A Comparative Study of Transfer Learning Strategies for Produce Freshness Classification

Advanced Machine Learning for Computer Vision

Maxwell Bernard
Copenhagen Business School
IT University of Copenhagen
Copenhagen, Denmark
maxbe@itu.dk

Johan Schommartz
Copenhagen Business School
IT University of Copenhagen
Copenhagen, Denmark
johsc@itu.dk

***Abstract*—Automated detection of food spoilage has the potential to reduce household food waste and improve smart-fridge functionality, yet remains largely unexplored in commercial systems. This project investigates the feasibility of computer-vision–based freshness classification under realistic data constraints using transfer learning. A balanced dataset of fresh and rotten fruits and vegetables is used to compare two prominent vision architectures, ResNet-50 and Vision Transformer ViT-B/16, under frozen, fine-tuned, and training-from-scratch regimes.**

Experimental results demonstrate that transfer learning is essential for reliable performance on small datasets. Fine-tuned models substantially outperform frozen and scratch-trained variants, with ResNet-50 achieving the highest accuracy and most stable training behaviour. Vision Transformer models exhibit greater instability and reduced performance when data are limited, highlighting the importance of inductive biases for this task. These findings indicate that transfer learning enables accurate freshness detection in constrained settings and that convolutional architectures remain the most robust choice for potential smart-fridge deployment.

***Index Terms*—computer vision, transfer learning, food freshness classification, deep learning, convolutional neural networks, vision transformers**

I. Introduction

Computer vision is increasingly added to the functionality of smart home devices. Commercial smart fridges integrate internal cameras to identify food items, track inventory levels, and support semi-automated shopping list generation, but none of them provide automated spoilage detection. An automated computer vision system capable of detecting spoilage could prevent cross-contamination, reduce unnecessary waste and enable proactive removal reminders or automatic re-ordering.

Early work on food image recognition demonstrated that convolutional neural networks can successfully learn discriminative visual features for food-related tasks, establishing the feasibility of applying deep learning to food imagery [1].

This project investigates whether deep learning models can reliably distinguish fresh from rotten fruits and vegetables based on fridge images. Because collecting large real-world, annotated fridge images is costly, a successful system must perform well on small datasets. This makes transfer learning a natural approach. Since we don't have access to even a small augmented real-world fridge image dataset, we train the system on a small, augmented non-fridge dataset, approximating fridge conditions.

We compare two prominent architectures, ResNet-50 [2] and Vision Transformer ViT-B/16 [3], under three training regimes. We use a dataset sourced from Kaggle, consisting of fresh and rotten images of eight fruit and vegetable types.

The research question addressed in this report is whether transfer learning with CNN and transformer architectures can reliably classify freshness under realistic constraints, and which architecture is best suited for potential deployment in a smart-fridge environment. Our project integrates insights from prior research on food recognition, freshness classification, industrial defect detection, and modern transfer learning.

II. Methods

The following section outlines the dataset used, the preprocessing pipeline, the architectural choices and the training configurations used in this project.

A. Dataset

This freshness detection is formulated as a 16-class classification problem. The dataset contains eight fruit and vegetable types, each split into fresh and rotten categories. The images exhibit substantial variation in lighting conditions, camera angles, and background settings, which helps approximate real-world imaging scenarios and promotes the generalizability of the model. While food freshness degrades along a continuous spectrum, the clear binary labelling between fresh and rotten states in the dataset enables models to learn discriminative features that separate late freshness from early stages of decay across different produce types.

B. Preprocessing and Balancing Strategy

The dataset originally contained PNG images of varying resolutions, many of which were unnecessarily large and slowed down data loading. Therefore, all images were converted to JPEG format with 95% quality, reducing the file size by

approximately a factor of three to five while preserving the visual quality [4].

The Kaggle dataset is pre-split into separate training and test folders and may contain variants of the same images across these folders. To avoid potential data leakage, this project only uses the training folder. Due to class imbalance within this folder, all classes were undersampled to match the smallest class. Rotten okra was identified as the smallest class with 338 images, and all other classes were downsampled accordingly, resulting in a balanced dataset of 5,408 images. The resulting dataset was then split into training, validation, and test sets using a 70/15/15 split, yielding approximately 3,776 training images and 816 images each for validation and testing. This balancing procedure ensures that performance metrics reflect genuine classification ability rather than prevalence by class.

