Computational and Structural Biotechnology Journal 24 (2024) 603–610

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

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

Research Article

ANN uncertainty estimates in assessing fatty liver content from ultrasound data

G. Del Corso [a,1], M.A. Pascali [a,1], C. Caudai [a,*,1], L. De Rosa [g,b], A. Salvati [c], M. Mancini [d], L. Ghiadoni [e], F. Bonino [d], M.R. Brunetto [c,d,f], S. Colantonio [a,2], F. Faita [g,2]

[a] Institute of Information Science and Technologies "A. Faedo" (ISTI) - National Research Council of Italy (CNR) - Pisa, Italy
[b] Department of Information Engineering and Computer Science, University of Trento, Trento, Italy
[c] Hepatology Unit, Pisa University Hospital, Pisa, Italy
[d] Institute of Biostructure and Bioimaging, National Research Council, Naples, Italy
[e] Emergency Medicine Unit, Department of Clinical and Experimental Medicine, University Hospital of Pisa, Pisa, Italy
[f] Department of Clinical and Experimental Medicine, Pisa University, Pisa, Italy
[g] Institute of Clinical Physiology - National Research Council of Italy (CNR) - Pisa, Italy

ARTICLE INFO

Keywords:
Bayesian uncertainty
Fatty liver content
Artificial neural networks
Ultrasound imaging
Uncertainty quantification

ABSTRACT

Background and objective: This article uses three different probabilistic convolutional architectures applied to ultrasound image analysis for grading Fatty Liver Content (FLC) in Metabolic Dysfunction Associated Steatotic Liver Disease (MASLD) patients. Steatosis is a new silent epidemic and its accurate measurement is an impelling clinical need, not only for hepatologists, but also for experts in metabolic and cardiovascular diseases. This paper aims to provide a robust comparison between different uncertainty quantification strategies to identify advantages and drawbacks in a real clinical setting.

Methods: We used a classical Convolutional Neural Network, a Monte Carlo Dropout, and a Bayesian Convolutional Neural Network with the goal of not only comparing the goodness of the predictions, but also to have access to an evaluation of the uncertainty associated with the outputs.

Results: We found that even if the prediction based on a single ultrasound view is reliable (relative RMSE [5.93%-12.04%]), networks based on two ultrasound views outperform them (relative RMSE [5.35%-5.87%]). In addition, the results show that the introduction of a "not confident" category contributes to increase the percentage of correctly predicted cases and to decrease the percentage of mispredicted cases, especially for semi-intrusive methods.

Conclusions: The possibility of having access to information about the confidence with which the network produces its outputs is a great advantage, both from the point of view of physicians who want to use neural networks as computer-aided diagnosis, and for developers who want to limit overfitting and obtain information about dataset problems in terms of out-of-distribution detection.

1. Introduction

Metabolic Dysfunction Associated Steatotic Liver Disease (MASLD) is one of leading cause of chronic liver disease [1]. If not identified and treated, it may lead to steatohepatitis, inflammation, cirrhosis and finally hepatocellular carcinoma [2,3]. Therefore, early diagnosis and continuous monitoring of the Fatty Liver Content (FLC) are essential aspects for the prevention and management of the disease. Historically, the most effective technique for determining the percentage of FLC was the liver biopsy [4], which however represented a method too invasive and laborious to be applied routinely. Accordingly, non-invasive techniques are preferred, including Magnetic Resonance imaging (MR) and Ultrasound imaging (US). MR is nowadays considered the non-invasive gold standard for quantifying FLC and it has been proven to be strongly correlated with biopsy [5]; however MR technique is very expensive and not largely/easily available, thus limiting the widespreading of the tech-

* Corresponding author.
E-mail address: claudia.caudai@isti.cnr.it (C. Caudai).
[1] GDC/MAP/CC share the first authorship in ascending order of age.
[2] SC/FF share the last authorship.

https://doi.org/10.1016/j.csbj.2024.09.021
Received 17 July 2024; Received in revised form 26 September 2024; Accepted 26 September 2024
Available online 1 October 2024

nique. US is also a non-invasive and non-ionizing technique, claimed as less sensitive in FLC assessment than MR, but extremely less expensive, and widely available in clinical settings [6]. Indeed, as US is considered highly operator dependent, a quantitative automatic system for US image analysis aimed at FLC measurement will fill a crucial clinical need.

In recent years, many quantitative indices [7,8], Machine Learning and Deep Learning techniques have been used for this purpose [9–12, 2,13–15] with promising results. With particular regards to steatosis quantification, recently few papers have been proposed [16,17], for classifying patients in steatosis severity classes. A model that can recognize a healthy case without uncertainty would already be a useful support for diagnosis. Model performance is highly dependent on the compositions of the training dataset. A model trained on a dataset with many healthy cases and few pathological cases will recognize a healthy case with little uncertainty and a pathological case with an high uncertainty, and vice versa. From this perspective it is very important that the network is able to provide an indication of the reliability of its prediction, so that the sonographer can then interpret the output with the right degree of confidence [18,19].

While most published works use the model's own implied confidence as an estimate of reliability, it has recently been shown that these values are poorly calibrated and that ad hoc techniques must be used and developed [20,21]. Possibilities include: (1) Intrusive techniques, such as Bayesian Neural Networks [22], which are expensive in terms of data requirements and computational resources, (2) Post-hoc techniques, such as Trust Score [18], and (3) Semi-intrusive techniques, such as Deep Ensemble [23] or Monte Carlo Dropout [24], which provide an approximation of the confidence interval at a reduced computational cost.

