Integrating Classification and Segmentation Models with Deep Learning for Fashion Image Analysis

***Abstract*— Due to the rapid development of e-commerce, especially the fashion industry, it is now becoming increasingly difficult for humans alone to analyze fashion images, classify clothing types, and predict or classify fashion styles, requiring the power of AI. Therefore, to solve these problems, a system framework that integrates a classification model and a segmentation model is proposed. This framework consists of two pipelines. Pipeline 1 utilizes segmentation to classify clothing parts, while Pipeline 2 classifies fashion styles (Feminine, Cool, etc.). In this study, model training and evaluation experiments using the DeepFashion2 dataset, the FashionStyle14 dataset, and an unseen dataset are conducted. ResNet, DenseNet, and a random forest classifier are used as models, and evaluated model performance through a 5-part cross-validation method. The performance of the models using a five-fold cross-validation method is evaluated using metrics such as precision, recall, and confusion matrix. As a result, DenseNet-121 achieved the highest accuracy of 82.10% for Pipeline 1, and ResNet-50 achieved 73.16% accuracy for Pipeline 2. These are very high results in accuracy for Multi-label Classification (13-label and 14-label). By integrating these best models and building a prediction pipeline, practical applications in the e-commerce field are now possible.**

Keywords*— *Fashion, Fashion Image Analysis, Image Classification, Deep Learning, ResNet, DenseNet, Random Forest, Semantic Segmentation,

I. INTRODUCTION

In recent years, the e-commerce industry has experienced unprecedented growth due to the proliferation of online shopping. The fashion industry in particular has become an important part of this growth, as customers select and purchase products primarily based on image data, and to boost e-commerce, more emphasis is being placed on fashion analysis to increase the number of users. However, in the past, fashion trends were mainly communicated by models and actors on TV and in magazines and were easy to analyze, whereas now, fashion is not limited to TV and magazines, and anyone can communicate trends through social media. This is also making it difficult to identify trends. There are many challenges in analyzing large amounts of fashion images, discovering trends in real time, and accurately categorizing products. These challenges are due to the diversity of fashion styles, the large variety of clothing types, and the rapid changes in trends [1].

Therefore, the following goals are set forth in this study:

- Build and appropriately adjust a model for classifying fashion styles and a model for segmenting clothing parts into smaller pieces, respectively.
- Integrate these models and propose a new framework for comprehensive classification and segmentation.
- Incorporate domain-specific preprocessing and feature engineering methods to address issues such as similarity between classes and data diversity.

Achieving these goals should allow us to integrate classification and segmentation of fashion images and provide new solutions for fashion analysis in the e-commerce industry.

II. PROPOSED METHOD

A. Proposed System Framework

The proposed system framework is displayed in Fig 1. The framework is divided into three parts, which are Pipeline 1, Pipeline 2, and Prediction. Each pipeline addresses a different aspect of the classification tasks. Two datasets, which are the garment parts dataset and fashion styles dataset, and a set of unseen data are used for this system.

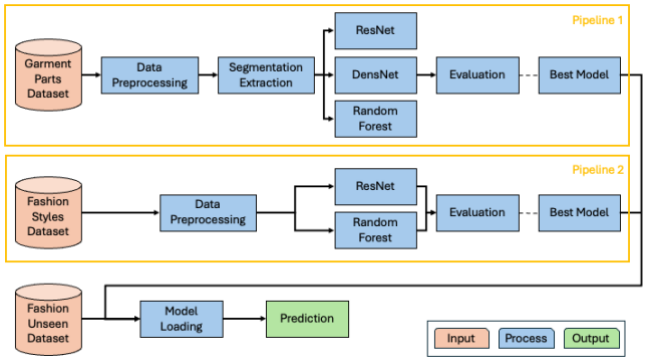

Figure 1: Proposed System Framework

Pipeline 1 focuses on the classification of garment parts with segmentation. The methodology consists of the following steps:

1) The data is preprocessed to normalize image inputs and prepare segmentation masks, which are used to isolate closing parts to improve model focus.
2) Three machine learning models, which are ResNet, DenseNet, and Random Forest, are employed for classification.
3) Each model is trained and evaluated on the dataset, and the best-performing model is selected for the prediction.

