Germinated Oil Palm Seed Quality Classification Using Transfer Learning with GoogleNet

Anshali Manoharan†, Carolina Lee Pei Qian†, Anshana Manoharan†

Faculty of Science and Engineering, University of Nottingham Malaysia, Jalan Broga, Semenyih, 43500, Selangor, Malaysia.

Contributing authors: hcyam5@nottingham.edu.my; hpcq5@nottingham.edu.my; hcyam4@nottingham.edu.my;
†These authors contributed equally to this work.

Abstract

The classification of oil palm seed germination is important for improving agricultural yield and resource efficiency. To address this, this study aims to apply transfer learning on GoogLeNet for automatic classification of germinated seeds. In this work, we assess GoogLeNet's feature extractor and classifier on various data transformations including translation, rotation, scaling, noise, and illumination changes. We design an offline augmentation strategy and generate a combined dataset in accordance to the weakness of GoogLeNet's performance on the mentioned transformations. We then finetune GoogLeNet and evaluate its performance under the same data transformations. Key findings show that the model's accuracy increases by 12% on unseen test batches with varying light conditions, reaches early convergence, and evaluation shows that GoogLeNet's features become invariant to the targeted transforms. These results suggest that the proposed approach can improve the generalisation ability of GoogLeNet in quality classification of germinating oil palm seeds.

Keywords: germinated oil palm seeds, quality classification, data augmentation, deep learning, GoogLeNet, transfer learning

1 Introduction

Elaeis guineensis, commonly known as the oil palm, is one of the world's leading sources of vegetable oil, playing a vital role in global agriculture. The lipids extracted from its mesocarp are widely used across numerous industrial sectors, with demand rising over the years [1].

With Malaysia, Thailand, and Indonesia accounting for over 88% of global palm oil production [2], the industry remains heavily dependent on manual labour, particularly for the careful gleaning of high-quality oil palm seeds after germination. At present, palm seed classification is predominantly carried out through manual visual assessment, where workers evaluate the germination quality of each seed. This approach is not only labour-intensive and costly but also demands a substantial workforce and significant time investment [3]. Furthermore, the rising labour shortages [2] accentuate the need to develop and adopt mechanised solutions to maintain productivity and ensure sustainability.

Automated seed classification presents a promising and scalable solution to improve efficiency and reduce reliance on manual labour. In line with this, the aim of this research is to develop a classifier capable of performing binary quality classification for germinated oil palm seeds.

To address the limitations of manual seed classification and improve model generalisability under varying conditions, this study is executed as follows. Firstly, a dataset analysis is performed. Next, a deep learning model to be used in this study is introduced with justifications provided for its selection. Next, the study then investigates the robustness of this model's feature extractor and classifier to common spatial and photometric transformations - including translation, rotation, scaling, illumination changes, and noise - through feature similarity analysis and classification consistency. Insights from this robustness evaluation guide the development of an augmentation strategy. Finally, the proposed approach is benchmarked using standard evaluation metrics, with a particular emphasis on its ability to generalise to unseen datasets collected under different imaging and lighting conditions.

This paper is organised as follows. Section 2 describes the dataset, including its composition and characteristics. Section 3 discusses transfer learning, while Section 4 covers the GoogLeNet model architecture [4], justification and performance of the model in other applications. Section 5 describes the set-up for the baseline model and analysis and modifications are discussed in Section 6. The proposed approach is executed in Section 7 and its effectiveness is discussed. Finally, Section 8 concludes the paper with a summary of findings and directions for future work.

2 Dataset

The dataset used is a private collection provided by Applied Agricultural Resources Sdn. Bhd., Malaysia [3], consisting of three batches of germinated oil palm seed images.

Batch-1 is split into training and test sets, each containing two class directories: GoodSeed and BadSeed. The test subset in Batch-1 is to be used for validation. Batch-2 and Batch-3 serve as test sets, captured under normal room lighting and lightbox conditions, respectively.

Table 1 Dataset image count per class and imbalance ratio.

Dataset	Good Seed	Bad Seed	Imbalance Ratio
Batch-1 (Train)	901	851	1.06
Batch-1 (Test)	201	200	1.0
Batch-2 (Normal Room Light)	450	450	1.0
Batch-3 (LightBox)	605	563	1.02

Table 1. summarises the class-wise image distribution and imbalance ratio for each batch. All batches maintain low-to-negligible class imbalance, supporting fair training and testing processes. The inclusion of data captured under varying lighting conditions (lightbox vs. natural room light) allows models to be evaluated on domain shifts. This setup guarantees a fair assessment of the model's generalisation ability across varying illumination scenarios.

3 Transfer Learning

In the field of deep learning, it is a well-established practice to transfer knowledge acquired from one task to accelerate learning in another domain. This approach, termed transfer learning, is particularly advantageous, when training from scratch would be computationally expensive or infeasible due to limited data availability [5].

Rather than initialising all model parameters randomly, transfer learning utilises the weights of models previously trained on large, diverse datasets such as ImageNet [6], MS-COCO [7], and Open Images Dataset V4 [8]. In several contexts [9], [10], the early convolutional layers of a pre-trained model—responsible for capturing universal low-level features such as edges, textures, and colour gradients—are typically preserved, based on the assumption that these features remain relevant for the new domain. Higher-level layers, encoding more task-specific representations, are then fine-tuned to better suit the new task [11].

However, for specialised classification problems, full fine-tuning (where all layers are updated starting from pre-trained weights) has shown to yield superior performance [12], [13]. This is because moving from a generalised feature space (e.g. ImageNet) to a highly specialised domain often requires the network to re-learn low-level features specific to the new data. Accordingly, we choose to unfreeze all weights in our model due to the fine-grained distinctions between the classes in our dataset, where both seed types exhibit highly similar visual traits. Unlike broader object classification tasks, the distinction between GoodSeed and BadSeed categories

depends on minute variations such as texture difference, and localised germination signs, which cannot be captured by retaining only high-level generic features through training the fully connected layers.

Although this strategy is more computationally intensive compared to partially freezing prior layers, it is essential for capturing the nuanced characteristics of the dataset, ensuring better predictive performance and improved generalisation. To this end, we employ GoogLeNet [4], originally trained on the ImageNet dataset, as the backbone network for supporting the classification of germinated palm seeds into GoodSeed and BadSeed categories.

