Journal of High School Science

A comparative study of Convolutional Neural Network and Vision Transformer models as classifiers of east Asian traditional clothing

Wong G

Submitted: June 24, 2025, Revised: version 1, August 10, 2025
Accepted: August 22, 2025

Abstract

Convolutional Neural Network (CNN) and Vision Transformer (ViT) are two models that have been at the center of discussion in the field of machine learning due to their accuracy in classifying images. Current research has shown that different image classification tasks lead to different performances for both models. One area of research that is still under-explored when comparing CNN and ViT is the task of classifying images of traditional clothing from various countries. In this study, a CNN and a ViT were constructed and trained on a dataset of Chinese, South Korean, and Japanese traditional clothing images. The two models were then quantitatively compared across different aspects of accuracy utilizing four evaluation metrics: accuracy, precision, recall, and F1 score. The results showed that the ViT–with an average score of 88%–outperformed the CNN–which had an average score of 73.15%–in every evaluation metric, underscoring the potential of the ViT for classifying traditional clothing. Additionally, the models were deliberately fed out-of-dataset data to find if there were patterns in the resulting obvious mis-classification of such images. Analysis of such patterns revealed features and attributes that the models used to classify data, which unlocked model explainability; the current Achilles' heel for stand-alone model adoption. Overall, these findings can prove useful to individuals in the field of cultural preservation by showcasing that the ViT can be used as a tool to accurately classify traditional clothing attire from China, Japan and Korea.

Keywords

Asian clothing classification, Traditional clothing, Machine Learning, Deep Learning, Image classification, Vision Transformer, Convolutional Neural Network, Model explainability, Black box, Cultural heritage preservation

Gavin Wong, Passaic County Technical Institute STEM Academy, 45 Reinhardt Rd, Wayne, NJ 07470, USA.
gk2008wong@gmail.com

https://doi.org/10.64336/001c.143597

1. Introduction

Computers can undertake increasingly complex tasks due to technological developments in the field of machine learning (ML). ML is a branch of computer science that utilizes statistical and mathematical methods to allow computers to acquire knowledge without the need for direct human programming (1). One of the most promising areas of ML is computer vision (CV), which involves the extraction of information from digital images through computational models (2). Essentially, CV enables computers to understand the 'meaning' behind images and is thus used for image classification tasks, which first involves training to correctly categorize images which have been previously categorized by humans (1). With the constant development of CV, many distinct ML models have been developed for image classification tasks, with two of the most prominent being the Convolutional Neural Network (CNN) and the Vision Transformer (ViT). Due to the ability of the CNN and ViT to capture complex patterns in images, these ML algorithms have been previously leveraged to classify images of clothing. However, one under-explored sector of such clothing classification tasks is classifying the traditional attire of different cultures. Traditional clothes contain rich colors, designs and styles that represent cultural heritage (3), and the aforementioned models can potentially be used to recognize them. Through the use of these two image classification tools to categorize traditional wear by culture, individuals in fields such as cultural preservation may benefit from the computational powers of ML algorithms, which could allow for faster and potentially more accurate recognition of cultural artifacts and ethnocultural differences and evolution. Thus, the CNN and ViT models have the potential to assist in classifying newly discovered traditional clothing artifacts, thereby helping understand, appreciate and promote cultural heritage.

2. Background

2.1 Deep Learning

Before deep learning was discovered, CV relied on manual feature detection, where humans explicitly programmed a ML model to recognize specific patterns. This made CV tasks less effective in real-world scenarios where conditions such as lighting and environment could vary. Deep learning (DL) addresses the issue of manual feature detection by enabling ML algorithms to discover patterns and learn what features to extract by themselves, making real-world CV tasks feasible (4). The two ML models presented in this work are, more specifically, DL models–a type of ML model that utilizes artificial neural networks (ANNs) to allow computers to "think" deeply (1).

ANNs are based on the human brain, where neurons connect together to transport information, as shown in Figure 1. However, in ANN architecture, these neurons direct the flow of information based on mathematical functions. These neurons are then connected together with more mathematical expressions to establish a deep layer of connections. This comparison can be seen visually in Figure 2. In a typical ANN, there are anywhere from just a few– as seen in Figure 2 –to millions of neurons that are stacked in layers that are fully connected to each other (7).

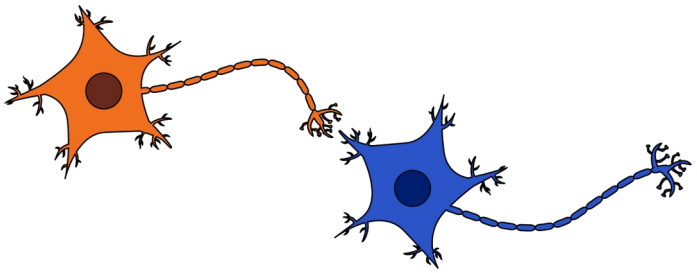

Figure 1. Diagram of two connected neurons (5)

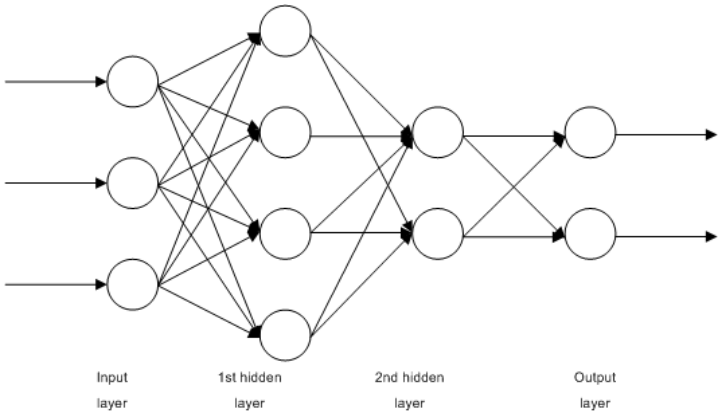

Figure 2. A simple Artificial Neural Network (6)

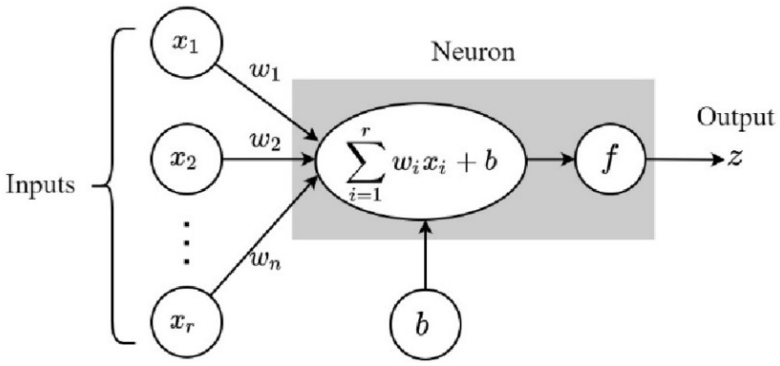

Figure 3. The Structure of a Neuron (8)

As shown in Figure 3, a singular neuron processes information by multiplying numerical inputs by a weight term, *w*. For non-numerical data to be input into the ANN, the data must first be transformed into a numerical format by the ML model. After summing the

products of the weights, a bias term, *b,* is added. By training on data, ANNs learn these weights and biases by minimization of a cost or loss function, where the higher a particular weight, the more influence it has on a model's understanding of the input data. To effectively "learn" from the input data, the ML model performs numerous mathematical calculations behind the scenes to adjust *w* and *b* to the optimal value to predict outcomes from the input data correctly.

