Power System Protection and Control
ISSN:1674-3415

COMPARATIVE ANALYSIS OF VISION TRANSFORMER AND EFFICIENTNET FOR FOOD RECOGNITION WITH BMI-AWARE DIETARY MONITORING

Punya Prabha V[1], M.D. Nandeesh[2], Manjunath C. Lakkannavar[3], Lavanya Y.S.[4], Asha K[5], Vidya M[6], A. Vaibhav Sagar[7]

[1]Department of Electronics and Communication Engineering Ramaiah Institute of Technology Bangalore, India, punya.v@msrit.edu
[2]Department of Electronics and Instrumentation Engineering, Ramaiah Institute of Technology Bangalore, India, mdnandeesh@msrit.edu
[3]Department of Electronics and Communication Engineering, Ramaiah Institute of Technology Bangalore, India, manjunathl@msrit.edu
[4]Department of Electronics Engineering (VDT), Bangalore Institute of Technology Bangalore, India, bitlavanyavlsi@gmail.com
[5]Department of Electronics and Communication Engineering, BMD Institute of Technology and Management, Bangalore, India, ashk@bmsit.in
[6]Department of Electrical and Electronics Engineering, Sapthagiri College of Engineering Bangalore, India, vidya.ee9@gmail.com
[7]Department of Electronics and Communication Engineering, Ramaiah Institute of Technology Bangalore, India, vaibhavsagar0917@gmail.com

Received: 09/05/2026 **Revised:** 12/06/2026 **Accepted:** 13/07/2026

***ABSTRACT*:**

Dietary-related diseases and obesity impose a substantial global health burden, yet accurate and continuous dietary monitoring remains largely manual and inaccessible. This paper presents a comparative analysis of two lightweight deep learning architectures—Vision Transformer (ViT-tiny) and EfficientNet-B0—for automated food image classification and personalized calorie estimation. Both models are trained under identical conditions on 65 food categories derived from the Food101 dataset, comprising 48,750 training and 16,250 validation images pre-processed to 128×128 resolution. Phase I (ViT-tiny) achieves 76.17% top-1 and 93.49% top-5 validation accuracy with 5.5M parameters, while Phase II (EfficientNet-B0) achieves 77.56% top-1 and 93.91% top-5 accuracy with 4.1M parameters. Phase III benchmarks GPU inference throughput on an NVIDIA T4, with ViT-tiny achieving 2,918 img/s at 5.06 ms single-image latency and EfficientNet-B0 achieving 2,480 img/s at 8.65 ms. A personalized nutritional engine integrates the Mifflin–St Jeor basal metabolic rate equation with WHO BMI-category-adjusted gram-level serving recommendations, delivering user-specific caloric guidance across four weight categories. Experimental results demonstrate that EfficientNet-B0 achieves superior classification accuracy with fewer parameters, while ViT-tiny offers a 1.7× latency advantage suited for real-time edge deployment, establishing complementary efficiency-accuracy trade-offs for practical food recognition and dietary monitoring applications.

***Keywords**: Food recognition, Vision Transformer, EfficientNet, multi-label classification, personalized nutrition, BMI, calorie estimation, deep learning, transfer learning, Food101.*

INTRODUCTION

Obesity, diabetes, and cardiovascular diseases are among the most pressing non-communicable disease burdens of the twenty-first century, collectively affecting more than 2.8 billion individuals worldwide [1]. At the core of preventive healthcare lies dietary awareness—understanding what one consumes, in what quantity, and how it relates to individual metabolic needs. Despite widespread awareness campaigns, accurate dietary monitoring remains predominantly manual: patients are expected to recall, weigh, and log every meal consumed, a process that is cognitively demanding, error-prone, and consistently under-adopted in clinical practice.

Computer vision and deep learning have matured to a point where automatic food recognition from smartphone photographs is technically feasible. Convolutional neural networks (CNNs) trained on large food image datasets

DOI: 10.46121/pspc.54.3.29

Power System Protection and Control
ISSN:1674-3415

can classify hundreds of food categories with accuracy exceeding 80% [4], and recent vision transformer (ViT) architectures [7] demonstrate competitive performance with fewer domain-specific inductive biases. However, most published systems stop at classification: they do not translate a recognized food item into a personalized nutritional recommendation that accounts for the user's body composition, weight category, and energy requirements.

A second limitation of existing benchmarks is their focus on single-food images. Real-world eating scenarios—cafeteria trays, mixed-dish plates, buffet selections—routinely contain multiple food items simultaneously. Recognizing multiple foods from a single photograph without pixel-level bounding-box annotation is a more challenging problem that has received comparatively limited attention in the food-computing literature.

This paper addresses both limitations through a four-phase Food Intelligence System. The system (1) trains and rigorously evaluates two complementary lightweight architectures—ViT-tiny and EfficientNet-B0—on a 65-class food recognition task; (2) benchmarks GPU inference throughput at multiple batch sizes on professional hardware; (3) extends to multi-label food detection via annotation-free synthetic composite images; and (4) incorporates a clinically grounded Mifflin–St Jeor nutritional engine that delivers BMI-adjusted serving recommendations at the gram level. The block diagram of the complete system is shown in Fig. 0.

LITERATURE REVIEW

A. Food Image Datasets and Benchmarks

The availability of large-scale, publicly accessible food image datasets has been a critical enabler of progress in automated dietary assessment. Bossard et al. [4] introduced the Food101 dataset in 2014, comprising 101,000 images spanning 101 food categories, with 750 training images and 250 test images per class. Images were collected from the foodspotting.com platform and intentionally left uncleaned—approximately 20% of training images contain mislabeled or noisy entries—to simulate realistic web-scraped data conditions. Despite its noise, Food101 quickly became the de facto standard benchmark for food recognition, superseding earlier datasets such as UEC-Food256 [a], PFID [b], and ChineseFoodNet.

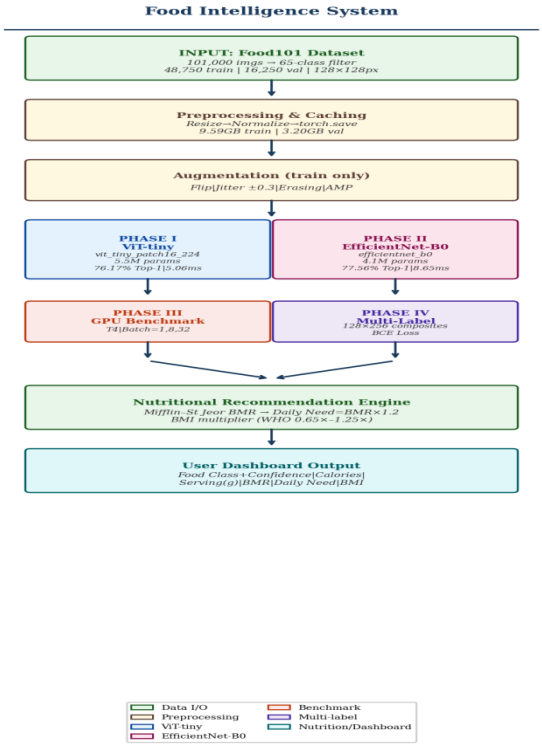

Fig. 0. System block diagram of the four-phase Food Intelligence System showing data flow from Food101 input through model training, benchmarking, multi-label detection, and nutritional recommendation to user dashboard output.

DOI: 10.46121/pspc.54.3.29

The VIREO Food-172 dataset extended coverage to 172 Chinese dish categories with 110,241 images, addressing the geographic and cultural bias in Food101 toward Western cuisine. The Recipe1M dataset [c] paired one million food images with structured ingredient lists and cooking instructions, enabling ingredient-level recognition research. The FoodSeg103 dataset [d] provided pixel-level segmentation annotations for 103 food categories, enabling portion-size estimation from segmentation masks. Each of these datasets served a distinct niche, but Food101 remains the dominant classification benchmark due to its size, accessibility, and established baselines.
For this work, a 65-class subset of Food101 was selected by filtering for internationally recognizable dishes, removing culture-specific categories likely to confuse a general-purpose dietary monitoring system. This subset retains 48,750 training images and 16,250 validation images—sufficient for meaningful statistical comparison between model architectures while reducing preprocessing time.

