Understanding the Black-box: Towards Interpretable and Reliable Deep Learning Models

Tehreem Qamar[1] and Narmeen Zakaria Bawany[1]

[1] Center for Computing Research, Department of Computer Science and Software Engineering, Jinnah university for Women, Karachi, Pakistan

Corresponding Author:

Tehreem Qamar[1]

Karachi, Sindh, 74600, Pakistan

Email address: tehreem.qamar@juw.edu.pk

Abstract

Deep learning (DL) has revolutionized the field of artificial intelligence by providing sophisticated models across a diverse range of applications, from image and speech recognition to natural language processing and autonomous driving. However, deep learning models are typically black-box models where the reason for predictions is unknown. Consequently, the reliability of the model becomes questionable in many circumstances. Explainable AI (XAI) plays an important role in improving the transparency and interpretability of the model thereby making it more reliable for real-time deployment. To investigate the reliability and truthfulness of DL models, this research develops image classification models using transfer learning mechanism and validates the results using XAI techniques. Thus, the contribution of this research is twofold, we employ three pre-trained models VGG16, MobileNetV2 and ResNet50 using multiple transfer learning techniques for a fruit classification task consisting of 131 classes. Next, we inspect the reliability of models, based on these pre-trained networks, by utilizing Local Interpretable Model-Agnostic Explanations (LIME), a popular XAI technique that generates explanations for the predictions. Experimental results reveal that transfer learning provides optimized results of around 98% accuracy. The classification of the models is validated on different instances using LIME and it was observed that each model predictions are interpretable and understandable as they are based on pertinent image features that are relevant to particular classes. We believe that this research gives an insight for determining how an interpretation can be drawn from a complex AI model such that its accountability and trustworthiness can be increased.

Introduction

Artificial Intelligence (AI) in the form of Deep Learning (DL) models [1] has gained significant advancement in recent years. We are increasingly dependent on artificial intelligence, more specifically on the deep learning for almost everything we do. From recommendations for shopping to self-driving cars, from loan approval to face detection, our lives are affected by these AI based systems. Deep Learning involves deep neural networks (DNNs) that learn in layers and resemble the human brain. They become the central technology and have been extensively utilized across various domains such as finance, medicine, natural language processing, cyber security, bioinformatics, robotics, etc. Unlike traditional machine learning models, DL models have the capability to automatically engineer features; therefore, there is no need for explicit feature extraction with human supervision. DL enables learning and classification in a single shot as they implicitly examine data and look for features that correlate and save days or even months of work of data scientists' and researchers' by identifying new, more complicated features that they would overlook [2]. Additionally, DL models are able to be trained on unstructured, unlabeled data and produce ample accuracy. On the other hand, DL models at the core are black-boxes i.e. their decisions are hidden in the thousands of simulated neurons, grouped into dozens or hundreds of highly interconnected layers. Further, the back-propagation method updates the computations made by individual neurons

58   so that the network may minimize the loss function thereby improving the performance of the model. Consequently,
59   the DL models become more complicated resulting in the high performing black-box models [3] [4]. Deep Learning
60   models are undergoing continuous development and recently surpassed human performance on tasks such as image
61   classification [5].
62   Image classification is one of the challenging and critical tasks in today's AI systems, implemented by
63   utilizing Convolutional Neural Network (CNN), the most popular DL model. CNN identifies visual patterns from
64   raw pixels. A CNN is often trained using methods like back propagation and gradient descent and has many layers
65   of activations and convolutions dispersed among pooling layers [6]. Training a CNN model requires huge amount of
66   data, computation time and processing power [7]. Moreover, they are designed for solving a single specific task and
67   have to rebuild from scratch once the feature space distribution changes. To overcome this isolated learning
68   paradigm and utilizing the knowledge acquired from training for a single task, CNNs have advanced the idea of
69   transfer learning: an optimization technique employing pre-trained models. A pre-trained model is a saved network
70   that has previously undergone extensive training on a large dataset, typically for a large image classification task.
71   They usually make use of ImageNet [8] and can be utilized directly in making predictions on new tasks or integrated
72   into the process of training a new model. By using pre-trained models, both, the training time and generalization
73   error are reduced.
74   Despite the remarkable success of deep learning, the lack of transparency and interpretability in their
75   models has become a major concern among stakeholders. The deep learning models achieve a high level of
76   accuracy, but being a black-box model it is almost impossible to identify the key features that led to the decision. In
77   order to build trust and ensure accountability of the DL based systems, that are widely being deployed, there is a
78   need for justification of decisions taken by models [9]. Thus, there is a growing demand for understanding and
79   interpreting the decisions undertaken by these black-box models. The need to verify the reliability of these networks
80   is more emphasized recently after several instances were reported of AI systems making decisions that resulted in
81   devastating effects. Fatal accident by Uber self-driving car [10], facial recognition systems evasion with a 3D
82   printed mask [11], gender biasing in Amazon Recruitment tool [12], bigotry in Correctional Offender Management
83   Profiling for Alternative Sanctions (COMPAS) [13] are some of the distressing automated decisions that accentuate
84   the need for explanations of AI system.
85   Thus, the motivation of this study comes from the fact that deep learning models created from complex pre-
86   trained models must be interpretable. Therefore, there is a need for a study in which a model is derived using a pre-
87   trained model as a base model and interpreting the modified model predictions using Explainable AI (XAI) tools.
88   Consequently, the hypothesis formulated for the research is "Complex deep learning models can be interpreted using
89   Explainable AI techniques." In order to validate this hypothesis, we intend to develop deep learning models for fruit
90   classification task and evaluate interpretability of the model using LIME. Fruit classification is a critical task in
91   many industries such as agriculture, food processing and retail. Accurate fruit classification can help improve the
92   efficiency of processes such as sorting, grading, and packaging, ultimately reducing waste and increasing
93   profitability. Numerous studies have proposed deep learning-based model for fruit classification [14] [15] [16] but
94   many of these models lack explainability, making it difficult to understand how the models arrived at their
95   classification decisions. This can be a significant barrier to the adoption of such models in practical applications,
96   particularly in industries such as agriculture or food processing. However, explaining the models making automated
97   fruit classification systems will improve transparency, identify biases, and meet regulatory requirements that will
98   eventually help users in trusting the model's output and increase their confidence in the system [17].
99   For the stated problem, we have used a popular dataset Fruits 360 [18, p. 360] that comprises 90K images
100 of fruits and vegetables with 131 different classes that makes it an ideal use case for our study. We have utilized
101 three pre-trained models VGG16, MobileNetV2, and ResNet50 as base network for three different classification
102 models specifically the type of fruit or vegetable. We fine-tuned and trained these models with Fruits 360 datasets to
103 get an accuracy of around 98%. To ensure the transparency and reliability of our models we performed experiments
104 by incorporating the widely used explainable AI tool; LIME.
105 The major contributions of this research are summarized below:

106 • We use transfer learning with three pre-trained models, VGG16, MobileNetV2, ResNet50 for classifying
107   fruits and vegetables using complete Fruits 360 dataset comprising 131 classes.
108 • We employ two transfer learning techniques and evaluate the performance of pre-trained models.
109 • We compare the effectiveness of our models derived from pre-trained models using various evaluation
110   metrics.
111 • We present a detailed explanation of our models using popular explainable AI tool LIME.

112   The remaining sections of the paper are structured as follows: Section 2 presents the background and literature
113 review. Section 3 explains the methodology adopted; Section 4 presents the details of experiments performed.
114 Section 5 discusses the results while Section 6 concludes the research with future directions.

115

Background and Related Work

117 Image classification has been around for decades. The conventional image classification models need large, labeled
118 dataset, hand-crafted features, huge computation resources, and enormous time for training [19]. This high cost of
119 developing image classification models compromises their robustness. In contrast, deep learning models exploit
120 multiple non-linear layers for feature extraction and classification, improve the efficiency of image classification
121 task and have achieved astonishing results [20]. Among several deep learning models, Convolutional Neural
122 Network (CNN) [21] has become the leading architecture for image classification and detection tasks such as facial
123 recognition [22], [23], medical image computing [24], [25], plant disease classification [26], remote sensed image
124 analysis [27].
125   Though CNN is the most popular deep learning model, its performance is largely dependent upon the volume of
126 data. CNN requires a sizeable amount of data that needs plentiful computation power and time for training.
127 Moreover, designing the CNN architecture from the scratch is exhausting as it needs a lot of time and effort in
128 finding suitable combination of layers and adjusted hyper-parameters. Further, the model learns for one specific task
129 on a very specific dataset and has to rebuilt if the features-space changes [19], [28]. To overcome these challenges,
130 researchers discovered the concept of transfer learning that allows transferring of knowledge gained by one task to
131 solve similar another task [29]. It reduces the cost of training as the model is readily trained for identifying low level
132 features and the last layer classifies the set of classes that were used during training. The following section discusses
133 the use of transfer learning in various domains and the popular CNN architectures used in this study.

