Mathematical Biosciences and Engineering

MBE, 22(3): 528–553.
DOI: 10.3934/mbe.2025020
Received: 20 November 2024
Revised: 01 February 2025
Accepted: 14 February 2025
Published: 20 February 2025

https://www.aimspress.com/journal/MBE

Research article

Uncertainty CNNs: A path to enhanced medical image classification performance

Vasileios E. Papageorgiou[1,*], Georgios Petmezas[2], Pantelis Dogoulis[3], Maxime Cordy[3] and Nicos Maglaveras[2]

[1] Department of Mathematics, Aristotle University of Thessaloniki, Thessaloniki, Greece
[2] School of Medicine, Aristotle University of Thessaloniki, Thessaloniki, Greece
[3] SerVal, University of Luxembourg, Luxembourg City, Luxembourg

* **Correspondence:** Email: vpapageor@math.auth.gr.

Abstract: The automated detection of tumors using medical imaging data has garnered significant attention over the past decade due to the critical need for early and accurate diagnoses. This interest is fueled by advancements in computationally efficient modeling techniques and enhanced data storage capabilities. However, methodologies that account for the uncertainty of predictions remain relatively uncommon in medical imaging. Uncertainty quantification (UQ) is important as it helps decision-makers gauge their confidence in predictions and consider variability in the model inputs. Numerous deterministic deep learning (DL) methods have been developed to serve as reliable medical imaging tools, with convolutional neural networks (CNNs) being the most widely used approach. In this paper, we introduce a low-complexity uncertainty-based CNN architecture for medical image classification, particularly focused on tumor and heart failure (HF) detection. The model's predictive (aleatoric) uncertainty is quantified through a test-set augmentation technique, which generates multiple surrogates of each test image. This process enables the construction of empirical distributions for each image, which allows for the calculation of mean estimates and credible intervals. Importantly, this methodology not only provides UQ, but also significantly improves the model's classification performance. This paper represents the first effort to demonstrate that test-set augmentation can significantly improve the classification performance of medical images. The proposed DL model was evaluated using three datasets: (a) brain magnetic resonance imaging (MRI), (b) lung computed tomography (CT) scans, and (c) cardiac MRI. The low-complexity design of the model enhances its robustness against overfitting, while it is also easily re-trainable in case out-of-distribution data is encountered, due to the reduced computational resources required by the introduced architecture.

Keywords: uncertainty quantification; convolutional neural networks; biomedical image classification; tumor detection; test-set augmentation; artificial intelligence

1. Introduction

The application of artificial intelligence (AI) in medicine is revolutionizing medical imaging and diagnostics, offering transformative capabilities for early disease detection and patient management [1]. Among various AI techniques [2,3], CNNs have emerged as a powerful tool for analyzing complex medical images [4]. These models excel in identifying patterns and features in high-dimensional data, enabling automated classification tasks that are critical in modern healthcare.

Despite their success, a significant challenge in deploying CNNs for clinical use lies in addressing uncertainties inherent in medical data, which can significantly impact the classification performance and reliability [5]. Uncertainty in medical image classification arises from multiple sources and is a critical aspect of AI systems in clinical settings [6]. Aleatoric uncertainty reflects the inherent noise in data, such as poor image quality, variations in imaging protocols, or ambiguous features that make interpretation challenging. On the other hand, epistemic uncertainty arises from model limitations, such as insufficient training data, biases in the dataset, or overfitting to specific patterns. These uncertainties can hinder the reliability of predictions, particularly in high-stakes scenarios where clinical decisions directly affect the patient outcomes. Addressing these uncertainties is especially important for diagnosing life-threatening conditions such as tumors, cancers and HF, where early detection significantly improves the patient outcomes [6–10].

Over the years, various UQ techniques have been proposed. Bayesian Neural Networks (BNNs) offer a principled approach by incorporating prior distributions over the model parameters, facilitating robust uncertainty estimation [11,12]. On the other hand, Monte Carlo Dropout introduces stochasticity during inference by randomly dropping units, allowing for uncertainty estimation through multiple forward passes [13,14]. Additionally, ensemble models enhance the reliability by training multiple models and aggregating their predictions to capture the output variability [15,16], while Gaussian Mixture Models (GMMs) can be utilized to model the distribution of predictions, providing insights into the inherent uncertainty in the data [17,18]. Collectively, these techniques contribute to a deeper understanding of the model confidence and the reliability of predictions in medical imaging applications.

However, these approaches share common limitations, including high computational costs, the need for large training datasets, and the complexity in implementation. These limitations are particularly challenging in medical imaging tasks, where the datasets are often small, annotated data is expensive to acquire, and real-time inference is essential. Consequently, there is a pressing need for computationally efficient and adaptable UQ techniques that can maintain a high performance and robustness without imposing significant resource demands.

In this study, we introduce a novel low-complexity uncertainty convolutional neural network (UCNN), optimized for efficient medical image classification with robust UQ. Our approach leverages test-time augmentation, a computationally efficient technique for quantifying aleatoric uncertainty. This method generates surrogate images by applying augmentations to the test samples, thus creating an empirical distribution of predictions that reflects the variability in the data. By analyzing this distribution, our framework improves the classification performance and enhances the model

confidence, therefore addressing the limitations of traditional UQ methods. Importantly, the proposed approach is adaptable to various DL architectures, making it a versatile tool for medical imaging tasks.

The effectiveness of the proposed methodology is demonstrated across three key diagnostic tasks: (i) brain tumor detection using MRI scans, focusing on identifying abnormal growths within the brain; (ii) lung cancer detection with CT imaging, aimed at classifying pulmonary nodules and early-stage malignancies; and (iii) HF classification via cardiac MRI, leveraging advanced imaging features to assess the heart function and tissue characteristics.

A thorough analysis is conducted to determine the optimal number of augmented samples that enhance the DL model's classification performance without significantly increasing the computational load from generating large numbers of augmented images. Additionally, the robustness of the approach is validated by the empirical distribution of the model's predictions across the entire test set, demonstrating that the selected cut-off point does not affect the performance. In conclusion, the variety and diversity of datasets, along with the inclusion of both cancer and HF classification tasks, validate the reliability of the model's detection capabilities, making it a dependable tool for modern oncology applications.

It should be noted that in the field of medical imaging, test-time augmentation has primarily been used for segmentation tasks [10], where the focus is on enhancing the performance of existing segmentation algorithms, particularly in identifying abnormalities such as tumor regions. As a result, this paper aims to highlight the positive contribution of test-time augmentation in classification tasks. In the Results and Discussion section, it becomes evident that this UQ approach not only provides uncertainty representations that can be used by experts for more informative diagnoses, but also significantly improves the performance of even low-complexity CNN architectures.

Compared to the aforementioned neural network (NN) UQ approaches, test-time augmentation presents notable advantages concerning the computational efficiency and simplicity of its operation. First of all, the idea of test-set augmentation is straightforward and its implementation is easily accessible, while it can be easily combined with any DL architecture. Moreover, in contrast to ensemble methods, test-time augmentation necessitates less memory consumption both during training and the inference phase. This is attributed to the fact that ensemble models require independent training of several models to reach satisfactory UQ, while test-time augmentation produces surrogates that are evaluated on a single DL model. Another advantage is that in contrast to Bayesian methods, the combination of test-time augmentation with transfer learning does not require any additional steps such as fine-tuning, making it more easily used [15]. Finally, both ensemble and Bayesian methods impose a notably higher computational burden during the training procedure.

The novelty of the present analysis can be summarized as follows:

- The presentation of a novel low-complexity UCNN for efficient tumor detection, leveraging aleatoric UQ via a generation of surrogates through test-set augmentation.
- The first use of test-set augmentation to medical imaging classification tasks, aiming to enhance the DL models' predictive performance.
- The introduction of a computationally efficient CNN, ideal for small medical datasets, minimizing overfitting and offering flexibility for easy retraining when out-of-distribution data may occur.
- The method's validation on brain, cardiac MRI, and lung CT scan datasets, showing robust tumor detection capabilities across different medical imaging tasks, including cancer and HF detection.

The remainder of the article is structured as follows: Section 2 reviews related work on the use of CNNs and uncertainty neural networks (UNNs) in biomedical data applications; Section 3 introduces the CNN architecture used in this study, along with a detailed explanation of the test-set augmentation process; Section 4 examines the classification performance of the UCNN, analyzes the stochastic attributes introduced by augmentation, and identifies the optimal number of augmented samples with respect to the maximization of the method's performance; Section 5 provides a discussion, concludes with a summary of the study's key findings, and offers directions for future work in the field.

2. Related work

Various AI-based and machine learning (ML) algorithms have been proposed for medical imaging tasks. Notable approaches include artificial neural networks (ANN) [19], K-nearest neighbors (KNN) [20], and support vector machines (SVM) [21,22]. However, CNNs have emerged as the most effective tools for processing MRI and CT images due to their superior performance in classification tasks. Numerous studies have addressed both binary and multiclass classification problems using state-of-the-art CNN architectures, often beginning with image preprocessing steps, such as augmentation or segmentation, to optimize classification accuracy.

Research efforts related to the binary classification of benign and malignant brain tumors frequently incorporate preprocessing and augmentation techniques alongside traditional CNN frameworks. For example, Seetha et al. [23], Babu et al. [24], Pathak et al. [25], and Kulkarni et al. [26] achieved classification accuracies ranging from 94.1% to 98% by combining CNNs with preprocessing steps such as image augmentation. Similarly, hybrid CNN-SVM models have been explored, as seen in [27–29], which achieved accuracies between 88.54% and 95.62%. Studies by Sert et al. [28] and Özyurt et al. [29] utilized ResNet as the backbone for the CNN-SVM hybrid approach, incorporating techniques such as resolution enhancement and entropy-based segmentation to further refine the classification of benign and malignant tumors. Moreover, the authors of [30] compared the CNN-SVM model to a CNN-KNN approach on the same dataset, yielding an accuracy of 90.62%.

The 95% accuracy of achieved by one of the three proposed CNNs was based on the ResNet50 architecture, albeit with obvious overfitting issues, as noted by Saxena et al. [31]. In contrast, some studies have focused on more complex multi-class tumor classification tasks, addressing multiple tumor types in brain cancer diagnoses. For example, [31–33] proposed CNN models for the classification of pituitary adenomas, meningiomas and gliomas, achieving accuracies between 90.89% and 98% by comparing the performances across cropped, uncropped and segmented image datasets. A hybrid DenseNet-LSTM model, developed in [34], achieved an accuracy of 92.13% for a four-class classification task on a public dataset.