The increasing interest towards the adoption of the Bayesian approach in the medical image domain is motivated by the fact that Bayesian Convolutional Neural Networks (BCNN) can be trained on both minor and massive datasets, and that the quantification of the epistemic and aleatoric uncertainty could be used as a key analytical tool for trustworthy learning [25,26]. Although the Bayesian approach has recently been adopted for several tasks in medical imaging (e.g., classification, segmentation, registration, denoising and tumour growth prediction [27–29]), to the best of our knowledge there are no works that directly address the regression task of FLC quantification using Bayesian methods.

In this paper we aim to compare the performances of three different convolutional architectures applied to the analysis of US imaging for quantifying FLC, with the aim not only of analyzing the different predictions, but also to quantify the level of uncertainty with which these architectures produce their outputs. In this study we do not only want to observe the numerical results and from these identify the model with best performances according to the classical metrics. Instead, we are interested in understanding whether the information we can obtain from the application of different architectures can provide complementary or redundant information relating to a complex problem such as the regression and staging of chronic degenerative diseases starting from medical images.

2. Materials and methods

In this section we will describe the data acquisition protocol and the characteristics of the population, the preprocessing algorithms adopted, the Deep Learning models used for data processing, their characteristics, and the types of uncertainty that can be defined and identified using the chosen models. We will also describe the training and validation scheme and the metrics for results evaluation.

2.1. Population study

The dataset we used for the study consists of 186 patients with different degrees of FLC (characteristics reported in Table 1). The patients were enrolled at two italian hospital centers: the Hepatology Unit, University Hospital of Pisa and the IRCCS SDN Foundation of Naples. All patients underwent MR and US examinations. All US image acquisitions were performed with two standard diagnostic ultrasound systems (Pisa: LogiQ E9, GE Healthcare, Bucking-hamshire, UK; Naples: Philips iU22, Philips Healthcare, Bothell, WA, USA), both equipped with a 1.8–5 MHz convex probe. Transmit frequency was set at 3.5 MHz with a sampling frequency of the RF signal typically 4-times greater. The focal depth (as well as the depth of the field of view) depends on the body mass index and liver size of each patient. However, the acquisition protocol requires focusing at approximately 2/3 of the liver parenchyma. The receive aperture is variable with the depth of the field of view, with typical values in the range of 64 piezoelectric elements.

Table 1
Characteristics of the study population: counts and mean ± standard deviation.

	Values	Range
SEX (M:F)	99 : 87	-
AGE (yrs)	51.95 (±13.40)	[17.0-75.3]
BMI (kg/m^2)	26.27 (±4.76)	[15.28-41.70]
Fat (%)	7.13% (±9.86%)	[0.27%-50.97%]

MR imaging was performed with two different MR scanners (Pisa: Philips Ingenia 3.0 T, Philips Healthcare, Best, The Netherlands; Naples: Philips Achieva 1.5 T, Philips Healthcare, Best, The Netherlands). Two commonly adopted algorithms have been used for FLC percentage quantification from MR images: single proton magnetic resonance spectroscopy (^{1}H-MRS) [5] and proton density fat fraction MRI-PDFF [30].

Each patient was labeled with its FLC value, used as the gold standard, while US images were used to train the DL architectures. For each patient, two different projections were used to acquire US images: an intercostal or subcostal longitudinal scan with the subject in the supine/left lateral position (called HR) and an oblique subcostal scan (called AR). The two views should contain complementary information for assessing the amount of fat; the HR view decodes the level of echogenicity of the ultrasound beam both within the renal and hepatic parenchyma, while the AR view shows the complete hepatic parenchyma and the diaphragm, providing a good representation of the attenuation of the ultrasound beam within the liver parenchyma. The ultrasound clips were acquired from two clips (AR and HR) and processed by extracting all frames as gray images, centered and cropped to 360×360 pixels. For each patient, 20 frames with AR view and 20 frames with HR view were taken, for a total of 3720 images with AR view and 3720 images with HR view.

The population study showed different levels of MR-FLC, ranging from 0.27% to 50.97%. According to the categorization of steatosis severity proposed by Karlaas et al. [5], it is possible to set the cutoff discriminating between healthy and pathological at 3.12% fat percentage. Our data-set counts 112 healthy patients (fat percentage $\leq 3.12\%$) and 74 pathological subjects (fat percentage $> 3.12\%$).

2.2. Data preprocessing

The acquired images are often annotated with reference scales and colored watermarks (see Fig. 1). Furthermore, the ultrasound cones can vary in terms of surface area and width of the cone angle. To mitigate these differences so that the model does not learn meta-information, the following preprocessing steps have been implemented:

1. **Filter color:** Non-grayscale components are removed from the image and, if overlapping regions of the cone, the missing pixels are reconstructed by bicubic interpolation.
2. **Watershed with jumps:** Starting from a non-zero central seed, all non-zero image components at a maximum distance of 5 pixels from the center (e.g., 5 jumps) are preserved. The algorithm is repeated

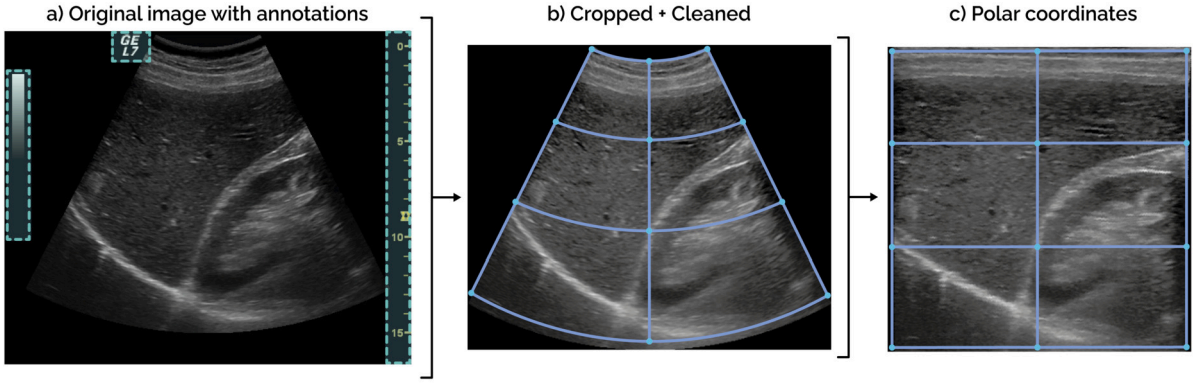

Fig. 1. Pre-processing steps: (a) a water-shed with jumps algorithm identifies and removes annotations on the original image; (b) the ultrasound cone is identified and the image is cropped; (c) the cone is transformed into polar coordinates.