134

Transfer Learning

136 Deep learning approaches are typically dependent on the dataset used to train the network. A large volume of
137 labeled data is needed to train a network to achieve desirable performance. Gathering a massive amount of labeled
138 data for a particular domain is not only exhausting but also quite challenging in most real-world applications such as
139 medical imaging [30]. As a result, the idea of transfer learning has been introduced. Transfer learning allows
140 utilizing the knowledge applied for solving one problem to resolve other relevant problems. The base network
141 (commonly referred to as pre-trained network) is initially trained on a large dataset and transfers its learning
142 parameters and weights to the target network. The last fully connected layer of the target network is then trained on
143 its respective dataset. The pre-trained network can also be fine-tuned by retraining some of its layers to further
144 increase performance. Transfer learning has been used widely in all machine learning applications, such as computer
145 vision [31], natural language processing [32], and speech recognition [33], and it has demonstrated excellent results
146 in terms of accuracy, training duration, and error rates.
147   In image classification, fine tuning a pre-trained model entails bootstrapping the top portion of the model,
148 freezing the pre-trained convolutional layers and un-freezing the last few pre-trained layers. The frozen layers
149 convolve visual features as usual while the un-frozen layers get trained on the custom dataset and updated according
150 to the Fully Connected Layer's predictions. ImageNet dataset is used for training for these pre-trained models as it
151 encompasses around one million images belonging to 1000 categories. Various CNN architectures have been

152 developed as pre-trained models for image classification tasks; however, this research employs three most popular
153 architectures variant VGGNet, MobileNet, and ResNet. The following section briefly introduces all three CNN
154 architecture used in this research.

155

VGGNet

156
157 VGGNet [34] is a convolutional neural network with 2 variants i.e. VGG16 and VGG19. VGG16 has 16 layers
158 including 13 convolution and 3 fully connected layers, while VGG19 has 16 convolution layers and 3 Fully
159 connected layers supported by MaxPool layers. It is one of the prominent architectures in image classification.
160 Rezende et al. [35] used VGG16 to classify malware family by converting malware executable to a byteplot
161 grayscale image and achieved 92.97% accuracy without any explicit feature engineering. Taranjit and Tapan [36]
162 applies similar methodology and classify MRI images as normal or abnormal with different neurological diseases by
163 using VGG16 acquiring 100% accuracy. Comparison of different pre-trained models (VGG, ResNet, DenseNet,
164 MobileNet, Inception, Xception) was presented by Himabindu and Kumar [37]. They evaluated each model on
165 accuracy, precision, f1-score, recall and reported that VGG outperforms all other models with 97% accuracy.

166

MobileNets

167
168 MobileNets [38] are convolutional neural networks designed by Google researchers. This CNN architecture is
169 popular for its adoptability on mobile phones as it has a low resource requirement. The MobileNet architecture is
170 developed using depthwise separable convolutions, which are lightweight deep neural networks that can have
171 minimal latency for embedded and mobile devices. Rabano et al. [39] uses MobileNet for classifying trash in an
172 android application. Shahi et al. [40] presented an attention-based MobileNetV2 architecture for fruit classification
173 and evaluate it against accuracy, f1score, Kappa-score, WAFI, MAFI, recall and precision. The authors compared
174 their architecture with other pre-trained models on three different datasets and their proposed framework surpassed
175 all with achieving more than 95% accuracy on all three datasets.

176

ResNet

177
178 ResNet [41] is a very deep residual network built by Microsoft and has a depth of 50 layers. ResNet combined
179 multiple sized convolution filters which manage the degradation problem and reduces the training time that occurs
180 due to its deep structures. Sarwinda et al. [42] proposed an image classification model for detection of colorectal
181 cancer in colon glands images using different variants of ResNet and found that ResNet50 provides the most reliable
182 performance for accuracy, sensitivity, and specificity. Precision classification for breast cancer histopathological
183 image was investigated by Jiang et al. [43] in which they proposed a customized version of ResNet. An accuracy of
184 99% was reported by the authors after the network was fined tuned.
185   The CNN architectures trained on large datasets have addressed the two major issues regarding the training
186 of deep learning networks. These pre-trained models have reduced the requirement of voluminous data and the need
187 for extensive computing environment to some extent. However, another critical concern with developing real-life
188 deep learning systems is the need for an explainable, interpretable, and transparent solutions. Deep learning models
189 remain black-box models and there is a growing demand for the explanation of their learning and prediction process.
190 The following section discusses the new research paradigm known as Explainable AI that is came into being to
191 provide explanations of black-box models predictions.

192

Explainable AI (XAI)

193
194 Deep neural networks are regarded as black-box models by both developers and users since they are comparatively
195 weaker in explaining their inference process and final decisions [44]. Explainable AI is a collection of methods and
196 techniques that allows end users to understand and trust the decisions AI systems make [45]. It has gained
197 significant attention recently among both industry and research community due to the fact that AI is now involved in
198 such real-world applications that demands explainability and transparency, for example, medical diagnosis,

199   investment recommendation, loan approvals, surveillance, autonomous vehicles, predictions for process
200   optimization etc. Evaluation metrics such as high accuracy may not be sufficient to ensure that the decisions taken
201   by these models are always correct, justified and without any bias. For example, COMPAS; an assistive tool used in
202   multiple states of the US to assess the likelihood that a criminal offender would reoffend has been proven to be
203   discriminatory, with results heavily biased towards white defendants [13]. Understanding the reasons behind the
204   decisions taken by autonomous models leads to more reliable and trustworthy systems. Therefore, the goal of XAI is
205   to make the reasoning behind the decision taken by AI systems that is understandable by humans [46]. A variety of
206   XAI tools have been introduced to explain the predictions made by AI systems, some of them are given below:

207   **LIME**
208   LIME [47] is an acronym of Local Interpretable Model-Agnostic Explanations. It helps users understand why a
209   machine learning model made a certain prediction by providing an explanation in terms of the most relevant features
210   that influenced the prediction. To create an explanation, LIME generates a local, interpretable model around the
211   instance being explained and weighs the contribution of each feature to the model's output. These features are then
212   presented as an explanation to the user, in order to help them understand how the complex machine learning model
213   arrived at the prediction. The explanations generated by LIME are intended to provide insights into the decision-
214   making process of the black-box machine learning model, and to enable users to validate and understand the model's
215   predictions.
216
217   LIME is one of the most popular XAI tools that has the capability to provide explanations for all machine learning
218   models. Many studies have been carried out that have used LIME for explaining the results irrespective of the nature
219   of data [48] [49] [50]. In this research, we have used LIME for evaluating the interpretability of three pre-trained
220   models i.e. VGG16, MobileNetV2 and ResNet50 on Fruits 360 dataset.
221

Methodology
223   The methodology applied in this research encompasses two phases, that is; development of three classification models
224   using transfer learning and explanation of these models using LIME.
225
226   (**Figure 1.** DL Architecture used in this study utilizing VGG16 Pre-trained model)

227   In the first phase, three classification models are created using pre-trained models (VGG16, MobileNetV2
228   and ResNet50) as the base model. The features extracted from the base model are used by the new layers introduced
229   in each model. However, the softmax layer is used as the last layer by initializing the number of neurons to total
230   number of classes in Fruits 360 dataset i.e. 131. Figure 1, Figure 2 and Figure 3 depict the architecture of the
231   classification models used in this study. The classification is performed by applying two transfer learning
232   techniques, (a) using pre-trained models with frozen layers and (b) using pre-trained models with fine-tuned layers.
233   Each classification model is evaluated through accuracy, precision, recall and f1-score.
234
235   (**Figure 2.** DL Architecture used in this study utilizing MobileNetV2 Pre-trained model)

236   (**Figure 3.** DL Architecture used in this study utilizing ResNet50 Pre-trained model)

237   In the second phase, the classification models are evaluated for their truthfulness. To accomplish this, five
238   specific examples or instances are studied using LIME. The purpose of using LIME is to understand the underlying
239   features that contribute to each decision made by the classification models. This helps in identifying what aspects of
240   the data are driving the models' predictions thereby making it easier to understand the decisions taken by DL
241   models.
242

Experiments

We analyzed the interpretability and understandability of three pre-trained deep learning models (VGG16, MobileNetV2 and ResNet50) by conducting experiments on fruit classification problem. We first developed fruit classification models using the pre-trained models and then employed explainable AI tool on classification models. For fruit classification problem, the Fruits 360 dataset has become a benchmark and has been utilized in various research studies. Ghosh et al. [51] used its 41 classes for detecting and classifying fruits using ShuffleNetV2. Similarly, Raheel Siddiqi [52] has used 101 classes of Fruits 360 dataset and presented a comparative analysis on the performance of various deep learning models. Furthermore, Sakib et al. [53] proposed a fruit recognition classifier by utilizing 17,823 images belonging to 25 categories. Rathnayake et al. [54] have used the complete Fruits 360 dataset for image identification and recognition but similar to aforementioned studies they also have not taken the interpretability or explainability factor for their proposed models.

