Computerized Medical Imaging and Graphics 125 (2025) 102630

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

Inference time correction based on confidence and uncertainty for improved deep-learning model performance and explainability in medical image classification

Joel Jeffrey [a], Ashwin RajKumar [a], Sudhanshu Pandey [a], Lokesh Bathala [b], Phaneendra K. Yalavarthy [a,*]

[a] Department of Computational and Data Sciences, Indian Institute of Science, Bangalore, Karnataka, 560012, India
[b] Aster-CMI Hospital, Bangalore, Karnataka, 560092, India

ARTICLE INFO

Dataset link: https://www.kaggle.com/competitions/rsna-intracranial-hemorrhage-detection, https://www.kaggle.com/datasets/paultimothymooney/kermany2018, https://www.kaggle.com/competitions/rsna-breast-cancer-detection, https://zenodo.org/records/10519652

Keywords:
Explainable artificial intelligence
Interpretable artificial intelligence
Confidence
Entropy
Uncertainty
Deep learning

ABSTRACT

The major challenge faced by artificial intelligence (AI) models for medical image analysis is the class imbalance of training data and limited explainability. This study introduces a Confidence and Entropy-based Uncertainty Thresholding Algorithm (CEbUTAl), which is a novel post-processing method, designed to enhance both model performance and explainability. CEbUTAl modifies model predictions during inference, based on uncertainty and confidence measures, to improve classification in scenarios with class imbalance. CEbUTAl's inference-time correction addresses explainability, while simultaneously improving performance, contrary to the prevailing notion that explainability necessitates a compromise in performance. The algorithm was evaluated across five medical imaging tasks: intracranial hemorrhage detection, optical coherence tomography analysis, breast cancer detection, carpal tunnel syndrome detection, and multi-class skin lesion classification. Results demonstrate that CEbUTAl improves accuracy by approximately 5% and increases sensitivity across multiple deep learning architectures, loss functions, and tasks. Comparative studies indicate that CEbUTAl outperforms state-of-the-art methods in addressing class imbalance and quantifying uncertainty. The model-agnostic, task-agnostic and post-processing nature of CEbUTAl makes it appealing for enhancing both performance and trustworthiness in medical image analysis. This study provides a generalizable approach to mitigate biases arising from class imbalance, while improving the explainability of AI models, thus increasing their utility in clinical practice.

1. Introduction

Medical imaging techniques have revolutionized modern medicine, enabling the visualization of internal structures and functions. These techniques are instrumental in early disease detection and contribute to improved patient outcomes (Panayides et al., 2020; Kalemaki et al., 2020). However, the medical image analysis (MIA) workflow relies heavily on human practitioners, who are constrained by subjectivity and fatigue (Alexander et al., 2022). While machine learning (ML) has aided decision making in MIA, it has a strong reliance on manual feature selection, which is a significant limitation (Jahangir et al., 2024). Advancements in information and communication technologies have radically increased data availability and computational capacity, paving the way for the adoption of deep learning (DL) models in MIA (Lee et al., 2017). Deep learning has shown promising results in several MIA applications (Leibig et al., 2017; Hamedani-KarAzmoudehFar et al., 2023), with convolutional neural networks, originally inspired by LeNet (LeCun et al., 1998) and AlexNet (Krizhevsky et al., 2012) architectures, demonstrating remarkable performance in the binary classification of breast cancer tumors (Balasubramaniam et al., 2023; Titoriya and Sachdeva, 2019). Current DL models have also surpassed human accuracy (Kim et al., 2019; McKinney et al., 2020; Iqbal et al., 2021), fueling excitement regarding the potential of artificial intelligence (AI) in MIA. This is reflected in the exponential increase of AI-based MIA investigations (Tang, 2019). Although DL models are mathematical frameworks (Higham and Higham, 2019), they are highly complex (Zhang et al., 2021) and lack decomposability (Lipton, 2018), operating as "black boxes" that lack insight into the underlying mechanisms (Muhammad and Bendechache, 2024). This complexity, coupled with the scarcity of data in positive disease cases (Yu et al., 2022), highlights the need for further research and development.

* Corresponding author.
E-mail address: yalavarthy@iisc.ac.in (P.K. Yalavarthy).

https://doi.org/10.1016/j.compmedimag.2025.102630
Received 12 November 2024; Received in revised form 25 April 2025; Accepted 5 August 2025
Available online 13 August 2025

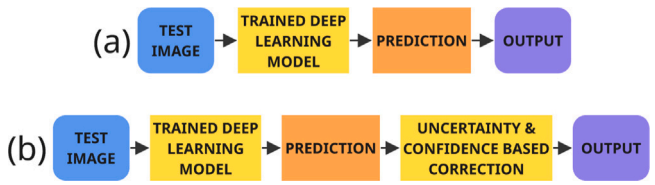

Fig. 1. Overview of deep learning (DL)-based approaches for medical image classification. (a) Conventional methods where a trained and validated DL model is employed to generate prediction. (b) The proposed approach, incorporating entropy-based uncertainty and confidence into the framework.

As AI is integrated into several areas, including healthcare, MIA, and clinical decision-making processes, the "black box" effect of DL models, particularly in critical clinical settings has triggered increased interest in eXplainable AI (XAI) (Gerlings et al., 2021; Dwivedi et al., 2023). Visual tools such as class activation maps (CAM), Grad-CAM (Panwar et al., 2020), and Grad-CAM++ (Chattopadhay et al., 2018) have been used to improve the explainability of AI. However, the reliability of these maps was questionable, emphasizing the need for caution when using them (Zhang et al., 2022). Alternatively, "non-visual" methods such as uncertainty evaluation (Abdar et al., 2021; Araújo et al., 2020; Dolezal et al., 2022; Angelopoulos et al., 2023; Huang et al., 2024; Rufibach, 2010) and uncertainty quantification (UQ) (Leibig et al., 2017; Hamedani-KarAzmoudehFar et al., 2023; Huang et al., 2024; Kurz et al., 2022; Lambert et al., 2024; Ayhan et al., 2020; Asgharnezhad et al., 2022; Gal and Ghahramani, 2016; Zou et al., 2023; Lakshminarayanan et al., 2017; Kimura, 2021) have been investigated to act as tools for explainability. While uncertainty evaluation estimates the uncertainty of model predictions (Abdar et al., 2021; Araújo et al., 2020; Dolezal et al., 2022; Angelopoulos et al., 2023; Huang et al., 2024; Rufibach, 2010), it offers only explainability without improving performance (Kendall and Gal, 2017). Conversely, UQ methods possess inherent characteristics that can enhance model confidence in its prediction. Existing UQ methods are contingent on the trained model architecture, producing an immutable framework that is not generalizable (Lahoti et al., 2023, 2021). Moreover, attempts to improve model explainability have resulted in a decline in performance (Wanner et al., 2021).

This work proposes "CEbUTAl" (Confidence- and Entropy-based Uncertainty Thresholding Algorithm), a post-hoc correction mechanism, that adaptively refines model predictions at inference time, based on UQ. CEbUTAl is specifically designed to address challenges posed by severe class imbalance, a prevalent issue in medical imaging datasets, which can lead to biased learning and suboptimal generalization. A high-level comparison of existing DL-based approaches with CEbUTAl is illustrated in Fig. 1. This study has two primary goals: (i) to enhance the model's explainability, and (ii) to maintain or improve the performance of the current model. The efficacy of CEbUTAl was evaluated across five medical image classification tasks: (i) intracranial hemorrhage (ICH) detection, (ii) optical coherence tomography (OCT) analysis, (iii) breast cancer detection (BCD), (iv) carpal tunnel syndrome (CTS) detection, and (v) multi-class skin lesion classification (SLC). Subsequently, to demonstrate its independence from model architectures and loss functions, CEbUTAl was evaluated by testing it across multiple models and loss functions. The key contributions of this study are as follows.

- Elaboration of the proposed CEbUTAl for correcting model predictions during post-processing, when the model has been trained on imbalanced data.
- Evaluation of the proposed CEbUTAl's model-agnostic characteristics across multiple DL models, including SqueezeNet 1.0, ResNet34, MobileNetV2, DenseNet201, InceptionV3, and ConvNeXt-small.
- Investigation into the proposed CEbUTAl's loss-agnostic behavior in classification tasks, comparing it with cross-entropy loss and focal loss.
- Analysis of the proposed CEbUTAl's task-agnostic performance, applying it to tasks with ICH, OCT, BCD, CTS, and multi-class SLC datasets.
- Comparison with common strategies for addressing class imbalance, such as (i) data augmentation and (ii) focal loss.
- Contrast of the proposed CEbUTAl with state-of-the-art (SOTA) methods for enhancing explainability, including (i) ensemble models, (ii) Monte Carlo Dropout (MCDO), and (iii) ensemble MCDO.

