VGG-19 and Vision Transformer Enabled Shelf-Life Prediction Model for Intelligent Monitoring and Minimization of Food Waste in Culinary Inventories

Bindhya Thomas, Dr Priyanka Surendran
College of Computer Studies, University of Technology, Bahrain

***Abstract*—Food waste, particularly in the prepared food industry, presents a serious worldwide concern with serious ethical, environmental and socioeconomic implications. In restaurants and catering contexts, traditional inventory and waste management systems frequently lack the versatility and granularity to mitigate spoilage in real-time. The study proposes a sophisticated deep learning framework that predicts the remaining shelf-life of prepared food items using visual input, enabling timely interventions to reduce food waste. The proposed hybrid architecture integrates VGG-19 (Visual Geometry Group 19-layer network) for fine-grained feature extraction with Vision Transformer (ViT) that models contextual degradation patterns and temporal cues. The model operates by analyzing food images at regular intervals and predicting the remaining time before spoilage, enabling proactive decision-making for consumption prioritization. Food images are categorized into four freshness states: Fresh, Fit for Consumption, About to Expire and Expired, enabling the model to monitor real-time conditions. An elaborate dataset with 34 distinct food categories was utilized in the study, achieving outstanding performance with 98% accuracy, 97.5% precision, 97.9% recall and an F1-score of 97.75% and yielded an estimated 84% reduction in food waste. The model stands out for its non-invasive, image-based decision-making and the potential scalability across various food service settings. By offering predictive insights into food degradation and by using only visual data, the study advances the integration of artificial intelligence into sustainable food management.**

Keywords*—Food waste reduction; shelf-life prediction; VGG-19; vision transformer; image-based freshness classification; sustainable food management

I. Introduction

Food waste represents one of the most critical and paradoxical challenges of the modern world, with a gigantic volume of edible resources being discarded while millions continue to experience food insecurity and hunger [1]. It is estimated that around one-third of all food produced, amounting to over 1.3 billion tons annually, is wasted across the supply chain, from farms and distribution hubs to retail outlets and households [2]. Production and logistics inefficiencies, overstocking, unpredictable consumer behavior, inadequate preservation technologies and poor inventory management are some of the many factors contributing to food waste generation [3], [4]. Food waste reduction is a vital social, environmental and ethical necessity that goes beyond efficiency. It is becoming more and more critical to reduce waste through more intelligent, flexible solutions, as the world's population continues to expand and the effects of climate change put more strain on food systems [5]. Procedural and policy-based strategies like enhanced inventory checks, manual expiration date monitoring, redistributive networks and consumer awareness campaigns were traditionally adopted to curb food waste generation [6]. These methods mostly depend on human judgement and intervention, making them vulnerable to frequent errors and overlooks. The product-level data and real-time environmental variables are generally ignored, relying singularly on human visual inspection to judge freshness, which is inappropriate and inefficient for a vast amount of food. Similarly, demand estimates in the catering and hospitality sectors are frequently based on historical patterns and managerial judgement that often results in overproduction and associated food wastage [7]. The absence of traceability and consistent monitoring in retail and agricultural supply chains leads to spoilage that could have been avoided. Despite being fundamental, these conventional approaches are ineffective at handling the complexity of contemporary food systems as they are reactive rather than proactive.

A revolutionary change in handling food waste was promised by the rise of data-driven technologies. Predictive analytics, real-time monitoring and intelligent inventory control were made possible by blockchain, artificial intelligence (AI), computer vision and Internet of Things (IoT) sensors [8]. Convolutional neural networks (CNNs), YOLO object detection and recurrent neural networks (RNNs) are examples of machine learning (ML) models that have shown great promise in detecting food products, forecasting demand and identifying spoilage [9]. Automated supply chain optimizers, dynamic pricing systems and smart refrigeration systems have all been tested in business and industrial settings. Despite the initial potential, high implementation costs, reliance on data, lack of interoperability and restricted adaptability in rural or low-resource environments hamper their adaptability, generalizability and scalability. The systemic efficiency is undermined by the siloed nature of most of the current solutions, which concentrate on discrete supply chain components.

The central research question addressed in this study is whether a hybrid deep learning framework that integrates convolutional neural networks and transformer-based self-attention mechanisms using only visual inputs can reliably predict the remaining shelf life of prepared food items, enable timely interventions and significantly reduce food waste in real-world service settings. This study proposes a novel hybrid architecture for remaining shelf-life prediction, combining the

strengths of deep learning (DL) with advanced image segmentation and self-attention mechanisms. By using visual cues from food items to infer freshness levels, the study enables more precise and automated decisions on redistribution, consumption prioritization and dynamic pricing. The primary objectives of the study are as follows:

- Develop a CNN-Transformer hybrid DL architecture for accurate classification of food items into different freshness categories based on visual cues extracted from prepared food images.
- Evaluate the model performance and compare it with conventional methods, demonstrating its practical impact through a measurable reduction in food waste by enabling timely consumption or disposal based on predicted freshness.

The further sections of the study are structured as follows: Section II offers a detailed analysis of the latest advances in food waste reduction research and highlights the current research constraints. Section III delineates the proposed methodology. Section IV presents the experimental findings accompanied by a comprehensive analysis of the model performance in Section V. Finally, Section VI concludes the study by encapsulating the principal findings and emphasizing prospective areas for further research.

II. Related Works

Chun et al. [10] suggested deep learning techniques in food image classification in Korean cuisine. The study employed a Korean food image dataset that contained diverse categories of dishes that were pre-processed and augmented to enhance model generalization. Several convolutional neural network architectures were tested through transfer learning, among which InceptionResNetV2 achieved the highest classification accuracy of 81.91%, though it required an extensive training time of 436,182 seconds due to its complex inception–residual structure. NasNetLarge and MobileNetV2 followed with accuracies of 77.91% and 75.36%, respectively. Traditional ResNet variants, including ResNet-101V2, ResNet-152V2, and ResNet-50V2, showed lower accuracies ranging from 73.7% to 68.27%. The computational intensity was a major limitation of the study.

Louro et al. [11] suggested convolutional neural network (CNN)-based approaches for food waste reduction through food recognition technology and promote sustainable eating practices. The study utilized Food-101 dataset, together with the ResNet-50 initially and later adopted transfer learning with ImageNet weights to avoid overfitting. The convolutional layers processed food images hierarchically, progressively learning low-level features such as edges and textures before combining them into higher-level representations of ingredients and dishes and achieved a 90% classification accuracy. However, the model was limited by its computational cost and training time, that hindered scalability and real-time deployment in resource-constrained environments.

Dey et al. [12] proposed SmartNoshWaste, a blockchain-based multi-layered system to reduce food waste within the Farm-to-Fork supply chain. Data System Architecture comprising blockchain technology, QR codes and cloud computing to digitize and store food-related data and the ML module employing Q-learning-based reinforcement learning to optimize decisions formed the two basic blocks of the framework. Production, processing, distribution, retailing and consumption were the five main supply chain phases that were functionally tracked and examined. Real-world potato waste data from the nosh app was used for experimental evaluations, and a 9.46% decrease in food waste compared to the baseline data was observed. The dependence on a single food item and context-specific evaluation limited the generalizability and scalability to larger agricultural or urban food systems.

