Information Sciences Letters
An International Journal

http://dx.doi.org/10.18576/isl/130317

Advanced Grocery Store Classification Using Deep Transfer Learning and CNNs

*Walid Dabour**

Department of Mathematics and Computer Science, Faculty of Science, Menoufia University, Shebin El Kom 32511, Menofia, Egypt

Received: 3 Feb 2024, Revised: 15 Feb 2024, Accepted: 29 Apr. 2024.
Published online: 1 May 2024.

Abstract: This paper presents a system employing advanced techniques, including convolutional neural networks (CNNs), transfer learning, fine-tuning, batch normalization, data augmentation, and dropout, to categorize grocery store images. Approximately 285 million visually impaired individuals worldwide face challenges in grocery shopping. The goal of this study is to aid visually impaired individuals in their shopping tasks. Unlike previous methods, which encountered limitations, our approach combines deep learning with data augmentation. This approach utilizes three CNN architectures (VGG19, MobileNet, and Extreme Inception Xception) trained on the ImageNet dataset and evaluate performance using a grocery store image dataset. The hybrid model, particularly Xception, achieves a remarkable F1-score of 98% for overall product recognition. Xception excels in fruit recognition (98% F1-score), MobileNet in vegetables (94% F1-score), and VGG19 in packages (97% F1-score). Our model outperforms existing methods in classification accuracy, precision, recall, and F1-score.

Keywords: Convolutional Neural Network, Data Augmentation, Transfer learning, Fine-tuning.

1 Introduction

The World Health Organization [1] has estimated that more than 2.2 billion people globally have either complete blindness or some form of visual impairment. Assistive technologies aim to enhance the independence and inclusion of individuals with disabilities [2]. Numerous computer vision and machine learning systems have been suggested to aid visually impaired individuals in various tasks such as navigation, text reading, currency identification, and face recognition. With ongoing advancements in information technology, mobile assistive solutions have the potential to enhance the independence, safety, and quality of life of visually impaired individuals [3].

One specific challenge faced by visually impaired individuals is shopping. Therefore, it is crucial to develop technological innovations to assist them in performing these tasks. This study focuses on classifying a dataset related to grocery stores with the aim of deploying a hybrid model on mobile devices to help visually impaired individuals with their daily shopping tasks, utilizing haptic (touch) and audio sensory channels [4].

Previous research in grocery store image classification, particularly aimed at assisting visually impaired individuals, has often encountered limitations that hindered its effectiveness. While deep learning methods have been utilized in these endeavors, they frequently struggled to achieve satisfactory levels of accuracy and robustness in categorizing grocery items. One notable gap in prior papers lies in the lack of integration of data augmentation techniques with deep learning approaches, which could potentially enhance model performance by addressing issues such as limited training data and overfitting. Additionally, existing studies have not thoroughly explored the effectiveness of different CNN architectures for specific product categories within grocery stores. This gap in literature motivated the novel approach presented in our paper. By combining advanced techniques such as convolutional neural networks (CNNs), transfer learning, fine-tuning, batch normalization, data augmentation, and dropout, we bridge this gap and introduce a comprehensive methodology that significantly improves upon the limitations of previous research. Our innovative approach not only addresses the shortcomings of prior papers but also sets a new standard in grocery store image classification, particularly in assisting visually impaired individuals with their shopping tasks. The novelty of our paper lies in the integration of cutting-edge deep learning techniques with data augmentation strategies to address the challenges faced by visually impaired individuals during grocery shopping. While previous research has attempted to utilize deep learning methods for product recognition in grocery stores, our approach introduces a significant advancement by incorporating data augmentation techniques. By doing so, we overcome limitations encountered in prior studies, such as limited training data and overfitting, thereby enhancing the accuracy and robustness of our classification model.

[*] Corresponding author e-mail: walid.dabour@science.menofia.edu.eg

Moreover, our research systematically evaluates the performance of different CNN architectures for specific product categories within grocery stores, providing valuable insights into optimal model selection. This novel approach not only improves upon existing methodologies but also sets a new standard in assistive technologies for the visually impaired, offering a practical solution to enhance their shopping experience.

In this research, our primary focus is on object recognition techniques to help visually impaired individuals identify products commonly found in grocery stores, such as supermarkets, refrigerators, and pantries. The ability to recognize these objects would enable visually impaired individuals to independently identify food products, making it easier to determine items like juice or milk in their refrigerator or pantry. However, object recognition presents several challenges, including variations in environmental conditions, lighting, background noise, object orientation, camera distance, and partial occlusions. Additionally, visually different objects that feel identical to the touch, like milk cartons and juice boxes, are difficult to distinguish [5].

To address these challenges, we propose a method for classifying grocery products using a convolutional neural network (CNN) that incorporates transfer learning, fine-tuning of pre-trained models, data augmentation, batch normalization, and dropout techniques. We suggest a client-server architecture for potential implementation on wearable devices such as smartphones, smart refrigerators, and smart glasses. This system is expected to assist visually impaired users in identifying items in various environments, including their homes, and facilitate independent shopping [6].

Mobile phones are becoming increasingly accessible to individuals with visual impairments, and assistive applications for the visually impaired are being developed for common devices that can be used on the go or in motion, such as smartphones, public transportation systems, and smart homes [7].

This paper's main contribution is improving accuracy and F1-score in classifying multiclass images within the grocery store dataset using hybrid transfer learning models. Initially, a base model categorizes the input image into one of three classes: fruits, vegetables, or packages. Subsequently, the image is re-evaluated to classify it into a specific subclass, such as the type of fruit or milk product. Our method outperforms existing grocery product classification methods, as detailed in the methodology section.

Novelty:

Our paper presents a novel approach to addressing the challenges visually impaired individuals face while grocery shopping by introducing a sophisticated system leveraging state-of-the-art techniques in convolutional neural networks (CNNs). Unlike previous endeavors in product recognition within grocery stores which have encountered limitations, our methodology pioneers the integration of deep learning with data augmentation strategies. This innovation enables our system to achieve remarkable accuracy and robustness in categorizing grocery store items.

Furthermore, our research introduces a unique framework for transfer learning in pre-trained deep learning architectures, specifically tailored to grocery store image classification. By fine-tuning pre-trained models such as VGG19, Xception, and MobileNetV2 on a real-world dataset, we demonstrate significant improvements in model performance, particularly with the Xception architecture achieving an outstanding F1-score of 98% for overall product recognition.

Our study also highlights the effectiveness of different CNN architectures for specific product categories, with MobileNet excelling in vegetable classification, VGG19 performing optimally for packages, and Xception demonstrating exceptional proficiency in fruit recognition. These findings not only contribute to the advancement of assistive technologies for the visually impaired but also hold promise for broader applications such as mobile phone-based assistance, facilitating unrestricted learning and enhanced consumer awareness of grocery store products.

In summary, our proposed methodology surpasses current state-of-the-art models in accuracy, precision, recall, and F1-score, thereby offering a significant leap forward in improving the grocery shopping experience for visually impaired individuals while paving the way for further research into alternative CNN architectures and one-stage detectors for enhanced classification capabilities.

The structure of this article is as follows: Section 2 discusses related publications and common grocery store terms and approaches. Section 3 outlines the methods, Section 4 presents the proposed methodology, and Section 5 analyzes the outcomes of the proposed model and compares them to other relevant models. Finally, Section 6 provides the conclusion and outlines future research directions.

2 Related Work

In a prior research study [8], deep learning was employed to construct a framework for classifying fruits. This framework encompassed two distinct deep learning architectures: a lightweight model consisting of six CNN layers and a 16-layer deep learning model pre-trained by the Visual Geometry Group (VGG16) and fine-tuned. To assess its performance, this framework was tested on two color image datasets, one of which is publicly available. The first dataset (dataset 1) comprised easily distinguishable fruit images, while the second dataset (dataset 2) contained more challenging ones. In dataset 1, the first and second models achieved classification accuracies of 99.49% and 99.75%, respectively. In dataset 2, the first model achieved an accuracy of 85.43%, while the second model reached 96.75%.