Pipeline 2 focuses on the classification of fashion styles:

1) The data is preprocessed to normalize image inputs.
2) Two machine learning models, which are ResNet, and Random Forest, are employed for classification.

3) Each model is trained and evaluated on the dataset, and the best-performing model is selected for the prediction.

The prediction pipeline integrates both best-performing models to make predictions on the unseen dataset to enable the practical application of the framework:

1) The fine-tuned models for both Pipeline 1 and Pipeline 2 are loaded into the system.
2) The garment type and fashion style is predicted based on the models.

B. Segmentation Mask

A segmentation mask is used to partition images into meaningful segments or objects [2]. Segmentation can be classified into three major tasks, which are semantic segmentation, instance segmentation and panoptic segmentation. Semantic segmentation assigns a class label to every pixel in the image, while instance segmentation extends semantic segmentation to identify individual objects, and panoptic segmentation combines both. The framework uses semantic segmentation methods to further reduce the size of the photos in the dataset and isolate regions of interest. By including a segmentation mask process in data processing, it makes the model focus on the target region and increases accuracy and efficiency in the classification task [3].

C. Random Forest

Random Forest creates multiple decision trees. For each tree, it begins with bootstrap sampling, which selects a random subset of the training dataset with replacement. At each node of the decision tree, the algorithm considers a random subset of features to decide the best split, which not only introduces diversity among the trees but also enables the model to assess the importance of features. This is achieved by measuring how much each feature contributes to reducing impurity across all splits in the forest. For classification tasks, the ensemble combines the predictions of all trees through majority voting, where the class most frequently predicted by the tree is chosen as the final output. The random forest classification algorithm is known for being fast, robust to noise, and capable of successfully identifying non-linear patterns in data [4]. Thus, it works well for datasets where feature extraction can be easily performed.

D. ResNet

Residual Network (ResNet) is a deep learning architecture that introduces residual connections, which connect non-neighboring layers, breaking the traditional chain-like structure of neural networks [5]. These connections create shortcut paths that allow information to bypass one or more layers, and it causes more effective training of very deep networks. Residual connections simplify the learning process by allowing the network to learn residual functions instead of full mappings, leading to improved training dynamics and generalization.

The Fig 2 shows the residual block of ResNet. It consists of a main path and a shortcut connection. The main path indicates batch normalization, 3x3 convolutional layers, and ReLU activation. Batch normalization normalizes the input to stabilize and accelerate training. This step makes sure the small changes in the network's weights have consistent effects on the outputs. 3x3 convolutional layers extract features from the input using learnable filters. These layers are used to learn localized patterns effectively. ReLU activation is an activation function that introduces non-linearity by applying the function. This activation function is used to model complex, which is essential to solve image classification. The shortcut connection directly passes the input to the output. This makes sure that the network can propagate gradients effectively during backpropagation. `

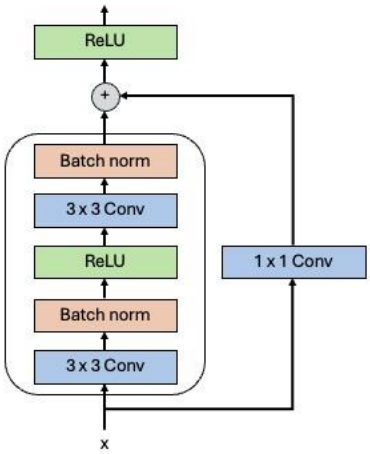

Figure 2: Residual Block

The advantage of ResNet is the shortcut connection makes gradients to flow directly to earlier layers, and it reduces the likelihood of vanishing gradients. It simplifies optimization because the network refines the input features instead of learning entirely new mapping by learning residuals.

E. DenseNet

Dense Convolutional Network (DenseNet) is another deep learning architecture that improves information flow and gradient propagation in very deep networks. However, unlike ResNet, it connects each layer to every other layer in a feed-forward manner [6]. The dense concatenation in DenseNet requires more GPU memory and longer training times.