CEbUTAl addresses the key challenge of class imbalance in DL and proposes a methodology based on uncertainty measures to identify and adjust less reliable predictions, thereby improving DL model performance. By correcting outputs using interpretable metrics, such as confidence and entropy, CEbUTAl enhances model explainability and trustworthiness. As a post-processing correction, CEbUTAl improves accuracy without retraining, which is particularly beneficial for imbalanced datasets. Model-agnostic, loss-function-agnostic and task-agnostic correction methods, such as the proposed CEbUTAl, offer broader applicability across various DL architectures and tasks. In medical imaging, correcting model outputs to account for uncertainty is essential for responsible clinical decision-making and AI tool adoption. Output correction also helps address biases in the original model predictions, especially when these models are trained on imbalanced data, ensuring more equitable and reliable results.

2. Related work

This section reviews recent approaches proposed to address class imbalance and uncertainty in model predictions. It is divided into three subsections: Section 2.1 class imbalance, Section 2.2 explainable AI, and Section 2.3 uncertainty quantification.

2.1. Class imbalance

Class imbalance, in which the abnormal or malignant class is significantly underrepresented compared with the normal class, is a pervasive challenge in medical imaging datasets. This imbalance adversely impacts the performance of AI algorithms, often manifesting as an increased rate of false negatives during classification. To mitigate these challenges, several methodological approaches have been proposed. Data augmentation techniques (Iqbal et al., 2021, 2025) are commonly employed to synthetically oversample minority classes, thereby improving their representation in a training set. Additionally, loss function modifications, such as focal loss (Ross and Dollár, 2017; Tran et al., 2019) have been introduced to dynamically adjust the contribution of hard-to-classify examples, effectively reweighting the loss to emphasize minority class instances, and improving model sensitivity.

2.2. Explainable AI

Explainability in MIA has traditionally centered around visual explanations, with saliency maps emerging as a popular tool (Itti et al., 2002). These maps highlight the regions of an image that influence the model's decision-making process, offering valuable insights into the areas of interest for a given prediction (Lundberg and Lee, 2017; Yosinski et al., 2015). However, saliency maps are vulnerable to perturbations (Tomsett et al., 2020) and adversarial attacks (Ghorbani et al., 2019), which distort their interpretations. Furthermore, previous studies (Adebayo et al., 2018) have shown that saliency maps are independent of the training data and trained model, thus making them unreliable.

These limitations have spurred interest in "non-visual" methods for XAI, which offer a more robust, although less intuitive, approach to explainability (Borys et al., 2023). A popular non-visual method to address explainability is the SHapley Additive exPlanations (SHAP), which generates scores for each feature, indicating its impact on the output. SHAP also ensures feature consistency and model stability (Meng et al., 2020). However, it is computationally complex and not applicable to all models (Van den Broeck et al., 2022). Furthermore, although perturbations to the input and approximate explanations contribute to explainability in SHAP, this approach is inconsistent and does not fully capture the behavior of the model's predictions (Slack et al., 2020).

Uncertainty evaluation is a key nonvisual XAI technique for knowing the model's confidence level, which is essential for assessing trustworthiness. A recent review (Huang et al., 2024) of uncertainty evaluation techniques highlighted the following widely used methods: (i) calibration metrics, (ii) Brier score, (iii) predictive entropy, and (iv) predictive variance. Calibration metrics (Huang et al., 2024; Kim et al., 2016) quantify the alignment between a model's predictions and true outcomes by assessing how closely the predicted probabilities correspond to the actual results (Wang et al., 2021). Conversely, the Brier score (Brier, 1950) is a comprehensive measure that evaluates both the calibration and accuracy of the probabilistic predictions. However, these methods require access to the ground truth, which is only available during testing and not in real-world deployment scenarios (Niculescu-Mizil and Caruana, 2005; Jewson, 2004; Assel et al., 2017). Predictive entropy (Malinin and Gales, 2018; Namdari and Li, 2019) measures the uncertainty linked to class probabilities, whereas predictive variance (Cawley et al., 2007) indicates the spread of the predictions. These tools that quantify uncertainty are not used to facilitate prediction correction, even though there are limited studies that have used this for test time adaptation (Ravishankar et al., 2025). Despite their potential, these methods have limited clinical adoption, as DL models rarely integrate uncertainty estimates (Gawlikowski et al., 2023).

2.3. Uncertainty quantification (UQ)

Uncertainty in DL models reflects a lack of confidence in their predictions stemming from various sources. Uncertainty is typically categorized into aleatoric and epistemic uncertainty. Aleatoric uncertainty arises from the noise and variability inherent in the data, whereas epistemic uncertainty results from the model's limitations and insufficient knowledge (Faghani et al., 2023). Among the numerous UQ methods, Monte Carlo Dropout (MCDO) and deep ensembles are popular for correcting uncertainty (Kurz et al., 2022; Lambert et al., 2024). These methods improve predictions by either introducing stochasticity during inference, or leveraging model diversity (Leibig et al., 2017; Hamedani-KarAzmoudehFar et al., 2023; Huang et al., 2024; Asgharnezhad et al., 2022; Gal and Ghahramani, 2016; Zou et al., 2023; Lakshminarayanan et al., 2017; Kimura, 2021).

2.3.1. Monte Carlo dropout

Monte Carlo dropout (MCDO) is an effective implementation of dropout during both training and inference. During inference, multiple forward passes are performed for a single batch, thereby quantifying the uncertainty (Gal and Ghahramani, 2016). MCDO is widely used for UQ in medical images, with applications in segmentation (Zou et al., 2023) and classification, particularly in diabetic retinopathy from fundus images (Leibig et al., 2017) and BCD (Hamedani-KarAzmoudehFar et al., 2023). However, MCDO is an approximation of Bayesian inference in which prior knowledge affects the performance of the model (Lakshminarayanan et al., 2017). Moreover, dropout is an inherent part of the model; therefore, MCDO cannot be model-agnostic and does not operate in a post hoc manner, limiting its generalizability.

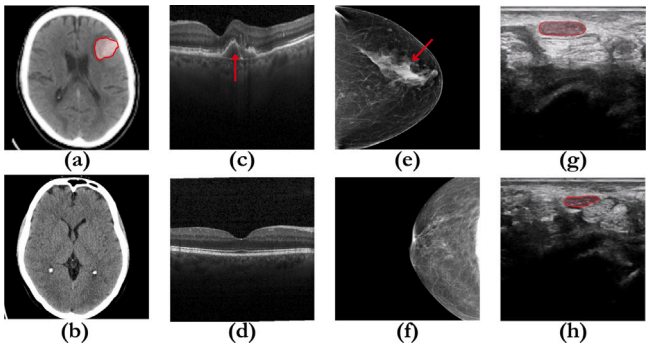

Fig. 2. Examples of ICH, OCT, BCD, and CTS images: (a), (c), (e), and (g) represent the positive/abnormal cases with arrows and masks highlighting regions of interest, while images (b), (d), (f), and (h) depict the normal class.

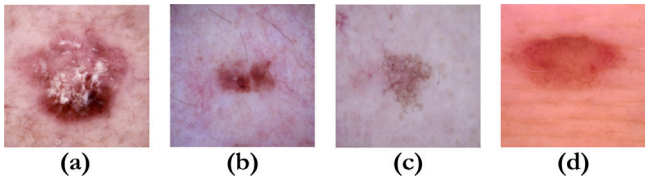

Fig. 3. Examples of dermatoscopic images used in the multi-class SLC task, as part of the DermaMNIST dataset (Yang et al., 2023). While (a) and (b) represent the cancerous lesions and melanocytic nevi, (c) and (d) represent benign and vascular lesions.

2.3.2. Deep ensembles

Deep ensembles leverage the weights of multiple DL models to improve their explainability (Lakshminarayanan et al., 2017). Aggregating predictions from different models improves the reliability of the model predictions. Deep ensembles have demonstrated significant performance in medical image classification tasks such as COVID19 detection (Müller et al., 2022) and Alzheimer's disease (Sreelakshmi et al., 2023). However, training multiple models is both time-consuming and computationally expensive (Ayhan et al., 2020). Additionally, determining the optimal hyperparameters is a heuristic and iterative task (Mohammed and Kora, 2023). Deep ensembles, similar to MCDO, are also not model agnostic and are not applied a posteriori, limiting their applicability.