until convergence, thus excluding all components that are too far from the central cone. The need to implement an ad hoc algorithm with jumps instead of a classical one is dictated by the fact that the cone does not necessarily consist of a single connected component, but rather of a plethora of gray areas separated by thin null pixel networks.

3. **Crop:** A bisection algorithm is used to determine the position of the symmetry axis and the width of the circular sector of the cone. The image is then cropped to maximize the cone size and rescaled (360x360 pixels) using bicubic interpolation (Fig. 1(b)).
4. **Final transform:** The resulting image is then transformed, using polar coordinates, to map exactly the US cone into a rectangular region (Fig. 1(c)).

2.3. Uncertainties

Uncertainty analysis in machine learning begins with the formal definition of the uncertain quantities involved in the problem (i.e., aleatoric, model, and approximation uncertainties) [31].

$$\begin{cases} \mathcal{R}(x) & = y + \varepsilon_{Al} \approx y \\ \mathcal{M}^{\infty}(x) & = y + \varepsilon_{M} \approx \mathcal{R}(x) \\ \mathcal{M}^{*}(x) & = y + \varepsilon_{Ap} \approx \mathcal{M}^{\infty}(x) \\ \hat{\mathcal{M}}(x) & = y + (\varepsilon_{Al} + \varepsilon_{M} + \varepsilon_{Ap}) = y + \varepsilon \approx y \end{cases} \tag{1}$$

With reference to the system of equations (1), given a relationship $\mathcal{R}$ that relates an input quantity x to an output quantity y, the results are usually non-deterministic due to the intrinsic randomness of the process and the random error added by measuring the quantities under investigation. This uncertainty, which cannot be reduced even by increasing the size of the data set, is called **aleatoric uncertainty** (ε_{Al}). Once the assumptions and hyperparameters characterizing the model (model type, loss, network geometry, etc.) are defined, the model is trained on a dataset D to emulate the relationship $\mathcal{R}$. The discrepancy that incurs between a model $\mathcal{M}^{\infty}$ trained on an arbitrary number of data and the relationship $\mathcal{R}$ is called **epistemic model uncertainty** (ε_{M}). This uncertainty is a product of the constructive assumptions projected by the experimenter on the problem under investigation and, although reducible, is very complex to analyze. In practice, however, the data set D has a finite size and the trained model $\mathcal{M}^{*}$ is just an approximation of $\mathcal{M}^{\infty}$. This discrepancy is called **epistemic approximation uncertainty** (ε_{Ap}) and is the simplest to reduce since it decreases as the size of the dataset D increases. The model $\hat{M}$ used in practical application, therefore, involves all the uncertainties ($\varepsilon = (\varepsilon_{Al} + \varepsilon_{M} + \varepsilon_{Ap})$) that concur in making the model's prediction inaccurate. Therefore, the methods presented here attempt not only to approximate y, but also to provide an estimate of $\varepsilon = (\varepsilon_{Al} + \varepsilon_{M} + \varepsilon_{Ap})$ in the form of an appropriate confidence interval.

2.4. Models description

To conduct the ablation study and the subsequent statistical comparison, we briefly present the three models developed to estimate FLC from AR/HR images. The starting model is a deterministic Convolutional Neural Network (CNN) with a reduced degree of freedom to prevent an overspecialization to the training set. Two probabilistic extensions are derived from this CNN: a MC Dropout CNN model and a Bayesian CNN with probabilistic output.

The models' hyperparameters (i.e., learning rate, number of layers, kernel size, etc.) have not been optimized for the given data set. In fact, reducing the difference to probabilistic components only (i.e., Dropout Layers/Distributional output) is essential for a fair comparison between models (ablation study). Therefore, as reported in Fig. 2, each network share the same hyperparameters: Kernel Size ($k = 3$), number of convolutional layers, learning rate $1e^{-4}$, and early-stopping criterion (patience $= 5$, evaluated on the validation fold to prevent model overfitting). In addition, each model is trained for a maximum of 50 epochs with a batch size of 30, using the same computational resources (CPU: Intel i7-12700; RAM: 64 GB; Graphic Card: NVIDIA RTX A4000-16 GB GDDR6). The software implementation is based on Python 3.9.16 using CUDA 11.6 for GPU acceleration and Keras (2.1), Tensorflow (2.1) and Tensorflow Probability (0.8) for probabilistic modeling.

2.4.1. Deterministic CNN

The CNN architecture is illustrated in Fig. 2. It is made up of two branches of 7 convolutional layers with Average Pooling. The first convolutional layer is followed by a Batch Normalization layer, to normalize the inputs by re-centering and re-scaling. The two branches are devoted to the paired processing of AR and HR views of the same samples. The features extracted by the convolution layers are combined in a unique flattened vector and then processed by two fully-connected layers of size 10 and 1 respectively (regression layer with linear activation), as illustrated in Fig. 2 The user-designed CNN structure is derived from a rigorous nested cross validation strategy [32].[3] The CNN architecture is able to evaluate the geometric and texture features of images very well, estimating distances, proportions and intensity ratios of the gray scales of the various elements of the image. The preprocessing strategy (illustrated in Fig. 1 and described in Section 2.2) has been studied in depth to try to minimize the acquisition biases, providing the network with

[3] A grid search on 27 different models (kernel size 3-5-7; number of layers 5-7-8; multiplier on number of nodes 1-2-3) was applied to the internal 4-folds, while the external 5-folds are kept as an unbiased estimate of the generalization capability (540 trained submodels). The best performing model is retained (kernel 3, layers 7, nodes 2), but it should be emphasized that there is no statistically significant difference between the performance of the models ($\alpha = 5\%$, paired t-test at patient level).