2.2 Convolutional Neural Networks

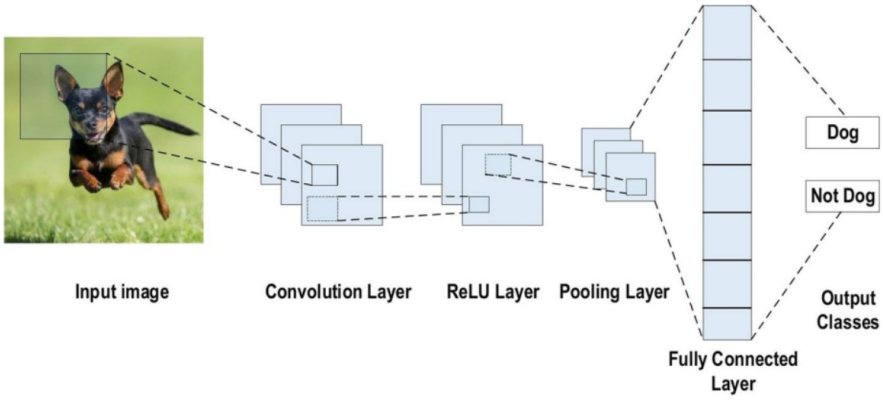

Figure 4. Structure of the convolutional neural network (9)

The first ML model in this work, the CNN, works by sliding a filter around the image, analyzing small regions of the image at a time, as seen in Figure 4. This mechanism of sliding the filter allows the model to pick up on patterns used to classify the image in the final layer, which consists of an ANN (10). Since these CNNs rely on fixed-sized filters to understand the image, CNNs lack global access to image features–meaning features from anywhere on the image–as the filters can only analyze the image in proximal regions at a time, giving it local access to image features (11).

2.3 Vision Transformers

The ViT was built off of the transformer architecture–a specific ML algorithm developed by Vaswani et al. (12), and harnessed the power of the transformer's self-attention mechanism, a feature that allows the ViT to capture increasingly complex relationships in data through complex math that is beyond the scope of this study.

As seen in Figure 5, the ViT processes images by splitting the input image into fixed-size patches. Next, mathematical operations are applied to the patches to make them suitable inputs for the transformer architecture. Finally, after passing through the transformer architecture and a multilayer perceptron (MLP)–a type of ANN–the image is classified (13). The self-attention mechanism within the transformer allows the ViT to adjust dynamically to capture global or local information in the image based on the type of input image (14).

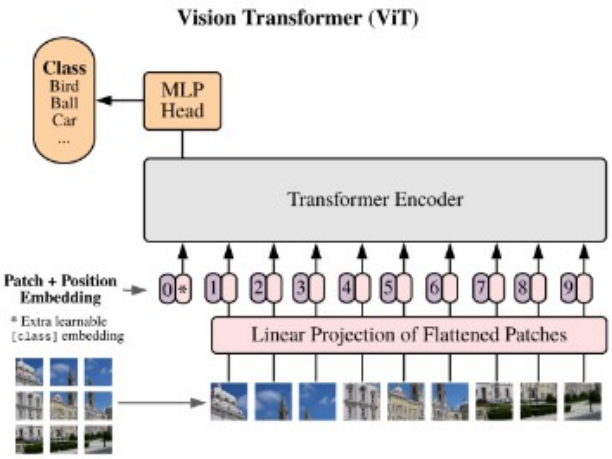

Figure 5. Structure of the Vision Transformer (13)

3. Related work

3.1 Current comparisons

In recent years, researchers have been attempting to find the most accurate, efficient, and reliable tools for image classification. There is ongoing debate between the relative performance of CNNs and ViTs, which have proven to be two of the most promising models. Specifically, some researchers find that the CNN outperforms the ViT when used for image classification. For instance, Gai et al. found that CNNs achieved greater accuracy when classifying a dataset of 212 lung cancer medical scans (15). However, the researchers concluded that the CNNs performed better due to the limited size of the dataset and predicted that the performance of ViTs would match, or even surpass, that of CNNs in an expanded dataset. Similarly, Zhao et al. found that CNNs classified a small dataset of 840 environmental microorganism images more effectively than ViTs (16).

While there are many arguments for the accuracy of CNNs, other researchers have found that ViTs classify images more effectively. Rachman et al. found that the ViT's self-attention mechanism allowed for a better understanding of complex relationships between parts of an image, therefore allowing it to classify two datasets of rice diseases better than the CNN (17). The study concluded that the gap in accuracy can be attributed to the fact that ViTs understand the context of the entire image compared to CNNs, which rely on local features. Alaziz et al. also supported the argument that ViTs were superior to CNNs in their study, which found that ViTs classified the MNIST fashion dataset with higher accuracy than the highest-performing CNN model, demonstrating the ViT's potential for clothing classification (18).

Table 1. Literature accuracy scores comparing CNNs and ViTs

Author	Dataset	Dataset Size	Best Model Accuracy
Gai et al. (2024)	Lung cancer medical scans	212	CNN: 95.3% | ViT: 87.5%
Zhao et al. (2022)	Environmental microorganisms	840	CNN: 44.76% | ViT: 34.60%
Rachmen et al. (2024)	Rice diseases	2,948	CNN: 93% | ViT: 97%
Alaziz et al. (2023)	MNIST fashion	60,000	CNN: 92.25% | ViT: 95.25%

As seen in Table 1, the accuracy of the CNN and ViT vary from study to study, eluding a universal conclusion on performance. The performance of the CNN and ViT is highly dependent on the specific dataset it is tasked with classifying, as certain datasets may contain patterns or characteristics better suited for one model than the other. Additionally, past research shows that CNNs generally classify better than ViTs on smaller datasets, while ViTs classify better than CNNs on larger datasets. Overall, previous work has demonstrated that the specific task given to the ML algorithm – which includes the type of dataset and its size – affects ML model performance and, consequently, which ML model is the better suited for that particular dataset.

3.2 Traditional wear classification

In the narrow scope of traditional wear classification, much more research is available regarding the performance of the CNN than the performance of the ViT, as there are few studies on the application of the latter for this purpose. Research conducted by Herwinsyah et al. found that the CNN performed well when identifying Indonesian traditional dances based on a dataset of 700 images of Indonesian traditional clothing (19). The researchers noted that the CNN's performance was achieved on a restricted dataset and that with a more diverse dataset, the capabilities of CNNs would be substantially greater. Likewise, Baalamash et al. found that CNNs demonstrated a high performance when classifying a dataset of Saudi traditional clothing (20). In a similar study conducted by Nawaz et al., CNNs were able to classify a dataset of Bangladesh's traditional clothing with high accuracy, further supporting the effectiveness of CNNs in classifying traditional wear (21).

