Convolutional Neural Networks for Fashion Classification

Muskan Singh*[1], Simone Shree Pathak[1], Anuska Basnet[1], Saishab Bhattarai[2]
[1]Department of Computer Science, Deerwalk Sifal School, Kathmandu, Nepal
[2]Department of Mathematics, Kathmandu University, Dhulikhel, Nepal
*Email: muskan.singh@sifal.deerwalk.edu.np

***Abstract*—Neural networks, inspired by the biological structure of the brain, have become a cornerstone in various machine learning tasks, including but not limited to those within the fashion domain. Convolutional Neural Networks (CNNs) are specialized neural networks designed for processing structured grid data - particularly images. Using these neural networks, The Virtual Wardrobe Fashion Advisor was created. Additionally programming languages like Python for algorithm implementation and Django,Html, Javascript and Cascading Style Sheets for web based application development were also used. The model was approached in the traditional CNN method to detect the type of clothes users prefer. Model detects the clothes with the help of selective gender and seasons; a setting in accordance to the user preference.**

Keywords: Convolutional Neural Networks, Fashion Advisor, Machine Learning, Data Collection.

I. Introduction

Over the past few years, technology has dramatically altered its relationship with fashion. Today, fashion designers are trying out new technologies that can create a difference in the fashion industry. Along with this tech which is a platform, machine learning, a part of artificial intelligence (AI), makes headlines as one of the effective ways of lifting the fashion field with all processes involved. Machine learning is the ability of computers to learn and get better with experience without being coded explicitly. Over time, it has experienced massive changes, leading to the growth of today's technology landscape and illustrating how robotics has impacted industries of such importance like healthcare or finance.

AI has its beginning in the fifties. But in the final stage of the 20th Century practically concentrated efforts were led to building plausible machine learning models. Arthur Samuel[1], the pioneer of this field, defined machine learning as the "study of computer systems that can learn without being explicitly programmed."

The machine learning development process has had its own set of milestones. Among them are the emergence of neural networks, the appearance of the statistical learning theory and the advancement of deep learning algorithms. Networks, along with mimicking the structure of the brains, and frequent applications within machine-learning tasks have been known in the fashion domain as well. In 1943, Warren McCulloch and Walter Pitts[2] were the first to succeed in doing the study of the neural networks, as they created a mathematical model that described the behavior of artificial neurons. It became the leading factor to the development of artificial neural networks for the next few years. Artificial Neural Networks (ANNs) represent computational models that consist of the interconnected nodes, or neurons, organized as layers. This results in having a single neuron that works as an input from one initial layer and that in turn serves as a result to the next layer. The idea of artificial neural networks (ANN) has led to the creation of many new architectural and training algorithmic revisions.

Within the field of neural networks, three prominent architectures stand out; the artificial neural networks (ANN), the convolutional neural networks (CNN), and the recurrent neural networks (RNN). Each architecture is represented by a group with its own attributes that underline the best method to particular challenges of data processing and is a part of a special-set.

Artificial Neural Networks (ANN) are the traditional neural networks or the perceptron as they were already called contain interconnected layers of neurons, so that every neuron in the one layer is connected to every neuron in the subsequent layer [3]. Convolutional Neural Networks (CNN) are a subclass of the neural networks which are made for the processing of structured grid data, mostly images. They use convolutional layers to model hierarchical features from input images, leading them to be able to rapidly depict the spatial links (LeCun and Juvan, 1998)[4]. Recurrent Neural Networks (RNN) have been designed for the sequential data processing tasks where the order of inputs is crucial. The same as feedforward networks, RNNs have a memory that allows information to be accessible at any moment of the flow spontaneously which, in turn, gives them the ability to capture temporal dependencies in sequential data [4]

In the context of fashion dress classification, these neural network architectures play a crucial role in leveraging the vast amounts of fashion data available to automate and enhance various processes, ranging from product categorization to trend analysis. In the generation where purchasers are faced with the assignment of navigating through an overwhelming pool of clothing objects to discover the proper outfit, such technologies can help tremendously to ease the people with a fashion advisory system.

The primary objective of this research is to streamline

and enhance the fashion decision-making procedure by implementing CNN algorithm to an advanced fashion advisory system. The main context of this paper are as follows, Section II provides Literature Review of previous works, Section III Methodology describes the mathematical use of CNN model, Section IV Results gives the summary of implementation of the CNN model and its results, Section V Discussion summarizes and gives the overview on the project and Section VI Conclusion and Future Works concludes the project with the future implications.

II. Literature Review

In the world of technology and advancement, machine learning and convolutional neural networks had been developed since the 1990s by Yann LeCun[3], along the colleagues. It got popular in the 2010s and many researchers had developed projects and given their knowledge and understanding about this topic.