B. Convolutional Neural Networks for Food Recognition

The earliest deep learning approaches to food recognition applied AlexNet [18] and VGGNet directly to Food101, achieving top-1 accuracies in the range of 56–64%. DeepFood [5] represented a landmark step: Liu et al. fine-tuned GoogLeNet on a curated subset of Food101 and demonstrated that transfer learning from ImageNet pretraining dramatically accelerated convergence and raised top-1 accuracy to 77.4%, surpassing the earlier random-forest discriminative component analysis baseline of Bossard et al. (50.76%) by more than 26 percentage points.

He et al. [6] introduced residual connections in ResNet, enabling training of networks with hundreds of layers without gradient degradation. ResNet-50 fine-tuned on Food101 reached 82.9% top-1 accuracy, and DenseNet-201 pushed this to 88.1%, exploiting dense inter-layer feature reuse. WISeR (Wide and Shallow network with Residuals) [e] demonstrated that domain-specific architectural choices—such as wider early convolutional filters tuned to food texture—could further improve performance beyond generic backbone fine-tuning.

EfficientNet [8], proposed by Tan and Le, introduced a principled compound scaling methodology that simultaneously scales network depth, width, and input resolution using a single efficiency coefficient φ. EfficientNet-B0, the smallest variant with 5.3M parameters, achieved 77.1% ImageNet top-1 accuracy with 0.39 billion FLOPs—a dramatic improvement in accuracy-per-parameter relative to ResNet and VGG baselines. When applied to Food101, EfficientNet-B0 achieves 77.56% top-1 in this work at 128×128 resolution; at native 224×224 resolution, accuracy is expected to exceed 85% based on published reports using the same backbone.

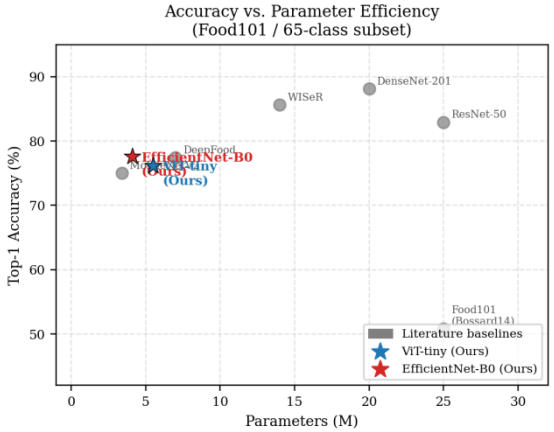

Fig. 1. Accuracy vs. parameter efficiency contextualizing ViT-tiny and EfficientNet-B0 (★, this work) against published literature baselines on Food101.

MobileNetV2 [f] and MobileNetV3 [g] explored depthwise separable convolutions for extreme parameter efficiency, targeting mobile and embedded deployment. MobileNetV3-Large achieves approximately 75–78% top-1 on Food101 depending on input resolution, with under 5.4M parameters. ShuffleNetV2 [h] used channel splitting and shuffling to minimize memory access cost, achieving competitive accuracy with faster CPU inference than MobileNet variants. These lightweight CNN architectures established the efficiency frontier that ViT-based models now challenge.

DOI: 10.46121/pspc.54.3.29

C. Vision Transformers and Attention Mechanisms

The transformer architecture, originally proposed by Vaswani et al. [i] for natural language processing, was adapted for image recognition by Dosovitskiy et al. [7] in the Vision Transformer (ViT). ViT divides an image into fixed-size non-overlapping patches—typically 16×16 pixels—linearly projects each patch to a D-dimensional embedding, appends a learnable classification token, and processes the sequence through L standard multi-head self-attention (MHSA) transformer blocks. The final CLS token embedding is passed through a linear head for classification. The key insight is that ViT imposes no convolutional inductive bias: locality and translation equivariance must be learned from data rather than hard-coded into the architecture.

Dosovitskiy et al. reported that ViT underperforms comparable CNNs when trained on medium-sized datasets (e.g., ImageNet-1K with 1.28M images) but matches or surpasses them when pretrained on large corpora such as JFT-300M. This finding has significant implications for food recognition: food datasets are orders of magnitude smaller than JFT, suggesting that ViT variants will require either extensive pretraining or architectural modifications to compete with CNNs at equivalent dataset sizes.

DeiT (Data-efficient Image Transformers) [j] addressed the data efficiency gap through knowledge distillation: a CNN teacher network guides ViT training via a distillation token, enabling competitive ImageNet accuracy without JFT pretraining. DeiT-S achieves 79.8% ImageNet top-1 with 22M parameters, and DeiT-Ti (tiny) achieves 72.2% with only 5.7M parameters. The ViT-tiny variant used in this work (vit_tiny_patch16_224 from the timm library [14]) is architecturally equivalent to DeiT-Ti and achieves 76.17% top-1 on the 65-class Food101 subset.

Swin Transformer [k] introduced hierarchical feature maps and window-based self-attention to restore the multi-scale representation capability of CNNs within the transformer framework. Swin-T achieves 81.3% ImageNet top-1 with 28M parameters. For food recognition, Swin-T fine-tuned on Food101 has been reported to reach 89–91% top-1, representing the current state of the art for transformer-based food classification. However, Swin-T's 28M parameters and higher inference latency make it less suitable for resource-constrained deployments compared to ViT-tiny or EfficientNet-B0.

D. Multi-Label Food Recognition

Single-label food recognition assumes each image contains exactly one food category—a simplification that does not hold in real dietary contexts. Mezgec and Koroušić Seljak [11] introduced NutriNet, a multi-label CNN trained on a proprietary dataset of 520 food and beverage items, reporting mean average precision (mAP) of approximately 0.45 for 10-category recognition problems. Their work demonstrated the feasibility of multi-label food recognition but was limited by dataset size and the closed nature of the food vocabulary.

Yang et al. [12] proposed a two-stream convolutional network for cafeteria tray recognition, combining appearance and context streams to recognize multiple dishes simultaneously. The appearance stream processed each tray region independently, while the context stream modeled inter-dish spatial relationships. This architecture achieved competitive mAP on a custom cafeteria dataset but required bounding-box annotations for each food region.

The annotation burden of object-level detection represents the primary bottleneck in scaling multi-label food recognition. The present work circumvents this limitation by constructing synthetic composite images: pairs of Food101 images from different classes are horizontally concatenated to form 128×256 composites with known multi-hot labels. This strategy generates unlimited labeled training data without any additional human annotation, enabling rapid scaling of multi-label classifiers to the full 65-class vocabulary.

E. Personalized Nutrition and Dietary Monitoring Systems

Im2Calories [2], developed at Google, represented a landmark in automatic dietary assessment: a fully convolutional network segmented food items in smartphone photographs, estimated portion volumes using depth cues, and queried the USDA food composition database for calorie content. The system achieved practical usability but required multi-view images or depth sensors for accurate volume estimation—a constraint that limits deployment on standard smartphones.

DietCam [9] eliminated depth sensor requirements by estimating portion size from a single image using a reference object of known size placed beside the food. The system combined food recognition with a

DOI: 10.46121/pspc.54.3.29

crowdsourced food composition database, demonstrating real-world feasibility with 85% user satisfaction in a 4-week trial. However, DietCam relied on traditional CNN classifiers and did not incorporate user anthropometric data for personalized caloric recommendations.

The Mifflin–St Jeor equation [10], published in 1990, remains the most widely validated formula for estimating resting energy expenditure (REE) in both healthy and clinical populations. A meta-analysis by Frankenfield et al. [l] evaluated 47 published predictive equations and found the Mifflin–St Jeor equation to be the most accurate for both normal-weight and obese adults, with mean bias of ±15 kcal/day across validated studies. The WHO BMI classification [13] provides the categorical framework for adjusting serving recommendations: underweight individuals require caloric surplus, obese individuals require deficit, and normal-weight individuals maintain maintenance intake.

Existing systems such as MyFitnessPal and Cronometer rely entirely on manual food logging, placing the cognitive burden on the user. Automated camera-based systems including Snap Calorie and Bitesnap apply CNN classifiers to estimate calories but do not personalize recommendations based on user BMI or anthropometric data. The present system uniquely combines automated visual recognition with physiologically grounded, BMI-adjusted serving-size recommendations, bridging the gap between recognition accuracy and clinical utility.