Nascimento et al. [13] developed an AI-driven model for reducing food waste through improved production planning of own-branded products in grocery stores in Brazil. Using historical daily sales data for a year, the study compared five ML algorithms: logistic regression (LR), multilayer perceptron (MLP), DT (J48), PART and random forest (RF). The algorithms forecasted revenue levels, convertible into product demand and rule-based models and RF outperformed others with 90.17% accuracy. A considerable reduction in total food waste across 312 days was observed: from 6,169 kg under aggressive strategies and 653 kg under balanced strategies to just 117.64 kg, representing an 82% reduction. The insights were limited, as the focus was on a single grocery store, restricting the applicability to broader retail settings or more complex supply chains.

Rasyidi et al. [14] aimed to create an elderly-friendly food recording application in Indonesia by overcoming the shortcomings of existing datasets and models that failed to capture the complexity of local cuisine. The study introduced a new food dataset of 24,427 images covering 160 Indonesian food categories and evaluated 67 models built on 16 state-of-the-art deep learning architectures. EfficientNet V2L achieved the best performance with an accuracy of 85.44% and a top-5 accuracy of 97.84%, outperforming models like ConvNeXt Large and Swin-S. The study was, however, limited by difficulties with single-label classification, variations in food presentation, and complex image compositions that hampered generalization across real-world food recognition scenarios.

Jacob et al. [15] utilized AI-driven optimization for food waste reduction in the cassava processing supply chain to transform farm-to-table operations for Garri production. The study utilized data collected from 4,200 respondents across seven states and 42 local government areas, with demand forecasting and inventory optimization performed by the AI framework incorporating regression analysis and DT algorithms, while natural language processing (NLP) analyzed qualitative interview data to extract stakeholder insights. The framework utilized data mining in detecting inefficiencies and waste points across the supply chain, enabling targeted interventions. A 40% reduction in food waste, 30% decrease in processing time and a 25% cut in transportation costs were observed due to optimized logistics and production scheduling. Scalability was a concern as implementation phase was hindered by poor data quality and inadequate infrastructure affecting the consistency of the AI deployment across the supply chain.

Hübner et al. [16] conducted a life cycle assessment (LCA) of "Foodforecast", a cloud-based ML service reducing bakery food waste comprising sales forecasting, cloud computing infrastructure and hardware resources. Historical sales data was utilized to generate optimized forecasts via ML algorithms to minimise overproduction and quantifying environmental impacts through life cycle modelling. The study evaluated four environmental impact categories: global warming, abiotic resource depletion, cumulative energy demand and freshwater eutrophication and achieved an average 30% reduction in bakery returns, equating to a total decrease of 2000 tons of waste on evaluation in real-world data from 175 bakeries. However, the methodological uncertainties and inadequate data constrained the generalizability of the study.

Sigala et al. [17] suggested a fully automated AI-based food waste tracking system deployed across various HORECA (Hotel + Restaurant + Cafe/Catering) establishments in Europe. The study integrated a hardware unit comprising a scale and an IoT-enabled camera device, positioned beneath and above food waste bins, respectively. The system detected each waste event by recording the weight of discarded food and simultaneously captured images, which were transmitted to a cloud-based system. The advanced image recognition and DL algorithms detected and segregated food waste into groups, specific items and categories of avoidability and source, enabling tailored operational changes, including portion control, menu redesign and improved food storage, achieving a 23 to 51% reduction in food waste across most sites. The absence of a detailed cost analysis hampered the study.

To improve inventory oversight and minimize food wastage in supermarkets, Li et al. [18] proposed a real-time inventory tracking framework leveraging computer vision techniques. YOLOv5 object detection algorithm formed the core of the method, that utilizes CNN to autonomously detect and enumerate items in shelf images submitted by users. These images are analyzed to extract item data, which is then synchronized with a Firebase database, while the front-end is dynamically updated using a Flutter-based application. The system achieved an image processing time of approximately 0.79 seconds, with a precision of around 50% and a recall rate of 80%. The accuracy declined in scenarios involving densely packed or multi-layered shelf views, suboptimal image angles, and a limited recognition vocabulary led to frequent miscounts and missed detections.

Min et al. [19] proposed a Stacked Global-Local Attention Network (SGLANet) for food recognition by capturing both global and local discriminative features of food images. The study employed the ISIA Food-500 dataset, comprising 399,726 images across 500 diverse categories. The architecture consisted of two complementary subnetworks: one applied hybrid spatial-channel attention to learn global-level characteristics such as texture and shape, while the other leveraged cascaded spatial transformers to identify ingredient-relevant local regions and aggregate regional information. By fusing these global and local representations, SGLANet produced a more comprehensive feature space for classification. The model achieved strong performance, recording a validation accuracy of 90.92%. However, the computational complexity and training time, made the model less practical for real-time deployment in lightweight environments such as mobile devices.

To curtail food waste within agricultural supply chains, Wang [20] introduced an AI-enhanced Decision Support System (DSS) for intelligent inventory management and optimized resource allocation. The system architecture consisted of three integral layers: real-time data acquisition, an AI-driven decision-making core and adaptive learning through continuous feedback loops. To forecast demand, NNs and support vector machines (SVM) were employed, achieving predictive accuracies of 90% and 85%, respectively. Inventory distribution across the supply chain was optimized using heuristic algorithms, including genetic algorithms (GA) and particle swarm optimization (PSO), which reduced spoilage by 20% and 22%, respectively. High dependency on uninterrupted, high-quality data streams presented challenges for deployment in rural or low-resource agricultural environments.

Rodrigues et al. [21] compared four ML models in forecasting demand within food catering services to minimize food waste arising from overproduction or underestimation. The study employed RF, LightGBM, Long Short-Term Memory (LSTM) networks and Transformer-based models. Two baseline approaches, a naïve model and a moving average method, were developed to simulate conventional forecasting practices and served as benchmarks across three distinct food service settings. The RF algorithm delivered the most accurate predictions in two of the cases, whereas the LSTM model demonstrated superior performance in the third, collectively contributing to food waste reductions ranging from 14% to 52%, and lowering unmet demand by up to 16% relative to the baselines. The study's practical scope was constrained as the focus was on a single dish, restricting applicability in more varied or complex menu environments.

Goh & Yann [22] suggested an Inception-V3 model with transfer learning for food image classification using the Food-101 dataset. The model leveraged convolutional layers for extracting deep spatial features, while transfer learning from ImageNet enhanced its recognition capability. The architecture operated by factorizing convolutions into smaller kernels and incorporating auxiliary classifiers, while retaining depth for complex feature extraction. The transfer learning setup adapted pre-trained ImageNet weights to food-specific features, enabling faster convergence and more accurate recognition. The study achieved an accuracy of 90%, but was limited by the intra-class variability and noise within the dataset, that reduced classification stability.