A. Hierarchical Convolutional Neural Networks

Yian Seo , Kyung-shik Shin [5] had introduced the concept of the Hierarchical Convolutional Neural Network in 2018 A.D . The concept of Hierarchical Convolutional Network was developed in the beginning of the CNN by other researchers. Although it was not utilized by the researchers who worked on the detection of the fashion clothes and apparel.

The fashion industry is a competitive business with a greater amount of profit. The CNN models had been used in this field through different organizations to decrease the workload and make more profit. Unfortunately those models were not perfect, as the result Yian Seo and Kyung-shik Shin developed the first hierarchical convoluted network. It had a high frequency with minimal error. It can work on a huge set of data at a particular time. The apparel classification had been major for this research. The most common problems during image detection are about apparel or objects. The CNN model assumes the pants and top to be the same thing and gives inaccurate results to the users. It also faces the problem of determining whether the clothes are worn by a human figure or not. So for these types of issues they have done research and gained 91.7 percent of accuracy. The hierarchical model reads and goes through each image and then classifies it into definite categories. They all are interconnected and it gives a perfect result. Each apparel can be classified and the object can be easily detected. In this methodology the user gets better experience and the business grows rapidly. The research made a significant impact on the detection of the clothes, apparel in the fashion field with accurate data and results.

B. Convolutional Neural Network Clothes Classification

Due to high accuracy and cost many businesses had not applied Convolutional Neural Networks for the clothes classification. The researchers of Gdansk University of Technology [6] have addressed this problem and use CNN to solve the problem. They had managed to get the basic information and category classification of the clothing through the input served to the model. The training and the test dataset had been a small number initially but as they got accuracy they increased the dataset to a higher number. Many ways are present for object detection, among them SDD300 was chosen. After almost 200 epochs they were able to achieve 42 percent accuracy. On increasing the training accuracy their test accuracy had a great improvement. There were many CNN architectures and layers applied in order to achieve the maximum accuracy. It contributed to a better and convenient search engine for the fashion and clothes market and business.

Alexander Schindler from the Austrian Institute of Technology had also researched on this topic of fashion and apparel classification using the deep learning neural networks [7]. In the research paper the major problem had been addressed as the inaccuracy or improper results on the classification of the apparel and the clothing materials. Due to large textures, shapes and colors the machines inability to correctly identify and give the results using the traditional model of neural networks had been clearly mentioned. Thus a smaller scale dataset with best performance which is certainly increased by the higher number of dataset for better classification.

C. Clothing Object Detection

A real-time clothing classification system facilitates automated fashion stylists, outfit recommendation, discovering similar fashion pieces, surveillance context, automatic annotation of images with tags or descriptions, context-aided people identification, occupation recognition, and improvement in information retrieval from various areas such as social media and consumer sites [8]

Hara, Kota et al. also researched clothing object detection, and discovered a new way of classifying fashion apparel using detection based spotters in the form of bounding boxes. These boxes worked more efficiently compared to other methods as it is faster and doesn't require multiple instances like semantic segmentation. In their research, they incorporated contextual information from body poses in order to improve the detection performance. By conducting multiple experiments, they drew a conclusion that the method they discovered, Pose-dependent Object Detection,is more effective than semantic categorization [9].

One of the earliest constructions of a framework for robust and extremely rapid visual detection was done by Pual Viola and Micheal J.Jones. It classified images based on the value of simple features. This architecture has been widely used to implement sliding window detectors for faces [10]

III. Methodology

The Virtual Wardrobe Fashion advisor had been created through the use of the programming languages like Python for algorithm implementation and Django, HTML, JavaScript and Cascading Style Sheets(CSS) for web based application development. The dataset had been created manually. The data ranged till fifteen thousand. The dataset of cloth was

categorized into male, female, and into the four seasons, namely winter, summer, spring and autumn. It also has the category of formal and informal dress. Before serving the data to the CNN model there were several preprocessing steps applied which involved resizing, and scaling the pixels of the images.

A. Data Collection

1) Gender: The dataset contains about 15000 images and among them 7000 were labeled as male and 8000 were labeled as female manually, as the ground truth annotations. It was done in order to check the predictability of the model to classify between male and females.

2) Seasons: At the beginning a dataset contains 40 thousand images. It has the categories of the four seasons with male, female, formal, informal and occasional. The dataset contained its ground truth annotations for each of the categories. Winter season consisted of the largest portion in our dataset. The smallest dataset was of the autumn male clothes for occasional. The complete dataset was not utilized for the training of the CNN model. About 10 percent of the data was utilized and the dataset containing less than 50 images were not taken into account. The dataset which contained an efficient amount of data was selected for the further step of the training dataset. The subset data only contained about 15000 images.