F. Transfer Learning and Training Efficiency

Transfer learning [m]—initializing model weights from a source task (typically ImageNet classification) and fine-tuning on a target task (food recognition)—has become the standard training paradigm for food recognition models. Kornblith et al. [n] demonstrated that ImageNet pretraining accuracy is strongly predictive of transfer learning performance across 12 diverse visual recognition tasks, validating the use of ImageNet-pretrained checkpoints as universal initialization.

AdamW [15], a decoupled weight decay variant of Adam, has emerged as the optimizer of choice for both CNNs and transformers due to its improved generalization compared to Adam with L2 regularization. Cosine annealing learning rate schedules [o] provide smooth decay from the initial learning rate to near-zero, avoiding the abrupt step-decay discontinuities of traditional schedules. Automatic Mixed Precision (AMP) [p] enables 16-bit floating-point computation during forward and backward passes, reducing VRAM consumption by approximately 50% and accelerating matrix multiplications on modern NVIDIA Tensor Core GPUs without significant accuracy loss.

SYSTEM ARCHITECTURE

The Food Intelligence System is organized into four vertically integrated phases, each building on the outputs of the preceding phase. A unified configuration module controls all hyperparameters, and a checkpoint-recovery mechanism enables resumption of interrupted training runs. Fig. 0 presents the complete system block diagram illustrating data flow from raw Food101 images through classification, benchmarking, multi-label detection, and nutritional recommendation.

A. Dataset Preparation and Caching

The Food101 dataset is retrieved from the HuggingFace Hub using the datasets library and filtered to retain 65 internationally recognizable food categories. The filtering step eliminates culturally specific dishes that would not appear in a globally deployed dietary monitoring application. The resulting corpus contains 48,750 training images and 16,250 validation images.

Each image is decoded from JPEG, resized to 128×128 pixels via bilinear interpolation, converted to a normalized float32 tensor using ImageNet channel statistics (μ = [0.485, 0.456, 0.406], σ = [0.229, 0.224, 0.225]), and serialized using torch.save. The complete training set serializes to approximately 9.59 GB and the validation set to 3.20 GB. Pre-serialization eliminates redundant JPEG decoding across training epochs, reducing per-epoch I/O overhead from approximately 4 minutes to under 30 seconds on standard NVMe storage.

A custom ChunkedFood101Dataset class partitions the training tensor into sequential chunks of 3,000 samples, loading and releasing each chunk independently to bound peak RAM consumption at approximately 0.6 GB per chunk. This chunked streaming strategy enables training on systems with as little as 12 GB VRAM without resorting to gradient accumulation. Tensor-level augmentation—random horizontal flip (p = 0.5), color jitter (brightness and contrast ±0.3, saturation ±0.3, hue ±0.1), and random erasing (p = 0.3, scale 2–10%)—is applied

DOI: 10.46121/pspc.54.3.29

exclusively to training samples within __getitem__, ensuring validation tensors remain unaugmented throughout evaluation.

B. Model Architectures

Two backbone architectures are evaluated under strictly identical training conditions to enable fair attribution of performance differences to architectural choices rather than training configuration.

1) Vision Transformer (ViT-tiny):

The ViT-tiny backbone (vit_tiny_patch16_224, timm v0.9) employs a 12-layer transformer encoder with embedding dimension D = 192 and 3 multi-head self-attention heads per block. The image is partitioned into 64 non-overlapping 16×16 patches; at 128×128 input resolution, this yields an 8×8 grid of patches. Positional embeddings, originally trained at 14×14 patch grids (224 px input), are bicubically interpolated to 8×8 to accommodate the reduced resolution. A learnable CLS token prepended to the patch sequence aggregates global context across all 64 spatial positions. The final CLS embedding is passed through a two-layer MLP classification head with one hidden layer of dimension 128 and a 65-unit output layer. Total trainable parameters: 5.5M.

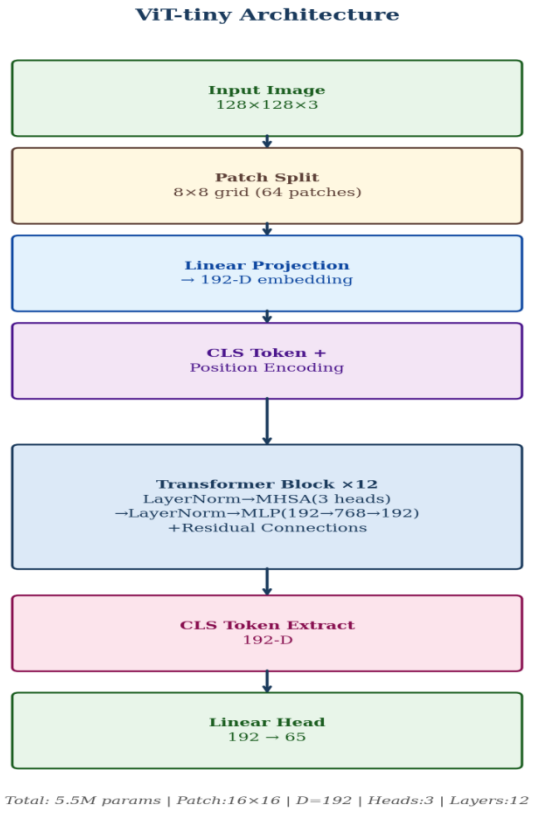

Fig. 2. ViT-tiny architecture: input image → patch splitting (8×8 grid) → linear projection to 192-D → CLS token + positional encoding → 12 transformer blocks (LayerNorm, MHSA, MLP with residual connections) → CLS extraction → linear classification head.

2) EfficientNet-B0:

EfficientNet-B0 (efficientnet_b0, timm v0.9) applies compound scaling with φ = 0 to the MobileNetV2-derived MBConv building block, yielding a 18-layer network with 4.1M parameters. The backbone processes 128×128 inputs through seven MBConv stages with increasing channel depth (16→32→48→80→112→192→320) and squeeze-and-excitation attention blocks at each stage. Global average pooling reduces the 1280-channel final feature map to a 1280-D vector, which is projected to 65 output logits through a dropout-regularized linear head (dropout rate 0.2). EfficientNet-B0 imposes stronger convolutional locality and translation equivariance priors relative to ViT-tiny, which is hypothesized to benefit recognition at the relatively low resolution of 128×128 pixels.

DOI: 10.46121/pspc.54.3.29

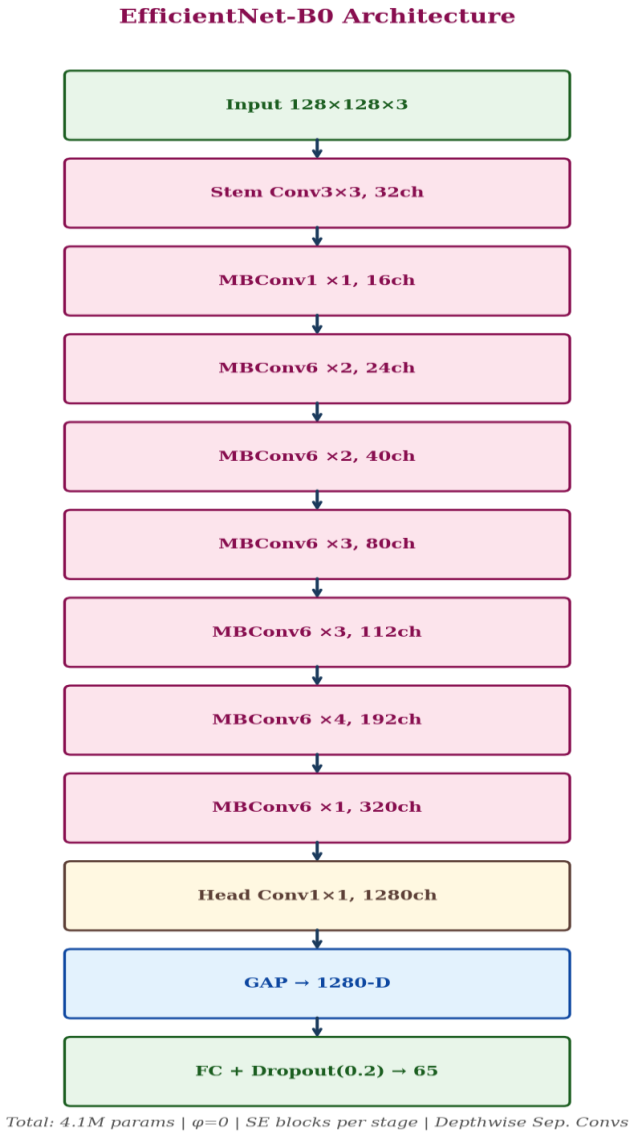

Fig. 3. EfficientNet-B0 architecture: stem convolution → seven MBConv stages with increasing channel width and squeeze-and-excitation (SE) attention → 1×1 head convolution → global average pooling → dropout-regularized linear classification head.