4 Rationale for Model Selection: GoogLeNet

Our choice of utilising a CNN model is supported by existing research in computer vision, where learned CNN features have proven to be highly task-specific and can out-perform traditional handcrafted features, particularly in seed classification [14]. In this work, we adopt the GoogLeNet architecture [4], originally developed for the ILSVRC 2014 competition by Szegedy et. al.

4.1 Applications of GoogLeNet in Agriculture

GoogLeNet and its derivatives have been successfully applied in various biological and agricultural contexts. For instance, Luo et al. used GoogLeNet for classifying weed seeds, achieving high classification accuracy of 94.61% [15]. Similarly, Hanafi et al. reported the effectiveness of GoogLeNet in the classification of rice seed growth and rice seed classification [16], a task analogous in complexity to palm seed germination classification. In addition, in a study classifying sunflower seeds [17], GoogLeNet obtained the best performance with 95% classification accuracy. Apart from seed classification tasks, GoogLeNet has been used in other agriculture based applications such as detecting and classifying plant diseases in tomato plants with 99.7% of AUC [18]. These precedents validate the model's suitability for biological object recognition tasks, particularly where high visual similarity exists between classes.

4.2 Architecture of GoogLeNet

This architecture is a deep CNN built around the Inception module [4], designed to optimise computational efficiency and model accuracy. The network is 22 layers deep when counting only layers with parameters, and 27 layers deep with pooling layers included. The input to GoogLeNet consists of RGB images resized to 224×224 with zero mean normalisation. Every convolutional layer, including those in Inception modules, uses rectified linear unit (ReLU) activation functions. To maintain computational efficiency, 1×1 convolutions are extensively used for dimensionality reduction before the more expensive 3×3 and 5×5 convolutions. Similarly, 1×1 convolutions follow pooling layers for projection. A key architectural choice in GoogLeNet is the use of average pooling before the final classification layer instead of fully connected layers (see Figure 1). This approach reduces the number of parameters and helps

overcome overfitting, while dropout remains crucial for regularisation. To address

Fig. 1 Architecture of GoogLeNet. Source: Original paper [4].

challenges with gradient flow in this deep network, Szegedy et. al. introduced auxiliary classifiers connected to intermediate layers after Inception modules 4a and 4d. These auxiliary classifiers act as regularisers during training, providing additional supervision to earlier layers and helping combat the vanishing gradient problem. Each auxiliary branch consists of an average pooling layer, a 1×1 convolution, a fully connected layer with ReLU activation, dropout, and a final SoftMax classification layer. Their losses are weighted with a discount factor of 0.3 and added to the main

loss during training but discarded at inference time.

GoogLeNet reduces computational cost while preserving essential spatial information using Inception modules, it maintains a relatively low parameter count (approximately 5 million parameters) compared to other models like AlexNet and VGGNet [19] . This enables faster training and inference with a reduced risk of overfitting. This is especially beneficial given the limited data scenarios often encountered in agricultural datasets [20].

4.3 GoogLeNet's Feature Extractor

When evaluating the suitability of GoogLeNet as a feature extractor, it is important to consider its robustness to common image variations, including scaling, translation, rotation, changes in illumination, and the presence of Gaussian noise. GoogLeNet introduced the Inception architecture, which uses parallel convolutional filters of different sizes (1×1, 3×3, 5×5) within the same layer. By capturing fine-grained and coarse features simultaneously, this design proves effective for tasks that rely on multi-scale feature representations. In the context of germinated palm seed classification, where key features vary in scale (e.g., the plumule, radicle, seed texture), this multi-scale capability is advantageous over traditional single-kernel architectures [21].

Regarding translational invariance in GoogLeNet's features, Myburgh et. al. [22] demonstrate that standard CNNs derive much of their translational invariance from fully connected layers, rather than the features itself. Our own empirical testing on GoogLeNet, detailed in the following sections, supports these observations. While GoogLeNet does not possess explicit rotational invariance, a study [23] has shown that GoogLeNet can learn to maintain relatively strong robustness to rotated inputs compared to other CNN architectures. It is demonstrated that applying rotation-based augmentation can improve CNN recognition performance, with GoogLeNet demonstrating superior performance.

In a study by Erik Rodner et al., the sensitivity of CNNs, including GoogLeNet, to image transformations and noise was analysed. The findings revealed that GoogLeNet's performance was affected by small intensity noise, leading to drops in classification accuracy. This suggests that GoogLeNet is not inherently robust to Gaussian noise without specific training or modifications [24]. Regarding illumination changes, GoogLeNet has been proven to be sensitive to brightness variations by Hu et al. [25].

4.4 GoogLeNet's Classifier

The GoogLeNet classifier was chosen based on existing literature where the complete GoogLeNet architecture, including its classification head, has been employed for end-to-end transfer learning. Al-Huseiny et. al. trained the GoogLeNet architecture end-to-end on a lung cancer dataset which resulted in 94.38% accuracy [26]. In addition, Barman et. al. fine-tuned the GoogLeNet model to classify skin cancer

and achieved the highest training and testing accuracy of 91.16% and 89.93% [27]. Another study involving lung cancer resulted with GoogLeNet performing 2% better than AlexNet [28].

These studies have shown that retaining GoogLeNet's original classifier yields high accuracy without requiring additional modifications, particularly in domains where distinguishing features are subtle. This evidence supports the use of GoogLeNet's built-in classifier as a reliable choice for the classification of germinated oil palm seeds as it has proven to perform well when fine-tuned in specialised scenarios. It involves softmax cross-entropy loss for its main classifier head, and it obtained strong performance across various image classification tasks. Additionally, it is observed that GoogLeNet's transformational invariance is derived from the classifier (as seen in the above section), therefore the feature extractor and classifier are jointly optimised for the task.

In summary, GoogLeNet's multi-scale feature extraction, built-in classifier and computational efficiency justifies our choice for the classification of germinated palm seeds. Its success in prior biological imaging tasks supports its applicability in the domain explored in this study. Our evaluations which are detailed in the following sections, provide insight into its performance under these conditions for germinating oil palm seed classification.

5 Baseline Methodology

This section shows the results of the performance of GoogLeNet with a default training set-up. This will serve as a baseline for comparison to our proposed modifications. Implementation details of GoogLeNet as the baseline model are presented.

5.1 Implementation Details

For the baseline experiment, the GoogLeNet architecture pre-trained on ImageNet was fine-tuned with all layers unfrozen. The auxiliary classifiers from the original GoogLeNet architecture were retained to preserve model structure during inference, with the optimisation focused solely on the main classifier.