B. The Mathematics of Convolutional Neural Network

1) Matrix Formation: The data served to the CNN model is represented in the form of a matrix. Let us assume a three by three matrix of X denoted as:

$$X = \begin{pmatrix} 10 & 20 & 30 \\ 15 & 25 & 35 \\ 5 & 10 & 15 \end{pmatrix}$$

In the matrix, each value represents specific weighting parameters in the CNN model. In CNN these parameters are required for the training of the model and to specify the features. These parameters are required in order to classify the image and predict which category this image belongs to. It goes through the binary classification in order for the detection.

2) Convolutional Layer: In the convolutional layer the images are transformed into matrix formation. It has multiple layers and can give output more accurately than other neuron networks. The convolutional is also called local connectivity as CNN connects the neurons to the local input. For example if 32 x 32 neurons are required in the next layer and 5 x 5 neurons are connected in the previous layer. It can be visualized through figure 1.

Now, let's introduce a simple 2x2 filter F for our convolution operation:

$$F = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

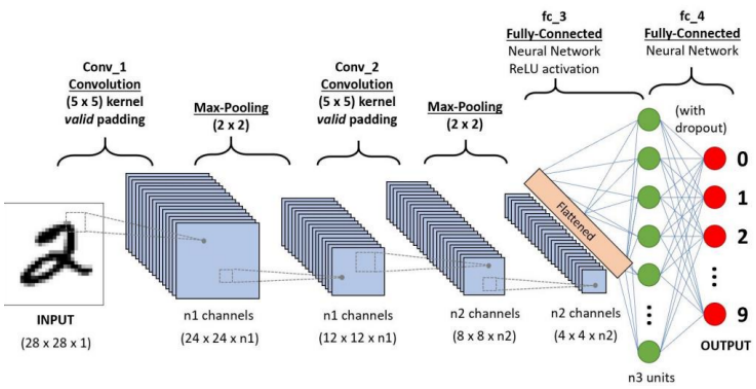

Fig. 1. Fully Connected Layer of CNN [12]

The convolution operation involves sliding this filter over our image matrix X and performing element-wise multiplication, producing a feature map Y:

$$Y = X*F = \begin{pmatrix} 0 & 20 \\ 5 & 10 \end{pmatrix}$$

The basic process of the training of the CNN model can be clearly visualized through the following diagram of figure 2. An image is passed through different layers which gives different output as the detection of edge, sharpened image, background, identity and blurred image. The augmented images represent the different features of the served image to the CNN model for classification.

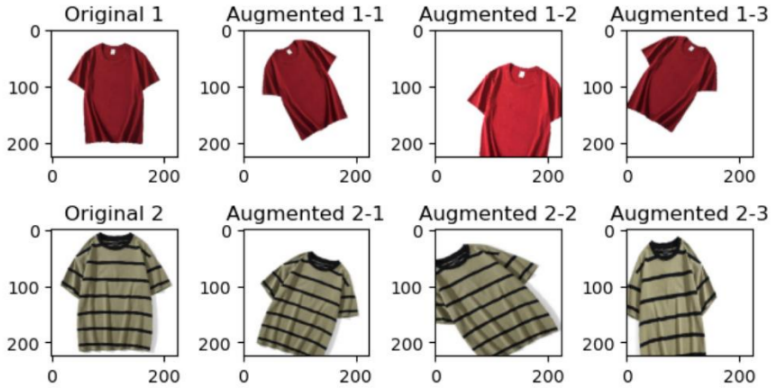

Fig. 2. Effects of different Convolution

3) Non-linearity: The non -linearity indicates an optimized output through the CNN. It consists of ReLU Activation, Sigmoid function and tan hyperbolic function. The ReLU Function can be defined as:

$$f(x) = \max(0, x)$$

In other words it only passes the positive value of the matrix and if there is a negative value then it converts them to zero. The ReLU layer can be represented through the following mathematical form. If the input neuron is z, the input to the

next layer is the sum of the inputs of the neuron of the previous layer. Its equation form is:

$$z = \sum_{i=1}^{n} w_i x_i + \epsilon$$

The w_i represents the weights and x_i are the inputs and b is the bias for the neuron of the inputs in this layer.

The sigmoid function similarly takes any real number as input and provides the output as 0 or 1. The hyperbolic function of tangent minimizes the output to range [-1,1]. The following graph represents the tan hyperbolic function, sigmoid and the ReLU activation of the Convolutional network in the object detection of the CNN model of fashion and clothes.