As for research on traditional clothing classification using ViT, only one notable study was found. In their study, Wang et al. found that the ViT performed worse than the CNN when classifying minority Chinese traditional clothing (22).

As shown in Table 2, CNNs have shown promising results utilizing the model for traditional clothing classification for various cultures. In the study conducted by Wang et al., the CNN outperformed the ViT in terms of accuracy, but not by a significant amount. CNNs are the primary favorite in terms of classifying traditional clothing, but the capabilities of ViTs for classifying traditional clothing remain relatively unknown.

Table 2. Literature accuracy scores classifying traditional clothing

Author	Dataset	Dataset Size	Best Model Accuracy
Herwinsyah & Jaya (2024)	Indonesian traditional clothing	700	CNN: 79.34%
Baalamash et al. (2024)	Saudi traditional clothing	339	CNN: 84.85%
Nawaz et al. (2018)	Bengali traditional clothing	389	CNN: 89.22%
Wang & Wen (2024)	Minority Chinese traditional clothing	2,222	CNN: 99.1% | ViT: 98.6%

3.3 Gap in existing research

Although a large body of research has been conducted to analyze the differences between the CNN and ViT models, existing studies have had contrasting opinions over which model performs better over a wide range of differing tasks. Specifically, there is still much uncertainty when selecting between the CNN and ViT for image classification tasks, such as classifying the traditional clothing of different cultures. When it comes to the task of classifying traditional attire, research is limited regarding the cultures included in the dataset of existing studies.

Additionally, since past research has shown that ViTs have outperformed CNNs in tasks such as fashion classification, ViTs could potentially excel at classifying traditional clothing. Due to the fact that only one notable study was found that analyzed the ViT's performance when classifying traditional attire, there is a knowledge gap surrounding whether the ViT or CNN is more accurate for classifying the traditional attire of different cultures. Therefore, a study comparing the CNN and ViT in classifying a novel dataset of traditional clothing is necessary to understand which model is better suited to recognize the unique styles, colors, designs, and patterns of the clothing of cultures that have not yet been explored. To contribute to the field of knowledge, this study explored the traditional attire of three cultures of East Asia which have not yet been subjected to this ML comparison. Specifically, the traditional clothes of South Korea and Japan, which have a large number of accessible images online, have surprisingly been left unexplored by scholarly work. Additionally, although Wang et al. explored the traditional clothing of specific minority Chinese groups, no studies have been found to analyze the traditional attire of China as a broader culture, which has a greater variety of designs and also has a significant number of available images online. Thus, this study aimed to quantitatively analyze the differences in effectiveness between the CNN and ViT models when classifying a dataset of Chinese, Japanese, and South Korean traditional clothing images.

4. Methods

To understand the differences between the CNN and ViT models when classifying a dataset of East-Asian cultural wear, two ML models were trained and tested on a dataset that contains images of the aforementioned clothing. To numerically analyze the difference in effectiveness between the two ML algorithms in this study, a quantitative comparative experimental research design was

employed to isolate the relationships between the ML models and their corresponding accuracy scores. Compared to other designs, such as a causal-comparative research design, experimental research allows for significantly greater control over the experimental setting, whence the differences in performance can be attributed directly to the differences in the ML model rather than potential confounding variables. Therefore, a comparison between quantitative metrics of the two ML models would potentially reveal insights into which model was better suited to classify a dataset of Chinese, Japanese, and South Korean traditional clothing.

4.1 Training environment

To develop ML models quickly and effectively, ML libraries and frameworks are utilized to "simplify the development, training, and launching of ML and DL models" (23). Therefore, the CNN and ViT models were coded and constructed with the Python library PyTorch, which is a flexible framework that allows rapid prototyping of a wide variety of DL models and offers support for image classification (24). Moreover, PyTorch is "widely used in image recognition" and popular in "cutting-edge DL research and development" (23).

To train larger and more complex ML algorithms, especially DL models, extensive computational power is required to handle the large amounts of data and mathematical computations necessary in a model's training stage. Specifically, graphics processing units (GPUs) are often utilized to make training as fast as possible, since large datasets often take anywhere from hours to days to train. Therefore, Google Colaboratory was used for this purpose. This is a well-used online service that allows users to run executable code online and, more importantly, allows free access to GPUs–something that very few services offer. The training and testing computational environment was kept constant between the comparison of the two models, as shown in Table 3.

Table 3. Training and testing environment

Hardware & Software	Characteristics
Platform	Google Colaboratory
Graphics	NVIDIA Tesla T4 GPU
Framework	PyTorch
Programming Language	Python

Due to the limited access to expensive GPUs, the chosen training environment setup was the best option in terms of meeting financial constraints and maximizing training efficiency. However, faster, more efficient, and more extensive training would likely be possible if higher-power GPUs could be accessed.

4.2 Dataset

One of the most significant factors for a ML model's performance is the dataset on which it

is trained since the model's "knowledge" is generated through learning patterns directly from its training data. As shown in Tables 1 and 2, larger dataset sizes tend to lead to greater accuracy from a ML model since there is more data to learn from. Thus, images were collected from the internet to generate a sufficiently large dataset of Chinese, Japanese, and South Korean traditional clothing. Specifically, search queries were entered into Google, and the resulting images were downloaded automatically via a Google Chrome extension termed 'Download All Images' (MeryDev). The search queries are shown in Table 4.

Table 4. Dataset Google search queries

Type of Traditional Clothing	Search Query
Chinese	"Chinese Traditional Clothing"
South Korean	"South Korean Traditional Clothing"
Japanese	"Japanese Traditional Clothing"

These images were manually parsed to eliminate any photos unrelated to the traditional clothing of the three countries within this study that could have been unintentionally downloaded into the dataset. Additionally, to mitigate bias and ensure fairness in the dataset, all three cultures within the dataset had equal numbers of images. However, due to disparities in the available images online, the number of male and female images in the dataset was not balanced (Table 5).

Slight alterations were applied to each image in the dataset, in order to artificially generate more images (the so called synthetic data), which "can help deep learning models perform better by…creating different and diverse samples" (25). The following alterations were carefully selected from the ones utilized by Herwinsyah et al. (18). Rotation allowed images to be randomly rotated to generate new data without altering the image's culturally significant colors and patterns. Brightness changes were used to alter the brightness level of the images within the dataset, generating new data. Brightness changes were set to ±15%, which was considered large enough to replicate variability in lighting conditions in the real world without distorting the culturally significant colors of the clothing. Horizontal and vertical shifts were used to randomly translate the input image and create new data. These shifts avoided distorting cultural colors or patterns. These three types of augmentations were each applied once to every image in the dataset, creating three augmented photos from each original image. The testing data of 600 images was stratified split from the training data of 2400 images before image augmentation. Subsequently, only the training set data was augmented to 9600 images (Table 5). A sample of the dataset is shown in Figure 6.

