026-03-30
DOI: 10.70454/JRIST.020102

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926, P- 3117-535X

Flipkart, Zalando and Myntra where there are millions of products across various categories. Due to the enormous number and diversity of these products, computer technologies that are complex enough for automated cataloging and personalized search are required. Conventional image recognition methods that largely relied on handcrafted features such as colour histograms, texture description and geometric comparison have proved to be inadequate in capturing the complexity and subtle differences inherent in fashion items. Identifying the difference between, say, a T-shirt and tank top or potentially noticing small details in prints and fabrics usually would require not just a worldwide frame of reference but also local consideration of features including collars, sleeves or cloth textures. Deep learning, in particular, has revolutionized computer vision by enabling models to directly learn how to represent structures from data (i.e. using neural networks).

This shift was ushered in with the emergence of Convolutional Neural Networks which performed extremely well on large image classification tasks such as ImageNet. Applied to fashion datasets they so far outperformed previous approaches. While convolutional neural networks (CNNs) excel in exploring local receptive fields, it is proved ineffective for fashion photos that contains long-range structural dependent and holistic contextual correlations. Over the past few years, the Vision Transformer (ViT) and its derivatives have completely transformed image classification. While these structures were first defined for natural language processing [1].

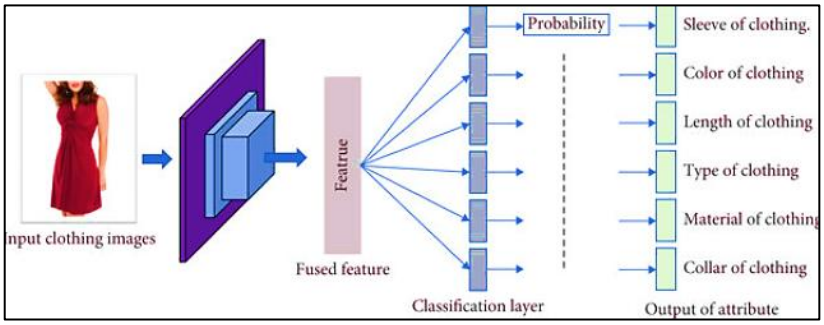

Fig. 1 Fashion Image Classification [2]

These models excel at addressing complex, intricate classification problems, such as those in the fashion industry, since they use self-attention mechanisms to depict global interactions between picture patches. Even with these improvements, the problems of class imbalance, noisy labels, inter-class similarity, and intra-class variability are still big problems when it comes to using neural networks for fashion image analysis. Fashion photographs typically include different poses, lighting, backgrounds, and obstructions. There are also many categories that overlap in meaning, which makes it hard to sort and find them. In e-commerce applications, the task goes beyond only classifying goods. It also includes visual search, which means finding items that look like a query image instead of just text metadata. Content-based visual search is very important in fashion since clients frequently have trouble putting styles, patterns, or designs into words, but they can simply show what they want using sample photographs. Text-based search or metadata tagging alone typically don't do a good job of capturing the subtle aesthetic elements of fashion items. This leads to poor retrieval performance and unhappy customers. In this case, neural networks that can pull out rich, discriminative, and semantically relevant visual embeddings are key to making good visual search engines. The addition of approximate nearest neighbour (ANN) indexing makes scalability even better, allowing for real-time retrieval from huge fashion catalogues. There are still gaps in striking the correct balance between accuracy and quality of retrieval, addressing dataset

Received: 2026-01-20
Accepted: 2026-03-11
Published Online: 2026-03-30
DOI: 10.70454/JRIST.020102

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926, P- 3117-535X

imbalances, limiting data leakage, and making sure that generalisation works well across different fashion categories, even though a lot of study has been done. This research introduces an extensive framework for fashion image classification and visual search, employing advanced neural network architectures, driven by these challenges. To extract global and local discriminative elements necessary for fashion understanding, the study uses a Vision Transformer backbone supplemented with a Spatial Attention Module. A balanced and well selected subset of the DeepFashion Multimodal dataset is used to train the model. By combining data augmentation, stratified product-level splitting, and weighted cross-entropy loss, it corrects class imbalance. This stops leaking. Annoy, an approximate closest neighbour search library, is utilised to index embeddings from the training model at the same time. This provides a scalable visual search engine that quickly detects images that are similar to each other based on cosine similarity[3]–[7].

Everyone take a holistic view of the design by combining various retrieval metrics with classification metrics, such as Precision and Mean Reciprocal Rank (MRR), and dependability, accuracy, memory, and F1-score. Through its four main contributions, this study shows that ViT-based architectures with spatial attention are well suited for fine-grained fashion classification, resolves the pervading class imbalance problem among fashion datasets by means of a multi-faceted preprocessing and augmentation approach, converts the classification framework into a usable semantic visual search system competing against text-reliant systems in state-of-the-art performance metrics, and presents an extensive empirical understanding on the advantages and disadvantages of neural networks in applications of interest to fashion. Computing vision, deep learning and fashion technology meets at the heart of this study which will ensure methodological contributions as well as real world impacts on e-commerce platforms, fashion merchants and customers. This proposed technique can help e-commerce firms by way of facilitating the creation of catalogues, making searches more relevant and easier. All these factors can make customers more engaged and purchase more[8]–[10]. For example, neural networks power visual search engines that intuitively assist people in recovering things. This reduces the stress and difficulty of shopping. Florida State University “Integrating Fashion Knowledge into Neural Networks” Deep learning is reshaping the fashion industry in many aspects, from design and production to marketing and e-commerce. As the fashion industry goes digital and global, the need for smart, automated, scalable solutions for sorting and finding fashion images has never been greater. These problems are well-recognised by neural networks, particularly transformer-based models that bridge the visual complexity with computational know-how. This interesting study contributes to that body of work, developing and validating both a robust attention-based neural network framework for classification and visual search tasks creating an important basis for future studies and applications at the crossroads between artificial intelligence and fashion [11]–[15].