Machado et al. [9] conducted a systematic literature review to identify state-of-the-art methods for recognizing grocery products. They analyzed and summarized five publicly accessible datasets used for classifying grocery products. Among the relevant papers, Rivera-Rubio et al. [10] proposed three different approaches that combined various techniques: (1) the combination of the scale-invariant feature transform (SIFT), k-means, and support vector machine (SVM); (2) the utilization of SIFT and locality-constrained linear coding (LLC); and (3) the incorporation of SIFT, principal component analysis, Fisher vector encoding, and an SVM. Additionally, Rivera-Rubio et al. introduced the SHORT-100 dataset and evaluated accuracy on two sets of images: one captured by a smartphone and the other extracted from videos. The SIFT+ K-Means+ SVM approach achieved the highest average accuracy, at 77.51%, for the first set, while SIFT+LLC performed the best on the second set, with an average accuracy of 69.41%.

Varol and Kuzu [11] addressed the recognition of specific commodities (cigarette packages) in images of stocked shelves across different establishments. They introduced a grocery dataset and proposed an approach that involved object detection using the Viola-Jones method. Only 40% of the image, located at the top of the package where the logo is present, was used for classification. The logo image was then represented using shape information (SIFT) and color information (HSV model) through the Bag-of-Words (BoW) approach. Visual vocabularies of color formations were created, and descriptors were clustered using the k-means clustering algorithm. Finally, the SVM classifier was used to identify the product's brand. The experimental results indicated accuracies of 85.9% using only SIFT descriptors, 60.5% using only HSV descriptors, and 92.3% using both descriptors.

Jund et al. [12] implemented transfer learning via the Caffe deep learning framework and proposed a fine-tuning CNN (FT-CNN). They introduced the Freiburg Groceries dataset for accuracy evaluation and achieved an average accuracy of 78.9% using their method.

Klasson et al. [13] utilized three pretrained CNNs, namely AlexNet, VGG16, and DenseNet-169, as feature descriptors for an SVM classifier. They extracted layers from the CNNs both with and without fine-tuning. Furthermore, they introduced a new grocery store dataset that facilitated fine-grained and coarse-grained classification of specific products. The FT-CNN DenseNet-169 approach combined with the SVM achieved an accuracy rate of 85.0% for classifying specific products. The CNN DenseNet-169 approach without fine-tuning, when paired with the SVM classifier, achieved a slightly better accuracy rate of 85.2% for classifying product classes. There are some applications that have been focused on IoT, ML and AI [34-44].

3 Methods

The proposed method uses different pretrained CNNs provided by the Keras library. In addition, we focus on a multiclass classification problem; thus, the SoftMax activation function is used in the final layer. Tests were performed to increase accuracy and reduce the loss function through transfer learning with fine tuning, data augmentation, batch normalization, and dropout layers. In addition, adjustments were made to the resolution of the images, and the domain-batch size value and photometric normalization functions from the Keras library were used for the proposed method.

3.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are widely employed in image processing tasks, effectively addressing the key limitations associated with deep neural networks. These networks play a vital role not only in image classification but also in tasks such as object recognition, image segmentation, Generative Adversarial Networks (GANs), and various other image-related applications. The construction of a CNN can follow various approaches, and there are numerous pretrained models that leverage CNNs to excel in diverse tasks.

In our research, we extensively utilized CNNs as a foundational component. The fundamental building blocks of a CNN consist of convolutions, filters, strides, padding, and pooling operations. Numerous pretrained CNN models have demonstrated exceptional performance in classification tasks, including well-known ones such as VGG, Inception, ResNet, AlexNet, GoogleNet, MobileNetV2, Xception, and DenseNet. In the context of our study, we specifically employed six pretrained models, namely AlexNet, VGG, ResNet, MobileNet, Xception, and Inception. Each of these models was configured with distinct depths, parameter counts, and input dimensions, but all were originally trained on the ImageNet dataset.

To adapt these architectures for our specific task of classifying grocery store images, we fine-tuned them. This fine-tuning process involved updating the fully connected layers responsible for multiclass classification while retaining the weights learned from ImageNet. Furthermore, we resized grocery store images to a consistent size of 224x224 pixels with three color channels to align with the requirements of the Keras framework for fine-tuning.

3.2 Deep Transfer Learning

Transfer learning (TL) is a well-established technique in machine learning, where a model is created based on a pretrained model [6]. The pretrained model is initially trained to perform a specific task using a substantial dataset, such as ImageNet. In this process, each layer of the deep neural network (DNN) learns data-related features, and these layers are connected to deeper layers through trainable weights. Consequently, the features learned in earlier layers serve as input data for subsequent deeper layers [23,24]. Additionally, the input data and learned weights undergo convolution to generate a new feature map, which is then passed through an activation function [25].

The primary objective of transfer learning is to enhance the performance of target learners in specific domains by transferring knowledge from distinct but related source domains. This approach can reduce the reliance on extensive target-domain data for training target learners. Transfer learning has gained significant prominence in the field of machine learning due to its extensive range of potential applications and promising prospects.

3.3 Data Normalization

Data normalization refers to the process of transforming the pixel values of image data into a predefined range, typically [0, 1] or [-1, 1]. Typically, images have pixel values ranging from 0 to 255, and these large integer values can potentially disrupt or slow down the learning process in a Deep Neural Network (DNN). Therefore, it is advisable to perform image normalization to bring the pixel values within the desired range of [0, 1].

In our dataset, we achieved image normalization, which involves rescaling the images, using the Python ImageDataGenerator function with the parameter rescale=1. / 255. This rescaling operation ensures that the pixel values of the images are adjusted to fall within the range [0, 1].

3.4 Data Augmentation

To ensure the effective training of a Deep Learning (DL) model, it is crucial that the model possesses a high degree of generalizability. Generalizability is demonstrated by the model's capacity to accurately categorize unseen test data and the diversity of the training data. In essence, a robust DL model should encompass various data variations, such as different sample orientations, color scales, or other factors. When a DL model misclassifies an image, it may be because it has not been exposed to different versions of that image with varying color ranges, orientations, or positions during its training. Models lacking good generalizability tend to suffer from overfitting, where they achieve high accuracy during training but exhibit poor accuracy when applied to validation data. Building an effective DL model requires that validation accuracy increases alongside training accuracy, and validation errors decrease in tandem with training errors.

Data augmentation [25], [26] is a valuable technique in developing comprehensive models that account for a wide range of possibilities and is particularly useful for mitigating overfitting. In our study, we performed data augmentation using Python, utilizing the ImageDataGenerator class from Keras. This process involved standardization, rotation, shifts, and adjustments in brightness, among other transformations.

The Keras ImageDataGenerator class is specifically designed to offer real-time data augmentation, offering a significant advantage. It applies image augmentation to the images in each epoch, contributing to the model's improved generalizability and overall performance.

4 The Proposed Methodology

The goal of our approach is to categorize the input image into its specific subclass, given that the dataset contains a total of 83 classes. Initially, we designed a Convolutional Neural Network (CNN) architecture to extract features from the images, and we employed a Support Vector Machine (SVM) as the classifier. However, this configuration yielded an accuracy of only 66%.

In an attempt to improve accuracy, we experimented with a pretrained model designed to handle all 83 classes within the dataset. Unfortunately, this approach did not yield satisfactory results. Recognizing that the dataset follows a hierarchical structure, we propose a multi-model approach that combines various pretrained models to effectively classify the multiclass images in the dataset.