Regarding lung cancer classification, several studies have employed CNN and SVM models to analyze CT images. Studies [35,36] used the IQ-OTHNCCD dataset, thereby applying SVM classifiers and GoogleNet-based CNN architectures. The authors of [35] applied Gaussian filtering, slicing and segmentation, and achieved an accuracy of 89.88%, while [36] implemented Gabor filters and region of interest (ROI) extraction to achieve an accuracy of 94.38%. Other ML approaches to lung cancer classification have utilized KNN [37], Naive Bayes [37], SVM [38] and Random Forests (RFs) [39]. In [40], CNN architectures, including VGG16, MobileNet, AlexNet, DenseNet, VGG19 and ResNet, were employed for the classification of normal, benign and malignant tumors, with accuracy rates ranging from 48% to 56%. Polat & Mehr [41] developed a hybrid 3D CNN-SVM model, and achieved

an accuracy of 91.81%. Additionally, traditional CNN methods applied to private CT image datasets [42,43] have incorporated image enhancement techniques such as median filtering, contrast stretching, and Otsu segmentation to improve the classification performance.

Beyond oncology, CNNs have also been applied in cardiovascular disease (CVD) detection, particularly in diagnosing HF using cardiac MRI [44]. More specifically, Wolterink et al. [45] trained a CNN for cardiac MRI feature extraction and combined the extracted image features with patient characteristics in an RF classifier for HF diagnosis, and reached an accuracy of 91%. Wang et al. [46] applied a CNN-LSTM model on cardiac MRIs from 9719 patients to diagnose 11 types of CVD with an area under the curve (AUC) of 0.95. Moreover, Sharma et al. [47] proposed a CNN to detect heart disease in a cardiac MRI dataset from 30 patients, and reached an accuracy of 95%, a sensitivity of 94.1% and a specificity of 94%.

However, medical applications involve high-risk tasks, and recent methodologies have been developed to effectively address the uncertainties associated with their outcomes. This falls under the umbrella of uncertainty estimation in ML, which has become an indispensable tool in enhancing the reliability of automated decision-making. In the field of uncertainty estimation, several models have been widely applied. BNNs incorporate uncertainty by assigning distributions to network weights, which enables more reliable confidence estimates [48]. Monte Carlo (MC) dropout is another common approach, where dropout is applied during inference to capture the model uncertainty [49]. Ensemble methods involve training multiple independent models and using the variance in their predictions to quantify the uncertainty, thus providing robust performance across tasks [15].

In medical applications, uncertainty-aware models play an essential role in improving the diagnostic accuracy and patient safety. For instance, in brain tumor segmentation, uncertainty estimation helps identify uncertain regions in MRI scans, which guides radiologists in focusing on ambiguous areas [50]. In lung nodule detection, uncertainty estimates reduce false positives in CT scan analyses, ensuring more reliable diagnosis outcomes [39]. Another application is in diabetic retinopathy detection, where uncertainty-aware models assist clinicians in identifying cases requiring further review, thus enhancing decision support in retinal imaging [51]. By identifying areas of uncertainty, these models contribute to more reliable and informed decision-making, ultimately supporting improved patient outcomes and safety.

It is important to mention that in medical imaging, test-time augmentation has mostly been applied to segmentation tasks in medical imaging [15], with an emphasis on improving the performance of existing segmentation algorithms. In two occasions, Wang et al. (2018, 2019) [52,53] combined deep CNN schemes, such as U-Net, V-Net, W-Net and cascaded networks, with the aim of increasing the segmentation precision of brain tumors using the BraTS open access dataset. Moshkov et al. (2020) [54] have employed augmentation both on training and test set, thereby focusing on cell segmentation of microscopy images. DL architectures such as U-net and Mask R-CNN were used. Additionally, Gailochet et al. (2024) [55] utilized a semi-supervised approach combined with test-set augmentation for active learning, to perform segmentation on cardiac images.

Based on the above, this paper aims to emphasize the positive impact of test-time augmentation on classification tasks. During the experimental phase, it becomes clear that this UQ approach not only generates uncertainty representations that can assist experts in making more informed diagnoses, but also significantly enhances the predictive performance of CNN models. In contrast to the aforementioned analysis on medical imaging, the proposed method focuses on developing a computationally efficient CNN architecture that is robust to overfitting and well-suited for small

datasets, which are common in the medical field. Additionally, its simple structure makes it ideal for re-training in the event of out-of-distribution data.

3. Methodology

In this section, we outline the comprehensive framework designed to train the proposed architecture and enhance its generalization performance in the context of efficient tumor detection and HF classification using medical images. First, we present a detailed explanation of the core principles of CNNs and the performance metrics used to evaluate the model during the inference phase. Next, we introduce the proposed low-complexity CNN architecture, followed by a discussion of the data augmentation strategies employed to improve the model robustness. Finally, we provide a detailed description of the datasets utilized, including their sources and the respective class distributions.

3.1. Convolutional neural networks

CNNs represent a subclass of NN architectures characterized primarily by the use of convolutional kernels [56]. These networks are predominantly applied to visual tasks, including video classification, image segmentation and medical image analyses. A typical CNN architecture is comprised of several key components: convolutional layers, which play a critical role in feature extraction; pooling layers, which reduce the dimensionality of the processed tensors; batch normalization layers, which enhance the computational stability during training; and fully connected layers, which serve as the mechanism for feature selection [57]. These layers are organized in a specific sequence, beginning with the feature extraction modules and followed by fully connected layers to produce the final classification output.

Convolutional layers consist of multiple kernels which are the layers' trainable parameters, modified after each iteration. Let $\boldsymbol{X}^k \in \mathbb{R}^{M^k \times N^k \times D^k}$ be the input of the k–th convolutional layer and $\boldsymbol{F} \in \mathbb{R}^{m \times n \times D^k \times S}$ be a 4-dimensional tensor representing the S kernels of k–th layer, of a spatial span $m \times n$. The output of the k–th convolutional layer will be a tensor of order 3 denoted by $\boldsymbol{Y}^k \in \mathbb{R}^{M^k - m + 1 \times N^k - n + 1 \times S}$. The elements of this tensor result from the following equation:

$$y_{i^k, j^k, s} = \sum_{i=0}^{m} \sum_{j=0}^{n} \sum_{l=0}^{D^k} F_{i,j,l,s} \times x^k_{i^k + i, j^k + j, l} \,. \tag{1}$$

Equation (1) is repeated for all $0 \leq s \leq S$ and for any spatial location satisfying $0 \leq i^k \leq M^k - m + 1$ and $0 \leq j^k \leq N^k - n + 1$ [58].

Let $\boldsymbol{X}^k \in \mathbb{R}^{M^k \times N^k \times D^k}$ be the input of the k – th layer that is now a pooling layer with a spatial span of $n \times m$. We assume that n divides M, m divides N, and the stride equals the aforementioned spatial span. The output is a tensor $\boldsymbol{Y}^k \in \mathbb{R}^{M^{k+1} \times N^{k+1} \times D^{k+1}}$, where

$$M^{k+1} = \frac{M^k}{n}, \quad N^{k+1} = \frac{N^k}{m}, \quad D^{k+1} = D^k, \tag{2}$$

while the polling layer operates independently upon $\boldsymbol{X}^k$ channel by channel. In our network, we utilize two max pooling layers, resulting in outputs produced based on the following:

$$y_{i^k,j^k,d} = \max_{0 \le i \le n, 0 \le j \le m} x^k_{i^k \times n + i, j^k \times m + j, d}, \tag{3}$$

where $0 \le i^k \le M^k$, $0 \le j^k \le N^k$ and $0 \le d \le D^k$. Polling operation reduces the tensor's dimension, although retaining the important detected patterns [59].

On the other hand, the fully connected layers belong to the second part of a CNN, which aims to efficiently select the most valuable features extracted by the convolutional layers. The input of the first fully connected layer is a high dimensional vector which contains all extracted features produced by a flattening operation. Finally, important transition mediums, which introduce non-linearity and connects the aforementioned layers, are the operations of ReLU and batch normalization. The ReLU function is defined as follows:

$$y_{i,j,d} = max\left(0, x^k_{i,j,d}\right), \tag{4}$$

for $0 \le i \le M^k, 0 \le j \le N^k$ and $0 \le d \le D^k$, aiming to only transfer the beneficial elements for the classification.

Since we focus on tumor-based regions, we are processing medical images. A medical image $\boldsymbol{X}$ can be represented as a three-channel tensor (in the RGB system), formally defined as $\boldsymbol{X} \in \mathbb{R}^{H \times W \times 3}$, where H and W are the height and the width of the image, respectively. The input tensor $\boldsymbol{X}^{(i)}$ corresponding to the $i-$th medical image is passed through the set of successive layers and a label $\hat{y}_i$ is produced. Then, an error is calculated using a predetermined loss function. In most cases, Cross-Entropy loss is utilized, which is denoted as L_{CE}. In our occasion, where we train our network in a binary classification scenario, we employ the Binary Cross-Entropy loss function (L_{BCE}), which is defined (for a single medical image) as follows:

$$L_{BCE}(y_i, \hat{y}_i) = -y_i * log(\hat{y}_i) - (1 - y_i) * log(1 - \hat{y}_i), \tag{5}$$

where $y_i = \{0, 1\}$ corresponds to the image's ground truth. Then, the produced error is utilized in the learning procedure that represents the modification of the trainable parameters of the network based on an optimization algorithm. The majority of the analyses in the literature use Adam or AdamW [60] algorithms as the optimizers.

3.2. Uncertainty quantification with UCNN

In this work, we propose a low-complexity CNN architecture which consists of seven key layers. The first four layers, including two convolutional and two max-pooling layers, are dedicated to the feature extraction. The remaining three fully connected layers leverage the extracted features to deliver strong classification results (Figure 1). The input to the CNN are two-dimensional grayscale images of size 100 × 100 pixels, a choice made after extensive experimentation. This smaller input size reduces the computational costs while maintaining the model effectiveness.

The architecture begins with a convolutional layer using 32 kernels, each with a 9 × 9 receptive field, followed by a 4 × 4 max-pooling layer. This process is repeated with a 5 × 5 convolutional layer and another 4 × 4 max-pooling layer. Both convolutional layers utilize the same padding and ReLU activation along with batch normalization, which is applied to enhance feature extraction.

The second section of the model is comprised of three fully connected layers, with dropout applied between the first two fully connected layers at a rate of 0.2. While batch normalization reduces the need for dropout, the inclusion of a dropout layer introduces stochasticity during training, which enhances the generalization by reducing the risk of overfitting, especially given the small size of medical imaging datasets used in this study. Moreover, this stochasticity contributes to epistemic UQ when combined with test-time augmentation.

Specifically, dropout-induced randomness during inference allows the model to better capture epistemic uncertainty by varying the active neurons during prediction, thereby reflecting the variability in the model outputs. This complements the aleatoric UQ achieved through test-time augmentation, which handles data variability (e.g., differences in posture or rotation of input images). Together, these mechanisms ensure a robust classification performance by addressing both types of uncertainties that commonly arise in medical imaging tasks.