Literature Review

Khan 2024 et.al Smartphone sensor technology advancements have opened up numerous new applications for human activity identification, including health monitoring and personal navigation. By examining data acquired from a variety of sensors, including accelerometers, microphones, gyroscopes, magnetometers, and GPS, the study utilised cutting-edge technology to investigate human movement and position identification. The following datasets were utilised: the Extra Sensory Dataset, the Huawei Locomotion Dataset, and the Continuous In-The-Wild Intelligent Watch Activity Monitoring Dataset of Huawei. Multilayer Perceptrons (MLPs) and Deep Polynomial Neural Networks (DPNNs) were used in that order. First, we employ DPNN algorithms for deep learning-based feature extraction; second, we use MLP (Multi Layer Perceptron) with

Received: 2026-01-20
Accepted: 2026-03-11
Published Online: 2026-03-30
DOI: 10.70454/JRIST.020102

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926, P- 3117-535X

manual feature extraction methods, including LPCC, step length, signal intensity area, spectral, and sound features, for activity recognition. Extensive study led to a high level of accuracy in DPNN's locomotion and localisation action detection, which regularly outperformed MLP. With 93% and 95% accuracy, respectively, DPNN accomplished the localisation and locomotor tasks on the Continuous In-The-Wild Dataset. Both 86% and 91% accuracy were attained by MLP in the same classes. The Huawei Locomotion Dataset was no match for DPNN, which achieved 95% localisation accuracy and 97% locomotion. Both MLP and the control group achieved 88% and 91% on the same tasks. In comparison to MLP's 90% localisation accuracy and 89% locomotor activity accuracy on the Extra Sensory Dataset, DPNN's rates were 92% and 89%, respectively. Our research shows that DPNN is the most accurate, but it is also the most expensive. Alternatively, MLP can be processed more quickly but with less precision. Using existing machine learning methods to examine sensor data has several advantages, as this study shows. It also details the costs and benefits of computing power vs accuracy when it comes to human activity detection. Our in-depth research reveals how the sensors in smartphones could improve activity detection systems, which could lead to new developments in mobile sensing technologies[16].

Skenderi 2024 et.al Since of the topic's complexity and multi-faceted nature, traditional methods of predicting sales of new fashion items are insufficient. Using multi-modal information about a new fashion item and external knowledge from Google Trends time series, this research analyses the effectiveness of correctly forecasting sales of the item without historical data. We provide a neural network-based method for sales forecasting that is more accurate; in this method, an encoder learns a representation of the external time series and a decoder uses the Google Trends encoding in addition to easily accessible visual and metadata insights. We avoid the cumulative effect of major early errors by designing our methods to not work in an autoregressive fashion. Next, we offer VISUELLE, a freely accessible dataset that tackles the issue of predicting sales of new fashion products. This dataset contains multimodal data for 5,577 actual, brand-new products that were sold by the Italian fast-fashion company Nunalie from 2016 to 2019. Included in the dataset are product images, product descriptions, sales figures, and data from Google Trends. By comparing our strategy to the best alternatives and several baselines, we utilise VISUELLE. With respect to absolute and percentage errors, our neural network-based approach is superior. Since it increases prediction accuracy by 1.5 percentage points in terms of Weighted Absolute Percentage Error (WAPE), using relevant external data is essential[17].

Dipu 2023 et.al Deep neural networks (DNNs) are on top of their game in a lot of different areas. The issue is that DNNs require a lot of processing power, and people are always seeking for ways to save time and effort without sacrificing quality. By studying the human somatosensory system, we were able to construct a neural network called SpinalNet that requires less effort to complete tasks. Traditional NNs involve an HL that takes in data from an earlier layer, applies an activation function to it, and then sends the result on to the subsequent layer. Each layer of the proposed SpinalNet consists of three parts: the input split, the intermediate split, and the output split. Each layer's input split receives a fraction of the inputs. Input data from the previous layer's split and output data from the layer above it are fed into each layer's intermediate split. In contrast to regular DNNs, there are a lot less incoming weights. The SpinalNet can function as the completely connected or classification layer of the DNN, and it is compatible with both traditional and transfer learning. The majority of the DNNs saw a decrease in computing costs and a substantial drop in error rates. We achieved SOTA performance on the QMNIST, Kuzushiji-MNIST, and EMNIST (Letters, Digits, etc.) datasets by applying classical learning on the VGG-5 network with SpinalNet classification layers. For the STL-10, Fruits 360, Bird225, and Caltech-101 datasets, the optimal use of ImageNet pretrained initial weights and SpinalNet classification layers was observed in traditional learning [18].

Santana 2022 [19] Everything that humans do with their eyes and brains relies on attention. Due to our limited cognitive capacity, our attention processes selectively filter, reorder, and zero down on the data that will have the most impact on our actions. Researchers in the fields of philosophy, psychology, neuroscience,

Received: 2026-01-20
Accepted: 2026-03-11
Published Online: 2026-03-30
DOI: 10.70454/JRIST.020102

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926, P- 3117-535X

and computer science have been exploring the concept and function of attention for many years. This feature of deep neural networks has been the subject of study for the previous six years. For many use cases, neural attention models currently represent the pinnacle of Deep Learning. A thorough summary and analysis of current advances in brain attention models are presented in this article. We looked for and analysed field structures where attention has shown a substantial influence after combing through hundreds of them. There is now a publicly available automatic system that can help with writing reviews in this field. After reviewing 650 papers, we identify the main recurrent, generative, and convolutional models that use attention. Additionally, categories of frequently used functions and applications are uncovered. Also covered are the consequences of attention on the comprehensibility of neural networks and its implications across different domains of application. We wrap off by looking ahead to possible patterns and avenues for further research. With any luck, this review will serve as a springboard for future research by offering a synopsis of the most popular attentional models in use today.

Kong 2022 Innovations in data and AI have made precision agriculture a dependable tool for keeping an eye on crops and warding off pests and illnesses. But in precision agriculture, identifying pests and diseases is really a matter of classifying images at a finer level. Ultimately, we want to find some practical discriminative features that can tell comparable visual samples apart. The issue is still hard to solve for traditional models that use overly parameterised and inefficient models. Developed for new agronomic methodologies, the feature-enhanced attention neural network (Fe-Net) can identify pests and illnesses in crops with high accuracy. An enhanced CSP-stage backbone network, which confers many dimension-and size-dependent channel-shuffled qualities on this model, serves as its foundation. Implementing a spatial feature-enhanced attention module is the next step in leveraging the spatial relationship between different semantic domains. To find more representative features, the suggested Fe-Net uses a higher-order pooling module that takes the square root of the elements' covariance matrices. There is no more work required to train complete architecture in an end-to-end method. Top-1 Accuracy was 85.29 percent and average recognition time was 71 milliseconds when the suggested Fe-Net was tested on the CropDP-181 Dataset, surpassing competing approaches. We find that our method achieves a good compromise between the model's parameters and performance, which makes it a good fit for precision agriculture applications[20].

