Object Recognition and Assistance System for Visually Impaired Shoppers

Shanelle Tennekoon[1], Nimsiri Abhayasinghe[1], Nushara Wedasingha[1]

[1]Department of Electrical and Electronic Engineering, Sri Lanka Institute of Information Technology, New Kandy Rd, Malabe, Sri Lanka

ABSTRACT

Shopping is indeed effortless for many individuals. However, it could certainly be a struggle and chaotic experience for the visually impaired. Visual impairment causes many societal stigma and inconvenience to visually impaired individuals. Although shopping may sound extremely easy, this is a crucial social activity for many visually impaired (VI) individuals. Visually impaired (VI) shoppers always require assistance when shopping for product identification purposes. This may lead to greater inconvenience as delays, lack of information and product familiarity of shop assistants may occur. Therefore, allowing visually impaired shoppers to independently perform shopping regardless of size and position of the shopping mall is essential. This encourages them to participate in enhanced social activities and perform their daily chores in independence. Although many products have been developed to assist visually impaired shoppers at shopping malls, due to their drawbacks, some of these have seem to undergo failures in producing accurate information to the visually impaired shopper for object identification and caused inconvenience. This project proposes a feasible solution for visually impaired shoppers to perform their shopping at ease and independently. Object recognition has been made possible in order to identify garment items while shopping with no assistance of another individual. The Convolutional Neural Network (CNN) has been used to obtain a sufficiently good accuracy and precision with a validation accuracy of 90%. Some of the novel techniques such as Ensemble Modelling has also been performed in order to reduce any generalization errors of the prediction and achieve a greater accuracy while overcoming all of the drawbacks of the currently existing products in the market. The overall product is proposed to attain maximum consumer population of visually impaired shoppers with satisfaction, reliability, and low cost.

KEYWORDS: *Convolutional Neural Networks (CNN), Ensemble Model, Visually Impaired (VI), Minimal Assistance while Shopping, Principal Component Analysis (PCA)*

1 INTRODUCTION

According to the collection of the most recent information about eye health that was published in 2021 by the official launch of the International Agency for the Prevention of Blindness's Vision Atlas, it reveals that there exists 43 million individuals globally living with blindness and another 295 million people living with moderate-to-severe visual impairment (Orbis, 2022). Furthermore, the According to the World Health Organization (WHO), uncorrected myopia and presbyopia alone are projected to have yearly global costs of lost productivity linked with visual impairment of US$ 244 billion and US$ 25.4 billion, respectively (Who.int, 2022). Therefore, it can be stated that visual impairment affects many major local and global activities.

Shopping may seem effortless for many individuals. However, with the development of the modern world, the complexity of modern supermarkets has seemingly increased. Gigantic shopping malls have claimed to possess stocks with an average of 45,000 products and a median store size of 4,529 square meters (Fmi.org, 2022). Shopping also involves a process of navigating through aisles, locating shelves with desired products, identifying the necessary products or items, and many more.

https://doi.org/10.54389/RRFZ2311

Individuals with visual impairments often rely on another individual for assistance when shopping. This compromises their independence at large. It is also a struggle as it is necessary to inform the store in order to reserve an individual for assistance beforehand or have their own means of assistance. This results in experiencing delays, inconvenience, irritation, and more. Assistants at supermarkets may also not possess adequate knowledge to read out ingredients, may be unfamiliar with aisles and products, and other inconveniences. Due to these difficulties faced by Visually Impaired (VI) shoppers, they may abandon shopping and choose distant alternatives. Although many apps have been especially designed and developed for VI shoppers, this reduces the ability of spontaneous shopping and reduces personal independence of VI shoppers. Therefore, means for VI individuals to shop by themselves independently and reliably is necessary. Many smart solutions such as assistive aids have been implemented to ease the life of VI shoppers and provide convenience to perform their daily activities. However, due to some of their drawbacks and failures, means to assist accurately, comfortably, and reliably prevails.

1.1 Problem Statement

Visually impaired citizens often deal with many challenges in their daily lifestyle and routine. The limited accessibility to many activities and information, societal stigma, requiring assistance of another individual to perform their daily chores can often be a struggle. The problem aimed at is the challenge faced by visually impaired individuals at shopping malls. The need for assistance is mandatory in this situation. However, this too is a struggle. The ability for VI individuals to shop with more convenience, safety, and independence is aimed through this project. It is essential that good quality of life is provided for VI individuals just as much as people with no impairments. It is also important that such solutions are implemented in order to encourage individuals with visual impairments to engage in more social activities. Therefore, affordable, and reliable means to assist individuals suffering from visual impairment has been implemented in this project while overcoming the drawbacks and research gaps of the products and research that have been conducted thus far.