Faezirad et al. [23] proposed a ML-based forecasting approach to reduce food waste in an Iranian university dining systems by predicting student attendance using both reservation and behavioral data and identified fluctuating student presence as the key contributor in food surplus. The two-stage prediction framework employed an artificial neural network (ANN) to model both deterministic and stochastic aspects of demand. In the first stage, the ANN generated a point estimate for student turnout and the system optimized the total operational cost, balancing the trade-off between food wastage and shortage penalties in the second stage. The model achieved accuracy rates between 73% and 75%, resulting in a reported 79.66% reduction

in food waste over a one-year period. The dependence on historical behavioral data from a single institution limited the model's adaptability to other educational environments.

Malefors et al. [24] proposed ML-based forecasting models to anticipate guest attendance in public catering facilities, for optimizing meal preparation and minimizing food waste under COVID-19 pandemic conditions. Attendance data from 18 primary schools and 16 preschool kitchens across Sweden, both pre- and during-pandemic periods, was collected and evaluated using several AI techniques, including ANN, Poisson autoregressive models and RF algorithms. RF model demonstrated superior accuracy with a conditional mean absolute error below 0.15 for training data and between 0.448 and 0.487 for kitchen-level forecasts. The implementation of forecasting yielded financial savings estimated between €921 and €1,298 compared to operations without predictive support. The models exhibited limited reliability during the initial phase of the pandemic and were challenged by abrupt surpluses, indicating reduced robustness under extreme uncertainty.

A. Research Gap

Despite the progress made in food image recognition and subsequent food waste recognition using deep learning architectures such as ResNet, InceptionResNetV2, MobileNetV2, EfficientNet and SGLANet on large-scale image datasets, these studies have primarily addressed the challenge of food categorization rather than the more critical issue of food waste reduction [11] [19] [22]. In parallel, approaches based on predictive analytics, regression models, neural networks, decision trees, reinforcement learning and optimization techniques, including genetic algorithms, random forests, LSTM networks, and Transformer-based forecasting have been widely applied for inventory management, demand forecasting and supply chain optimization [13] [15] [21]. These models have demonstrated measurable reductions in waste but are fundamentally dependent on structured transactional or sensor-based data streams, overlooking the immediate and non-invasive potential of image-based monitoring. This divergence highlights a critical research gap: existing studies either focus on food recognition without waste-oriented outcomes or address waste reduction without leveraging visual cues. Thus, there is a pressing need for an integrated framework with freshness-annotated image data to deliver accurate shelf-life prediction and provide actionable strategies for minimizing food waste in real-world service operations.

III. Materials and Methods

The proposed hybrid model utilizes image data to precisely estimate the remaining shelf-life of prepared foods by combining convolutional and transformer-based architectures. The study makes use of Food Image Classification dataset that has undergone rigorous preprocessing and augmentation and an average shelf-life table is employed for reference in the training stage. Rich spatial characteristics are extracted from input images by the feature extractor, VGG-19, and then used by the Vision Transformer (ViT) to represent temporal and contextual dependencies for accurate shelf-life prediction. Fig. 1 illustrates the basic architecture of proposed model.

A. Dataset Description

The dataset utilised in this study is the Food Image Classification Dataset [25], from a publicly accessible Kaggle repository, comprising around 24,000 high-resolution images spanning 34 diverse food categories, covering a wide range of both Indian and Western cuisines, representing a realistic mix of freshly prepared and commonly consumed dishes. The images were collected from various sources and curated to ensure diversity in lighting conditions, presentation styles and angles, with each food image labelled according to its class. In restaurant and catering inventory analysis, where food items are prepared in bulk and require regular monitoring, the dataset serves as an essential analytical tool. The construction of sophisticated structures, expiration trends and utilisation strategy optimisation are made possible by the visual diversity and class annotations. This makes the dataset perfect for food service operations' waste minimisation and real-time shelf-life estimate algorithms. Fig. 2 represents the sample images in the dataset over different food categories.

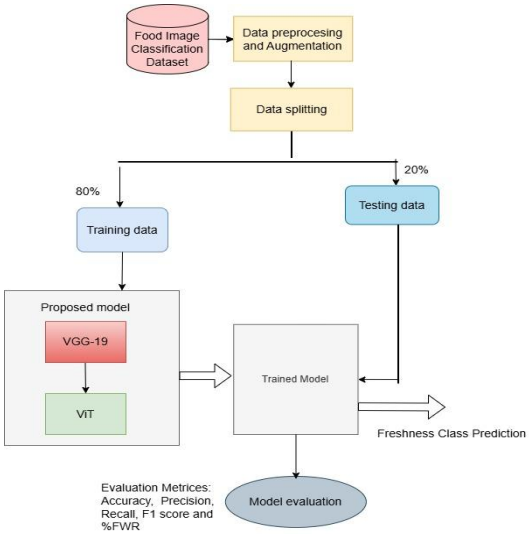

Fig. 1. Basic architecture of the proposed model.

Fig. 2. Sample images in the food image classification dataset.

B. Exploratory Data Analysis

In order to find patterns, anomalies and underlying structures in the dataset, exploratory data analysis or EDA, is an essential step. EDA enables to assess the quality and variability of the images, identify class imbalances and evaluate the distribution of classes in image classification tasks. Additionally, it offers information on possible redundancy, noise or labelling irregularities that have the potential to impair model performance. EDA guarantees well-informed judgments being taken prior to preprocessing and model building by visualizing and encapsulating the data. Fig. 3 illustrates the distribution of images across the 34 food categories in the dataset. Certain subgroups, such as Baked Potato, Hot Dog, Donut and Crispy Chicken, are well-represented, while others, such as Paani puri, Samosa and Kulfi, have relatively fewer samples. Addressing the imbalance is of vital importance as it may bring about in biased learning when the model is being trained. By using EDA to identify this distribution, balanced data augmentation and sampling techniques can be developed to enhance model generalization.

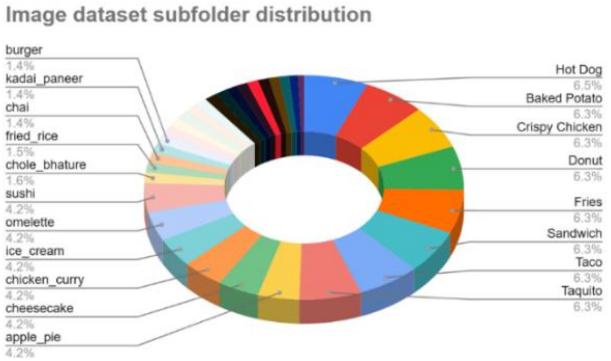

Fig. 3. Image dataset subfolder distribution.

C. Data Preprocessing and Augmentation

A structured data preprocessing and augmentation pipeline is designed for the Food Image Classification dataset to ensure clear and standardized input for the subsequent hybrid architecture. The stage ensures uniformity in image dimensions, address class imbalance and enrich the training set with diverse transformations to reduce overfitting and a compatible data input for the proposed models is generated.