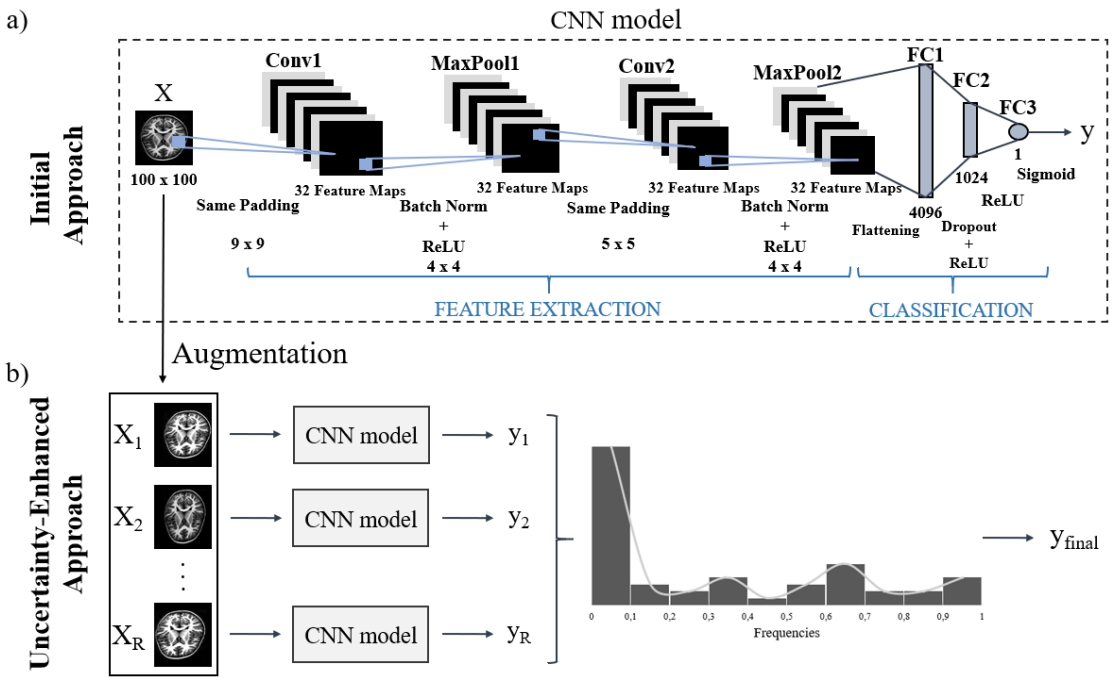


Figure 1. Diagrammatic representation of the low-complexity UCNN architecture considering test-set augmentation.

After completing the training process, during which the final weight values are determined, we employ a test-set augmentation strategy to quantify the aleatoric uncertainty of the system. Our approach involves generating a cardinality of R surrogate samples $\left\{\boldsymbol{X}_i^{(1)}, \ldots, \boldsymbol{X}_i^{(R)}\right\}$ for each test set image ($\boldsymbol{X}_i$) which, in turn, produce a set of model's estimates of the same cardinality $\left\{\hat{y}_i^{(1)}, \ldots, \hat{y}_i^{(R)}\right\}$

using data augmentation techniques. Then, the augmented samples are then evaluated to generate a predictive distribution that captures the uncertainty. The idea behind this method is that the augmented samples provide varied perspectives, helping to better quantify the uncertainty.

Typically, the same augmentation techniques used for training regularization are applied during test-time augmentation. These techniques have been shown to improve the calibration for both in-distribution and out-of-distribution data detection [61,62]. As mentioned above, each surrogate passes through the trained CNN, providing an estimate

$$\hat{y}_i^{(r)} = NN_{\boldsymbol{\theta}}\left(\boldsymbol{X}_i^{(r)}\right), \quad r = 1, \dots, R, \quad i = 1, \dots, N_D, \tag{6}$$

where $\hat{y}_i^{(r)} \in [0, 1]$ based on the sigmoid activation. Scalar N_D denotes the cardinality of the test set images. Based on the set of surrogates, an empirical distribution is derived for each medical image as follows:

$$\hat{g}_i(x) = \frac{1}{R}\sum_{r=1}^{R} \delta_{\hat{y}_i^{(r)}}(x), \tag{7}$$

where $\delta_y(x)$ represents the Dirac delta function. From this, the median estimates and confidence intervals can be produced.

We prefer the median over the mean statistic for these estimates, since the empirical distributions tend to be non-Gaussian. Ideally, by indefinitely increasing the cardinality of produced surrogate images, the empirical distribution tends to estimate the theoretical one ($g_i(x)$) of the model's test set estimates, namely:

$$\hat{g}_i(x) = \frac{1}{R}\sum_{r=1}^{R} \delta_{\hat{y}_i^{(r)}}(x) \ \rightarrow \ g_i(x), \quad R \rightarrow \infty, \tag{8}$$

considering that the employed augmentations efficiently represent the phenomenon's uncertainty. However, the determination of the optimal R remains an issue of interest for the present analysis, as we aim to enhance both the computational time and the model's classification performance. This task is extensively investigated in the Results section.

Except from the determination of the empirical distribution corresponding to each original test set image, it is feasible to identify each test set image's class according to the acquired UCNN scheme. Specifically, by taking the median of the R samples produced after the passing through the CNN scheme, the image's class ($\hat{c}_i$) can be determined through the following formula:

$$\hat{c}_i = H\left(median\left(\left\{\hat{y}_i^{(r)}\right\}_{r=1,\dots,R}\right) - \ c_{th}\right). \tag{9}$$

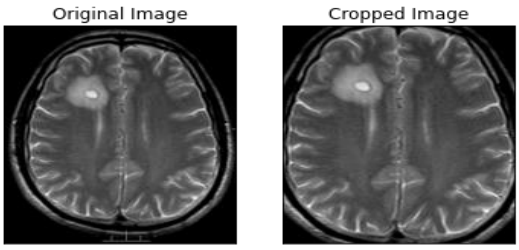
where $H(\cdot)$ denotes the Heaviside step function and c_{th} the predetermined threshold value which separates the two classes (in case of a binary classification task).

Various image-based operations were incorporated into our proposed augmentation pipelines. These include Gaussian blur, minor adjustments to the contrast, hue, brightness, and zoom -referred to as color jittering- as well as rotation and translation. Gaussian blur serves as a noise-reduction tool by

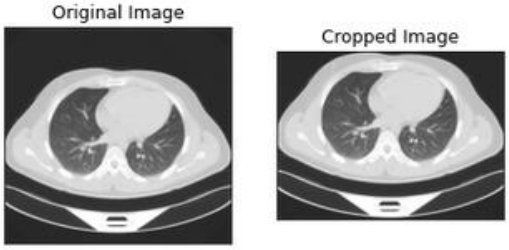
removing high-frequency details from the image regions, effectively cleaning the image. Random resizing ensures the model focuses on infectious areas regardless of the image's dimensions. Additionally, rotation and translation encourage the model to identify the tumorous regions in different parts of the image, while color jittering helps the model learn features based on the shape of infectious and healthy areas rather than relying on color, particularly given the variability in pixel values across grayscale images.

3.3. Image preprocessing

The augmentation process is also used as a preprocessing step. A series of transformations, similar to those described in Subsection 3.2, is applied. The goal of this step is to create balanced training sets and enhance the classification performance [63,64]. Following augmentation, the MRI and CT scans which show brain and lung tumors are automatically cropped to remove the outer black regions, which do not contain valuable information related to the phenomenon being studied. This helps isolate the relevant organ tissue, while it provides inputs of reduced dimensions, which decrease the required computational time. Figure 2 illustrates some examples of the cropping process. Finally, all images are resized to 100 × 100 pixels to ensure uniformity. The cardiac MRI images were already cropped. As a result, no further preprocessing was required.



a) Brain MRI displaying the existence of tumor.



b) Normal lung CT image.

Figure 2. Instances of original and cropped brain MRI and lung CT images.

3.4. Evaluation metrics

During the inference phase, we evaluate the performance of our trained model using widely accepted metrics for classification problems, including accuracy, recall, specificity and the F1 score. These are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{10}$$

$$Recall = \frac{TP}{TP + FN}, \tag{11}$$

$$Specificity = \frac{TN}{TN + FP}, \tag{12}$$

$$F1\ score = \frac{2TP}{2TP + FP + FN}. \tag{13}$$

Accuracy refers to the proportion of cases correctly classified by the model and is the simplest and most intuitive measure, though it may be misleading in cases of imbalanced data, where one class dominates. On the other hand, precision measures how many of the model's positive predictions (e.g., tumorous regions) are actually correct, thus reflecting the reliability of these predictions. Recall assesses the model's ability to identify actual positive cases, making it critical when missing positive cases (such as tumors) would have severe consequences. Specificity evaluates how well the model identifies negative cases, such as healthy (non-tumorous) regions. Lastly, the F1 score balances precision and recall by calculating their harmonic mean, with high values indicating that the model effectively identifies both tumorous and healthy regions without favoring one over the other.

3.5. Datasets

The brain cancer dataset contains 3000 images, which are used to both train and test the CNN model. It is available as an open-access resource on Kaggle [65]. For the lung cancer dataset, the images were collected over three months in 2019 at the Iraq-Oncology Teaching Hospital/National Center for Cancer Diseases (IQ-OTH/NCCD) [66]. It is comprised of 1190 CT scan slices from 110 patients, categorized into 50 malignant, 15 benign and 55 normal cases. Of these, 1097 images are labeled, including 416 normal, 120 benign and 561 malignant cases. Since our task focuses on binary classification for tumor detection, all benign and malignant images were combined and labeled as tumorous.

Finally, the cardiac MRI dataset used in our experiments is the Automated Cardiac Diagnosis Challenge (ACDC) dataset [67]. The data acquisition was held at the University Hospital of Dijon in France over a 6-year period. The dataset comprises 150 patient exams, with 30 normal controls, and 120 HF patients categorized into four groups: 30 with hypertrophic cardiomyopathy (HCM), 30 with dilated cardiomyopathy (DCM), 30 with previous myocardial infarction, and 30 with abnormal right ventricle function. To simplify the classification task into a binary format, we combined the four HF classes into a single HF category.

4. Results

The proposed low complexity scheme is employed to investigate the overall classification efficiency on all three datasets containing lung and brain tumors. For each dataset, we utilized a stratified cross-validation splitting strategy of ratio 3:1:1. This resulted in test sets that contained 600, 220 and 925 images for the brain, lung and HF dataset, respectively. Several learning rates were utilized during the

training phase, namely $\eta = \{0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001\}$, while the best classification results were generated for $\eta = 0.001$. The classification threshold was set to $c_{th} = \frac{1}{2}$. We trained our model for 50 epochs based on the Adam optimizer, which seems to have a more unstable but more effective training process compared to other optimizers such as the stochastic gradient descent (SGD). The augmentation step has been applied to both the training set and the test set. The former signifies a standard preprocessing step, which aims to improve the model's generalization to new data instances, while the latter enables the production of uncertainty estimations, which represents a new method to improve the classification efficacy of medical images.