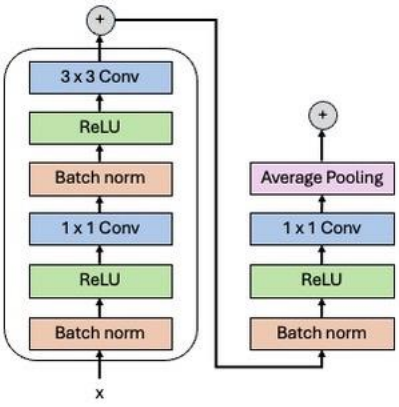

Figure 3: Dense Block and Transition Layer

The Fig 3 shows the dense block of DenseNet. Within the dense block, each layer performs a sequence of batch normalization, ReLU activation, and 3x3 convolution, 1x1 convolution like ResNet. The transition layer includes batch normalization, ReLU activation, and 1x1 convolution and average pooling. Average pooling is to reduce the spatial dimensions of the feature maps, control model complexity, and prepare the data for the next dense block. While ResNet focuses on simplifying the residual learning process by directly passing the input and applying transformations afterward, DenseNet, on the other hand, prioritizes efficient gradient flow and feature reuse by normalizing first, activating the normalized features, and then applying convolution to the already stabilized inputs.

III. EXPERIMENTS

A. Datasets

In this study, DeepFashion2 [7] and FashionStyle14 [8] were used as datasets.

1) **DeepFashion2:** 801K clothing items with annotations such as garment parts (13 garment labels), dense landmarks and segmentation masks, and 873K commercial-consumer image pairs sourced from both commercial shopping stores and consumer uploads. However, available dataset which was possible to download was 364,676 images. Since dataset is imbalanced across its 13 clothing categories, oversampling was applied for underrepresented classes to match the number of samples, and random subsampling was applied for overrepresented classes to balance the dataset. As a result, each category includes 1,985 images with segmentation masks. Categories are “short sleeve top”, “long sleeve top”, “short sleeve outwear”, “long sleeve outwear”, “vest”, “sling”, “shorts”, “trousers”, “skirt”, “short sleeve dress”, “long sleeve dress”, “vest dress” and “sling dress”.
2) **FashionStyle14:** 13,126 images distributed across 14 fashion styles, which focus on Japanese fashion styles. Each image is a single individual in a fully visible pose. Although the dataset is imbalanced a little bit, no resampling techniques were applied. Thus, approximately 1,000 images per class. The categories are “conservative”, “dressy”, “ethnic”, “fairy”, “feminine”, “gal”, “girlish”, “kireime-casual”, “lolita”, “mode”, “natural”, “retro”, “rock”, and “street”.
3) **Unseen Dataset for Testing:** Unseen fashion images are collected from the internet by the author to evaluate the proposed framework.

B. Feature Engineering Techniques

Segmentation masks are used in pipeline 1, and data augmentation such as random rotation, color jitter, and horizontal flips, are applied.

C. Cross-Validation

5-Fold Cross-validation is applied for pipeline 1 to make sure the model is evaluated on multiple splits of the data, reduce the bias, and compare several models’ performance.

D. Model Selection

1) **Pipeline 1:** ResNet 18 and DenseNet-121 were fine-tuned, and Random Forest Classifier was trained to compare the performance. DenseNet is more complex in terms of the number of layers compared to ResNet-18. The performance of all models was evaluated using cross-validation and the best-performing models were saved for use in the final system framework evaluation.
2) **Pipeline 2:** ResNet 50 was fine-tuned, and Random Forest Classifier was trained to compare the performance. The best-performing models were saved.

E. Training

A CrossEntropyLoss function was used as the loss function. The AdamW and Adam optimizer were applied for pipelines 1 and 2, and the learning rate scheduler and early stopping were implemented.

F. Evaluation

Training and validation sets for each dataset were evaluated as accuracy and loss. As an evaluation matrix, precision and recall metrics were used. Additionally, a confusion matrix was computed and visualized to analyze the classification performance for each class.

IV. RESULTS AND DISCUSSIONS

The confusion matrix and accuracy and loss graph were plot every fold, but the best performance graphs were selected to discuss.