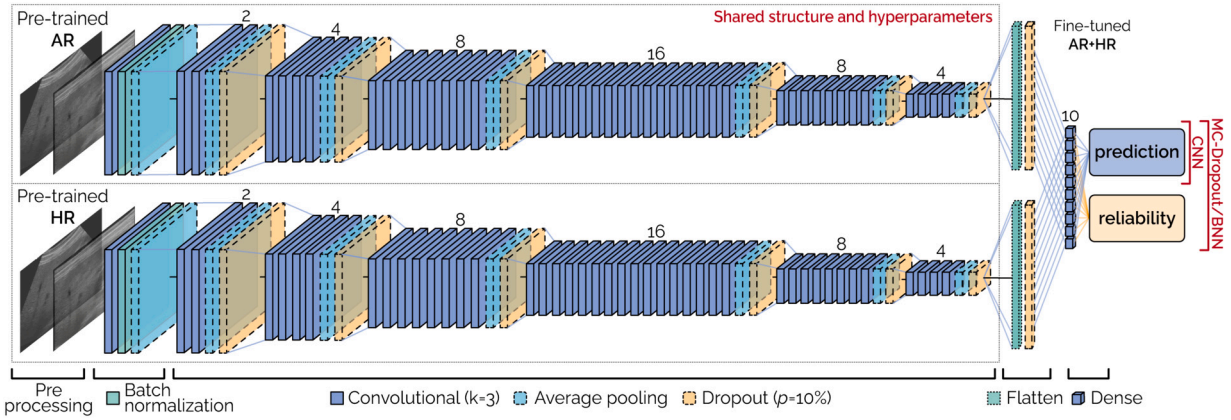

Fig. 2. Scheme of the three models that are used for the evaluation of the degree of steatosis from the AR and HR images. The standard model (deterministic CNN) takes as input AR and HR images and, after a pre-processing phase described in Section 2.2, it applies: a batch normalization, 6 convolutional layers (2,4,8,16,8,4) followed by an average pooling. The AR and HR are pre-trained and then combined (flatten layer followed by a dense one) which is fine-tuned to produce a deterministic prediction (regression). The other two models (MC-Dropout and Bayesian Neural Network) have the same structure and hyperparameters, but provide two outputs: a prediction (regression) and a reliability score. The MC-Dropout includes a dropout layer ($p = 10\%$) after each average pooling and before the combined dense layer, and during inference uses these layers to define several predictions whose variability is a measure of reliability. Instead, the Bayesian model has two outputs instead of one (mean and standard deviation), which when combined with an ad hoc loss, allows both a prediction and the corresponding reliability score to be trained directly.

images that decode in the most homogeneous way possible the information suitable for the CNN elaboration.

2.4.2. Monte Carlo dropout

A dropout model consists of a standard Artificial Neural Network which allows to randomly delete selected neurons. Dropout has been developed as a regularization technique (the so-called Dropout Regularization) to avoid overfitting of the models [33]. In fact, at each training epoch, the dropout layer sets each neuron to 0 with probability p (Bernoulli noise) and applies backpropagation to a pruned version of the ANN. The pruned neurons change randomly at each epoch, allowing all weights to be trained after sufficient time. This strategy avoids premature convergence to a local minimum and thus reduces the risk of overfitting at the cost of an increased number of epochs before achieving convergence [34].

However, once the model is fully trained, dropout can still be used as a strategy to generate confidence estimates of the results. Indeed, we can define a Forward Dropout Estimate (FDE) as a prediction made by the model after some neurons have randomly dropped out. While the performance of a single FDE tends to underestimate the regression value and generally produces a low quality prediction, by applying the estimate several thousand times, the average regressor can outperform the original [35]. This procedure mimics the training of a Deep Ensamble [23] (i.e., several variations of the same model to produce both a point estimate and a confidence estimate, reported as the interval between the first and third quartiles [Q1-Q3]) without increasing the computational training burden.

Referring to Fig. 2, we added a dropout layer (with probability $p = 10\%$ [35]) to our models after each average pooling and before the last dense layer. After the training, we collected the output by applying 10^3 FDE for the given input. The Dropout models results are reported as the mean output and the confidence estimate [Q1-Q3]. This estimate is related to the differences in predictions between different dropout models and thus approximates epistemic uncertainty.

2.4.3. Bayesian CNN

Bayesian neural networks (BNNs) are a reformulation of standard neural networks from a probabilistic point of view designed to investigate both aleatoric and epistemic uncertainties [36]. Furthermore, BNNs have a reduced risk of overfitting compared to their deterministic counterparts [37]. The paradigm underlying the development of Bayesian networks is the application of Bayes' formula to describe the distribution of parameters characterizing the model based on the available data:

$$p(\theta|D) = \frac{p(D|\theta) \cdot p(\theta)}{p(D)} \tag{2}$$

where D is the available data (training set), θ are the trainable model parameters (e.g., weights), $p(\theta|D)$ is the **posterior distribution** (i.e., the optimal distribution of the model parameters once the training dataset is fixed), $p(\theta)$ is the **prior distribution** of the model parameters, and $p(D) = \int_\Theta p(\theta) \cdot p(D|\theta)\, d\theta$ is the **marginal likelihood/evidence**.

