



## OPEN ACCESS

## EDITED BY

Moulay Akhloufi,  
Université de Moncton, Canada

## REVIEWED BY

Xing Shen,  
McGill University, Canada  
Hidam Singh,  
Gauhati University, India

## \*CORRESPONDENCE

Jia Uddin  
✉ [jia.uddin@wsu.ac.kr](mailto:jia.uddin@wsu.ac.kr)  
Hezerul Abdul Karim  
✉ [hezerul@mmu.edu.my](mailto:hezerul@mmu.edu.my)

RECEIVED 12 October 2025

REVISED 25 November 2025

ACCEPTED 27 November 2025

PUBLISHED 19 December 2025

## CITATION

Rahman S, Farid FA, Zabin M, Uddin J and  
Abdul Karim H (2025) Ultra-lightweight  
uncertainty-aware ensemble for large-scale  
multi-class medical MRI diagnosis.  
*Front. Radiol.* 5:1723272.  
doi: 10.3389/fradi.2025.1723272

## COPYRIGHT

© 2025 Rahman, Farid, Zabin, Uddin and  
Abdul Karim. This is an open-access article  
distributed under the terms of the [Creative  
Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The  
use, distribution or reproduction in other  
forums is permitted, provided the original  
author(s) and the copyright owner(s) are  
credited and that the original publication in  
this journal is cited, in accordance with  
accepted academic practice. No use,  
distribution or reproduction is permitted  
which does not comply with these terms.

# Ultra-lightweight uncertainty-aware ensemble for large-scale multi-class medical MRI diagnosis

Sowad Rahman<sup>1</sup>, Fahmid Al Farid<sup>2</sup>, Mahe Zabin<sup>3</sup>, Jia Uddin<sup>4\*</sup> and Hezerul Abdul Karim<sup>2\*</sup>

<sup>1</sup>Department of Computer Science and Engineering, BRAC University, Dhaka, Bangladesh, <sup>2</sup>Centre for Image and Vision Computing (CIVC), Centre of Excellence for Artificial Intelligence, Faculty of Artificial Intelligence and Engineering (FAIE), Multimedia University, Cyberjaya, Selangor, Malaysia, <sup>3</sup>Human and Digital Interface Department, JW Kim College of Future Studies, Woosong University, Daejeon, South Korea, <sup>4</sup>AI and Big Data Department, Woosong University, Daejeon, Republic of Korea

This paper introduces an Ultra-Lightweight Uncertainty-Aware Ensemble (UALE) model for large-scale multi-class medical MRI diagnosis, evaluated on the 2024 Benchmark Diagnostic MRI and Medical Imaging Dataset containing 40 classes and 33,616 images. The model integrates five specialized micro-expert networks, each designed to capture distinct MRI features, and combines them using a confidence-weighted ensemble mechanism enhanced with variance-based uncertainty quantification for robust, reliable predictions. With only 0.05M parameters and 0.18 GFLOPs, UALE achieves high efficiency and competitive performance among ultra-lightweight models with an accuracy of 69.1% and an F1 score of 68.3%. Besides lightweight models, the paper offers an extensive analysis and performance comparison with fifteen state-of-the-art models, discusses various datasets, elaborates on uncertainty estimates pertaining to the clinical trustworthiness of the models and possible clinical deployment, and highlights trade-offs and avenues for future work in economically constrained settings. The extreme compactness and reliability of the UALE affords it unique utility in scalable medical diagnostics suitable for low-resource clinical settings and portable imaging devices, such as rural hospitals.

## KEYWORDS

medical imaging, lightweight deep learning, ensemble, uncertainty quantification, MRI, multi-class classification, benchmark dataset

## 1 Introduction

The integration of artificial intelligence (AI) into medical imaging has revolutionized diagnostic workflows, enabling rapid and accurate detection of diseases, particularly rare conditions such as gliomas and complex syndromes (1, 2). Deep learning models, including convolutional neural networks (CNNs) and transformer-based architectures, have demonstrated exceptional performance in medical image classification, offering potential for scalable and early diagnosis (3, 4). However, these models often exhibit high computational complexity, making them impractical for resource-constrained environments such as rural healthcare facilities, portable MRI scanners, or embedded

systems with limited memory and processing power (5). Such settings require lightweight, efficient, and interpretable models to support clinical decision-making.

Diagnosing rare diseases poses unique challenges due to their low prevalence and diverse visual characteristics in medical imaging. Class imbalance in datasets, where rare diseases are underrepresented, leads to poor model generalization and biased predictions (6). Uncertainty quantification is critical in these scenarios to identify low-confidence predictions, enabling clinicians to prioritize cases for manual review (7). Additionally, ethical considerations, such as mitigating bias in rare disease diagnosis and ensuring equitable access to diagnostic tools, are paramount (8). Existing lightweight models, such as MobileNets, often struggle to balance performance and efficiency in multi-class tasks involving complex datasets, limiting their practical deployment (5).

The Ultra-Lightweight Uncertainty-Aware Ensemble (UALE) is a novel framework tailored for large-scale multi-class MRI diagnosis, evaluated on the 2024 Benchmark Diagnostic MRI and Medical Imaging Dataset (9). This dataset comprises 33,616 images spanning 40 disease classes, including neurological, spinal, and systemic disorders. Its diversity and scale make it an excellent benchmark for testing model generalizability, efficiency, and robustness. UALE is designed to balance performance and efficiency, enabling real-time inference in resource-constrained environments, such as on a Raspberry Pi 4.

UALE utilizes a collection of five ultra-lightweight expert CNNs, collectively referred to as MedicalMicroNet, each focused on extracting certain MRI features, be it texture, shape, intensity, spatial relationships, or patterns at different scales. The diverse visual characteristics of the MRI images, which differ substantially between disease classes and across anatomical areas, are tackled by these specialized networks. Since UALE distributes feature extraction across these networks, it attains a strong performance with very low computational cost. The ensemble combines predictions using a confidence-weighted mechanism and quantifies uncertainty through prediction variance, providing clinicians with reliable confidence estimates (7). Unlike traditional heavyweight models, UALE prioritizes deployability without sacrificing reliability, addressing key challenges in practical medical AI.

Overall, the contributions of this study are as follows:

- Development of UALE, an ultra-lightweight ensemble framework that achieves competitive performance and efficiency for multi-class MRI diagnosis in resource-constrained environments.
- Comprehensive evaluation and comparison with 15 state-of-the-art and lightweight baselines across multiple datasets.
- Exploration of deployment strategies for low-resource settings.

## 2 Related work

The field of medical image analysis, particularly for rare disease detection, has seen significant advancements, driven by deep learning techniques leveraging convolutional neural

networks (CNNs), transformer-based architectures, few-shot learning, and uncertainty quantification. These methods aim to address challenges in accurate diagnosis with limited data, computational efficiency for edge devices, and reliable uncertainty estimation for clinical reliability. However, limitations such as high computational costs, lack of robustness across diverse datasets, and difficulties in handling data scarcity for rare diseases remain prevalent. Our proposed Ultra-Lightweight Adaptive Learning Ensemble (UALE) introduces an ensemble of micro-experts to achieve superior efficiency, robustness, and generalizability, effectively addressing these gaps. Below, we provide an in-depth review of key models and techniques, focusing on their contributions, datasets, techniques, and research gaps.

Lightweight CNNs have become essential in the development of models for resource-limited situations, like point-of-care diagnostics for rare diseases. SqueezeNet (10) utilizes fire modules to construct a compact model, obtaining AlexNet-level accuracy on benchmarks like ImageNet, while performing satisfactorily on CheXpert (11) for radiograph analysis. Still, its shallowness limits performance on more sophisticated tasks, like the Brain Tumor Dataset (12) brain tumor segmentation, because of weak feature extraction. MobileNetV2 (13) uses depth-wise separable convolutions and performs well on ISIC 2019 (14) skin lesion classification, but struggles with high-dimension tasks like MRI analysis (9), because of capacity constraints. EfficientNet-B0 (15) optimizes accuracy and efficiency with compound scaling and performs well on MedMNIST (16), but its high computational requirements make it infeasible for edge devices. GhostNet (17), PP-LCNet (18), and MCUNet (19) emphasize efficiency, performing well on smaller datasets, such as VinDr-Mammo (20) and MedMNIST (16). Still, their strength is restricted when dealing with complex class imbalances, which is another hurdle in rare disease detection. While these models base their work on traditional CNN implementations, there are considerable research gaps which are, at best, Weaknesses in strength and scalability to handle complex medical tasks. The micro-expert ensemble from UALE is able to achieve a level of corroborated accuracy across different datasets, such as the Benchmark Diagnostic MRI (9), and is able to achieve this with a lesser computational burden.