Table 5. Dataset

Class	Number of original Images	Number of Male Images	Number of testing images split before augmentation (20% of original)	Augmented Number of training Images (80% of original after split)*	Total Dataset Size
Chinese	1,000	177	200	3200	3400
South Korean	1,000	125	200	3200	3400
Japanese	1,000	217	200	3200	3400
Total	3,000	519	600	9600	10200

*The remaining 800 images of each category (after split for testing) were multiplied by 4 (1 original + 3 augmented)

Figure 6. Sample of 30 images from the dataset

4.3 Transfer Learning

ML researchers often use transfer learning to improve model performance without needing extra data. Transfer learning allows a new ML model to leverage the knowledge that another ML model has learned from being trained on another dataset. This saves significant time and computational power, as the new ML algorithm will possess some preexisting knowledge when training (26). The chosen transfer learning models were ResNet50 for the CNN and the ViT-B/16 because both were trained on ImageNet-21K, a large dataset of assorted images that provided both models with a wide and extensive variety of information. Due to both models being pre-trained on the same dataset, the CNN and ViT possessed the same pre-existing knowledge and thus ensured a fair comparison between the two models. Additionally, ResNet50 is known for

"compelling accuracy" and "saving computing resources and training time" (27), and ViT-B/16 is known for being effective on different dataset sizes and being computationally efficient (16). Thus, these transfer learning models would effectively improve both model performances. To conduct transfer learning, PyTorch's built-in functionality was used to load the pre-existing algorithms and configure them for transferring their knowledge to the new CNN and ViT models constructed in this study.

4.4 Model training

To best compare the effectiveness of the two ML models in this study, settings were chosen to optimize model training. Therefore, both models would be compared at their optimal performance, providing more insights into their utility for classifying East-Asian traditional clothing.

As mentioned in section 2 (Background), ML models analyze input data to recognize patterns and "learn" how to classify them. For ML models to perform better through training, they should iterate through the entire dataset a certain number of times, which is controlled by a parameter known as the epoch parameter (18). In other words, one epoch means the ML algorithm has viewed the dataset once. By viewing the dataset several times, the ML model adjusts its weights and biases to optimize classification accuracy. Additionally, to balance computational power, training time, and results, the optimal number of epochs was found to be 30, as it provided a large number of iterations for the model to learn from while also not consuming too much time to train. However, with an increased number of epochs and a more extensive training environment, the accuracy of the two algorithms could potentially be increased.

Another training parameter chosen manually was the learning rate. The learning rate determines the rate at which new information learned by the ML model replaces old information. A learning rate of 0 will prevent the ML model from learning any new information, while a learning rate of 1 will prevent the ML model from preserving old information previously learned in an earlier training iteration (28). The default learning rate of 0.0001 was chosen for this study, since extensive testing requires techniques that were not feasible given the time constraints of the study. Since this learning rate was kept as a control variable between the training of both ML models, both models were equally impacted by the parameter and the comparison remained fair. Thus, the choice of not altering the learning rate was acceptable for the study's design. However, more extensive experimentation could lead to a more optimal learning rate and potentially increased accuracy in both models.

An additional training parameter that was adjusted was the batch size of training. Batch size determines the number of images an ML model sees before its weights are updated (29). Since the ML algorithm can see more data with a larger batch size, increasing the batch size can lead to faster training. However, as a result of utilizing a larger batch size, more computational power is typically needed. Therefore, a batch size of 32 was chosen as it allowed the ML model to see a significant number of images before updating weights,

which decreased training times to fit within the study's time constraint, and also did not require extensive computation.

The optimizer used to train the ML models was also chosen manually. Optimizers are utilized to improve model accuracy through specialized algorithms (30). The Adam optimization algorithm was chosen due to its "easy implementation" and "efficient computing" (30) that kept the study within time constraints.

A fixed train-test split of the dataset was chosen, which refers to the amount of training data and testing data. K-fold cross-validation was not performed due to the time and computational restraints present and since a single train-test split was deemed sufficient for the scope and resources of the study. The chosen split for the dataset was 80/20, meaning that a random 80% of the dataset would be for training data while the remaining 20% of the data would be for testing. This split allows for a large portion of the data to be used to train the ML models, which optimizes learning while also saving enough data to be used to test the models and actually evaluate their performance. The train-test split was stratified along cultural lines, allowing each culture clothing to be represented equally in both the ML models' training and testing. However, stratification did not account for gender distribution.

The final variable chosen was the loss function. The loss function calculates the "error or loss between the actual and desired output" and guides the ML model to update its weights to improve model performance (30). The Cross-Entropy loss function was used since this loss function is the most commonly used in ML tasks that involve multi-class classification (9). The study's dataset utilized three classes: Chinese traditional clothing, South Korean traditional clothing, and Japanese traditional clothing.

To provide a fair and unbiased comparison between the CNN and the ViT, the parameters shown in Table 6 were kept the same between the training of the two ML models.

Table 6. Training Configuration

Parameter	Value
Epochs	30
Learning Rate	0.0001
Batch Size	32
Optimizer	Adam
Train-Test Split	80/20
Loss Function	Cross Entropy

4.5 Model evaluation

Evaluation metrics–tools that assess the effectiveness of an ML model from different perspectives (32)–were used to directly compare various aspects of performance of the CNN and ViT models when classifying the

dataset of East Asian cultural wear and therefore identify the difference in effectiveness between the two models to reach the research goal. The following evaluation metrics were used in this study after being carefully selected from Vujovic (33). Accuracy measures the proportion of all images that were correctly classified out of the total number of images. Recall, also called the true positive rate or sensitivity measures the model's ability to correctly identify all positive instances out of all the actual positive cases present in a certain class of the dataset. Recall = True Positives / (True Positives + False Negatives). Then, the recall of each class is averaged together to provide a total recall score for the ML model. Precision, also called specificity, measures how many predictions for a certain class are actually correct. Precision = True Positives / (True Positives + False Positives). Then, the precision of each class is averaged together to provide a total precision score for the ML model. The F1 score provides a balance of both the recall and precision of an ML model. It is the harmonic mean of precision and recall.

To calculate these metrics, the Python library scikit-learn was used, which provides built-in functionality to calculate all of the aforementioned metrics automatically. These metrics were used to numerically compare the CNN and ViT to provide an unbiased and well-rounded comparison of the effectiveness of the two models, where a score of 0% indicates the worst performance and 100% indicates the best performance for each selected metric. Finally, the accuracy and loss of the two ML models over time was graphed to analyze how the models learn, which is key to understanding how the model's effectiveness changed as the models trained.

4.6 Model explainability

To further explore the performance and enhance the explainability of both ML models, two additional analyses were conducted. First, to determine the influence of background elements that were within some of the traditional clothing images–such as buildings, plants, and landscapes–on accuracy, the ML models were also tested on the same test split with the image backgrounds removed using the rembg Python library.