Since both ResNet-50 and ViT-B/16 are pretrained on ImageNet, the training images were adapted to match the data these models were originally trained on. All images were first resized and center-cropped to 224×224 pixels, corresponding to the standard ImageNet input resolution, and subsequently normalized using the ImageNet statistics with channel means $\mu = (0.485, 0.456, 0.406)$ and standard deviations $\sigma = (0.229, 0.224, 0.225)$. This preprocessing ensures that the images are presented in a format and scale familiar to the models, helping them transfer their learned features effectively to the new task.

C. Data Augmentation

Fridge environments exhibit strong visual variability due to uneven lighting, parts of objects being blocked by shelves and food, and blur from condensation. To simulate these conditions, we applied data augmentations, including random resized crops, horizontal flips, colour jitter, Gaussian blur, and occasional random perspective distortions. These augmentations were designed to preserve semantic integrity while increasing the diversity of training samples. Comparable augmentation strategies are recommended in the industrial food-quality literature [5].

D. Model Architectures

Each pre-processed and augmented image is represented as $x \in \mathbb{R}^{224\times224\times3}$ and mapped by the model to one of 16 output classes. Both model architectures are trained using the categorical cross-entropy loss, defined as

$$\mathcal{L} = -\sum_{c=1}^{16} y_c \log \hat{y}_c, \tag{1}$$

which minimizes the divergence between the true label distribution y and the predicted class probabilities $\hat{y}$. This objective corresponds to the standard maximum-likelihood training criterion for multi-class image classification [6].

We evaluate two vision architectures, ResNet-50 and Vision Transformer ViT-B/16, both imported from Hugging Face and pretrained on ImageNet. A high-level comparison of the two architectures is illustrated in Figure 1.

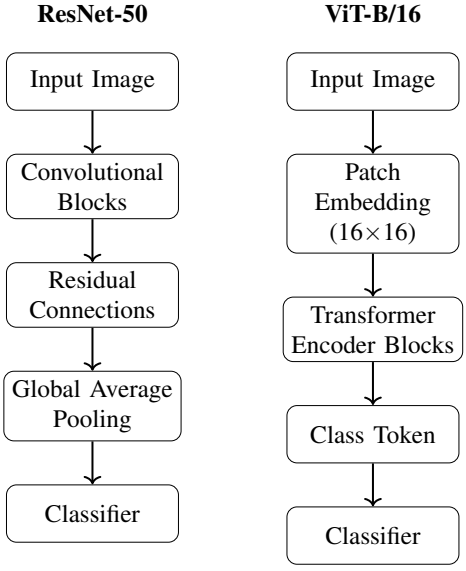

Fig. 1. High-level comparison of the ResNet-50 and Vision Transformer (ViT-B/16) architectures, highlighting convolutional processing with residual connections versus patch-based global self-attention, based on He et al. [2] and Dosovitskiy et al. [3].

As illustrated in Figure 1, ResNet-50, introduced by He et al. [2], is a deep convolutional neural network that employs residual connections of the form

$$\mathbf{h}_{l+1} = \mathcal{F}(\mathbf{h}_l) + \mathbf{h}_l, \tag{2}$$

allowing each block to learn a residual mapping added to its input. These skip connections stabilize gradient flow and enable effective training of very deep networks. More importantly for this application, ResNet-50 embeds strong inductive biases such as locality and translation equivariance, making it data-efficient and well suited for recognizing fine-grained local visual patterns. These properties align with the detection of mold speckling, surface bruising, and texture collapse, explaining the widespread use of ResNet variants in surface defect detection and food-quality inspection tasks.

In contrast, Vision Transformer ViT-B/16, proposed by Dosovitskiy et al. [3], adopts a transformer-based architecture in which images are divided into non-overlapping 16×16 patches. Each patch x_p^i is linearly embedded and combined with a positional encoding p_i to form

$$\mathbf{z}_0^i = \mathbf{E}x_p^i + \mathbf{p}_i, \tag{3}$$

followed by a sequence of transformer encoder blocks applying multi-head self-attention and feed-forward layers:

$$\mathbf{z}' = \mathbf{z} + \mathrm{MSA}(\mathrm{LN}(\mathbf{z})), \tag{4}$$

$$\mathbf{z}^+ = \mathbf{z}' + \mathrm{MLP}(\mathrm{LN}(\mathbf{z}')). \tag{5}$$

Unlike convolutional networks, Vision Transformers rely primarily on global self-attention and lack strong spatial inductive biases. As a result, they typically require large-scale

pretraining to learn robust visual representations. This characteristic makes ViT-B/16 an informative contrast to ResNet-50 in the smart-fridge setting, where only limited data are available and robustness under data scarcity is critical.