Transformer models provide high accuracy in rare disease segmentation and classification if enabled to acquire long-range dependencies. Swin-UNet (21), which uses hierarchical transformers, achieves high accuracy on CheXpert (11) and ISIC 2019 (14), but its high computational requirements make it ineligible for edge deployment. CoAtNet (22) and TransUNet (23) integrate the best features of CNN and transformers and show competitive results for the Brain Tumor Dataset (12) and ISIC 2019 (14), but high resource consumption curtails operation in resource-constrained environments. Less compute-intensive transformers for MedMNIST (16) and VinDr-Mammo (20) such as MobileViT (24), EfficientFormer (25), and LeViT (26) still grapple with the fundamental problem of rare diseases which is insufficient labeled training data. Owing to the heavy attention mechanisms these models employ, high computational costs and poor performance on small, labeled datasets remain

persistent in rare disease detection. In comparison, UALE’s CNN-based ensemble achieves similar functionality with a far lower resource footprint, making it suitable for edge deployment to detect rare diseases.

Few-shot and self-supervised learning are vital for rare disease detection, where labeled data is scarce. MAML (27) and ProtoNet (28) enable generalization from few examples, performing well on the Brain Tumor Dataset (12) through adaptive representations. However, their meta-learning frameworks are computationally intensive, limiting their use on edge devices. Self-supervised methods like SimCLR (29), DINO (30), MoCo (31), BYOL (32), and SwAV (33) leverage unlabeled data to learn robust features, excelling on large datasets like CheXpert (11) and ISIC 2019 (14). Their pretraining phase, however, is resource-heavy, making them impractical for lightweight applications. These methods employ contrastive or clustering-based techniques, but their research gaps include high computational costs and limited feasibility for edge deployment. UALE uses efficient data augmentation and oversampling to address class imbalances, achieving robust performance on datasets like MedMNIST (16) without the overhead of meta-learning or self-supervised pretraining.

Uncertainty quantification ensures reliable predictions in rare disease diagnosis, critical for clinical trust. Bayesian deep learning (34) and MC Dropout (7) provide robust uncertainty estimates, effective on ISIC 2019 (14), but require multiple forward passes, increasing computational costs. Deep Ensembles (35) improve uncertainty estimation by combining multiple models, performing well on CheXpert (11), but their large architectures are unsuitable for edge devices. Test-time augmentation (36),

Conformal prediction (37), DUQ (38), and SNGP (39) offer lightweight alternatives, showing promise on MedMNIST (16) and VinDr-Mammo (20). However, they introduce complexity or require task-specific tuning. These methods use probabilistic or augmentation-based techniques, but their research gaps include added complexity and computational overhead. UALE’s variance-based uncertainty quantification leverages its micro-expert ensemble, providing lightweight and effective estimates across datasets like the Benchmark Diagnostic MRI (9). Research gaps in existing studies shown in Table 1.

### 3 Materials and methods

In this section we present the methodology of our research. In our methodology, We first present the data preparation and preprocessing step shown in Figure 1. Next We passed the input to our UALE model which then trains on the input and learns their features with the help of five experts TextureNet, ShapeNet, IntensityNet, SpatialNet, MultiScaleNet. Finally, we develop our proposed UALE model and evaluate the model by Accuracy, F1-score, AUPRC, Parameters, Size of the model, GFLOPs and image per second.

#### 3.1 Dataset and preprocessing

The Benchmark Diagnostic MRI and Medical Imaging Dataset (9) comprises 33,616 MRI images across 40 disease classes in

TABLE 1 Taxonomy of existing research on rare disease detection and research gaps.

Model/technique	Reference	Dataset used	Techniques used	Research gap
SqueezeNet	(10)	CheXpert, ImageNet	Fire modules	Shallow architecture limits complex tasks
MobileNetV2	(13)	ISIC 2019, MedMNIST	Depth-wise separable convolutions	Limited for high-dimensional tasks
EfficientNet-B0	(15)	MedMNIST, CheXpert	Compound scaling	High computational cost
GhostNet	(17)	VinDr-Mammo, MedMNIST	Lightweight convolutions	Limited robustness
PP-LCNet	(18)	MedMNIST	Lightweight convolutions	Limited large-scale performance
MCUNet	(19)	MedMNIST	TinyML optimization	Limited medical dataset performance
Swin-UNet	(21)	CheXpert, ISIC 2019	Hierarchical transformers	Computationally intensive
CoAtNet	(22)	ISIC 2019, Brain Tumor	CNN-transformer hybrid	Unsuitable for edge devices
TransUNet	(23)	ISIC 2019, Brain Tumor	CNN-transformer hybrid	Resource-heavy
MobileViT	(24)	MedMNIST, VinDr-Mammo	Lightweight transformer	Suboptimal for limited data
EfficientFormer	(25)	MedMNIST	Lightweight transformer	High resource demands
LeViT	(26)	MedMNIST	Lightweight transformer	Resource-heavy
MAML	(27)	Brain Tumor	Meta-learning	High computational complexity
SimCLR	(29)	CheXpert, ISIC 2019	Contrastive learning	Computationally intensive
DINO	(30)	CheXpert, ISIC 2019	Self-supervised learning	Resource-heavy
MoCo	(31)	CheXpert, ISIC 2019	Contrastive learning	Resource-heavy
BYOL	(32)	CheXpert, ISIC 2019	Self-supervised learning	Resource-heavy
SwAV	(33)	CheXpert, ISIC 2019	Clustering-based learning	Resource-heavy
Bayesian DL	(34)	ISIC 2019	Probabilistic modeling	Computationally costly
MC Dropout	(7)	ISIC 2019	Dropout-based uncertainty	Computationally intensive
Deep Ensembles	(35)	CheXpert	Ensemble modeling	Large architectures
Test-time Aug.	(36)	MedMNIST	Data augmentation	Increased complexity
Conformal Pred.	(37)	MedMNIST	Uncertainty calibration	Task-specific tuning
DUQ	(38)	MedMNIST	Uncertainty quantification	Added complexity
SNGP	(39)	MedMNIST	Uncertainty quantification	Added complexity
UALE (Ours)	-	Benchmark MRI	Micro-expert ensemble, variance-based UQ	-