For research purposes, a few YouTubers with visual impairments were followed and gathered information with regard to their shopping experiences. VI YouTuber, 'Katy's Eyes' mentions that shopping requires pre reservation of assistance at the shopping mall where she shops. The assistant is required to navigate the VI shopper along the mall to isles of products, avoiding obstacles, and help identify the necessary products. An instance of navigation along with object identification was captured as follows,

Figure 1. Object Recognition with Assistance while Shopping (YouTube, 2022)

2 DESIGN METHODOLOGY

Firstly, it was necessary to build and train the Convolutional Neural Network Model that comprised of several hidden layers along with the input dataset that was used, namely, the Fashion MNIST dataset comprising of 60,000 training images and 10,000 testing images. Thereafter, the results were visualized, and the Ensemble Model was implemented to prevent any generalization errors of the prediction. Principal Component Analysis (PCA) along with a qualitative analysis was also performed in order to compare the implemented model with all other existing models utilized for object identification.

2.1 Convolutional Neural Network Model

The CNN Algorithm is one of the primary algorithms that can be considered for object recognition. Before reaching the output layer, the data travels through many hidden layers after entering the CNN through the input layer. The output of the network is compared to the actual labels in terms of loss and error. One of the several backpropagation techniques is used to update the trainable weights, and the partial derivatives of this loss with respect to the trainable weights are then calculated. In many CNN models, most of the hidden layers are often common and follow a pattern that can be identified.

The model implemented in this research can be summarized as follows,

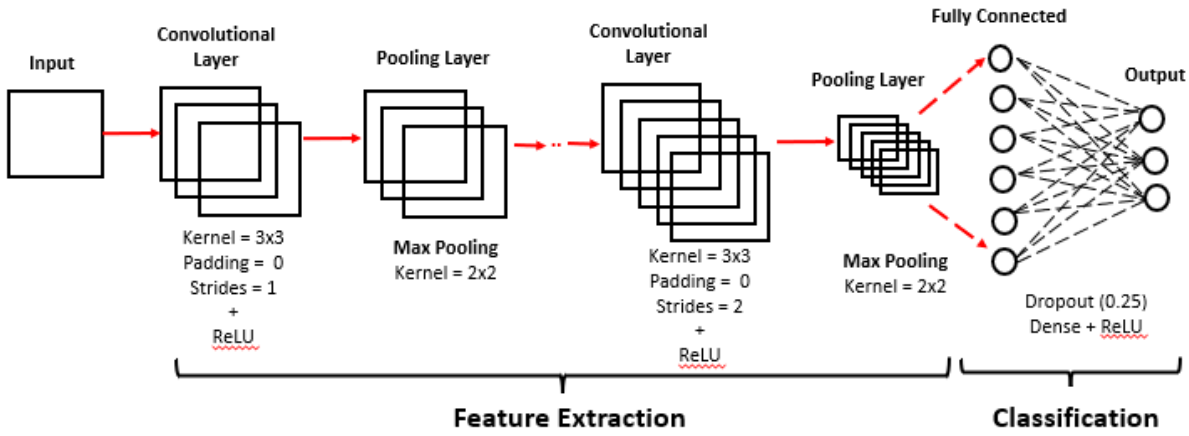

Figure 2. Architecture of Implemented Model

1. Layer Function: Basic function of transformation (convolutional/fully connected layer)
   a) Fully Connected: Functions are linear between the input and the output (For i input nodes and j output nodes, the trainable weights are wij and bj)
   b) Convolutional Layers: Applied to the input feature maps that are 2D (and 3D). The trainable weights are made up of a 2D (or 3D) kernel/filter that iterates through the input feature map and creates dot products using the overlapped areas of the input feature map. The three variables that define a convolutional layer are as follows:
      - **Kernel Size K:** Filter size of 2x2 and 3x3
      - **Stride Length S:** Specifies the amount of kernel sliding that must occur before the dot product may produce the output pixel – used stride 1 and 2
      - **Padding P:** not used in this implementation
   c) Transposed Convolutional (DeConvolutional) Layer: Typically utilized to improve the output feature map size (Upsampling). With the transposed convolutional layer, the input feature map is modified. If the provided stride and padding to the output along with the convolutional kernel of the required size is applied, it will result in the input.