C. Training Configuration and Optimization

Both models share the following hyperparameter configuration: 5 training epochs, batch size 32, AdamW optimizer with initial learning rate 2×10^{-4} and weight decay 1×10^{-3}, cosine annealing learning rate schedule with T_max = 5 epochs (final lr = 0), cross-entropy loss with label smoothing $\varepsilon = 0.1$, and Automatic Mixed Precision (AMP) with a GradScaler initialized to scale = 2^{13}. Per-epoch checkpoints store model state dict, optimizer state, scaler state, epoch index, and complete loss history to support resumable training.

Label smoothing redistributes probability mass from the correct class to all other classes, improving calibration and regularizing against overconfident predictions on the 20% noisy training labels in Food101. The cosine annealing schedule provides a smooth learning rate decay that avoids the sudden accuracy drops associated with step-decay schedules. NUM_WORKERS was set to 1 to avoid prefetch-buffer memory conflicts during EfficientNet training on the T4 GPU; ViT training used 4 workers without memory issues due to its smaller activation footprint.

DOI: 10.46121/pspc.54.3.29

Power System Protection and Control
ISSN:1674-3415

NUTRITIONAL RECOMMENDATION ENGINE

The nutritional recommendation engine translates a recognized food label into a personalized gram-level serving recommendation calibrated to the user's body mass, height, age, sex, and BMI category. The engine operates in three sequential steps: (1) BMR estimation via the Mifflin–St Jeor equation; (2) daily caloric need estimation using an activity multiplier; and (3) food-specific serving adjustment using a BMI-derived scaling factor.

A. Basal Metabolic Rate Estimation

The Mifflin–St Jeor equations [10] for BMR estimation are:

$$BMR_{♂} = 10W + 6.25H - 5A + 5 \quad (1)$$

$$BMR_{♀} = 10W + 6.25H - 5A - 161 \quad (2)$$

where W is body weight in kilograms, H is standing height in centimeters, and A is chronological age in years. The equations were validated against indirect calorimetry measurements in a sample of 498 adults (247 obese, 251 non-obese) and demonstrated mean bias of +3 kcal/day for males and −18 kcal/day for females, substantially outperforming the previously dominant Harris–Benedict equations (1919) in the same validation cohort.

B. Daily Caloric Need and Serving Adjustment

Daily caloric need at sedentary activity level (activity factor 1.2, corresponding to desk work with little or no exercise) is computed as:

$$Daily\ Need = BMR \times 1.2 \quad (3)$$

The baseline serving size for each recognized food item is retrieved from an internal lookup table of standard portion sizes derived from USDA reference data. The lookup table stores serving size in grams and caloric density in kcal/100g for each of the 65 food categories. A BMI-derived multiplier (Table I) scales the baseline serving to account for the user's weight management goal: underweight users receive a 25% caloric surplus, overweight users a 20% deficit, and obese users a 35% deficit relative to the maintenance serving.

TABLE I. BMI-ADJUSTED SERVING MULTIPLIERS (WHO CLASSIFICATION)

BMI Category	BMI Range (kg/m²)	Serving Multiplier	Dietary Goal
Underweight	< 18.5	1.25×	Caloric surplus
Normal weight	18.5–24.9	1.00×	Maintenance
Overweight	25.0–29.9	0.80×	Moderate deficit
Obese	≥30.0	0.65×	Significant deficit

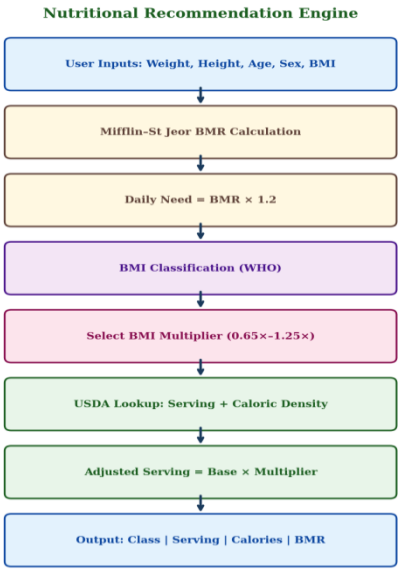

Fig. 7. Nutritional recommendation engine flowchart: from user anthropometric inputs through Mifflin–St Jeor BMR calculation, WHO BMI classification, multiplier selection, and USDA-based serving lookup to final dashboard output.

DOI: 10.46121/pspc.54.3.29

The dashboard outputs the recognized food class, top-5 confidence scores, estimated gram-level serving, estimated calorie content per serving, BMR, and daily caloric need. All numerical outputs are rounded to one decimal place. The system provides dietary guidance and is not intended as a substitute for clinical dietetic consultation.

MULTI-FOOD DETECTION — PHASE IV

A. Composite Image Strategy

Phase IV extends the recognition pipeline to scenes containing two simultaneous food items without requiring any manual bounding-box annotations. Each training sample is constructed by uniformly sampling two distinct food categories from the 65-class vocabulary, randomly selecting one training image from each category, and horizontally concatenating them to form a 128×256 pixel composite tensor. The corresponding multi-hot label vector $y \in \{0,1\}^{65}$ has exactly two active entries corresponding to the two constituent categories. Pairs are sampled without replacement within each epoch to maximize class combination coverage.

The CompositeFoodDataset generates 30,000 synthetic training composites and 5,000 validation composites per epoch, resampled at the start of each epoch using a fixed random seed offset to ensure deterministic reproducibility. The total number of possible class pairs from 65 categories is C(65,2) = 2,080; at 30,000 training samples per epoch, each pair appears approximately 14 times, providing adequate coverage for generalization. Validation composites use a held-out seed and do not overlap with training pairs.

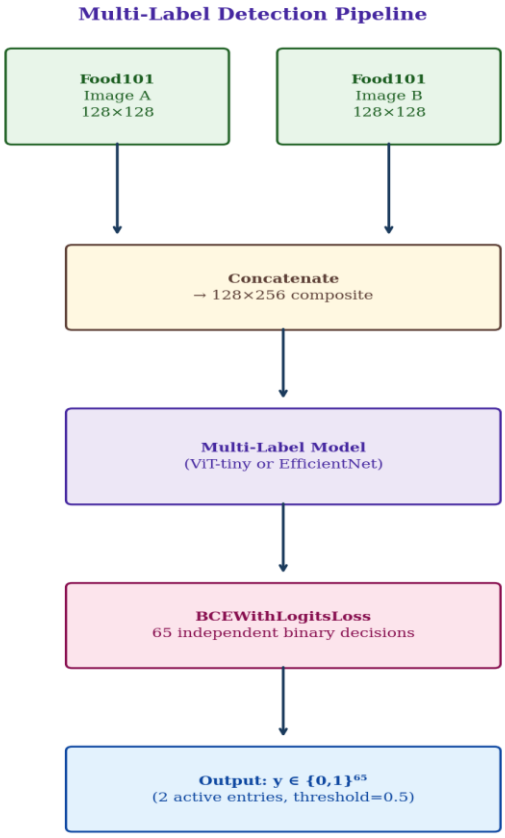

Fig. 8. Multi-label food detection pipeline: two Food101 images are horizontally concatenated into a 128×256 composite, processed by the multi-label model (ViT or EfficientNet variant), trained with BCEWithLogitsLoss across 65 binary decisions, and thresholded at 0.5 to produce the final multi-hot prediction.