In Table 1, we display some comparisons between the introduced architecture and respective variations, with respect to the classification accuracy, the number of parameters and the training time for one epoch. We examine how the addition or removal of one convolutional and two fully connected layers influences the above measures. We observe that the number of fully connected layers is responsible for the main difference regarding the number of parameters. In parallel, an increasing number of convolutional layers does not notably alter after the number of parameters, but is responsible for higher computational times. According to Table 1, the proposed model seems to provide the best balance between computational efficiency and performance. The models with three convolutional, and three and five fully connected layers, respectively, seem to provide slightly higher performance on some occasions. However, they require an increase of more than 100% in running time for each epoch. Apparently, this percentage is significant when we refer to multiple epochs or larger datasets. These observations led to the selection of the model with two convolutional and three fully connected layers.

Table 1. Comparisons concerning number of parameters, running time and accuracy of structure variations of the introduced CNN architecture.

Structure	Params (Million)	Time (sec)	Brain MRI Accuracy (%)	Lung CT Accuracy (%)	Cardiac MRI Accuracy (%)
1CN + 1FC	0.01	9.28	88.86 ± 4.88	96.45 ± 1.75	83.75 ± 4.03
1CN + 3FC	8.72	9.95	73.29 ± 6.81	98.92 ± 1.03	85.78 ± 0.99
1CN + 5FC	8.92	10.04	90.42 ± 4.04	97.75 ± 1.72	86.92 ± 1.25
2CN + 1FC	0.04	9.94	90.83 ± 1.67	92.68 ± 3.34	82.60 ± 5.86
2CN + 5FC	9.59	11.38	90.82 ± 0.95	95.39 ± 3.64	84.01 ± 2.08
3CN + 1FC	0.13	11.52	92.00 ± 0.82	99.38 ± 0.48	87.07 ± 1.70
3CN + 3FC	9.04	21.41	93.40 ± 1.56	99.17 ± 0.65	87.21 ± 2.23
3CN + 5FC	9.69	22.25	93.85 ± 1.13	99.50 ± 0.41	85.84 ± 2.71
2CN + 3FC	**8.87**	**10.78**	$\mathbf{93.13 \pm 0.30}$	$\mathbf{99.07 \pm 0.50}$	$\mathbf{87.29 \pm 1.56}$

In Tables 2–4 we present the tumor and HF detection performance of the employed CNN architecture. We note that the standard deviations displayed derive from the application of the cross-validation method. In these Tables, the third row refers to the original test set where still no augmentation strategy is applied. We observe a higher classification performance for the lung dataset for the training, validation and testing phase. Specifically, we receive an accuracy of 99.07 ± 0.50, a specificity of 99.07 ± 1.46, a sensitivity of 99.07 ± 1.01 and an F1-score of 99.07 ± 0.49, while we achieve an accuracy of 93.13 ± 0.30, a specificity of 95.77 ± 1.66, a sensitivity of 90.64 ± 1.56 and an F1-score of 93.03 ± 0.54 for the brain tumor test set. Finally, we receive the lowest performance with an accuracy 87.29 ± 1.56, a specificity 71.62 ± 11.07, a sensitivity of 89.88 ± 1.13 and an F1-score of 92.51 ± 1.02 for the cardiac MRI images.

Table 2. Classification performance based on brain MRI images.

	Accuracy (%)	Specificity (%)	Recall (%)	F1 score (%)
Training	99.52 ± 0.25	99.28 ± 0.40	99.50 ± 0.25	99.39 ± 0.33
Validation	93.84 ± 0.91	95.08 ± 2.06	92.71 ± 1.76	93.83 ± 2.77
Test	93.13 ± 0.30	95.77 ± 1.66	90.64 ± 1.56	93.03 ± 0.54

Table 3. Classification performance based on lung CT scans.

	Accuracy (%)	Specificity (%)	Recall (%)	F1 score (%)
Training	99.59 ± 0.12	99.31 ± 0.21	99.62 ± 0.19	99.46 ± 0.19o
Validation	98.74 ± 0.98	99.35 ± 0.52	98.28 ± 1.85	98.64 ± 1.04
Test	99.07 ± 0.50	99.07 ± 1.46	99.07 ± 1.01	99.07 ± 0.49

Table 4. Classification performance based on cardiac MRI images.

	Accuracy (%)	Specificity (%)	Recall (%)	F1 score (%)
Training	99.96 ± 0.02	99.78 ± 0.20	99.97 ± 0.02	99.96 ± 0.18
Validation	87.33 ± 1.59	84.62 ± 4.03	89.11 ± 1.98	87.63 ± 3.17
Test	87.29 ± 1.56	71.62 ± 11.07	89.88 ± 1.13	92.51 ± 1.02

Next, we apply the previously mentioned test-set augmentation methodology to generate 5, 10, 20, 50, and 100 surrogate images for each test set instance in the brain, lung and cardiac datasets. Experimenting these varying sample sizes will help identify the optimal number of samples concerning the classification performance and computational efficiency. For all datasets, the classification performance improves as the sample sizes increase. Specifically, for the brain tumor and cardiac MRI datasets, our model achieves the highest detection efficiency using $R = 50$ samples. In contrast, for the lung tumor dataset – where the CNN model performed well even with the initial test set – the performance peaks at $R = 20$ and then plateaus. This indicates that while $R = 50$ surrogates present a suitable choice for all datasets with respect to the performance, continuously increasing the number of samples does not necessarily lead to an improved accuracy.

Based on Tables 5–7, for $R = 50$, the brain tumor dataset leads to an accuracy of 98.13%, a specificity of 97.67%, a sensitivity of 98.60%, and an F1-score of 98.16%. For the lung tumor dataset, the accuracy is 99.87%, the specificity at 100%, the sensitivity at 99.73%, and the F1-score of 99.87%. Finally, for the cardiac MRI dataset, the UCNN with 50 surrogates yields an average accuracy of 91.83%, a specificity of 92.65%, a sensitivity of 91.31%, and an F1-score of 95.26%.

Figures 3–5 display examples of confusion matrices from one of the five iterations of the brain, lung and cardiac MRI test sets. Another significant observation is that the use of test-set augmentation results in smaller differences between the evaluation metrics for specificity and recall. Moreover, the test-set augmentation procedure prevents the model from being affected by the imbalance between the normal and HF classes.

In Table 8, we assess the suitability of central statistical measures–mean and median–based on the performance of UCNN. After applying the median to the test set of brain scans, we achieved an accuracy of 98.13 ± 0.74, compared to the mean value, which yielded an accuracy of 96.67 ± 1.21. For the lung CT scan dataset, we observe accuracies of 99.87 ± 0.18 and 99.77 ± 0.27, respectively, while the UCNN yield accuracies of 89.24 ± 1.78 and 91.83 ± 0.96 for the cardiac MRI images, respectively. Therefore, this experiment reinforces the choice of using the median, demonstrating its appropriateness both theoretically and empirically.

Table 5. Assessing the performance of test-set augmentation using brain MRI scans.

	Accuracy (%)	Specificity (%)	Recall (%)	F1 score (%)
Test set	93.13 ± 0.30	95.77 ± 1.66	92.64 ± 1.56	93.03 ± 0.54
5 samples	97.03 ± 0.73	96.78 ± 0.90	97.29 ± 1.29	97.07 ± 0.71
10 samples	97.50 ± 0.90	96.61 ± 1.93	98.40 ± 1.00	97.55 ± 0.83
20 samples	97.70 ± 0.86	97.06 ± 1.52	98.61 ± 0.92	97.88 ± 0.69
50 samples	98.13 ± 0.74	97.67 ± 1.23	98.60 ± 0.76	98.16 ± 0.71
100 samples	96.73 ± 0.69	94.27 ± 1.02	99.46 ± 0.76	96.64 ± 0.84

Table 6. Assessing the performance of test-set augmentation using lung CT scans.

	Accuracy (%)	Specificity (%)	Recall (%)	F1 score (%)
Test set	99.07 ± 0.50	99.07 ± 1.46	99.07 ± 1.01	99.07 ± 0.49
5 samples	99.46 ± 0.19	99.46 ± 0.56	99.20 ± 1.10	99.33 ± 0.34
10 samples	99.47 ± 0.38	99.47 ± 0.87	99.47 ± 0.73	99.45 ± 0.39
20 samples	99.87 ± 0.18	100.00 ± 0.00	99.73 ± 0.37	99.87 ± 0.18
50 samples	99.87 ± 0.18	100.00 ± 0.00	99.73 ± 0.37	99.87 ± 0.18
100 samples	99.87 ± 0.18	100.00 ± 0.00	99.73 ± 0.37	99.87 ± 0.18

Table 7. Investigating the performance of test-set augmentation based on cardiac MRI images.

	Accuracy (%)	Specificity (%)	Recall (%)	F1 score (%)
Test set	87.29 ± 1.56	71.62 ± 11.07	89.88 ± 1.13	92.51 ± 1.02
5 samples	90.31 ± 1.00	88.23 ± 3.41	90.74 ± 1.20	94.43 ± 0.69
10 samples	90.82 ± 0.75	89.18 ± 6.58	91.26 ± 1.30	94.71 ± 0.47
20 samples	91.38 ± 1.12	90.01 ± 0.50	92.35 ± 2.05	94.99 ± 0.61
50 samples	91.83 ± 0.96	92.65 ± 1.89	91.31 ± 1.10	95.26 ± 0.52
100 samples	91.67 ± 0.93	93.02 ± 2.99	91.23 ± 1.15	95.17 ± 0.50

Table 8. Assessing the performance of test-set augmentation using lung CT scans.