The images are first resized to a fixed dimension of 224×224 pixels, as shown in Eq. (1), conforming to the input requirements of the VGG-19 architecture.

$$I_{resized} = Resize(I_{original}, (H, W)) \quad (1)$$

where, $I_{original}$ is the raw image and $H, W = 224$, the target height and width. Image normalization is further applied to scale pixel intensity values to a standard range, to stabilize the training process and accelerate convergence. The pixel values are normalized to the range [0, 1], as shown in Eq. (2):

$$I_{norm} = \frac{I_{resized}}{255} \quad (2)$$

As the pretrained weights are leveraged, mean subtraction and division by standard deviation were performed as per Eq. (3):

$$I_{norm} = \frac{I_{resized} - \mu}{\sigma} \quad (3)$$

where, μ and σ are the channel-wise mean and standard deviation, respectively. This standardization ensures that each input image contributes uniformly during gradient updates, reducing internal covariate shift. To enable supervised classification, the images are labelled not only by food category but also by freshness state, using a four-class scheme: Fresh, Fit for Consumption, About to Expire and Expired. To numerically represent these classes, label encoding was applied, assigning each category a unique integer value to ensure compatibility with the proposed hybrid model.

Let $C = \{C_0, C_1, C_2, C_3\}$ be the set of freshness categories, where C_0, C_1, C_2, C_3 represent Fresh, Fit for Consumption, About to Expire and Expired, respectively. Each image is tagged based on its freshness level determined by the food's time since preparation t and its standard shelf-life x, as shown in Eq. (4):

$$y_i = \begin{cases} 0\ ;if\ t < 0.2x\ (fresh) \\ 1\ ;\ if\ 0.2x \le t \le 0.8x\ (Fit\ for\ consumption) \\ 2\ ;\ If\ 0.8x < t < x\ (About\ to\ expire) \\ 3\ ;\ If\ t \ge x\ (Expired) \end{cases} \quad (4)$$

The encoded label $yi \in \{0,1,2,3\}$ allows the model to learn the visual differences associated with freshness states, thereby supporting classification tasks aligned with expiration forecasting. Table I illustrates the average shelf-life of various food items.

TABLE I. AVERAGE SHELF LIFE OF VARIOUS FOOD ITEMS

Food Item	Room Temp (hours)	Refrigerated (days)	Frozen (months)
Hot Dog	2	7	1-2
Baked Potato	2	3-5	Not Recommended
Crispy Chicken	2	3-4	4
Donut	24-48	5-7	2-3
Fries	2	2-3	1
Sandwich	3	3	1-2
Taco	2	2-3	1-2
Taquito	2	3-4	2
Apple pie	48	4-5	6-8
Cheese cake	2	5-7	2
Chicken curry	6	3-4	2-3
Ice cream	N/A	30-60	2-4
Omelet	2	2-3	Not Recommended
Sushi	2	1	Not Recommended
Chole bhature	2	2-3	1
Fried rice	6	3-4	1
Tea (Chai)	1	1-2	Not Recommended
Kadai paneer	2	3-4	2
Burger	2	2-3	1
Chapati	24	3-	1
Momos	2	2-3	1-2

Butter naan	24	3-4	1
Pav bhaji	2	2-3	1-2
Idli	24	3-4	1
Dal makhani	4	4-5	2-3
Jalebi	48-72	7	2
Kaathi rolls	24	2-3	1
Pizza	2	3-4	1-2
Masala dosa	2	2-3	Not Recommended
Pakode	2	2-3	1
Dhokla	24	3-4	1-2
Samosa	8	4-5	1
Kulfi	N/A	7-14	2-3
Paani puri	2	1-2	Not Recommended

The dataset is now partitioned into training and testing subsets using an 80:20 split in a randomized fashion. This stratification ensures that the model learns from a broad spectrum of data, while being evaluated on unseen instances. Overall class distribution is maintained to avoid sampling bias and class imbalance. As dataset exhibited class imbalance identified through EDA, the potential of this disparity to create bias during training hampering generalization is considered. Class weights are calculated and added to the loss function to counteract this, allowing the model to penalize minority class misclassification more severely. The weight w_c assigned to class c is determined as in Eq. (5):

$$w_c = \frac{N}{n_c \times C} \tag{5}$$

where, N is the total number of samples and n_c, the number of samples in class c and C is the total number of classes, the four freshness classes. These weights are passed to the training loop so that the model learns more effectively from all classes regardless of frequency.

Data augmentation is applied to the training images to simulate real-world variability in food images, such as differences in angle, lighting and scale, enabling the model to learn invariant features. It includes a number of operations, including random horizontal and vertical flipping, slight rotation, width and height shifts, zooming and brightness variation and the augmented image I' is obtained from the original image I, as shown in Eq. (6):

$$I' = T(I) = R(\theta) \circ Z(s) \circ S(x, y) \circ F(I) \tag{6}$$

where, $R(\theta)$, $Z(s)$ and F denote rotation by angle θ, zooming by scale factor s and flipping transformations, respectively, and $S(x, y)$ represents a shift in width and height. The augmentation steps are applied only to the training dataset to avoid data leakage into test sets. After augmentation, the training dataset increased from 19,200 images to approximately 38,400 images, resulting in a total dataset size of about 43,200 images. By synthetically expanding the dataset, the model is better equipped to handle diverse food presentations and mitigate overfitting.

The images are further converted into tensor format to be efficiently processed by the hybrid model and also to support GPU acceleration. For optimal speed and generalization, the dataset is further segmented into smaller batches, enabling the model to handle multiple samples at once. Prefetching lowers data loading latency and offers more seamless training cycles by preparing the subsequent batch while the current one is being processed.

D. Model Development

1) VGG-19 (Visual Geometry Group-19): VGG-19 is a deep CNN developed to perform robust image classification tasks by learning hierarchical representations of visual data [26]. The VGG-19 network model depicted in Fig. 4, showcases a deep CNN architecture consisting of 19 weight layers. The input image with dimensions 224×224×3 is initially passed through multiple stacked convolutional layers employing small 3×3 kernels with a stride of 1 and padding, followed by ReLU activation functions to introduce non-linearity.

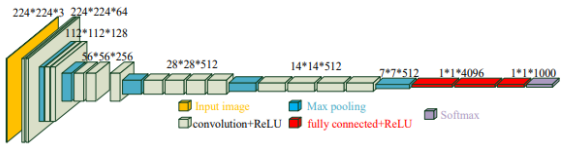

Fig. 4. VGG-19 network model architecture.

As the data progresses through the network, the number of feature maps increases from 64 to 512, while spatial dimensions reduce due to the interleaved max pooling layers which halve the resolution using 2×2 windows. Following the last convolutional block, the flattened feature maps are fed into three fully connected layers, two of which are 4096 in size and one final layer, which is 1000 in size. All of these layers are triggered by ReLU, with the exception of the last one, which outputs class probabilities using Softmax. This particular layer is excluded in the proposed hybrid model. By repeating tiny filters, this design highlights depth and aids in capturing hierarchical representations and fine-grained information. Each convolution layer performs feature extraction, as shown in Eq. (7):

$$X_l = \sigma(W_l * X_{l-1} + b_l) \tag{7}$$

where, X_{l-1} is the input feature map to layer l and W_l, b_l are trainable weights and biases of l. σ and $*$ denotes the ReLU activation function and convolution operation, respectively. The change in feature map size after convolution can be expressed as in Eq. (8) and Eq. (9):

$$H_{out} = \left\lfloor \frac{H_{in} - K + 2P}{S} \right\rfloor + 1 \tag{8}$$

$$W_{out} = \left\lfloor \frac{W_{in} - K + 2P}{S} \right\rfloor + 1 \tag{9}$$

where, H_{in} and W_{in} are the input height and width, respectively. K, P and S represent the kernel size, padding and stride, respectively. The max-pooling operation reduces spatial dimensions while preserving critical information and is defined as in Eq. (10):

$$X_{pooled}(i, j) = \max_{(m,n) \in \Omega} X(i + m, j + n) \tag{10}$$

where, Ω represents the pooling window, (i, j) are the top-left coordinates of the pooling window and m, n are the indices over the local region. The fully connected operation that follows convolution and pooling is defined as in Eq. (11):

$$z = W_{fc} \cdot x + b_{fc} \tag{11}$$

where, x is the flattened input feature vector and z is the linear transformation output. For the proposed hybrid model, the final classification layer is removed, and the intermediate feature representation is fed into downstream models for task-specific inference. The final feature maps are flattened and passed through fully connected layers, producing a high-level embedding.