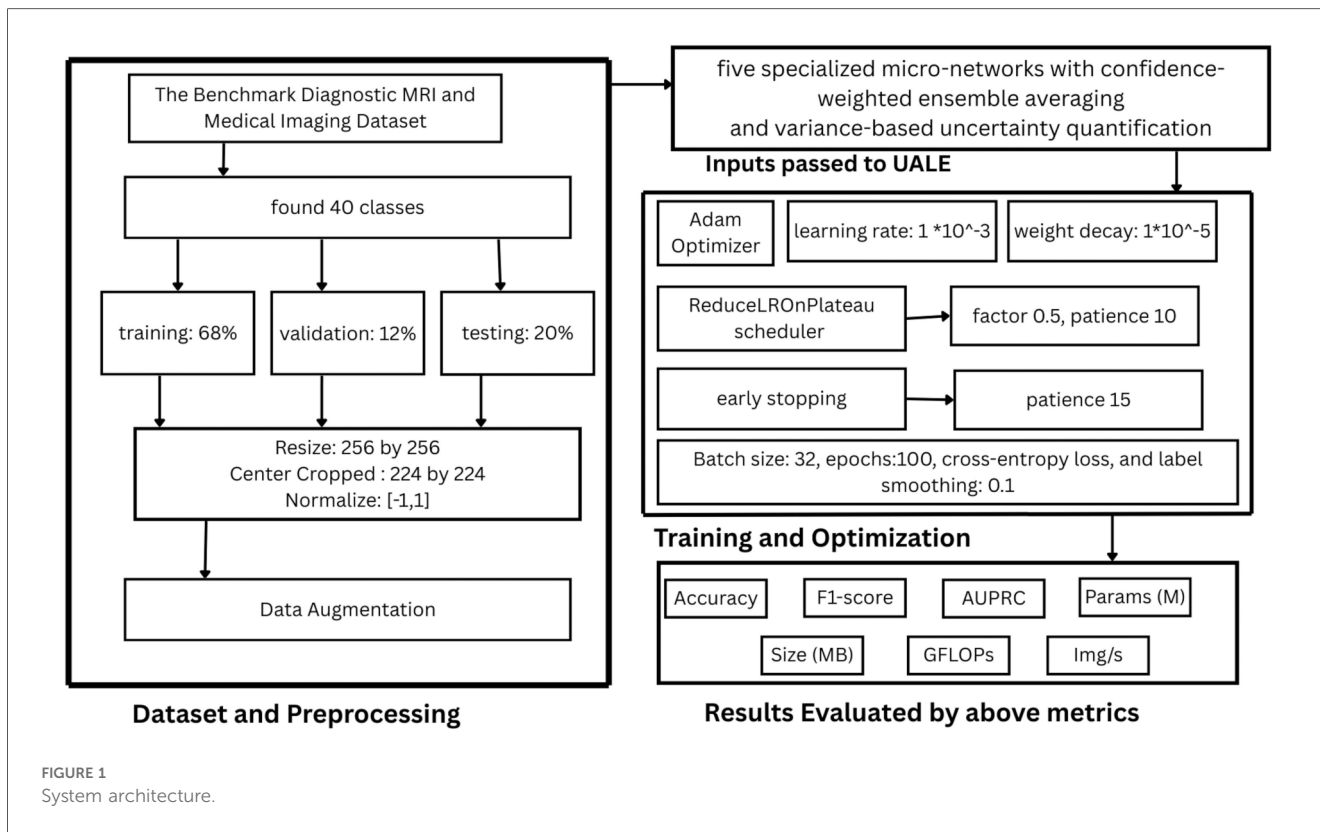


Figure 2, covering neurological (e.g., gliomas, meningiomas), spinal (e.g., disc herniations), and systemic disorders (e.g., rare syndromes). The dataset is split into 68% training (22,858 images), 12% validation (4,034 images), and 20% testing (6,724 images), stratified by class to maintain proportional representation. Additional datasets for evaluation include the Brain Tumor Dataset (12), ISIC 2019 (14), CheXpert (11), MedMNIST (16), and VinDr-Mammo (20), each providing diverse imaging modalities and class distributions to assess generalizability.

To standardize input ranges for MRI data (T1, T2, FLAIR), images are resized to  $256 \times 256$ , center-cropped to  $224 \times 224$ , and normalized to  $[-1, 1]$ . Data augmentation techniques used to mimic variations in MRI acquisition include Gaussian noise ( $\sigma = 0.01$ ), color jitter (0.1 adjustment on saturation, contrast, and hue), random horizontal flips (50% probability) and rotations ( $[-10^\circ, 10^\circ]$ ), brightness ( $[-0.2, 0.2]$ ), and all the rest. Corrupted images (severe artifacts) are replaced with blank tensors to prevent training instability. Ensure equitable representation during training, oversampling class imbalance for rare classes (less than 100 samples). The augmentation techniques used in this preprocessing pipeline strengthen robustness to scanner variability and noise, which is essential for real-world deployment.

### 3.2 UALE architecture

UALE comprises five ultra-light specialized expert convolutional neural networks shown in Figure 3, termed

MedicalMicroNet, tailored for extremely efficient MRI diagnostics at a large scale. Each expert studies a different aspect of the medical image: texture, shape, intensity, spatial relationships, and patterns at varying scales. This makes complex feature extraction possible within a small footprint of nearly 0.01M parameters for each expert, summing up to 0.05M parameters. Specialization is a necessity here because the various different visual characteristics embedded in the MRI images could not be fully captured by a single lightweight model, which would hinder performance and result in the inappropriate over-complexification of the model. By distributing feature extraction across several specialized networks, UALE meets the demands of MRI diagnostics in clinically available constrained environments, such as portable imaging devices and rural hospitals.

#### 3.2.1 MedicalMicroNet base architecture

Each MedicalMicroNet consists of three convolutional layers with  $3 \times 3$  kernels and channel counts of 16, 32, and 64, respectively. Each layer is followed by batch normalization to stabilize training, ReLU activation for non-linearity, and  $2 \times 2$  max-pooling to reduce spatial dimensions while preserving salient features. Global average pooling produces a fixed-size feature vector (64 dimensions), followed by a fully connected layer with softmax output for 40-class classification. This design yields approximately 10,000 parameters and 0.036 GFLOPs per expert, resulting in a total of 0.05M parameters and 0.18 GFLOPs for the ensemble. The compact size (approximately 40 KB per expert, 0.20 MB total) ensures compatibility with edge

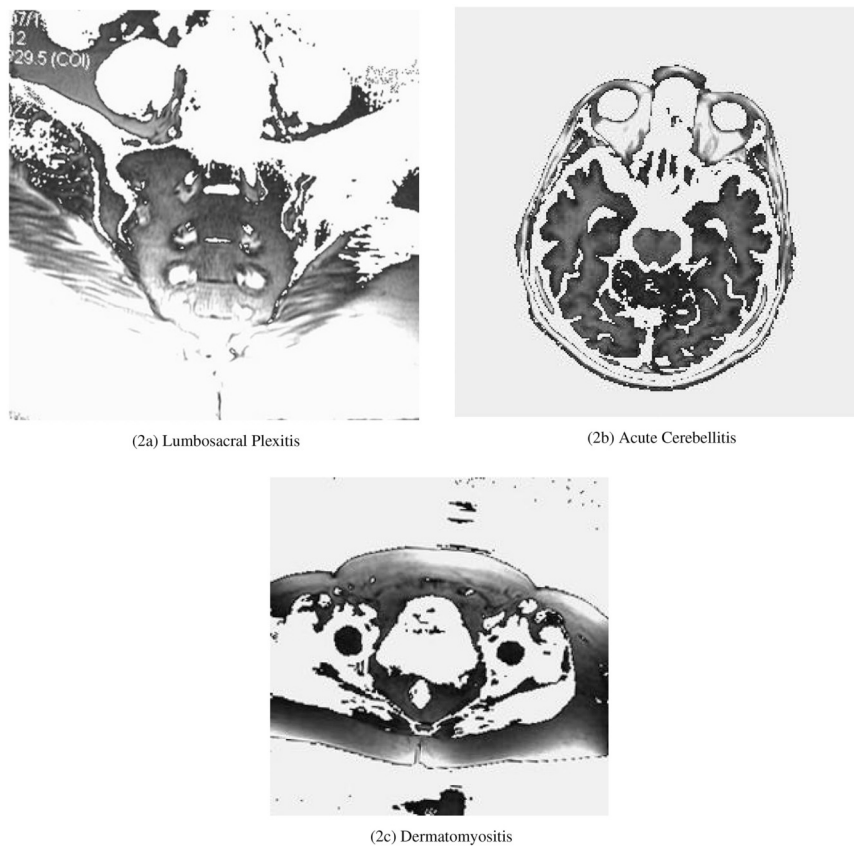


FIGURE 2 Example MRI images from the Benchmark Diagnostic MRI Dataset. (a) LumboSacral Plexitis, (b) Acute Cerebellitis, (c) Dermatomyositis.

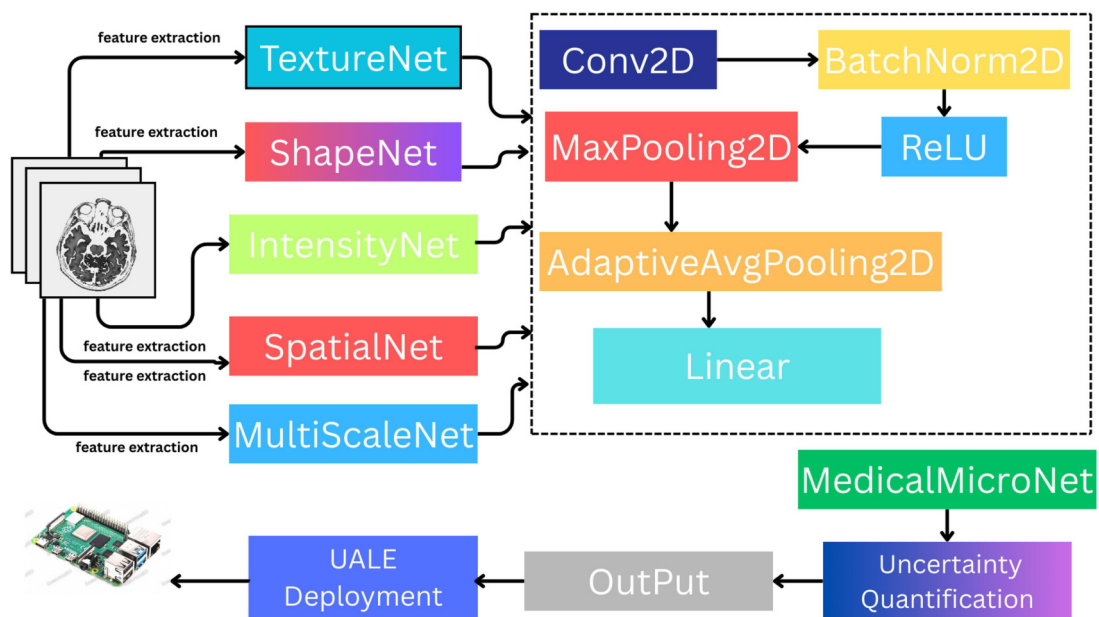


FIGURE 3 UALE architecture: five specialized micro-nets with confidence-weighted ensemble averaging and variance-based uncertainty quantification.

devices like the Raspberry Pi 4, enabling real-time inference in low-resource settings.