Additionally, to investigate the processes the ML models use to identify each clothing culture, a supplemental evaluation dataset of Bengali, Indian, Indonesian, and Saudi Arabian traditional clothing images was curated to be used exclusively as a test set. This dataset was also generated using the 'Download All Images' Google Chrome extension. Each clothing culture in this dataset had 250 images, resulting in a total supplemental evaluation dataset size of 1,000 images. It was thought that, based on how the ML models misclassified these new cultural clothes as Chinese, Japanese or Korean, the visual features that the models associated with the Chinese, Japanese, and South Korean traditional clothes could be deciphered.

5. Results

5.1 Classification performance

Table 7 compares the final evaluation metric scores for the CNN and ViT. The results show that although the CNN performed relatively well with a classification accuracy of 73.17%,

the ViT outperformed the CNN in every evaluation metric, with a notable classification accuracy of 88%. Specifically, the ViT had a high recall, precision, and F1-score, indicating that the model could effectively distinguish the cultures of the various traditional clothes with minimal mis-classification. However, the CNN did not achieve significantly low scores in those metrics either, with slightly lower but still promising scores of over 73% for recall, precision, and F1. Overall, the results show that the ViT was able to more accurately classify the different countries of origin of the traditional clothing dataset.

Table 7. Final CNN and ViT training results

Model	Accuracy	Recall	Precision	F1-Score
CNN	73.17%	73.17%	73.15%	73.10%
ViT	88.00%	88.00%	88.01%	87.99%

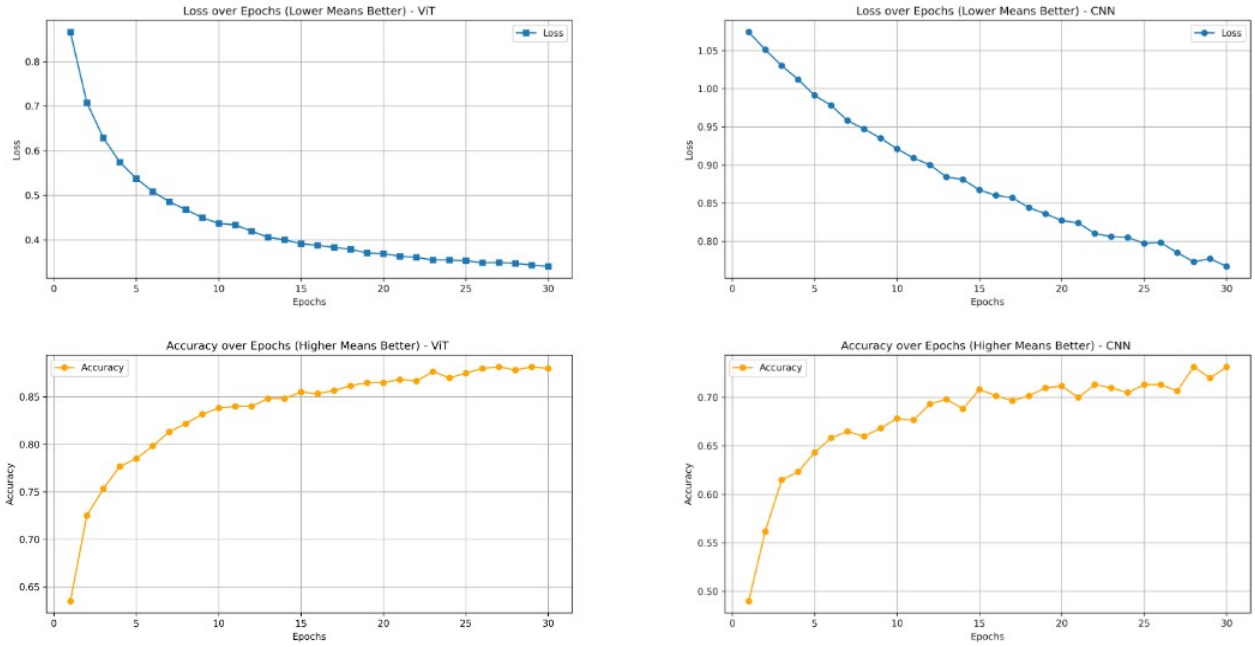

Figure 7. ViT and CNN loss and accuracy curves

Figure 7 shows an overview of how the performance of the CNN and ViT changed over the course of model training. By comparing the loss graphs of both models, it is evident that the ViT's loss score decreased significantly more and faster than the CNN, indicating that the ViT trained more effectively than the CNN. Moreover, the accuracy graphs show that the ViT had a more consistent and stable increase in accuracy, while the CNN had larger and more frequent fluctuations. Thus, the ViT also showed more promising ability to learn information from datasets.

Table 8. Best Accuracy Scores for the CNN and ViT models

Model	Best Accuracy	Epoch of Best Accuracy
CNN	73.17%	28
ViT	88.17%	27

Table 8 showcases the peak performance of both models. The results show that the ViT reached its peak accuracy slightly earlier than the CNN, once more indicating faster training. Figure 7 shows that both models began to plateau after their peak accuracies, but also that the CNN's accuracy remained significantly lower than the ViT's. Thus, the CNN is unlikely to close the performance gap even if it were trained for more epochs. Moreover, by analyzing Figure 7, it is evident that the ViT surpassed the CNN's peak accuracy on its third epoch, once again showing the ViT's ability to classify East-Asian traditional clothing at a faster rate. Overall, the results show that despite both models converging near the end of training, the ViT consistently outperformed the CNN.

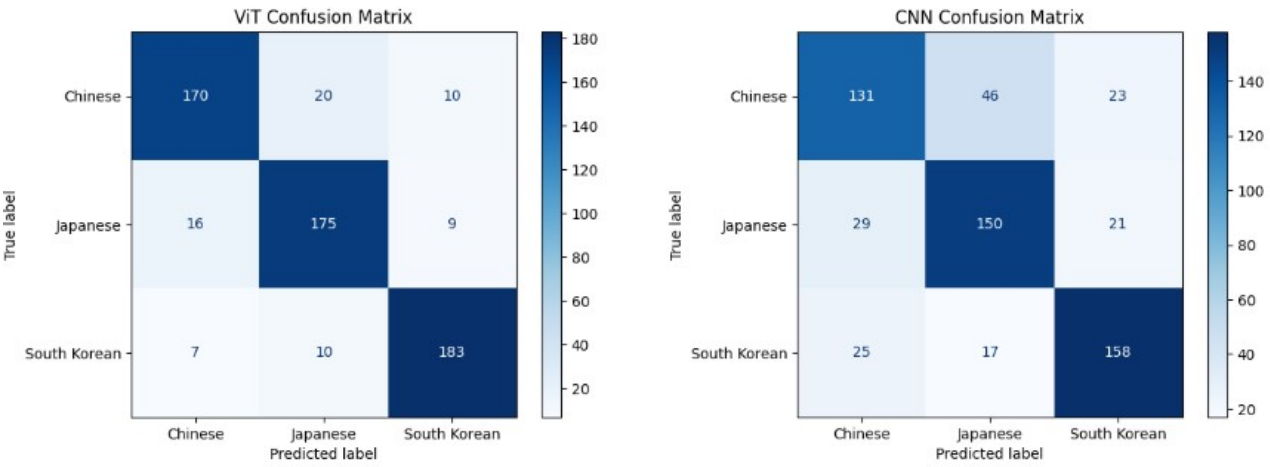

Figure 8. ViT and CNN confusion matrices

Figure 8 provides confusion matrices that compare how both ML models performed on the task of classifying the study's dataset. It is evident that the CNN struggled notably more with classifying Chinese traditional clothing than the ViT did. Specifically, the CNN was found to misclassify Chinese traditional clothes as Japanese far more often, whereas the ViT showed a smaller and more balanced distribution of errors. Moreover, the classification of both Japanese and South Korean traditional clothing was marginally worse for the CNN, demonstrating the ViT's superiority in classifying every clothing culture within the study.