The difference from a deterministic network is that θ is a random variable and not a deterministic value (uniquely determined based on the characteristics of the training set). In practice, a subset of the neural network deterministic weights are replaced by appropriate distributions. These distributions, initialized on the basis of *a priori* knowledge $p(\theta)$, are updated on the basis of the available data D through a process called Bayesian inference, which can be pursued using several approaches, including Monte Carlo methods [38] and Variational Inference [39].

These networks integrate epistemic uncertainty because each prediction $p(y|x,\theta)$ is a realization of the random variable $y|x, D$ (the output conditioned the input and the available data) given by randomly sampling the posterior distribution network weights.

Probabilistic BNN further extends this uncertainty estimation by replacing a point output with a distributional one. The probabilistic output also allows the modeling of aleatoric uncertainty, providing a complete view of model reliability.

As a remark, fully disentangling the source of uncertainty is difficult and often unnecessary [40]. A distributional regression output requires an appropriate loss function. In this work, we chose to use a **negative Log-Likelihood loss**.

Referring to Fig. 2, we defined a probabilistic BNN during fine-tuning. Therefore, we added just one last Bayesian layer and a distributional Gaussian output, whose a priori distribution is still a Gaussian (with mean 0 and variance 1). The model thus returns a prediction (the mean μ) along with a confidence interval ($[\text{Q1-Q3}] = [\mu - 0.675 \cdot \sigma; \mu + 0.675 \cdot \sigma]$) that includes both the aleatoric and epistemic uncertainty components.

2.5. Training and validation scheme

The training and validation scheme implemented is a rigorous 5-fold cross-validation: each time 3 sets were used for training, one to

Table 2
Evaluation metrics: Median and IQR (reported as quartiles Q1-Q3) of (normalized) RMSE, coefficient of variation, and epochs of convergence between the 5 folds.

Classic CNN	AR	HR	Combined
RMSE Training	1.31% [1.09-2.35]	2.29% [1.90-2.86]	1.76% [1.73-1.85]
RMSE Test	6.91% [5.31-6.96]	**7.24% [6.63-7.70]**	5.87% [5.80-6.79]
Convergence Epochs	**12 [9-14]**	10 [7-14]	**11 [7-23]**
MC Dropout	AR	HR	Combined
RMSE Training	1.99% [1.87-3.09]	2.50% [1.91-2.54]	1.45% [1.38-1.69]
RMSE Test	**5.93% [5.33-5.94]**	7.38% [6.76-7.71]	**5.35% [5.23-6.40]**
CoV Training	0.33 [0.28-0.35]	0.40 [0.32-0.46]	0.32 [0.31-0.38]
CoV Test	**0.29 [0.28-0.34]**	**0.32 [0.28-0.35]**	**0.32 [0.29-0.37]**
Convergence Epochs	14 [9-15]	**9 [8-14]**	19 [12-25]
Bayesian CNN	AR	HR	Combined
RMSE Training	5.66% [5.30-5.83]	6.40% [5.56-6.43]	4.82% [4.53-5.87]
RMSE Test	12.04% [11.30-12.11]	11.92% [11.56-12.27]	**5.82% [5.03-6.76]**
CoV Training	0.55 [0.48-0.74]	0.56 [0.46-0.67]	0.55 [0.46-0.68]
CoV Test	0.53 [0.44-0.60]	0.55 [0.42-0.56]	0.50 [0.44-0.57]
Convergence Epochs	16 [12-21]	16 [12-29]	29 [17-30]

implement early stopping, and one to obtain a low bias assessment of the generalization ability. To ensure homogeneity between the 5 folds, patients are stratified according to their FLC. In addition, to prevent data leakage between the folds, and thus positively biased results, there is further stratification by patient (i.e. all images relating to a patient are in the same fold). The same validation scheme is applied to each architecture (Ablation Study), so when a given patient is evaluated in the test set, each model has seen exactly the same data to make the prediction.

To avoid over-specialization on the training set (thus reducing the generalization ability of the models), we used a two-step training procedure (Fig. 2). For each test fold, we trained two sub-models on only one view (AR/HR). These sub-models were trained on the same 3 training folds to avoid data contamination on the test fold. After training the sub-models, their weights were fixed, the convolutional parts of the two networks were merged, and the combined model was fine-tuned (training only the dense layers) on the same 3 folds and evaluated on the unseen test fold.

2.6. Evaluation metrics

The comparison between deterministic (CNN) and probabilistic (MC Dropout/Bayesian CNN) models requires several ad hoc evaluation metrics. These scores are calculated for each fold and then reported as the median (over the 5 folds) and the corresponding [Q1-Q3] interquartile range.

2.6.1. Regression metric

We adopted the normalized Root-Mean-Square Error as the main regression metric:

$$\mathrm{RMSE} = c \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{3}$$

where n is the number of data (training, validation, or test), (x_i, y_i) the views and actual value, $c = 1/(100\% - 0\%)$ (ranges of FLC) is the normalization coefficient, and $\hat{y}_i = \mathcal{M}(x_i)$ is the prediction.

For the deterministic CNN, $\hat{y}$ is given directly by the predicted value provided by the model. For probabilistic models that return a distribution instead of a point estimate, $\hat{y}$ is obtained as the average of point estimates over 10^3 runs. As a remark, the distribution over the weights for the Bayesian CNN and the random pruning induced by the dropout layers imply that each of these runs yields a slightly different prediction.