3. Methodology

This section describes proposed CEbUTAl development in two subsections: Section 3.1 - the "Baseline Model Learning", outlining the steps pursued for model training; and Section 3.2 - the "Proposed CEbUTAl approach", detailing the algorithm's ability to enhance the model's performance.

3.1. Baseline model learning

The baseline model learning phase centers around training different neural network (NN) architectures to generate probabilities for each class.

3.1.1. Datasets

The evaluation of CEbUTAl's task-agnostic capabilities utilized four open-source and one locally curated (CTS) dataset. Figs. 2 and 3 illustrate the visually distinctive features across the classification tasks considered in this study.

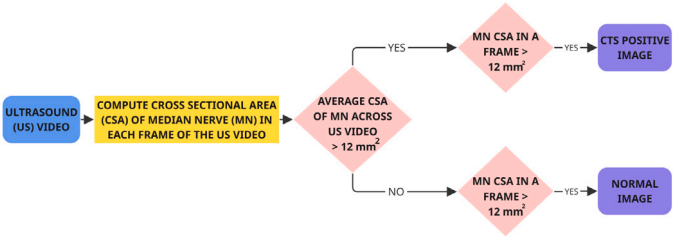

Fig. 4. Data flow for the classification of frames (images) into CTS positive or normal from ultrasound videos of the median nerve at the wrist region, based on cross-sectional area analysis.

Intra-cranial hemorrhage (ICH). Non-contrast computed tomography (CT) images of the brain were obtained from the Radiological Society of North America's (RSNA) ICH detection challenge held in 2019 (Flanders et al., 2020). The dataset consists of six classes: normal and five hemorrhage subtypes - epidural, intraparenchymal, intraventricular, subarachnoid, and subdural. In this study, all five hemorrhage subtypes were considered as ICH positive class. The dataset comprised of $\sim 4,500,000$ images from $\sim 25,000$ patients, with $\sim 4,260,600$ images in the normal class. As part of preprocessing, images were converted to the portable network graphics (PNG) format and uniformly resized to 512×512

Optical coherence tomography (OCT). The UCSD dataset (Kermany et al., 2018) contains retinal cross-section images categorized as: choroidal neovascularization (CNV), diabetic macular edema (DME), drusen, and normal. For binary classification, the CNV and DME were considered as severe class and drusen and normal as mild class. This resulted in a data distribution of $\sim 48,000$ severe cases to $\sim 59,000$ mild cases. Images used were in PNG format and resized to 512×512

Breast cancer detection (BCD). The mammograms for BCD were sourced from the RSNA Challenge 2023 (Carr et al., 2022). Since most patients have images from both the mediolateral oblique and craniocaudal views, the dataset was split at the patient level, resulting in a severe class imbalance with $\sim 53,000$ normal cases and ~ 1200 malignant cases. The mammograms were standardized by converting to PNG and resizing them to 512×512 dimension.

Carpal tunnel syndrome (CTS). To assess model generalizability to real-world clinical data, we used a proprietary (locally curated) CTS dataset (Gujarati et al., 2023) consisting of ultrasound (US) images collected at Aster-CMI Hospital, Bangalore, India, under ethical approval (Approval No. Aster/IEC/049/2020-21, Dated June 27, 2020). Written informed consent was obtained from all participants. The detailed description of the data is provided in Gujarati et al. (2023). US video sweeps of the upper limb (from wrist to elbow) were acquired using a Philips CX50 US machine. Frames from the wrist region were used to classify CTS, where patients were labeled CTS-positive if both their average cross-sectional area (CSA) and per-frame CSA of the Median Nerve (MN) exceeded 12 mm^2 and normal if both were below 12 mm^2 (as shown in Fig. 4). The dataset considered for this task, comprised 13 CTS positive patients, contributing to 500 frames, and 73 normal patients, accounting for 2803 frames.

Multi-class skin lesion classification (SLC). For multi-class classification, dermatoscopic images were obtained from the DermaMNIST dataset (Yang et al., 2023), a subset of the MedMNIST2D collection of medical images. The dataset comprises seven distinct skin lesion classes, which can be broadly grouped into four major diagnostic categories: (i) cancerous lesions - actinic keratoses and intraepithelial carcinoma (akiec), basal cell carcinoma (bcc), and melanoma (mel), (ii) melanocytic nevi (nv), (iii) benign lesions - dermatofibroma (df) and benign keratosis-like lesions (bkl), and (iv) vascular lesions (vasc). Class imbalance was prominent with 983 images for cancerous lesions, 1113 for nv, 1214 for benign lesions, and 6705 for vascular lesions.

Table 1
Number of images considered for training, validation, and testing (split at patient level) for the tasks (ICH, OCT, BCD, CTS, and multi-class SLC).

Task		Training	Validation	Testing
ICH	ICH+	2500	500	800
	NORMAL	47 500	500	800
OCT	SEVERE	10 000	1000	2000
	MILD	40 000	1000	500
BCD	MALIGNANT	881	76	200
	NORMAL	40 473	3369	25
CTS	CTS+	260	120	120
	NORMAL	2563	120	120
Multi-class SLC	CANCEROUS	400	250	300
	NEVI	400	250	300
	BENIGN	400	250	300
	VASCULAR	6000	250	300

To ensure robust generalization, datasets were split patient-wise into training, validation, and test subsets, thereby preventing data leakage. Since the ICH, OCT, BCD, and CTS datasets were grayscale, the first convolutional layer of the model was adapted accordingly; no such change was needed for the multi-class SLC dataset, which contained RGB images. Each training dataset retained its natural class imbalance to reflect real-world conditions. In contrast to the training datasets, the validation datasets were curated to be perfectly balanced to ensure the selection of the best-performing model across all classes. The data distribution used for the training (D_{train}), validation (D_{val}), and testing (D_{test}) datasets for each task has been listed in Table 1.

As is common in real-world medical datasets, the healthy or normal class constitutes the majority and is represented by C_{maj}, whereas the abnormal or severe cases, often rarer and clinically significant, are grouped under C_{min}. The symbols x, y, and n denote images, labels, and the number of classes, respectively.

3.1.2. Experimentation

The training set of images was subjected to random transformations i.e. rotation, flipping, zooming, and the addition of Gaussian noise to result in a robust trained model. A consistent batch size of eight, with two workers, was used throughout the process. The model was trained using cross-entropy loss with the "Adam" optimizer with a learning rate of 1e-4 for 100 epochs. The experiments were conducted on a system with an NVIDIA RTX A5000 GPU (Compute Capability 8.6, 8192 CUDA cores, 24 GB GDDR6 VRAM). Validation on D_{val} is performed after each epoch, and the model with the best validation loss is saved. The saved model (M) is applied to D_{test}, and the softmax function is performed to obtain the probabilities of each class, providing the predictions. For each image x and $i \in [0, n]$, the softmax score (p_i) for the ith class is obtained from the predicted raw scores ($z_i = M(x) : i \in [0, n]$) as follows:

$$p_i = \frac{e^{z_i}}{\sum_{i=1}^{n} e^{z_i}} \tag{1}$$

3.2. Proposed CEbUTAl approach

The baseline model M, is typically biased towards C_{maj} in instances of class imbalance, resulting in inaccurate classifications. Furthermore, model bias produces a higher number of false negatives, which is a major challenge in medical diagnoses (Petticrew et al., 2001). Prior studies have utilized the entropy of softmax probabilities to evaluate uncertainty (Hamedani-KarAzmoudehFar et al., 2023; Asgharnezhad et al., 2022). Similarly, this study computes entropy to measure the uncertainty in the predictions. Once entropy is computed, the uncertainty estimates are obtained from the predicted entropy. Subsequently, the confidence score was computed from the probabilities. Finally, CEbUTAl utilizes the computed confidence scores and uncertainty estimates to refine the model's predictions during inference time.

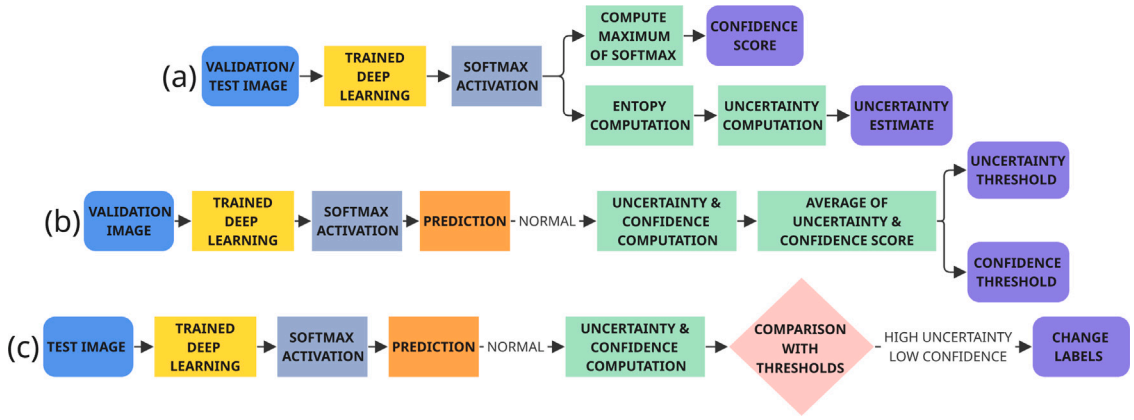

Fig. 5. Design of proposed CEbUTAl. (a) Uncertainty estimate and confidence score computation from softmax output. (b) Computation of threshold from D_{val}. (c) Correction by CEbUTAl during inference (testing).