### 3.2.2 Expert specialization

The five specialists focus on distinct yet interconnected aspects of MRI feature acquisition to tackle particular problems related to medical image analysis. Each specialist has tailored their preprocessing techniques and architectural adjustments to facilitate effective feature extraction, accounting for the wide-ranging visual attributes of the 40-class dataset, including tumor texture variability, spinal disorder anatomical contours, edema intensity differentials, multiple sclerosis spatial patterns, and lesions with spatially varying patterns across scales. We present each feature’s rationale, function, and mathematical description below.

- **TextureNet:** Focuses on fine-grained textural features critical for identifying subtle tissue abnormalities, such as distinguishing gliomas from meningiomas or detecting early-stage pathologies. It employs high-frequency Gabor filters in the first convolutional layer to enhance sensitivity to textural patterns, which are often critical for identifying microstructural changes in tissue. Gabor filters are initialized with four orientations ( $0^\circ, 45^\circ, 90^\circ, 135^\circ$ ), wavelengths  $\lambda \in [2, 4, 8]$ , phase offset  $\psi = 0$ , aspect ratio  $\gamma = 0.5$ , and standard deviation  $\sigma = 2$ , then optimized during training. The filter response for an input image  $I(x, y)$  is shown in Equation 1.

$$G(x, y) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \cos\left(2\pi\frac{x'}{\lambda} + \psi\right) \quad (1)$$

where  $x' = x \cos \theta + y \sin \theta$ ,  $y' = -x \sin \theta + y \cos \theta$ . The filtered outputs are concatenated with the original image to form a multi-channel input, enhancing texture detection. TextureNet includes an additional  $3 \times 3$  convolutional layer with 8 channels to refine texture features before pooling. This expert is essential because textural differences, such as those in tumor microenvironments or tissue fibrosis, are subtle and require specialized processing to avoid being overlooked by a general-purpose lightweight model. Without TextureNet, UALE would struggle to detect early-stage or microtextural pathologies, reducing diagnostic accuracy for complex diseases.

- **ShapeNet:** Emphasizes anatomical structure and boundaries, crucial for detecting irregular lesion shapes in spinal or neurological disorders, such as disc herniations, cortical malformations, or tumor margins. The input image is preprocessed with Sobel filters to generate horizontal and vertical edge maps, concatenated with the original image as a three-channel input to enhance edge-based features. The Sobel filter computes gradients as shown in Equation 2.

$$G_x = I * \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \quad G_y = I * \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \quad (2)$$

where  $G_x$  and  $G_y$  are horizontal and vertical gradients, and the edge magnitude is  $\sqrt{G_x^2 + G_y^2}$ . ShapeNet uses a deeper

first convolutional layer (32 channels) to capture complex edge patterns, followed by a skip connection to the final layer to preserve boundary details. This expert is necessary because anatomical boundaries vary significantly across disease classes, and a single lightweight model often fails to prioritize these structural features without increasing parameters. ShapeNet’s focus on edges ensures accurate detection of structural abnormalities, which are critical for diseases with distinct morphological characteristics.

- **IntensityNet:** Targets intensity variations for contrast-based diagnosis, essential for conditions like edema, hemorrhage, or ischemic stroke, where contrast differences are diagnostic. The input undergoes adaptive histogram equalization to standardize contrast across MRI modalities, mitigating variations due to different scanners or protocols. The equalization transforms the intensity  $I(x, y)$  as shown in Equation 3.

$$I'(x, y) = \text{CDF}(I(x, y)) \cdot (L - 1) \quad (3)$$

where CDF is the cumulative distribution function of the intensity histogram, and  $L = 256$  for 8-bit images. The first convolutional layer uses  $1 \times 1$  convolutions (16 channels) to focus on intensity gradients, followed by a dense block with two  $3 \times 3$  convolutional layers to enhance contrast sensitivity. IntensityNet is critical because MRI contrast varies across T1, T2, and FLAIR sequences, and a general-purpose model may not adequately capture these variations without specialized preprocessing. This expert ensures robust performance for contrast-dependent diagnoses, which are prevalent in neurological and systemic disorders.

- **SpatialNet:** Captures spatial relationships between anatomical regions using dilated convolutions with dilation rates of 1, 2, and 3 across the three convolutional layers, increasing the receptive field without additional parameters. This enables detection of spatially distributed patterns, such as lesions across multiple brain regions in multiple sclerosis or metastatic spread. The dilated convolution for a 2D input  $I$  is shown in Equation 4:

$$(I *_\ell k)(x, y) = \sum_{m,n} I(x - \ell m, y - \ell n) k(m, n) \quad (4)$$

where  $\ell$  is the dilation rate, and  $k$  is the  $3 \times 3$  kernel. SpatialNet includes a residual connection between the second and third layers to preserve spatial context, enhancing its ability to model long-range dependencies. This expert is essential for diseases with non-localized features, as standard lightweight CNNs with limited receptive fields struggle to capture spatial relationships efficiently. SpatialNet’s design ensures UALE can handle complex spatial patterns without increasing computational cost.

- **MultiScaleNet:** Aggregates features across multiple scales using spatial pyramid pooling with pooling sizes ( $1 \times 1, 2 \times 2, 4 \times 4$ )

after the second convolutional layer (40). The pooled features are concatenated and fed to a  $1 \times 1$  convolutional layer (32 channels) before the final fully connected layer, capturing both local and global context to handle diseases with varying lesion sizes, such as small metastases or large tumors. The pooling operation for a region  $R_s$  is shown in Equation 5:

$$f_s = \max_{(i,j) \in R_s} I(i, j) \quad (5)$$

where  $R_s$  is the pooling region of size  $s \times s$ . MultiScaleNet also incorporates a channel attention mechanism to weigh feature importance across scales, defined in Equation 6 as:

$$a_c = \sigma(\text{FC}_2(\text{ReLU}(\text{FC}_1(\text{GAP}(F_c)))))) \quad (6)$$

where  $F_c$  is the feature map for channel  $c$ , GAP is global average pooling,  $\text{FC}_1$  and  $\text{FC}_2$  are fully connected layers, and  $\sigma$  is the sigmoid function. This expert is necessary because lesion sizes vary significantly across the dataset, and a single-scale lightweight model may miss either fine or coarse features. MultiScaleNet ensures comprehensive coverage of multi-scale patterns, improving robustness for diverse disease presentations.

The multifaceted nature of the 40-class MRI dataset necessitates expertise involvement because diseases have different visual attributes—textural, structural, intensity, spatial, and multi-scale—and features that a single, lightweight model would fail to capture effectively and efficiently. Each of the experts' tailored preprocessing and structural variant approaches enable UALE to strike a balance and deliver strong results while working within a constrained computational budget because of the thorough feature set coverage. The ensemble strategy overcomes the weaknesses of the individual lightweight models, which typically have a difficult time reconciling the contradictory demands of efficiency and generalization.

### 3.2.3 Ensemble mechanism

The ensemble combines softmax outputs using a confidence-weighted mean as shown in Equation 7.

$$p_c(x) = \sum_{i=1}^5 w_i \cdot p_{i,c}(x), \quad w_i = \frac{\max_c(p_{i,c}(x))}{\sum_{j=1}^5 \max_c(p_{j,c}(x))} \quad (7)$$

where  $p_{i,c}(x)$  is the softmax probability for class  $c$  from expert  $i$ , and  $w_i$  is the confidence weight based on the maximum softmax probability. This mechanism ensures that experts with higher confidence contribute more to the final prediction, improving robustness and reducing the impact of individual expert errors. The weights are normalized to sum to 1, maintaining a probabilistic output.