A. Pipeline 1

1) Random Forest Classifier

Table 1: Train and Val Accuracy of all folds

	Train Accuracy (%)	Val Accuracy (%)
Fold 1	100.00%	38.60%
Fold 2	100.00%	39.24%
Fold 3	100.00%	38.60%
Fold 4	100.00%	37.14%
Fold 5	100.00%	39.41%

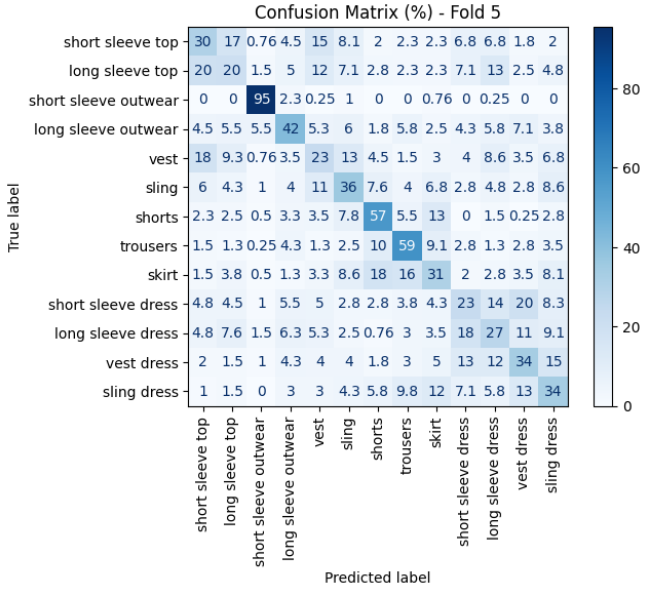

Figure 4: Best-performing Fold's Confusion Matrix

2) *ResNet 18*

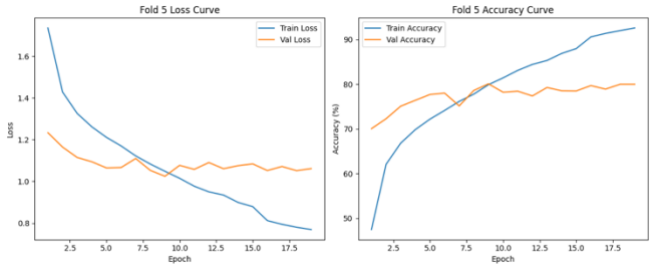

Figure 5: Accuracy and Loss Plot

Table 2: Train and Val Accuracy of all folds

	Train Accuracy (%)	Val Accuracy (%)
Fold 1	92.30%,	78.86%
Fold 2	92.03%	78.80%
Fold 3	92.04%	78.82%
Fold 4	92.23%	77.85%
Fold 5	92.57%	79.97%

Table 3: Train and Val Loss of All Fold

	Train Loss	Val Loss
Fold 1	0.7715	1.0979
Fold 2	0.7794	1.0910
Fold 3	0.7794	1.0830
Fold 4	0.7743	1.1144
Fold 5	0.7692	1.0612