3.2.1. Predictive entropy

The level of uncertainty in each image's prediction by the model is quantified by the predictive entropy, denoted as H (Namdari and Li, 2019; Ovadia et al., 2019), and is computed as follows:

$$H = -\sum_{i=1}^{n}(p_i \times log(p_i)) \tag{2}$$

In Eq. (2) log denotes logarithm to the base 2.

3.2.2. Uncertainty estimation

A higher H indicates greater uncertainty, whereas a lower H indicates M is confident about its prediction (Malinin and Gales, 2018). For binary classification, maximum H occurs when the classes are equiprobable ($p_i = 0.5$), and minimum H when the model is absolutely sure of its prediction ($p_0 = 0$ *and* $p_1 = 1$ or vice-versa). A min–max normalization on H was performed using the maximum and minimum entropy to deduce the uncertainty measure (U) for each image, as illustrated below:

$$U = \frac{H - H_{min}}{H_{max} - H_{min}} \equiv \frac{\sum_{i=1}^{n}(p_i \cdot log(p_i))}{log(0.5)} \tag{3}$$

3.2.3. Confidence score

The confidence score (CS) for each sample is a measure of the certainty of the model in its prediction, which can be given by

$$CS = \max_{1 \le i \le n} p_i \tag{4}$$

Algorithm 1 Major steps of proposed CEbUTAl

Require: trained model(M), datasets D_{val}, D_{test} : N_{val}, N_{test} lengths, n classes : majority class (C_{maj}), minority class (C_{min})

$\sigma_{unc} = 0, \sigma_{con} = 0, x_{count} = 0$
for $j = 1$ *to* N_{val} **do**
  if $M(x_j) = C_{maj}$ **do**
    $H_j = -\sum_{i=1}^{n}(M(j)_i \cdot ln(M(j)_i))$
    $U_j = H_j \times 100$
    $\sigma_{unc} += U_j$
    $C_j = \arg\max M(j)$
    $\sigma_{con} += C_j$
    $x_{count} += 1$
  end if
end for
$\tau_{unc} = \frac{\sigma_{unc}}{x_{count}}, \tau_{con} = \frac{\sigma_{con}}{x_{count}}$
for $k = 1$ *to* N_T **do**
  $H_k = -\sum_{i=1}^{n}(M(k)_i \cdot ln(M(k)_i))$
  $U_k = H_k \times 100$
  $C_k = \arg\max M(k)$
  if $U_k > \tau_{unc}$, $C_k < \tau_{con}$ *and* $M(k) = C_{max}$ **do**
    $M(k) = C_{min}$
  end if
end for

3.2.4. Thresholding

In this study, D_{val} did not have a class imbalance, and the computed uncertainty and confidence metrics were used to establish the thresholds. For each $x_j \in D_{val} : j \in [0, N]$ where N is the number of samples in D_{val} and $\hat{y}_j$ denotes the prediction of $M(x_j)$, the average uncertainty μ_{unc} and average confidence μ_{con} was computed as follows:

$$\mu_{con} = \frac{1}{N}\sum_{x_j \in D_{val}} CS_j \tag{5}$$

$$\mu_{unc} = \frac{1}{N}\sum_{x_j \in D_{val}} U_j \tag{6}$$

When $N \equiv N_V$, where N_V represents the number of samples with $\hat{y}_j = C_{maj}$, the uncertainty threshold (τ_{unc}) and confidence threshold (τ_{con}) are equal to average uncertainty μ_{unc} and confidence μ_{con} respectively. The average uncertainty and confidence measures from the validation dataset, μ_{unc} and μ_{con}, establish concrete independence of the test data thus eliminating the necessity of heuristics and data leakage (Dolezal et al., 2022; Syrykh et al., 2020; Senousy et al., 2021). Finally, predictions with high uncertainty and low confidence were converted into C_{min}. A schematic representation of CEbUTAl was presented in Fig. 5. Subsequently, the steps of CEbUTAl are described in Algorithm 1. In the multi-class classification setting, decision thresholds were derived using the validation set following same procedure as binary classification task. Specifically, the threshold for uncertainty-based correction was computed across all correctly classified validation samples that did not belong to C_{maj}. During inference, if the model prediction is majority class (C_{maj}) then these predictions with high uncertainty and low confidence were changed to the class with the second-highest predicted probability.

3.2.5. Evaluation metrics

The performance of models was evaluated using the following metrics: precision (P), sensitivity (S_e) or true positive rate, specificity (S_p) or true negative rate, F1-score (F_1), and accuracy (Y) in binary classification based on the counts of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). Fig. 6 pictorially represents CEbUTAl's ability to correct the model's predictions during inference time. CEbUTAl decreases FN by converting it to TP. However, because of the bias in the dataset, a higher FP was observed, which requires additional effort from the clinician to correct it manually.

4. Results

4.1. Evaluating agnostics of CEbUTAl

4.1.1. Model agnostic evaluation

The evaluation metrics were computed for six trained models: (i) SqueezeNet 1.0, (ii) ResNet34, (iii) DenseNet201, (iv) MobileNetV2,

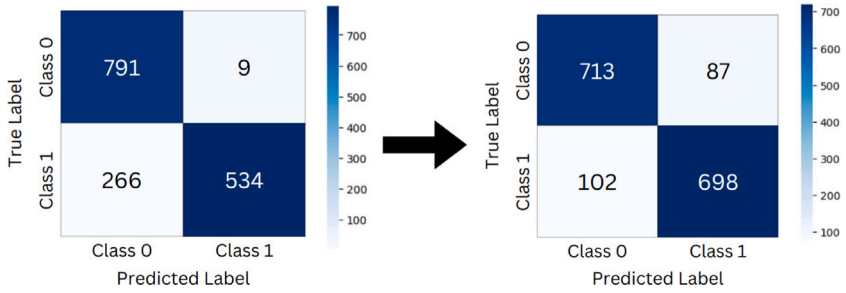

Fig. 6. (a) Confusion matrix obtained from the InceptionV3 model for the ICH classification task (Section 3.1). (b) Confusion matrix obtained after implementing CEbUTAl (Section 3.2).

Table 2
Comparison of deep learning (DL) models utilized in this study in terms of parameters and model size.

Model	Parameters (in million)	Model size (in MB)
SqueezeNet 1.0	0.73	2.79
ResNet34	21.28	81.31
DenseNet201	18.09	70.34
MobileNetV2	2.23	8.74
InceptionV3	21.79	83.45
ConvNeXt-small	49.45	188.77

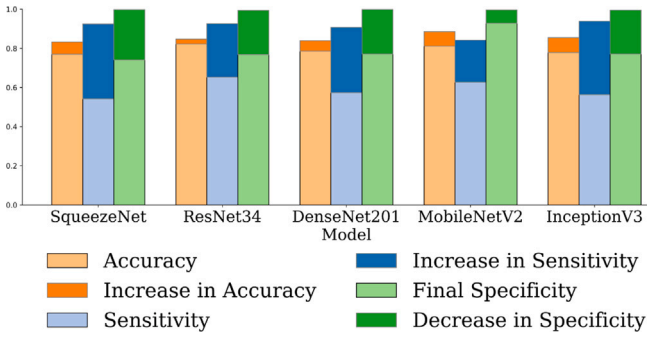

Fig. 7. Y, S_e, and S_p for models using baseline and after CEbUTAl. Note that the loss function used in this experiment is focal loss, and the task performed is ICH classification.