2.6.2. Uncertainty metric

To estimate the amount of uncertainty estimated by the probabilistic models (MC Dropout/Bayesian CNN), we introduced the Coefficient of Variation (CoV). CoV is defined as the average over each prediction $\hat{y}_i$ as the ratio between the estimated standard deviation ($\sigma(\hat{y}_i)$) and the corresponding output mean ($\mu(\hat{y}_i)$):

$$\mathrm{CoV} = \frac{1}{n}\sum_{i=1}^{n}\frac{\sigma(\hat{y}_i)}{\mu(\hat{y}_i)} \tag{4}$$

2.6.3. Convergence rate metric

Models which implement probabilistic components are usually characterized by a lower convergence rate. Therefore, for each model $\mathcal{M}$ and each fold, we calculated the number of epochs necessary to fulfill the early stopping criterion.

2.6.4. Bland Altman analysis

Bland-Altman analysis was performed to compare the assessment of FLC by neural networks against MR gold standard. Bias and Intervals of Agreement were evaluated.

3. Results

3.1. Regression results

Regression results, summarized in Table 2, show good prediction performance for all architectures on the 5-fold test sets (Normalized RMSE 5.87%, 5.35%, and 5.82% for Combined CNN, MC Dropout, and Bayesian CNN respectively).

Classic CNN and MC Dropout show a higher risk of overfitting (Normalized RMSE 1.76% and 1.45% on training data, respectively) compared to Bayesian CNN (4.82%), demonstrating their increased ability to avoid overspecialization even on medium-size databases.

The probabilistic models provide a confidence estimate of the uncertainty associated with the prediction. As expected, both single-view (0.53/0.55 for AR/HR) and combined (0.50) Bayesian CNN have higher CoV values compared to MC Dropout (0.33/0.29/0.32 AR/HR/Combined respectively). Indeed, Bayesian CNN approximates both epistemic and aleatoric uncertainty and hence requires more training data, while MC Dropout focuses mainly on epistemic uncertainty.

The advantage to estimate uncertainties is balanced by the increasing computational cost. Indeed, the median number of epochs to reach convergence ranges from 11 for Combined CNN to 19 for MC Dropout, and 29 for Combined Bayesian CNN. Given a comparable training time per epoch, this means that adding uncertainties requires 2/3 the computational effort of its deterministic counterpart. However, since dropout

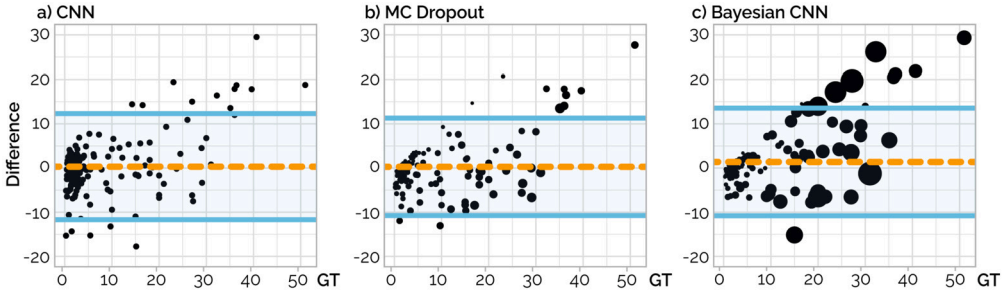

Fig. 3. Bland Altman plots of the three different models against the ground truth (GT). The graph shows the mean difference and the 95% CI.

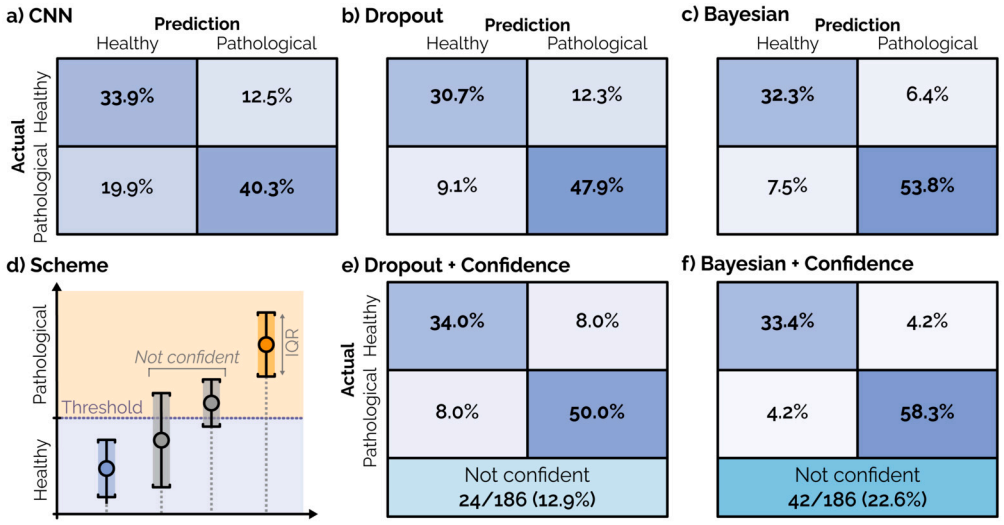

Fig. 4. Confusion matrices of the three architectures used. a) Classical CNN Confusion matrix. b) e) Confusion matrices of the MC Dropout, respectively without and with the "Non Confident" category. c) f) Confusion matrices of the Bayesian CNN, respectively without and with "Non Confident" category. d) Legenda for e) and f) Figures: all outputs for which the confidence band overlaps the value of 3.12 (threshold value of fat percentage between healthy and pathological subjects) are defined "Non Confident". All outputs for which the confidence band is entirely below the value of 3.12 are defined as Healthy. The outputs for which the confidence band is entirely above the value of 3.12 are defined as Pathological.

layers are usually also used as a regularization strategy, this cost is usually handled implicitly.

Another hidden cost of using probabilistic approaches is that the non-deterministic output requires thousands of simulations. While the computation time is negligible compared to the training time, in some practical applications this cost (approximately 1 minute for probabilistic networks, compared to less than 10^{-1} seconds for deterministic CNN) can influence the choice of model.