Although GoogLeNet was originally designed for 224 $\times$ 224 input images, in this study the images were resized to 299 $\times$ 299. Empirically, this adjustment resulted in better classification performance. The reason is that seed classification relies on capturing very minute differences, such as slight variations in texture, surface structure, and localised patterns. A higher input resolution preserves these spatial details, allowing the network to learn these determinant features more effectively.

The final fully connected layers were modified for binary classification. Training and validation were performed on Batch-1 of the dataset, with images resized to 299$\times$299 pixels and normalised using dataset-specific mean and standard deviation values computed beforehand. Data was loaded using a batch size of 32. The Adam

optimiser was used with a default learning rate of 0.001, and the loss function was categorical cross-entropy. Early stopping with a patience of 5 epochs was employed to prevent overfitting. The best-performing model based on validation accuracy was saved for final evaluation. Testing was conducted on Batch-2 and Batch-3, and performance was assessed using accuracy, precision, recall, F1-score, AUC, and confusion matrix analysis.

5.2 Performance of Baseline Model

The following performance metrics show how well GoogLeNet performed with our baseline training set-up. Table 2 shows the performance on the train and test batches.

Table 2 Performance Metrics for Baseline Model.

Dataset	Accuracy	Precision	Recall	F1 Score	AUC Score
Batch-1 (Validation set)	0.953	0.917	0.995	0.955	0.995
Batch-2 (Normal Room Light)	0.752	0.671	0.989	0.800	0.875
Batch-3 (LightBox)	0.732	0.661	0.962	0.784	0.829

Below in Figure 2. is the confusion matrices for the train and test batches.

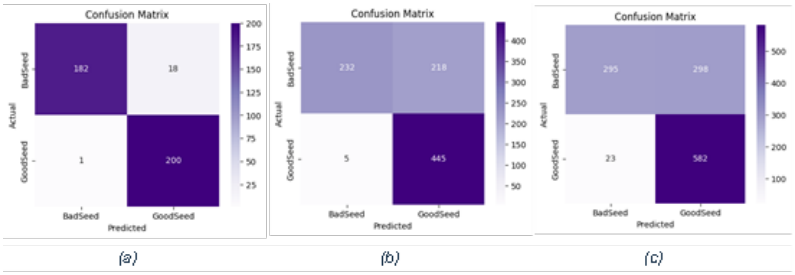

Fig. 2 Confusion matrices for (a) the validation batch, (b) Batch-2, (c) Batch-3

5.3 Discussion

The baseline retrained GoogLeNet model demonstrated strong performance on the validation set (Batch 1), achieving an accuracy of 95.3%, a high precision of 91.7%, and an exceptionally high recall of 99.5%, alongside an AUC score of 0.995. The near-perfect recall indicates that the model correctly identified almost all "GoodSeed" instances in the validation set, minimizing false negatives. This is critical for agricultural applications, where missing viable seeds (false negatives) could lead to wasted resources. However, the precision of 91.7% suggests that some "BadSeed" instances were misclassified as good, which could result in lower-quality seeds advancing to

cultivation.

When evaluated on unseen test sets (Batch 2 and Batch 3), captured under different lighting conditions, the model maintained high recall (98.9% on Batch 2 and 96.2% on Batch 3), reaffirming its sensitivity to detecting good seeds. However, precision dropped significantly (67.1% on Batch 2 and 66.1% on Batch 3), indicating a rise in false positives (bad seeds labeled as good). This trade-off suggests that while the model is highly adept at finding good seeds, its ability to exclude bad seeds degrades under domain shifts.

Overall, the model struggles with generalisation to the test batches, and these findings accentuate the need for further analysis of the robustness of the model to various input transforms to assess what can be included to improve its robustness. Class activation maps (CAMs) will also be shown to verify whether the model's attention remains focused on the biologically meaningful regions (the plumule and radicle) especially in cases of misclassification. In the following section, the baseline model is further analysed in this context, including an examination of its feature extractor and classifier components to identify potential weaknesses.

6 Baseline Model Analysis

This section will identify and analyse the strengths and weaknesses of GoogLeNet's feature extractor and classifier using the experimental set-ups described below, and the results are analysed.

6.1 Robustness of the Feature Extractor to Input Variations

To assess the robustness of the feature extractor, the following experimental set-up is executed. The experimental setup focuses on evaluating the consistency of internal representations (feature vectors) generated by GoogLeNet, retrained on the AAR dataset, in response to various image transformations. The analysis is conducted on the validation dataset from Batch-1 and employs Cosine Similarity to quantify the similarity between feature vectors before and after transformation.

The transformations applied include translation (image shift by ±10%, ±20%, ±30%), rotation (±15°, ±30°, ±45°, ±90°, ±180°, ±360°), scaling (resizing by 0.8×, 1.2×, 1.5×), noise addition (Gaussian noise with $\sigma = 0.01, 0.05, 0.1$), and illumination changes (brightness adjustments to 0.5 and 1.5, representing dimmer and brighter images respectively). Each transformation is applied independently on the validation batch and compared to the original validation batch to isolate its effect on the feature representations. This approach aims to provide empirical insight into the transformation invariance of GoogLeNet's 1024-dimensional feature vector before classification. The results are as shown below in Table 3.

Across all types of transformations, the model demonstrated a reasonable degree of stability in its internal feature representations. The model demonstrated moderate

Table 3 Cosine similarity between original images and transformed versions from the validation set.

Comparison	Cosine Similarity (mean)	Cosine Similarity (std)
Original vs. Brightness 0.5x	0.76	0.07
Original vs. Brightness 1.5x	0.69	0.15
Original vs. Gaussian Noise $\sigma = 0.01$	0.80	0.01
Original vs. Gaussian Noise $\sigma = 0.05$	0.77	0.05
Original vs. Gaussian Noise $\sigma = 0.1$	0.73	0.08
Original vs. Scale 0.8x	0.75	0.05
Original vs. Scale 1.2x	0.77	0.04
Original vs. Scale 1.5x	0.70	0.11
Original vs. Rotation 15°	0.76	0.05
Original vs. Rotation 30°	0.74	0.07
Original vs. Rotation 45°	0.74	0.08
Original vs. Rotation 90°	0.73	0.07
Original vs. Rotation 180°	0.71	0.10
Original vs. Rotation 360°	0.71	0.10
Original vs. Translation 10%	0.78	0.04
Original vs. Translation 20%	0.75	0.06
Original vs. Translation 30%	0.74	0.07

stability under geometric and photometric distortions but showed notable degradation under more aggressive perturbations. For brightness adjustments, reducing illumination to 0.5x resulted in a cosine similarity of 0.76, while increasing it to 1.5x further decreased similarity to 0.69, indicating significant sensitivity to global intensity changes. Similarly, Gaussian noise had a progressive impact: at low noise levels ($\sigma = 0.01$), similarity remained relatively high (0.80), but increasing noise ($\sigma = 0.1$) reduced it to 0.73, suggesting that the model's feature representations are susceptible to high-frequency perturbations.