In this study, we have used a complete Fruits 360 dataset having 131 classes of fruits and vegetables images that is publicly available on Kaggle and developed classification models followed by their explanation. The dataset holds 90,483 images in total and each image has one fruit or vegetable. The size of each image is 100x100 and is captured with white background. Fig. 4 shows some of the images from the Fruits 360 dataset.

(**Figure 4**. Sample Images from Fruits 360 Dataset)

The experiments were performed on the dataset using Keras API of tensorflow employing three pre-trained architectures, VGG16, MobileNetV2 and ResNet50. Two techniques of transfer learning, frozen and fine-tuned, were applied during the course of experiments to achieve the best possible results. In the frozen case, the pre-trained models convolve the data according to the ImageNet weights leaving their top portion, that is fully connected layer of the model, and transfer fixed features to the customized fully connected layer. While in fine tuning, the pre-trained models bootstrap the top portion, freeze the pre-trained convolutional layers and un-freeze the last few pre-trained layers. The frozen layers convolve visual features as usual while the un-frozen layers get trained on the custom dataset and updated according to the last fully connected layer. We used complete Fruits 360 dataset with 131 classes containing around 90K images; divided into training, validation, and test sets with a ratio of 65:10:35 respectively. Thus, the training set comprises 57,612 images; validation set contains 10,080 images while 22,688 images are part of the test set. Torres et al. [55] and Azarmdela et al. [56] have utilized the similar distribution of dataset for fruit image processing. All images in each set are pre-processed using the pre-processing function defined in the Keras Library for the respective pre-trained model. All experiments were carried out on a machine having an Intel Core i5-1135G7 @ 2.40GHz processor with 32GB RAM and 2 GB NVIDIA GeForce MX350 GPU. The experiments performed in this study are presented in Table 1.

(**Table 1.** Experiments performed in the Study)

For frozen layer case, the training and validation batch size used is 64. The models are evaluated by training them iteratively with varying epochs that is 10 and 20 with training step size of 900 and validation step size of 157. The Adam optimizer is used from the Keras Optimizers and the initial learning rate was set to 10^{-3}. Early stopping is applied in the validation loss with the patience value of 10 that is the model stops training if the validation loss keeps increasing till 10 successive epochs. For fine-tune case, the last two layers of each mode left un-freeze and the models are trained for 10 epochs with initial learning rate of 10^{-2}. The hyperparameters used for the development of classification models are summarized in Table 2.

286

287   **(Table 2.** Hyperparameters used for development of Classification Models)

288 Results and Discussion

289 This section describes the results in terms of phases as mentioned in the methodology section. In the first phase, we
290 generate classification models by utilizing three pre-trained models (VGG16, MobileNetV2, ResNet50) using two
291 transfer learning techniques. While in the second phase, LIME interpretations are generated that highlight the key
292 features used in the respective prediction. Table 3 reflects how our research objectives have been achieved with our
293 corresponding research contribution.
294

295   **(Table 3.** Research Objectives to Research Contribution mapping**)**

296

297 Phase I - Development of Classification Models

298 The two techniques of transfer learning are frozen layers and fine-tuned layers as explained in the background and
299 related work section. We have developed classification models by incorporating three pre-trained models VGG16,
300 MobileNetV2 and ResNet50 whose results are described in successive sections.
301

302 Frozen Layers

303 In frozen layers technique, where each pre-trained model served as a feature extractor with all layers freeze,
304 ResNet50 outperformed VGG16 and MobileNetV2 models by achieving 97.3% test accuracy in 10 epochs as shown
305 in Fig. 5.

306

307   **(Figure 5.** Results of Frozen Layers technique of Transfer Learning)

308 The ResNet50 also outclasses the two models in terms of execution time as it takes comparatively less time to
309 achieve highest accuracy. It is observed that VGG16 and ResNet50 start overfitting when set to train for 20 epochs
310 while MobileNetV2 shows improvement by acquiring 100% training accuracy and 93.62% test accuracy. The results
311 of experiments achieved by each model are presented in Table 4.
312

313   **(Table 4.** Classification models with Frozen Layers)

314 Fine-tuned Layers

315 In fine-tuned technique, all models were trained for 10 epochs on two variations of unfreeze layers. First, all models
316 are trained by tuning just one layer whose results. Then, the models were trained by tuning their two layers in which
317 all models got 100% training accuracy. The overall experiment results are presented in Table 5.

318   **(Table 5.** Classification models with Fine-tuned Layers)

319 VGG16 and MobileNetV2 show relatively improved performance by achieving 97.09% and 92.25% test accuracy
320 respectively after unfreezing their last two layers. However, ResNet50 dominance is witnessed again on VGG16 and
321 MobileNetV2 as it attains test accuracy of 98.08% by unfreezing one last layer only as depicted in Fig. 6.
322

323
324   **(Figure 6.** Results of Fine-tuned Layers technique of Transfer Learning)

325 It is evident from analyzing the experimental results that pre-trained models offer greater accuracy in comparatively
326 smaller training time. ResNet50 exhibits highest accuracy in both cases while VGG16 and MobileNetV2 have also
327 shown greater than 90% accuracy which is acceptable in classification problems. The highest accuracy produced
328 version of each pre-trained model is used for the investigation of their interpretation capability using LIME.
329

330 **Phase II - Interpretation using LIME**

331 LIME produces instance level explanations, therefore; five correctly predicted instances were selected for
332 interpretations' study. Table 6 presents the instances and their top 5 predictions made by each model.
333
334   **(Table 6.** Instances evaluated for LIME Interpretations)

335 LIME works by creating perturbed images of the instance being predicted; therefore, 150 perturbed images were
336 generated by turning on and off the super pixels of the instance as depicted in Fig. 7. Super pixels are used in LIME
337 to help generate explanations for image classification models by simplifying the image and reducing the number of
338 features used to explain the model's prediction. Instead of considering every individual pixel in an image, LIME
339 groups pixels into super pixels and treats each super pixel as a feature. The left of the Figure shows the image with
340 all super pixels on while the right of the image shows some of the perturbations generated by the ResNet50 model.
341 Similar perturbed images were generated by MobileNet and ResNet.
342
343   **(Figure 7.** Perturbed images)

344 Each perturbed image is then predicted and the distance between the perturbed image and the original image is
345 calculated. Cosine metric was used to find the distance with kernel size of 0.25. Finally, the linear regression model
346 is fitted to find out the top feature in the model prediction. Figure 8 shows the top feature selected in green color by
347 the models to make prediction of the instance chosen by each model. It is clearly evidenced that VGG16 and
348 ResNet50 select the most suitable position of the image to be predicted as banana. Whereas MobileNetV2's feature
349 selection is abstruse.
350
351   **(Figure 8.** Top Feature selected by each classification model for the chosen instance)

352 Table 7 shows the top feature selected for the rest of the instances declared in Table 6 by each model. Nevertheless, it
353 is apparent that the predictions of the pre-trained model are explainable, interpretable and can be trusted as the top
354 feature chosen by each model is suitable for the respective data instance. Hence, our hypothesis that complex deep
355 learning models can be interpreted using explainable AI techniques has been successfully validated by these findings.
356
357   **(Table 7.** Top feature selection for the prediction for each instance)

358

359 **Conclusions**
360 In this study, our objective is to investigate the interpretability of deep learning models due to their black-box nature
361 that leads to a lack of transparency and interpretability in their decision-making process. Although these models
362 have shown exceptional accuracy, the absence of justification raises concerns related to trust, accountability, biases,
363 and transparency. Therefore, we aim to address this issue by exploring the interpretability of deep learning models
364 using explainable AI techniques. We assessed the truthfulness of the deep learning models specifically pre-trained
365 models by generating interpretations of their predictions. To do this, we performed experiments in two settings.
366 First, we have utilized three pre-trained models (VGG16, MobileNetV2 and ResNet50) with two techniques of

367 transfer learning. Second, we produced interpretations of these models using LIME. Extensive experiments have
368 been carried out to obtain the classification model with highest accuracy. ResNet50 has outperformed VGG16 and
369 MobileNetV2 by attaining 98.08% accuracy with one last layer unfreeze. While VGG16 and MobileNetV2 have
370 shown significant performances with 97.09% and 93.62% accuracy respectively. Few instances that are correctly
371 predicted by the models are selected to evaluate the interpretability of each pre-trained model. LIME explanations
372 are generated that mark the top feature selection in the underlying prediction and it is observed that all pre-trained
373 models used in this research are interpretable.