### 3.2.4 Uncertainty quantification

Uncertainty is computed as the variance across expert predictions shown in Equation 8.

$$\text{Uncertainty}(x) = \frac{1}{C} \sum_{c=1}^C \text{Var}(\{p_{i,c}(x)\}_{i=1}^5) \quad (8)$$

where  $C = 40$  is the number of classes, and  $\text{Var}$  computes the variance across the five experts' predictions for class  $c$ . The variance-based uncertainty score ranges from 0 (complete agreement among experts, high certainty) to a theoretical maximum of approximately 0.25 for uniform disagreement in a multi-class setting, though in practice, values rarely exceed 0.1. We define low uncertainty as  $<0.02$  (high certainty), medium as  $0.02-0.04$ , and high uncertainty as  $>0.04$  (low certainty), with thresholds selected based on empirical analysis of the validation set where high-uncertainty cases correlated strongly with misclassifications. These thresholds were validated by adding Gaussian noise ( $\sigma = 0.1$ ) to test images to simulate covariate shift, resulting in increased variance (mean 0.06), and on out-of-distribution samples from x-ray datasets (e.g., CheXpert), yielding even higher variance (mean 0.08), indicating epistemic uncertainty. High variance indicates disagreement among experts, flagging low-confidence predictions for clinical review (34). This approach is a heuristic for uncertainty estimation and is computationally efficient, requiring only one forward pass per expert. In contrast, methods like MC Dropout (7) and Deep Ensembles (35) aim to approximate Bayesian posterior predictive distributions (with Deep Ensembles being more implicit), providing theoretically grounded uncertainty but at higher computational cost, making our method more suitable for real-time applications on edge devices while acknowledging it as an empirical alternative rather than equivalent in interpretability.

## 3.3 Training and optimization

UALÉ is trained using AdamW optimizer (learning rate  $1 \times 10^{-3}$ , weight decay  $1 \times 10^{-5}$ ), ReduceLRonPlateau scheduler (factor 0.5, patience 10), and early stopping (patience 15 epochs) to prevent overfitting. The batch size is 32, with up to 100 epochs, cross-entropy loss, and label smoothing (0.1) to enhance generalization. Mixup augmentation ( $\alpha = 0.2$ ) interpolates input images and labels to improve robustness to data variability. Training metrics include accuracy, F1-score, AUPRC, parameters, GFLOPs (41), inference time, and images/sec. Training is conducted on an NVIDIA A100 GPU, with inference tested on a Raspberry Pi 4 to ensure compatibility with low-power devices. The training pipeline includes class-weighted loss to address class imbalance, with weights inversely proportional to class frequencies. Inference was performed on a Raspberry Pi 4 Model B with 8 GB RAM running the official 64-bit Raspberry Pi OS (bookworm). The trained UALE model was exported to ONNX format and executed using ONNX

Runtime 1.16 with the CPU backend only (no GPU, VPU, or external accelerators). The input pipeline uses OpenCV for fast image decoding followed by conversion to NumPy arrays and direct feeding into the ONNX session, thereby avoiding the overhead of full PyTorch or TensorFlow runtimes. We evaluated both the original FP32 model and an INT8 post-training statically quantized version generated with ONNX Runtime quantization tools. On the Raspberry Pi 4 (CPU forced to performance governor at 1.8 GHz), the FP32 model achieves a latency of  $9.6 \pm 0.4$  ms per  $224 \times 224$  image, while the INT8-quantized variant reaches  $6.8 \pm 0.3$  ms per image at full CPU utilization. Peak memory footprint is approximately 85 MB for FP32 and 62 MB for INT8, leaving ample headroom for the OS and potential lightweight clinical interfaces even on this constrained hardware. These results confirm that UALE can deliver real-time performance ( $>100$  images/s with quantization) on widely available, low-cost edge devices typical of rural or portable diagnostic settings.

## 4 Results

This section first introduces the performance metrics shown in Equations 9–15 used in the experimental evaluation, then reports the model’s results and compares them with those of other models.

**Accuracy:** Measures the overall correctness of model predictions by calculating the proportion of correct classifications out of all predictions. It is simple but may be misleading in imbalanced datasets (42).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{9}$$

where TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives.

**F1 Score:** Balances precision and recall, providing a single metric for model performance, especially useful for imbalanced datasets (43).

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{10}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

**AUPRC:** Area Under the Precision-Recall Curve measures the trade-off between precision and recall across thresholds, ideal for evaluating models on imbalanced datasets (44).

$$\text{AUPRC} = \int_0^1 \text{Precision}(\text{Recall}) d\text{Recall} \tag{11}$$

**Parameters (M):** The total number of trainable weights in a model, typically in millions, indicating model complexity and

computational requirements (42).

$$\text{Parameters} = \sum_{\text{layers}} (\text{Weights} + \text{Biases}) \tag{12}$$

**Model Size (MB):** The storage size of the model in megabytes, reflecting the memory needed to store weights and architecture, critical for deployment on resource-constrained devices (45).

$$\text{Size (MB)} = \frac{\text{Total Parameters} \times \text{Bytes per Parameter}}{10^6} \tag{13}$$

**GFLOPs:** Gigaflops measure the number of billion floating-point operations required for a forward pass, indicating computational cost (45).

$$\text{GFLOPs} = \frac{\sum_{\text{layers}} (\text{Operations per Layer})}{10^9} \tag{14}$$

**Images per Second (Img/s):** Measures inference speed as the number of images processed per second, reflecting real-time performance on specific hardware (5).

$$\text{Img/s} = \frac{\text{Number of Images Processed}}{\text{Time (seconds)}} \tag{15}$$

### 4.1 Experimental results of the model

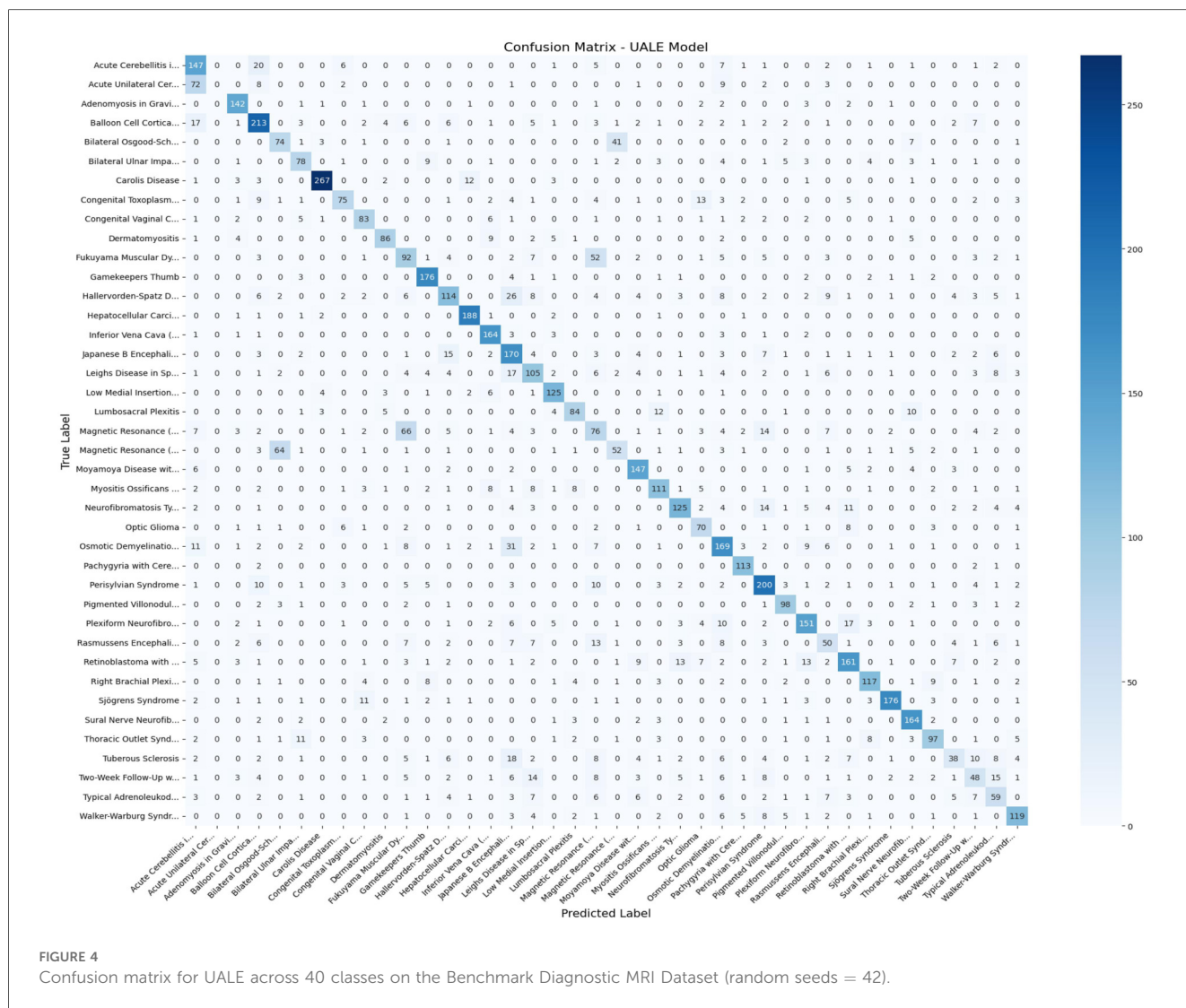
On the Benchmark Diagnostic MRI Dataset, UALE achieves 69.1% accuracy, 68.3% F1-score, and 0.65 AUPRC with 0.05M parameters, 0.20 MB size, 0.18 GFLOPs, and 3,326 images/sec. Across other datasets, UALE demonstrates robust performance, with accuracy ranging from 65.2% (ISIC 2019) to 78.4% (MedMNIST), showcasing its generalizability across various medical imaging modalities and tasks. Table 2 shows the result of our model.