2) Vision Transformer (ViT): ViT adapts the self-attention-based architecture of transformers from NLP to computer vision tasks by processing images as sequences of patches instead of relying on convolutional filters [27]. It segments an image into fixed-size patches, embeds positional information in these patches and uses self-attention mechanisms to model global contextual relationships throughout the visual field. ViT achieves better performance in applications requiring holistic image processing by capturing long-range relationships through stacked transformer layers, in contrast to convolutional networks that rely on local receptive fields. Fig. 5 illustrates the architecture of ViT.

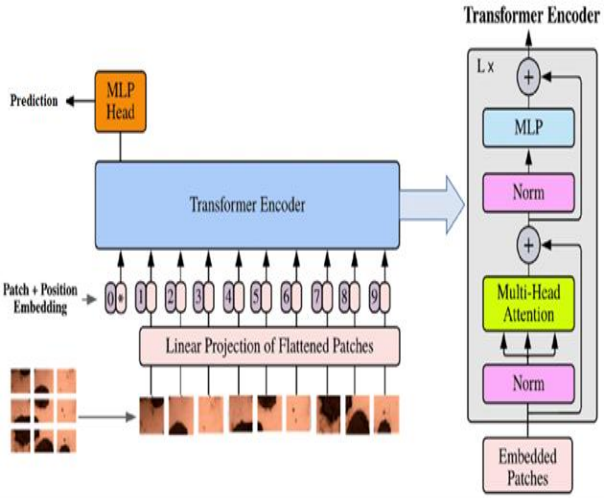

Fig. 5. Basic architecture of ViT.

The process begins with patch embedding, where the input image $X \in \mathbb{R}^{H \times W \times C}$ is divided into $N = \frac{HW}{P^2}$ non-overlapping patches of size $P \times P$. Each patch is then flattened and projected into a latent vector using a trainable linear projection, as shown in Eq. (12):

$$z_0^i = x^i \cdot E \; ; for \; i = 1,2, \dots, N \tag{12}$$

where, $x^i \in \mathbb{R}^{P^2 \cdot C}$ is the i-th patch and $E \in \mathbb{R}^{(P^2 \cdot C) \times D}$ is the embedding matrix. Spatial information lost in flattening is retained by adding positional embeddings $E_{pos} \in \mathbb{R}^{(N+1) \times D}$ to the patch embeddings, including a special learnable CLS token used for classification, as shown in Eq. (13):

$$z_0 = [x_{cls}; x^1 E; x^2 E; \dots ; x^N E] + E_{pos} \tag{13}$$

The embedded sequence is then passed through L identical transformer encoder blocks, each comprising a Multi-Head Self-Attention (MSA) layer and a Multi-Layer Perceptron (MLP) block, both followed by Layer Norm and residual connections, as shown in Eq. (14) and Eq. (15):

$$z_l' = MSA(LN(z_{l-1})) + z_{l-1} \tag{14}$$

$$z_l = MLP(LN(z_l')) + z_l' \tag{15}$$

where, $l = 1,2, \dots, L$ and LN denotes Layer Normalization. The core of ViT lies in self-attention, that models dependencies across all patches. For each attention head, the attention scores are computed, as shown in Eq. (16):

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{16}$$

where, $Q = zW^Q, the \; query$, $K = zW^K$, the key and $V = zW^V$, the value projection. d_k represents the dimension of the key vectors. The output corresponding to the CLS token after the final transformer block is passed through an MLP head to produce the final class probabilities, as shown in Eq. (17):

$$y = softmax\ (W_{head} \cdot z_L^{(cls)}) \tag{17}$$

3) Proposed VGG 19-ViT hybrid model: The proposed hybrid model integrates the robust spatial feature extraction capabilities of VGG-19 with the temporal reasoning and attention mechanisms of ViT to effectively classify food items based on their visual freshness condition. The VGG-19 network processes the input image by passing it through a deep convolutional stack, extracting hierarchical spatial features from local edges to complex textures. These learned features, represented as a dense feature map, are then flattened and embedded into a sequential patch representation suitable for transformer architecture input.

Once patch embeddings are generated, they are forwarded to the ViT module, which applies a series of transformer encoder blocks to model inter-patch dependencies and positional relationships. Through multi-head self-attention and feed-forward networks, ViT generates a high-dimensional latent representation that captures both the visual condition and aging indicators of the food item. The model is trained to classify each food image into one of four predefined freshness classes: Fresh, Fit for Consumption, About to Expire and Expired. This prediction assists in intelligent inventory decisions such as prioritizing consumption or removal, without the need for an explicit expiration timestamp. By leveraging both spatial precision and temporal reasoning, the hybrid VGG-19–ViT model shows strong potential in real-time food quality monitoring, dynamic freshness tracking and minimizing food waste in high-throughput environments like restaurants, canteens or retail kitchens.

E. Simulation Setup

The proposed hybrid VGG-19-ViT model was implemented in a high-performance computing environment configured with an Intel Core i7 processor, 32 GB RAM and an NVIDIA Tesla T4 GPU to ensure robust training and inference performance.

The model was developed using the TensorFlow framework with the Keras API, leveraging its modularity and advanced capabilities for implementing custom DL architectures. Model training, evaluation and prototyping were executed on Google Colaboratory Pro, which provided accelerated GPU support and cloud infrastructure optimizing both training time and memory efficiency. To ensure optimal learning, convergence speed and generalization capability, key hyperparameters were carefully selected through empirical tuning and preliminary examinations. Table II shows the hyperparameter specifications utilized in the proposed model.