374   As future perspective, we intent to extend our research on the interpretability of deep learning models into
375 more critical and sensitive applications such as healthcare and finance. The interpretation of these models is
376 essential to ensure transparency and enhance trust in these domains. Additionally, we aim to investigate the root
377 cause of any misclassifications made by the models using various explainable AI techniques. This will allow us to
378 identify the areas that need improvement and develop strategies to reduce misclassifications in classification models.
379 By doing so, we can improve the performance and accuracy of deep learning models, making them more reliable
380 and trustworthy in critical applications.
381

References

[1] Huang Yi, Sun Shiyu, Duan Xiusheng, Chen Zhigang, "A study on deep neural networks framework," in *2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, IEEE, 2016, pp. 1519–1522.

[2] Yann Lecun, Yoshua Bengio, Geoffrey Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.

[3] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, Jianfeng Gao, "Deep Learning-Based Text Classification," *ACM Computing Surveys*, vol. 54, no. 3, Jun. 2021, doi: 10.1145/3439726.

[4] Fred Hohman, Minsuk Kahng, Robert Pienta, Duen Horng Chau, "Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers," *IEEE Trans. Visual. Comput. Graphics*, vol. 25, no. 8, pp. 2674–2693, Aug. 2019, doi: 10.1109/TVCG.2018.2843369.

[5] Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, J Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, Laith Farhan, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, no. 1, p. 53, 2021, doi: 10.1186/s40537-021-00444-8.

[6] Waseem Rawat, Zenghui Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural Computation*, vol. 29, no. 9, pp. 2352–2449, Sep. 2017, doi: 10.1162/NECO_a_00990.

[7] Risto Miikkulainen, Jason Liang, Elliot Meyerson, Aditya Rawal, Daniel Fink, Olivier Francon, Bala Raju, Hormoz Shahrzad, Arshak Navruzyan, Nigel Duffy, others, "Evolving deep neural networks," in *Artificial intelligence in the age of neural networks and brain computing*, Elsevier, 2019, pp. 293–312.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.

[9] Wojciech Samek, Thomas Wiegand, Klaus-Robert Müller, "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models." arXiv, Aug. 28, 2017. Accessed: May 07, 2023. [Online]. Available: http://arxiv.org/abs/1708.08296

[10] Puneet Kohli, Anjali Chadha, "Enabling pedestrian safety using computer vision techniques: A case study of the 2018 uber inc. self-driving car crash," in *Future of Information and Communication Conference*, Springer, 2019, pp. 261–279.

415   [11] Aaron Holmes, "Facial Recognition Fooled at Airport Using Masks, Researchers Found,"
416   Feb. 07, 2020. https://www.businessinsider.com/facial-recognition-fooled-with-mask-
417   kneron-tests-2019-12 (accessed Jun. 16, 2022).
418   [12] Jeffrey Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women,"
419   in *Ethics of Data and Analytics*, Auerbach Publications, 2018, pp. 296–299.
420   [13] Christina Wadsworth, Francesca Vera, Chris Piech, "Achieving fairness through
421   adversarial learning: an application to recidivism prediction," *arXiv preprint
422   arXiv:1807.00199*, 2018.
423   [14] Khurram Hameed, Douglas Chai, Alexander Rassau, "A comprehensive review of fruit and
424   vegetable classification techniques," *Image and Vision Computing*, vol. 80, pp. 24–44,
425   2018, doi: 10.1016/j.imavis.2018.09.016.
426   [15] Mehenag Khatun, Forhad Ali, Nakib Aman Turzo, Julker Nine, Pritom Sarker, "Fruits
427   Classification using Convolutional Neural Network," *GRD Journals-Global Research and
428   Development Journal for Engineering*, vol. 5, no. 8, 2020.
429   [16] Achanta Jyothi Prakash, P Prakasam, "An intelligent fruits classification in precision
430   agriculture using bilinear pooling convolutional neural networks," *The Visual Computer*, pp.
431   1–17, 2022.
432   [17] Arun Rai, "Explainable AI: from black box to glass box," *J. of the Acad. Mark. Sci.*, vol. 48,
433   no. 1, pp. 137–141, Jan. 2020, doi: 10.1007/s11747-019-00710-5.
434   [18] Mihai Oltean, *Fruits 360 dataset: new research directions*. 2021.
435   [19] Chitra Desai, "Image Classification Using Transfer Learning and Deep Learning,"
436   *International Journal of Engineering and Computer Science*, vol. 10, Jul. 2021, doi:
437   10.18535/ijecs/v10i9.4622.
438   [20] Tom Vermeire, Dieter Brughmans, Sofie Goethals, Raphael Mazzine Barbossa de Oliveira,
439   David Martens, "Explainable image classification with evidence counterfactual," *Pattern
440   Analysis and Applications*, vol. 25, no. 2, pp. 315–335, 2022, doi: 10.1007/s10044-021-
441   01055-y.
442   [21] Keiron O'Shea, Ryan Nash, "An Introduction to Convolutional Neural Networks," no.
443   December, 2015, [Online]. Available: http://arxiv.org/abs/1511.08458
444   [22] Suleman Khan, M Hammad Javed, Ehtasham Ahmed, Syed A A Shah, Syed Umaid Ali,
445   "Facial recognition using convolutional neural networks and implementation on smart
446   glasses," in *2019 International Conference on Information Science and Communication
447   Technology (ICISCT)*, 2019, pp. 1–6.
448   [23] Shraddha Mane, Gauri Shah, "Facial recognition, expression recognition, and gender
449   identification," in *Data management, analytics and innovation*, Springer, 2019, pp. 275–
450   290.
451   [24] Qishuo Gao, Samsung Lim, Xiuping Jia, "Hyperspectral image classification using
452   convolutional neural networks and multiple feature learning," *Remote Sensing*, vol. 10, no.
453   2, p. 299, 2018.
454   [25] Jianpeng Zhang, Yutong Xie, Qi Wu, Yong Xia, "Medical image classification using
455   synergic deep learning," *Medical image analysis*, vol. 54, pp. 10–19, 2019.
456   [26] Jinzhu Lu, Lijuan Tan, Huanyu Jiang, "Review on convolutional neural network (CNN)
457   applied to plant leaf disease classification," *Agriculture*, vol. 11, no. 8, p. 707, 2021.
458   [27] Lei Ma, Yu Liu, Xueliang Zhang, Yuanxin Ye, Gaofei Yin, Brian Alan Johnson, "Deep
459   learning in remote sensing applications: A meta-analysis and review," *ISPRS journal of
460   photogrammetry and remote sensing*, vol. 152, pp. 166–177, 2019.
461   [28] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui
462   Xiong, Qing He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*,
463   vol. 109, no. 1, pp. 43–76, 2020.