		Accuracy (%)	Specificity (%)	Recall (%)	F1 score (%)
Lung CT scans	Mean	99.77 ± 0.27	100.00 ± 0.00	99.64 ± 0.42	99.82 ± 0.21
	Median	99.87 ± 0.18	100.00 ± 0.00	99.73 ± 0.37	99.87 ± 0.18
Brain MRI Images	Mean	96.67 ± 1.21	96.01 ± 4.21	94.57 ± 2.57	96.41 ± 1.62
	Median	98.13 ± 0.74	98.67 ± 1.23	98.60 ± 0.76	98.16 ± 0.71
Cardiac MRI Images	Mean	89.24 ± 1.78	84.35 ± 5.35	90.29 ± 1.80	93.27 ± 0.96
	Median	91.83 ± 0.96	92.65 ± 1.89	91.31 ± 1.10	95.26 ± 0.52

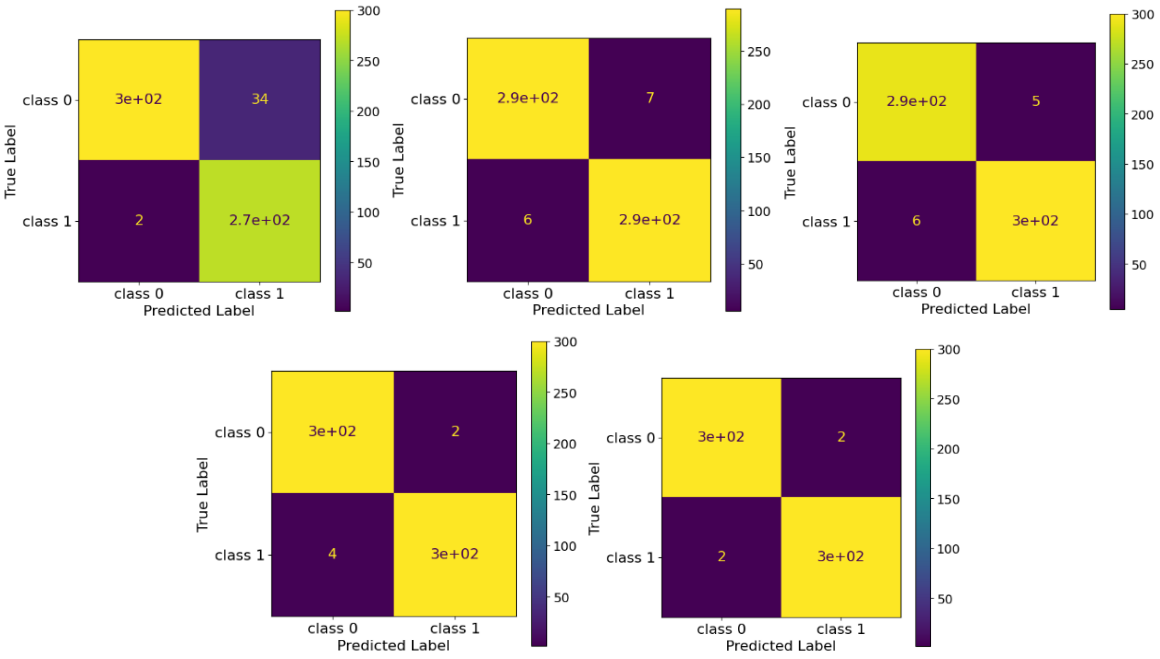


Figure 3. Confusion matrix instances corresponding to the tumor detection performance of the UCNN based on MRI brain images for the original test set and set containing 5, 10, 20 and 50 augmented samples.

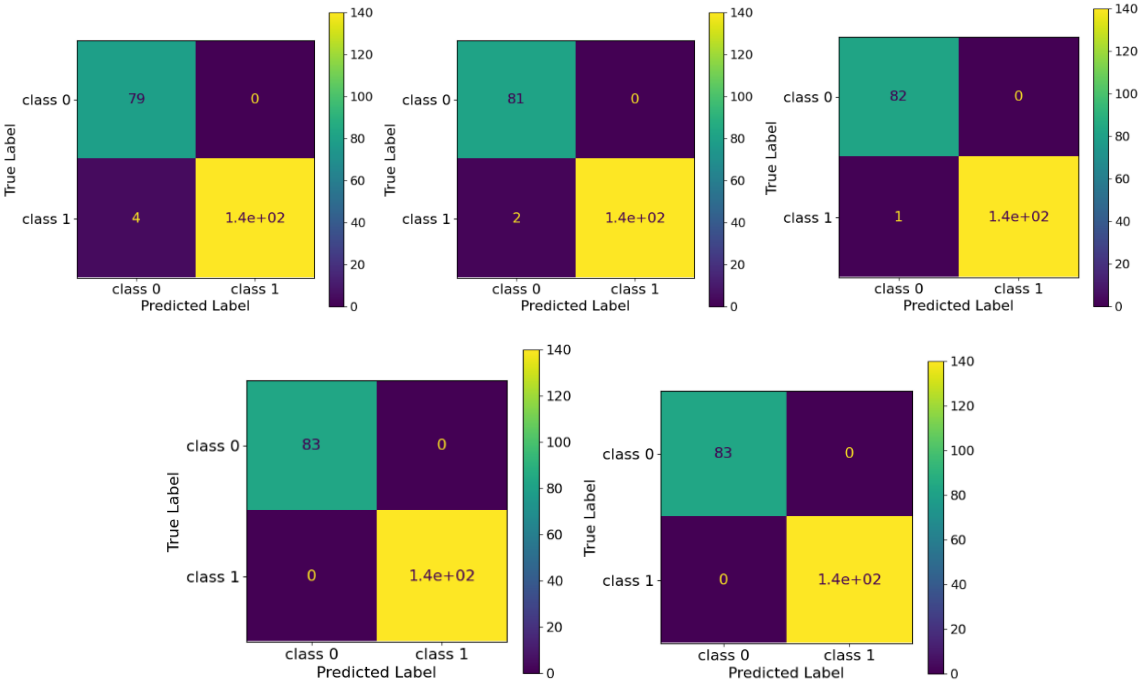


Figure 4. Confusion matrix instances corresponding to the tumor detection performance of the UCNN based on CT lung scans for the original test set and set containing 5, 10, 20 and 50 augmented samples.

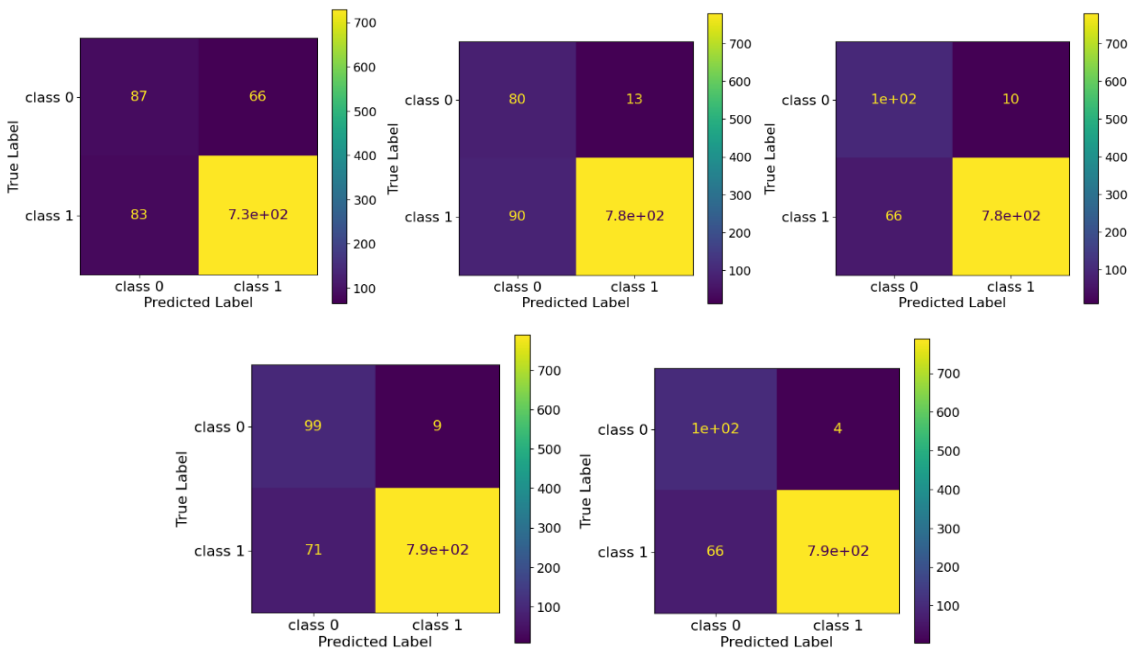


Figure 5. Confusion matrix instances corresponding to the HF detection performance of the UCNN based on cardiac MRI scans for the original test set and set containing 5, 10, 20 and 50 augmented samples.

Figures 6–8 illustrate the empirical distributions of predictions for three test set instances from the brain, lung and cardiac MRI datasets, respectively. In both Figures, image (A) represents a normal case, image (B) depicts an abnormal-like case, and image (C) shows an abnormal scan. The inclusion of these instances demonstrates that the UCNN model is well-calibrated and that the integrated UQ procedure yields reliable estimates.

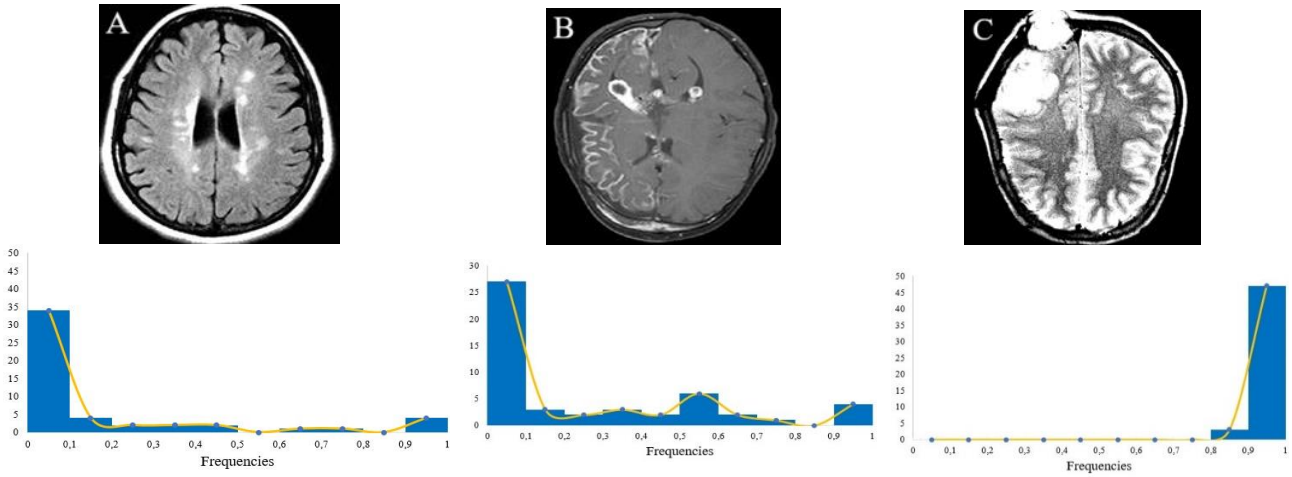


Figure 6. UQ of three brain MRI scans showing the original test image along with the empirical distribution. (A) shows a normal MRI, (B) a tumor-like normal MRI, and (C) a tumorous MRI image.