Geometric transformations also affected feature consistency, though to varying degrees. Small rotations (15°) retained a similarity of 0.76, but more extreme rotations (180°) led to a drop to 0.71, indicating that the model lacks full rotational invariance. Similarly, spatial translations of 10% resulted in a similarity of 0.78, while larger shifts (30%) reduced it to 0.74, highlighting a limited tolerance to positional changes. Scaling transformations followed a comparable trend: moderate resizing (1.2x) preserved a similarity of 0.77, but aggressive scaling (1.5x) caused a decline to 0.70, suggesting that the model's feature extraction is scale dependent.

In summary, the model exhibits partial robustness to common image transformations, with cosine similarity typically ranging between 0.69 and 0.80 under perturbations. While it maintains reasonable stability under mild distortions, its performance degrades with stronger variations, particularly in brightness, large rotations, and significant scaling.

6.2 Semantic Representation in Feature Extraction

It is visible through the Class Activation Maps (CAMs) as shown below in Figure 5., that the baseline model captures the semantic meanings of the input images well. The CAMs across different layers - inception3a, inception4a, and inception5b, demon-

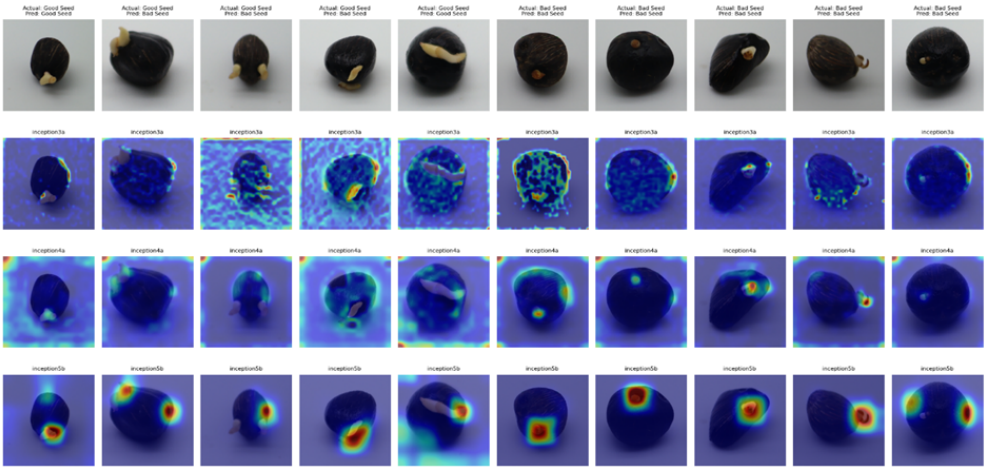

Fig. 3 CAMs for 5 different good and bad seeds, for three different layers of the GoogLeNet model: Inception3a, Inception4a and Inception5b.

strate how the trained model evolves its focus from low-level to high-level semantic regions in germinated oil palm seed images.

The CAMs at the Inception3a show relatively diffused activations. The model appears to be responding broadly to texture, edges, and surface contrast, especially around the seed perimeter or background noise. These patterns are indicative of low-level features, which are essential for basic shape and edge detection. In Inception4a (Mid Layer) the activations begin to sharpen and localise. The focus shifts towards meaningful regions like the plumule and radicle, which are critical biological indicators of germination quality. This layer captures a blend of low- and mid-level features, encoding texture alongside structural patterns.

At Inception5b (Final Layer), the CAMs from this deeper layer exhibit strongly localised and discriminative activations concentrated on highly relevant regions such as the emergent sprouts or surface abnormalities.

These responses indicate the model is using high-level semantic features that correlate closely with domain-specific cues used by human experts as previously mentioned. The CAMs confirm that the model progressively transitions from capturing low-level visual patterns to high-level semantic concepts. This representation aligns well with

the biological structure of germinated seeds and suggests that the network effectively captures semantic meaning necessary for accurate classification.

6.3 Feature Discriminability

Regarding how discriminant the features are to the classes GoodSeed and BadSeed, it can be seen through the AUC score in the performance of the baseline model (0.995, 0.875, 0.829 on the validation batch, Batch-2 and Batch-3) that the features are fairly well discriminated.

To further evaluate the discriminative capability of the features, a t-Distributed Stochastic Neighbour Embedding (t-SNE) visualisation experiment was conducted. First, the validation set (Batch-1 test set) was loaded with consistent resizing, normalisation, and tensor conversion applied during preprocessing. The final fully connected layer of the re-trained baseline GoogLeNet model was replaced with an identity mapping, allowing direct extraction of 1024-dimensional feature vectors from the penultimate layer. These high-dimensional feature representations were then projected into two dimensions using t-SNE, with a perplexity value of 30 and a fixed random seed to ensure reproducibility. Finally, the resulting 2D features were plotted separately for GoodSeed and BadSeed classes, allowing for visual analysis of class separability. This is shown in Figure 4 below.

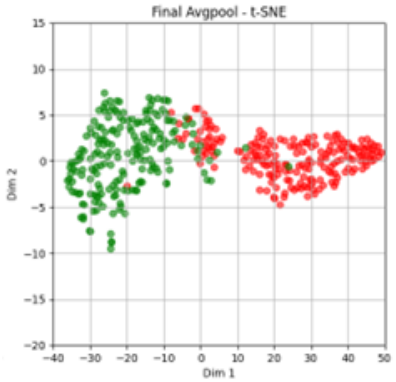

Fig. 4 The T-SNE plot for the 1024-dimensional feature vectors from the last layer of GoogLeNet, of the images from the validation batch.

It is observed that the features are discriminant, but there is some overlap between the two classes.