5.2 Model explainability analyses

Table 9. CNN and ViT performance with removed backgrounds from images

Model	Accuracy	Recall	Precision	F1-Score
CNN	68.33%	70.89%	68.83%	68.25%
ViT	82.17%	82.83%	82.17%	82.24%

Table 9 compares the classification performance of the CNN and the ViT on the test split with the image backgrounds removed. Compared to the results in Table 7, the CNN's accuracy decreased by 4.84%, while the ViT's accuracy decreased by 5.83%. Moreover, the ViT's recall, precision, and F1-score all decreased by a greater percentage than the CNN. Overall, the results show that both models relied on background elements to classify the East-Asian traditional clothing, but still primarily utilized clothing features for classification since the metrics did not decrease significantly.

Table 10. CNN and ViT performance on the supplemental evaluation dataset

Culture	Model	Predicted as Chinese	Predicted as Japanese	Predicted as South Korean
Bengali	CNN	145	52	53
	ViT	204	25	21
Indian	CNN	142	57	51
	ViT	218	15	17
Indonesian	CNN	115	102	33
	ViT	171	56	23
Saudi Arabian	CNN	102	85	63
	ViT	133	95	22

Table 10 shows the breakdown of the CNN's and ViT's mis-classification of the supplemental evaluation dataset. Evidently, both models exhibited a strong bias towards classifying the unseen cultural attire as Chinese clothing, with the CNN being marginally less biased. Overall, the CNN's mis-classifications are more evenly distributed, while the ViT demonstrated a stronger tendency towards labeling these out-of-dataset culture clothes as Chinese traditional clothing. Both models predicted Bengali and Indian clothing as Chinese more often than they predicted Indonesian and Saudi Arabian clothing as being Chinese. Both models also predicted Indonesian and Saudi Arabian as Japanese more often than they predicted Bengali and Indian clothing as being Japanese. Both models ~ predicted all four cultural clothing as being equally South Korean.

6. Discussion

6.1 New understanding

This experiment demonstrated the superiority of the ViT over CNN when used to classify a

dataset of traditional clothing images from China, South Korea, and Japan. Unlike the previous research discussed, which mainly focused on the utility of the CNN for classifying traditional clothing, this study revealed the under-explored yet promising ability of the ViT model. The ViT outperforms the CNN model in every metric measured –as shown in Table 7–when classifying a dataset of East-Asian traditional clothes. Additionally, the ViT's ability to learn information faster was highlighted, which may have contributed to allowing the model to classify images of traditional clothing with greater accuracy than the CNN.

6.2 Explanation of findings

The ViT likely outperformed the CNN due to various differences in the underlying structure of the models. Firstly, as noted by Maurício et al., the ViT has access to global features in an image due to its self-attention mechanism. In contrast, Rachman et al. note that the CNN can only access local features due to its filter mechanism. This difference in access to image features could limit the CNN on the task of classifying East-Asian traditional clothing, as viewing all aspects of the image–with its various colors and patterns–may be necessary to differentiate between the clothes of these three cultures. Moreover, the difference in feature access could explain how the ViT achieved faster and more stable training results, as it was able to take into account more features of the images at a time.

Past research (Table 1) has generally shown that ViTs outperform CNNs on larger datasets of over 1,000 images, which aligns with this study's results. Since this study utilized a dataset of a total of 10,200 images, the ViT may have had the advantage of being better suited for larger datasets than the CNN, which has been generally shown to outperform ViTs on smaller datasets of less than 1,000 images.

6.3 Model robustness and explainability

The two additional analyses performed to investigate the two ML models also provide insights into the robustness of the algorithms in real-world use. Evaluating the model classifications on background-removed images demonstrated that both models are moderately robust, with marginal losses in all metrics tested in the study. Specifically, the ViT experienced greater losses than the CNN. This shows that the ViT's global feature access potentially enabled the model to detect key features in the background elements of the traditional clothing images that the CNN may not have utilized. On the other hand, the CNN's greater resistance to the backgrounds being removed demonstrates how its local feature access may have enabled it to have greater focus on distinctive characteristics found on the traditional clothing itself, such as fabrics, textures, and patterns. Overall, the results of the first analysis demonstrate that the CNN may be more robust to stark changes in background and may generalize better to images of different backgrounds.

The analysis of both models on the supplemental evaluation dataset provided insights into how well both models handled unexpected data. The results demonstrated that both models exhibited a strong tendency to classify the out-of-distribution traditional clothes as Chinese, with the ViT showing a stronger bias. These results show how the ViT

may overgeneralize some detected features as Chinese, while the CNN has more balance in terms of classifying based on detected features. This underscores the difference between how the ViT's self-attention mechanism classifies images in comparison to the CNN's filter mechanism. Overall, both models demonstrated low robustness to unexpected data.

Visual comparison between the supplemental evaluation dataset and the East-Asian dataset showed one clear pattern: many Chinese, Bengali, Indian, and Indonesian traditional clothes shared a primary color of red. When comparing this observation with Table 10, it is evident that both models associated Bengali and Indian clothes with the Chinese culture more than Saudi Arabian clothes, which did not share the primary color of red. This demonstrates that both models likely relied on color as one of many attributes to classify the East-Asian traditional clothes.

Figure 9. Left panel: Bengali traditional dress, Middle panel: Indian traditional dress, Left panel: Chinese traditional dress

In addition, as can be seen in Figure 9, the traditional Chinese Hanfu and/or Qipao dress is similar to the Bengali or Indian Saree in that the former has loose flowing silhouettes which look similar to the 'pallu' of the Bengali or Indian Saree, The Chinese garment has a cross-collar, one of whose lapels is replicated by the Indian or Bengali Saree. The lapel in all three cases is diagonal and generally of a different color than the rest of the garment. All three dresses give the appearance of being draped around the hands and/or torso. In all cases, the garments reach all the way almost to the floor.