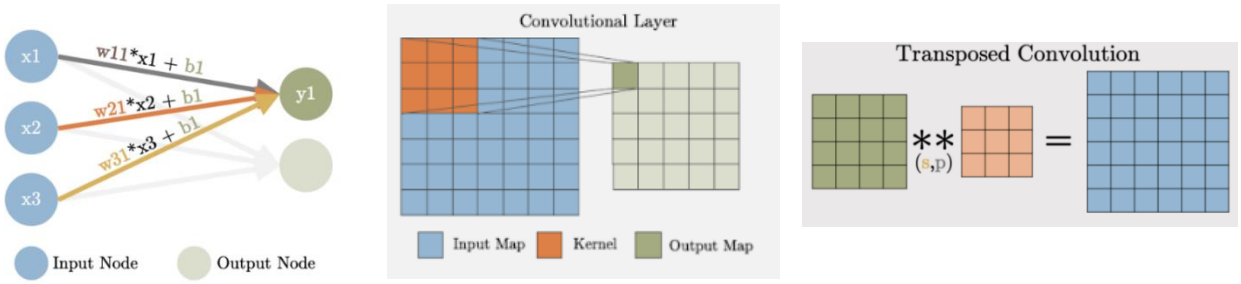

Figure 3. Fully connected layer, Convolutional Layer, and Transposed Convolutional Layer (Anwar, 2020)

2. Pooling: This layer cannot be trained and is used to modify the feature map's size.
   a) Max/Average Pooling: Was used to reduce the input layer's spatial size by choosing the field's maximum or average value, as indicated by the kernel.
   b) UnPooling: this is a non-trainable layer that increases the input layer's spatial area by positioning the input pixel at a specific index inside the kernel's designated output field.

3. Normalization: To prevent the unbounded activation from raising the output layer values excessively, it is typically employed just before the activation functions.
   a) Local Response Normalization (LRN): An untrainable layer square-normalizes the pixel values in a feature map of a particular neighborhood.

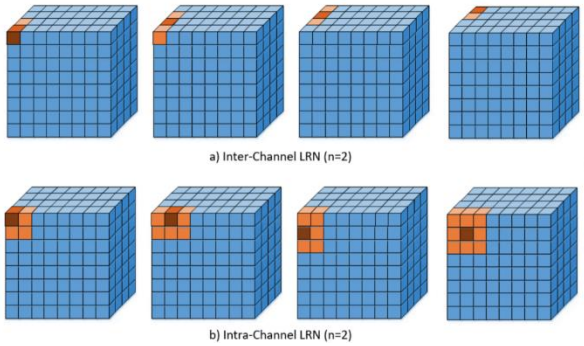

Figure 4. Local Response Normalization (LRN) (Anwar, 2020)

$$b_{x,y}^{k} = a_{x,y}^{k} / (k + \alpha \sum_{0}^{\min\left(W,\ x+\frac{n}{2}\right)} \sum_{0}^{\min\left(H,y+\frac{n}{2}\right)} (a_{i,j}^{k})^{2})^{\beta} \quad (1)$$

$$b_{x,y}^{i} = a_{x,y}^{i} / (k + \alpha \sum_{j=\max(0,i-\frac{n}{2})}^{\min\left(N-1,i+\frac{n}{2}\right)} (a_{x,y}^{j})^{2})^{\beta^{-n/2)}} \quad (2)$$

Equation (1) gives the Intra-channel LRN whereas equation (2) gives the Inter-channel LRN when performing Local Response Normalization.

   b) Batch Normalization: A method of normalizing data that is trainable by learning the scale and shift variables.

4. Activation: So that CNN can efficiently map non-linear complex mapping, introduce non-linearity.
   a) Non-parametric/Static Functions: Linear ReLU
   b) Parametric Functions: tanh, ELU, Leaky ReLU, sigmoid
   c) Bounded Functions: sigmoid, tanh

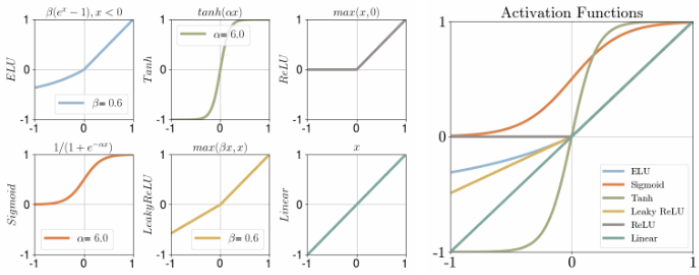

Figure 5. Activation Functions in CNN (Anwar, 2020)

ReLU activation was used in this implementation.

2.2 Implementation of the Model

The CNN Model was implemented in order to identify objects in real-time by following the simple procedure and methodology where firstly, the Fashion MNIST dataset was utilized, and the CNN Model was implemented. Next, Ensemble Modelling was performed and thereafter was tested on a real-time video that was obtained at a shopping mall in a VI shopper's point of view.