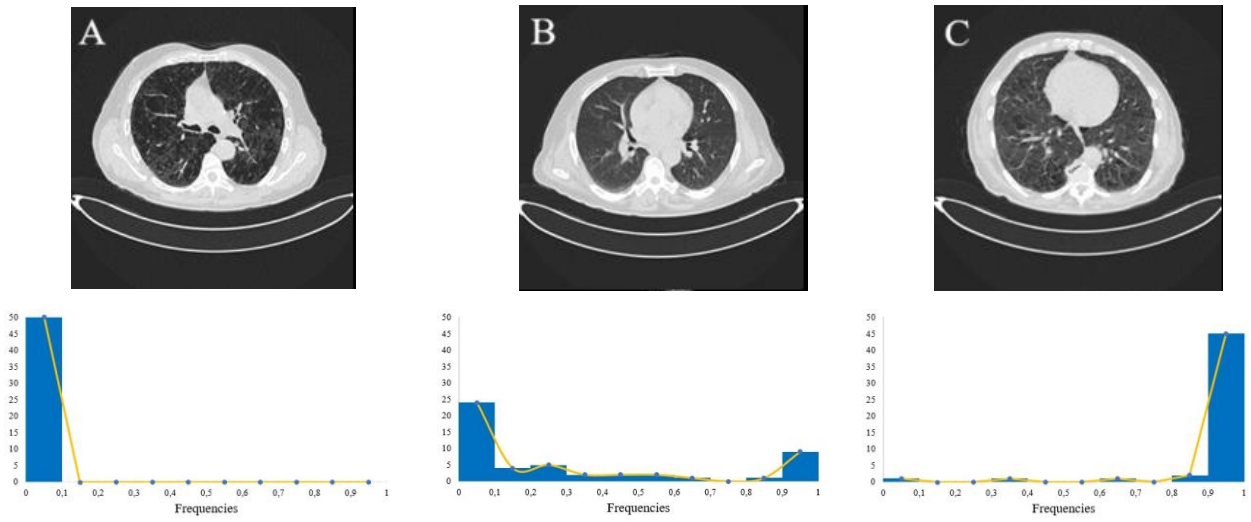


Figure 7. UQ of three lung CT scans showing the original test image along with the empirical distribution. (A) displays a normal CT scan, (B) a tumor-like normal CT scan, and (C) a lung tumor case.

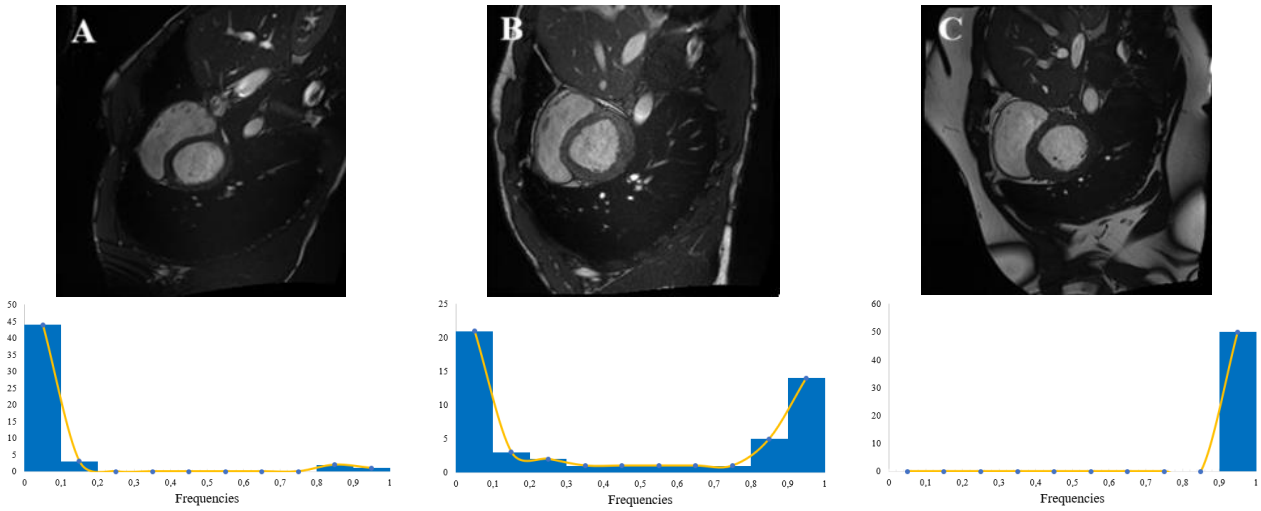


Figure 8. UQ of three cardiac MRIs showing the original test image along with the empirical distribution. (A) displays a normal case, (B) presents a HF-like normal case, and (C) shows an HF case characterized by left ventricular wall thickening.

In Figures 6 and 7, we notice that the DL model accurately classifies most surrogate images into class 0 and 1 for the non-tumorous (A) and tumorous scan (C), respectively. Additionally, the model demonstrates a strong confidence in these predictions, with only a few cases notably deviating from 0 or 1. Of particular importance is scan (B), which reveals tumor-like instances. As previously noted, tumors typically present as white, round regions within the organ. In Figure 6, a small white spot in the upper left part suggests the presence of a benign brain tumor, leading to surrogate predictions that deviate from 0. Consequently, the introduced UCNN generates an empirical distribution with an increased uncertainty, thus highlighting this suspicious area. A similar pattern is observed in image (B) of Figure 7, which experts annotated as non-tumorous. The circular region at the center of the scan

could be mistakenly interpreted as tumorous. However, our model not only correctly identifies this as non-tumorous, but also produces estimates of heightened uncertainty, thus reinforcing the significance of this characteristic.

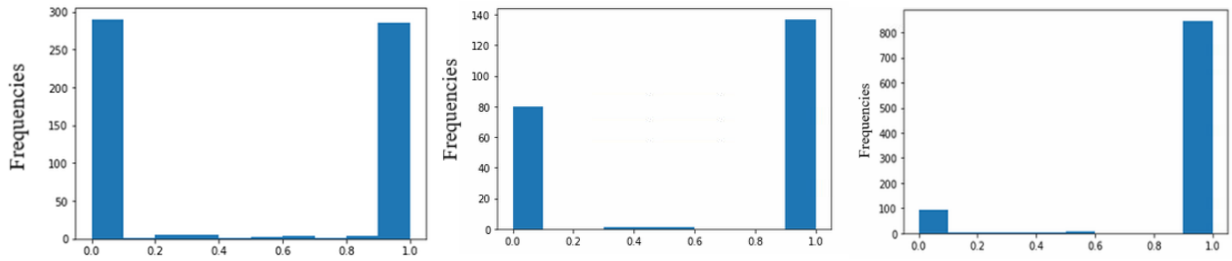


Figure 9. Empirical distributions corresponding to the median estimates produced by the UCNN for the test set's 600 brain MRIs (left panel), 220 lung CT scans (middle panel), and 925 cardiac MRIs (right panel) using $R = 50$ augmented samples.

Figure 8 illustrates the empirical distributions obtained from the median predictions of the UCNN. Image (A) represents a non-HF case, with median predictions clustered around 0, showing the model's high confidence in this classification. Image (B) presents an HF-like instance, with a slight deviation from 0 and an increased prediction uncertainty, indicating that this instance shares characteristics with HF cases, prompting the model to exhibit caution in its classification. Image (C), an HF-positive case, shows predictions tightly clustered around 1, thus reflecting a high confidence in this positive classification.

Finally, Figure 9 shows the empirical distributions derived from the median predictions of the CNN model for the medical instances in the test set, utilizing $R = 50$ surrogate images generated through augmentation. The histograms indicate that the DL architecture is well-calibrated, demonstrating a strong confidence in its class predictions. Importantly, since only a few images have median values that diverge from 0 or 1, the selection of a cut-off point (c_{th}) for the class distinction does not affect the outcomes.

Overall, these findings demonstrate the efficacy of the proposed methodology in achieving a high classification performance while maintaining the computational efficiency. The integration of test-set augmentation, along with the optimization of model configurations, enables robust predictions across diverse medical imaging datasets, showcasing the approach's adaptability and scalability to different clinical scenarios. This not only highlights the benefits of the proposed method over conventional approaches, but also reinforces its suitability for small datasets typically encountered in medical research.

5. Discussion and conclusions

In this paper, we present a low-complexity UCNN tailored for effective medical image classification, specifically targeting tumor and HF detection. The quantification of the system's aleatoric uncertainty is achieved through a test-set augmentation process applied to medical imaging. This paper marks the first attempt to demonstrate that test-set augmentation can enhance the DL classification performance in the field of medical imaging. The effectiveness of the proposed methodology is assessed using datasets of brain MRI, lung CT and cardiac MRI scans. It should be

noted that augmentations have been applied to both the training and test sets. The (standard) training set augmentation focuses on enhancing the model's ability to generalize to new data, while the (introduced) test-set augmentation step facilitates the generation of uncertainty estimations.

A significant finding of this uncertainty-based DL approach is that generating augmented samples substantially enhances the classification accuracy during testing. By producing surrogate data from each original instance, the model gains additional information that is incorporated to improve the prediction accuracy. This process boosts the model's confidence in its output, thereby reducing the impact of noisy inputs. Additionally, the simplistic structure of the proposed network makes it particularly suitable for analyzing small datasets often encountered in medical research, in contrast to many studies that depend on more complex and computationally expensive architectures. This design minimizes the risk of overfitting, but also enhances the method's adaptability and re-trainability to out-of-distribution data.

A detailed analysis shows that using $R = 50$ augmented (surrogate) images is the optimal sample size to enhance the classification performance of the DL model, for the majority of experiments. A number of $R = 20$ surrogates is also satisfactory to achieve the maximum efficiency concerning the lung CT scans dataset. This finding indicates that generating an unlimited number of surrogates isn't necessarily beneficial. Determining the ideal sample size also helps lessen the computational demands of our method. For instance, for the brain tumor dataset, using 50 surrogate images improves the average accuracy by 5%, the specificity by 1.9%, the sensitivity by 7.96%, and the F1-score by 5.13% compared to the original test set. A similar trend is observed in the lung tumor and HF datasets, supporting the effectiveness of our approach. It should be noted that the test-set augmentation procedure also enhances the model's robustness against class imbalance, too, even when a small number of surrogates is generated (Tables 5–7). However, it is important to note that while the use of 50 surrogates appears to be suitable for the datasets under consideration, a smaller number of samples could still achieve a satisfactory predictive accuracy, potentially further reducing the computational resources required.

Additionally, the proposed UCNN effectively captures the inherent uncertainty present in the datasets analyzed, as illustrated in Figures 6–8. Based on the analysis of these images, medical professionals can focus more on scans that exhibit higher levels of uncertainty. This uncertainty can subsequently be reduced through more targeted screening, aiding experts in making better-informed decisions.

Furthermore, as shown in Figure 9, the effectiveness of our approach is bolstered by examining the empirical distribution of the model's predictions across the entire test set. This analysis verifies that the selected cut-off point does not impact the classification performance. This feature adds value to the proposed model, as establishing a classification threshold is often a difficult challenge in such tasks [68]. Additionally, Table 8 demonstrates the validity of using median statistics. With the model producing 50 surrogate images, the UCNN's classification performance improves, which can be attributed to the median's robustness: it remains unaffected by outliers and is statistically more appropriate given the presence of non-Gaussian empirical distributions.