Table 1: Literature Summary

Author/Year	Methodology	Findings	Research Gaps	Limitations
Hou et al. (2024) [21].	Proposed Conv2Former, a convolution-modulation network tested on ImageNet, COCO, ADE20K.	Outperformed ConvNets and Transformers; larger kernels improved global feature capture.	Needs validation on domain-specific datasets like fashion.	High computational cost for high-resolution images.
Peng et al. (2023) [22].	Introduced Conformer, hybrid CNN-Transformer, with ConformerDet for detection.	Preserved local details and global dependencies; strong results on recognition/detection.	Limited application to multimodal or fashion tasks.	Complex dual structure reduces deployment efficiency.

Received: 2026-01-20
Accepted: 2026-03-11
Published Online: 2026-03-30
DOI: 10.70454/JRIST.020102

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926, P- 3117-535X

Dobs et al. (2022) [23]	Used neural nets to test functional specialization in face vs object recognition.	Networks segregated into specialized modules, similar to brain function.	Application to fine-grained fashion categories unexplored.	Focused on neuroscience, limited direct fashion use.
Solanki et al. (2022) [24].	Proposed Π-Nets, polynomial neural networks with tensor factorization.	Achieved expressive learning and strong results in image generation and verification.	Not tested on fashion datasets like DeepFashion.	Scalability and interpretability issues in large tasks.
Baldrati et al. (2022) [25]	Built CLIP-based multimodal CBIR system for image + text retrieval.	State-of-the-art on FashionIQ and CIRR; enabled refined search via text.	Integration with newer architectures unexplored.	Struggles with ambiguous queries; scalability untested.

Research Methodology

In this research project, a systematic design is made, up to the evaluation of a neural network-based architecture for style picture classification or visual search. The process involved five main steps, collecting the data, preprocessing and augmenting it, building the model, training and testing it as well making visual search system. The entire pipeline is carefully crafted to tackle the challenges that arise during image retrieval on fashion datasets, including but not limited to class imbalance, overlaps in annotations and need for accurate search results. The subsections below provide a full account of each.

A. Data Collection: The DeepFashion multi-modal Dataset was used in this study. It is available to the public on Kaggle and is often used as a standard for computer vision research in the fashion industry. There were originally 12,278 entries, each with seven critical attributes: image details, descriptive descriptions, gender, product type, and product ID. This collection has a total of 7,644 individual item IDs, and each product has between 1 and 42 photographs (an average of 1.6 images per product).
https://www.kaggle.com/datasets/silverstone1903/deep-fashion-multimodal. A preliminary integrity evaluation was performed to ascertain dataset fidelity. This examination showed that there were no missing photos or duplicate entries, which proved that it was good for deep learning applications. We looked at a random sample of 500 photographs more closely and found that the average resolution was 750×1101 pixels in RGB format. This indicated that the pictures in the collection were always good.
There were several complications during the data investigation, even though it was strong. First, there was a considerable variance in the number for classes, with a ratio of 582.83 among product groupings. Second, there were labels that were the same for more than one product, which meant that 19 products were put into more than one category. When you split the train and testing sets, this could cause data loss. Lastly, several product groups, including Tees_Tanks, were over-represented compared to others.

Received: 2026-01-20
Accepted: 2026-03-11
Published Online: 2026-03-30
DOI: 10.70454/JRIST.020102

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926, P- 3117-535X

B. Data Filtering: To ensure the dataset was prepared for testing and training, a rigorous approach of removing data was adopted. The first DeepFashion Multimodal Dataset had categories that were relevant and balanced for fashion grouping and retrieval. We removed classes like Shirts_Polos, Denim, Jackets_Vests, Leggings, Suiting, Graphic T-Shirts, Sweatshirts_Hoodies, Cardigans and Skirts_Rompers_Jumpsuits to reduce noise and irrelevant classes. This adjustment reduced the dataset to size and realism, allowing the model to hone in on more dynamically interesting fashion items. Dataset was filtered and re-indexed to have the labels and indices in the same sequence. This prevented inconsistencies during training. However, one major issue discovered regarding Tees_Tanks was the overrepresentation of entries compared to other categories. If this imbalance is not corrected on the data, the model will be more likely to predict this majority class, which can make it less correct overall. To correct this, 1550 random samples through Tees_Tanks were systematically removed which led to a more balanced class distribution across the categories. This filtering process reduced repeated entries and also decreased potential bias, resulting in the dataset being more representative of clothing items. This led to a dataset that is cleaner, more balanced and aligned with the objective of study, it helped in subsequent steps like preprocessing, augmentation and model training.

C. Data Preprocessing: A full data pretreatment procedures used to prepare the dataset for deep learning operations after filtering. For product-level data, the stratified splitting method was used to create test, validation, and training sets. This ensured that photographs of the same products were in the same split and made certain that data was evenly distributed across categories, stopping them from leaking. Finally, we applied label encoding to convert categorical product labels into numbers for use in the neural network. Class weights were also calculated and included in the training process to further mitigate class imbalance. This ensured that under-represented classes received adequate focus during optimisation.

Data augmentation techniques were applied at the training stage to diversify the dataset by randomly flipping, changing brightness and rotation it. This made model stronger and more like real-life changes. Just for validation and testing, the resizing and normalisation were only applied so that the evaluation would remain consistent. In addition, offline augmentation was employed to increase the sample size of minority classes ensuring each class was adequately represented. For all photos we create a custom Fashion Dataset class which resizes, normalises (using ImageNet mean and standard deviation), and converts them to tensors so that they are the same dimensions. Eventually, Data Loaders were introduced to enable batching, shuffling and parallelised loading that made training a lot more efficient. This preprocessing workflow resulted in a clean, balanced and diverse dataset suitable for robust model training.