Figure 10. Left panel: Japanese traditional dress, Middle panel: Indonesian traditional dress, Right panel: Saudi Arabian traditional dress

As seen in figure 10, the Kimono; the traditional dress of Japan is worn with a broad sash around the waist and stomach and gives the appearance of a tighter fit than the Chinese Hanfu or Qipao. The dress also does not go all the way to the floor. The Indonesian and Saudi Arabian dresses are designed stylistically similar to the Kimono in these features. The kimono does not have as visible or as ornate a cross-collar as does the Chinese Hanfu; and neither the Indonesian nor the Saudi Arabian dresses have prominent cross-collars as well. Furthermore, the cross-collars are usually the same color as the rest of the garment. In all three dresses, there is not as much appearance of extraneous or superfluous cloth as there is in the dresses shown in Figure 9. The kimono and the Saudi Arabian Abaya do have flowing motifs around the hands and as a draped robe respectively so that it is possible for the ML models (especially the global feature intensive ViT model) to (mis)classify both these dresses as Chinese, depending on the image attributes of angle, lighting, etc. added to the various regional and spatio-temporal variations in traditional cultural clothing.

This feeding of out-of-dataset and unseen data into classification models hence represents a deliberate method to discern the workings of ML models; and has not explicitly been reported as being such in the literature. The method should be required as a standard procedure in reporting ML model results because it can provide information on the otherwise ‘blackbox’ non-explainable nature of

ML; *the* primary reason why ML models have not yet been used as 'stand alone' discriminators or classifiers. It is hoped that more scholars will adopt this paradigm.

6.4 Implications

The ViT's ability to classify the traditional clothing of cultures from China, South Korea, and Japan more effectively than the CNN in this study has significant implications for the field of cultural preservation. Since the ViT may be able to recognize subtle differences in cultural clothing due to the vast number of details and patterns it learns through training, museums and cultural organizations could construct software that integrates the ViT's accurate classification of such traditional clothes to document and categorize these countries' artifacts via pictures taken by individuals. Additionally, due to the computational efficiency of ML models, the ViT can potentially help identify and verify the traditional garments of China, South Korea, and Japan more rapidly, allowing for easier generation of large databases of traditional clothing images. Thus, the ViT has the potential to be a significant aid in effectively identifying and categorizing these traditional clothes and, therefore, benefit the preservation of the Chinese, South Korean, and Japanese cultures.

Another potential use of the ViT's accuracy is in the fashion industry. Businesses that sell the traditional clothes of China, South Korea, and Japan may benefit from utilizing the efficient categorization of the ViT within their websites to make product sorting potentially faster and more accurate. Moreover, search and recommendation engines can implement the ViT into their software to potentially provide more relevant images and sources of such cultural clothing, thus improving user experience.

6.5 Limitations

Several limitations of this study may influence the generalizability and validity of aspects within the paper. The study used a dataset consisting of images sourced from Google. Since there was no feasible way, given time constraints, to validate that the traditional clothing images were truly representative of the cultures analyzed in the study, the ML models in this study may have learned unrepresentative patterns that would limit their real-world use. Thus, this study's findings may not fully represent the ViT and CNN's performance in classifying Chinese, South Korean, and Japanese traditional clothing. Moreover, the dataset has significantly more female images, potentially leading the models in this study to develop a bias towards features more prominent in female traditional clothing. Future research could involve a dataset of traditional clothing that is fully curated and verified by experts and that balances gender, thus helping minimize the potential for cultural misrepresentation and gender bias.

Another limitation of this study is that it used specific training parameters, as shown in Table 6. Since all of these settings alter the training configuration of the ML models, the performance of the CNN and ViT might differ if the parameters were to be changed. Thus, the results of this study can only be generalized to CNNs and ViTs specifically trained with the parameters used in this study. Future studies can explore different training parameters and

determine which parameters optimize model performance. Obviously, the image angle, lighting, the spatio-temporal differences in traditional clothes of the same culture, whether the image represents a frontal or profile view, all influence the ML result to various extents. The larger the number of original (not synthetic) images, the better the prediction of the ML models is likely to be.

An additional limitation of this study is that, due to time constraints, only two transfer learning algorithms were trained and evaluated. Since these transfer learning models were pre-trained on a specific dataset, other transfer learning models not used in this study–which have different pre-existing knowledge derived from other datasets–may yield different results from this study. Thus, future research can test various transfer learning models or even train models from scratch to get a more well-rounded idea of whether the CNN or ViT is more effective for classifying traditional clothing.

Future studies can also investigate the performance of other image classification models that have not been widely explored. Additionally, datasets of traditional clothing from different countries can be developed and used to train ML algorithms. These studies would provide a broader understanding of the capabilities of ML models to identify traditional clothing.

Finally, to improve the explainability of how ML models classify traditional clothing, future studies can use tools such as Grad-CAM and attention heatmaps to analyze what features are actually being detected. Feeding out-of-dataset information – as has been done in this work – represent a useful method to achieve explainability.

7. Conclusion

In this study, a CNN and ViT were trained on a dataset of 3200 traditional clothing images each from China, South Korea, and Japan to evaluate the effectiveness of these models in classifying culturally significant artifacts. Moreover, the study hoped to examine the effects of architectural differences between models on performance.

The results of this paper showed that the ViT (scores of > 87%) outperformed the CNN (with scores of > 73%) across all the tested metrics (accuracy, recall, precision and F1 score). This was likely attributed to its self-attention mechanism, which allowed for global image feature detection compared to the local focus of the CNN.

Moreover, to evaluate the robustness of the models trained in the study, two additional datasets were used for testing. Testing found that the CNN demonstrated stronger robustness to image background differences than the ViT did, highlighting the effect of structural differences in both models. Both models were obviously inaccurate in classifying out-of-distribution/dataset data, but provided clues with regard to which features were utilized for classification. Such deliberate feeding of out-of-distribution/dataset data can therefore increase the explainability of these ‘blackbox’ models, providing a path for stand-alone applications.

Overall, the study demonstrated promise for the use of ML models in traditional clothing classification, with the ViT in particular providing accurate results. With the rapid development of ML, these models may prove to be valuable tools for cultural preservation organizations and online businesses in the future.

Acknowledgements

I would like to thank Mr. Sanchez and Dr. Taite for their guidance and expertise during the literature review and methodology development process.