The combined information content of AR and HR images leads to models with slightly better results than those trained on one view alone, particularly for Bayesian CNN (12.04/11.92 AR/HR RMSE vs 5.82 combined RMSE). As a remark, MC dropout proves to be the most efficient approach to obtain valuable 1-view-only models (5.93/7.38 AR/HR RMSE).

The Bland-Altman plots (Fig. 3) describe the differences between the predictions of the three architectures and the ground truth. The uncertainty of the predictions is represented by the radius of the points. Bayesian CNN provides predictions with greater uncertainty compared to MC Dropout and CNN (point estimate), especially for subjects with more advanced pathology. In addition, the predictions provided by the three methods are not affected by any bias (the mean values are centered close to zero).

3.2. Classification results

The benefit of having estimated uncertainties can be appreciated especially when applied to a related classification problem. In fact, in computer-aided diagnosis, the main objective is often to distinguish between the healthy class (percentage of fat $\leq 3.12\%$) and the pathological one.

Referring to Fig. 4(a,b,c), the three models prove to successfully identify pathological patients even with a medium-sized database (accuracy 74.2%, 78.6%, 86.1% for CNN, MC Dropout and Bayesian CNN, respectively). Probabilistic networks are more effective at correctly identifying false negatives (pathological patients classified as healthy). Indeed, the percentage of false negatives improves from 19.9% for deterministic CNN to 9.1% for MC Dropout and only 7.5% for Bayesian CNN.

The confidence estimate can be used to define a further non-confident class. In computer-aided diagnosis, patients classified as not confident are those who require more detailed analysis and further (human) expert evaluation. In this study, we assign a patient to the non-confident class if the confidence interval ([Q1-Q3]) overlaps the medical threshold of 3.12%, see Fig. 4(d).

Referring to Fig. 4(e,f), adding the non-confident class to the MC Dropout improves model performance (accuracy increases from 78.6% to 84%) at the cost of a limited number of patients requiring further monitoring (12.9%), demonstrating the strength of this approach. Similarly, the accuracy of Bayesian CNN increases from 86.1% to 91.7%, with most of the misclassified cases correctly identified as not confident. However, several well classified patients are erroneously assigned to the not confident class (22.6% of the total dataset requires further monitoring).

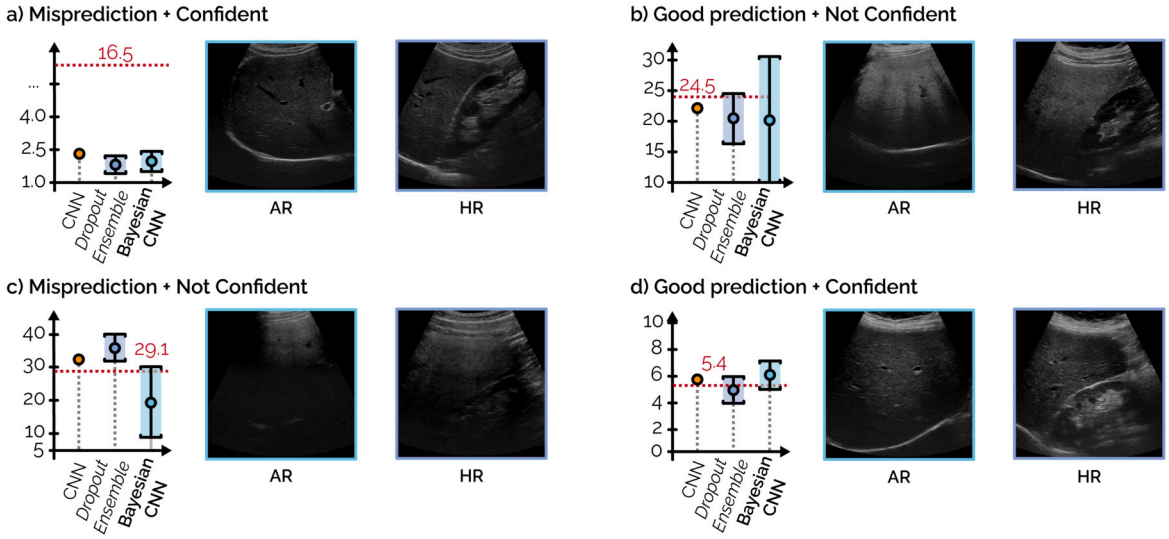

Fig. 5. Comparison of predictions and level of uncertainty for 4 cases with different percentage of FLC. a) Case predicted with low precision and small amount of uncertainty. b) Case predicted with good precision and big amount of uncertainty. c) Case predicted with low precision and big amount of uncertainty. d) Case predicted with good precision and small amount of uncertainty.

4. Discussions and conclusions

In this work, we developed a systematic comparison between deterministic CNN, MC dropout, and Bayesian CNN for regression and classification of fatty liver content (FLC) on a novel and multicentric dataset including AR and HR views of 186 patients. The purpose of this comparison was to determine the optimal strategy for FLC values regression, not only in terms of prediction, but also in terms of result reliability.

We found that CNN-based approaches show a good capability to predict FLC both for regression (CNN, Dropout, and Bayesian RMSE of 5.87%, 5.35%, and 5.82%, respectively) and classification (74.2%, 78.6%, 86.1% of correctly classified patients). It is interesting to note that predictions based on AR/HR only views still have a good predictive capability for simpler models (CNN, Dropout), while the most complex one (Bayesian CNN) requires fine tuning to avoid overfitting phenomena.

Overall, these results show that the US-based artificial intelligence-calculated FLC is a reliable method in good agreement with gold standard MR assessment, thus suggesting its adoption in Point Of Care Ultrasound (POCUS) applications or in supporting sonographers with relatively limited experience in liver analysis.