6.4 Robustness of the Classifier to Input Variations

To assess the classification robustness of the re-trained GoogLeNet model, we conducted a classification consistency analysis on Batch-1 of the validation dataset. The evaluation involved applying a series of controlled image transformations—each introduced independently—to measure the model's predictive stability against its original

performance on the validation batch. The transformations included: spatial translations (±10%, ±20%, ±30%), rotations (±15°, ±30°, ±45°, ±90°, ±180°, ±360°), scaling (0.8×, 1.2×, 1.5×), additive Gaussian noise (σ = 0.01, 0.05, 0.1), and illumination variations (brightness factors of 0.5 and 1.5). Classification outputs were recorded post-transformation to analyse the model's resilience and sensitivity to each perturbation in Table below.

Table 4 Performance metrics of classification consistency under various image transformations.

Transformation	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**AUC**
Without Transforms	0.953	0.917	0.995	0.955	0.995
Brightness=0.5	0.8803	0.9581	0.7960	0.8696	0.9691
Brightness=1.5	0.8180	0.9000	0.7164	0.7978	0.9281
Noise=0.01	0.8554	0.9931	0.7164	0.8324	0.9901
Noise=0.05	0.5112	1.0000	0.0249	0.0485	0.9350
Noise=0.1	0.4988	0.0000	0.0000	0.0000	0.7347
Scale=0.8	0.9526	0.9691	0.9353	0.9519	0.9901
Scale=1.2	0.8529	0.9931	0.7114	0.8290	0.9910
Scale=1.5	0.8379	0.9857	0.6866	0.8094	0.9842
Rotate=15	0.9277	0.9886	0.8657	0.9231	0.9908
Rotate=30	0.9377	0.9783	0.8955	0.9351	0.9931
Rotate=45	0.9501	0.9738	0.9254	0.9490	0.9905
Rotate=90	0.9551	0.9841	0.9254	0.9538	0.9942
Rotate=180	0.9476	0.9737	0.9204	0.9463	0.9850
Rotate=360	0.9202	0.9519	0.8856	0.9175	0.9864
Translate=0.1	0.9302	0.9943	0.8657	0.9255	0.9928
Translate=0.2	0.9352	0.9834	0.8856	0.9319	0.9904
Translate=0.3	0.9451	0.9786	0.9104	0.9433	0.9893

It is seen that Noise = 0.1 causes heavy deterioration to the classification performance – the classifier is not robust to heavy levels of Gaussian noise. The transform which had the highest similarity in validation performance was Scale = 0.8, suggesting the model's invariance to scale can be attributed to the classifier head. Notably, despite the moderate cosine similarity scores between original and scaled feature representations, the classifier maintained strong performance similar to the original validation performance. This suggests that the model's ability to generalise under such transformations is largely attributed to the robustness of the classifier head itself. Additionally, it is observed that although the features' cosine similarity decreases under increasing levels of translation, the classifier still maintains a constant level of accuracy under increasing levels of translation. Conclusively, while this experiment suggests that the classifier is exhibits moderate invariance to the rotation, scale and translation, it is observed that none match the initial validation performance, suggesting that there needs to be modifications in the training strategy.

6.5 Impact of Feature Semantics on Classification Performance

To evaluate how different levels of semantic feature representations affect classification performance, feature vectors were extracted from multiple layers of the trained GoogLeNet model and visualised using t-SNE projections. The resulting embeddings are illustrated in Figure 5.

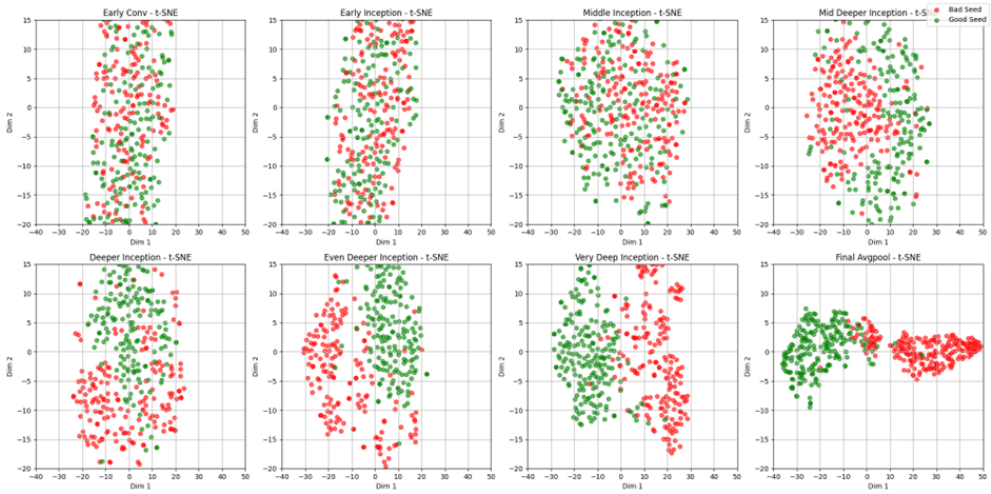

Fig. 5 The T-SNE plot for the 1024-dimensional feature vectors of the images from the validation batch, visualised from different layers of GoogLeNet, earliest to deepest (from top left to bottom right).

Feature embeddings from early convolutional and initial inception layers are highly entangled, with poor separation between GoodSeed and BadSeed classes. These layers capture low-level visual cues such as texture and edges, which are insufficient for reliable class distinction in this context. As we progress deeper, we observe increasing class separation, suggesting that mid-level layers begin encoding more abstract and semantically relevant representations. The final layers show strong class separability, with GoodSeed and BadSeed instances forming distinct clusters. Although intermediate layers capture progressively abstract representations, only the final layers exhibit strong class discriminability. Earlier convolutional and inception modules lack sufficient separation between GoodSeed and BadSeed features, as evidenced by t-SNE clustering. Therefore, features extracted from earlier layers are suboptimal for classification tasks compared to those from the final layer.

6.6 Proposed Modifications

The proposed modification involves an augmentation approach. This augmentation strategy is informed by the prior classifier and feature extractor analyses, as well as established findings in the literature. Rotation-based augmentation was selected based

on both empirical observations and prior research. Qu et al. [23] demonstrated that rotation enhanced the recognition performance of GoogLeNet. Our feature extractor analysis showed a 5% reduction in cosine similarity of extracted features with increasing rotation levels. This indicates feature sensitivity to rotational changes, prompting the inclusion of 360-degree rotation in the training data to improve rotational robustness of the features. Gaussian noise with a standard deviation of 0.01 was applied to enhance the model's robustness to minor noise perturbations.