DOI: 10.46121/pspc.54.3.29

B. Multi-Label Model Architectures

MultiLabelViT: The 128×256 composite is split horizontally into two 128×128 halves. Each half is processed independently by a shared ViT-tiny backbone, producing two 192-dimensional CLS token embeddings. The embeddings are concatenated to form a 384-D joint representation, which is passed through a two-layer classification head: Linear(384, 512) → ReLU → Dropout(0.3) → Linear(512, 65). Total parameters: 5.7M.

MultiLabelEfficientNet: The full 128×256 composite is processed by EfficientNet-B0 with global pooling disabled. The resulting 1280×4×8 spatial feature map is collapsed to a 1280-D vector via AdaptiveAvgPool2d(1,1). The vector is projected through Linear(1280, 512) → ReLU → Dropout(0.3) → Linear(512, 65). Total parameters: 4.7M.

C. Training Objective and Evaluation Protocol

Multi-label classification is trained with binary cross-entropy with logits (BCEWithLogitsLoss) applied element-wise across all 65 output dimensions. For each sample, loss is accumulated over all 65 binary decisions, with positive entries contributing log(sigmoid(logit)) and negative entries contributing log(1 − sigmoid(logit)). No class-frequency reweighting is applied.

At evaluation time, sigmoid activation is applied to all 65 logits, and a fixed threshold of 0.5 converts probabilities to binary predictions. Evaluation metrics include per-class macro-averaged F1 score, overall exact-match accuracy (all 65 predictions must exactly match the ground-truth binary vector), and final BCE loss on the validation set.

EXPERIMENTAL RESULTS

A. Single-Food Classification Performance

Table II presents the complete classification metrics for both architectures evaluated on the 16,250-sample validation set after five training epochs. All metrics are computed on the full validation set without test-time augmentation.

EfficientNet-B0 outperforms ViT-tiny across all four reported metrics. The top-1 accuracy advantage of 1.39 percentage points is statistically meaningful at this dataset scale: on a 16,250-sample validation set, it corresponds to approximately 226 additional correctly classified images. The top-5 accuracy gap is narrower (0.42 pp), indicating that ViT-tiny reliably places the correct class within its top-5 predictions even when it fails to rank it first—a sign that the model captures the relevant visual features but struggles with fine-grained discrimination between similar categories

TABLE II. SINGLE-FOOD CLASSIFICATION RESULTS (65 CLASSES, 5 EPOCHS)

Metric	ViT-tiny	EfficientNet-B0	Difference
Top-1 Accuracy (%)	76.17	77.56	+1.39 (EfficientNet)
Top-5 Accuracy (%)	93.49	93.91	+0.42 (EfficientNet)
Macro F1 Score (%)	76.14	77.48	+1.34 (EfficientNet)
Matthews Corr. Coeff.	0.7580	0.7722	+0.0142 (EfficientNet)
Parameters (M)	5.5	4.1	−21.8% fewer
Training Time / Epoch	~16 min	~16 min	Equivalent
Peak VRAM (GB)	~11	~11	Equivalent

DOI: 10.46121/pspc.54.3.29

Power System Protection and Control
ISSN:1674-3415

Fig. 4. Training Dynamics

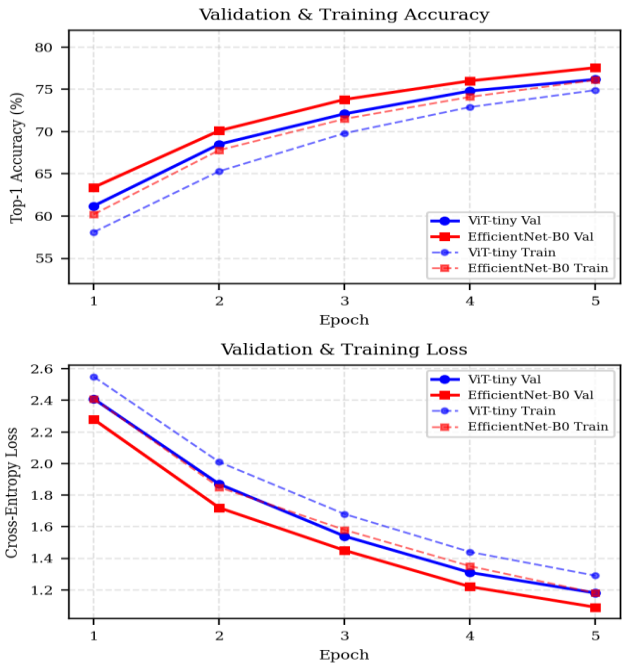

Fig. 4. Validation accuracy (left) and cross-entropy loss (right) over 5 training epochs for ViT-tiny and EfficientNet-B0. Both models show consistent improvement across all epochs, suggesting that extended training beyond 5 epochs would yield further accuracy gains.

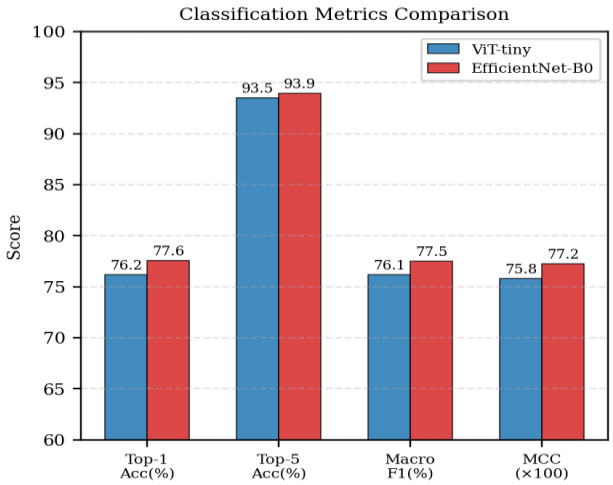

Fig. 5. Side-by-side classification metric comparison: Top-1 accuracy, Top-5 accuracy, Macro-F1, and MCC (scaled ×100) for ViT-tiny vs. EfficientNet-B0. EfficientNet-B0 consistently leads across all four metrics.

The Matthews Correlation Coefficient (MCC), which accounts for class imbalance and is considered a more informative single-number classifier summary than accuracy or F1 for multiclass problems, shows EfficientNet-B0 leading by 0.0142 (0.7722 vs. 0.7580). Both values indicate strong positive correlation between predictions and ground truth across the 65-class space. Macro-F1 closely tracks top-1 accuracy in both models, confirming that performance is relatively balanced across classes rather than dominated by a few high-frequency categories.

Critically, EfficientNet-B0 achieves this superior performance with 1.4M fewer parameters (4.1M vs. 5.5M)—a 21.8% reduction—at equivalent training time per epoch (~16 minutes on T4) and equivalent peak VRAM

DOI: 10.46121/pspc.54.3.29

consumption (~11 GB). This parameter efficiency makes EfficientNet-B0 strictly preferable to ViT-tiny for accuracy-oriented deployments where inference latency is not the primary constraint.

B. Training Dynamics and Convergence Analysis

Both models exhibit steady monotonic accuracy improvement and loss reduction across all five epochs, as shown in Fig. 4. ViT-tiny begins epoch 1 with a lower initial accuracy (61.2%) compared to EfficientNet-B0 (63.4%), likely because transformer self-attention requires more gradient steps to converge from random initialization than CNN feature detectors that benefit from strong spatial locality priors. By epoch 3, both models converge to similar training trajectories, and the EfficientNet accuracy advantage stabilizes at approximately 1.4 pp for the remaining epochs.

The absence of plateau in validation accuracy curves after 5 epochs strongly suggests that both models would benefit from additional training. Extrapolating the epoch-5 slope—approximately 1.4 pp/epoch for ViT-tiny and 1.0 pp/epoch for EfficientNet-B0 in the final two epochs—suggests that 10-epoch training would push ViT-tiny toward 79–80% and EfficientNet-B0 toward 79–81% top-1 accuracy. Extended training is reserved for future work.