TABLE II. HYPERPARAMETER SPECIFICATIONS

Hyperparameters	Values
Optimizer	ADAM
Activation Function	Softmax
Loss Function	Categorical Cross-entropy
Batch Size	32
Epochs	50
Learning Rate	0.0001
Dropout Rate	0.2
Number of Transformer Layers	12
Attention Heads (ViT)	12

IV. RESULTS

A set of standard evaluation metrics has been utilized for in depth performance evaluation of the proposed model, as shown in Eq. (18) to Eq. (22). True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) are mathematically computed using confusion matrix core elements. Different metrics offer unique insights of the model performance, overall correctness by accuracy, precision and recall highlight the ability of model to correctly predict food freshness stage without excessive misses or false alarms.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (18)$$

$$Arecision = \frac{TP}{TP+FP} \quad (19)$$

$$Recall = \frac{TP}{TP+FN} \quad (20)$$

$$F1\ score = 2 \times \frac{precision \times Recall}{Precision+Recall} \quad (21)$$

The % Food Waste Reduction calculates the percentage of food waste that is reduced as a result of the proposed model's timely forecasts, especially when products are identified as "About to Expire". It is calculated based on the weight of food consumed before expiration, via predictions by the proposed model, relative to the total food that would have been expired and wasted otherwise. The analysis considered standard portion sizes and average weight estimates for each food category, allowing a precise quantification of rescued food in kilograms. This indicator evaluates how the model helps food service settings manage waste and inventory in a sustainable way.

$$\%FWR = \frac{W_{baseline} - W_{model}}{W_{baseline}} \times 100 \quad (22)$$

where, $W_{baseline}$ and W_{model} are the total waste generated without and with the proposed model, respectively. A higher $\%FWR$ indicates greater effectiveness in reducing waste, showcasing the model's practical utility in inventory decision-making.

The accuracy plot of the proposed model, illustrated in Fig. 6, reveals a consistently improving learning trajectory over 50 epochs, underscoring the model's effective optimization. Initially, both training and validation accuracy rise sharply during the first 10 epochs, indicating rapid feature learning. The training accuracy steadily improves and reaches near-saturation beyond epoch 20, surpassing 99%, while the validation accuracy stabilizes just below 98%, suggesting robust generalization on unseen data. The gap between the two curves remains minimal, indicating low overfitting and well-regularized model behavior. This convergence confirms the proposed model's ability to learn meaningful representations across diverse food freshness categories and maintain high predictive reliability across both training and validation phases.

Fig. 6. Accuracy plot of the proposed model.

The loss plot of the proposed model, illustrated in Fig. 7, demonstrates the convergence behavior of the proposed model across 50 epochs. Both training and validation losses exhibit a sharp decline during the initial epochs, indicating effective learning and a rapid reduction in prediction error. The training loss continues to decrease steadily, approaching near-zero values by the final epochs, reflecting the model's high confidence on seen data. Meanwhile, the validation loss also follows a decreasing trend and stabilizes around 0.08, suggesting good generalization without signs of significant overfitting. The small and consistent gap between the two curves affirms the robustness and stability of the training process. Overall, the minimized loss values validate the model's capability to distinguish food freshness categories accurately and reliably.

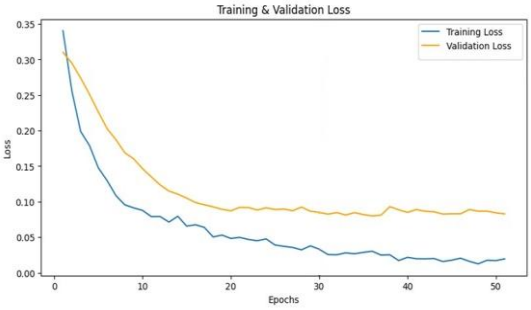

Fig. 7. Loss plot of the proposed model.

The classification report for the proposed model, illustrated in Table III, reflects strong overall performance across all food freshness categories. The model achieves excellent precision and recalls scores for the Fresh and Fit for Consumption classes (0.99/0.99 and 0.99/0.98, respectively), demonstrating excellent ability to consistently and accurately distinguish consumable items. The About to Expire class shows a precision of 0.97 and recall of 0.98, indicating reliable early identification of items nearing spoilage and the Expired class, often difficult to detect due to visual ambiguity, still attains solid performance with a precision of 0.94 and recall of 0.99, minimizing false negatives in disposal-critical cases.

TABLE III. CLASSIFICATION REPORT OF THE PROPOSED MODEL

	Precision	Recall	F1-score
Fresh	0.99	0.99	0.98
Fit for Consumption	0.99	0.98	0.98
About to expire	0.97	0.98	0.97
Expired	0.94	0.99	0.96
Accuracy	0.98		
macro avg	0.97	0.98	0.98
weighted avg	0.98	0.98	0.98

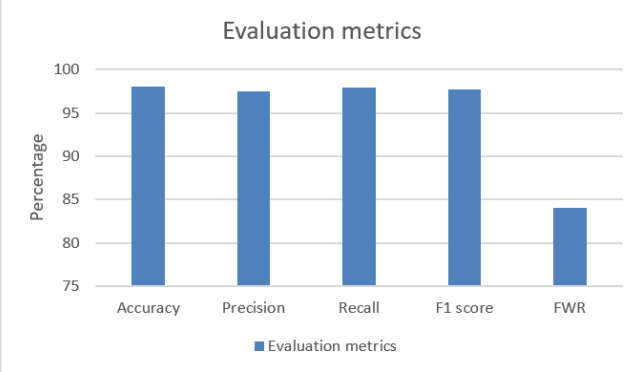

Fig. 8. Evaluation metrics of proposed model.

The evaluation metrics illustrated in Fig. 8 shows a 98% overall accuracy rate confirming that the model can accurately identify most food categories in the test dataset. With a precision score of 97.5%, the model minimizes false positives and proves to be quite dependable, while the 97.9% recall indicates how well the model captures almost all real, pertinent events, crucial for promptly identifying perishable or expired goods. Furthermore, the F1-score of 97.75% verifies a balanced performance, guaranteeing that predictions are accurate and comprehensive. The proposed model significantly enhanced food sustainability by enabling real-time freshness detection and shelf-life prediction, achieving a remarkable 84% reduction in food waste. The proposed model's real-time applicability in minimizing food loss across dynamic food service operations were demonstrated by its excellent capability to detect early signs of deterioration that allowed for timely consumption or redirection.

The confusion matrix illustrated in Fig. 9 analyses the performance of the proposed model across four food freshness classes. The model shows strong predictive accuracy, with particularly high true positive counts for Fresh (1443), Fit (for consumption) (1880), About to Expire (930) and Expired (448) categories. Misclassifications are minimal across all classes, with most errors occurring between neighboring freshness stages, reflecting the natural difficulty in distinguishing borderline cases. The overall accuracy of 98% confirms the model's robustness and high generalization capability in real-world food freshness classification.

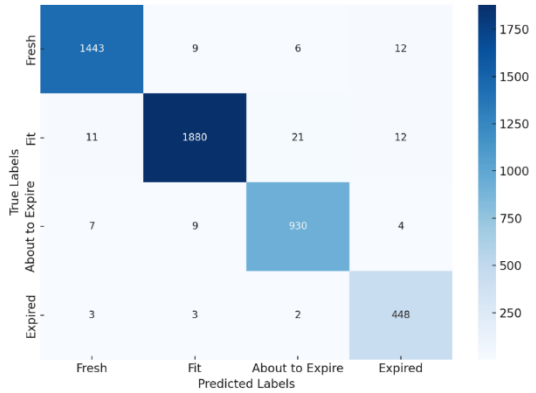

Fig. 9. Confusion matrix of the proposed model.

Fig. 10 represents the results of the proposed model, showcasing a variety of correctly predicted samples across all four freshness classes. Each image includes ground truth and predicted labels, demonstrating the model's ability to generalize across different food types and states with high visual accuracy.