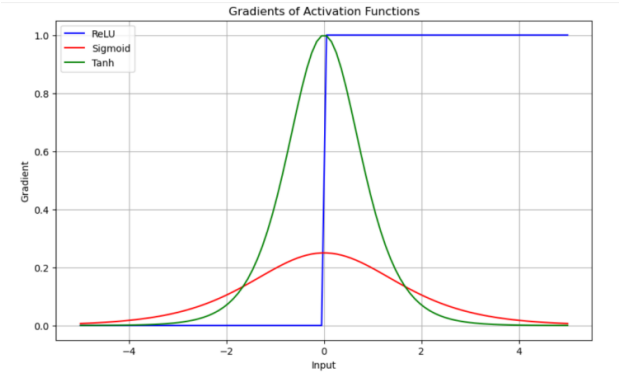

Fig. 3. Uses of Non Linearity in Object Detection

4) Padding and Max-Pooling: One of the major drawbacks of the Convolution Network is that it does not include the objects at the border if we don't use zero padding. Thus we have used zero padding for the object detection of the CNN model.

This reduces the size of the image and shrinks it after it passes through each layer. Similarly the kernel touches the middle of the image more than the corners which reduces its accuracy. In order to solve this problem Padding was introduced [20]. In padding when a image of m x m matrix is utilized with f x f filter with padding p then the size of the output image is:

$$\textit{Output size} = (m + 2p - f + 1) \times (m + 2p - f + 1)$$

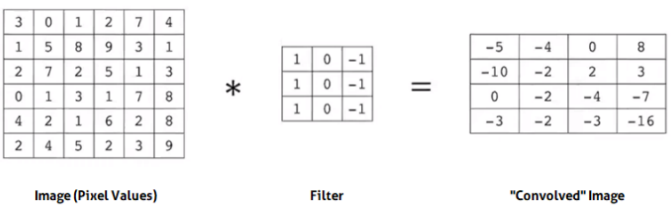

Fig. 4. Convolved Image using Padding

Figure 4, shows the stride of the filter that hovers along the image pixel. The stride simply can be defined as the filter over the input matrix during convolution. After the strides is done on the input matrix then the output image dimension can be described through the following equation:

$$\textit{Output size} = \left(\frac{n + 2p - f + 1}{s} + 1\right) \times \left(\frac{n + 2p - f + 1}{s} + 1\right)$$

where:

- p is the padding, which adds a border of p pixels around the image.
- $f \times f$ is the filter size, which determines the size of the kernel used for convolution.
- s is the stride, which specifies the number of pixels that the filter moves after each step.
- $n \times n$ is the input size, which represents the dimensions of the original image.

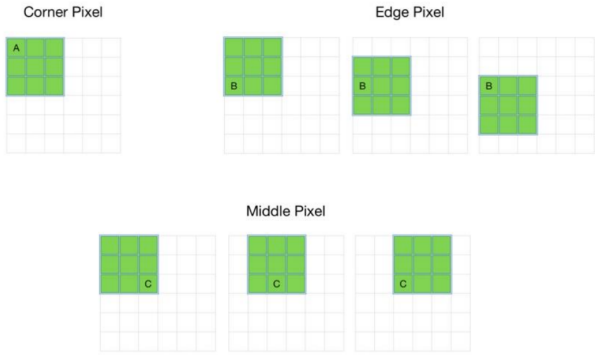

Fig. 5. Padding and Sliding Filter over the image[14]

Pooling can be defined as one of the important parts of CNN as it reduces the parameters of the model and simplifies it. The maximum pooling method contributes to this work. In this process a non overlapping filter goes over the image. It only selects that matrix which has the maximum value and discards the rest[15]. It controls the overfitting of the input matrix / image. The output size of the matrix can be determined through following equation:

$$\text{Output size} = \frac{n - f}{s} + 1$$

where:

- n is the input size,
- f is the filter size,
- s is the stride.

5) Fully Activated Layer: In the fully activated and connected layer each neuron is connected to the previous layer and each neuron of the previous layer is connected to the one before them. It gives output through sending the input neuron in softmax activation. In softmax activation the exponential of the neuron is divided by the sum of all the neurons exponential. The output of this activation predicts the object of the certain category [11].

$$p_i = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}}$$

The above equation represents the softmax activation of the fully connected neurons. The image of the fully connected neurons is in diagram figure 1.

C. Model Development

During the prediction one of the essential parts is how the model will be able to detect the gender and seasons. The background detection was a crucial part of the research. We have mostly preferred those images which had white background and studio lights so that it would be easier and more convenient for the model to be trained. The process of recognition of gender through the CNN model is presented through the flowchart of fig 6.

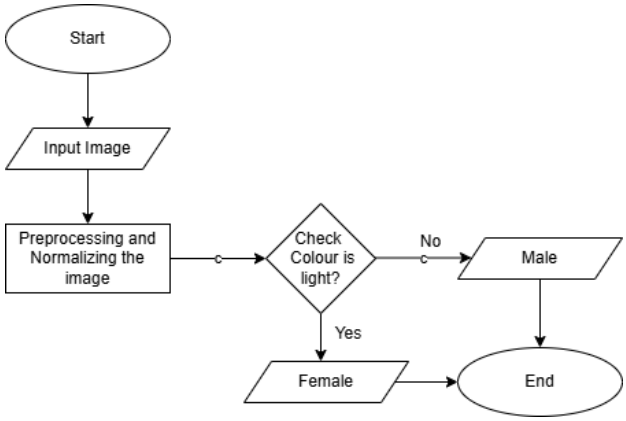

Fig. 6. Flowchart of Gender Detection

In the gender detection first the input image goes through processing and normalizing the image. Then the color is checked whether it is light or not. As it's a basic model so it does not go through vivid properties for gender classification. If the color is light it considers the cloth to be of female and if the color is dark then male is considered. After the gender detection it goes through the following process shown in the flowchart as shown in Fig 6. It depicts the season of the input image through it.