The comparison between convolutional and transformer-based architectures is motivated by the comparative study of Jawarneh et al. [7], which evaluates fruit classification performance under limited data conditions. Assessing these architectures under identical training settings provides insight into trade-offs in data efficiency, training stability, and the ability to learn complex visual patterns in the food domain.

E. Training Regimes and Optimization

Each architecture was trained under three distinct regimes. In the *frozen pretrained* regime, ImageNet-pretrained backbones were kept fixed and only the classifier head was trained. In the *fine-tuned* regime, higher-level layers were unfrozen to enable limited domain adaptation: layers three and four in ResNet-50 and the final transformer block in ViT-B/16 were made trainable. In the *scratch* regime, all weights were randomly initialized and trained end-to-end.

All models were optimized using AdamW [8], which decouples weight decay from gradient-based parameter updates. Weight decay was set to 10^{-4}, following the training configuration used in the original Vision Transformer study [3] and common practice in ResNet fine-tuning, as this value provides effective regularisation without destabilising transfer learning. A `ReduceLROnPlateau` learning-rate scheduler was employed to stabilize training by automatically reducing the learning rate when validation loss plateaued.

The overall training procedure follows the standard transfer-learning workflow described in the PyTorch Transfer Learning tutorial [9], including freezing the pretrained backbone during initial training and fine-tuning later layers with a reduced learning rate.

III. Experiments

A. Exploratory Data Analysis

The dataset is visually inspected to validate its diversity and identify potential challenges. Fresh images typically exhibit uniform coloration and smooth textures, whereas rotten items display a range of different colour and textures. Some classes, particularly tomatoes and capsicum, show subtle distinctions between fresh and slightly rotten stages, making them naturally difficult. Other classes like okra and cucumber just look very similar in both stages, making them easily confusable even for the human eye. Representative examples of unprocessed images appear in Figure 2, while pre-processed, normalized and augmented samples appear in Figure 3.

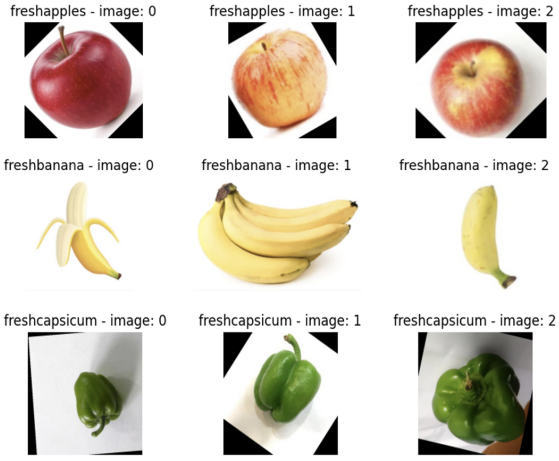

Fig. 2. Representative examples of unprocessed images from the dataset.

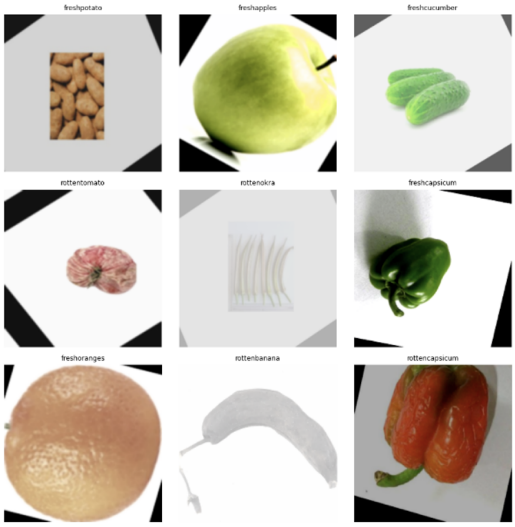

Fig. 3. Pre-processed, normalized, and augmented image samples.

B. Training Behaviour

Training behaviour varied significantly across architectures and training regimes as seen in Figure 4. The pretrained frozen models showed the smoothest convergence, with low variance in both training and validation loss. They did not exhibit any signs of underfitting or overfitting, which suggests that ImageNet features transfer well to freshness classification.

The pretrained fine-tuned models achieved better overall accuracy but displayed slightly more unstable training. Both ViT and ResNet showed small oscillations in validation loss and indications of mild overfitting. This behaviour is expected, since fine-tuning high-level layers increases model capacity for domain-specific adaptation while making the training more sensitive to noise.