8. References

1. Alnuaimi, A. F. A. H., Albaldawi, T. H. K. (2024). An overview of machine learning classification techniques. BIO Web of Conferences, 97, 00133. https://doi.org/10.1051/bioconf/20249700133

2. Khan, A., Laghari, A., Awan, S. (2018). Machine Learning in computer vision: A Review. EAI Endorsed Transactions on Scalable Information Systems, 8(32), e4. https://doi.org/10.4108/eai.21-4-2021.169418

3. Amirova, A., Aghamaliyeva, Y., Salehzadeh, G. (2024). Cultural and historical significance of the traditional attire of the Far East. InterConf, 43(193), 186–191. https://doi.org/10.51582/interconf.19-20.03.2024.020

4. Yao, Y. (2024). The Impact of Deep Learning on Computer Vision: From Image Classification to Scene Understanding. Int. J Scientific Res. & Management, 12(9), 1428-1433. https://doi.org/https://doi.org/10.18535/ijsrm/v12i09.ec02

5. Zabaleta, D. S. (2019). Drawing of neuron synapse. https://commons.wikimedia.org/wiki/File:Two_neurons_connected.svg.

6. Salatas, J. (2011). Multilayer Neural Network. https://commons.wikimedia.org/wiki/File:Multilayer_Neural_Network.png.

7. Islam, M., Chen, G., Jin, S. (2019). An overview of neural network. American Journal of Neural Networks and Applications, 5(1), 7-11. http://dx.doi.org/10.11648/j.ajnna.20190501.12

8. Luo, X., Yan, R., Wang, S., Zhen, L. (2023). A fair evaluation of the potential of machine learning in Maritime Transportation. Electronic Research Archive, 31(8), 4753–4772. https://doi.org/10.3934/era.2023243

9. Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., et al. (2021). Review of Deep Learning: Concepts, CNN Architectures, challenges, applications, Future Directions. Journal of Big Data, 8(1), 53. https://doi.org/10.1186/s40537-021-00444-8

10. Bbouzidi, S., Hcini, G., Jdey, I., Drira, F. (2024). Convolutional Neural Networks and Vision Transformers for Fashion MNIST Classification: A Literature Review. https://doi.org/10.48550/arXiv.2406.03478

11. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A. (2021). Do vision transformers see like convolutional neural networks?. Advances in neural information processing systems, 34, 12116-12128.

12. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. (2023a). Attention is all you need. https://doi.org/10.48550/arXiv.1706.03762

13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. https://arxiv.org/abs/2010.11929v2

14. Maurício, J., Domingues, I., Bernardino, J. (2023). Comparing vision transformers and convolutional neural networks for Image Classification: A Literature Review. Applied Sciences, 13(9), 5521. https://doi.org/10.3390/app13095521

15. Gai, L., Xing, M., Chen, W., Zhang, Y., Qiao, X. (2024). Comparing CNN-based and transformer-based models for identifying lung cancer: which is more effective?. Multimedia Tools and Applications, 83(20), 59253-59269. https://doi.org/10.1007/s11042-023-17644-4

16. Zhao, P., Li, C., Rahaman, M. M., Xu, H., Yang, H., Sun, H., Jiang, T., & Grzegorzek, M. (2022). A comparative study of deep learning classification methods on a small environmental microorganism image dataset (EMDS-6): From Convolutional Neural Networks to visual transformers. https://doi.org/10.48550/arXiv.2107.07699

17. Rachman, R. K., Susanto, A., Nugroho, K., Islam, H. M. M. (2024). Enhanced Vision Transformer and Transfer Learning Approach to Improve Rice Disease Recognition. Journal of Computing Theories and Applications, 1(4), 446-460. http://dx.doi.org/10.62411/jcta.10459

18. Abd Alaziz, H. M., Elmannai, H., Saleh, H., Hadjouni, M., Anter, A. M., Koura, A., Kayed, M. (2023). Enhancing fashion classification with Vision Transformer (VIT) and developing recommendation fashion systems using DINOVA2. Electronics, 12(20), 4263. https://doi.org/10.3390/electronics12204263

19. Herwinsyah, H., Jaya, D. Y. (2024). Accuracy potential of the Convolutional Neural Network (CNN) in recognizing traditional clothing. IT Journal Research and Development, 8(2), 95-106. https://doi.org/10.25299/itjrd.2023.12690

20. Baalamash, F., Debis, R., Al-Barhamtoshy, H. (2024). Classification of Saudi traditional costumes using deep learning technique. International Design Journal, 14(2), 50–60. https://doi.org/10.21608/idj.2024.254693.1106

21. Nawaz, M. T., Hasan, R., Hasan, M. A., Hassan, M., Rahman, R. M. (2018). Automatic categorization of traditional clothing using convolutional neural network. In 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS) (pp. 98-103). IEEE. http://dx.doi.org/10.1109/ICIS.2018.8466523

22. Wang, T., Wen, B. (2024). Research on image recognition of ethnic minority clothing based on improved Vision Transformer. Mathematical Foundations of Computing, 7(1), 84–97. https://doi.org/10.3934/mfc.2022054

23. Rane, N. L., Mallick, S. K., Kaya, Ö., Rane, J. (2024). Tools and frameworks for machine learning and Deep Learning: A Review. In Applied Machine Learning and Deep Learning: Architectures and Techniques (pp. 80-95). Deep Science Publishing. https://doi.org/10.70593/978-81-981271-4-3_4

24. NVIDIA. (n.d.). What is PyTorch? NVIDIA Data Science Glossary. https://www.nvidia.com/en-us/glossary/pytorch/

25. Alomar, K., Aysel, H. I., Cai, X. (2023). Data Augmentation in classification and segmentation: A survey and new strategies. Journal of Imaging, 9(2), 46. https://doi.org/10.3390/jimaging9020046

26. Ali, A. H., Yaseen, M. G., Aljanabi, M., Abed, S. A. (2023). Transfer learning: A new promising techniques. Mesopotamian Journal of Big Data, 2023, 29–30. https://doi.org/10.58496/mjbd/2023/004

27. Maeda-Gutiérrez, V., Galván-Tejada, C. E., Zanella-Calzada, L. A., Celaya-Padilla, J. M., Galván-Tejada, J. I., Gamboa-Rosales, H., et al. (2020). Comparison of convolutional neural network architectures for classification of Tomato Plant Diseases. Applied Sciences, 10(4), 1245. https://doi.org/10.3390/app10041245

28. Shaw, R. N., Ghosh, A., Balas, V. E., Bianchini, M. (Eds.). (2021). Artificial Intelligence for Future Generation Robotics. Elsevier.

29. Hwang, J.-S., Lee, S.-S., Gil, J.-W., Lee, C.-K. (2024). Determination of Optimal Batch Size of Deep Learning Models with Time Series Data. Sustainability, 16(14), 5936. https://doi.org/10.3390/su16145936

30. Hassan, E., Shams, M. Y., Hikal, N. A., Elmougy, S. (2022). The effect of choosing optimizer algorithms to improve computer vision tasks: A comparative study. Multimedia Tools and Applications, 82(11), 16591–16633. https://doi.org/10.1007/s11042-022-13820-0

31. Kerkhof, M., Wu, L., Perin, G., Picek, S. (2023). No (good) loss no gain: Systematic evaluation of loss functions in deep learning-based side-channel analysis. Journal of Cryptographic Engineering, 13(3), 311–324. https://doi.org/10.1007/s13389-023-00320-6

32. Hossin, M., Sulaiman, M. N. (2015). A review on evaluation metrics for Data Classification Evaluations. International Journal of Data Mining & Knowledge Management Process, 5(2), 1–11. https://doi.org/10.5121/ijdkp.2015.5201

33. Vujovic, Ž. (2021). Classification Model Evaluation Metrics. International Journal of Advanced Computer Science and Applications, 12(6). https://dx.doi.org/10.14569/IJACSA.2021.0120670