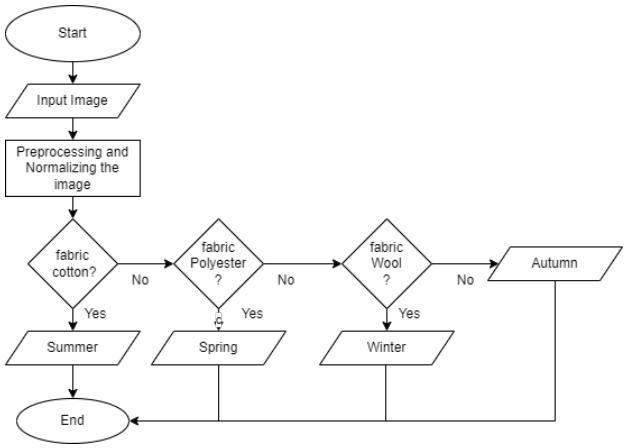

Fig. 7. Flowchart of Season Detection

After the gender detection the image is used for the season detection. In that as shown in the Fig 7 the conditions are applied. According to the satisfaction with the condition the results are displayed. The CNN model predicts the season for the input image. Now the occasion is left which is also vividly presented through the following flowchart.

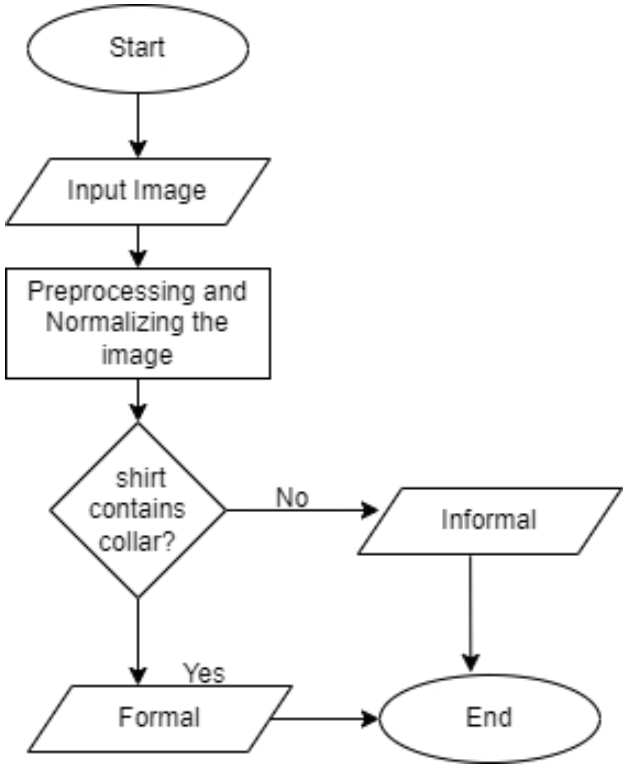

Fig. 8. Flowchart of Occasion Detection

When the CNN model has classified among the gender and season then it checks for occasion whether its formal or informal. It checks whether the image contains a shirt or collar and if there is a shirt or collar then it predicts the image to be formal or informal. This mechanism is also visualized using the flowchart diagram.

IV. Results

After the formation of the CNN model, we took in account about 200 images for the test dataset and 500 images for the training dataset. At the beginning due to imbalanced dataset, model overfitting, limited data size and data quality there were several negative impacts on the accuracy of the model. It had about 67.11% of accuracy on the dataset. It was not practically applicable model due to loss of accuracy.