D. Feature Extraction: Image classification and visual search depend on feature extraction, which turns raw pixels into high-dimensional embeddings that show important patterns. The method in this study was conducted with a Vision Transformer (ViT) backbone, augmented with a Spatial Attention Module. The ViT, on the other hand, breaks each image into patches and encodes them as tokens. This lets it represent global context. Traditional CNNs rely on local receptive fields. The Spatial Attention Module made these embeddings better by emphasising on areas of the clothing images that were different from the rest, like collars, textures, and sleeve shapes. This allowed the network to not only see the object's overall shape, but also pay attention to little features that are critical for classifying garments. The combination of a global context (CLS token) and a localised emphasis (spatial maps) made for exceptionally powerful embeddings. This method, each processed image was saved as a dense feature vector that maintained both the

Received: 2026-01-20
Accepted: 2026-03-11
Published Online: 2026-03-30
DOI: 10.70454/JRIST.020102

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926, P- 3117-535X

category-level meaning and the detailed information. After that, these vectors were used in two ways: (1) a classifier was used to estimate the fashion category, and (2) they were put into the visual search engine so that they could be located by similarity. This consistent and discriminative feature extraction method was used to build both the classification pipeline and the visual search system. This made sure that the embeddings were strong and functional.

E. Exploratory Data Analysis: Researchers developed a very good intuition about the structure of the dataset, its distribution and complexities through extensive Exploratory Data Analysis (EDA) while training the model in person. The dataset contained 12,278 rows and a product ID uniquely identified each row. 1 to 42 Number of photos per product The representation was poor, as each product had an average of 1.6 photos. Visual inspection of a sample of 500 images showed that the RGB quality was consistent and the average resolution was 7501101 pixels, indicating that these images were suitable for deep learning. “But there were several significant issues.” The class imbalance ratio of 582.83 proved that certain groups, such as Tees_Tanks were far too represented compared to others, say Dresses or Outerwear. Also, 19 products appeared in multiple categories, which makes labels that could overlap a little too much & lead to data leak through the splitting if not cleaned. These results demonstrated the need for diligent preprocessing and filtering. Visualisation tools, particularly with category count bar graphs and image montage sampling, assisted in identifying dominant types of products sold and differences in colour and shape between categories. Captions were also analysed to provide us with useful multimodal insights as text descriptions here tended to add information on top of the images. The EDA not only displayed the dataset's strengths, such as its quality and variety,: it also highlighted issues that required resolution -- especially in terms of controlling imbalances, avoiding leaks and optimizing categories. These all had an impact on the downstream workflow.

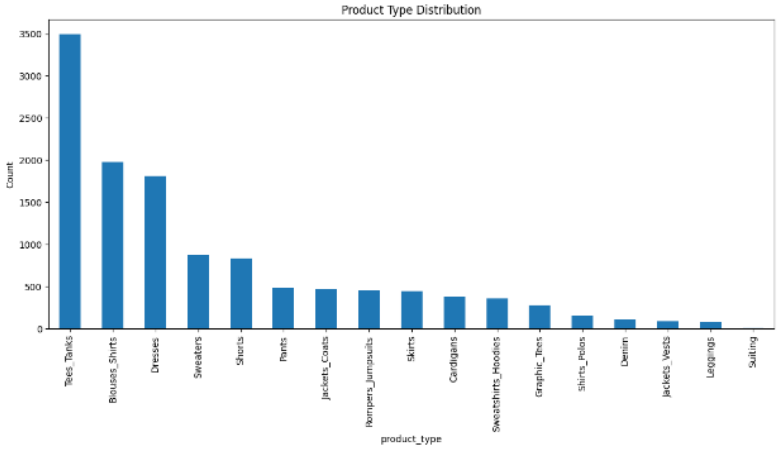

Fig. 2 Product Type Distribution

This figure shows how the different types of products are spread out in the dataset. It shows that there is an imbalance in the classes, with some categories, such Tees/Tanks, being over-represented and others being under-represented. This kind of skewness makes it hard for models to generalise because training may not include minority classes. The examination of the distribution shows that we need to use weighted loss functions and do careful preprocessing to make sure that all classes are represented equally. It is important to be aware of these differences in order to make sure that fashion image categorisation tasks are fair and strong.

Received: 2026-01-20
Accepted: 2026-03-11
Published Online: 2026-03-30
DOI: 10.70454/JRIST.020102

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926, P- 3117-535X

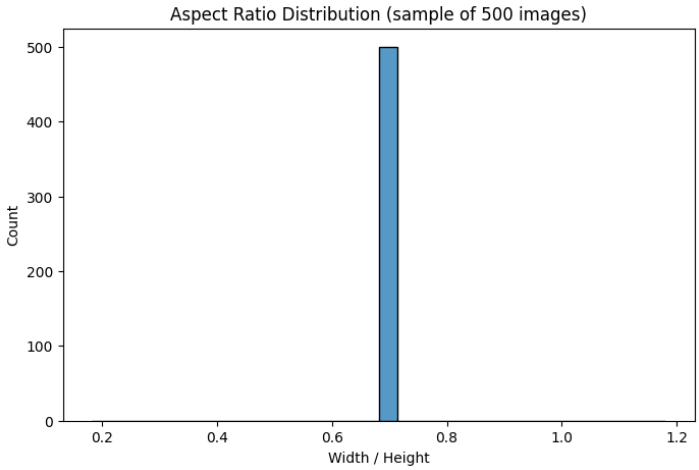

Fig. 3 Aspect Ratio Distribution

This graph shows the distribution of aspect ratios for fashion photos, which is found by dividing the width of an image by its height. The plot indicates that most of the pictures stay upright, which is in line with the rules for taking pictures of clothes. But differences in ratios show that there are several ways to frame and crop. These inconsistencies could make it harder to extract features since neural networks might learn false representations. Knowing how aspect ratios change can help you choose preprocessing methods, like scaling with aspect-ratio preservation, that keep the look of the image the same while reducing distortion.