Choosing appropriate Dataset

60,000 training images and 10,000 test images of clothing from 10 classes, including t-shirts, tops, pants, pullovers, dresses, coats, sandals, shirts, sneakers, bags, and ankle boots, make up the Fashion-MNIST dataset. Each grayscale image has a uniform dimension of 28x28 (784 total pixels). The visual representation of the data (each class containing in three-rows), is depicted in the figure below,

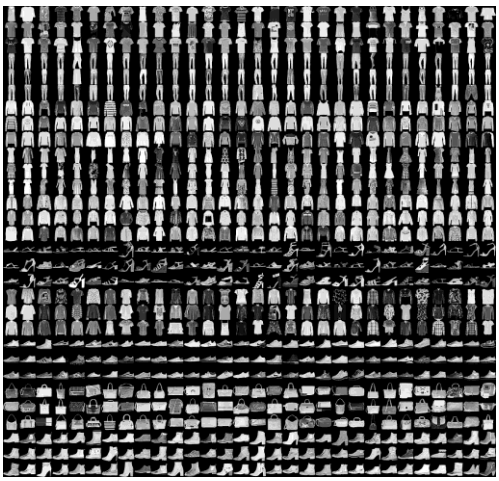

Figure 6. Fashion MNIST Dataset (GitHub, 2021)

Building the Convolutional Neural Network

The architecture of the CNN model that was implemented is depicted in Figure 7 and comprises of 2 convolutional layers, 2 pooling layers, and fully connected layers as seen.

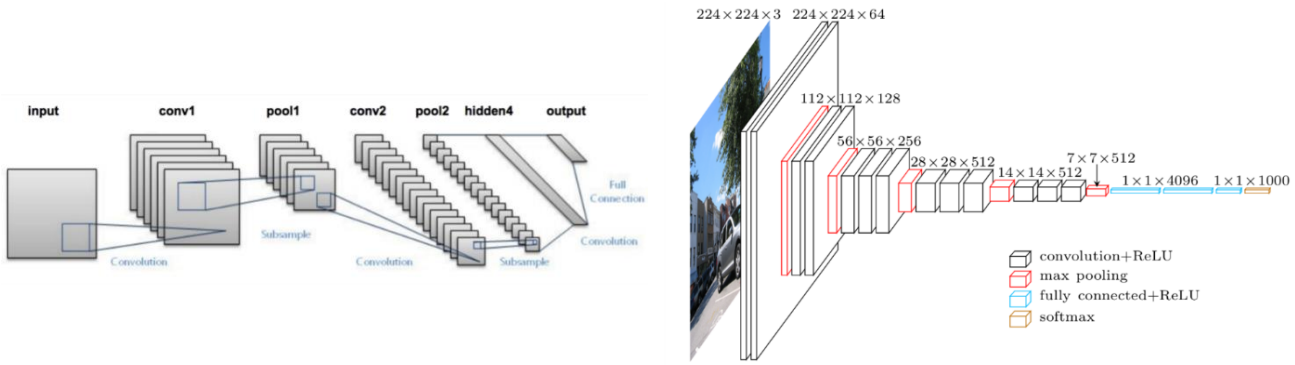

Figure 7. Architecture of the CNN Model Used (Gupta, 2017)

Building the CNN Model was performed as follows,

Build the Convolutional Network

```
In [20]: #Building CNN Model
cnn_model = keras.models.Sequential([
                        tf.keras.layers.Conv2D(filters=32, kernel_size=3, strides=(1,1), padding='valid',activation= 'relu'
                        tf.keras.layers.MaxPooling2D(pool_size=(2,2)),
                        tf.keras.layers.Conv2D(filters=64, kernel_size=3, strides=(2,2), padding='same', activation='relu')
                        tf.keras.layers.MaxPooling2D(pool_size=(2,2)),
                        tf.keras.layers.Flatten(),
                        tf.keras.layers.Dense(units=128, activation='relu'),
                        tf.keras.layers.Dropout(0.25),
                        tf.keras.layers.Dense(units=256, activation='relu'),
                        tf.keras.layers.Dropout(0.25),
                        tf.keras.layers.Dense(units=128, activation='relu'),
                        tf.keras.layers.Dense(units=10, activation='softmax')
                        ])
```