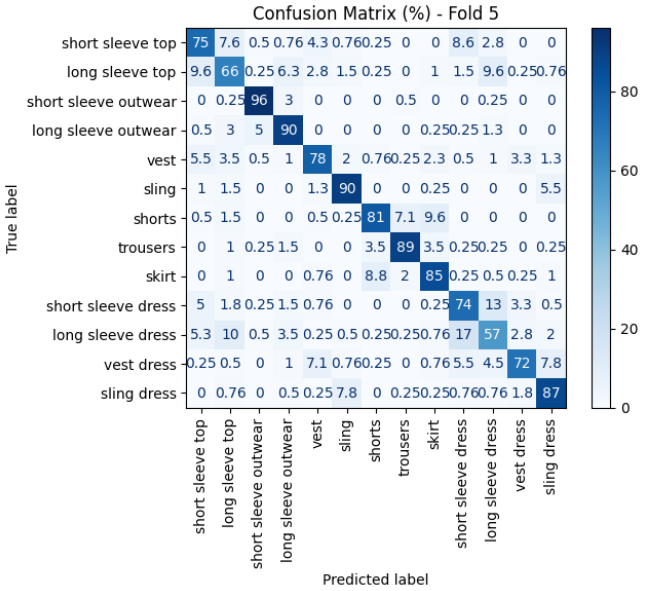

Figure 6: Best-performing Fold's Confusion Matrix

3) *DenseNet 121*

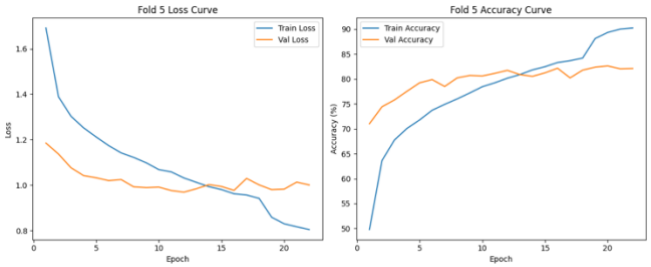

Figure 7: Accuracy and Loss Plot

Table 4: Train and Val for All Folds

	Train Accuracy (%)	Val Accuracy (%)
Fold 1	93.22%	81.69%
Fold 2	89.66%	81.71%
Fold 3	90.34%	80.68%
Fold 4	90.76%	80.82%
Fold 5	90.23%	82.10%

Table 5: Train and Val Loss for All Folds

	Train Loss	Val Loss
Fold 1	0.7361	1.0256
Fold 2	0.8257	1.0118
Fold 3	0.8053	1.0445
Fold 4	0.7930	1.0608
Fold 5	0.8047	1.0008

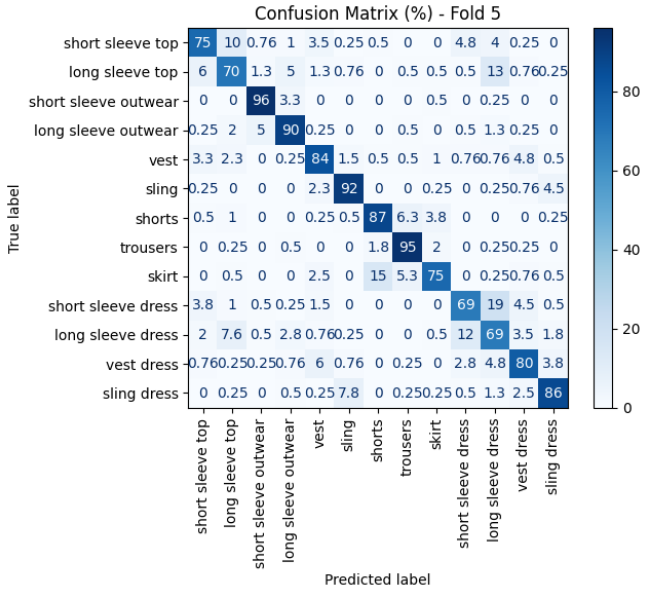

Figure 8: Best-performing Fold's Confusion Matrix

4) Discussions

Although the random Forest Classifier achieved 100 % accuracy in the training set across all folds, the validation accuracy fell between 37% to 40%. Additionally, for the confusion matrix, the diagonal line is supposed to be dark color, but the variations can be seen. Short sleeve outwear achieved the highest accuracy (95%) among labels, while long sleeve top achieved the lowest accuracy (20%). As insights, the model seems to be overfitting since the training accuracy is perfect, but the validation accuracy is low. Also, the random forest struggled to generalize to the validation dataset, which points to limitations in feature representation and model complexity for this task. On the other hand, the deep learning model, ResNet-18, achieved quite a high accuracy in both training and validation (train: approximately 92%, validation: approximately 78%). For the confusion matrix, Figure 6 shows a better result than the random forest confusion matrix. However, the highest accuracy label in the confusion matrix was 96%, same as the random forest. Fig 5 shows the accuracy and loss graph over epochs. Since the early stopping was implemented, the training stopped around epoch 18. Although there was no dramatic decrease in validation loss, it decreased little by little. Similar to this, there was no dramatic increase in validation accuracy; it increased little by little. In contrast, DensNet-121 achieved the best performance, with training accuracy ranging approximately from 89% to 93% and validation accuracy between 80% and 83%. As shown in Fig 8, the model achieved the most accurate prediction among all models, with reduced misclassifications across categories. Thus, it can be concluded that DenseNet-121 outperformed both the Random Forest and ResNet-18 in validation accuracy and generalization. The reason could be ResNet-18 has less hidden layers than DenseNet, and while it can learn hierarchical features better than Random Forest, it may limit the model's capacity to completely capture complex patterns.