Test-set augmentation is not only a straightforward approach that enhances both the uncertainty estimation and the model performance, but also offers significant advantages in terms of the computational efficiency. It demands far less memory than ensemble methods during both training and inference. Additionally, when combined with transfer learning, test-set augmentation avoids the need

for extra steps, unlike Bayesian methods. Overall, both ensemble and Bayesian approaches typically incur much higher computational costs during training.

Aleatoric and epistemic uncertainty are two distinct types of uncertainty encountered in DL and statistical modeling [69,70]. Aleatoric uncertainty arises from inherent randomness or noise in the data. It reflects the variability present in the observations themselves and cannot be reduced by gathering more data, as it is tied to the stochastic nature of the environment or the measurement process. On the other hand, epistemic uncertainty is related to the model's lack of knowledge or limitations in learning. This uncertainty stems from insufficient or incomplete data and can be reduced by either improving the model or acquiring more representative data [71].

A practical way to reduce aleatoric uncertainty is by obtaining higher-quality data or more precise measurements [72]. For instance, enhancing sensor resolution in imaging tasks or using more accurate instruments can help minimize noise at the source. However, because improving the sampling process can be expensive and not always feasible, an alternative and cost-effective approach is to apply data augmentation techniques. These help the model become robust to variations in the data, thus reducing the impact of noise or outliers. Therefore, this method offers a low-cost solution for uncertainty reduction, particularly valuable in medical imaging.

In contrast, epistemic uncertainty can be reduced by collecting more diverse and representative data. Expanding the training dataset, especially by including underrepresented or previously unseen cases, helps the model generalize better in new situations. Another strategy is to employ more advanced or better-calibrated models, such as BNNs or deep ensembles, which can capture the uncertainty by modeling distributions over possible parameters [15, 73,74]. These methods hold promise for future research in this area.

Although the proposed method focuses on enhancing UQ and improving the performance of a relatively simple and widely recognized CNN model, we acknowledge that comparisons with state-of-the-art architectures, such as VGGs, U-Nets and GoogLeNet, could meaningfully contribute to the broader evaluation of our approach. Future research could explore integrating the proposed uncertainty estimation framework with these more complex architectures to evaluate its performance across a wider spectrum of model designs. Such comparisons would not only help establish the generalizability of the methodology, but also offer a deeper understanding of its strengths and limitations in diverse clinical imaging scenarios.

Additionally, while learnable test time augmentation techniques help in selecting appropriate augmentations, a key limitation is understanding how different types of augmentations impact uncertainty. For instance, a straightforward augmentation such as reflection may not account for much uncertainty, whereas more specialized techniques such as stretching and shearing might better capture it. Moreover, evaluating the robustness of X-ray classification is challenging and yields significantly different results depending on the dataset, architecture, and robustness metric used [75].

In future work, exploring advanced UQ methods, such as Bayesian DL, could further refine the model reliability for critical medical imaging tasks. These methods not only provide a measure of the uncertainty, but also enhance the interpretability of the model predictions, which is vital in clinical settings. Also, investigating optimized augmentation strategies tailored to specific image types may enhance the uncertainty estimation accuracy. By customizing augmentations based on the unique characteristics of different medical images, we can improve the robustness of the model against variations and artifacts commonly encountered in clinical practice.

Moreover, it would be valuable to conduct a more in-depth analysis of the individual samples, particularly challenging ones, to confirm the real-world effectiveness of the test-set augmentation strategy. Finally, incorporating a feedback loop where the model performance is continuously assessed against real-world data could lead to ongoing improvements in both the accuracy and the reliability. Collaborative efforts with medical professionals to validate these advancements will ensure that the tools developed are aligned with clinical needs, ultimately contributing to better patient outcomes and more effective decision-making in healthcare.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

We would like to thank you for following the instructions above very closely in advance. The utilized datasets exist in public repositories and can be accessed through the links cited in references [62–64].

Conflict of interest

The authors declare there is no conflict of interest.

References

1. S. A. Alowais, S. S. Alghamdi, N. Alsuhebany, T. Alqahtani, A. I. Alshaya, S. N. Almohareb, et al., Revolutionizing healthcare: The role of artificial intelligence in clinical practice, *BMC Med. Educ.*, **23** (2023), 689.
2. D. Saligkaras, V. E. Papageorgiou, On the detection of patterns in electricity prices across European countries: An unsupervised machine learning approach, *AIMS Energy*, **10** (2022), 1146–1164. https://doi.org/10.3934/energy.2022054
3. D. Saligkaras, V. E. Papageorgiou, Seeking the truth beyond the data. An unsupervised machine learning approach, *AIP Conf. Proc.*, **2812** (2023), 020106. https://doi.org/10.1063/5.0161454
4. S. S. Kshatri, D. Singh, Convolutional neural network in medical image analysis: A review, *Arch. Comput. Methods Eng.*, **30** (2023), 2793–2810. https://doi.org/10.1007/s11831-023-09898-w
5. K. Zou, Z. Chen, X. Yuan, X. Shen, M. Wang, H. Fu, A review of uncertainty estimation and its application in medical imaging, *Meta-Radiol.*, **1** (2023), 100003. https://doi.org/10.1016/j.metrad.2023.100003
6. L. Huang, S. Ruan, Y. Xing, M. Feng, A review of uncertainty quantification in medical image analysis: Probabilistic and non-probabilistic methods, *Med. Image Anal.*, **97** (2024), 103223. https://doi.org/10.1016/j.media.2024.103223
7. M. M. Jassim, Systematic review for lung cancer detection and lung nodule classification: Taxonomy, challenges, and recommendation future works, *J. Intell. Syst.*, **31** (2022), 944–964. https://doi.org/10.1515/jisys-2022-0062

8. H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, et al., Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, *CA Cancer J. Clin.*, **71** (2021), 209–249. https://doi.org/10.3322/caac.21660
9. G. Savarese, P. M. Becher, L. H. Lund, P. Seferovic, G. M. C. Rosano, A. J. S. Coats, Global burden of heart failure: A comprehensive and updated review of epidemiology, *Cardiovasc. Res.*, **119** (2023), 1453. https://doi.org/10.1093/cvr/cvad026
10. A. Inamdar, A. Inamdar, Heart failure: Diagnosis, management and utilization, *J. Clin. Med.*, **5** (2016), 62. https://doi.org/10.3390/jcm5070062
11. L. Li, J. Chang, A. Vakanski, Y. Wang, T. Yao, M. Xian, Uncertainty quantification in multivariable regression for material property prediction with Bayesian neural networks, *Sci. Rep.*, **14** (2024), 1783. https://doi.org/10.1038/s41598-024-61189-x
12. A. Olivier, M. D. Shields, L. Graham-Brady, Bayesian neural networks for uncertainty quantification in data-driven materials modeling, *Comput. Methods Appl. Mech. Eng.*, **386** (2021), 114079. https://doi.org/10.1016/j.cma.2021.114079
13. M. Malmström, I. Skog, D. Axehill, F. Gustafsson, Uncertainty quantification in neural network classifiers—A local linear approach, *Automatica*, **163** (2024), 111563. https://doi.org/10.1016/j.automatica.2024.111563
14. A. Kendall, Y. Gal, What uncertainties do we need in Bayesian deep learning for computer vision?, *Adv. Neural Inf. Process. Syst.*, **30** (2017).
15. J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, et al., A survey of uncertainty in deep neural networks, *Artif. Intell. Rev.*, **56** (2023), 1513–1589. https://doi.org/10.1007/s10462-023-10562-9
16. B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, *Adv. Neural Inf. Process. Syst.*, **31** (2017), 6405–6416.
17. A. R. Tan, S. Urata, S. Goldman, J. C. B. Dietschreit, R. Gómez-Bombarelli, Single-model uncertainty quantification in neural network potentials does not consistently outperform model ensembles, *npj Comput. Mater.*, **9** (2023), 148. https://doi.org/10.1038/s41524-023-01180-8
18. G. Gao, H. Jiang, J. C. Vink, C. Chen, Y. El Khamra, J. J. Ita, Gaussian mixture model fitting method for uncertainty quantification by conditioning to production data, *Comput. Geosci.*, **24** (2019), 663–681. https://doi.org/10.1007/s10596-019-9823-3
19. S. Manjunath, M. B. S. Pande, B. N. Raveesh, G. K. Madhusudhan, Brain tumor detection and classification using convolution neural network, *Int. J. Recent Technol. Eng.*, **8** (2019), 34–40.
20. R. H. Ramdlon, E. M. Kusumaningtyas, T. Karlita, Brain tumor classification using MRI images with K-nearest neighbor method, *2019 Int. Electron. Symp.*, (2019), 660–667. https://doi.org/10.1109/ELECSYM.2019.8901560
21. N. Vani, A. Sowmya, N. Jayamma, Brain tumor classification using support vector machine, *Int. Res. J. Eng. Technol*., **4** (2017), 1724–1729.
22. A. R. Mathew, P. B. Anto, Tumor detection and classification of MRI brain image using wavelet transform and SVM, in *2017 International Conference on Signal Processing and Communication (ICSPC)*, (2017). https://doi.org/10.1109/CSPC.2017.8305810
23. J. Seetha, S. S. Raja, Brain tumor classification using convolutional neural networks, *Biomed. Pharmacol. J.*, **11** (2018), 1457–1461. https://dx.doi.org/10.13005/bpj/1511
24. K. R. Babu, U. S. Deepthi, A. S. Madhuri, P. S. Prasad, S. Shammem, Comparative analysis of brain tumor detection using deep learning methods, *Int. J. Sci. Technol. Res.*, **8** (2019), 250–254.