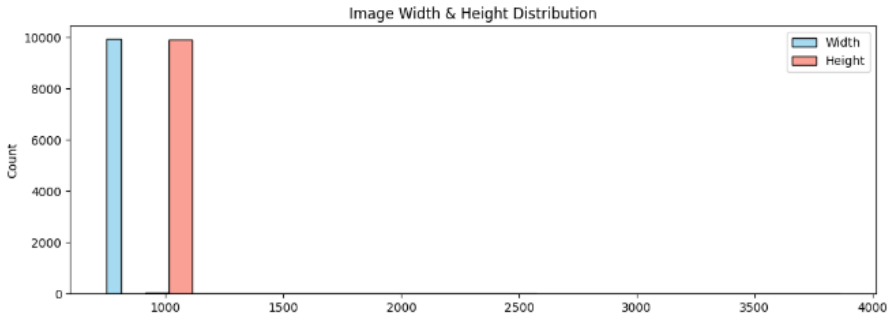

Fig. 4 Image Size Distribution

This graphic demonstrates how the photos in the dataset are spread out. Most of them are around 750×1101 pixels in size. Most of them still have excellent resolution, which is good for extensive visual inspection, but differences in dimensions signal that the source data might not be consistent. To make sure that neural networks can work with all the different input sizes, preprocessing needs to resize them all to the same size. This makes sure that the computer works quickly, that feature maps don't get misaligned, and that the general quality of fashion-related visual signals that are important for model performance stays high.

Received: 2026-01-20
Accepted: 2026-03-11
Published Online: 2026-03-30
DOI: 10.70454/JRIST.020102

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926, P- 3117-535X

Fig. 5 Sample images in the dataset

This image shows a few examples from the dataset that highlight how different fashion categories, product orientations, materials, and styles may be. It shows how rich the dataset is in terms of visual variety, which is important for teaching models to spot small deviations. Sample photographs also show problems such backgrounds that overlap, lighting that isn't consistent, and stances that change, all of which could affect feature extraction. Researchers can check the quality of a dataset and see if preprocessing methods are needed to cut down on noise and make classification more reliable by looking at samples.

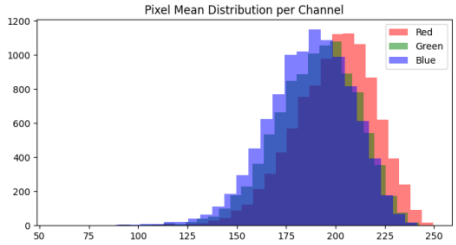

Fig. 6 Pixel mean Distribution

This metric represents the average number of pixels in photographs, broken down per channel. The picture depicts how the brightness and colour intensity alter because of varied lighting and background settings. Normalisation techniques, essential for maintaining stability in deep learning pipelines, are grounded in pixel mean statistics. Normalising picture pixels stops models from fitting too closely to changes that don't matter, like changes in illumination. So, looking at pixel distributions helps create preprocessing tactics that make fashion image recognition more reliable.

F. **Data Splitting:** Everyone used a stratified strategy to split the dataset into subsets for testing, validation, and training. This prevented the model from overfitting and ensured an accurate evaluation. This made sure that the class proportions stayed the same in all subgroups, keeping things fair and stopping bias towards the more common categories. It was important to separate at the product ID level so that different subsets didn't get different photos of the same product. This step got rid of the possibility of data leakage, which may happen if the model "memorised" product features instead of learning general patterns. A well-calculated split ratio was employed to guarantee a robust training set while simultaneously allowing for sufficient data for impartial testing and validation. It was common practice to allocate 70% of data to training, 15% to validation, and 15% to testing. The class balance was maintained, nonetheless, with little adjustments.

Received: 2026-01-20
Accepted: 2026-03-11
Published Online: 2026-03-30
DOI: 10.70454/JRIST.020102

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926, P- 3117-535X

G. Visual Search System: In addition to classification, a visual search system was employed in this study to retrieve visually similar fashion products. To get representational embedding of global and fine-grained garment features, system used Vision Transformer with Spatial Attention. These embeddings were stored in an Annoy-based approximate nearest neighbour (ANN) index. This index was designed to accelerate and scale similarity searches. This means that all photos went through some sort of preprocessing before being indexed. Data processing steps which include resizing, normalising and conversion to tensors. This ensured that the same was true for query and gallery images, thereby reducing the possibility of variance when they were retrieved. Annoy to index each gallery image using cosine similarity after extracting embeddings. It was a simple method for how close image attributes were to one another.

Given a query image, its embedding was computed with the same procedure and used to search in the gallery index for top-K most similar things. Content-based retrieval like this eliminated the need for human tagging or text-based metadata. The results were displayed alongside the query image with corresponding rated matches — an illustration of how well those retrievals performed. We tested the system performance using Precision and Mean Reciprocal Rank (MRR). This ensured that both accuracy and ranking were evaluated. It was quick and easy to search for similar-looking products, so the technique held a lot of potential for e-commerce and fashion recommendation engines.

H. Model Implementation: FashionViT-SA is the name of the suggested categorisation framework. It combines a Vision Transformer backbone (ViT-Base, Patch16, pretrained on ImageNet) with a Spatial Attention Module. This hybrid architecture lets the model use the transformer to efficiently capture global feature dependencies while also employing spatial attention to focus on localised garment properties. To avoid overfitting, dropout layers were used at different points in the process, and the final classifier head put the learnt embeddings into seven pre-defined fashion categoriesA weighted cross-entropy loss was used to train FashionViT-SA in order to address the issue of class imbalance. As a result, minority classes were guaranteed a higher optimisation priority. Training was conducted with the usage of the Adam-W optimiser and a cosine annealing scheduler in tandem to ensure consistent convergence. The learning process became even more steady with the addition of gradient clipping, which prevented gradients from exploding. Employing a batch size of 32 for mini-batch training, we evaluated the model's performance on the validation set after every epoch. Among several evaluation metrics, the F1-score was deemed most important due to its ability to demonstrate a model's performance on datasets that are imbalanced. For the final test, we retained the checkpoint with the highest validation F1-score.