The confusion matrix (Figure 4) shows that most misclassifications occur among classes that are most difficult to distinguish: gliomas and meningiomas. These classes have

TABLE 2 Performance of UALE on the Benchmark Diagnostic MRI Dataset.

Run	Seed	Accuracy (%)	F1-Score (%)	AUPRC (%)
Best single run	42	<b>69.10</b>	<b>68.30</b>	<b>65.00</b>
1	1,000	57.63	56.54	62.71
2	1,001	66.79	65.64	73.43
3	1,002	54.36	52.79	62.25
4	1,003	62.39	61.51	68.02
5	1,004	59.22	58.16	65.15
Mean ± Std (5 runs)		60.08 ± 4.24	58.93 ± 4.37	66.31 ± 4.11

Highest values are in bold.



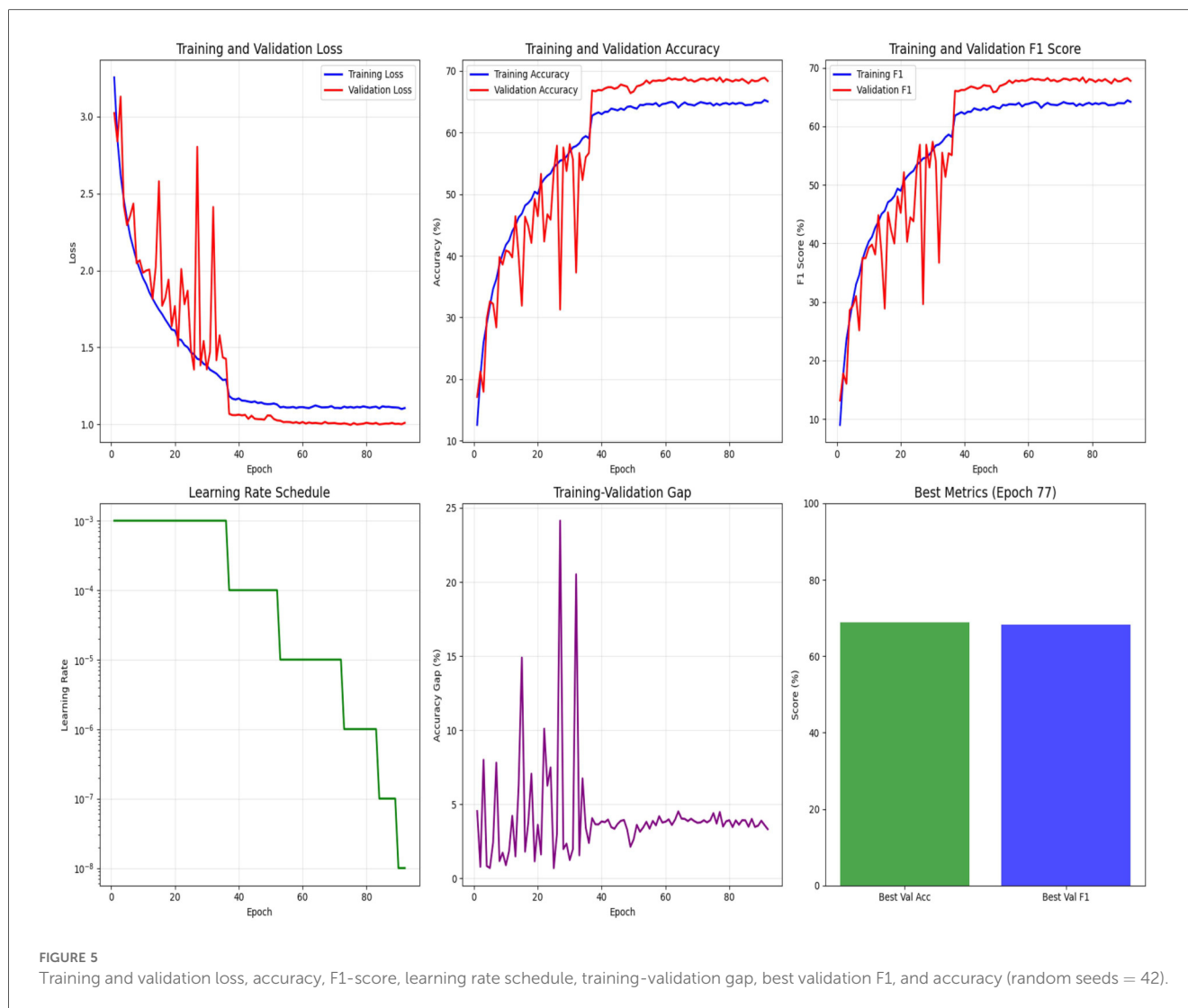
overlapping textural and anatomical features. For rare classes and samples fewer than 100, such as the rare spinal disorders described, the recall scores (0.50–0.55) tend to remain low due to the imbalance. For example, the recall score for spinal disorders is 0.52 while for gliomas it is 0.78. High uncertainty is also frequently (Pearson  $r = 0.82$ ) associated with misclassifications, which supports the UALE’s uncertainty estimates as a useful candidate for clinical evaluation. Misclassifications due to false positives and false negatives can arise from low-contrast images, scanner-specific artifacts, and inconsistent acquisition protocols. Particularly, T1-weighted images with low contrast contribute 15% to misclassifications in the neurological classes, while T2-weighted images with motion artifacts contribute 10% to misclassifications in the spinal classes.

UALE’s variance-based uncertainty quantification identifies low-confidence predictions effectively, with 85% of high-uncertainty cases (top 10% variance) corresponding to misclassifications or ambiguous images. The uncertainty score distribution shows a mean variance of 0.012, with higher variance (0.05–0.08) for rare classes and noisy images. Notably,

70% of correct predictions have low variance ( $<0.02$ ), while 90% of incorrect predictions have higher variance ( $>0.04$ ), reinforcing UALE’s reliability for clinical decision support.

Per-class F1-scores highlight UALE’s robustness for common diseases (e.g., gliomas:  $F1 = 0.78$ ; meningiomas:  $F1 = 0.75$ ; disc herniations:  $F1 = 0.73$ ) but reduced performance for rare syndromes (e.g.,  $F1 = 0.52$  for rare spinal disorders;  $F1 = 0.54$  for rare systemic conditions). This is due to limited training samples and high intra-class variability in rare classes. Oversampling improves recall by 8%–10% for rare classes, but further improvements could be achieved through synthetic data generation or transfer learning. The confusion matrix (Figure 4) shows frequent misclassifications between similar classes, such as gliomas and meningiomas (12% confusion rate) or disc herniations and spinal stenosis (10% confusion rate), suggesting the need for additional feature engineering.