The layers comprised of the following features and hidden layers,

ReLU (Rectified Linear Unit) – As a result, the neural network's computing requirements do not grow exponentially. As the size of the CNN scales, the computational cost of adding more ReLUs increases linearly. **Softmax** – This function serves as the activation function in the output layer of neural network models that predict a multinomial probability distribution. Softmax is used as the activation function in multi-class classification problems when class membership is required on more than two class labels. **Dense Layer** – Each neuron in this basic, dense layer of neurons receives information from every neuron in the layer below it. Based on the results of the convolutional layers, a dense layer is utilized to categorize the images. single neuron in action. Such neurons are found in large numbers inside a layer. **Dropout layer -** This mask leaves all other neurons unaltered while nullifying specific neurons' contribution to the subsequent layer. **Max pooling** is a pooling procedure that chooses the largest element from the feature map region that the filter has covered. The result of the max-pooling layer would thus be a feature map that is smaller and contains the most noticeable features from the preceding feature map. In order to save calculations, the pooling layer basically shrinks the input image's spatial dimension. **Flatten Layer –** to convert multidimensional data into a vector

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 26, 26, 32)        320
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 13, 13, 32)        0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 7, 7, 64)          18496
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 3, 3, 64)          0
_________________________________________________________________
flatten (Flatten)            (None, 576)               0
_________________________________________________________________
dense (Dense)                (None, 128)               73856
_________________________________________________________________
dropout (Dropout)            (None, 128)               0
_________________________________________________________________
dense_1 (Dense)              (None, 256)               33024
_________________________________________________________________
dropout_1 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_2 (Dense)              (None, 128)               32896
_________________________________________________________________
dense_3 (Dense)              (None, 10)                1290