TABLE 2: Hyperparameter Values

Hyperparameter	Value / Description
Backbone Model	Vision Transformer (ViT-Base, Patch16, Pretrained)
Attention Mechanism	Spatial Attention (7×7 convolution, sigmoid)
Number of Classes	7
Batch Size	32
Learning Rate	3e-5
Optimizer	Adam-W
Loss Function	Weighted Cross-Entropy
Scheduler	Cosine Annealing LR
Dropout Rate	0.3
Gradient Clipping	1.0

Received: 2026-01-20
Accepted: 2026-03-11
Published Online: 2026-03-30
DOI: 10.70454/JRIST.020102

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926, P- 3117-535X

Evaluation Metrics	Accuracy, Precision, Recall, F1-score
Early Stopping Criterion	Best Validation F1-score

Table 2 illustrates the hyperparameter settings that were utilised to train FashionViT-SA. The backbone model, ViT-Base, Patch16 pretrained on ImageNet, was the main feature extractor. It was improved by a spatial attention block (7×7 convolution followed by a sigmoid function) to make clothing images more detailed. The classification challenge included seven categories, hence the output layer had to have seven classes. Training was done in mini-batches of 32 samples, and the learning rate was set at 3e-5 to make sure that convergence was balanced. We chose the Adam-W optimiser because it works well with transformer-based designs and because cosine annealing learning rate scheduling makes the decay process smooth. Weighted cross-entropy loss made sure that classes that weren't represented enough got more weight. Dropout (0.3) kept overfitting to a minimum, and gradient clipping at 1.0 kept gradient updates from getting too unstable. The accuracy, precision, recall, and F1-score were the four metrics utilised for performance evaluation. To ensure the model was robust and did not overtrain, early halting was based on the validation F1-score.

I. **Performance Metrics:** In order to verify that FashionViT-SA model is strong and fair across all classes, it evaluated using multiple criteria. Accuracy assessed how correct all the predictions were; precision examined how many elements in each category were properly recognized. Recall measured how well the model was able to identify all of the relevant cases. This is crucial especially for minority classes. The dataset was unbalanced, therefore the F1-score, which balances precision and recall, was under pressure. The combined indicators revealed how well the classification performed and its general applicability under various situational contexts.

Results and Discussion

In this section, we demonstrate the test results of our proposed FashionViT-SA model and discuss what its performance for vision-based sorting and retrieval of fashion images. The findings show that data pretreatment, augmentation, and attention-based design are essential for improving the classification performance of imbalance classes. Essential metrics we review are loss, accuracy, precision, recall and F1-score. E.g. product distributions, picture properties, and embedding spaces visualisations provide more proof of dataset qualification characteristics as well as model working. The commentary discusses these outcomes; compares them to other methodologies (such as those reliant on a more general melting point term); and emphasizes how crucial the system is for fashion applications.

- **Accuracy:** The accuracy shows how many samples are classified correctly. It shows how well the model predictions of accuracy. While it is widely used, this metric can be misleading as it can inflate the perceived accuracy in situations where a dataset may already be imbalanced or include under-represented minority groups.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

- **Precision:** One measure of accuracy is the ratio between expected positive occurrences and actual positive occurrences. Instead, here we will evaluate a model's ability to 'avoid' false positives. A system that has high precision is more detailed and accurate when describing objects. Grabbing the incorrect item type may allow to make tracking down or recommending

Received: 2026-01-20
Accepted: 2026-03-11
Published Online: 2026-03-30
DOI: 10.70454/JRIST.020102

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926, P- 3117-535X

just about anything less reliable — something that's especially important for fashion classification.

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

- **Recall:** Recall, also known as sensitivity, is the percentage of true positives out of all positive predictions. This indicates the model has managed to sample all and every on-demand samples. Minimising false negatives for products poorly represented in the training data is important, and a high recall ensures that the system retrieves or detects most fashion items correctly.

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

- **F1 Score:** The F1-score is a well-rounded statistic that takes both false positives and false negatives into consideration; it is the harmonic mean of recall and precision. It shines most in unbalanced datasets, when precision could be misleading about actual performance. As F1-score guarantees both completeness and correctness, it becomes a more dependable indication of the overall effectiveness of the model in fashion classification.

$$F1 - score = \frac{2}{\frac{1}{precision}+\frac{1}{recall}} \quad (5)$$

- **Loss:** During training, the loss function guides weight updates by quantifying the discrepancy between the model's predictions and the actual labels. A weighted cross-entropy loss was employed in Fashion ViT-SA to tackle class imbalance. This means that misclassifications of minority categories are penalised more heavily.

Table 3 Performance Evaluation of **Fashionvit-Sa** Model

Model	Accuracy	Precision	Recall	F1-score	Loss
FashionViT-SA	0.8308	0.8121	0.8433	0.8249	1.2993

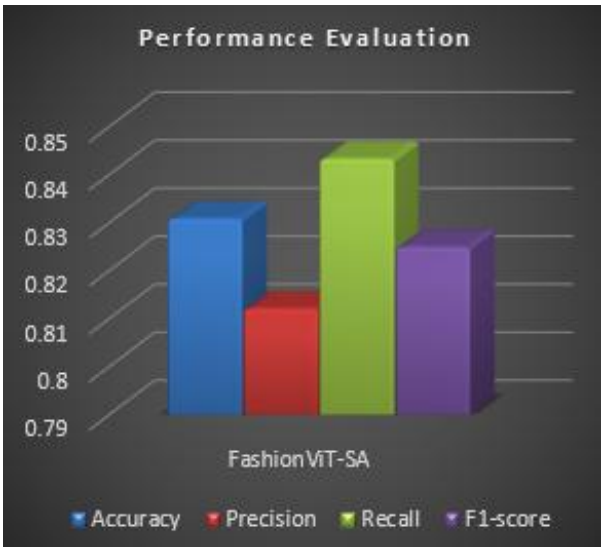

Fig. 7 Evaluation of Proposed FashionViT-SA Model

Received: 2026-01-20
Accepted: 2026-03-11
Published Online: 2026-03-30
DOI: 10.70454/JRIST.020102