B. Pipeline 2

1) Random Forest Classifier

Table 6: Train and Val Accuracy

	Train Accuracy	Val Accuracy
Average	-	31.79%

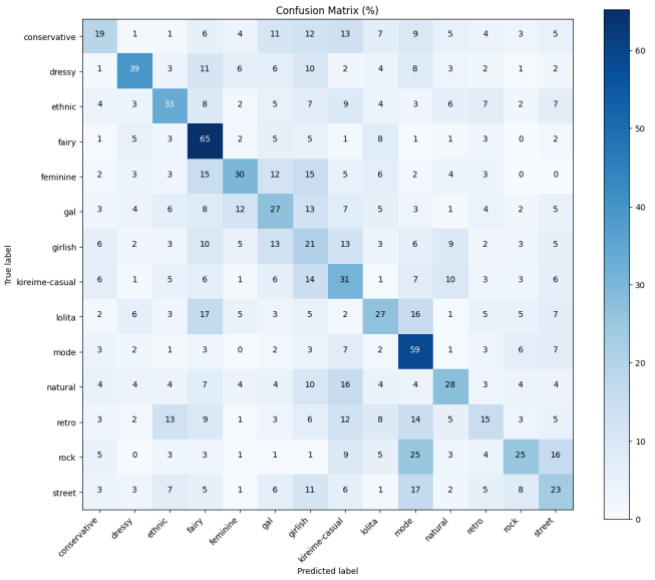

Figure 9: Confusion Matrix

2) ResNet 50

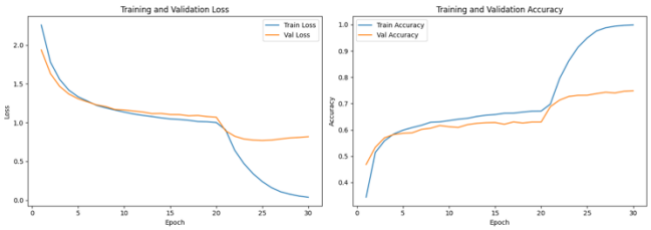

Figure 10: Loss and Accuracy Plot

Table 7: Train and Val Accuracy

	Train Accuracy	Val Accuracy
Average	94.92%	73.16%

Table 8: Train and Val Loss

	Train Loss	Val Loss
Average	0.2402	0.7702

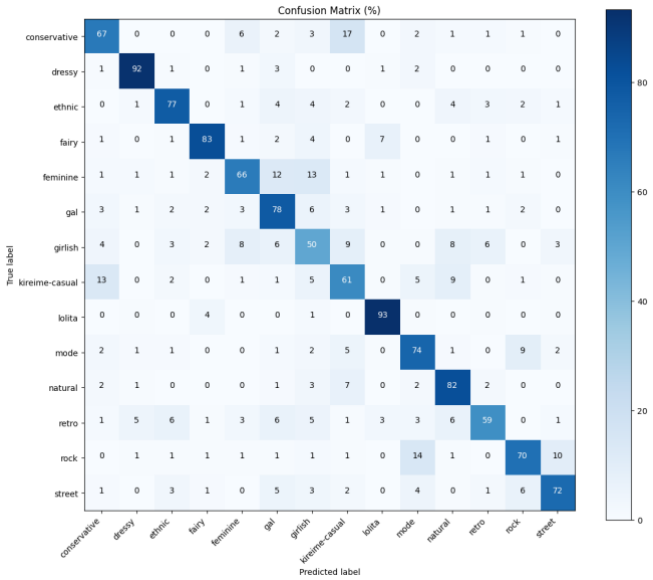

Figure 11: Confusion Matrix

3) *Discussions*

For the second pipeline, 5-fold cross-validation was not applied, so it shows only one accuracy and loss.
The Random Forest Classifier achieved an average validation accuracy of 31.79%, as shown in Table 6. This is significantly lower than expected for this task. From Fig 9, there are many misclassifications across multiple categories. Fairly label has 65% accuracy, while retro has 15% accuracy. And overall, the diagonal in the confusion matrix lacks a strong dark color. On the other hand, ResNet-50 achieved an average training accuracy of 94.92% and a validation accuracy of 73.16%. The accuracy looks low compared to the best model of Pipeline 1, but considering 14 labels, this accuracy is high enough. To conclude, since the Random Forest struggled with the complexity of the dataset, it could not achieve high accuracy. ResNet-50's convolutional layers were able to learn hierarchical features directly from the images, while Random Forest relied on predefined features, and it limited its performance as a result.