=================================================================
Total params: 159,882
Trainable params: 159,882
Non-trainable params: 0
_________________________________________________________________
```

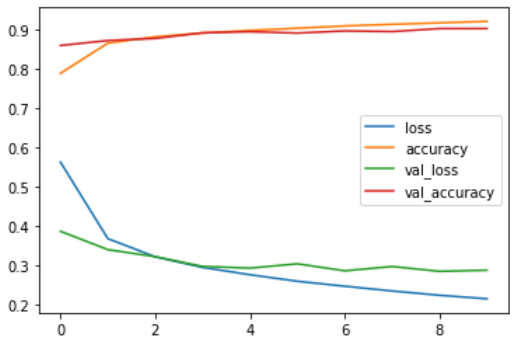

Figure 8. Visualized Architecture and Losses Plot of the Trained Model

2.3 Implementing Ensemble Model

Ensemble Modelling is one of the primary techniques used to lower the prediction's generalization errors in Machine learning and Neural Networks. It uses a variety of modelling techniques or training datasets and build numerous varied models parallelly to predict a result. Thereafter, it combines the forecast in order to evaluate the data accuracy, the data are combined into a single overall forecast. If the base models are varied and independent, the models prediction error will be an all-time low (Kotu et al., 2020).

Ensemble modelling has thus far not been applied to any of the implementations of object recognition of fashion items (especially at shopping malls). With this step being performed, it was able to improve the accuracy of the CNN Model while evaluating it through multiple models with multiple dimensions that captured the images.

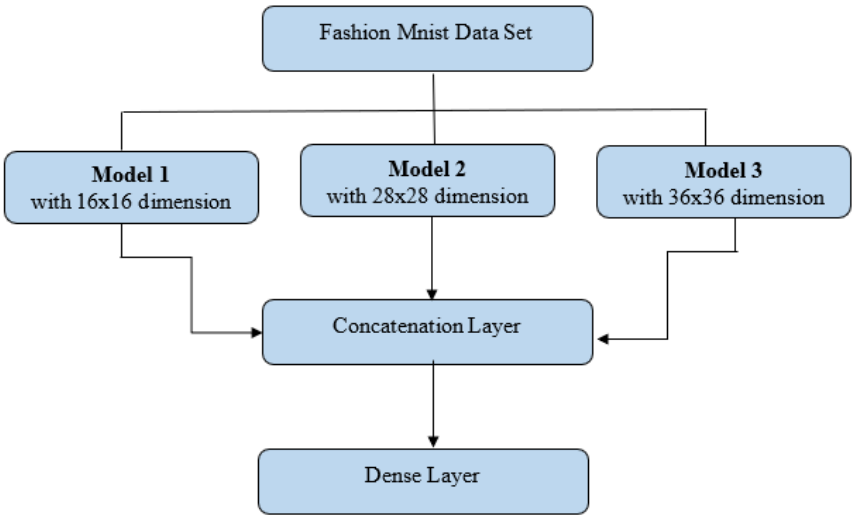

Figure 9. Architecture of the Ensemble Model

The dataset (fashion_mnist) was split and then changed into multiple dimensions of each 16x16, 28x28, and 36x36 after resampling, in order to have 3 models for evaluation. These models were then concatenated using the 'concatenated layer' along with the dense layer in order to produce the output. The dimensions captured the input images and can be visualized as follows,

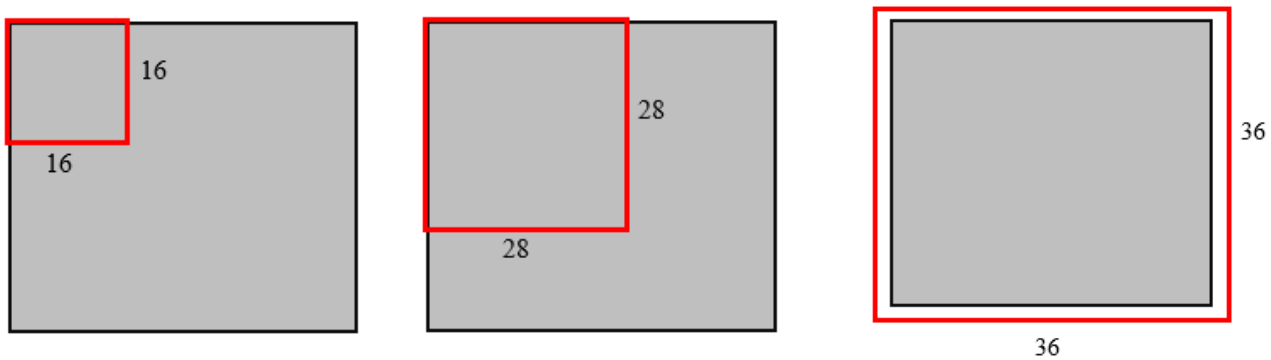

Figure 10. Dimensions of the 3 models implemented by resampling

2.4 Implementing Qualitative Analysis and Principal Component Analysis (PCA)

The Principal Component Analysis (PCA) method was also utilized in order to analyze the large number of datasets with multiple features while preserving the ability to withhold the maximum number of information and features of the multidimensional data. PCA was applied to the data, and it was seen that by taking just 24 features, an 80% of explained variance was obtained for the model as depicted in the plot below,

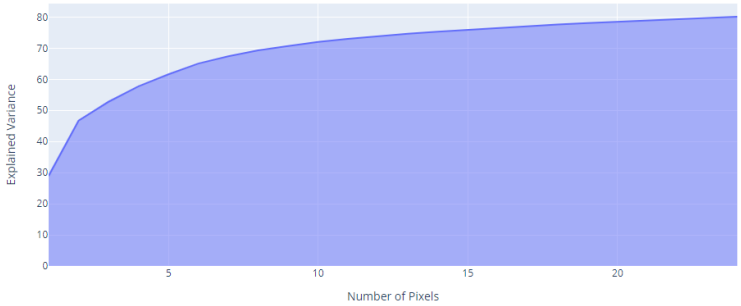

Figure 11. Explained variance as a function of the number of dimensions after PCA

The distribution of data in the 3D space was observed with an explained variance of 80.11%

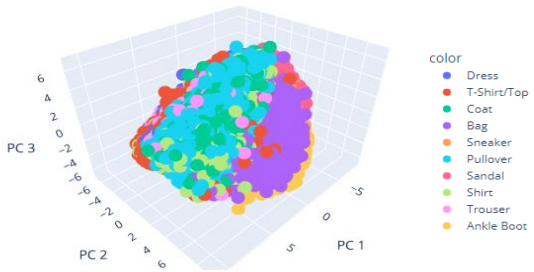

Figure 12. Distribution of data in 3D space