4.1 Dataset Description

The dataset consists of images captured within the fruit, vegetable, and refrigerated product sections of 18 different grocery stores. It comprises a total of 5125 photos, which are categorized into 81 distinct fine-grained groups. Each class contains a varying number of images, ranging from 30 to 138.

Our objective was to collect natural images under conditions resembling those encountered in a mobile phone assistive application. All the images were taken using a 16-megapixel Android smartphone camera, capturing items from various distances and angles. It's worth noting that some images may include background elements or items placed on the wrong shelf alongside the intended item. These instances are common in real grocery store environments, and image classifiers used for assistive purposes must be capable of effectively handling such extraneous information. Additionally, lighting conditions within the images may vary based on the item's location within the store. Some images are taken while the photographer holds the item in their hand, especially for refrigerated products, which are often tightly packed in refrigerators. To ensure diversity, we intentionally varied the positioning of the objects in these images, avoiding a consistent centered or complete view.

For further insights and a comprehensive description of the dataset used in this study, readers can refer to the work conducted by Klasson et al. [13].

4.2 Data Preprocessing

Following the training of our models on the dataset introduced by Klasson et al., we identified certain issues with the dataset. Notably, there were inaccuracies in the class arrangement, resulting in instances where classes in the test set did not correspond to those in the training or validation sets, and vice versa. Furthermore, the initial split of the dataset into training, validation, and testing sets lacked a logical and effective ratio.

To address these concerns and align the dataset with our objectives, we performed a new split. We combined all the images and divided them into training, validation, and testing sets, maintaining a ratio of 70:10:20, respectively. This allocation was chosen to strike a balance between computational costs during model training and evaluation and the representativeness of the training and test sets. This revised split led to an improvement in accuracy.

To further enhance the size of our training set, we employed various data augmentation techniques. Given that the target (i.e., a captured image of a product) can appear under diverse conditions, including different orientations, positions, scales, and lighting conditions, we applied realistic transformations. These transformations included image rotation (up to a maximum of 40 degrees), scaling of width and height (by a factor of 0.2), shearing (with a small value of 0.2), simulating motion blur, zooming (with a factor of 0.2), and horizontal flipping (to enable recognition of the object from both sides). Figure 1 illustrates samples of the images generated using these data augmentation methods.

It's important to note that these data augmentation techniques were exclusively applied to the training data, as indicated in Table 1.

Table 1: the augmentation techniques and the parameter of each technique

Argument	Parameter value
Rescale	1/255
Rotation range	40
Zoom range	0.2
Shear range	0.2
Horizontal shift range	0.2
Vertical shift range	0.2
Horizontal flip	True

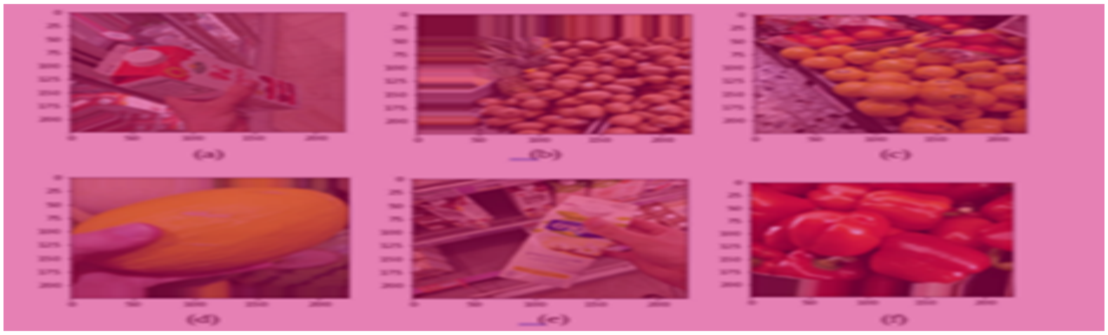


Fig. 1: Samples

To implement data augmentation, we employed the Keras Image Data Generator class, a technique that operates dynamically during the training process. The procedure can be outlined as follows:

1. Take a batch of input images intended for training.
2. Apply a predefined set of transformations to each individual image within the batch.
3. Substitute the original batch with this newly generated batch, which now includes randomly transformed images.
4. Proceed to train the Convolutional Neural Network (CNN) on this batch containing the randomly transformed images.

4.3 Base Model

Our approach to model creation involved two distinct phases. Initially, we developed and fine-tuned a foundational CNN model. Subsequently, we applied Transfer Learning (TL) by utilizing various pretrained models, including VGG, InceptionV2, ResNet, AlexNet, and MobileNetV2. During the construction and tuning of the model, we made several noteworthy observations:

1. Increasing the number of training epochs led to improved accuracy.
2. The implementation of early stopping proved effective in reducing execution time.
3. Modifying the batch size had a positive impact on execution time and memory requirements.
4. Employing data augmentation was instrumental in preventing overfitting of the data.

Our proposed model, based on a hybrid of pretrained models, was designed for the prediction and classification of the grocery store image dataset. This model leveraged pretrained architectures to handle the dataset, with the training process involving four distinct models. Initially, we established a base model to classify input images into one of three broad categories: fruits, vegetables, or packages. Subsequently, based on this initial classification outcome, we further classified the images into their respective subclasses, such as identifying specific types of fruits or dairy products. We extracted relevant features, and the architecture employed a categorical cross-entropy loss function tailored for classification tasks. The performance of this hybrid model was evaluated using the test set. The flow of this framework is illustrated in Figure 2, and the proposed method is divided into four phases, as depicted in Figure 3. Detailed descriptions of each phase can be found in the Method and Experiments and Results sections.

Model creation approach was divided into two stages. First, we constructed and tuned a base CNN model, and then we applied TL utilizing several pretrained models: VGG [14], InceptionV2 [28], ResNet[16], AlexNet [17], MobileNetV2 [19]. In the model construction and tuning process, we discovered that increasing the number of epochs improved accuracy, early stopping helped reduce execution time, changing the batch size improved the execution time and memory requirements, and data augmentation helped avoid data overfitting.

The proposed model, which is based on a hybrid of the pretrained models, was designed to predict and classify the grocery store image dataset. The pretrained architecture utilized in the proposed train the dataset as we will train four models. First, constructed the base model to classify the input image into one of three classes (i.e., fruits, vegetables, or packages), and then, based on this initial classification result, we classified the image into its subclass (e.g., which type of fruit or milk product). We retrieved extracted features, and this architecture calculates the categorical cross-entropy loss function for classification problems. Then, the hybrid model was assessed using the test set. Fig. 2 shows a flow diagram for this framework. The proposed method is divided into four phases, as shown in Fig. 3. Each phase of the proposed method is described in the method and experiments and results section. Data-driven models that can be trained to learn essential features from raw Input are becoming increasingly popular, particularly when used in conjunction with feature learning, which is a core advantage of DL [4], [5].

A comparison of several TL models is presented in the following sections. DL models require a large dataset to function effectively [29], and model performance can be improved by supplementing the available source images. This may also be accomplished by image preprocessing.

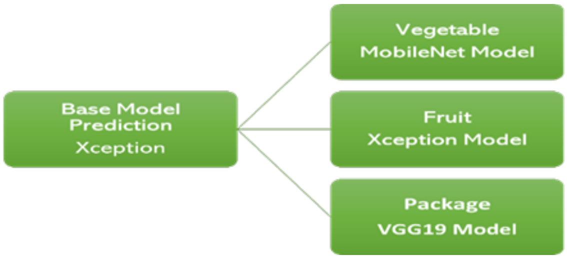


Fig. 2: Framework flow diagram.