Higher noise levels (standard deviations of 0.05 and 0.1) were found to significantly degrade classification accuracy on the validation batch (to approximately 50%), and heavy noise application to the train dataset can result to the model misclassifying images with low noise. Translation augmentation was limited to 10% of the image dimensions. This decision was based on the observation that the model's translational invariance appeared to be learned primarily through classifier training, necessitating explicit exposure to translated examples. Larger translation values (e.g., 20%–30%) were avoided to prevent occlusion of key seed structures, such as the plumule and radicle, which are critical for accurate classification.

Scaling augmentation was informed by the classification consistency analysis. Classification consistency dropped by 12% with increasing scale. To address this sensitivity, a scaling range of 0.9 to 1.4 was adopted, allowing the model to generalise to both zoomed-in and zoomed-out views of the seed images. Finally, random horizontal and vertical flips were incorporated. These transformations are widely recognised for enhancing model generalisation in image classification tasks and have shown effectiveness in datasets such as ImageNet and CIFAR-10 [29].

The proposed augmentation strategy will be implemented as an offline augmentation strategy, whereby transformed versions of the training images are generated prior to training and appended to the original dataset. This approach ensures that the model consistently encounters a diverse set of inputs during each epoch, without introducing runtime variability.

7 Proposed Modified Model

The proposed modification was implemented, and the performance is reported and analysed.

7.1 Implementation Details

The implementation remains consistent with that of the baseline model, with the sole modification being the incorporation of data augmentation.

7.2 Results

It is observed that while the validation performance is reduced by 1% from the baseline model, its performance on Batch-2 and Batch-3 increase by 12%. This suggests that

Table 5 Model performance metrics per batch.

Batch	Accuracy	Precision	Recall	F1 Score	AUC Score
Batch 1 (Validation Set)	0.945	1.000	0.891	0.942	0.992
Batch 2 (NormalRoomLight)	0.871	0.917	0.816	0.864	0.942
Batch 3 (LightBox)	0.851	0.813	0.917	0.862	0.916

the model significantly generalises better to the test batches compared to the baseline model. The confusion matrices are in Figure 6 below.

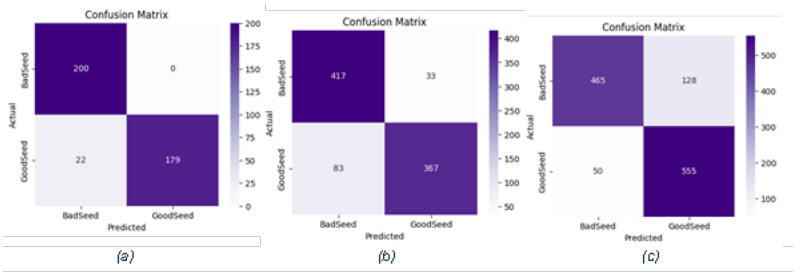

Fig. 6 Confusion matrices for (a) the validation batch, (b) Batch-2, (c) Batch-3

7.3 Effectiveness of the Proposed Modification

The analysis for the baseline model is repeated, and our observations validate the effectiveness of the offline augmentation technique. The feature extractor analysis was repeated to assess the effectiveness of the offline feature augmentation strategy. The results demonstrate that the extracted features have become *nearly invariant to scale, rotation, and translation* as shown in Table 6 below. Additionally, there was a notable improvement in cosine similarity with respect to brightness variations, and the model exhibited *complete invariance* to Gaussian noise with a standard deviation of 0.01.

The invariance to these features imply its effectiveness on the generalisability of the model, which led to the increase in performance on the test batches.

Furthermore, to further assess the effectiveness of the proposed modifications, the t-SNE plots of the features extracted from both the baseline model and the proposed model on the validation batch are compared in Figure 7. below.

It is shown in the t-SNE plot for the proposed approach that the features become more discriminant, the features belonging to each class become more similar to each other resulting in the tighter clusters. For the proposed model's plot, the two clusters show better separation, but with fewer points lying near and across the decision boundary compared to the baseline model. There's increased intra-class variance, particularly in the GoodSeed cluster, suggesting that the proposed model has learned

Table 6 Cosine similarity between original and augmented images.

Comparison	Cosine Similarity (mean)	Cosine Similarity (std)
Original vs. Brightness 0.5x	0.97	0.09
Original vs. Brightness 1.5x	0.81	0.28
Original vs. Gaussian Noise $\sigma = 0.01$	1.00	0.01
Original vs. Gaussian Noise $\sigma = 0.05$	0.86	0.20
Original vs. Gaussian Noise $\sigma = 0.1$	0.63	0.35
Original vs. Scale 0.8x	0.97	0.08
Original vs. Scale 1.2x	0.98	0.04
Original vs. Scale 1.5x	0.95	0.10
Original vs. Rotation 15°	0.99	0.02
Original vs. Rotation 30°	0.99	0.03
Original vs. Rotation 45°	0.99	0.03
Original vs. Rotation 90°	0.97	0.09
Original vs. Rotation 180°	0.96	0.09
Original vs. Rotation 360°	0.97	0.07
Original vs. Translation 10%	0.99	0.01
Original vs. Translation 20%	0.99	0.02
Original vs. Translation 30%	0.97	0.08

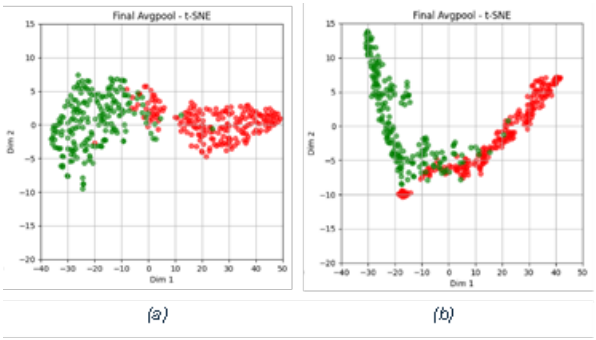

Fig. 7 T-SNE plots of the 1024-dimensional feature vectors from the validation batch for the (a) baseline model, (b) modified model.

a broader, more flexible representation of each class.

Lastly, the classifier consistency experiment to test the sensitivity to the individual transforms is also executed, the following Table 7. shows the results.