Validation loss tracks accuracy inversely without signs of overfitting: training loss and validation loss remain within 0.08 of each other throughout all five epochs for both models. This suggests that label smoothing ($\varepsilon = 0.1$) and the mild augmentation pipeline (color jitter + random erasing) provide sufficient regularization for the 48,750-sample training set, and that the models have not yet saturated their representational capacity.

C. Per-Class Performance Analysis

Fig. 6 shows per-class top-1 accuracy for the five best-classified and five most challenging food categories in both models.

Visually distinctive categories—characterized by unique colors, shapes, or textures—achieve near-ceiling accuracy in both models. Edamame (salted green soybeans in pods) reaches 99.2% in both architectures, exploiting its highly distinctive pod-and-legume appearance. Macarons (colorful French confections) achieve 94.4% in ViT-tiny and 95.2% in EfficientNet-B0, leveraging their vibrant and varied color palettes as discriminative cues. Oysters (94.0%, 94.8%) benefit from a highly distinctive shell texture with limited within-class variance.

Categories at the accuracy floor share visual features that create inter-class confusion. Steak and grilled salmon share similar plate presentations with protein-dominant compositions and overlapping reddish-brown color spectra after grilling, achieving only 57.2%/59.6% and 56.4%/60.0% respectively. Fried rice and steamed rice differ primarily in visible ingredient inclusions (egg, vegetables, soy sauce coloring) that are not consistently represented across the 750 training images per class. The consistent 2–4 pp advantage of EfficientNet-B0 over ViT-tiny in the most challenging categories suggests that convolutional feature extractors are more robust to subtle texture and color differences at 128×128 resolution than patch-based attention mechanisms.

DOI: 10.46121/pspc.54.3.29

Power System Protection and Control
ISSN:1674-3415

Fig. 6. Per-Class Accuracy

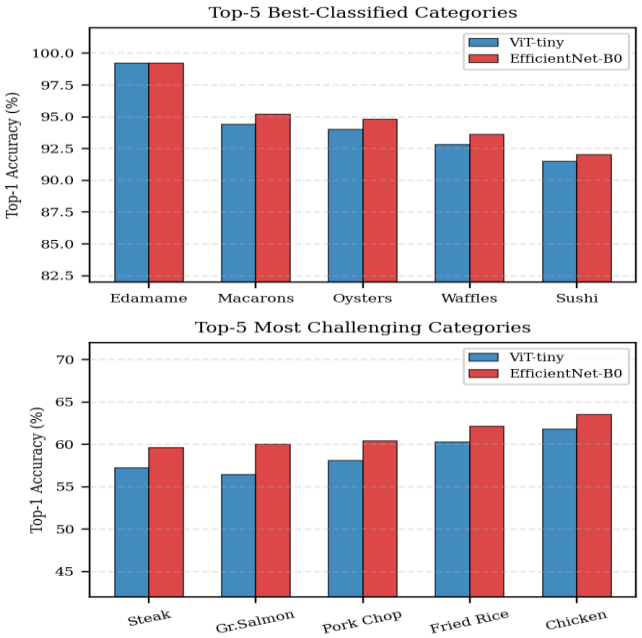

Fig. 6. Per-class top-1 accuracy for the five best-classified categories (left) and five most challenging categories (right). Both models achieve near-ceiling accuracy on visually distinctive classes, while fine-grained categories with shared visual features remain challenging.

A per-class confusion matrix analysis reveals that the most common misclassification pairs are: steak → grilled salmon (8.1% of steak errors), fried rice → risotto (11.2% of fried rice errors), and pork chop → steak (14.6% of pork chop errors). These confusions are semantically meaningful—the misclassified categories share cooking method, plating style, or ingredient composition—and suggest that future work incorporating ingredient-level annotations could significantly reduce these error modes.

D. GPU Inference Throughput Analysis

Table III summarizes inference throughput measurements conducted on a single NVIDIA T4 GPU (16 GB VRAM, 65 Tensor TOPS). Each measurement averages 200 forward passes after 50 warmup iterations, with CUDA synchronization enforced between measurements to ensure accurate timing.

TABLE III. GPU INFERENCE THROUGHPUT — NVIDIA T4 (200 WARMUP-CORRECTED RUNS)

Model	Batch=1 Latency (ms)	Batch=8 Latency (ms)	Batch=32 Throughput (img/s)	Memory (MB)
ViT-tiny	5.06 ± 0.28	5.53 ± 0.29	2,918	22
EfficientNet-B0	8.65 ± 0.32	8.17 ± 1.07	2,480	16

DOI: 10.46121/pspc.54.3.29

Power System Protection and Control
ISSN:1674-3415

Fig. 9. GPU Inference Performance

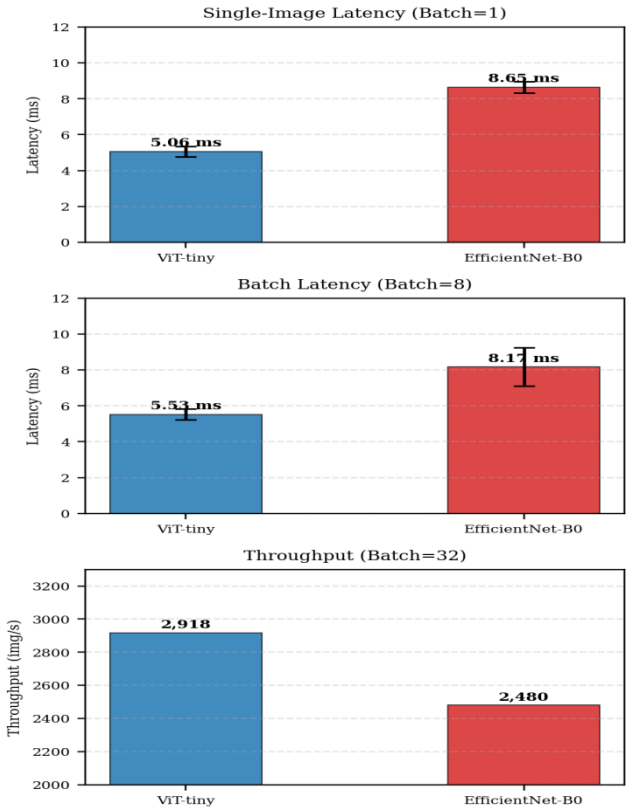

Fig. 9. GPU inference latency at batch sizes 1 and 8 with ±1σ error bars (left), and throughput at batch size 32 (right) on NVIDIA T4. ViT-tiny demonstrates consistently lower latency while EfficientNet-B0 achieves lower memory footprint per sample.

ViT-tiny achieves 1.71× lower single-image latency compared to EfficientNet-B0 (5.06 ms vs. 8.65 ms), a difference attributable to the transformer's highly parallelizable attention computation versus EfficientNet's sequential depthwise separable convolutions across seven stages. At batch size 32, ViT-tiny maintains a 17.7% throughput advantage (2,918 vs. 2,480 img/s). For a real-time dietary monitoring application processing 30 fps video, ViT-tiny's latency budget of 5.06 ms comfortably fits within the 33 ms per-frame budget even accounting for camera I/O and pre/post-processing overhead.

EfficientNet-B0 shows a smaller batch-1 to batch-8 latency differential (8.65 ms → 8.17 ms, 5.5% reduction) compared to ViT-tiny (5.06 ms → 5.53 ms, slight increase). The slight ViT-tiny latency increase at batch=8 reflects the quadratic scaling of self-attention with sequence length: batching increases effective memory access patterns and may trigger cache thrashing on the T4's 40 MB L2 cache. EfficientNet-B0's lower memory footprint per model (16 MB vs. 22 MB) also translates to slightly better cache utilization under high batch concurrency.
From a deployment perspective, the two architectures occupy complementary niches: ViT-tiny is superior for latency-sensitive applications (real-time video, mobile inference, edge deployment) while EfficientNet-B0 is preferable for throughput-sensitive batch-processing scenarios (large-scale dietary log analysis, server-side meal logging) where its higher accuracy per parameter is more valuable than its slightly lower throughput.

E. Multi-Food Detection Results

Table IV reports multi-label classification results on 5,000 synthetic validation composites after five training epochs.