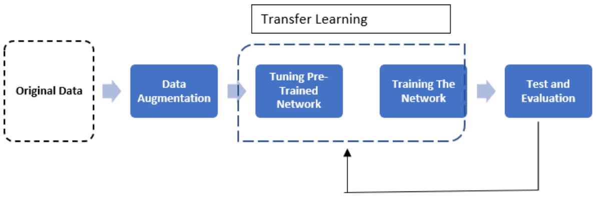


Fig. 3: The framework of the proposed model

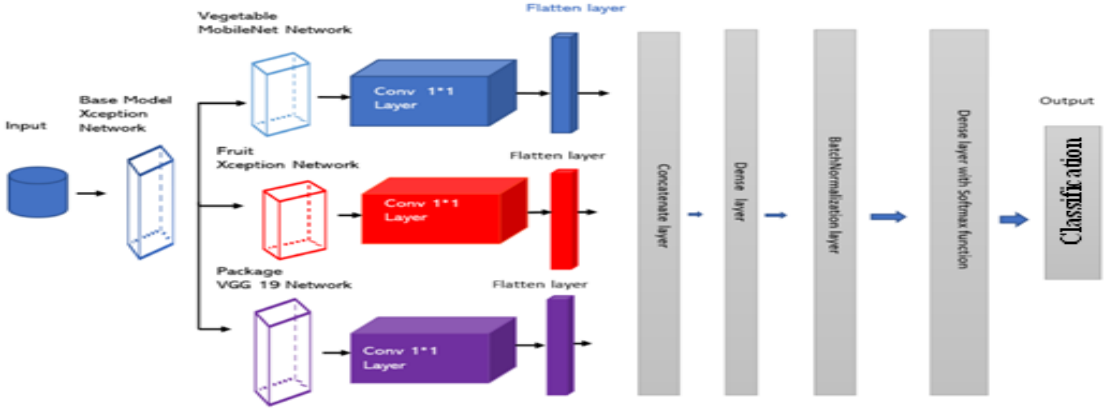


Fig. 4: The overall architecture of end-to-end model

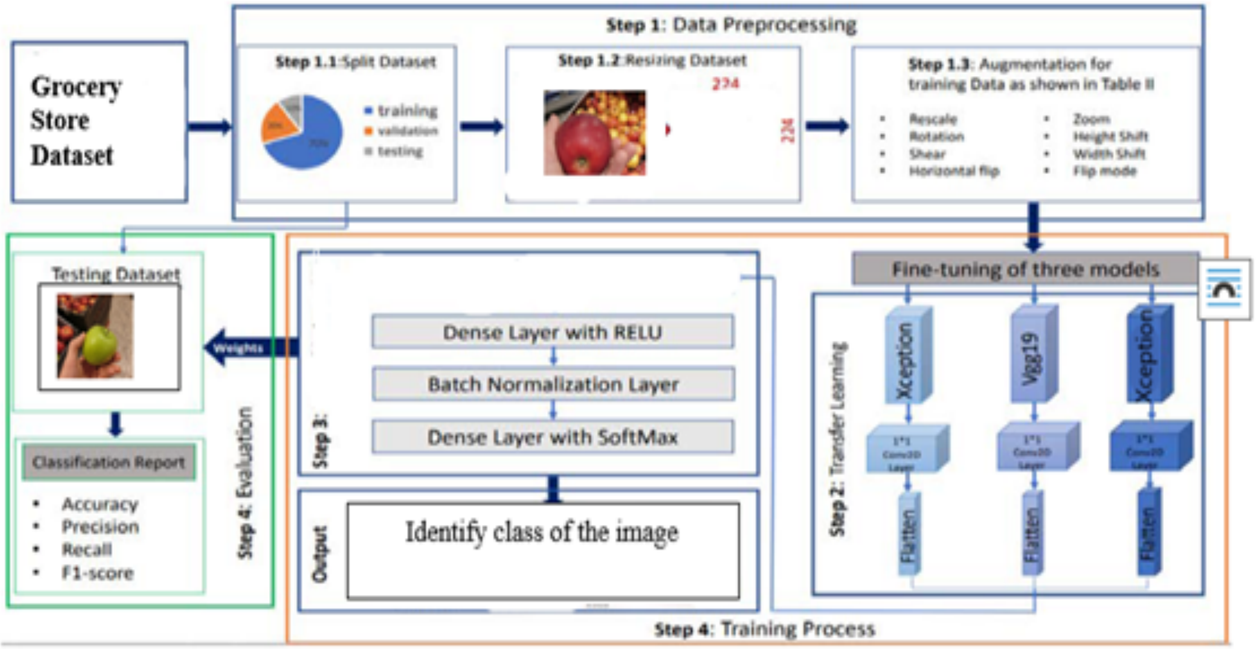


Fig. 5: Block diagram of the proposed approach for the classification of different grocery dataset.

Table 2: performance equations summary

Assessments	Equation	Equ.No
Accuracy (Acc)	$\frac{TP+TN}{TP+TN+FP+FN}$	(1)
F1-Score (F)	$2 * \frac{P*R}{P+R}$	(2)
Precision (P)	$\frac{TP}{TP+FP}$	(3)
Recall (R)	$\frac{TP}{TP+FN}$	(4)

The proposed model is based on a modified CNN that incorporates training weights from the MobileNet, VGG19, and Xception models. The learning scenario begins with the input layer getting images of grocery store products. Three pre-trained networks share the input layer. By freezing all of the layers of three models, they are reshaped. However, in order to add the required output layers, we dropped the top layer (output layer) from each model.

A Conv 1x1 layer with 1024 filters with padding-zero and stride-one was added to gather the most essential features and allow for reduced dimensionality, followed by a flattening layer to convert the matrix to a one-dimensional tensor (vector). As seen in Fig. 4 and Fig. 5 shows a block diagram of the proposed method.

5 Experiments and Results

5.1 Experimental Settings

This study illustrates that transfer learning (using VGG19 Model, MobileNet, and Xception) achieves the maximum accuracy using a grocery store dataset Table 3 shows how the dataset is partitioned in a 7:2:1 ratio. All tests are carried out using Google's Colab [30], the Keras framework [31], which can operate on top of TensorFlow, and the Python programming language. All experiments were carried out on an NVIDIA Tesla K80 GPU (Graphical Processing Unit) with 12 GB of RAM.

5.2 Experimental Evaluation

In the testing phase equations provided in Table 2, the performance of the proposed model is assessed using accuracy (1), F1-score (2), precision (3), recall (4), and confusion matrix. Furthermore, as illustrated in Fig. 9-12, loss, accuracy, validation loss, and validation accuracy are determined at various epochs throughout the training phase. In addition, we compared the performance of our proposed model (shown in Table 4-6) with that of other models (e.g. [13]) that operate on the same dataset (shown in Table 8) and with each architecture utilized in the proposed model separately.

The Adam optimizer was used to compile the proposed model, which is a stochastic gradient descent approach with a learning rate of 2e-5. The loss function is a categorical cross-entropy that is used to estimate the loss in the multi-class classification task.

Where: True Positive (TP), the model properly predicts the positive class. The True Negative (TN) model classifies the negative class properly. The model predicts the positive class incorrectly in a false positive (FP). The model predicts the negative class incorrectly in false negative (FN).

5.3 Experimental Results and Analysis

Here, we present the outcomes of six distinct experiments conducted on the base model, as well as the comprehensive results of an experimental study involving multiclass images from grocery stores. We employed a pretrained CNN model and a hybrid model, and a comparative analysis of various models is included. The results obtained from these models are compared against existing state-of-the-art methods, ultimately leading to the identification of the best-performing model.

In contrast to conventional neural networks used for image recognition, where image features are manually defined, Convolutional Neural Networks (CNNs) enable the extraction of higher-level image representations. Due to the scarcity of adequately large datasets, only a limited number of individuals opt to train an entire CNN from scratch, given the extensive resources or complex datasets required for such training. Instead, it is common practice to leverage a pretrained ConvNet, initially trained on a substantial dataset (e.g., ImageNet, containing 1.2 million images), and utilize it as either initialization or a fixed feature extractor for solving the target problem. This practice is the essence of Transfer Learning (TL).