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926, P- 3117-535X

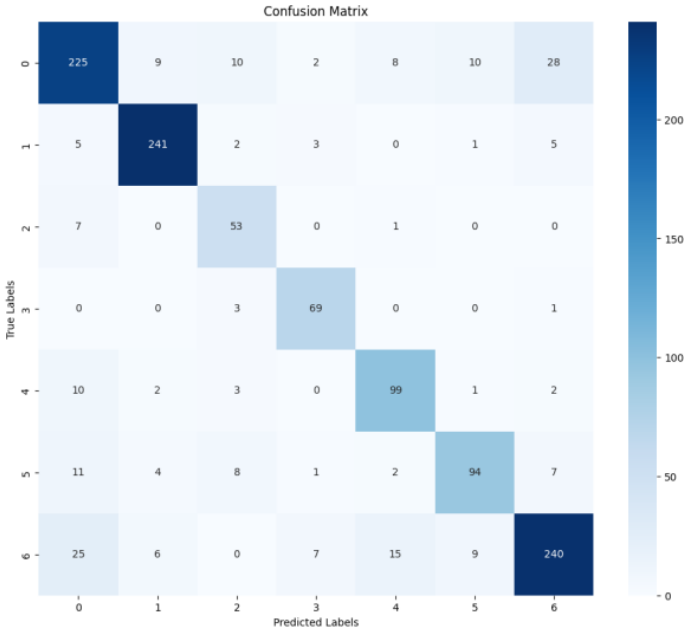

Fig. 8 Confusion Matrix

The confusion matrix for the FashionViT-SA from our first experiment is depicted in Figure 8 oriented to demonstrate an overview of how well it performed against classifying items in all seven fashion classes. It indicates that recognition performance is strong in the main categories, as shown by numbers along the diagonal which reflect correctly classified samples. Off-diagonal entries tell us where things were misclassified, particularly with regard to clothes that look identical. This study highlights what the model does well and where it performs poorly so that you have some actionable insights to help improve balance in classes and overall predictive accuracy.

Comparative Analysis

Comparative analysis is essential to highlight the effectiveness of the proposed FashionViT-SA model against existing approaches in fashion image classification. Traditional convolutional models have demonstrated the ability to capture local patterns but often struggle with global feature representation, leading to moderate performance. As shown in Table 4, the existing CNN model [26] achieved 70% accuracy, reflecting these limitations. In contrast, the proposed FashionViT-SA model significantly improves accuracy to 83%, showcasing the advantages of Vision Transformer architecture combined with spatial attention.

Table 4 Comparative Analysis Between Proposed And Existing Model

Model	**Accuracy**
Existing CNN Model [26]	70%
Proposed FashionViT-SA Model	**83%**

Received: 2026-01-20
Accepted: 2026-03-11
Published Online: 2026-03-30
DOI: 10.70454/JRIST.020102

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926, P- 3117-535X

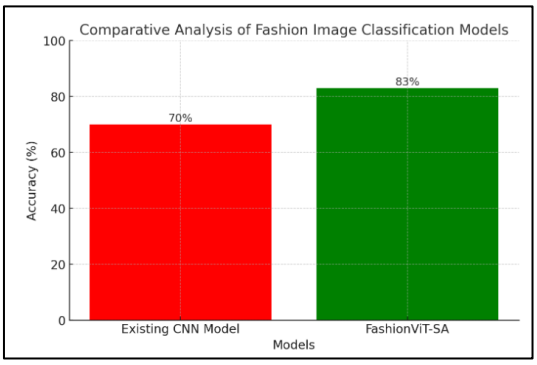

Fig. 9 Comparative Analysis Graph

Table 4 compares the baseline CNN model with the proposed FashionViT-SA model. The CNN baseline attained only 70% accuracy, indicating its restricted ability to generalize across diverse fashion categories. Meanwhile, FashionViT-SA achieved 83% accuracy, marking a 13% improvement. This performance gain shows that the model can efficiently include both global and local features, which lowers the number of misclassifications and makes it more stable for use in the real world.

Conclusion

FashionViT-SA — a hybrid Vision Transformer framework with a Spatial Attention Module does well at classifying fashion images and for content-based visual search. By combining global modelling using the transformer backbone with localised attention mechanisms, the model is able to capture both coarse silhouette as well as fine-grained clothing details. This addresses issues found in fashion datasets that include class imbalance, overlapping categories, differences in products representation. Much preprocessing in files like filtering, stratified splitting, label encoding and full data augmentation helped making the dataset balanced and of good quality. This made the model a lot more robust and generalized. With 83% accuracy and F1-score of 0.8249, the model clearly outperformed traditional CNN-based methods in both classifying fashion items correctly and retrieving relevant images. This was demonstrated by evaluating it across various metrics, such as accuracy, precision, recall, F1 score and weighted cross-entropy loss. The confusion matrix showed how well the model distinguishes similar looking categories and it identified minor misclassifications that can be improved in future iterations of the model. Applied the model within a visual search system where we utilized an Annoy-based approximation method to allow it to be invoked in business use cases like e-commerce and recommendation systems while providing fast and reproducible retrieval of visually similar products without any reliance on textual metadata. While the study's results are promising, it is essential to note the evolution showcasing a new paradigm in fashion image analysis as it fits into information retrieval within content-based paradigms and builds on advances that fuse visual recognition generated by deep learning methods with hybrid highlighting similarity scores capable of handling beyond-2DT decisions, all using transformers intertwined with attention mechanisms for image data pursued towards revolutionizing cross-scene impacts.

Acknowledgment

Write you acknowledgment here using the same text format.

Authors’Declaration

- Conflicts of Interest: None.

Received: 2026-01-20
Accepted: 2026-03-11
Published Online: 2026-03-30
DOI: 10.70454/JRIST.020102

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926, P- 3117-535X

References

[1] R. Jabbar, M. Shinoy, M. Kharbeche, K. Al-Khalifa, M. Krichen, and K. Barkaoui, "Driver Drowsiness Detection Model Using Convolutional Neural Networks Techniques for Android Application," *2020 IEEE Int. Conf. Informatics, IoT, Enabling Technol. ICIoT 2020*, pp. 237–242, 2020, doi: 10.1109/ICIoT48696.2020.9089484.