(v) InceptionV3, and (vi) ConvNeXt-small on D_{test}. The architectural details, including the number of parameters and storage requirements for each model, are summarized in Table 2. ConvNeXt (Liu et al., 2022) was utilized in this study to showcase CEbUTAl's compatibility with hybrid models, in addition to CNN-based models. All trained models were processed with CEbUTAl during the inference time for evaluation. The computed P, S_e, S_p, F_1, and Y for the baseline approach and the CEbUTAl approach have been presented in Table 3. The baseline performance is poor in ConvNeXt, due to a higher number of parameters, optimization dynamics, and loss minimization favoring C_{maj}. This observation aligns with recent studies suggesting that transformer models may overfit to frequency-based priors in class imbalance (Kunstner et al., 2024; Xu et al., 2023; Li et al., 2023). A detailed comparison of model predictions using the baseline approach and after applying CEbUTAl for the ICH task has been presented in Table 4, highlighting shifts in classification accuracy, estimated miss rate, and overcall rate across ICH and normal cases. For the loss- and task-agnostic studies (BCD and OCT), SqueezeNet 1.0 and InceptionV3 models were employed, while ResNet34 was utilized for the CTS and multi-class SLC tasks.

4.1.2. Loss agnostic evaluation

Cross entropy and focal loss (Ross and Dollár, 2017) are common loss functions for classification tasks (Tran et al., 2019). The parameters

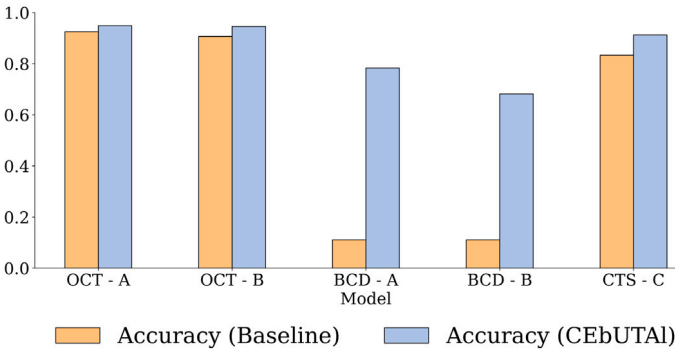

Fig. 8. Accuracy of models SqueezeNet 1.0 (A), InceptionV3 (B), and ResNet34 (C) for OCT analysis, BCD, and CTS detection using baseline and after applying CEbUTAl. Note that Tables 3 and 5 provide the results for ICH classification and multi-class SLC.

α, γ, and *reduction* for focal loss are set to default values of 0.25, 2, and "*mean*", respectively. The evaluation metrics computed using baseline and after CEbUTAl for all models trained with cross-entropy loss and focal loss were tabulated in Table 3 and Fig. 7 respectively.

4.1.3. Task agnostic evaluation

In addition to ICH, the proposed approach was evaluated on binary classification tasks—including BCD, OCT, and CTS as well as on multi-class classification using dermatoscopic images from the multi-class SLC dataset. Notably, CTS is a clinical and proprietary dataset, included to demonstrate CEbUTAl's performance on real-world clinical data. While Fig. 8 plots the accuracy of the model using the baseline approach and after the application of CEbUTAl for OCT analysis, BCD, and CTS detection, Table 5 illustrates the improvement in multi-class SLC.

4.2. Comparison with SOTA data imbalance mitigation

State-of-the-art (SOTA) methods for class imbalance mitigation include data augmentation (Yang et al., 2022) and focal loss (Lin et al., 2017). The data augmentation strategy involves implementing rotations, flipping, brightness adjustments, contrast enhancements, hue and saturation modifications, Gaussian blur, and affine transformations. CEbUTAl achieved SOTA performance in comparison with models trained with data augmentation and focal loss, and the results have been presented in Fig. 9.

4.3. Comparison with SOTA UQ methods

CEbUTAl was evaluated against several state-of-the-art (SOTA) methods, including MCDO, deep ensembles of trained models, and ensemble MCDO. As shown in Fig. 10, the proposed approach demonstrates superior performance compared to these methods in terms of UQ and correction. CEbUTAl consistently demonstrated superior results

Table 3
Model agnostic evaluation metrics for the ICH task using baseline and after application of CEbUTAl. Dataset details are provided in Table 1. While the ConvNeXt model was trained on the same training set, the validation set was also imbalanced with ICH positive containing 100 and normal containing 900 sample images, to showcase CEbUTAl's generalizability. The metric precision presented here is an average of both ICH positive and normal classes.

Model	Baseline					CEbUTAl				
	Precision	Sensitivity	Specificity	F1-Score	Accuracy	Precision	Sensitivity	Specificity	F1-Score	Accuracy
SqueezeNet 1.0	0.8586	0.6388	0.9900	0.7748	0.8144	**0.8658**	**0.8762**	0.8550	**0.8670**	**0.8656 (5.1%)**
ResNet34	0.8608	0.6500	0.9888	0.7825	0.8194	**0.8756**	**0.8800**	0.8712	**0.8762**	**0.8756 (5.6%)**
DenseNet201	0.8605	0.6887	0.9738	0.8032	0.8313	**0.8662**	**0.8862**	0.8450	**0.8683**	**0.8656 (3.4%)**
MobileNetV2	0.8748	0.6875	0.9925	0.8112	0.8400	**0.8872**	**0.9000**	0.8738	**0.8883**	**0.8869 (4.7%)**
InceptionV3	0.8658	0.6675	0.9888	0.7952	0.8281	**0.8820**	**0.8725**	0.8912	**0.8808**	**0.8819 (5.4%)**
ConvNeXt-small	0.2500	0.0000	1.0000	0.0000	0.5000	**0.6962**	**0.4787**	0.8575	**0.5906**	**0.6681 (16.8%)**

Table 4
Quantitative comparison of ResNet34 predictions using baseline approach and after applying CEbUTAl for the ICH task, highlighting significant reductions in miss rate and improvements in correct ICH classification.

Criteria	Baseline	CEbUTAl
Correctly classified as ICH	520 of 800 (65%)	704 of 800 (88%)
Both (Baseline and CEbUTAl)	520 of 520 (100%)	520 of 704 (73.9%)
Exclusively (Baseline and CEbUTAl)	0 of 520	184 of 704 (26.1%)
Correctly classified as NORMAL	791 of 800 (98.9%)	697 of 800 (87.1%)
Both (Baseline and CEbUTAl)	697 of 791 (88.2%)	697 of 697 (100%)
Exclusively (Baseline and CEbUTAl)	94 of 791 (11.8%)	0 of 697
Estimated miss rate	280 of 800 (35%)	96 of 800 (12%)
Estimated overcall rate	9 of 800 (1.1%)	103 of 800 (12.8%)

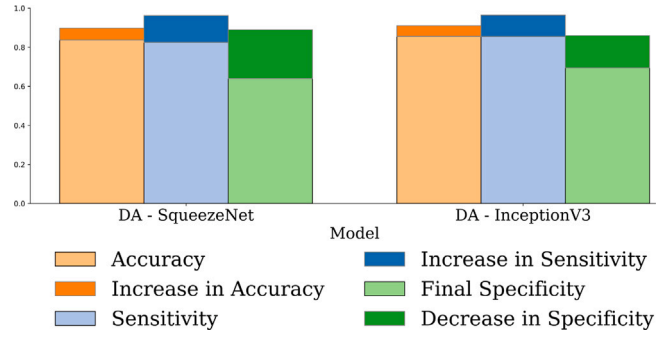

Fig. 9. Y, S_e, and S_p for SqueezeNet 1.0 and InceptionV3 models using data augmentation (DA) using baseline and after applying CEbUTAl for the task of ICH classification.

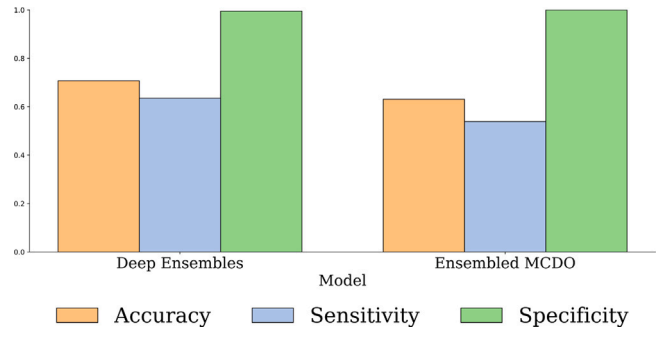

Fig. 10. Y, S_e, and S_p using UQ methods Deep Ensembles (left) and Ensembled MCDO (right). Deep Ensembles utilized the five models discussed in this work (listed also in Table 3). Ensembled MCDO was computed using the ResNet34 model. Note that the task here was ICH classification.

against several iterations of MCDO, implemented with dropout rates of 0.01, 0.05, 0.1, 0.2, and 0.3. The ensembled MCDO method averaged the predictions from these models, whereas the deep ensembles involved five models that were trained a priori.