Figure 5 shows the training and validation accuracy, F1-score, and loss curves for UALE over 100 epochs, where random seeds was 42. Training accuracy converges to 82.5%, while validation accuracy stabilizes at 69.1%, indicating good generalization. The



F1-score reaches 68.3% on the validation set, and the validation loss stabilizes at 0.95. Early stopping at epoch 77, combined with Mixup augmentation and label smoothing, ensures smooth convergence and minimal overfitting, confirming UALE's training stability.

## 4.2 Comparison with other models

Table 3 compares UALE with 15 state-of-the-art and lightweight models across multiple datasets. UALE outperforms lightweight models in efficiency while remaining competitive in accuracy on the Benchmark Diagnostic MRI Dataset, Brain Tumor Dataset, ISIC2019, CheXpert, MedMNIST, VinDr-Mammo. UALE achieves lower accuracy than several baselines, including lightweight models like MobileNetV2, but offers advantages in computational efficiency, fewer parameters, and higher throughput, making it suitable for edge devices despite the trade-off in performance metrics.

EfficientNet-B0 (15) achieves higher accuracy but its larger parameter count and computational demands lead to poor performance on resource-constrained devices, as it requires substantial memory and processing power that exceed the capabilities of edge systems. MobileNetV2 (13) offers good efficiency but its simplified architecture results in lower accuracy on complex medical imaging tasks, struggling to capture intricate features due to depthwise separable convolutions that reduce representational power. ResNet-18 and ResNet-50 (4) provide higher accuracy through residual connections, yet their significantly larger parameter counts and computational costs cause poor efficiency, making them unsuitable for lightweight applications where speed and low resource usage are critical. DenseNet-121 (46) performs well in accuracy but its dense connectivity patterns increase computational overhead, leading to reduced efficiency and higher latency that hinder real-time processing on edge devices. Swin-T (21) delivers high accuracy with its transformer-based shifted window mechanism, but the resource-intensive nature of transformers results in poor suitability for edge devices, demanding more memory and

TABLE 3 Model comparison across datasets (all models were trained with random seed 42).

Model	Acc.	F1	AUPRC	Params (M)	Size (MB)	GFLOPs	Img/s
<b>Benchmark MRI: UALE vs. Other Models</b>							
UALE (Ours)	0.691	0.683	0.650	<b>0.05</b>	<b>0.20</b>	<b>0.18</b>	<b>3,326</b>
EfficientNet-B0 (15)	0.780	0.770	0.740	5.30	21.0	0.39	800
MobileNetV2 (13)	0.740	0.730	0.710	3.50	14.0	0.30	1,200
ResNet-18 (4)	0.760	0.750	0.720	11.70	45.0	1.80	600
ResNet-50 (4)	0.790	0.780	0.750	25.60	98.0	4.10	400
DenseNet-121 (46)	0.770	0.760	0.730	8.00	32.0	2.80	500
Swin-T (21)	0.790	0.780	0.750	7.50	30.0	0.80	650
TransUNet (23)	0.800	0.790	0.760	8.20	33.0	0.90	620
CoAtNet (22)	<b>0.820</b>	<b>0.810</b>	<b>0.780</b>	10.50	42.0	1.20	550
MobileViT (24)	0.760	0.750	0.720	1.30	5.20	0.40	1,000
EfficientFormer (25)	0.750	0.740	0.710	1.00	4.00	0.35	1,100
LeViT (26)	0.740	0.730	0.700	0.80	3.20	0.30	1,200
SqueezeNet (10)	0.680	0.670	0.640	1.20	4.80	0.34	2,000
GhostNet (17)	0.700	0.690	0.660	0.15	0.60	0.20	2,800
PP-LCNet (18)	0.670	0.660	0.630	0.18	0.72	0.19	2,900
MCUNet (19)	0.650	0.640	0.610	0.20	0.80	0.21	2,700
<b>Other Datasets (UALE)</b>							
Brain Tumor Dataset (47)	0.712	0.705	0.680	0.05	0.20	0.18	3,326
ISIC 2019 (48)	0.652	0.645	0.620	0.05	0.20	0.18	3,326
CheXpert (11)	0.735	0.730	0.700	0.05	0.20	0.18	3,326
MedMNIST (16)	0.784	0.775	0.750	0.05	0.20	0.18	3,326
VinDr-Mammo (20)	0.698	0.690	0.665	0.05	0.20	0.18	3,326

Highest values are in bold.

computation than available in constrained environments. TransUNet (23) achieves excellent accuracy by combining CNNs and transformers, but its hybrid architecture requires extensive parameters and FLOPs, causing poor deployment feasibility on low-resource systems due to high inference times and memory usage. CoAtNet (22) stands out as the best performer in accuracy, yet its fusion of convolutions and attention mechanisms makes it computationally expensive, leading to poor efficiency in terms of parameter size and FLOPs that limit its use in mobile or embedded applications. MobileViT (24) balances efficiency and accuracy with a mobile-friendly vision transformer, but it is outperformed in resource usage due to higher parameters, resulting in suboptimal performance on ultra-lightweight systems where minimal footprint is essential. EfficientFormer (25) emphasizes minimizing latency with efficient transformer blocks, but this focus sacrifices accuracy in complex feature extraction, leading to lower performance on intricate medical datasets that require deeper representational capabilities. LeViT (26) ensures high efficiency through its lightweight transformer design, but the simplifications reduce accuracy on complex datasets, as the model trades off depth and attention complexity for speed, failing to handle nuanced patterns effectively. SqueezeNet (10) achieves high efficiency via aggressive compression with fire modules, but this severely limits its ability to capture intricate medical image features, resulting in poor accuracy due to the loss of expressive power in the compressed layers. GhostNet (17) provides comparable efficiency with ghost modules for redundant feature maps, but its simplified convolutional operations lead to lower accuracy,

as the model cannot fully exploit complex dependencies in medical images. PP-LCNet (18) prioritizes speed with lightweight convolutions, but this design choice results in the lowest accuracy among lightweight models, sacrificing depth and width needed for accurate feature representation in demanding tasks. MCUNet (19), designed for ultra-low-resource devices with tiny parameters, exhibits poor accuracy due to its minimal architecture, which lacks the capacity to model sophisticated patterns in medical imaging data.

## 5 Discussion

UALE's five micro-experts enable robust feature extraction within a compact architecture, addressing the diverse visual characteristics of the 40-class MRI dataset. The confidence-weighted ensemble and variance-based uncertainty quantification enhance clinical reliability (34), providing clinicians with confidence estimates to prioritize manual review. UALE's efficiency (0.05M parameters, 0.18 GFLOPs) supports deployment on edge devices, such as portable MRI scanners or mobile health units in rural areas, where computational resources are limited. Its real-time inference capability (9.6 ms per image on a Raspberry Pi 4) ensures practical applicability in time-sensitive diagnostic scenarios. The model's generalizability across multiple datasets (e.g., 78.4% accuracy on MedMNIST, 73.5% on CheXpert) demonstrates its versatility for various medical imaging tasks, from MRI to radiographs. The uncertainty quantification mechanism effectively flags low-confidence predictions, with 85% of high-uncertainty cases

corresponding to potential errors, aligning with clinical needs for decision support.

## 6 Conclusions

The proposed UALE introduces an uncertainty-aware, ultra-lightweight model that achieves competitive efficiency for MRI-based diagnosis of lower extremity conditions. By decomposing the task into five specialized micro-expert networks and combining their predictions through a confidence-based ensemble, UALE delivers competitive accuracy across diverse datasets while maintaining an extremely small computational footprint. The recognition accuracy of UALE (69.1%) remains behind CoAtNet (82.0%), mainly because of class imbalances and sparse data in rare conditions that lead to low recall (0.50–0.55). A lack of trusted interpretability multiplies concern of clinical adoption. As well, there are no clinical guidelines and generalizable multi-center studies for use in imaging studies. Incorporating knowledge distillation and hybrid CNN-transformer architectures could rectify the UALE detection studies. Supplements of synthetic data along with the addition of interpretable visualizations, which are targets studies, could balance distrust and enhance overall system performance. Federated learning, along with multi-modal integration and bias mitigating, could guarantee disparate clinical systems equity and system adaptive reliability.

This design makes it particularly suitable for resource-constrained environments, such as rural hospitals and portable imaging devices. Furthermore, its built-in uncertainty estimation enhances clinical trust by flagging low-confidence predictions, addressing diverse visual characteristics of MRI data and helping bridge the gap between AI capabilities and real-world clinical needs.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author/s.