Classifier consistency declined on individually transformed inputs, with accuracy dropping from the 90% to 70-80% range. This suggests the model, trained on combined augmentations (e.g., rotation, noise, brightness, scaling), lacked exposure to isolated transformations, reducing its ability to handle them independently. However, real-world distortions rarely occur in isolation; thus, training on combined perturbations fostered more generalised, combined distortion-invariant representations, enhancing performance on mixed-condition test sets. While this improved robustness, the reduced

Table 7 Performance Metrics for Data Augmentations.

Augmentation	Accuracy	Precision	Recall	F1 Score	AUC Score
Brightness=0.5	0.778	0.983	0.567	0.719	0.917
Brightness=1.5	0.713	0.989	0.433	0.602	0.875
Noise=0.01	0.813	1.000	0.627	0.771	0.960
Noise=0.05	0.651	1.000	0.303	0.466	0.925
Noise=0.1	0.526	1.000	0.055	0.104	0.815
Scale=0.8	0.815	0.992	0.637	0.776	0.950
Scale=1.2	0.838	1.000	0.677	0.807	0.963
Scale=1.5	0.860	1.000	0.721	0.838	0.967
Rotate=15	0.830	0.993	0.667	0.798	0.961
Rotate=30	0.830	0.993	0.667	0.798	0.961
Rotate=45	0.840	0.979	0.697	0.814	0.965
Rotate=90	0.850	1.000	0.701	0.825	0.968
Rotate=180	0.848	0.993	0.701	0.822	0.967
Rotate=360	0.845	0.973	0.711	0.822	0.963
Translate=0.1	0.813	0.992	0.632	0.772	0.963
Translate=0.2	0.825	0.985	0.662	0.792	0.962
Translate=0.3	0.843	0.973	0.706	0.818	0.955

specialisation to individual distortions remains a limitation, potentially addressable through a hybrid augmentation strategy including both isolated and combined augmentations in future works.

8 Further Discussion

Our experimental results demonstrate how a systematically designed augmentation strategy significantly improved the generalisability of the GoogLeNet classifier for germinated oil palm seed quality assessment. The modifications, carefully selected through empirical analysis of the baseline model's weaknesses and supported by literature evidence, addressed specific challenges in agricultural image classification.

8.1 Performance Improvements and Trade-offs

8.1.1 Accuracy and Generalisation

The modified model achieved higher accuracy on both Batch-2 (87.1% vs. 73.2%) and Batch-3 (85.1% vs. 75.2%) compared to the baseline. This improvement can be attributed to the offline augmentation strategy, which introduced controlled variations (rotation, scaling, noise, and translation) during training. By exposing the model to a broader range of transformations, we enhanced its ability to generalise to unseen lighting conditions. The t-SNE visualisations further confirmed that the modified model learned more discriminative feature representations, as evidenced by tighter intra-class clustering.

8.1.2 Precision and Recall Trade-offs

Precision improved markedly on both test batches, rising to 91.7% (from 67.1%) on Batch-2 and 81.3% (from 66.1%) on Batch-3, indicating a substantial reduction in false positives where bad seeds were incorrectly classified as good. This improvement came with a moderate decrease in recall, which declined to 81.6% (from 98.9%) on Batch-2 and 91.7% (from 96.2%) on Batch-3. This trade-off reflects the model's more conservative approach to labelling seeds as "GoodSeed", which is beneficial for agricultural applications where avoiding false positives (misclassifying bad seeds as good) is often more critical than maximizing recall.

8.1.3 AUC and F1 Score Enhancements

The AUC scores improved for Batch-2 (94.2% vs. 87.5%) and for Batch-3 (91.6% vs. 82.9%), indicating better class separability in the modified model. Similarly, the F1 score increased for both test batches (Batch-2: 86.4% vs. 80.0%; Batch-3: 86.2% vs. 78.4%), demonstrating a more balanced performance. These improvements confirm that the augmentation strategy successfully mitigated some of the baseline model's weaknesses, particularly in handling photometric and geometric variations.

8.2 Effectiveness of the Proposed Modification

Figure 8 illustrates sample seed images from three datasets: Batch-1, Batch-2 (NormalRoomLight), and Batch-3 (LightBox).

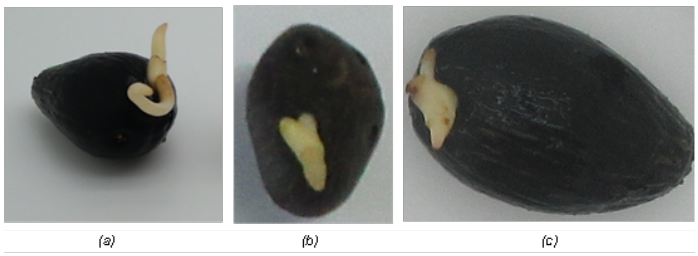

Fig. 8 Seed images from (a) Batch-1, (b) Batch-2, (c) Batch-3

Substantial variation is observed across batches in terms of image scale, clarity, and illumination. Batch-1 comprises high-resolution images that are generally zoomed out, providing a broader view of the seeds. Batch-2, on the other hand, contains images with increased noise levels and a closer crop around the seed objects. Batch-3 images exhibit uniform lighting conditions, eliminate shadows but introduce differences in both scale and resolution. These inter-batch discrepancies show the domain shift between training and test data, thereby explaining the baseline model's limited generalisation performance when evaluated across batches with differing visual characteristics. Our carefully designed offline augmentation strategy significantly enhanced model robustness through combining rotation (360°) to address orientation sensitivity, minimal

Gaussian noise ($\sigma = 0.01$) to improve noise resilience without performance degradation, and controlled scaling (0.9×-1.4×) to maintain size invariance across batches. Each augmentation was selected to address specific baseline weaknesses, such as sensitivity to noise and scaling inconsistencies. Although performance slightly declined on the validation set (Batch-1), significant gains on unseen test data (Batch-2 and Batch-3) indicate improved generalisation—crucial for practical agricultural deployment. This targeted strategy effectively balanced model refinement with real-world applicability. In addition, early convergence is observed as the model stops training at 8 epochs than 11 from the baseline model (see Figure 9).

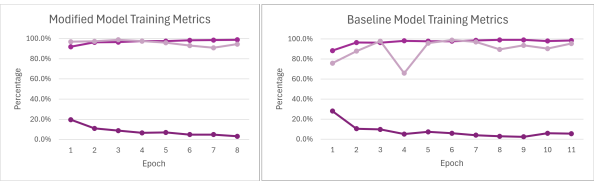

Fig. 9 Training metrics for proposed and baseline models.