To identify the optimal model for our task, we pretrained several networks and assessed their performance. The broad framework for the TL phases we applied is as follows:

1. Load a pretrained CNN model trained on an extensive dataset.
2. Freeze the parameters of the model's lowest convolutional layers (i.e., the weights).
3. Add a custom classifier with multiple layers of trainable parameters to the model.
4. Train the classifier layers using the available training data.
5. Adjust hyperparameters such as the learning rate and number of epochs and unfreeze additional layers as necessary.

Fine-tuning a pretrained model involves the following steps:

- Configure new fully connected and output layers.
- Freeze the pretrained layers.
- Unfreeze and retrain the last few pretrained layers.

Base Model Networks: Commencing with the base model, we evaluated several pretrained networks that were originally trained on the ImageNet dataset for the purpose of classifying our dataset into three primary classes: fruits, vegetables, and packages.

I. **AlexNet**: AlexNet, a prominent architecture among pretrained models in computer vision, was implemented due to its unavailability in Keras. However, after multiple runs with different optimizers and epochs, the results in terms of training-validation accuracy, loss, and test set evaluation were not favorable (refer to Figure 7).

II. **ResNet**: We explored ResNet-50 and ResNet101-V2, which introduced the concept of skip connections to address the vanishing gradient problem. Training-validation accuracy and loss over 30 epochs for ResNet50 with RMSprop optimizer, ResNet50 with SGD optimizer, and ResNet101-V2 with RMSprop optimizer were examined. However, the results did not align well with the dataset.

III. **VGG**: VGG models exhibited a good fit with the data, displaying stable accuracy curves (see Figure 7). Training-validation accuracy and loss were observed using the Adam optimizer with a learning rate of 0.001 over 50 epochs for both VGG16 and VGG19 networks. Notably, precision and recall metrics on the test set revealed certain model errors and misclassifications, particularly in the vegetable class.

IV. **Inception**: We evaluated Inception-V2 and Inception-V3, with Inception-V2 delivering superior results compared to Inception-V3. Nevertheless, misclassification issues were noted, particularly concerning the vegetable class. Training was performed using RMSprop optimizer with a learning rate of 0.0001 over 30 epochs for both Inception-V2 and Inception-V3 networks.

V. **MobileNet**: Two versions of MobileNet were assessed in the base model. Although confusion matrices from test set evaluation displayed promising results with minimal errors, the learning curves exhibited significant noise, including oscillations, indicating difficulties in fitting many batches of the validation set and a lack of generalization (see Figure 7).

VI. **Xception**: Xception, an "Extreme" version of Inception, utilizing modified deep-wise separable convolutional layers, was considered. Despite the associated training costs, Xception demonstrated significant improvements over the Inception network. Stable learning curves and favorable evaluation results on the test set were observed, as depicted in Figures 7 and 8. Consequently, Xception was selected as the ultimate base model.

5.4 Final model results

We observed that the dataset required reorganization. Originally, the dataset was divided into three folders: training, validation, and test. During our examination, we identified certain issues with the dataset introduced by Klasson et al. [13]. Specifically, we noted that some classes were incorrectly organized. In other words, certain classes present in the test set did not correspond to classes in the training or validation sets, and vice versa. Additionally, the division of images into training, validation, and testing sets did not adhere to a logical distribution with the correct proportions.

As a remedy, we amalgamated all the images and restructured the dataset into three distinct sets: training, validation, and testing, with a distribution ratio of 70:10:20, respectively, as outlined in Table 3. Consequently, the training set comprised 3780 images, the validation set included 508 images, and the testing set encompassed 1156 images. Our next step involved initiating the training of the second-level models, as depicted in the hierarchy diagram (refer to Fig. 2). These models were designed to further classify each category into its respective subclasses.

In this phase, we commenced training the same pretrained models, previously employed, on the subclasses within each category. During this process, we identified the models that exhibited the best fit for our revised dataset arrangement. Moreover, in pursuit of precision, we retrained the Xception base model once again, this time using the reorganized dataset.

Table 3: description of the utilized dataset after augmentation

Category	Training set	Validation set	Testing set
fruit	1649	224	497
vegetables	913	122	281
packages	1218	162	378
Total	3780	508	1156

Following, we list the best-resulting model for each category along with its learning curves, confusion matrix, and evaluation metrics on the test set:

5.4.1 **Base model (Xception)**

Fig. 8 shows the learning curves of the xception network for our base model after retraining, Fig. 9 shows the confusion matrix of the test set, and Table 4 summarizes the test evaluation metrics.

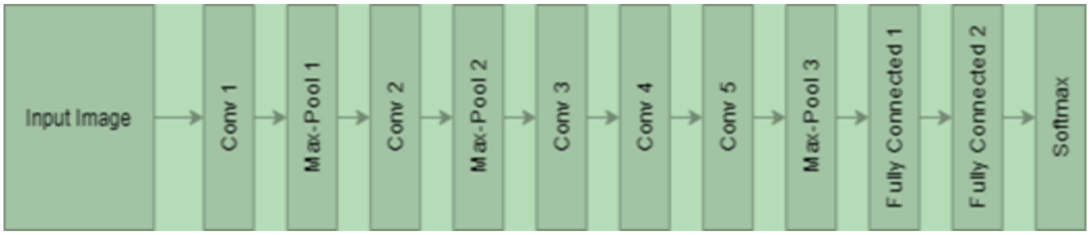
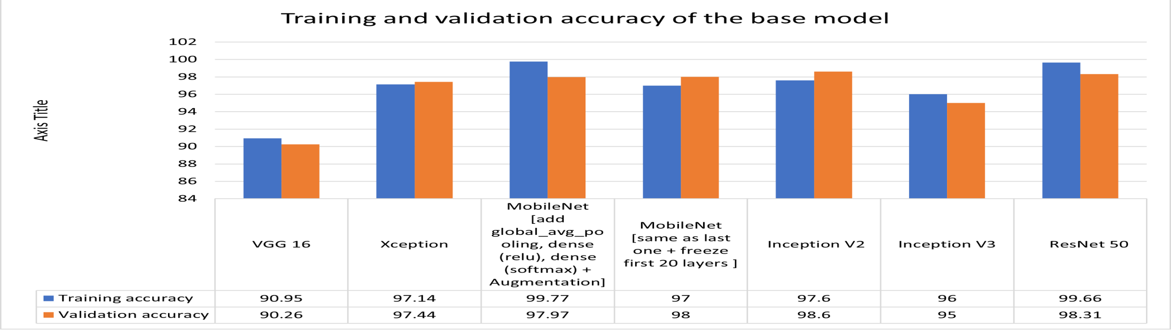


Fig. 6: Alex architecture



	VGG 16	Xception	MobileNet [add global_avg_pooling, dense (relu), dense (softmax) + Augmentation]	MobileNet [same as last one + freeze first 20 layers]	Inception V2	Inception V3	ResNet 50
🟦 Training accuracy	90.95	97.14	99.77	97	97.6	96	99.66
🟧 Validation accuracy	90.26	97.44	97.97	98	98.6	95	98.31

Fig. 7: Training and validation accuracy of the base model

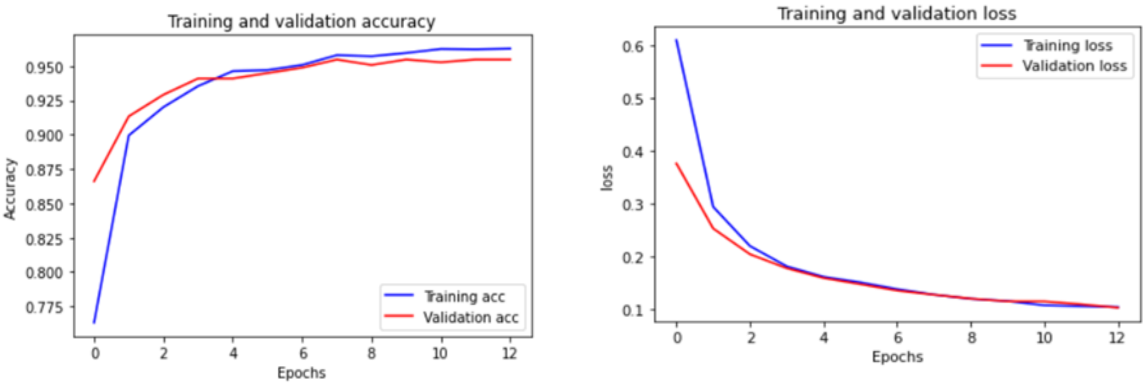


Fig. 8: Training-validation accuracy and loss graphs of the Xception model over 50 epochs.