Fig. 10. Results of the proposed model.

Table IV and Fig. 11 illustrate the comparative performance of different architectures applied to various food image datasets. InceptionResNetV2 trained on a Korean food image dataset achieved an accuracy of 81.91%, reflecting the challenges posed by diverse regional cuisine representations. ResNet-50 and Inception-V3, both evaluated on the Food-101 dataset, attained accuracies of 90%, but were constrained by dataset-level variability. EfficientNet V2L on the Indonesian food dataset recorded an accuracy of 85.44%, indicating limitations posed by complex food presentation styles. The SGLANet model applied to the ISIA Food-500 dataset demonstrated a validation accuracy of 90.92%. In contrast, the proposed VGG-19–ViT hybrid model applied to the Food Image Classification dataset achieved a substantially higher accuracy of 98%, clearly outperforming all other architectures. This result underscores the effectiveness of integrating convolutional feature extraction with transformer-based attention mechanisms and highlights the advantage of employing a freshness-annotated dataset tailored for food waste reduction.

TABLE IV. ACCURACY COMPARISON OF THE PROPOSED MODEL WITH EXISTING METHODS

Author [Ref]	Food Image Dataset Used	Method	Accuracy (%)
Chun et al. [10]	Korean Food Image Dataset	InceptionResNetV2	81.91
Louro et al. [11]	Food-101	ResNet-50	90
Rasyidi et al. [14]	Indonesian Food Dataset	EfficientNet V2L	85.44
Min et al. [19]	ISIA Food-500	SGLANet	90.92
Goh & Yann [22]	Food-101	Inception-V3	90
Proposed model	**Food Image Classification**	**VGG-19-ViT hybrid model**	**98**

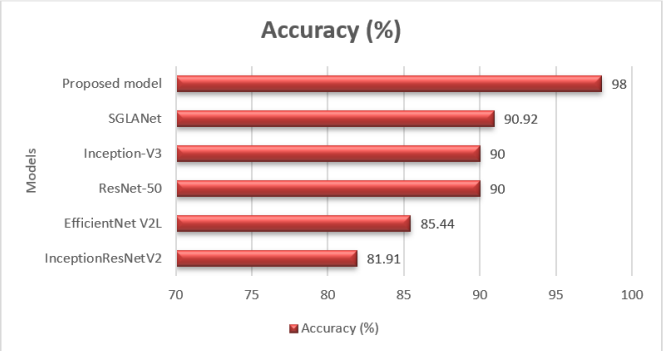

Fig. 11. Accuracy comparison of the proposed model with existing methods.

Table V and Fig. 12 illustrate the comparative performance of different approaches to food waste reduction, evaluated solely on the basis of the percentage of waste reduction achieved, irrespective of whether the methods relied on traditional survey-based strategies, machine learning forecasting, or real-time monitoring systems. Reported reductions vary widely, from 9.46% with SmartNoshWaste to intermediate values of 22% with neural networks and heuristic optimization, 30% through cloud-based machine learning with life cycle assessment, and 40% with AI-driven supply chain optimization. Higher reductions were observed in food service and institutional settings, with tracking systems in HORECA achieving 51% and forecasting models combining random forests and LSTM networks achieving 52%. Advanced predictive models showed further promise, with artificial neural networks reaching 79.66% and random forest–based planning attaining 82%. In comparison, the proposed VGG-19–ViT hybrid model delivered the highest performance with an 84% reduction in food waste, demonstrating the effectiveness of image-based freshness prediction in directly supporting sustainability goals.

TABLE V. % FWR COMPARISON OF THE PROPOSED MODEL WITH EXISTING METHODS

Author [Ref]	Model	% FWR
Dey et al. [12]	SmartNoshWaste	9.46%
Nascimento et al. [13]	RF	82%
Jacob et al. [15]	AI-Driven Optimization	40%
Hübner et al. [16]	Cloud ML + LCA	30%
Sigala et al. [17]	Waste Tracking in HORECA	51%
Wang [20]	NN, SVM with GA & PSO	22%
Rodrigues et al. [21]	RF, LSTM	52%
Faezirad et al. [23]	ANN	79.66%
Proposed model	**VGG-19-ViT hybrid model**	**84%**

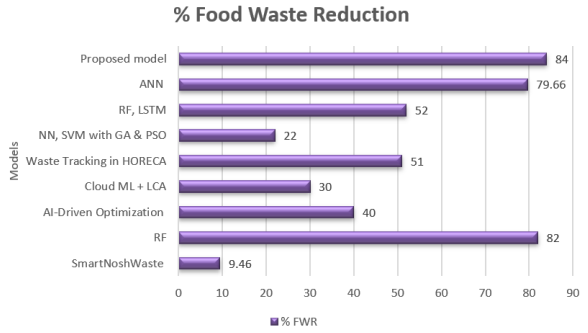

Fig. 12. % FWR comparison of the proposed model with existing methods.

V. DISCUSSION

The results clearly demonstrate that the proposed model delivers strong and reliable performance in food freshness classification. The very high precision and recall scores in the Fresh and Fit for Consumption classes confirm the model's ability to accurately identify items suitable for use, minimizing the risk of unnecessary disposal. The slightly lower precision observed in the Expired class, despite its excellent recall, indicates that the model tends to prioritize capturing all true expired items, which is advantageous for avoiding false negatives in disposal-critical cases. Similarly, the About to Expire class results reflect the model's strength in early spoilage detection, a key factor in reducing waste through timely redirection or consumption. The overall accuracy of 98% and the balanced F1-score highlight the stability of the model across categories. Beyond numerical accuracy, the observed 84% reduction in food waste illustrates the real-world impact of integrating such a system into food service operations, where early and dependable freshness prediction can transform management practices and contribute to sustainability goals.

VI. CONCLUSION

The study proposed a novel hybrid deep learning model integrating VGG-19 and Vision Transformer (ViT) architectures for intelligent monitoring of food freshness and predicting the remaining shelf-life of prepared food items. A meticulously selected food image dataset with 34 distinct food categories that represented actual restaurant and catering settings was employed in training the model. The framework successfully categorized food images into four freshness classes: Fresh, Fit for Consumption, About to Expire and Expired, by employing VGG-19 to extract deep visual characteristics and ViT to predict temporal degradation trends. In doing so, the study makes a theoretical contribution by demonstrating how convolution-based fine-grained feature extraction can be effectively fused with transformer-based self-attention to capture both spatial and temporal degradation cues, offering a new direction for freshness prediction models in food technology. The classification greatly reduced food waste and enabled prompt consumption decisions. The proposed model achieved 98% accuracy, 97.5% precision, 97.9% recall, 97.75% F1-score, and an estimated 84% reduction in food waste, underscoring its strong potential for real-world deployment. The method remains scalable and lightweight by integrating freshness estimation without requiring a large amount of sensor data. Nevertheless, the study is not without limitations. The reliance on a single image modality may restrict generalization under varied lighting or presentation conditions, and the dataset, while diverse, is smaller in scale compared to global benchmarks. Additionally, external factors such as storage temperature or humidity were not integrated into the model. Despite these constraints, the study contributes to the domain by establishing that purely image-based hybrid architectures can reliably predict freshness levels and directly translate to measurable sustainability outcomes in food service operations. Future research could explore integration with real-time kitchen inventory systems, multi-modal learning using sensors, and deployment in edge devices for low-resource settings. Further studies may also validate the framework in larger-scale, multi-institutional datasets and investigate explainability mechanisms to improve model transparency for end-users. Such extensions would further enhance the impact and global applicability of the proposed model in minimizing food waste.