The model was also compared against the Naive Bayes, KNN, Logistic Regression, CatBoost, AdaBoost, XGBoost, and Random Forest models with qualitative analysis. It was observed that the CNN Model depicted the highest accuracy of 90% as visualized follows,

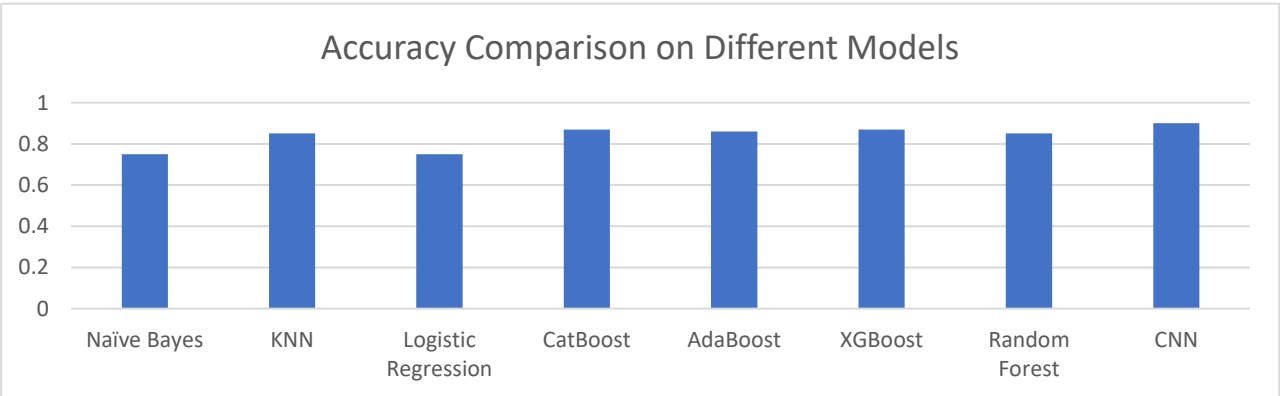

Figure 13. Accuracy Comparison between CNN Model and other Models

3 RESULTS & DISCUSSION

It was observed that the CNN Model depicted a 90% validation accuracy which is thus far the highest number recorder for any model used for object identification. It was also visualized by the aid of a heat map which depicted instances where the identification was performed accurately. The generated heat-map is as follows,

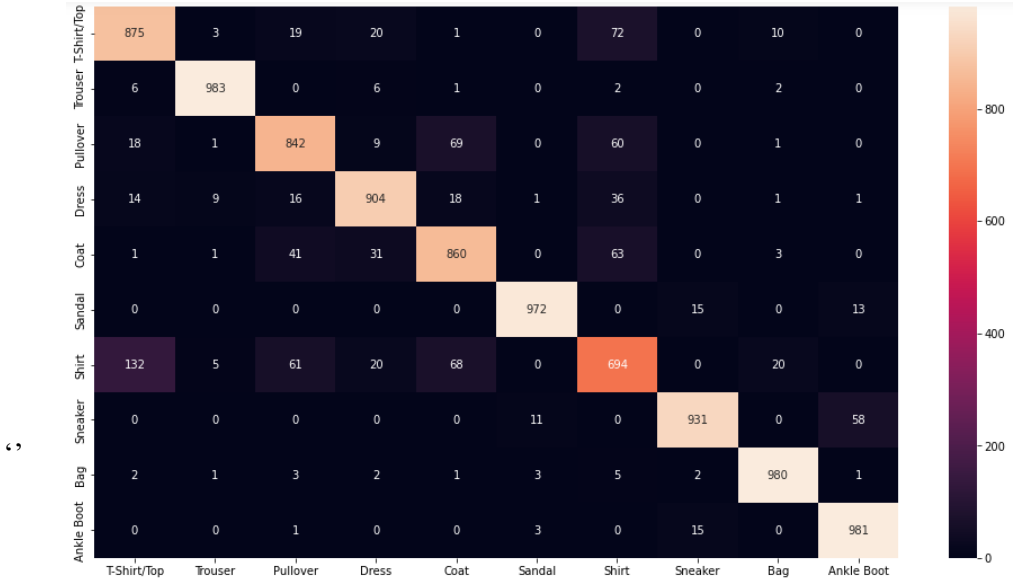

Figure 14. Heatmap generated to visualize accuracy of predictions

Furthermore, the trained model was also implemented on a real-time video obtained at a shopping mall in a VI Shopper's point of view. This video was processed using the Open CV platform and audio output was obtained using Text-to-Speech by converting the produced label texts along with the bounding boxes around the identified objects to speech. This was performed in order to instruct the visually impaired shopper to assist in identifying the garment objects. A few instances of the identified garments are as follows,

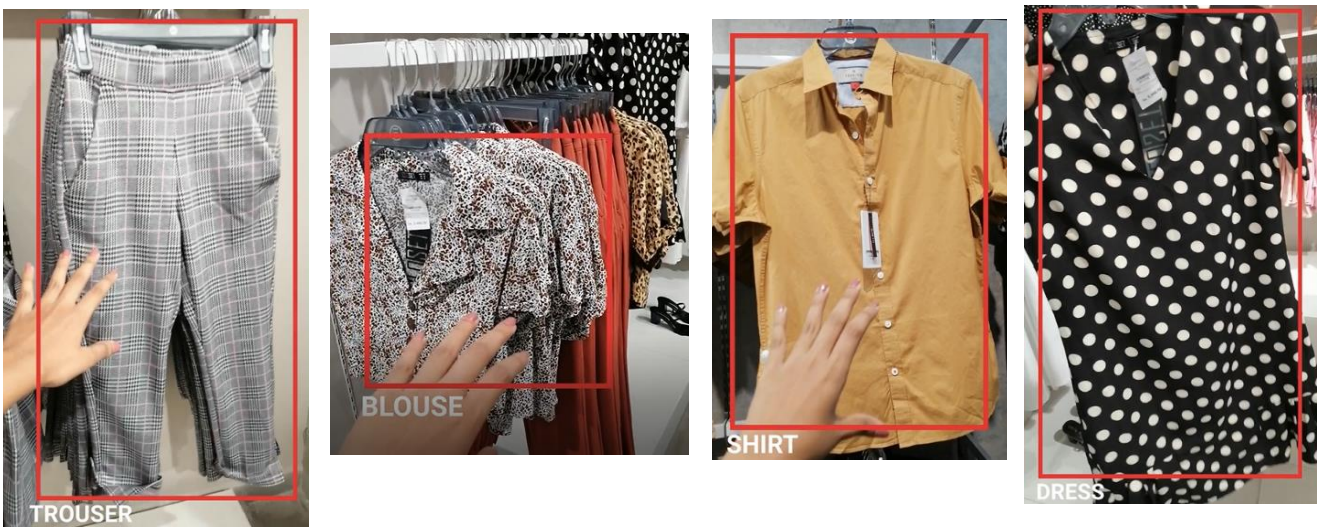

Figure 15. Detected Clothing items in shopping mall

4 CONCLUSION

As mentioned previously, the primary constraint of this design and product is to allow and assist visually impaired individuals to shop at their convenience without the help of any other individual. This is intended to save time and efforts of the VI shopper and provide them with comfort and avoid societal stigma. Due to many such devices already existing in the modern world, it was important that a qualitative analysis of such products along with the proposed solution was performed. As discussed in this paper, it was evident that due to the drawback of the devices that currently exists, a solution to avoid these drawbacks was necessary. The product was then designed specifically to achieve this purpose and provide the users with comfort and accuracy. The basic CNN Model was implemented with a validation accuracy of 90% and was then further evaluated using Ensemble Modelling as well as carrying out a qualitative analysis with other models that is currently used for object recognition.