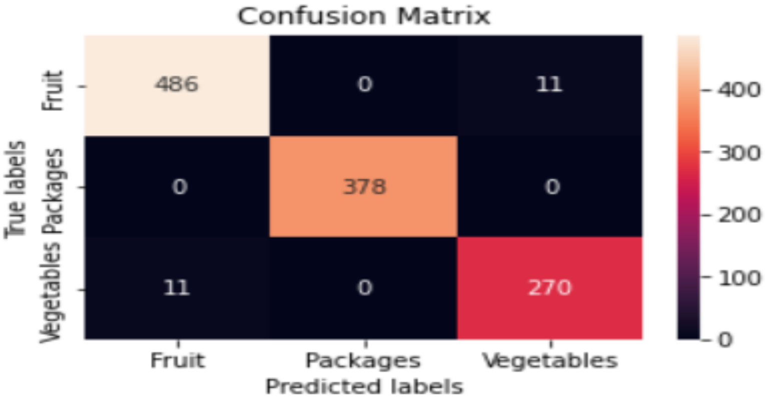


Fig. 9: Confusion matrix of the test set (Base Model).

Table 4: evaluation metrics of a base model

Testing set	
	Base model Xception
Precision (weighted avg)	98.0
Recall (weighted avg)	98.0
F1-score (weighted avg)	98.0
Accuracy	98.0

5.4.2 Fruits model (Xception)

Xception network once again outstanding the fruits data and was by far the best model. We can notice this in the learning curves (Fig. 10) and the test set evaluation metrics (Table 5).

Table 5: evaluation metrics of the fruits model

Testing set	
	Fruits Xception
Precision (weighted avg)	98.0
Recall (weighted avg)	98.0
F1-score (weighted avg)	98.0
Accuracy	98.0

Table 6: evaluation metrics of the vegetable model

Testing set	
	Vegetables MobileNet
Precision (weighted avg)	95.0
Recall (weighted avg)	95.0
F1-score (weighted avg)	94.0
Accuracy	95.0

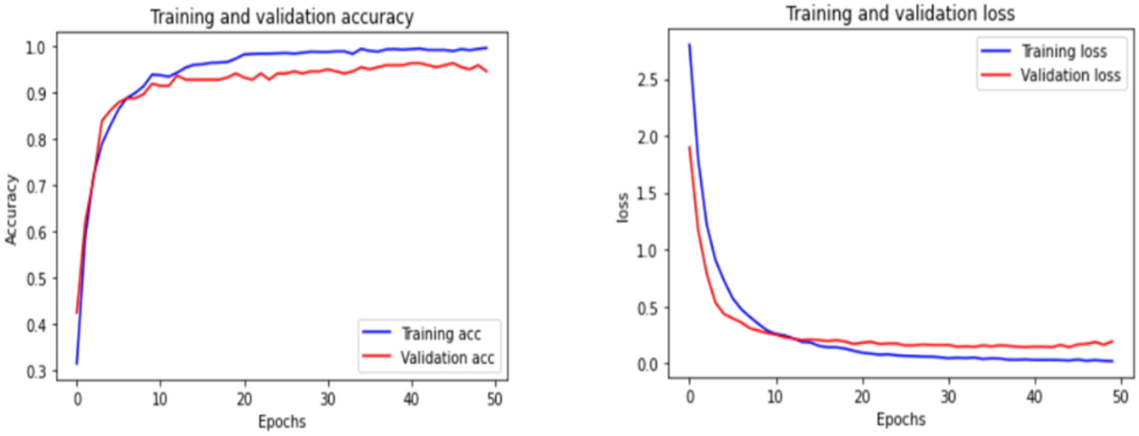


Fig. 10: Training-validation accuracy and loss graphs of fruits model for the Xception network

5.4.3 Vegetables model (MobileNet)

Vegetable classes were very challenging, and we evaluated several networks with multiple fine-tuning and different parameters; at last, MobileNet was the best network for the vegetable data. Fig. 11 and Table 6 show the train-validation accuracy, loss graphs, and evaluation metrics.

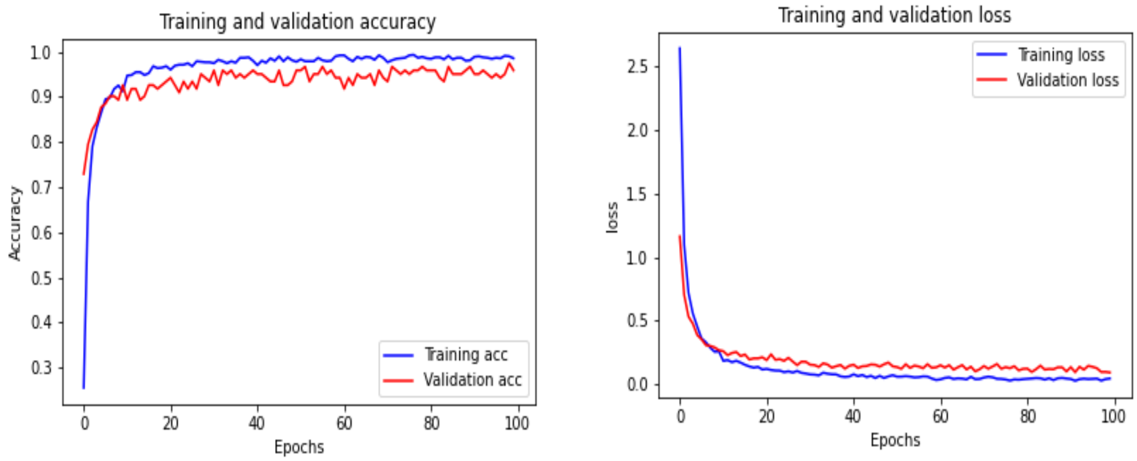


Fig. 11: Training-validation accuracy and loss graphs of vegetables model with MobileNet network

5.4.4 Packages model (VGG19)

VGG-19 was the best network for package data, and the results are shown in Fig. 12 for learning curves and Table 7 for evaluation metrics on the test set.

The proposed approach based on hybrid transfer learning models achieved superior accuracy on the approaches based on CNNs and on the approaches found in the literature (see Table 8).