4.4. Comparison with balanced datasets

CEbUTAl was also evaluated with an InceptionV3 model trained for a balanced subset of the ICH dataset, containing 25,000 images per class. The validation and test splits remained the same as illustrated in Table 1. The results of this experiment were tabulated in Table 6.

4.5. Time complexity analysis

While CEbUTAl addresses explainability and improves classification performance, it is also important to highlight its time efficiency and minimal computational overhead. Table 7 provides the run time taken for testing and the additional time introduced by each step in CEbUTAl for the CTS detection task.

5. Discussion

Class imbalance in datasets biases the model predictions and favors the majority class, thereby reducing trustworthiness. Additionally, deep learning (DL) models are highly complex, and associated with a "black-box" analogy. Prior approaches have used explainability to understand a model's decision-making. This study proposed "CEbUTAl": a Confidence and Entropy-based Uncertainty Thresholding Algorithm, implemented at inference time, addressing the issue of class imbalance and improves explainability. Detailed evaluation including the generalizability of CEbUTAl with state-of-the-art (SOTA) methods was conducted in this study for class imbalance mitigation and uncertainty quantification (UQ). This study included investigations of the generalizability by observing CEbUTAl's performance across (i) different models: SqueezeNet 1.0, ResNet34, DenseNet201, MobileNetV2, InceptionV3, and ConvNeXt-small; (ii) different loss functions: Cross-Entropy and Focal loss; and (iii) different tasks: intracranial hemorrhage (ICH) detection, breast cancer detection (BCD), optical coherence tomography (OCT) analysis, carpal tunnel syndrome (CTS), and skin lesion classification (SLC). From Table 3, one can observe that when tested with six different DL models for classification, CEbUTAl improves the overall accuracy as high as $\sim$ 16% and increases the sensitivity and F1-score. Fig. 7 presents the results using an alternate loss function, demonstrating similar improvements in accuracy and sensitivity. Furthermore, the evaluation of CEbUTAl's performance across different tasks improved accuracy, as illustrated in Fig. 8 and Table 5. Overall, the accuracies achieved using CEbUTAl for ICH detection, OCT analysis, BCD, CTS detection, and multi-class SLC increased to 84%, 94%, 73%, 91%, and 63% respectively. The results indicate that the developed approach is agnostic to the model, loss function, and task and is therefore generalizable to a wide range of classification tasks.

Subsequently, the CEbUTAl performance was compared with SOTA methods for class imbalance mitigation such as focal loss and data

Table 5
Evaluation metrics for the multi-class SLC task using ResNet34 baseline and after applying CEbUTAl.

Class	Baseline				CEbUTAl			
	Sensitivity	Specificity	F1-Score	Accuracy	Sensitivity	Specificity	F1-Score	Accuracy
Cancerous Lesions	0.5833	0.9455	0.6679	0.8550	**0.6967**	0.9122	**0.7109**	**0.8583 (0.3%)**
Melanocytic nevi	0.1633	0.9644	0.2572	0.7642	**0.5433**	0.8889	**0.5778**	**0.8017 (3.8%)**
Benign Lesions	0.3767	0.8678	0.4248	0.7450	**0.6167**	0.7767	**0.5395**	0.7367
Vascular Lesions	0.9467	0.5789	0.5898	0.6708	0.6533	**0.9267**	**0.6978**	**0.8583 (18.8%)**

Table 6
Evaluation metrics for InceptionV3 trained on a balanced ICH dataset with and without application of proposed CEbUTAl.

Model	Sensitivity	Specificity	F1-Score	Accuracy
InceptionV3 Baseline	0.9537	0.9212	0.9385	0.9375
InceptionV3 with CEbUTAl	0.9975	0.7650	0.8936	0.8813

Table 7
Runtime analysis for each step of proposed CEbUTAl applied for the ResNet34 model in CTS detection.

Task	Time
Testing (including test-data loading and classification report generation)	6.38 s
Testing on validation set (for obtaining predictions for thresholds)	2.77 s
Entropy computation for test and validation set (the right predictions from validation)	8.56 ms
Uncertainty computation	4.85 ms
Confidence computation	2.32 ms
Computing threshold	1.65 ms
Changing model prediction based on threshold	2.72 ms

augmentation. The results of the comparative study are shown in Fig. 9, illustrating that CEbUTAl outperforms other techniques for any model. This plot reveals that CEbUTAl outperforms focal loss by $\sim 5\%$. While the model showed improvements with data augmentation, it was limited by data variability, which is challenging to address in class-imbalance scenarios. This study included an investigation of comparing CEbUTAl with the widely used UQ methods: Monte Carlo Dropout (MCDO), deep ensembles, and ensembled MCDO. The results of the comparative study of the accuracy, sensitivity, and specificity are presented in the bar chart in Fig. 10. The results demonstrate that CEbUTAl provides increased accuracy ($\sim 25\%$ compared with ensemble MCDO) and sensitivity compared with all other UQ approaches. Although these methods improve robustness and uncertainty estimation, they do not inherently correct the model bias, leading to suboptimal performance in cases of class imbalance. By contrast, CEbUTAl specifically addresses and corrects these biases, enabling it to achieve accurate predictions and explainability. In the absence of a class imbalance, an InceptionV3 model trained on 50,000 images produced an accuracy of 94%. The application of CEbUTAl minimally reduces the accuracy to 88%. However, a notable observation is the steep decline in false negatives, which is emphasized by the increase in sensitivity tabulated in Table 6. More importantly, the detailed analysis presented in Table 4 for the ICH task highlights that the miss rate for ICH detection was improved by 23% compared to the baseline. Minimizing miss rates in AI models is essential for advancing their reliability and clinical utility. Lower miss rates enhance patient outcomes by enabling earlier and more reliable detection of diseases. Reducing miss rates also supports healthcare efficiency by decreasing the need for repeat tests and accelerating clinical decision-making. Although the proposed approach results in an 11.8% increase in overcall rate, this cautious strategy is appropriate in clinical settings where AI serves as an assistive and recommendatory tool, ensuring fewer true positive cases are overlooked. Such a balance between sensitivity and specificity aligns with current clinical AI implementations that prioritize patient safety by reducing missed diagnoses while managing acceptable false positive rates (Krupinski, 2000).

This study included experiments to evaluate CEbUTAl's performance when the validation set was imbalanced. While the train and test split remained as described in Table 1, the validation set contained 900 images for the normal class and 100 ICH positive images, reflecting real world imbalance. While the baseline ResNet34 model gave an accuracy of 79.31%, application of CEbUTAl increased it to 87.56% (balanced validation dataset results are presented in Table 3), signifying that CEbUTAl performs well in scenarios of imbalanced validation set. Note that ConvNeXt model result with imbalanced validation dataset was presented in the last row of Table 3. Studies were also conducted comparing CEbUTAl to other ways of obtaining thresholds such as the usage of isotonic regression, followed by thresholding from the precision–recall curve. However, this approach did not yield desirable results, leading to a decline in accuracy. Additionally, the runtime analysis of CEbUTAl, as outlined in Table 7, demonstrates minimal overhead—adding approximately 5 ms per image compared to the baseline.

Uncertainty-based corrections such as the proposed method are valuable for increasing the utility of DL models in medical image classification. CEbUTAl was demonstrated to address class imbalance, improve sensitivity, enhance explainability, and handle data variability. The proposed method adapts to less represented class, balances precision and recall, and improves the overall accuracy, especially in imbalanced datasets. Its model-agnostic nature allows for versatile applications across different architectures, complementing other techniques and has the capability of accounting for both aleatoric and epistemic uncertainty. Note that the proposed CEbUTAl is a post-processing method, and approaches such as active learning that can perform uncertainty-based corrections can also provide a comprehensive approach for improving model performance and reliability in medical imaging applications. In addition, the formulation of entropy-based uncertainty as a loss function embedded in the training process helps in the development of uncertainty-aware frameworks that are more trustworthy and reliable in the clinical setting.

6. Conclusion

This study has introduced "CEbUTAl": a Confidence and Entropy-based Uncertainty Thresholding Algorithm designed to adapt the model predictions by incorporating uncertainty and confidence. This algorithm provides a robust method for correcting model predictions at inference time, simplifying implementation, and minimizing reliance on complex loss functions and hyperparameters. Experiments on clinical problems and results demonstrate (i) the generalizability of CEbUTAl, (ii) performance improvements, and (iii) the explainability of AI systems, thereby increasing the trustworthiness of model predictions. Integration of entropy-based uncertainty with confidence contributes to reducing false negatives and improves the overall accuracy of model predictions. Furthermore, compared to state-of-the-art (SOTA) methods used for explainability and class imbalance mitigation, CEbUTAl demonstrated superior performance. The model, loss function, and task-agnostic nature make CEbUTAl versatile and easily adaptable, making it a valuable tool for a wide range of AI applications, including image classification, segmentation, natural language processing, and reinforcement learning. The code utilized to generate the results in this study has been provided here: https://github.com/Joel-Jeffrey/CEbUTAl.