A. Analysis of the Challenges and Limitations Graph

The data format of figure 9 is reflected in tabular form in Table I, the limited data size had the lowest accuracy about 60.75% on the test data set and the data quality had the minimum accuracy in the training dataset. Similarly the maximum accuracy was received in the limited data size in the training dataset. It gave us the idea on which portions we need

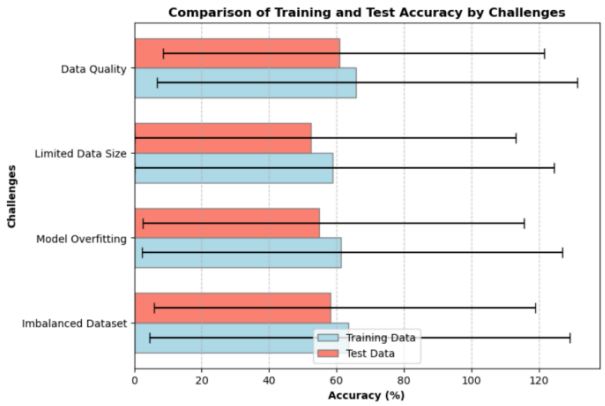

Fig. 9. Challenges and Limitations Graph

to improve in order to overcome the challenges and limitations to make the project more efficient. It also represented the stability of the model with the trend of accuracy when such drawbacks were not removed through the dataset. It was the initial dataset that manually contained such issues and which decreased the accuracy of the model.

As the accuracy was not as expected the model was remodified several times, for each modification the size of data was changed. We had a different accuracy rate for the prediction of the clothing images through the CNN model. The most accurate results were shown by Model C which had the highest accuracy of 89% on the detection of the images.

TABLE I
IMPACT OF VARIOUS DATA CHALLENGES ON MODEL ACCURACY

Challenges/Limitations	**Training Accuracy (%)**	**Test Accuracy (%)**
Imbalanced Dataset	63.50	61.25
Model Overfitting	65.75	62.50
Limited Data Size	64.00	60.75
Data Quality	62.25	59.50

B. Analysis of Comparison of Different Models

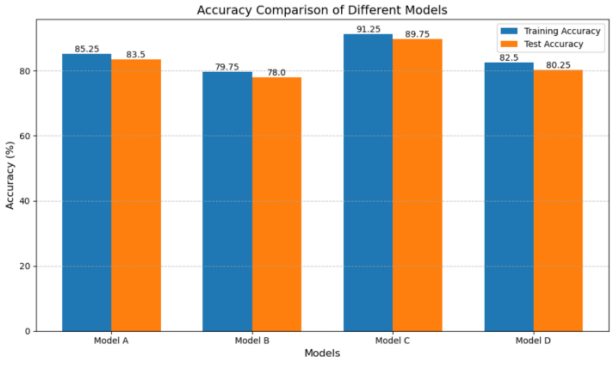

Fig. 10. Comparision of the different models

From Figure 10 and Table II, we determined that the model C had the maximum accuracy of 91.25% and was highly efficient. The Model B had the lowest accuracy of 79.75% and was not as efficient as other models. The challenges and limitations were overcome in each model and the dataset was modified. The best version of the model was C which was finalized for the final product.

TABLE II
COMPARISION OF DIFFERENT MODELS ON BASIS OF VALIDATION AND TRAINING DATA RATIO

Model	Initial Data	Data Type	Validation Data%	Training Data %
A	40,000	Uncategorized	83.50	77.05
B	30,000	Uncategorized	77.1	92.86
C	15,000	Uncategorized	89.75	91.25
D	10,000	Uncategorized	80.25	93

Although the training data has the maximum accuracy in comparison to the test data. The loss of training data and test data was significant at the beginning. Although the prediction for the training data improved in the model improvement. The model C had almost 90% accuracy for the training data and 89% accuracy for the test data. The training and validation accuracy for model C with the training and validation loss is presented in the figure 11.

C. Analysis of CNN Model Loss Metrics

The Fig 11 and 12, illustrates the training and validation accuracy of the final Model C, which achieved the highest accuracy in object detection. From the graph, it is evident that the training accuracy consistently surpasses the validation accuracy. While the validation accuracy remains stable across epochs, the training accuracy demonstrates noticeable variability. Additionally, the training loss is initially higher than the validation loss during the early epochs. Over time, the validation loss exhibits a steady trend, whereas the training loss fluctuates across epochs. To gain further insights into the patterns and trends of prediction, the training and validation loss of other models, namely Model A, B, and C, were also analyzed and plotted in a line graph. A table summarizing the data for training and validation loss and accuracy over 20 epochs, among the total 155 epochs completed during the training process, provides a comparative overview of the models' performance.

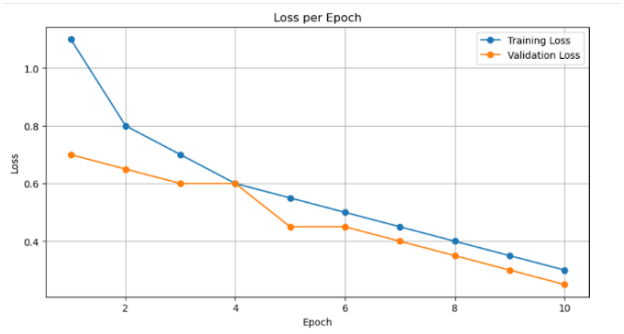

Fig. 11. Loss Per Epoch: Model C

The training and validation loss of other models as Model A, B and C were also plotted in the line graph to learn the