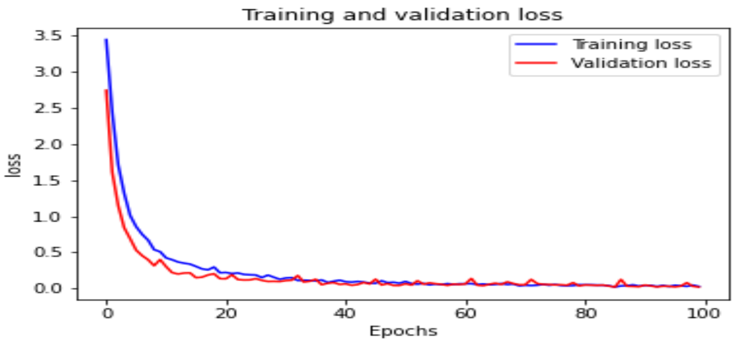


Fig. 12: Training-validation accuracy and loss graphs of packages model with VGG19 network

Table 7: evaluation metrics of the packages model

Testing set	
	Packages VGG19
Precision (weighted avg)	97.0
Recall (weighted avg)	96.0
F1-score (weighted avg)	97.0
Accuracy	97.0

Table 8: Literature accuracy results versus the best approaches of this work

Database		Accuracy (in percentage)				
		State-of-Art	Proposed model			
			Base_model	Fruit model	Vegetable model	Package model
SHORT-100 [9]	ST-SG	77.51				
	VF-SG	69.41				
Grocery [10]		92.3				
Freiburg Groceries [11]		78.9				
Grocery Store [13]		85.0	98.0	98.0	94.0	97.0
		85.2				

Key Findings and Contributions:

1. **Innovative Methodology for Grocery Store Image Classification:** Our paper introduces a novel approach that combines advanced deep learning techniques with data augmentation strategies to accurately categorize grocery store items. This methodology significantly improves upon existing methods, particularly in assisting visually impaired individuals with their shopping tasks.
2. **Effective Transfer Learning Framework:** We propose a unique framework for transfer learning in pre-trained deep learning architectures, specifically tailored to grocery store image classification. By fine-tuning pre-trained models such as VGG19, Xception, and MobileNetV2, we achieve remarkable improvements in classification accuracy, demonstrating the efficacy of our approach.
3. **Optimal CNN Architectures for Product Categories:** Our study identifies the most effective CNN architectures for specific product categories within grocery stores. Xception emerges as the top-performing model for fruit recognition, MobileNet excels in vegetable classification, and VGG19 demonstrates proficiency in package identification. These findings provide valuable insights for developing specialized models tailored to different types of grocery store products.
4. **Superior Performance Metrics:** Our proposed hybrid model surpasses current state-of-the-art models in terms of classification accuracy, precision, recall, and F1-score. Particularly noteworthy is the achievement of an impressive F1-score of 98% for overall product recognition using the Xception architecture, underscoring the effectiveness of our methodology.
5. **Broader Applications and Implications:** Beyond aiding visually impaired individuals, our research has broader implications, including the potential integration of our model into mobile phone applications to provide accessible assistance and facilitate unrestricted learning for consumers. This contributes to greater consumer awareness of grocery store products, benefiting both visually impaired individuals and the general public.

Overall, this study presents a significant advancement in assistive technologies for the visually impaired and contributes to the field of computer vision by addressing real-world challenges in grocery shopping accessibility. Through innovative methodologies, effective transfer learning frameworks, and superior model performance, we provide valuable insights and pave the way for future research in this domain.

In future directions, while the exploration of alternative CNN architectures and one-stage detectors holds promise, providing more concrete plans or hypotheses for future research directions would strengthen the paper. Specifically, investigating the efficacy of emerging CNN architectures beyond those examined in this study could lead to further improvements in classification performance. Additionally, exploring the integration of advanced techniques such as attention mechanisms or reinforcement learning into the classification pipeline may enhance the model's ability to discern finer details and improve overall accuracy. Moreover, considering the dynamic nature of grocery store layouts and product offerings, developing adaptive learning algorithms that can continually update and refine the classification model in real-time could be beneficial. Furthermore, extending the scope of the research to encompass multi-modal approaches, such as incorporating text recognition or audio cues in conjunction with image classification, could provide a more comprehensive solution for assisting visually impaired individuals during grocery shopping. By outlining these concrete plans and hypotheses for future research directions, the paper can not only inspire further inquiry but also contribute to the advancement of assistive technologies in this domain.

6 Conclusion

To identify the grocery store dataset, we proposed a unique framework of transfer learning in pre-trained deep learning architectures based on VGG19, Xception, and MobileNetV2. We made use of a real-world dataset. Fine-tuning pre-trained models weighted on the ImageNet dataset was used in our proposed methodology. The experiment findings show that pre-trained Xception on the base model is effective, with an F1-score of 98% on the testing set. The Xception network performed the best for the fruits model, achieving an F1-score of 98% on the testing set. MobileNet was found to be the top model for the vegetables model, with an F1-score of 94% on the testing set. The VGG19 network performed best for the packages model, achieving an F1-score of 97% on the testing set. The experiment results could be relevant in applying the model obtained in mobile phones. People can benefit from learning without restrictions. An essential element is that consumers are more knowledgeable of the many types of grocery store products, particularly for the visually impaired. In future work, we will research and test alternative CNN architectures as well as other well-known one-stage detectors such as a YOLO or Siamese neural network to identify an appropriate one for classifying grocery store species.

Declarations

Conflicts of Interest Statement

No conflict of interest. The author declares that there is no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical Approval

No applicable

Availability of data and materials

Data will be available on request by contacting the corresponding author, Dr. Walid Dabour at

walid.dabour@science.menofia.edu.eg

References

[1] R. R.A. Bourne et al., "Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: The Right to Sight: An analysis for the Global Burden of Disease Study," Lancet Glob Health, vol. 9, pp. e144–e160, 2021. [Online]. Available: https://doi.org/10.1016/S2214-109X(20)30489-7.

[2] D. Bal, M.M. Islam Tusher, M. Rahman, M.S. Rahman Saymon, "NAVIX: A Wearable Navigation System for Visually Impaired Persons," in 2020 2nd International Conference on Sustainable Technologies for Industry 4.0, STI 2020, 2020. [Online]. Available: https://doi.org/10.1109/STI50764.2020.9350480.

[3] G. Vaidya, K. Vaidya, K. Bhosale, "Text recognition system for visually impaired using portable camera," in 2020 International Conference on Convergence to Digital World - Quo Vadis, ICCDW 2020, 2020. [Online]. Available: https://doi.org/10.1109/ICCDW45521.2020.9318706.

[4] L. P. Sousa, R.M.S. Veras, L.H.S. Vogado, L.S. Britto Neto, R.R.V. Silva, F.H.D. Araujo, F.N.S. Medeiros, "Banknote Identification Methodology for Visually Impaired People," in International Conference on Systems, Signals, and Image Processing, 2020-July, 2020, pp. 261–266. [Online]. Available: https://doi.org/10.1109/IWSSIP48289.2020.9145294.

[5] L. de Sousa Britto Neto, V.R.M.L. Maike, F.L. Koch, M.C.C. Baranauskas, A. de Rezende Rocha, S.K. Goldenstein, "A wearable face recognition system built into a smartwatch and the visually impaired user," in ICEIS 2015 - 17th International Conference on Enterprise Information Systems, Proceedings, vol. 3, 2015, pp. 5–12. [Online]. Available: https://doi.org/10.5220/0005370200050012.

[6] K. Weiss, T.M. Khoshgoftaar, D.D. Wang, "A survey of transfer learning," J Big Data, vol. 3, pp. 1–40, 2016. [Online]. Available: https://doi.org/10.1186/S40537-016-0043-6/TABLES/6.

[7] L. Hakobyan, J. Lumsden, D. O'Sullivan, H. Bartlett, "Mobile assistive technologies for the visually impaired," Surv Ophthalmol, vol. 58, pp. 513–528, 2013. [Online]. Available: https://doi.org/10.1016/J.SURVOPHTHAL.2012.10.004.

[8] M. Shamim Hossain, M. Al-Hammadi, G. Muhammad, "Automatic Fruit Classification Using Deep Learning for Industrial Applications," IEEE Trans Industr Inform, vol. 15, pp. 1027–1034, 2019. [Online]. Available: https://doi.org/10.1109/TII.2018.2875149.

[9] "A Systematic Review on Product Recognition for Aiding Visually Impaired People," IEEE Latin America Transactions. [Online]. Available: https://latamt.ieeer9.org/index.php/transactions/article/view/3760 (accessed April 11, 2022).