25. K. Pathak, M. Pavthawala, N. Patel, D. Malek, V. Shah, B. Vaidya, Classification of brain tumor using convolutional neural network, in *2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)*, (2019), 128–132. https://doi.org/10.1109/ICECA.2019.8821931
26. S. M. Kulkarni, G. Sundari, Brain MRI classification using deep learning algorithm, *Int. J. Eng. Adv. Technol.*, **9** (2020), 1226–1231. https://doi.org/10.35940/ijeat.C5350.029320
27. R. Lang, K. Jia, J. Feng, Brain tumor identification based on CNN-SVM model, in *Proceedings of the 2nd International Conference on Biomedical Engineering and Bioinformatics*, (2018), 31–35. https://doi.org/10.1145/3278198.3278209
28. E. Sert, F. Özyurt, A. Doğanteklin, A new approach for brain tumor diagnosis system: Single image super-resolution-based maximum fuzzy entropy segmentation and convolutional neural network, *Med. Hypotheses*, **133** (2019), 109438. https://doi.org/10.1016/j.mehy.2019.109413
29. F. Özyurt, E. Sert, E. Avci, E. Doğanteklin, Brain tumor detection on convolutional neural networks with neutrosophic expert maximum fuzzy sure entropy, *Measurement*, **147** (2019), 106830. https://doi.org/10.1016/j.measurement.2019.07.058
30. P. Saxena, A. Maheshwari, S. Maheshwari, Predictive modeling of brain tumor: A deep learning approach, in *Innovations in Computational Intelligence and Computer Vision: Proceedings of ICICV 2020*, (2020), 275–285.
31. S. Das, O. F. M. Riaz, R. Aranya, N. N. Labiba, Brain tumor classification using convolutional neural network, *Int. Conf. Adv. Sci. Eng. Robot. Technol.*, 2019.
32. P. Afshar, K. N. Plataniotis, A. Mohammadi, Capsule networks for brain tumor classification based on MRI images and coarse tumor boundaries, in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (2019), 1368–1372. https://doi.org/10.1109/ICASSP.2019.8683759
33. A. M. Alqudah, H. Alquraan, I. A. Qasmieh, A. Alqudah, W. Al-Sharu, Brain tumor classification using deep learning technique–A comparison between cropped, uncropped, and segmented lesion images with different sizes, *Int. J. Adv. Trends Comput. Sci. Eng.*, **8** (2019), 3684–3691.
34. Y. Zhou, Z. Li, H. Zhu, C. Chen, M. Gao, K. Xu, et al., Holistic brain tumor screening and classification based on DenseNet and recurrent neural network, in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4.*, **11383** (2019), 208–217. https://doi.org/10.1007/978-3-030-11723-8_21
35. H. F. Kareem, M. S. Al-Husieny, F. Y. Mohsen, E. A. Khalil, Z. S. Hassan, Evaluation of SVM performance in the detection of lung cancer in marked CT scan dataset, *Indones. J. Electr. Eng. Comput. Sci.*, **21** (2021), 1731–1738. http://doi.org/10.11591/ijeecs.v21.i3.pp1731-1738
36. M. S. Al-Huseiny, A. S. Sajit, Transfer learning with GoogLeNet for detection of lung cancer, *Indones. J. Electr. Eng. Comput. Sci.*, **22** (2021), 1078–1086. http://doi.org/10.11591/ijeecs.v22.i2.pp1078-1086
37. S. M. Naqi, M. Sharif, I. U. Lali, A 3D nodule candidate detection method supported by hybrid features to reduce false positives in lung nodule detection, *Multimed. Tools Appl.*, **78** (2019), 26287–26311. https://doi.org/10.1007/s11042-019-07819-3
38. W. Abbas, K. B. Khan, M. Aqeel, M. A. Azam, M. H. Ghouri, F. H. Jaskani, Lungs nodule cancer detection using statistical techniques, in *2020 IEEE 23rd International Multitopic Conference (INMIC)*, (2020), 1–6. https://doi.org/10.1109/INMIC50486.2020.9318181

39. K. Roy, S. S. Chaudhury, M. Burman, A. Ganguly, C. Dutta, S. Banik, et al., A comparative study of lung cancer detection using supervised neural network, in *2019 International Conference on Opto-Electronics and Applied Optics (Optronix)*, (2019), 1–5. https://doi.org/10.1109/OPTRONIX.2019.8862326
40. A. Mohite, Application of transfer learning technique for detection and classification of lung cancer using CT images, *Int. J. Sci. Res. Manag.*, **9** (2021), 621–634.
41. H. Polat, H. D. Mehr, Classification of pulmonary CT images by using hybrid 3D-deep convolutional neural network architecture, *Appl. Sci.*, **9** (2019), 940. https://doi.org/10.3390/app9050940
42. S. Mukherjee, S. U. Bohra, Lung cancer disease diagnosis using machine learning approach, in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, (2020), 207–211. https://doi.org/10.1109/ICISS49785.2020.9315909
43. A. Hoque, A. K. M. A. Farabi, F. Ahmed, M. Z. Islam, Automated detection of lung cancer using CT scan images, in *2020 IEEE Region 10 Symposium (TENSYMP)*, (2020), 1030–1033. https://doi.org/10.1109/TENSYMP50017.2020.9230861
44. G. Petmezas, V. E. Papageorgiou, V. Vassilikos, E. Pagourelias, G. Tsaklidis, A. K. Katsaggelos, et al., Recent advancements and applications of deep learning in heart failure: A systematic review, *Comput. Biol. Med.*, **176** (2024), 108557. https://doi.org/10.1016/j.compbiomed.2024.108557
45. J. M. Wolterink, T. Leiner, M. A. Viergever, I. Išgum, Automatic segmentation and disease classification using cardiac cine MR images, *Lect. Notes Comput. Sci.*, (2018), 101–110.
46. Y. R. Wang, K. Yang, Y. Wen, P. Wang, Y. Hu, Y. Lai, et al., Screening and diagnosis of cardiovascular disease using artificial intelligence-enabled cardiac magnetic resonance imaging, *Nat. Med.*, **30** (2024), 1471–1480. https://doi.org/10.1038/s41591-024-02971-2
47. A. Sharma, R. Kumar, V. Jaiswal, Classification of heart disease from MRI images using convolutional neural network, in *2021 6th International Conference on Signal Processing, Computing and Control (ISPCC)*, (2021), 1–6. https://doi.org/10.1109/ISPCC53510.2021.9609408
48. M. Magris, A. Iosifidis, Bayesian learning for neural networks: An algorithmic survey, *Artif. Intell. Rev.*, **56** (2023), 11773–11823. https://doi.org/10.1007/s10462-023-10443-1
49. Y. Gal, Z. Ghahramani, Dropout as a Bayesian approximation: Representing model uncertainty in deep learning, in *International Conference on Machine Learning*, **48** (2016), 1050–1059.
50. T. Nair, D. Precup, D. L. Arnold, T. Arbel, Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation, *Med. Image Anal.*, **59** (2020), 101557. https://doi.org/10.1016/j.media.2019.101557
51. P. Seeböck, J. I. Orlando, T. Schlegl, S. M. Waldstein, H. Bogunović, S. Klimscha, et al., Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal OCT, *IEEE Trans. Med. Imaging*, **39** (2020), 87–98. https://doi.org/10.1109/TMI.2019.2919951
52. G. Wang, W. Li, S. Ourselin, T. Vercauteren, Automatic brain tumor segmentation using convolutional neural networks with test-time augmentation, in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018*, (2018), 61–72.
53. G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks, *Neurocomputing*, **338** (2019), 34–45. https://doi.org/10.1016/j.neucom.2019.01.103

54. N. Moshkov, B. Mathe, A. Kertesz-Farkas, R. Hollandi, P. Horvath, Test-time augmentation for deep learning-based cell segmentation on microscopy images, *Sci. Rep.*, **10** (2020), 5068. https://doi.org/10.1038/s41598-020-61808-3
55. M. Gaillochet, C. Desrosiers, H. Lombaert, TAAL: Test-time augmentation for active learning in medical image segmentation, *Lect. Notes Comput. Sci.*, (2022), 43–53.
56. O. Berezsky, P. Liashchynskyi, O. Pitsun, I. Izonin, Synthesis of convolutional neural network architectures for biomedical image classification, *Biomed. Signal Process. Control*, **95** (2024), 106325.
57. D. Pessoa, G. Petmezas, V. E. Papageorgiou, B. Rocha, L. Stefanopoulos, V. Kilintzis et al., Pediatric respiratory sound classification using a dual input deep learning architecture, in *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, (2023), 1–5. https://doi.org/10.1109/BioCAS58349.2023.10388733
58. J. Wu, *Introduction to Convolutional Neural Networks*, National Key Lab for Novel Software Technology, **5** (2017), 495.
59. D. Scherer, A. Müller, S. Behnke, Evaluation of pooling operations in convolutional architectures for object recognition, in *International Conference on Artificial Neural Networks*, (2010), 92–101. https://doi.org/10.1007/978-3-642-15825-4_10
60. J. C. Duchi, E. Hazan, Y. Singer, Adaptive subgradient methods for online learning and stochastic optimization, *J. Mach. Learn. Res.*, **12** (2011), 2121–2159.
61. A. Ashukha, A. Lyzhov, D. Molchanov, D. Vetrov, Pitfalls of in-domain uncertainty estimation and ensembling in deep learning, preprint, arXiv:2002.06470.
62. A. Lyzhov, Y. Molchanova, A. Ashukha, D. Molchanov, D. Vetrov, Greedy policy search: A simple baseline for learnable test-time augmentation, in *Conference on Uncertainty in Artificial Intelligence*, (2020), 1308–1317.
63. V. E. Papageorgiou, T. Zegkos, G. Efthimiadis, G. Tsaklidis, Analysis of digitalized ECG signals based on artificial intelligence and spectral analysis methods specialized in ARVC, *Int. J. Numer. Methods Biomed. Eng.*, **38** (2022), e3644. https://doi.org/10.1002/cnm.3644
64. V. Papageorgiou, Brain tumor detection based on features extracted and classified using a low-complexity neural network, *Trait. Signal*, **38** (2021), 547–554. https://doi.org/10.18280/ts.380302
65. A. Hamada, Br35H: Brain tumor detection 2020, 2020.
66. A. Mahimkar, IQ-OTH/NCCD-Lung cancer dataset, 2021.
67. O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P. A. Heng, et al., Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?, *IEEE Trans. Med. Imaging*, **37** (2018), 2514–2525. https://doi.org/10.1109/TMI.2018.2837502
68. I. Unal, Defining an optimal cut-point value in ROC analysis: An alternative approach, *Comput. Math. Methods Med.*, **2017** (2017), 3762651. https://doi.org/10.1155/2017/3762651
69. A. Kiureghian, O. Ditlevsen, Aleatory or epistemic? Does it matter?, *Struct. Saf.*, **31** (2009), 105–112. https://doi.org/10.1016/j.strusafe.2008.06.020
70. E. Hüllermeier, W. Waegeman, Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods, *Mach. Learn.*, **110** (2021), 457–506. https://doi.org/10.1007/s10994-021-05946-3

71. X. Zhou, H. Liu, F. Pourpanah, T. Zeng, X. Wang, A survey on epistemic (model) uncertainty in supervised learning: Recent advances and applications, *Neurocomputing*, **489** (2022), 449–465. https://doi.org/10.1016/j.neucom.2021.10.119
72. S. A. Singh, N. C. Krishnan, D. R. Bathula, Towards reducing aleatoric uncertainty for medical imaging tasks, in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, (2022), 1–4. https://doi.org/10.1109/ISBI52829.2022.9761638
73. J. M. Caicedo, B. A. Zarate, Reducing epistemic uncertainty using a model updating cognitive system, *Adv. Struct. Eng.*, **14** (2016), 1–12. https://doi.org/10.1260/1369-4332.14.1.55
74. A. Kendall, Y. Gal, What uncertainties do we need in Bayesian deep learning for computer vision? *Adv. Neural Inf. Process. Syst.*, **30** (2017), 5574–5584.
75. S. Ghamizi, M. Cordy, M. Papadakis, Y. Le Traon, On evaluating adversarial robustness of chest X-ray classification: pitfalls and best practices, preprint, arXiv:2212.08130.

AIMS Press