REFERENCES

[1] M. Lai, A. Rangan, and A. Grech, “Enablers and barriers of harnessing food waste to address food insecurity: A scoping review,” Nutr. Rev., vol. 80, no. 8, pp. 1836–1855, Aug. 2022.

[2] A. Pandey, “Food wastage: Causes, impacts and solutions,” Sci. Herit. J., vol. 5, pp. 17–20, 2021.

[3] N. Aloysius, J. Ananda, A. Mitsis, and D. Pearson, “Why people are bad at leftover food management? A systematic literature review and a framework to analyze household leftover food waste generation behavior,” Appetite, vol. 186, p. 106577, Feb. 2023.

[4] L. Wei, G. Prabhakar, and L. N. Duong, “Usage of online food delivery in food waste generation in China during the crisis of COVID-19,” Int. J. Food Sci. Technol., vol. 58, no. 10, pp. 5602–5608, Oct. 2023.

[5] Y. Shigetomi, A. Ishigami, Y. Long, and A. Chapman, “Curbing household food waste and associated climate change impacts in an ageing society,” Nat. Commun., vol. 15, no. 1, p. 8806, Jan. 2024.

[6] C. Sundgren, “Circular supply chain relationships for food redistribution,” J. Clean. Prod., vol. 336, p. 130393, Jan. 2022.

[7] V. Amicarelli, A. C. Aluculesei, G. Lagioia, R. Pamfilie, and C. Bux, “How to manage and minimize food waste in the hotel industry: An exploratory research,” Int. J. Cult. Tour. Hosp. Res., vol. 16, no. 1, pp. 152–167, Jan. 2022.

[8] A. Režek Jambrak, M. Nutrizio, J. Dukić, I. Djekić, M. Vinceković, S. Jurić, and F. Donsi, “Digitalisation, bioinformatics, and delivery systems in sustainable nonthermal extraction of proteins,” Int. J. Food Sci. Technol., vol. 60, no. 1, p. vvae038, Jan. 2025.

[9] R. N. Arshad, Z. Abdul-Malek, C. Parra-López, A. Hassoun, M. I. Qureshi, A. Sultan, and G. Garcia-Garcia, “Food loss and waste reduction by using Industry 4.0 technologies: Examples of promising strategies,” Int. J. Food Sci. Technol., vol. 60, no. 1, p. vvaf034, Jan. 2025.

[10] M. Chun, H. Jeong, H. Lee, T. Yoo, and H. Jung, “Development of Korean food image classification model using public food image dataset and deep learning methods,” IEEE Access, vol. 10, pp. 128732–128741, 2022.

[11] J. Louro, F. Fidalgo, and Â. Oliveira, “Recognition of food ingredients—dataset analysis,” Appl. Sci., vol. 14, no. 13, p. 5448, Jul. 2024.

[12] S. Dey, S. Saha, A. K. Singh, and K. McDonald-Maier, “SmartNoshWaste: Using blockchain, machine learning, cloud computing and QR code to reduce food waste in decentralized web 3.0 enabled smart cities,” Smart Cities, vol. 5, no. 1, pp. 162–176, Mar. 2022.

[13] A. M. Nascimento, A. Queiroz, V. V. de Melo, and F. S. Meirelles, “Applying Artificial Intelligence to reduce food waste in small grocery stores,” Unpublished.

[14] M. A. Rasyidi, Y. S. Mardhiyyah, Z. Nasution, and C. H. Wijaya, “Performance comparison of state-of-the-art deep learning model architectures in Indonesian food image classification,” Bull. Electr. Eng. Inform., vol. 13, no. 5, pp. 3355–3368, Oct. 2024.

[15] L. A. Jacob and O. L. U. F. E. M. I. Omitoyin, “AI-driven farm-to-table supply chain optimization: A case study of reducing food waste,” J. Entomol. Agron. Stud., 2024.

[16] N. Hübner, J. Caspers, V. C. Coroamă, and M. Finkbeiner, “Machine-learning-based demand forecasting against food waste: Life cycle environmental impacts and benefits of a bakery case study,” J. Ind. Ecol., vol. 28, no. 5, pp. 1117–1131, May 2024.

[17] E. G. Sigala, P. Gerwin, C. Chroni, K. Abeliotis, C. Strotmann, and K. Lasaridi, “Reducing food waste in the HORECA sector using AI-based waste-tracking devices,” Waste Manag., vol. 198, pp. 77–86, Jan. 2025.

[18] T. Li and Y. Sun, “An intelligent food inventory monitoring system using machine learning and computer vision,” Unpublished, 2022.

[19] W. Min, L. Liu, Z. Wang, Z. Luo, X. Wei, X. Wei, and S. Jiang, “ISIA Food-500: A dataset for large-scale food recognition via stacked global-local attention network,” in Proc. 28th ACM Int. Conf. Multimedia, Oct. 2020, pp. 393–401.

[20] H. Wang, “Agricultural product supply chain inventory control and allocation decision support system integrating artificial intelligence technology,” Int. J. High Speed Electron. Syst., p. 2540556, 2025.

[21] M. Rodrigues, V. Migueis, S. Freitas, and T. Machado, “Machine learning models for short-term demand forecasting in food catering services: A solution to reduce food waste,” J. Clean. Prod., vol. 435, p. 140265, Jan. 2024.

[22] A. M. Goh and X. L. Yann, “Food-image classification using neural network model,” Int. J. Electron. Eng. Appl., vol. 9, no. 3, pp. 12–22, 2021.

[23] M. Faezirad, A. Pooya, Z. Naji-Azimi, and M. A. Haeri, “Preventing food waste in subsidy-based university dining systems: An artificial neural network-aided model under uncertainty,” Waste Manag. Res., vol. 39, no. 8, pp. 1027–1038, Aug. 2021.

[24] C. Malefors, L. Secondi, S. Marchetti, and M. Eriksson, “Food waste reduction and economic savings in times of crisis: The potential of machine learning methods to plan guest attendance in Swedish public catering during the COVID-19 pandemic,” Socio-Econ. Plan. Sci., vol. 82, p. 101041, Jan. 2022.

[25] https://www.kaggle.com/datasets/gauravduttakiit/food-image-classification

[26] T. H. Nguyen, T. N. Nguyen, and B. V. Ngo, "A VGG-19 model with transfer learning and image segmentation for classification of tomato leaf disease," AgriEngineering, vol. 4, no. 4, pp. 871–887, Dec. 2022.

[27] C. Xia, X. Wang, F. Lv, X. Hao, and Y. Shi, "ViT-COMER: Vision Transformer with convolutional multi-scale feature interaction for dense predictions," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2024, pp. 5493–5502.