The scratch models experienced the greatest difficulty. The ResNet-50 model trained from scratch converged to a moderate accuracy, but its training curve was notably noisier than

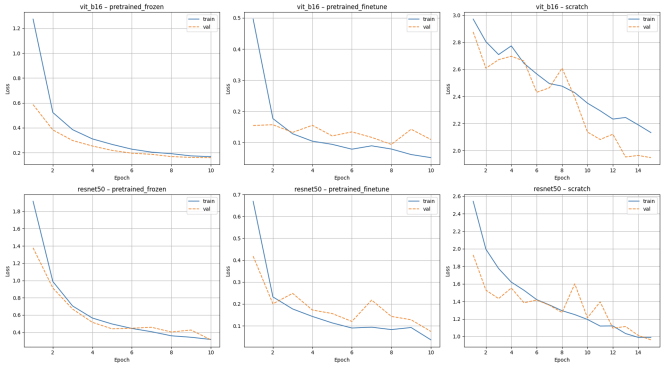

Fig. 4. Training and validation loss curves across architectures and training regimes.

the pretrained scenarios, and it underfitted relative to the fine-tuned variant. The ViT-B/16 model trained from scratch performed poorly, showing highly unstable loss curves. Across all regimes, ViT exhibited significantly less stable behaviour than ResNet, since transformers require far larger datasets to develop meaningful representations [3].

C. Quantitative Results

Quantitative classification performance across architectures and training regimes is summarized in Table I.

TABLE I
CLASSIFICATION ACCURACY (%) ACROSS MODEL ARCHITECTURES AND TRAINING REGIMES.

Model	Frozen	Fine-tuned	Scratch
ResNet-50	90.6	97.1	63.1
ViT-B/16	94.2	96.2	30.2

The fine-tuned ResNet-50 achieved the highest overall accuracy at 97.1%, outperforming the fine-tuned ViT-B/16, which reached 96.2%. Under frozen-backbone conditions, ViT slightly outperformed ResNet, with accuracies of 94.2% and 90.6% respectively. When trained from scratch, ResNet converged to a moderate accuracy of 63.1%, whereas the ViT-B/16 model collapsed to approximately 30.2%, although still performing above random guessing.

To further analyse class-level performance, Figures 5 and 6 present confusion matrices for the fine-tuned ResNet-50 and ViT-B/16 models, respectively. The fine-tuned ResNet-50 exhibits highly localized errors, with most misclassifications occurring between fresh and rotten variants of the same produce or between visually similar classes such as okra and cucumber. This suggests that the model effectively learns local texture cues associated with spoilage.

In contrast, the ViT-B/16 confusion matrix shows more diffuse error patterns, particularly among rotten produce classes. This indicates that, under limited data conditions, the transformer struggles to learn stable patch-level representations, leading to increased confusion across visually similar categories.

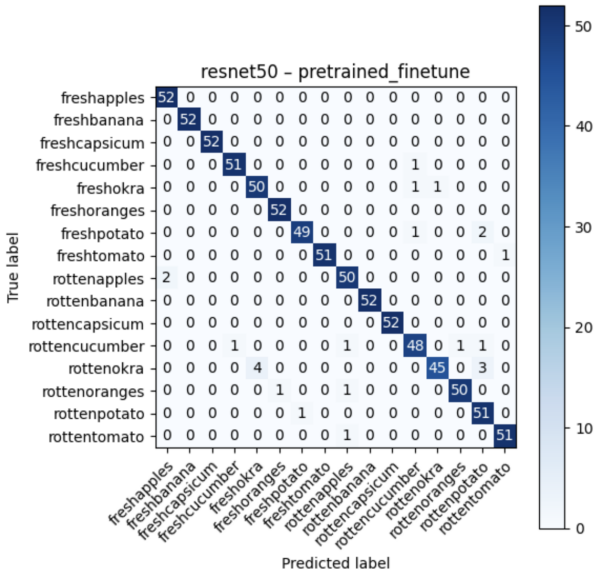

Fig. 5. Confusion matrix for the fine-tuned ResNet-50 model.