C. *Final Prediction*

Since the best model in Pipeline 1 was DenseNet-121, and the best model in Pipeline 2 was ResNet-50, these models were saved and tested by unseen data.

1) *Case A*

Predicted Clothing Type: long sleeve dress
Predicted Fashion Style: dressy

2) *Case B*

Predicted Clothing Type: short sleeve top
Predicted Fashion Style: natural

3) *Case C*

Predicted Clothing Type: long sleeve dress
Predicted Fashion Style: dressy

4) Discussions

For the first and second cases, the models successfully identified and classified the clothing type and fashion style. However, for the third case, the fashion style was predicted correctly, but the clothing type was misclassified.

I. CONCLUSIONS

In this study, a whole framework for classifying clothing types and predicting fashion styles of people in photographs by leveraging deep learning architectures, machine learning models, and then segmentation techniques was built. This time, two pipelines were designed to address this specific task, with Pipeline 1 focusing on clothing part classification with segmentation and Pipeline 2 on fashion style classification using models such as ResNet, DenseNet, and Random Forest. Models were Fine-tuned and the model with the best performance in each pipeline was selected as the best model in that pipeline.

The following are the results of the experiments.

1) Results for pipeline 1: DenseNet-121 achieved a validation accuracy of 82.10%, which is higher than the other two models (Deep Learning model and Machine Learning Classifier).

2) Results for Pipeline 2: ResNet-50 achieved a validation accuracy of 73.16%, slightly less accurate than Pipeline 1, but still highly effective in classifying Fashion style.

3) The final prediction as a system was also shown to be very high, thanks in part to the high accuracy of the models in Pipeline 1 and Pipeline 2.

In conclusion, this system proved to classify images into labels almost accurately and is able to be used for fashion trend analysis in real-life applications. As a future goal, it is necessary to increase the number of labels for classification and further train models.

REFERENCES

[1] S. Guercini, P. M. Bernal, and C. Prentice, “New marketing in fashion e-commerce,” *Journal of Global Fashion Marketing*, vol. 9, no. 1, pp. 1–8, Jan. 2018, doi: 10.1080/20932685.2018.1407018.

[2] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz and D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 7, pp. 3523-3542, 1 July 2022, doi: 10.1109/TPAMI.2021.3059968.

[3] T. Mallavarapu, L. Cranfill, E. H. Kim, R. M. Parizi, J. Morris, and J. Son, “A federated approach for fine-grained classification of fashion

[4] apparel,” *Machine Learning With Applications*, vol. 6, p. 100118, Jul. 2021, doi: 10.1016/j.mlwa.2021.100118.

[5] A. Chaudhary, S. Kolhe, and R. Kamal, “An improved random forest classifier for multi-class classification,” *Information Processing in Agriculture*, vol. 3, no. 4, pp. 215–222, Sep. 2016, doi: 10.1016/j.inpa.2016.08.002.

[6] F. He, T. Liu and D. Tao, "Why ResNet Works? Residuals Generalize," in IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 12, pp. 5349-5362, Dec. 2020, doi: 10.1109/TNNLS.2020.2966319.

[7] C. Zhang *et al.*, “ResNet or DenseNet? Introducing dense shortcuts to ResNet,” *arXiv.org*, Oct. 23, 2020. https://arxiv.org/abs/2010.12496

[8] Y. Ge, R. Zhang, L. Wu, X. Wang, X. Tang, and P. Luo, “DeepFashion2: a versatile benchmark for detection, pose estimation, segmentation and Re-Identification of clothing images,” *arXiv.org*, Jan. 23, 2019. https://arxiv.org/abs/1901.0797

[9] M. Takagi, E. Simo-Serra, S. Iizuka and H. Ishikawa, "What Makes a Style: Experimental Analysis of Fashion Prediction," 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 2017, pp. 2247-2253, doi: 10.1109/ICCVW.2017.263.