[2] X. Wang *et al.*, "Pedestrian attribute recognition: A survey," *Pattern Recognit.*, vol. 121, pp. 1–32, 2022, doi: 10.1016/j.patcog.2021.108220.

[3] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1195–1215, 2022, doi: 10.1109/TAFFC.2020.2981446.

[4] S. Abbas *et al.*, "Convolutional neural network based intelligent handwritten document recognition," *Comput. Mater. Contin.*, vol. 70, no. 3, pp. 4563–4581, 2022, doi: 10.32604/cmc.2022.021102.

[5] A. M. Obeso, J. Benois-Pineau, M. S. García Vázquez, and A. Á. R. Acosta, "Visual vs internal attention mechanisms in deep neural networks for image classification and object detection," *Pattern Recognit.*, vol. 123, 2022, doi: 10.1016/j.patcog.2021.108411.

[6] M. Zhuge *et al.*, "Kaleido-Bert: Vision-Language Pre-training on Fashion Domain," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 12642–12652, 2021, doi: 10.1109/CVPR46437.2021.01246.

[7] B. Yan, H. Peng, K. Wu, D. Wang, J. Fu, and H. Lu, "LightTrack: Finding lightweight neural networks for object tracking via one-shot architecture search," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 15175–15184, 2021, doi: 10.1109/CVPR46437.2021.01493.

[8] M. Lin *et al.*, "Zen-NAS: A Zero-Shot NAS for High-Performance Image Recognition," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 337–346, 2021, doi: 10.1109/ICCV48922.2021.00040.

[9] Y. He, D. Yang, H. Roth, C. Zhao, and D. Xu, "DiNTS: Differentiable Neural Network Topology Search for 3D Medical Image Segmentation," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 5837–5846, 2021, doi: 10.1109/CVPR46437.2021.00578.

[10] H. Alqahtani, M. Kavakli-Thorne, and G. Kumar, "Applications of Generative Adversarial Networks (GANs): An Updated Review," *Arch. Comput. Methods Eng.*, vol. 28, no. 2, pp. 525–552, 2021, doi: 10.1007/s11831-019-09388-y.

[11] H. Wu *et al.*, "Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 11302–11312, 2021, doi: 10.1109/CVPR46437.2021.01115.

[12] S. J. Wang, Y. He, J. Li, and X. Fu, "MESNet: A Convolutional Neural Network for Spotting Multi-Scale Micro-Expression Intervals in Long Videos," *IEEE Trans. Image Process.*, vol. 30, pp. 3956–3969, 2021, doi: 10.1109/TIP.2021.3064258.

[13] H. Kumar and Y. Hasija, *Machine Learning in Medical Image Processing*, vol. 195. 2021. doi: 10.1007/978-981-15-7078-0_35.

Received: 2026-01-20
Accepted: 2026-03-11
Published Online: 2026-03-30
DOI: 10.70454/JRIST.020102

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926, P- 3117-535X

[14] J. M. Wolfe, *Guided Search 6.0: An updated model of visual search*, vol. 28, no. 4. Psychonomic Bulletin & Review, 2021. doi: 10.3758/s13423-020-01859-9.

[15] G. Castellano and G. Vessio, "Deep learning approaches to pattern extraction and recognition in paintings and drawings: an overview," *Neural Comput. Appl.*, vol. 33, no. 19, pp. 12263–12282, 2021, doi: 10.1007/s00521-021-05893-z.

[16] D. Khan *et al.*, "Advanced IoT-Based Human Activity Recognition and Localization Using Deep Polynomial Neural Network," *IEEE Access*, vol. 12, no. July, pp. 94337–94353, 2024, doi: 10.1109/ACCESS.2024.3420752.

[17] Geri, C. Joppi, M. Denitto, and M. Cristani, "Well googled is half done: Multimodal forecasting of new fashion product sales with image-based google trends," *J. Forecast.*, vol. 43, no. 6, pp. 1982–1997, 2024, doi: 10.1002/for.3104.

[18] H. M. Dipu Kabir *et al.*, "SpinalNet: Deep Neural Network With Gradual Input," *IEEE Trans. Artif. Intell.*, vol. 4, no. 5, pp. 1165–1177, 2023, doi: 10.1109/TAI.2022.3185179.

[19] A. de Santana Correia and E. L. Colombini, *Attention, please! A survey of neural attention models in deep learning*, vol. 55, no. 8. Springer Netherlands, 2022. doi: 10.1007/s10462-022-10148-x.

[20] J. Kong, H. Wang, C. Yang, X. Jin, M. Zuo, and X. Zhang, "A Spatial Feature-Enhanced Attention Neural Network with High-Order Pooling Representation for Application in Pest and Disease Recognition," *Agric.*, vol. 12, no. 4, 2022, doi: 10.3390/agriculture12040500.

[21] Q. Hou, C. Z. Lu, M. M. Cheng, and J. Feng, "Conv2Former: A Simple Transformer-Style ConvNet for Visual Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 8274–8283, 2024, doi: 10.1109/TPAMI.2024.3401450.

[22] Z. Peng *et al.*, "Conformer: Local Features Coupling Global Representations for Recognition and Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9454–9468, 2023, doi: 10.1109/TPAMI.2023.3243048.

[23] K. Dobs, J. Martinez, A. J. E. Kell, and N. Kanwisher, "Brain-like functional specialization emerges spontaneously in deep neural networks," *Sci. Adv.*, vol. 8, no. 11, pp. 1–11, 2022, doi: 10.1126/sciadv.abl8913.

[24] G. G. Chrysos, S. Moschoglou, G. Bouritsas, J. Deng, Y. Panagakis, and S. Zafeiriou, "Deep Polynomial Neural Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 8, pp. 4021–4034, 2022, doi: 10.1109/TPAMI.2021.3058891.

[25] A. Baldrati, M. Bertini, T. Uricchio, and A. Del Bimbo, "Effective conditioned and composed image retrieval combining CLIP-based features," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2022-June, pp. 21434–21442, 2022, doi: 10.1109/CVPR52688.2022.02080.

[26] B. Lao and K. Jagadeesh, "Convolutional Neural Networks for Fashion Classification and Object Detection," *CCCV Comput. Vis.*, pp. 120–129, 2015.