464 [29] Lisa Torrey, Jude Shavlik, "Transfer learning," in *Handbook of research on machine*
465   *learning applications and trends: algorithms, methods, and techniques*, IGI global, 2010,
466   pp. 242–264.
467 [30] Awwal Muhammad Dawud, Kamil Yurtkan, Huseyin Oztoprak, "Application of Deep
468   Learning in Neuroradiology: Brain Haemorrhage Classification Using Transfer Learning,"
469   2019, doi: 10.1155/2019/4629859.
470 [31] Xuhong Li, Yves Grandvalet, Franck Davoine, Jingchun Cheng, Yin Cui, Hang Zhang,
471   Serge Belongie, Yi Hsuan Tsai, Ming Hsuan Yang, "Transfer learning in computer vision
472   tasks: Remember where you come from," *Image and Vision Computing*, vol. 93, p. 103853,
473   Jan. 2020, doi: 10.1016/J.IMAVIS.2019.103853.
474 [32] Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, Thomas Wolf, "Transfer
475   Learning in Natural Language Processing," *Proceedings of the 2019 Conference of the*
476   *North*, pp. 15–18, 2019, doi: 10.18653/V1/N19-5004.
477 [33] Chu Xiong Qin, Dan Qu, Lian Hai Zhang, "Towards end-to-end speech recognition with
478   transfer learning," *Eurasip Journal on Audio, Speech, and Music Processing*, vol. 2018, no.
479   1, pp. 1–9, Dec. 2018, doi: 10.1186/S13636-018-0141-9/TABLES/4.
480 [34] Karen Simonyan, Andrew Zisserman, "Very deep convolutional networks for large-scale
481   image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
482 [35] Edmar Rezende, Guilherme Ruppert, Tiago Carvalho, Antonio Theophilo, Fabio Ramos,
483   Paulo de Geus, "Malicious software classification using VGG16 deep neural network's
484   bottleneck features," in *Information technology-new generations*, Springer, 2018, pp. 51–
485   59.
486 [36] Taranjit Kaur, Tapan Kumar Gandhi, "Automated brain image classification based on VGG-
487   16 and transfer learning," in *2019 International Conference on Information Technology*
488   *(ICIT)*, 2019, pp. 94–98.
489 [37] D Dakshayani Himabindu, S Praveen Kumar, "A Comprehensive Analytic Scheme for
490   Classification of Novel Models," in *2020 3rd International Conference on Intelligent*
491   *Sustainable Systems (ICISS)*, IEEE, 2020, pp. 564–569.
492 [38] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, Liang-Chieh Chen,
493   "MobileNetV2: Inverted Residuals and Linear Bottlenecks." pp. 4510–4520, 2018.
494 [39] Stephenn L Rabano, Melvin K Cabatuan, Edwin Sybingco, Elmer P Dadios, Edwin J
495   Calilung, "Common garbage classification using mobilenet," in *2018 IEEE 10th*
496   *International Conference on Humanoid, Nanotechnology, Information Technology,*
497   *Communication and Control, Environment and Management (HNICEM)*, 2018, pp. 1–4.
498 [40] Tej Bahadur Shahi, Chiranjibi Sitaula, Arjun Neupane, William Guo, "Fruit classification
499   using attention-based MobileNetV2 for industrial applications," *PLoS ONE*, vol. 17, no. 2
500   February, pp. 1–21, 2022, doi: 10.1371/journal.pone.0264586.
501 [41] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Deep residual learning for image
502   recognition," in *Proceedings of the IEEE conference on computer vision and pattern*
503   *recognition*, 2016, pp. 770–778.
504 [42] Devvi Sarwinda, Radifa Hilya Paradisa, Alhadi Bustamam, Pinkie Anggia, "Deep Learning
505   in Image Classification using Residual Network (ResNet) Variants for Detection of
506   Colorectal Cancer," *Procedia Computer Science*, vol. 179, pp. 423–431, Jan. 2021, doi:
507   10.1016/J.PROCS.2021.01.025.
508 [43] Yun Jiang, Li Chen, Hai Zhang, Xiao Xiao, "Breast cancer histopathological image
509   classification using convolutional neural networks with small SE-ResNet module," *PloS*
510   *one*, vol. 14, no. 3, p. e0214587, 2019.
511 [44] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, Jun Zhu, "Explainable
512   AI: A brief survey on history, research areas, approaches and challenges," in *CCF*
513   *international conference on natural language processing and Chinese computing*,
514   Springer, 2019, pp. 563–574.

515 [45] Andreas Holzinger, Anna Saranti, Christoph Molnar, Przemyslaw Biecek, Wojciech Samek,
516 "Explainable AI methods-a brief overview," in *International Workshop on Extending*
517 *Explainable AI Beyond Deep Models and Classifiers*, Springer, 2022, pp. 13–38.
518 [46] Judea Pearl, "The limitations of opaque learning machines," *Possible minds: twenty-five*
519 *ways of looking at AI*, pp. 13–19, 2019.
520 [47] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, "'why should i trust you?' explaining
521 the predictions of any classifier," in *NAACL-HLT 2016 - 2016 Conference of the North*
522 *American Chapter of the Association for Computational Linguistics: Human Language*
523 *Technologies, Proceedings of the Demonstrations Session*, 2016, pp. 97–101. doi:
524 10.18653/v1/n16-3020.
525 [48] Subrata Bhattacharjee, Yeong Byn Hwang, Kobiljon Ikromjanov, Rashadul Islam Sumon,
526 Hee Cheol Kim, Heung Kook Choi, "An Explainable Computer Vision in Histopathology:
527 Techniques for Interpreting Black Box Model," *4th International Conference on Artificial*
528 *Intelligence in Information and Communication, ICAIIC 2022 - Proceedings*, pp. 392–398,
529 2022, doi: 10.1109/ICAIIC54071.2022.9722656.
530 [49] Mingzhe Zhu, Bo Zang, Linlin Ding, Tao Lei, Zhenpeng Feng, Jingyuan Fan, "LIME-Based
531 Data Selection Method for SAR Images Generation Using GAN," *Remote Sensing 2022,*
532 *Vol. 14, Page 204*, vol. 14, no. 1, p. 204, Jan. 2022, doi: 10.3390/RS14010204.
533 [50] Nicholas Hamilton, Adam Webb, Matt Wilder, Ben Hendrickson, Matt Blanck, Erin Nelson,
534 Wiley Roemer, Timothy C. Havens, "Enhancing Visualization and Explainability of
535 Computer Vision Models with Local Interpretable Model-Agnostic Explanations (LIME),"
536 *2022 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 604–611, Dec.
537 2022, doi: 10.1109/SSCI51031.2022.10022096.
538 [51] Sourodip Ghosh, Md. Jashim Mondal, Sourish Sen, Soham Chatterjee, Nilanjan Kar Roy,
539 Suprava Patnaik, "A novel approach to detect and classify fruits using ShuffleNet V2," in
540 *2020 IEEE Applied Signal Processing Conference (ASPCON)*, Oct. 2020, pp. 163–167.
541 doi: 10.1109/ASPCON49795.2020.9276669.
542 [52] "Effectiveness of Transfer Learning and Fine Tuning in Automated Fruit Image
543 Classification | Proceedings of the 2019 3rd International Conference on Deep Learning
544 Technologies." https://dl.acm.org/doi/abs/10.1145/3342999.3343002 (accessed May 07,
545 2023).
546 [53] Shadman Sakib, Zahidun Ashrafi, Md Siddique, Abu Bakr, "Implementation of fruits
547 recognition classifier using convolutional neural network algorithm for observation of
548 accuracies for various hidden layers," *arXiv preprint arXiv:1904.00783*, 2019.
549 [54] Namal Rathnayake, Upaka Rathnayake, Tuan Linh Dang, Yukinobu Hoshino, "An Efficient
550 Automatic Fruit-360 Image Identification and Recognition Using a Novel Modified
551 Cascaded-ANFIS Algorithm," *Sensors*, vol. 22, no. 12, Art. no. 12, Jan. 2022, doi:
552 10.3390/s22124401.
553 [55] José Naranjo-Torres, Marco Mora, Ruber Hernández-García, Ricardo J. Barrientos,
554 Claudio Fredes, Andres Valenzuela, "A Review of Convolutional Neural Network Applied to
555 Fruit Image Processing," *Applied Sciences*, vol. 10, no. 10, Art. no. 10, Jan. 2020, doi:
556 10.3390/app10103443.
557 [56] Hossein Azarmdel, Ahmad Jahanbakhshi, Seyed Saeid Mohtasebi, Alfredo Rosado
558 Muñoz, "Evaluation of image processing technique as an expert system in mulberry fruit
559 grading based on ripeness level using artificial neural networks (ANNs) and support vector
560 machine (SVM)," *Postharvest Biology and Technology*, vol. 166, p. 111201, Aug. 2020,
561 doi: 10.1016/j.postharvbio.2020.111201.
562

Experiments performed in the Study

1   **Table 1.** Experiments performed in the Study

Experiments	Results
Phase I – Generation of Classification Models	
Classification models development using pre-trained models with Frozen Layers	Table 3
Classification models development by fine tuning the pre-trained models	Table 4
Phase II - Interpretation of Classification Models	
Instances evaluated for LIME interpretations and their top 5 predictions	Table 5
Top features selected by classification models for the chosen instances	Table 6