[10] J. Rivera-Rubio, S. Idrees, I. Alexiou, L. Hadjilucas, A.A. Bharath, "Small Hand-held Object Recognition Test (SHORT)," in 2014 IEEE Winter Conference on Applications of Computer Vision, WACV 2014, 2014, pp. 524–531. [Online]. Available: https://doi.org/10.1109/WACV.2014.6836057.

[11] G. Varol, R.S. Kuzu, "Toward retail product recognition on grocery shelves," in Sixth International Conference on Graphic and Image Processing (ICGIP 2014), vol. 9443, 2015, p. 944309. [Online]. Available: https://doi.org/10.1117/12.2179127.

[12] P. Jund, N. Abdo, A. Eitel, W. Burgard, "The Freiburg Groceries Dataset," 2016. [Online]. Available:

https://doi.org/10.48550/arxiv.1611.05799.

[13] M. Klasson, C. Zhang, H. Kjellström, "A Hierarchical Grocery Store Image Dataset with Visual and Semantic Labels," 2019. [Online]. Available: http://arxiv.org/abs/1901.00711 (accessed March 9, 2022).

[14] K. Simonyan, A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 2014. [Online]. Available: https://doi.org/10.48550/arxiv.1409.1556.

[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-December, 2016, pp. 2818–2826. [Online]. Available: https://doi.org/10.1109/CVPR.2016.308.

[16] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-December, 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90.

[17] A. Krizhevsky, I. Sutskever, G.E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," 2010, pp. 248–255. [Online]. Available: https://doi.org/10.1109/CVPR.2009.5206848.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, "Going deeper with convolutions," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 07-12-June-2015, 2015, pp. 1–9. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7298594.

[19] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520. [Online]. Available: https://doi.org/10.1109/CVPR.2018.00474.

[20] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017, pp. 1800–1807. [Online]. Available: https://doi.org/10.1109/CVPR.2017.195.

[21] G. Huang, Z. Liu, L. van der Maaten, K.Q. Weinberger, "Densely connected convolutional networks," in Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017, pp. 2261–2269. [Online]. Available: https://doi.org/10.1109/CVPR.2017.243.

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2010, pp. 248–255. [Online]. Available: https://doi.org/10.1109/CVPR.2009.5206848.

[23] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, "A Comprehensive Survey on Transfer Learning," Proceedings of the IEEE, vol. 109, pp. 43–76, 2021. [Online]. Available: https://doi.org/10.1109/JPROC.2020.3004555.

[24] L. Alzubaidi, J. Zhang, A.J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M.A. Fadhel, M. Al-Amidie, L. Farhan, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," J Big Data, vol. 8, 2021. [Online]. Available: https://doi.org/10.1186/s40537-021-00444-8.

[25] W. Rawat, Z. Wang, "Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review," Neural Comput, vol. 29, pp. 2352–2449, 2017. [Online]. Available: https://doi.org/10.1162/NECO_A_00990.

[26] K. Maharana, S. Mondal, B. Nemade, "A Review: Data Pre-Processing and Data Augmentation Techniques," Global Transitions Proceedings, 2022. [Online]. Available: https://doi.org/10.1016/J.GLTP.2022.04.020.

[27] A. Asperti, C. Mastronardo, "The Effectiveness of Data Augmentation for Detection of Gastrointestinal Diseases from Endoscopical Images," 2017. [Online]. Available: http://arxiv.org/abs/1712.03689 (accessed June 7, 2022).

[28] S. Ioffe, C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," 2015, pp. 448–456. [Online]. Available: https://proceedings.mlr.press/v37/ioffe15.html (accessed September 24, 2022).

[29] Y. Lecun, Y. Bengio, G. Hinton, "Deep learning," Nature 2015 521:7553, vol. 521, pp. 436–444, 2015. [Online]. Available: https://doi.org/10.1038/nature14539.

[30] E. Bisong, "Building Machine Learning and Deep Learning Models on Google Cloud Platform," Building Machine Learning and Deep Learning Models on Google Cloud Platform, 2019. [Online]. Available: https://doi.org/10.1007/978-1-4842-4470-8.

[31] Keras Team, "Getting started," Keras.io. [Online]. Available: https://keras.io/getting_started/. [Accessed: 24-Sep-2022].

[32] M.Z. Alom, T.M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M.S. Nasrin, B.C. van Esesn, A.A.S. Awwal, V.K. Asari, "The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches," Information Sciences Letters, vol. 6, no. 3, pp. 221-235, 2018. [Online]. Available: https://doi.org/10.48550/arxiv.1803.01164

[33] D.P. Kingma, L.J. Ba, "Adam: A Method for Stochastic Optimization," Information Sciences Letters, vol. 2, no. 1, pp. 45-56, 2015.

[34] B. Prabha, J. Thangakumar, K. Ramesh, "Reinforcement Learning Based Energy Consolidation Model for Efficient Cloud Computing System," Applied Mathematics & Information Sciences, vol. 17, no. 1, pp. 67-77, 2023. [Online]. Available: http://dx.doi.org/10.18576/amis/170109

[35] S. Saravanan, M. Sivabalakrishnan, N. Duraimurugan, D. Divya, "Artificial Intelligence Security Model For Privacy Renitence In Big Data Analytics," Applied Mathematics & Information Sciences, vol. 16, no. 6, pp. 919-927, 2022. [Online]. Available: http://dx.doi.org/10.18576/amis/160608

[36] H. H. El-Sayed, S.K. Refaay, S.A. Ali, M.T. El-Melegy, "Chain based Leader Selection using Neural Network in Wireless Sensor Networks protocols," Applied Mathematics & Information Sciences, vol. 16, no. 4, pp. 643-653, 2022. [Online]. Available: http://dx.doi.org/10.18576/amis/160418

[37] S. Aldossary, N. Noura, R. Zagrouba, "Authentication Solutions in Industrial Internet of Things: A Survey," Applied Mathematics & Information Sciences, vol. 17, no. 6, pp. 953-965, 2023. [Online]. Available: https://dx.doi.org/10.18576/amis/170602

[38] A. Alhaj, N.I. Zanoon, A. Alrabea, H.I. Alnatsheh, O. Jawabreh, M. Abu-Faraj, B.J.A. Ali, "Improving the Smart Cities Traffic Management Systems using VANETs and IoT Features," Journal of Statistical Applications & Probability, vol. 12, no. 2, pp. 405-414, 2023. [Online]. Available: http://dx.doi.org/10.18576/jsap/120207

[39] M. E. Karar, F. Alotaibi, A. Al Rasheed, O. Reyad, "A Pilot Study of Smart Agricultural Irrigation using Unmanned Aerial Vehicles and IoT-Based Cloud System," Information Sciences Letters, vol. 10, no. 1, pp. 131-140, 2021. [Online]. Available: http://dx.doi.org/10.18576/isl/100115

[40] M. Malkawi, Z. Al-Ghazawi, Z. Alshboul, A. Al-Yamani, "Internet of Things Based Monitoring System of Leaks in Water Supply Networks Using Pressure-Based Model," Information Sciences Letters, vol. 11, no. 2, pp. 495-500, 2022. [Online]. Available: http://dx.doi.org/10.18576/isl/110219

[41] R. Radhika, K. Kulothungan, "Mitigation of Distributed Denial of Service Attacks on the Internet of Things," Applied Mathematics & Information Sciences, vol. 13, no. 5, pp. 831-837, 2019. [Online]. Available: http://dx.doi.org/10.18576/amis/130517

[42] P. Varun, K. Ashokkumar, "Intrusion Detection System in Cloud Security using Deep Convolutional Network," Applied Mathematics & Information Sciences, vol. 16, no. 4, pp. 581-588, 2022. [Online]. Available: http://dx.doi.org/10.18576/amis/160411