It was noted that the Ensemble Model was not implemented for fashion object recognition thus far and therefore, the results obtained were of immense success in order to enhance the accuracy of the Model. By using an ensemble model with 3 different models that analyses the spatial features in different dimensions, the Model depicted a 92% accuracy for the dataset. This robust classification model is essential to overcome the drawbacks and challenges faced when classifying basic garments for VI shoppers to shop at their convenience.

The qualitative analysis performed amongst the Naive Bayes, KNN, Logistic Regression, CatBoost, AdaBoost, XGBoost, and Random Forest models claimed that the CNN Model was of the highest accuracy with 90% and 92% after Ensemble Modelling. Therefore, it can be claimed that this research has successfully extended the current research and existing products for assisting VI individuals to recognize objects. The audio output to help instruct the VI Shopper and identify objects was one of the greatest advantages of the model. The model is suggested to be implemented on an app that can be utilized by VI shoppers all over the world in order to obtain assistance and independence while shopping.

REFERENCES

Anwar, A. (2020). A visualization of the basic elements of a Convolutional Neural Network. Retrieved from https://towardsdatascience.com/a-visualization-of-the-basic-elements-of-a-convolutional-neural-network-75fea30cd78d

Food Marketing Institute. (2022). The Food Retailing Industry Speaks 2006 (PDF Download). Retrieved from https://www.fmi.org/forms/store/ProductFormPublic/the-food-retailing-industry-speaks-2006-pdf-download

GitHub. (2021). zalandoresearch/fashion-mnist: A MNIST-like fashion product database. Retrieved from https://github.com/zalandoresearch/fashion-mnist

Gupta, D. (2017). Architecture of CNN | CNN Image Recognition. Analytics Vidhya. Retrieved from https://www.analyticsvidhya.com/blog/2017/06/architecture-of-convolutional-neural-networks-simplified-demystified/

Kotu, V. K., & Deshpande, B. (2020). Ensemble modeling. In Ensemble Modeling - an overview ScienceDirect Topics. Retrieved from https://www.sciencedirect.com/topics/computerscience/ensemblemodeling

Orbis. (2022). New data shows 33 million people living with avoidable blindness. Retrieved from https://www.orbis.org/en/news/2021/new-global-blindness-data#:~:text=2021%20has%20seen%20the%20official,%2Dto%2Dsevere%20visual%20impairment

World Health Organization. (2022). Vision impairment and blindness. Retrieved from https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment

YouTube. (2022). [Video] Retrieved from https://www.youtube.com/watch?v=wTsaWYVeeDg