Hyperparameters used for development of Classification Models

1   **Table 2. Hyperparameters used for development of Classification Models**

Hyperparameter	Value
Epochs	10, 20
Batch Size (Training and Validation)	64
Training step size	900
Validation step size	157
Learning Rate	10^{-2} , 10^{-3}
Optimizer	Adam
Regularization	Early Stopping

Research Objectives to Research Contribution mapping

1   **Table 3. Research Objectives to Research Contribution mapping**

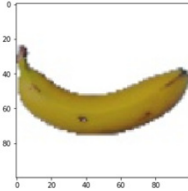
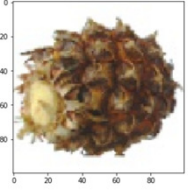
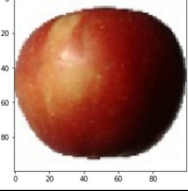
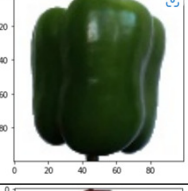
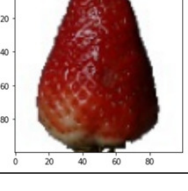
Research Objectives	Contributions
To develop fruit classification models using pre-trained models	Three pre-trained models namely VGG16, ResNet50 and MobileNetV2 are utilized for the development of fruit classification models. See Figure 5 and Figure 6
To explore transfer learning mechanism in the development of fruit classification models	Two transfer learning mechanisms 1) frozen layers and 2) fine-tuned layers are employed in the development of classification models. Refer to Section Phase I - Development of Classification Models for further details
To compare performance of fruit classification models using different metrics	The classification models employing pre-trained models are evaluated using accuracy, precision, recall and f1 score. See Table 4 and Table 5
To interpret fruit classification models using explainable AI tool	Explainable AI tool LIME is used to interpret the results of pre-trained models based classification models. See Table 7 and Figure 8

Classification models with Frozen Layers

Table 4. Classification models with Frozen Layers

Parameters	VGG16		MobileNetV2		ResNet50	
Epochs	10	20	10	20	10	20
Execution Time (Seconds)	4517	5251	1438	3203	3101	4133
Early Stopped?	No	Yes, at Epoch 13	No	No	No	Yes, at Epoch 14
Training Accuracy (%)	99.8	99.9	98.2	99.99	99.8	99.9
Validation Accuracy (%)	97.4	97.6	87.6	98.98	99	99.2
Testing Accuracy (%)	94.03	94.1	80.84	95.20	97.3	96.43
Precision	0.95	0.96	0.83	0.95	0.97	0.97
Recall	0.94	0.94	0.8	0.95	0.97	0.96
F1-Score	0.94	0.94	0.8	0.95	0.97	0.96

Classification models with Fine-tuned Layers

1   **Table 5. Classification models with Fine-tuned Layers**

Parameters	1-layer Unfreeze			2-layers Unfreeze		
	VGG16	**MobileNetV2**	**ResNet50**	**VGG16**	**MobileNetV2**	**ResNet50**
Execution Time (Seconds)	3742	1337	3012	4020	1398	3128
Training Accuracy (%)	99.8	99.6	100	100	99.8	100
Validation Accuracy (%)	97.9	96.6	99.9	99.1	96.1	99.8
Testing Accuracy (%)	96.12	91.49	98.08	97.09	92.25	97.56
Precision	0.97	0.92	0.98	0.97	0.93	0.98
Recall	0.96	0.92	0.98	0.97	0.93	0.98
F1-Score	0.96	0.91	0.98	0.97	0.92	0.98

Instances evaluated for LIME Interpretations

1 **Table 6. Instances evaluated for LIME Interpretations**

Instance Chosen	Top 5 Predictions		
	VGG16	MobileNetV2	ResNet50
	Banana Peach 2 Physalis Banana Red Tomato Maroon	Banana Kaki Banana Red Carambula Banana Lady Finger	Banana Banana Lady Finger Carambula Lemon Cucumber Ripe 2
	Pineapple Mini Tomato Yellow Tomato Maroon Apple Red Yellow 1 Huckleberry	Pineapple Mini Physalis with Husk Mulberry Rambutan Pitahaya Red	Pineapple Mini Pineapple Mangostan Mulberry Rambutan
	Apple Braeburn Apple Red 2 Apple Red 3 Tomato Yellow Apricot	Apple Braeburn Nectarine Apple Red 2 Apple Pink Lady Tamarillo	Apple Braeburn Apricot Potato Red Washed Nut Forest Cherry Wax Yellow
	Pepper Green Tomato not Ripened Watermelon Eggplant Grape Pink	Pepper Green Apple Red Yellow 2 Tomato not Ripened Tomato Heart Pepper Orange	Pepper Green Pepper Red Eggplant Dates Tomato Heart
	Strawberry Cucumber Ripe Grape Pink Raspberry Dates	Strawberry Mandarine Nectarine Lemon Strawberry Wedge	Strawberry Cucumber Ripe Avocado ripe Strawberry Wedge Carambula