8.3 Limitations and Future Work

Despite observed improvements, challenges remain. While this study used a balanced dataset, real-world scenarios often involve class imbalance, which may necessitate weighted loss functions or advanced resampling. Future work should systematically evaluate individual and combined augmentations to ensure robustness under diverse perturbations, enhancing model generalisability in complex agricultural settings.

9 Conclusion

This study presented a comprehensive approach to germinated oil palm seed quality classification using transfer learning with GoogLeNet. Through detailed analysis of the baseline model, we identified key weaknesses in the feature extractor and classifier when exposed to common image transformations such as rotation, translation, scaling, noise, and illumination changes. These findings guided the design of a targeted offline data augmentation strategy, which significantly improved the model's robustness and generalisation to real-world imaging conditions.By retraining GoogLeNet with augmented data, we observed a 12% performance improvement on unseen test batches captured under different lighting environments. Feature similarity analysis confirmed enhanced invariance to the targeted transformations, while t-SNE visualisations demonstrated improved intra-class variance. Interestingly, our results show that the augmentation strategy strengthened the feature extractor more than the classifier, contrary to our initial expectations. Lastly, our findings highlight the value of conducting pre-modification analysis on both the feature extractor and classifier to design effective augmentation pipelines.

References

[1] Cui, J., Lamade, E., and Tcherkez, G. (2020). Seed germination in oil palm (Elaeis guineensis Jacq.): A review of metabolic pathways and control mechanisms. *International Journal of Molecular Sciences*, 21(12), 4227.

[2] Murphy, D. J., Goggin, K., and Paterson, R. R. M. (2021). Oil palm in the 2020s and beyond: Challenges and solutions. *CABI Agriculture and Bioscience*, 2, 1–22.

[3] Liao, I. Y., Shen, B. S. K., Chen, Z. Y., Jelani, M. F., Wong, C. K., and Wong, W. C. A preliminary study on germinated oil palm seeds quality classification with convolutional neural networks.

[4] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9.

[5] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., et al. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1), 43–76.

[6] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255.

[7] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, Zurich, Switzerland, pp. 740–755.

[8] Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., et al. (2020). The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7), 1956–1981.

[9] Ma, Y., Zhang, P., and Tang, Y. (2018). Research on fish image classification based on transfer learning and convolutional neural network model. In *14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, Huangshan, China, pp. 850–855.

[10] Chen, Z. (2022). Study of transferability of ImageNet-based pretrained model to brain tumor MRI dataset. In *International Conference on Machine Learning and Intelligent Systems Engineering (MLISE)*, Guangzhou, China, pp. 87–91.

[11] Khotsyanovsky, V. (2022). Comparative characteristics of the ability of convolutional neural networks to the concept of transfer learning. *Technology Audit and Production Reserves*, 1(2(63)), 10–13.

[12] Oh, K., Chung, Y. C., Kim, K. W., Kim, W. S., and Oh, I. S. (2019). Classification and visualization of Alzheimer's disease using volumetric convolutional neural network and transfer learning. *Scientific Reports*, 9(1), 18150.

[13] Raj, J. A., Qian, L., and Ibrahim, Z. (2024). Fine-tuning—A transfer learning approach. *arXiv preprint* arXiv:2411.03941.

[14] Loddo, A., and Di Ruberto, C. (2021). On the efficacy of handcrafted and deep features for seed image classification. *Journal of Imaging*, 7(9), 171.

[15] Luo, T., Zhao, J., Gu, Y., Zhang, S., Qiao, X., Tian, W., and Han, Y. (2023). Classification of weed seeds based on visual images and deep learning. *Information Processing in Agriculture*, 10(1), 40–51.

[16] Hanafi, F. S., Dewanta, F., and Budiman, G. (2023). Performance evaluation of rice seed classification system based on the CNN with VGG-GoogLeNet architecture. In *10th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, pp. 231–236.

[17] Kurtulmuş, F. (2021). Identification of sunflower seeds with deep convolutional neural networks. *Journal of Food Measurement and Characterization*, 15(2), 1024–1033.

[18] Maeda-Gutiérrez, V., Galván-Tejada, C. E., Zanella-Calzada, L. A., Celaya-Padilla, J. M., Galván-Tejada, J. I., Gamboa-Rosales, H., et al. (2020). Comparison of convolutional neural network architectures for classification of tomato plant diseases. *Applied Sciences*, 10(4), 1245.

[19] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826.

[20] Li, J., et al. (2023). Label-efficient learning in agriculture: A comprehensive review. *Computers and Electronics in Agriculture*, 215, 108412.

[21] Ng, J., Liao, I. Y., Jelani, M. F., Chen, Z. Y., Wong, C. K., and Wong, W. C. (2024). Multiview-based method for high-throughput quality classification of germinated oil palm seeds. *Computers and Electronics in Agriculture*, 218, 108684.

[22] Myburgh, J. C., Mouton, C., and Davel, M. H. (2020). Tracking translation invariance in CNNs. In *Southern African Conference on Artificial Intelligence Research*, pp. 282–295. Springer.

[23] Qu, J. (2016). An improved image classification method considering rotation based on convolutional neural network. In *Big Data Computing and Communications (BigCom 2016)*, Lecture Notes in Computer Science, vol. 9784, Springer, pp. 369–378.

[24] Rodner, E., Simon, M., Fisher, R. B., and Denzler, J. (2016). Fine-grained recognition in the noisy wild: Sensitivity analysis of convolutional neural networks approaches. *arXiv preprint* arXiv:1610.06756.

[25] Hu, C., et al. (2023). Impact of light and shadow on robustness of deep neural networks. *arXiv preprint* arXiv:2305.14165.

[26] Al-Huseiny, M. S., and Sajit, A. S. (2021). Transfer learning with GoogLeNet for detection of lung cancer. *Indonesian Journal of Electrical Engineering and Computer Science*, 22(2), 1078–1086.

[27] Barman, S., Biswas, M. R., Marjan, S., Nahar, N., Hossain, M. S., and Andersson, K. (2022). Transfer learning based skin cancer classification using GoogLeNet. In *International Conference on Machine Intelligence and Emerging Technologies*, Springer, pp. 238–252.

[28] Fang, T. (2018). A novel computer-aided lung cancer detection method based on transfer learning from GoogLeNet and median intensity projections. In *IEEE International Conference on Computer and Communication Engineering Technology (CCET)*, pp. 286–290.

[29] Shorten, C., and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), 1–48.