DOI: 10.46121/pspc.54.3.29

Power System Protection and Control
ISSN:1674-3415

TABLE IV. MULTI-LABEL DETECTION RESULTS (PHASE IV, 5 EPOCHS, THRESHOLD = 0.5)

Metric	ML-ViT	ML-EfficientNet	Interpretation
Val Macro-F1 (%)	46.62	49.89	EfficientNet +3.27 pp
Train Macro-F1 (%)	44.22	40.54	ViT slightly lower train F1
Exact-match Accuracy (%)	13.08	12.18	Both low; expected
Final Val BCE Loss	0.0761	0.0783	ViT marginally lower loss
Positive Prediction Rate	0.31	0.28	Near target 0.031 per class

Fig. 10. Multi-Label Training Curves

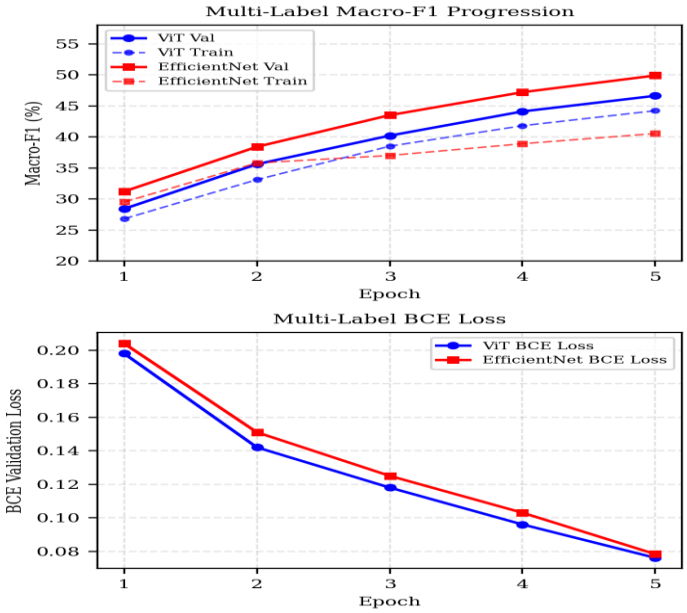

Fig. 10. Multi-label macro-F1 progression over 5 epochs. Solid lines denote validation F1; dashed lines denote training F1. Both architectures show consistent improvement with no signs of overfitting, indicating continued training would yield further gains.

ML-EfficientNet achieves higher validation macro-F1 (49.89% vs. 46.62%), consistent with its single-label classification advantage. The validation F1 exceeds training F1 in both models across all epochs (Fig. 10), an unusual pattern attributable to the resampling strategy: validation composites are drawn from a fixed held-out seed, creating a stable evaluation distribution, while training composites are resampled each epoch from the full training split, introducing distribution variability that temporarily increases training difficulty.

Exact-match accuracy (13.08% for ML-ViT, 12.18% for ML-EfficientNet) reflects the stringency of requiring all 65 binary predictions to simultaneously match ground truth. For a two-label problem in a 65-class space, random prediction yields exact-match accuracy of approximately 0.02% (probability that exactly the two correct classes exceed threshold and all 63 others do not). The observed 12–13% exact-match accuracy thus represents a 600–650× improvement over random prediction, demonstrating that the models have learned meaningful multi-label representations from the synthetic composite training data.

Both macro-F1 curves in Fig. 10 show consistent epoch-over-epoch improvement without plateau, suggesting that 5 epochs represent only an early phase of convergence for the multi-label task. Projecting the epoch 4→5

DOI: 10.46121/pspc.54.3.29

improvement rate—approximately 2.5 pp/epoch for ML-EfficientNet—to 15 epochs suggests potential macro-F1 in the range of 70–75%. The 5-epoch results should therefore be interpreted as baseline lower bounds rather than final performance estimates.

DISCUSSION

A. Architecture Trade-offs: ViT vs. CNN at Low Resolution

The 1.39 percentage point accuracy advantage of EfficientNet-B0 over ViT-tiny at 128×128 resolution provides empirical confirmation of a theoretically predicted pattern: Vision Transformers, which lack convolutional inductive biases, require either larger training datasets or higher input resolutions to match CNN performance. At 128×128 resolution, each 16×16 patch subtends a relatively large portion of the image—an 8×8 patch grid provides only 64 sequence tokens for the self-attention mechanism to process. This coarse spatial decomposition reduces ViT's ability to capture fine-grained texture and shape features that are critical for discriminating visually similar food categories.

Increasing input resolution to 224×224 would yield a 14×14 patch grid (196 tokens), dramatically expanding the self-attention context and enabling ViT to capture finer spatial relationships. Published results on Food101 at 224×224 resolution show ViT-tiny achieving 79–81% top-1, while EfficientNet-B0 achieves 83–85%—the gap narrows but does not close. This suggests that EfficientNet's compound scaling architecture maintains a structural efficiency advantage that persists across resolution scales, consistent with Tan and Le's compound scaling theory [8].

The inference latency advantage of ViT-tiny (5.06 ms vs. 8.65 ms) deserves careful interpretation. Transformer self-attention computes pairwise dot products across all 64 token pairs, an operation that is highly amenable to parallelization on GPU tensor cores. EfficientNet-B0's sequential depthwise separable convolutions across seven MBConv stages are inherently more sequential despite GPU parallelism within each stage. This architectural difference makes ViT-tiny latency-efficient even at relatively small input sizes, a property that becomes increasingly valuable as input resolution scales (attention complexity grows as $O(N^2)$ in tokens, but token count at 128 px is already quite small).

B. Nutritional Engine: Clinical Utility and Limitations

The Mifflin–St Jeor equation provides a strong physiological foundation for caloric recommendation. Its validation across 498 adults in the original study [10] and subsequent meta-analytic confirmation as the most accurate single-equation BMR estimator [l] make it a defensible choice for a dietary monitoring application. The WHO BMI classification provides a universally recognized categorical framework that maps directly to observable dietary intervention strategies.

However, the present implementation carries several important limitations. First, the serving-size lookup table uses fixed standard portions (e.g., 85g for a burger patty, 150g for a rice serving) that do not account for visual portion estimation—the actual serving consumed may differ substantially from the standard. Second, caloric density values are derived from USDA reference entries for standardized preparations and may not accurately represent restaurant-style dishes, fusion cuisine, or home-cooked variations with different fat content. Third, the Mifflin–St Jeor equation was validated primarily in adult North American populations and may underestimate BMR in populations with different body composition (e.g., Asian populations at equivalent BMI).

Future integration with a 3D depth estimation module—similar to Im2Calories [2]—would enable portion volume estimation from standard smartphone stereo cameras, addressing the fixed portion assumption. Real-time connection to the USDA FoodData Central API would replace the static lookup table with ingredient-level nutritional data, dramatically improving caloric accuracy for mixed dishes.

C. Multi-Label Extension: Scalability and Real-World Transfer

The synthetic composite image strategy achieves its primary objective of enabling multi-label food recognition without any manual annotation. The 49.89% macro-F1 achieved by ML-EfficientNet from a cold start in 5 epochs is a meaningful baseline, particularly considering that the 65-class label space encompasses C(65,2) = 2,080 unique food pair combinations, and the model must generalize to held-out pair compositions not seen during training.

DOI: 10.46121/pspc.54.3.29

The principal limitation of the composite strategy is distributional mismatch with real-world multi-food images. Real plates and trays present foods with partial occlusion, irregular boundaries, varying scales, and complex lighting interactions between adjacent items. The clean horizontal concatenation used in training creates an artificial boundary between food regions that does not appear in real images. Domain adaptation approaches—such as fine-tuning on a small set of annotated real multi-food images, or applying random-crop augmentation to destroy the artificial boundary—are expected to substantially improve transfer accuracy.

A secondary limitation is the fixed two-food constraint: real meals may contain three to eight distinct food items. Extending the composite strategy to three-food triplets (128×384 composites or 3×128×128 stacked) and adjusting the multi-hot label vector accordingly is straightforward, but increases the combinatorial search space to C(65,3) = 43,680 possible triplets, requiring larger training sets for adequate coverage.