CRediT authorship contribution statement

Joel Jeffrey: Writing – original draft, Software, Methodology, Investigation. **Ashwin RajKumar:** Methodology, Writing – review & editing, Validation, Investigation, Formal analysis. **Sudhanshu Pandey:** Data curation, Resources, Visualization. **Lokesh Bathala:** Data curation, Project administration, Resources, Supervision, Validation. **Phaneendra K. Yalavarthy:** Writing – review & editing, Supervision, Resources, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the ARG grant# ARG01-0524-230330 from the Qatar National Research Fund (a member of the Qatar Foundation).

Data availability

ICH dataset is publicly available at https://www.kaggle.com/competitions/rsna-intracranial-hemorrhage-detection (Flanders et al., 2020). OCT dataset is publicly available at https://www.kaggle.com/datasets/paultimothymooney/kermany2018 (Kermany et al., 2018). BCD dataset is publicly available at https://www.kaggle.com/competitions/rsna-breast-cancer-detection (Carr et al., 2022). Multi-class SLC dataset is publicly available as dermamnist.npz at https://zenodo.org/records/10519652 (Yang et al., 2023).

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., et al., 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Inf. Fusion 76, 243–297.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B., 2018. Sanity checks for saliency maps. Adv. Neural Inf. Process. Syst. 31.

Alexander, R., Waite, S., Bruno, M.A., Krupinski, E.A., Berlin, L., Macknik, S., Martinez-Conde, S., 2022. Mandating limits on workload, duty, and speed in radiology. Radiology 304 (2), 274–282.

Angelopoulos, A.N., Bates, S., et al., 2023. Conformal prediction: A gentle introduction. Found. Trends textregistered Mach. Learn. 16 (4), 494–591.

Araújo, T., Aresta, G., Mendonça, L., Penas, S., Maia, C., Carneiro, Â., Mendonça, A.M., Campilho, A., 2020. Dr| GRADUATE: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. Med. Image Anal. 63, 101715.

Asgharnezhad, H., Shamsi, A., Alizadehsani, R., Khosravi, A., Nahavandi, S., Sani, Z.A., Srinivasan, D., Islam, S.M.S., 2022. Objective evaluation of deep uncertainty predictions for covid-19 detection. Sci. Rep. 12 (1), 815.

Assel, M., Sjoberg, D.D., Vickers, A.J., 2017. The brier score does not evaluate the clinical utility of diagnostic tests or prediction models. Diagn. Progn. Res. 1, 1–7.

Ayhan, M.S., Kühlewein, L., Aliyeva, G., Inhoffen, W., Ziemssen, F., Berens, P., 2020. Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection. Med. Image Anal. 64, 101724.

Balasubramaniam, S., Velmurugan, Y., Jaganathan, D., Dhanasekaran, S., 2023. A modified LeNet CNN for breast cancer diagnosis in ultrasound images. Diagnostics 13 (17), 2746.

Borys, K., Schmitt, Y.A., Nauta, M., Seifert, C., Krämer, N., Friedrich, C.M., Nensa, F., 2023. Explainable AI in medical imaging: An overview for clinical practitioners–beyond saliency-based XAI approaches. Eur. J. Radiol. 162, 110786.

Brier, G.W., 1950. Verification of forecasts expressed in terms of probability. Mon. Weather Rev. 78 (1), 1–3.

Van den Broeck, G., Lykov, A., Schleich, M., Suciu, D., 2022. On the tractability of SHAP explanations. J. Artificial Intelligence Res. 74, 851–886.

Carr, C., FelipeKitamura, MD, PhD, Partridge, G., inversion, Kalpathy-Cramer, J., Mongan, J., Andriole, K., Lavender, Vazirabad, M., Riopel, M., Ball, R., Dane, S., Chen, Y., 2022. RSNA screening mammography breast cancer detection. Kaggle, https://kaggle.com/competitions/rsna-breast-cancer-detection.

Cawley, G.C., Janacek, G.J., Haylock, M.R., Dorling, S.R., 2007. Predictive uncertainty in environmental modelling. Neural Netw. 20 (4), 537–549.

Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N., 2018. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision. WACV, IEEE, pp. 839–847.

Dolezal, J.M., Srisuwananukorn, A., Karpeyev, D., Ramesh, S., Kochanny, S., Cody, B., Mansfield, A.S., Rakshit, S., Bansal, R., Bois, M.C., et al., 2022. Uncertainty-informed deep learning models enable high-confidence predictions for digital histopathology. Nat. Commun. 13 (1), 6572.

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., et al., 2023. Explainable AI (XAI): Core ideas, techniques, and solutions. ACM Comput. Surv. 55 (9), 1–33.

Faghani, S., Moassefi, M., Rouzrokh, P., Khosravi, B., Baffour, F.I., Ringler, M.D., Erickson, B.J., 2023. Quantifying uncertainty in deep learning of radiologic images. Radiology 308 (2), e222217.

Flanders, A.E., Prevedello, L.M., Shih, G., Halabi, S.S., Kalpathy-Cramer, J., Ball, R., Mongan, J.T., Stein, A., Kitamura, F.C., Lungren, M.P., et al., 2020. Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge. Radiol.: Artif. Intell. 2 (3), e190211.

Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning. PMLR, pp. 1050–1059.

Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al., 2023. A survey of uncertainty in deep neural networks. Artif. Intell. Rev. 56 (Suppl 1), 1513–1589.

Gerlings, J., Shollo, A., Constantiou, I., 2021. Reviewing the need for explainable artificial intelligence (xAI). In: Proceedings of the 54th Hawaii International Conference on System Sciences. HICSS, pp. 1284–1293.

Ghorbani, A., Abid, A., Zou, J., 2019. Interpretation of neural networks is fragile. In: Proceedings of the AAAI Conference on Artificial Intelligence. 33, (01), pp. 3681–3688.

Gujarati, K.R., Bathala, L., Venkatesh, V., Mathew, R.S., Yalavarthy, P.K., 2023. Transformer-based automated segmentation of the median nerve in ultrasound videos of wrist-to-elbow region. IEEE Trans. Ultrason. Ferroelectr. Freq. Control 71 (1), 56–69.

Hamedani-KarAzmoudehFar, F., Tavakkoli-Moghaddam, R., Tajally, A.R., Aria, S.S., 2023. Breast cancer classification by a new approach to assessing deep neural network-based uncertainty quantification methods. Biomed. Signal Process. Control. 79, 104057.

Higham, C.F., Higham, D.J., 2019. Deep learning: An introduction for applied mathematicians. Siam Rev. 61 (4), 860–891.

Huang, L., Ruan, S., Xing, Y., Feng, M., 2024. A review of uncertainty quantification in medical image analysis: Probabilistic and non-probabilistic methods. Med. Image Anal. 103223.

Iqbal, I., Ullah, I., Peng, T., Wang, W., Ma, N., 2025. An end-to-end deep convolutional neural network-based data-driven fusion framework for identification of human induced pluripotent stem cell-derived endothelial cells in photomicrographs. Eng. Appl. Artif. Intell. 139, 109573.

Iqbal, I., Younus, M., Walayat, K., Kakar, M.U., Ma, J., 2021. Automated multi-class classification of skin lesions through deep convolutional neural network with dermoscopic images. Comput. Med. Imaging Graph. 88, 101843.

Itti, L., Koch, C., Niebur, E., 2002. A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. 20 (11), 1254–1259.

Jahangir, Z., Saeed, F., Shiwlani, A., Shiwlani, S., Umar, M., 2024. Applications of ML and DL algorithms in the prediction, diagnosis, and prognosis of alzheimer's disease. Am. J. Biomed. Sci. Res. 22 (6), 779–786.

Jewson, S., 2004. The problem with the brier score. ArXiv Preprint Physics/0401046.

Kalemaki, M.S., Karantanas, A.H., Exarchos, D., Detorakis, E.T., Zoras, O., Marias, K., Millo, C., Bagci, U., Pallikaris, I., Stratis, A., et al., 2020. PET/CT and PET/MRI in ophthalmic oncology. Int. J. Oncol. 56 (2), 417–429.

Kendall, A., Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for computer vision? Adv. Neural Inf. Process. Syst. 30.

Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al., 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 172 (5), 1122–1131.

Kim, B., Khanna, R., Koyejo, O.O., 2016. Examples are not enough, learn to criticize! criticism for interpretability. Adv. Neural Inf. Process. Syst. 29.