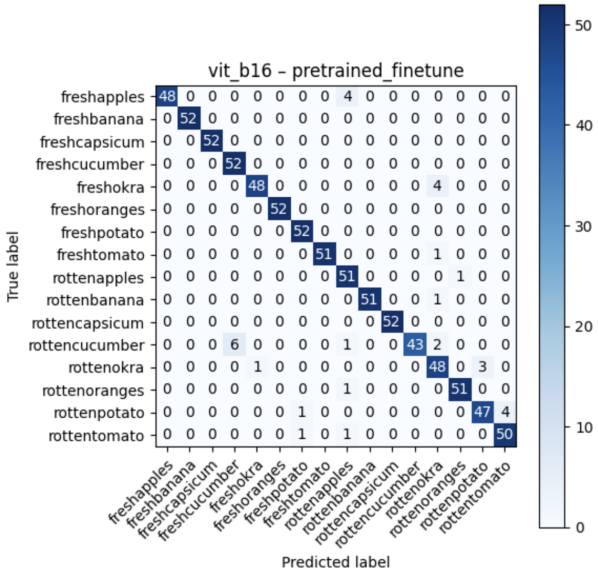

Fig. 6. Confusion matrix for the fine-tuned ViT-B/16 model.

IV. DISCUSSION

The experimental results demonstrate that transfer learning is essential for effective freshness classification on small datasets. Both architectures benefitted substantially from pretrained ImageNet weights, and fine-tuning high-level layers provided further improvements by enabling models to specialize in identifying spoilage-related patterns.

ResNet-50 outperformed ViT-B/16 in the fine-tuned regime primarily due to differences in inductive biases and sample efficiency. ResNet's convolutional filters excel at detecting localized textural and colour variations. These properties allow ResNet to extract meaningful features from even small datasets. ViT, in contrast, relies on global self-attention without

the structural priors that guide convolutional networks. ViT requires a large number of images to learn robust patch embeddings and attention patterns, and its performance deteriorates when these conditions are not met. This explains the instability and underperformance of ViT in the scratch regime as well as the higher variance observed during fine-tuning.

The project's main limitations include the reliance on studio-like images rather than true fridge environments, the absence of multi-object scenes, and the binary nature of the freshness labels that ignore intermediate stages. Future work could address these issues by collecting real-world fridge datasets, integrating object detection modules such as YOLOv8 to localize items before classification, and exploring modelling of gradual spoilage progression.

V. Conclusion

This project evaluated the feasibility of automated freshness detection using deep learning and found that transfer learning enables highly reliable classification even with a relatively small balanced dataset. Both ResNet-50 and ViT-B/16 achieved strong performance when pretrained on ImageNet. ResNet's strong inductive biases, greater sample efficiency, and more stable optimization dynamics explain its superior performance relative to ViT under the constraints of this dataset.

These findings indicate that smart-fridge freshness detection is technically feasible with current deep learning models, although additional work is required to adapt the system to real-world fridge implementation. Overall, convolutional architectures remain the most robust and practical choice for freshness-detection applications in limited-data scenarios.

References

[1] Y. Kawano and K. Yanai, "Food image recognition with deep convolutional features," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, ser. UbiComp '14 Adjunct. New York, NY, USA: Association for Computing Machinery, 2014, p. 589–593. [Online]. Available: https://doi.org/10.1145/2638728.2641339

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015, published at CVPR 2016. [Online]. Available: https://arxiv.org/abs/1512.03385

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020, published at ICLR 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[4] A. Shukla, "Efficient and optimized usage of various image formats (jpeg/jpg vs png) in applications," *Journal of Mathematical & Computer Applications*, vol. 2, no. 3, pp. 1–3, 2023. [Online]. Available: https://srcpublishers.com/index.php/mathematical-computer-applicatio/article/view/4412

[5] K. Shehzad, U. Ali, and A. Munir, "Computer vision for food quality assessment: Advances and challenges," *Global Journal of Machine Learning and Computing*, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:276586764

[6] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951. [Online]. Available: http://www.jstor.org/stable/2236703

[7] M. Jawarneh, A. Marwanto, D. Syamsuar, and M. Kusnandar, "A comparative study of convolutional neural networks and vision transformers for fruit classification," *International Journal of Advances in Artificial Intelligence and Machine Learning*, vol. 2, no. 2, pp. 104–115, 2025. [Online]. Available: https://doi.org/10.58723/ijaaiml.v2i2.435

[8] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam," *CoRR*, vol. abs/1711.05101, 2017. [Online]. Available: http://arxiv.org/abs/1711.05101

[9] PyTorch Team, "Transfer learning for computer vision," https://docs.pytorch.org/tutorials/beginner/transfer_learning_tutorial.html, 2023, accessed: 2025-03-17.