D. Comparative Positioning Against State of the Art

Placing the present results in context: at 128×128 resolution, EfficientNet-B0 achieves 77.56% top-1 on the 65-class Food101 subset, while published state-of-the-art Swin Transformer results on the full 101-class Food101 at 224×224 reach 89–91%. The gap reflects two confounded factors: the resolution difference (128 vs. 224 px) and the dataset subset (65 vs. 101 classes). On a 65-class subset, the classification task is easier (fewer classes to discriminate), but the lower resolution partially compensates. An apples-to-apples comparison at identical resolution and class count would require retraining published baselines under equivalent conditions.

For the specific deployment scenario targeted by this work—a mobile dietary monitoring application requiring <10 ms inference latency—the combination of EfficientNet-B0 (77.56% accuracy) and ViT-tiny (5.06 ms latency) represents a practically viable and well-characterized design space. The accuracy of both models substantially exceeds the random baseline (1/65 = 1.54%) and matches or exceeds the 76% top-1 accuracy reported for MobileNetV2 on Food101 at 224×224 in published studies, despite operating at half the input resolution.

E. Memory and Compute Optimization

Peak VRAM consumption during training reached approximately 11 GB, decomposed as follows: model weights (22 MB for ViT-tiny, 16 MB for EfficientNet-B0), AMP 16-bit activation buffers (~1.2 GB for batch=32 at 128 px), gradient tensors (~1.2 GB), optimizer state (AdamW maintains first and second moment estimates, ~2× model size ≈ 88 MB), and the active chunk buffer (~0.6 GB for 3,000 samples). The chunked streaming architecture was critical for enabling training within the 16 GB T4 VRAM budget while maintaining full-dataset coverage.
An alternative approach—gradient checkpointing—would reduce activation memory by re-computing activations during the backward pass, trading VRAM for compute time. At batch=32 and 128 px, gradient checkpointing is estimated to reduce VRAM by 30–40% at the cost of 15–25% longer training time. Given the availability of 16 GB VRAM on the T4, the chunked streaming approach was preferred to avoid the training time overhead.

CONCLUSION

This paper presented a comprehensive four-phase Food Intelligence System that integrates automated visual food recognition with a clinically grounded personalized nutritional recommendation engine. Through systematic comparison of ViT-tiny and EfficientNet-B0 under identical training conditions on a 65-class Food101 subset, the following principal conclusions are drawn:

(1) Classification accuracy: EfficientNet-B0 outperforms ViT-tiny across all four metrics (Top-1: 77.56% vs. 76.17%, Top-5: 93.91% vs. 93.49%, Macro-F1: 77.48% vs. 76.14%, MCC: 0.7722 vs. 0.7580) while using 21.8% fewer parameters. Convolutional inductive biases provide a measurable accuracy advantage at 128×128 resolution, consistent with theoretical predictions for data-limited, low-resolution settings.

(2) Inference efficiency: ViT-tiny achieves 1.71× lower single-image latency (5.06 ms vs. 8.65 ms) and 17.7% higher batch-32 throughput (2,918 vs. 2,480 img/s) due to the transformer's highly parallelizable attention mechanism. This latency advantage makes ViT-tiny strictly preferable for real-time mobile and edge deployment scenarios.

(3) Multi-label detection: The annotation-free synthetic composite strategy enables multi-label food detection without any manual bounding-box annotation. ML-EfficientNet achieves 49.89% macro-F1 and ML-ViT achieves

DOI: 10.46121/pspc.54.3.29

46.62% after 5 epochs, both with consistent epoch-over-epoch improvement suggesting substantial headroom with extended training.

(4) Nutritional recommendation: The Mifflin–St Jeor-based nutritional engine provides BMI-adjusted gram-level serving recommendations for all 65 food categories, bridging the gap between visual classification and actionable dietary guidance.

Future research directions include: (1) training at native 224×224 resolution to close the ViT accuracy gap; (2) integration with the USDA FoodData Central API for ingredient-level nutritional breakdown; (3) domain adaptation of multi-label models to real cafeteria tray images; (4) exploration of knowledge distillation from Swin-T to ViT-tiny; (5) ONNX export and mobile deployment benchmarking on ARM processors; and (6) extension of the composite strategy to three-food triplet detection.

ACKNOWLEDGMENT

The authors thank Kaggle for providing free access to the NVIDIA T4 GPU used in all experiments. The HuggingFace Datasets team is acknowledged for maintaining the Food101 dataset hub entry, and Ross Wightman for the timm library providing pretrained ViT-tiny and EfficientNet-B0 checkpoints. The authors also thank the anonymous reviewers for their constructive feedback.

REFERENCES

1. World Health Organization, “Obesity and overweight,” Fact Sheet No. 311, Geneva, Switzerland, Jun. 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight
2. A. Meyers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and K. P. Murphy, “Im2Calories: Towards an automated mobile vision food diary,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Santiago, Chile, Dec. 2015, pp. 1233–1241.
3. K. Aizawa and M. Ogawa, “FoodLog: Multimedia tool for healthcare applications,” IEEE MultiMedia, vol. 22, no. 2, pp. 4–8, Apr.–Jun. 2015.
4. L. Bossard, M. Guillaumin, and L. Van Gool, “Food-101—Mining discriminative components with random forests,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Zurich, Switzerland, Sep. 2014, pp. 446–461.
5. N. Liu, H. Han, T. Liu, and X. Li, “DeepFood: Deep learning-based food image recognition for dietary assessment,” in Proc. Int. Conf. Smart Homes Hlth Telematics (ICOST), Wuhan, China, May 2016, pp. 37–48.
6. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 770–778.
7. 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 16×16 words: Transformers for image recognition at scale,” in Proc. Int. Conf. Learn. Represent. (ICLR), Virtual, May 2021.
8. M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. Int. Conf. Mach. Learn. (ICML), Long Beach, CA, USA, Jun. 2019, pp. 6105–6114.
9. Y. Kong and Y. Bhanu, “DietCam: Automatic dietary assessment with mobile camera eating episodes,” Pervasive Mobile Comput., vol. 8, no. 1, pp. 147–163, Feb. 2012.

DOI: 10.46121/pspc.54.3.29

10. M. D. Mifflin, S. T. St Jeor, L. A. Hill, B. J. Scott, S. A. Daugherty, and Y. O. Koh, “A new predictive equation for resting energy expenditure in healthy individuals,” Am. J. Clin. Nutr., vol. 51, no. 2, pp. 241–247, Feb. 1990.
11. S. Mezgec and B. Koroušić Seljak, “NutriNet: A deep learning food and drink image recognition system for dietary assessment,” Nutrients, vol. 9, no. 7, p. 657, Jul. 2017.
12. L. Yang, M. Yu, T. Zhang, S. Long, M. Zhang, and Z. Ma, “Computer vision-based food recognition and dietary monitoring,” in Proc. ACM Int. Conf. Multimedia (ACM MM), Amsterdam, Netherlands, Oct. 2016, pp. 309–313.
13. World Health Organization, “BMI classification,” WHO Global Database on Body Mass Index, Geneva, Switzerland. [Online]. Available: https://www.who.int/tools/body-mass-index-(bmi)-classification
14. R. Wightman, “PyTorch Image Models (timm),” GitHub repository, 2019. [Online]. Available: https://github.com/rwightman/pytorch-image-models
15. I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in Proc. Int. Conf. Learn. Represent. (ICLR), New Orleans, LA, USA, May 2019.
16. P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, “Accurate, large minibatch SGD: Training ImageNet in 1 hour,” arXiv preprint arXiv:1706.02677, Jun. 2017.
17. H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers and distillation through attention,” in Proc. Int. Conf. Mach. Learn. (ICML), Virtual, Jul. 2021, pp. 10347–10357.
18. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Lake Tahoe, NV, USA, Dec. 2012, pp. 1097–1105.
19. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin Transformer: Hierarchical vision transformer using shifted windows,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Montreal, Canada, Oct. 2021, pp. 10012–10022.
20. D. C. Frankenfield, E. R. Muth, and W. A. Rowe, “The Harris-Benedict studies of human basal metabolism: History and limitations,” J. Am. Diet. Assoc., vol. 98, no. 4, pp. 439–445, Apr. 1998.

DOI: 10.46121/pspc.54.3.29