Kim, M., Yun, J., Cho, Y., Shin, K., Jang, R., Bae, H.-j., Kim, N., 2019. Deep learning in medical imaging. Neurospine 16 (4), 657.

Kimura, M., 2021. Understanding test-time augmentation. In: International Conference on Neural Information Processing. Springer, pp. 558–569.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25.

Krupinski, E.A., 2000. The importance of perception research in medical imaging. Radiat. Med. 18 (6), 329–334.

Kunstner, F., Milligan, A., Yadav, R., Schmidt, M., Bietti, A., 2024. Heavy-tailed class imbalance and why adam outperforms gradient descent on language models. Adv. Neural Inf. Process. Syst. 37, 30106–30148.

Kurz, A., Hauser, K., Mehrtens, H.A., Krieghoff-Henning, E., Hekler, A., Kather, J.N., Fröhling, S., von Kalle, C., Brinker, T.J., et al., 2022. Uncertainty estimation in medical image classification: systematic review. JMIR Med. Informatics 10 (8), e36427.

Lahoti, P., Gummadi, K.P., Weikum, G., 2021. Detecting and mitigating test-time failure risks via model-agnostic uncertainty learning. In: 2021 IEEE International Conference on Data Mining. ICDM, IEEE, pp. 1174–1179.

Lahoti, P., Gummadi, K., Weikum, G., 2023. Responsible model deployment via model-agnostic uncertainty learning. Mach. Learn. 112 (3), 939–970.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. Adv. Neural Inf. Process. Syst. 30.

Lambert, B., Forbes, F., Doyle, S., Dehaene, H., Dojat, M., 2024. Trustworthy clinical AI solutions: a unified review of uncertainty quantification in deep learning models for medical image analysis. Artif. Intell. Med. 102830.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86 (11), 2278–2324.

Lee, J.-G., Jun, S., Cho, Y.-W., Lee, H., Kim, G.B., Seo, J.B., Kim, N., 2017. Deep learning in medical imaging: general overview. Korean J. Radiol. 18 (4), 570–584.

Leibig, C., Allken, V., Ayhan, M.S., Berens, P., Wahl, S., 2017. Leveraging uncertainty information from deep neural networks for disease detection. Sci. Rep. 7 (1), 1–14.

Li, K., Duggal, R., Chau, D.H., 2023. Evaluating robustness of vision transformers on imbalanced datasets (student abstract). In: Proceedings of the AAAI Conference on Artificial Intelligence. 37, (13), pp. 16252–16253.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988.

Lipton, Z.C., 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.. Queue 16 (3), 31–57.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30.

Malinin, A., Gales, M., 2018. Predictive uncertainty estimation via prior networks. Adv. Neural Inf. Process. Syst. 31.

McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.S., Darzi, A., et al., 2020. International evaluation of an AI system for breast cancer screening. Nature 577 (7788), 89–94.

Meng, Y., Yang, N., Qian, Z., Zhang, G., 2020. What makes an online review more helpful: an interpretation framework using xgboost and SHAP values. J. Theor. Appl. Electron. Commer. Res. 16 (3), 466–490.

Mohammed, A., Kora, R., 2023. A comprehensive review on ensemble deep learning: Opportunities and challenges. J. King Saud University- Comput. Inf. Sci. 35 (2), 757–774.

Muhammad, D., Bendechache, M., 2024. Unveiling the black box: a systematic review of explainable artificial intelligence in medical image analysis. Comput. Struct. Biotechnol. J..

Müller, D., Soto-Rey, I., Kramer, F., 2022. An analysis on ensemble learning optimized medical image classification with deep convolutional neural networks. Ieee Access 10, 66467–66480.

Namdari, A., Li, Z., 2019. A review of entropy measures for uncertainty quantification of stochastic processes. Adv. Mech. Eng. 11 (6), 1687814019857350.

Niculescu-Mizil, A., Caruana, R., 2005. Predicting good probabilities with supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning. pp. 625–632.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J., 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. Adv. Neural Inf. Process. Syst. 32.

Panayides, A.S., Amini, A., Filipovic, N.D., Sharma, A., Tsaftaris, S.A., Young, A., Foran, D., Do, N., Golemati, S., Kurc, T., et al., 2020. AI in medical imaging informatics: current challenges and future directions. IEEE J. Biomed. Heal. Informatics 24 (7), 1837–1857.

Panwar, H., Gupta, P., Siddiqui, M.K., Morales-Menendez, R., Bhardwaj, P., Singh, V., 2020. A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-scan images. Chaos Solitons Fractals 140, 110190.

Petticrew, M., Sowden, A., Lister-Sharp, D., 2001. False-negative results in screening programs: Medical, psychological, and other implications. Int. J. Technol. Assess. Health Care 17 (2), 164–170.

Ravishankar, H., Paluru, N., Sudhakar, P., Yalavarthy, P.K., 2025. Information geometric approaches for patient-specific test-time adaptation of deep learning models for semantic segmentation. IEEE Trans. Med. Imaging 1.

Ross, T.-Y., Dollár, G., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2980–2988.

Rufibach, K., 2010. Use of brier score to assess binary predictions. J. Clin. Epidemiol. 63 (8), 938–939.

Senousy, Z., Abdelsamea, M.M., Mohamed, M.M., Gaber, M.M., 2021. 3E-net: Entropy-based elastic ensemble of deep convolutional neural networks for grading of invasive breast carcinoma histopathological microscopic images. Entropy 23 (5), 620.

Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H., 2020. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. pp. 180–186.

Sreelakshmi, S., Malu, G., Sherly, E., Mathew, R., 2023. M-net: An encoder-decoder architecture for medical image analysis using ensemble learning. Results Eng. 17, 100927.

Syrykh, C., Abreu, A., Amara, N., Siegfried, A., Maisongrosse, V., Frenois, F.X., Martin, L., Rossi, C., Laurent, C., Brousset, P., 2020. Accurate diagnosis of lymphoma on whole-slide histopathology images using deep learning. NPJ Digit. Med. 3 (1), 63.

Tang, X., 2019. The role of artificial intelligence in medical imaging research. BJR| Open 2 (1), 20190031.

Titoriya, A., Sachdeva, S., 2019. Breast cancer histopathology image classification using AlexNet. In: 2019 4th International Conference on Information Systems and Computer Networks. ISCON, IEEE, pp. 708–712.

Tomsett, R., Harborne, D., Chakraborty, S., Gurram, P., Preece, A., 2020. Sanity checks for saliency metrics. In: Proceedings of the AAAI Conference on Artificial Intelligence. 34, (04), pp. 6021–6029.

Tran, G.S., Nghiem, T.P., Nguyen, V.T., Luong, C.M., Burie, J.-C., 2019. Improving accuracy of lung nodule classification using deep learning with focal loss. J. Heal. Eng. 2019 (1), 5156416.

Wang, X., Liu, H., Shi, C., Yang, C., 2021. Be confident! towards trustworthy graph neural networks via confidence calibration. Adv. Neural Inf. Process. Syst. 34, 23768–23779.

Wanner, J., Herm, L.-V., Heinrich, K., Janiesch, C., 2021. Stop ordering machine learning algorithms by their explainability! an empirical investigation of the tradeoff between performance and explainability. In: Conference on E-Business, E-Services and E-Society. Springer, pp. 245–258.

Xu, Z., Liu, R., Yang, S., Chai, Z., Yuan, C., 2023. Learning imbalanced data with vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15793–15803.

Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B., 2023. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Sci. Data 10 (1), 41.

Yang, S., Xiao, W., Zhang, M., Guo, S., Zhao, J., Shen, F., 2022. Image data augmentation for deep learning: A survey. arXiv preprint arXiv:2204.08610.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., Lipson, H., 2015. Understanding neural networks through deep visualization. In: Deep Learning Workshop, International Conference on Machine Learning. ICML.

Yu, X., Wang, J., Hong, Q.-Q., Teku, R., Wang, S.-H., Zhang, Y.-D., 2022. Transfer learning for medical images analyses: A survey. Neurocomputing 489, 230–254.

Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O., 2021. Understanding deep learning (still) requires rethinking generalization. Commun. ACM 64 (3), 107–115.

Zhang, J., Chao, H., Dasegowda, G., Wang, G., Kalra, M.K., Yan, P., 2022. Overlooked trustworthiness of saliency maps. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 451–461.

Zou, K., Chen, Z., Yuan, X., Shen, X., Wang, M., Fu, H., 2023. A review of uncertainty estimation and its application in medical imaging. Meta- Radiol. 1 (1), 100003.