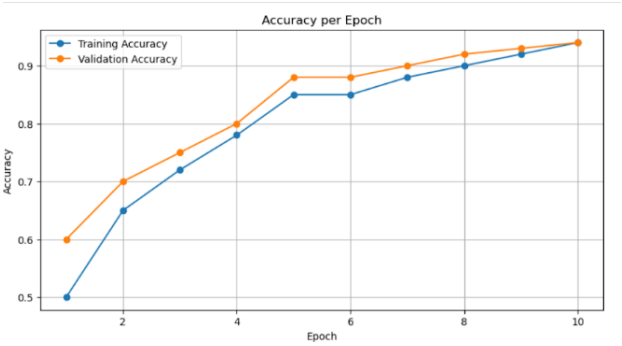

Fig. 12. Accuracy per Epoch: Model C

patterns and trend of prediction for those models. Table III shows their data of average training and validation loss and accuracy over 10 epochs among the 155 that the model went through during the training process.

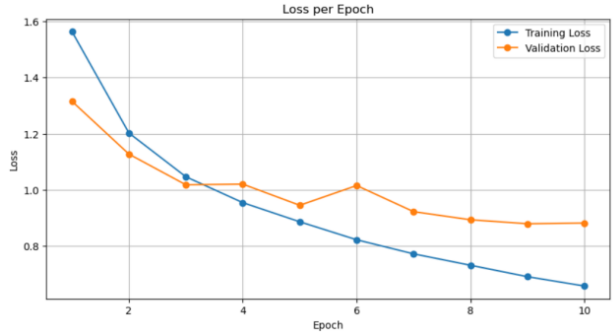

Fig. 13. Loss Per Epoch: Model B

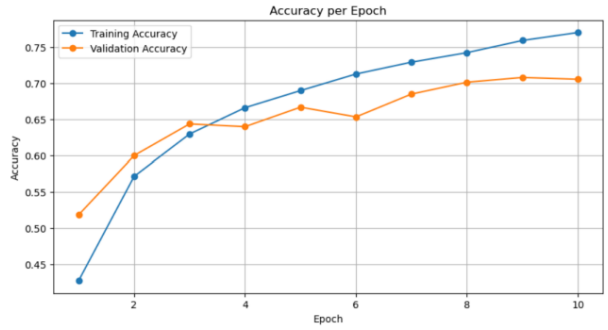

Fig. 14. Accuracy Per Epoch: Model B

The graphs represented in Fig 12 - 17, represents the evolution of training and validation accuracy as well as training and validation loss across 155 epochs, offering a comprehensive overview of the model's learning dynamics. The insights gained from these visualizations further inform potential adjustments and refinements to enhance the model's overall predictive capabilities. The following were the results shown by the model after the completion of the steps and accuracy improvement. We can also observe that the model can accurately predict the male female and seasons as well as