Top feature selection for the prediction for each instance

1   **Table 7. Top feature selection for the prediction for each instance**

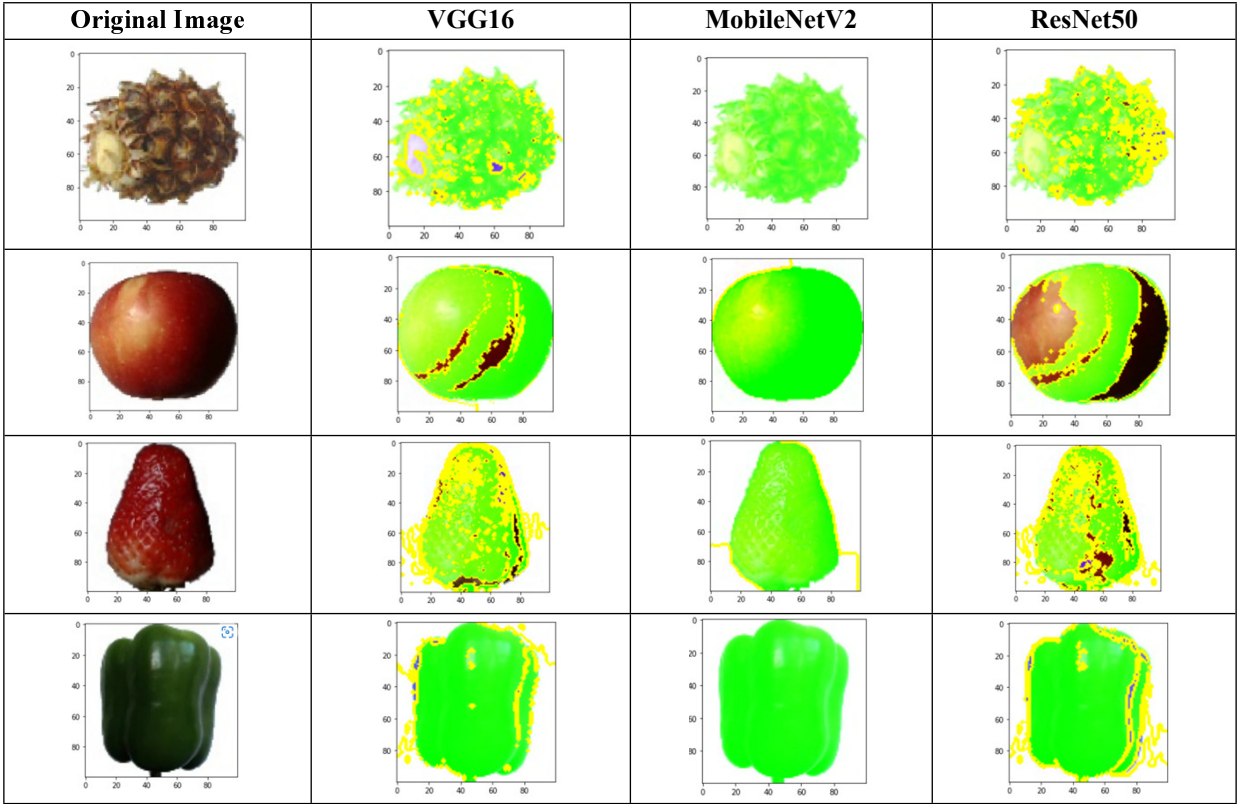



Figure 1

DL Architecture used in this study utilizing VGG16 Pre-trained model

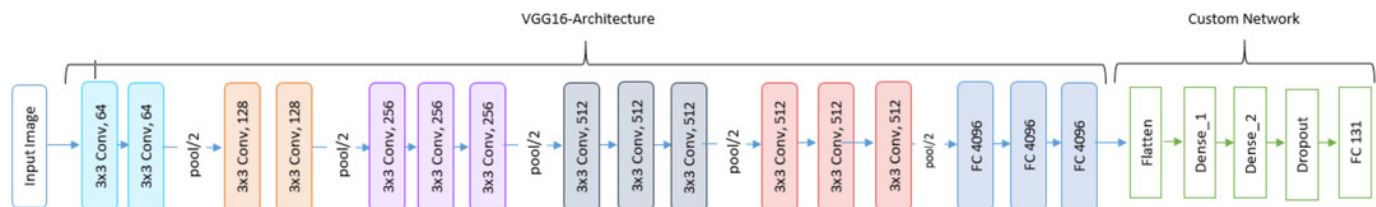



Figure 2

DL Architecture used in this study utilizing MobileNetV2 Pre-trained model

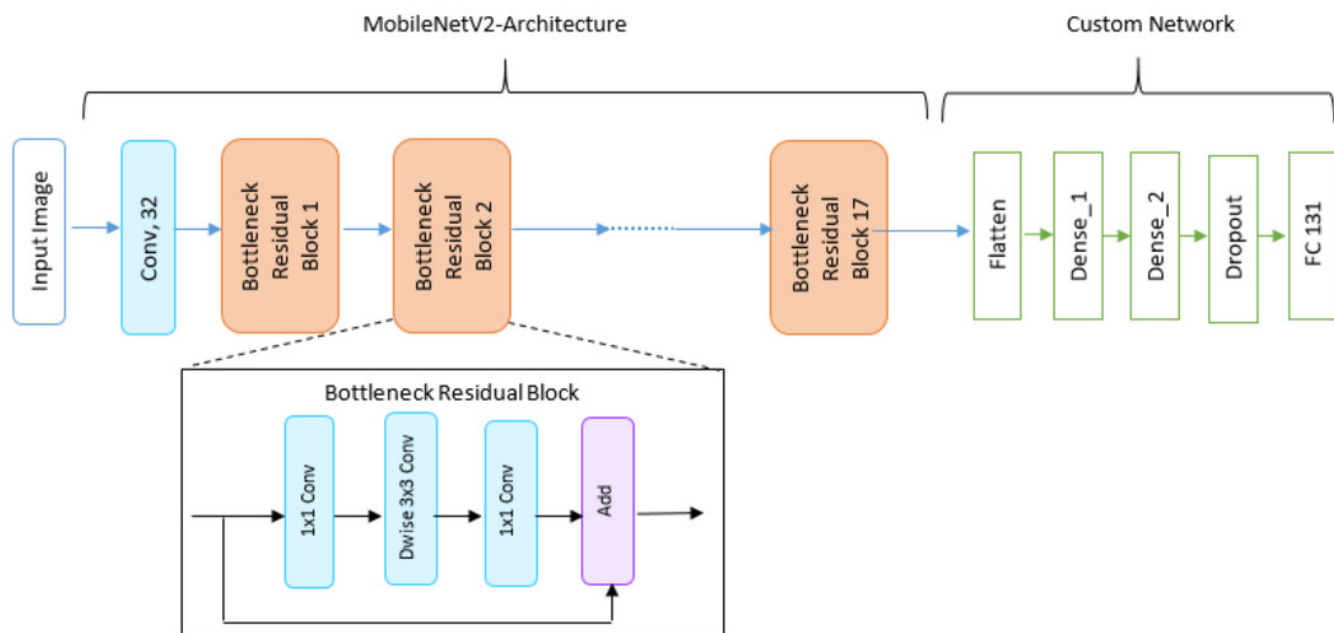



Figure 3

DL Architecture used in this study utilizing ResNet50 Pre-trained model