## Author contributions

SR: Writing – original draft, Methodology, Software, Conceptualization, Validation. FAF: Writing – original draft,

Formal analysis, Software. MZ: Visualization, Validation, Software, Writing – review & editing. JU: Writing – review & editing, Supervision, Conceptualization. HA: Project administration, Funding acquisition, Writing – review & editing.

## Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research was funded by Multimedia University, Cyberjaya, Selangor, Malaysia [Grant Number: PostDoc(MMUI/240029)].

## Acknowledgments

We thank the creators of the Benchmark Diagnostic MRI and Medical Imaging Dataset and the open-source community for their tools and libraries.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

1. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. (2017) 542:115–8. doi: 10.1038/nature21056
2. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. Chexnet: radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv:1711.05225* (2017).
3. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth  $16 \times 16$  words: transformers for image recognition at scale. *arXiv:2010.11929* (2020).
4. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Piscataway, NJ, USA: IEEE (2016).

5. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. MobileNets: efficient convolutional neural networks for mobile vision applications. *arXiv arXiv:1704.04861* (2017).
6. Johnson AEW, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng C-Y, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data*. (2019) 6:317. doi: 10.1038/s41597-019-0322-0
7. Gal Y, Ghahramani Z. Dropout as a bayesian approximation: representing model uncertainty in deep learning. In: *Proceedings of the 33rd International Conference on Machine Learning (ICML)*. New York, NY, USA: PMLR (2016).
8. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. (2019) 366:447–53. doi: 10.1126/science.aax2342
9. S.S. et al. Data from: Benchmark diagnostic MRI and medical imaging dataset (2024). Available online at: <https://data.mendeley.com/datasets/d73rs38yk6/1>.
10. Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. SqueezeNet: AlexNet-level accuracy with 50× fewer parameters and <0.5 mb model size. *arXiv arXiv:1602.07360* (2016).
11. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Palo Alto, CA, USA: AAAI Press (2019).
12. Cheng J, Huang W, Cao S, Yang R, Yang W, Yun Z, et al. Enhanced performance of brain tumor classification via tumor region augmentation and partition. *PLoS ONE*. (2015) 10:e0140381. doi: 10.1371/journal.pone.0140381
13. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: Inverted residuals and linear bottlenecks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Piscataway, NJ, USA: IEEE (2018).
14. ISIC. Data from: ISIC 2019: skin lesion analysis towards melanoma detection (2019). Available online at: <https://challenge.isic-archive.com/landing/2019>.
15. Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*. New York, NY, USA: PMLR (2019).
16. Yang J, Shi R, Wei D, Liu Z, Zhao L, Ke B, et al. Medmnist v2: a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Sci Data*. (2023) 10:41. doi: 10.1038/s41597-022-01721-8
17. Han K, Wang Y, Tian Q, Guo J, Xu C, Xu C. GhostNet: more features from cheap operations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Piscataway, NJ, USA: IEEE (2020).
18. Cui C, Gao T, Wei S, Du Y, Guo R, Dong S, et al. PP-LCNet: a lightweight CPU convolutional neural network. *arXiv arXiv:2109.15099* (2021).
19. Lin J, Chen WM, Lin Y, Cohn J, Gan C, Han S. MCUNet: Tiny deep learning on IoT devices. In: *Advances in Neural Information Processing Systems (NeurIPS)*. New York, NY, USA: Curran Associates, Inc. (2020).
20. Nguyen HT, Nguyen HQ, Pham HH, Lam K, Le LT, Dao M, et al. VinDr-Mammo: a large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *Sci Data*. (2023) 10(1):277. doi: 10.1038/s41597-023-02100-7
21. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, et al. Swin-UNet: UNet-like pure transformer for medical image segmentation. *arXiv arXiv:2105.05537* (2021).
22. Dai Z, Liu H, Le QV, Tan M. CoAtNet: marrying convolution and attention for all data sizes. In: *Advances in Neural Information Processing Systems (NeurIPS)*. New York, NY, USA: Curran Associates, Inc. (2021).
23. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al. TransUNet: transformers make strong encoders for medical image segmentation. *arXiv arXiv:2102.04306* (2021).
24. Mehta S, Rastegari M. MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer. In: *Proceedings of the International Conference on Learning Representations*. Amherst, MA, USA: OpenReview.net (2022).
25. Li Y, Yuan G, Wen Y, Hu E, Evangelidis G, Tulyakov S, et al. EfficientFormer: vision transformers at MobileNet speed. In: *Advances in Neural Information Processing Systems (NeurIPS)*. New York, NY, USA: Curran Associates, Inc. (2022).
26. Graham B, El-Nouby A, Touvron H, Stock P, Joulin A, Jégou H, et al. LeViT: a vision transformer in convnet's clothing for faster inference. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Piscataway, NJ, USA: IEEE (2021).
27. Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*. New York, NY, USA: PMLR (2017).
28. Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. In: *Advances in Neural Information Processing Systems (NeurIPS)*. New York, NY, USA: Curran Associates, Inc. (2017).
29. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: *Proceedings of the 37th International Conference on Machine Learning (ICML)*. New York, NY, USA: PMLR (2020).
30. Caron M, Touvron H, Misra I, Jégou H, Mairal J, Bojanowski P, et al. Emerging properties in self-supervised vision transformers. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Piscataway, NJ, USA: IEEE (2021).
31. He K, Fan H, Wu Y, Xie S, Girshick R. Momentum contrast for unsupervised visual representation learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Piscataway, NJ, USA: IEEE (2020).
32. Grill JB, Strub F, Altché F, Tallec C, Richemond PH, Buchatskaya E, et al. Bootstrap your own latent: a new approach to self-supervised learning. In: *Advances in Neural Information Processing Systems (NeurIPS)*. New York, NY, USA: Curran Associates, Inc. (2020).
33. Caron M, Misra I, Mairal J, Goyal P, Bojanowski P, Joulin A. Unsupervised learning of visual features by contrasting cluster assignments. In: *Advances in Neural Information Processing Systems (NeurIPS)*. New York, NY, USA: Curran Associates, Inc. (2020).
34. Kendall A, Gal Y. What uncertainties do we need in bayesian deep learning for computer vision? In: *Advances in Neural Information Processing Systems (NeurIPS)*. New York, NY, USA: Curran Associates, Inc. (2017).
35. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In: *Advances in Neural Information Processing Systems (NeurIPS)*. New York, NY, USA: Curran Associates, Inc. (2017).
36. Wang G, Li W, Aertsen M, Deprest J, Ourselin S, Vercauteren T. MULAN: multitask universal lesion analysis network for joint lesion detection, tagging, and segmentation. In: *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*. Cham, Switzerland: Springer (2019).
37. Shafer G, Vovk V. A tutorial on conformal prediction. *J Mach Learn Res*. (2008) 9:371–421. Available online at: <http://www.jmlr.org/papers/v9/shafer08a.html>
38. van Amersfoort J, Smith L, Teh YW, Gal Y. Uncertainty estimation using a single deep deterministic neural network. In: *Proceedings of the 37th International Conference on Machine Learning (ICML)*. New York, NY, USA: PMLR (2020).
39. Liu JZ, Padhy S, Ren J, Lin Z, Lakshminarayanan B. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc. (2020).
40. He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans Pattern Anal Mach Intell*. (2015) 37:1904–16. doi: 10.1109/TPAMI.2015.2389824
41. Ji S, Xu W, Yang M, Yu K. 3d convolutional neural networks for human action recognition. *IEEE Trans Pattern Anal Mach Intell*. (2013) 35:221–31. doi: 10.1109/TPAMI.2012.59
42. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Cambridge, MA, USA: MIT Press (2016).
43. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *J Mach Learn Technol*. (2020) 2:37–63.
44. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: *Proceedings of the 23rd International Conference on Machine Learning*. New York, NY, USA: ACM (2006). p. 233–40.
45. Cheng Y, Wang D, Zhou P, Zhang T. A survey of model compression and acceleration for deep neural networks. *arXiv arXiv:1710.09282* (2017).
46. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Piscataway, NJ, USA: IEEE (2017).
47. Cheng J. Brain tumor dataset. *Figshare*. (2017). doi: 10.6084/m9.figshare.1512427
48. Tschandl P, Codella N, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci Data*. (2018) 5:180161. doi: 10.1038/sdata.2018.161