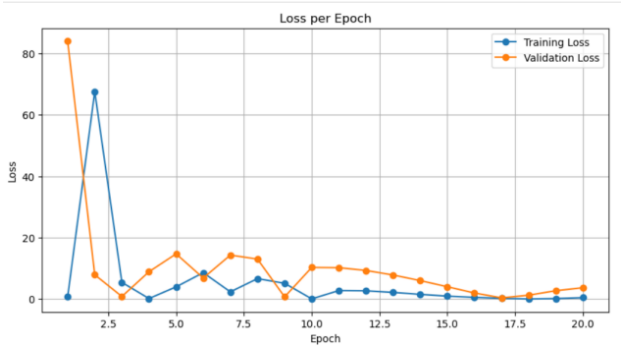

Fig. 15. Loss Per Epoch: Model A

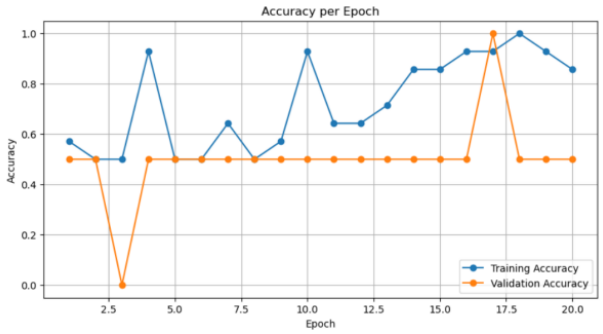

Fig. 16. Accuracy Per Epoch: Model A

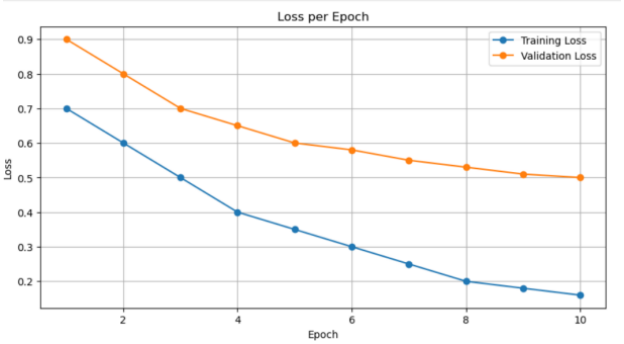

Fig. 17. Loss Per Epoch: Model D

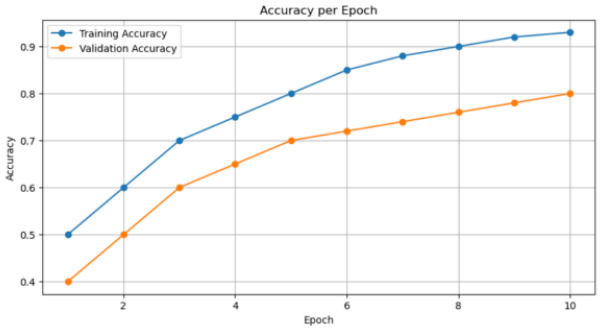

Fig. 18. Accuracy Per Epoch: Model D

the occasion of the input image. The accessories are mostly present in the female outfit. While designing the model the accessories were also kept as the base logic for the model to identify between male and female. The vibrant colors are mostly present in the female outfits. The fabric of the clothes are checked in order to detect the season. On the basis of shirt and collar formal and informal occasion is detected.

TABLE III
MODEL EVALUATION METRICS (SWAPPED LABELS)

	Model A	Model B	Model C	Model D
Epochs	10	10	10	10
Train Loss	0.93258	0.93255	0.575	0.364
Train Accuracy	0.67	0.709	0.799	0.783
Val Loss	1.00169	1.00169	0.475	0.632
Val Accuracy	0.65235	0.68303	0.83	0.665

V. CONCLUSION

This study demonstrated how deep learning models like CNN may be used in real-world scenarios like virtual wardrobe systems and clothing recognition. Model C was the most successful of the models that were assessed; it had the lowest validation loss (0.475) and the best validation accuracy (83%) showing excellent generalization. Furthermore, Model C effectively identified characteristics such as occasion, gender, and season from input photographs with a training accuracy of 91.25%. It was integrated into a virtual wardrobe application and trained on a dataset of over 15,000 photos spanning 155 epochs, allowing users to get tailored fashion recommendations depending on their tastes. The system solved typical fashion-related problems by offering precise and user-centric recommendations using CNN's conventional techniques and ground truth annotations.

By growing the dataset, adding a variety of clothing designs, and applying cutting-edge techniques like Generative Adversarial Networks (GANs) for increased accuracy and customisation, future research can aim to improve the model's capabilities. In order to guarantee improved accessibility and usefulness, future advancements will additionally concentrate on real-time picture processing for mobile devices. These developments will expand the capabilities and scope of virtual wardrobe systems, increasing their resilience, interactivity, and influence in the rapidly changing digital fashion market.

REFERENCES

[1] G. Wiederhold and J. McCarthy, "Arthur samuel: Pioneer in machine learning," *IBM Journal of Research and Development*, vol. 36, no. 3, pp. 329–331, 1992.

[2] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Journal of Symbolic Logic*, vol. 9, no. 2, pp. 49–50, 1943.

[3] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proceedings of 2010 IEEE international symposium on circuits and systems*. IEEE, 2010, pp. 253–256.

[4] Y. Ma, Y. Ding, X. Yang, L. Liao, W. K. Wong, and T.-S. Chua, "Knowledge enhanced neural fashion trend forecasting," in *Proceedings of the 2020 international conference on multimedia retrieval*, 2020, pp. 82–90.

[5] Y. Seo and K. Shin, "Hierarchical convolutional neural networks for fashion image classification," *Expert Systems with Applications*, 2018.

[6] J. Cychnerski, A. Brzeski, A. Boguszewski, M. Marmolowski, and M. Trojanowicz, "Clothes detection and classification using convolutional neural networks," in *2017 22nd IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, 2017.

[7] A. Schindler, T. Lidy, S. Karner, and M. Hecker, "Fashion and apparel classification using convolutional neural networks," *arXiv preprint arXiv:1811.04374*, 2018.

[8] M. Shajini and A. Ramanan, "An improved landmark-driven and spatial–channel attentive convolutional neural network for fashion clothes classification," *The Visual Computer*, vol. 37, no. 6, pp. 1517–1526, 2021.

[9] K. Hara *et al.*, "Fashion apparel detection: The role of deep convolutional neural network and pose-dependent priors," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1–9.

[10] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[11] S. S. Basha *et al.*, "Impact of fully connected layers on performance of convolutional neural networks for image classification," *Neurocomputing*, vol. 378, pp. 112–119, 2020.