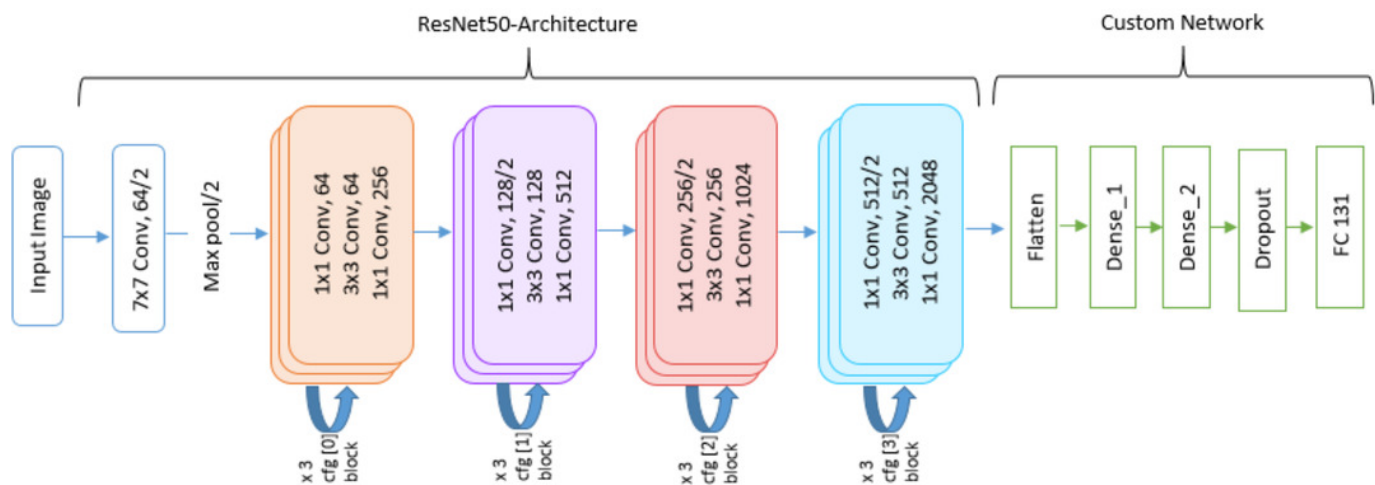



Figure 4

Sample Images from Fruits 360 Dataset

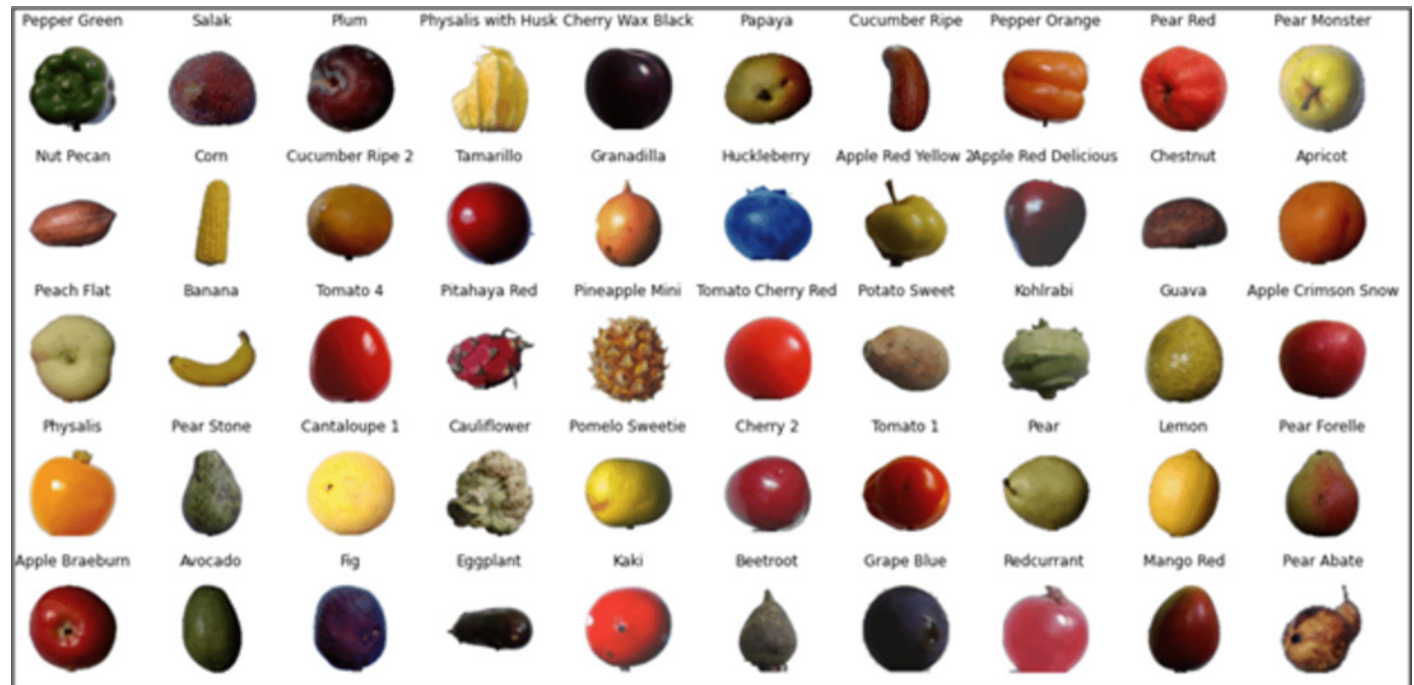



Figure 5

Results of Frozen Layers technique of Transfer Learning

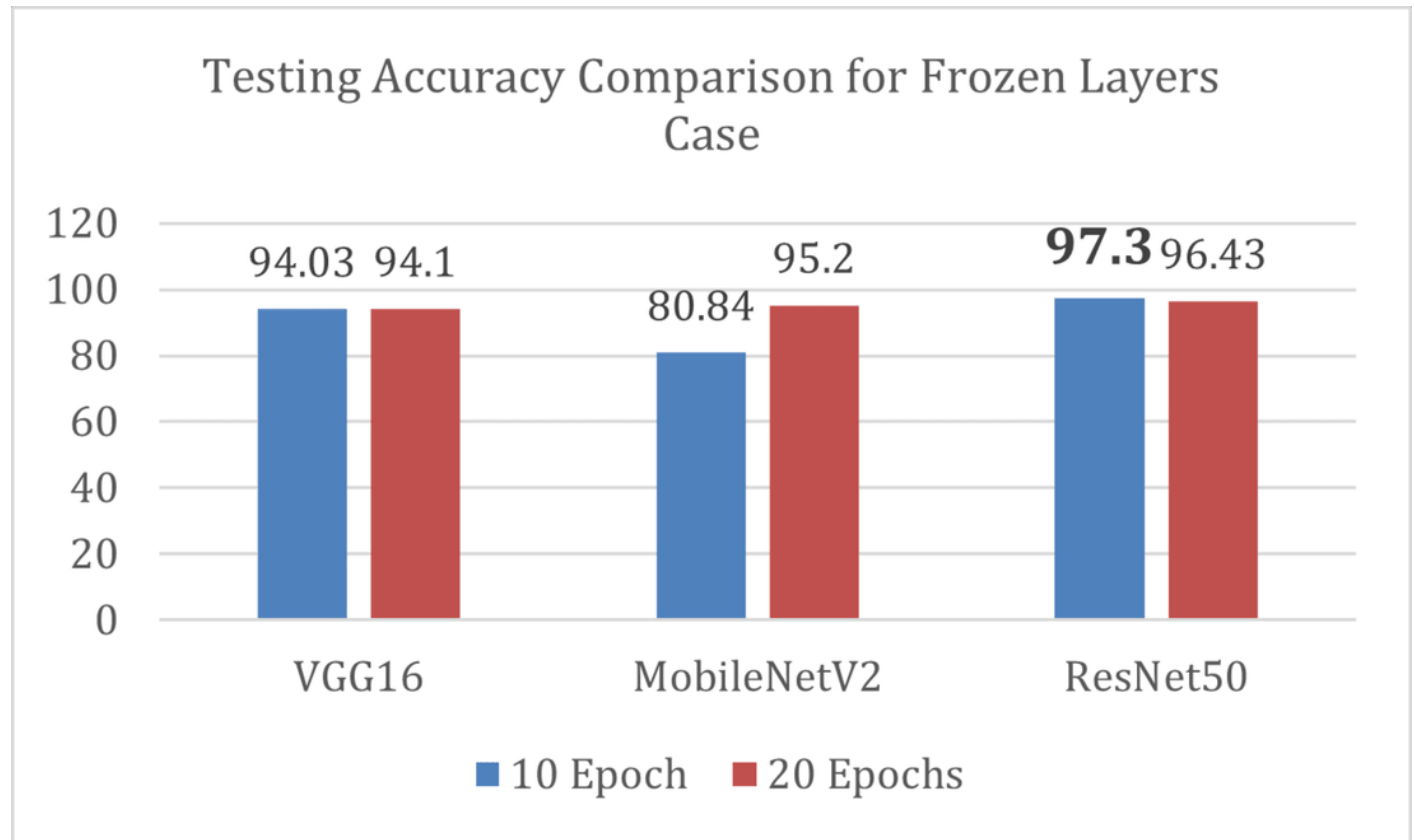



Figure 6

Results of Fine-tuned Layers technique of Transfer Learning

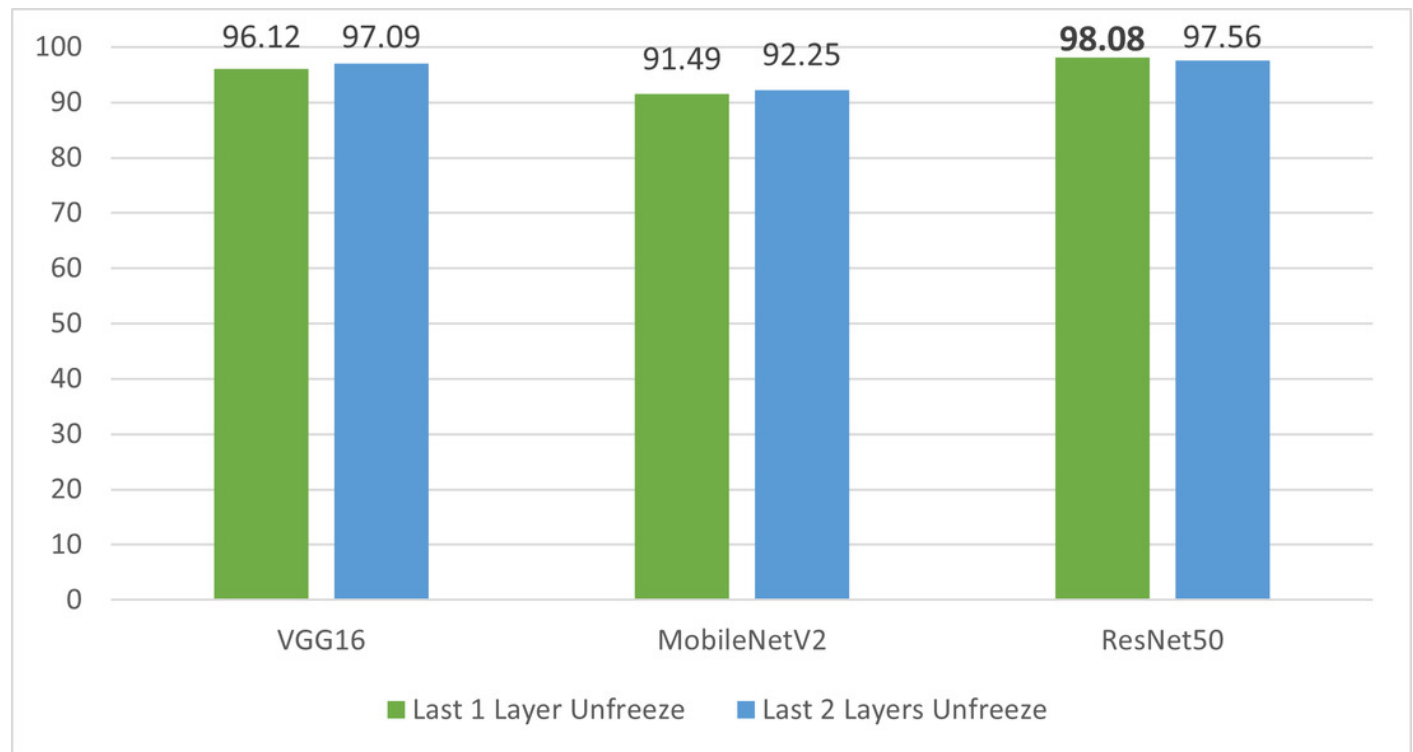



Figure 7

Perturbed images

Figure 8

Top Feature selected by each classification model for the chosen instance