Regarding reliability, MC Dropout proves to provide smaller confidence interval (CoV 0.32) compared to Bayesian CNN (CoV 0.50). By using those intervals to identify unreliable predictions, MC Dropout classifies 12.9% of the patients as unreliable, thus reducing the misclassified cases from 21.4% to 16%. Bayesian CNN further improves this reduction with only 8.4% of misclassified cases, but at the cost of an increased amount of patients classified as unreliable (22.6%). However, the addition of a reliability score significantly increases the median number of epochs to convergence (11, 19, and 29 for CNN, Dropout, and Bayesian, respectively).

Providing levels of credibility for AI decisions could facilitate the adoption of AI systems, as physicians can more readily accept AI results when they are accompanied by high reliability score. However, it should be recognized that the AI-generated reliability score should only suggest that a re-evaluation by a human expert is requested. Translating the contribution of AI architectures into clinical practice remains challenging due to inadequate levels of explainability and interpretability of most of machine learning tools. However, the addition of reliability scores allows AI-based medical devices to be used as a valid diagnostic aid under human supervision in the decision-making and interpretation phases.

Nevertheless, the clinician's interpretation during the routine objective ultrasound examination of the liver would be eased particularly in the screening of FLC when there is less than 20% of fat.

As a consequence, while adding a reliability score proves to be an efficient way to improve model predictions [18,19], it should be done considering the significant increase in computational cost and the risk of classifying too many cases as unreliable, especially for Bayesian CNN.

Referring to Fig. 5(d), most cases are correctly predicted (with a small confidence interval). These cases are usually characterized by a clear and well-defined image with a low amount of artifacts. Conversely, there are images that show a particularly low contrast or a high level of noise (e.g., US images with not well-defined kidney contours and/or with diaphragm barely visible), which the model correctly identifies as an uncertain prediction, allowing the intervention of a physician to further evaluate the case (see Fig. 5(c)).

The main limitation of this work is the amount of data available, which prevents probabilistic fine-tuning and thus the use of pre-trained state-of-the-art networks (such as VGG, AlexNet, etc.) or more complex models (attention/recursion mechanism [41] or explainable machine learning). Further studies will focus on extending the dataset (increasing both size and variability, collecting data from multiple centers) to fully unleash the power of Bayesian CNN, and comparing deep approaches with classical ones, integrating more advanced model and attention-based analyses [9–11,2,14,15]. Furthermore, we will compare these techniques with reliability scores produced by post-hoc methods, such as Topological UQ and Trust Score [18].

Declarations

Ethical approval

The study protocol was approved by the Ethic Committee of the University Hospital of Pisa (19/02/2016) in accordance with the Declaration of Helsinky. All the subjects signed a written informed consent. Protocol n. 1179/2016.

Code availability

The pre-processing code is freely available under the MIT license (https://github.com/GDelCorso/EchoLocator) [42]. A technical report

describing the use of the code and detailing the preprocessing step can be downloaded at https://hdl.handle.net/20.500.14243/495122.

CRediT authorship contribution statement

G. Del Corso: Writing – review & editing, Software, Methodology, Formal analysis, Conceptualization. **M.A. Pascali:** Writing – review & editing, Software, Conceptualization. **C. Caudai:** Writing – review & editing, Software, Methodology, Formal analysis, Conceptualization. **L. De Rosa:** Writing – review & editing, Formal analysis, Data curation. **A. Salvati:** Writing – review & editing, Data curation. **M. Mancini:** Writing – review & editing, Formal analysis. **L. Ghiadoni:** Writing – review & editing, Project administration, Funding acquisition. **F. Bonino:** Writing – review & editing, Conceptualization. **M.R. Brunetto:** Writing – review & editing, Project administration, Funding acquisition. **S. Colantonio:** Writing – review & editing, Conceptualization. **F. Faita:** Writing – review & editing, Project administration, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Data availability

The informed consent signed by patients enrolled in this study, did not explicitly allow the sharing of echographic data (even if anonymized). Trained models can be provided upon request as long as the privacy conditions are met.

References

[1] Rinella Mary E, Lazarus Jeffrey V, Ratziu Vlad, Francque Sven M, Sanyal Arun J, Kanwal Fasiha, et al. A multisociety delphi consensus statement on new fatty liver disease nomenclature. J Hepatol 2023;79(6):1542–56.

[2] Han A, Byra Michal, Heba E, Andre M, Erdman J, Loomba R, et al. Noninvasive diagnosis of nonalcoholic fatty liver disease and quantification of liver fat with radiofrequency ultrasound data using one-dimensional convolutional neural networks. Radiology 2020:191160.

[3] Loomba R, Sanyal A. The global nafld epidemic. Nat Rev Gastroenterol Hepatol 2013;10:686–90.

[4] Bravo AA, Sheth S, Chopra S. Liver biopsy. N Engl J Med 2001;344(7):495–500.

[5] Karlas T, Petroff D, Garnov N, Böhm S, Tenckhoff H, Wittekind C, et al. Non-invasive assessment of hepatic steatosis in patients with nafld using controlled attenuation parameter and 1h-mr spectroscopy. PLoS ONE 2014;9.

[6] Mancini M, Prinster A, Annuzzi G, Liuzzi R, Giacco R, Medagli Carmela, et al. Sonographic hepatic-renal ratio as indicator of hepatic steatosis: comparison with (1) h magnetic resonance spectroscopy. Metab Clin Exper 2009;58(12):1724–30.

[7] Di Lascio N, Avigo C, Salvati A, Martini N, Ragucci M, Monti S, et al. Non-invasive quantitative assessment of liver fat by ultrasound imaging. Ultrasound Med Biol 2018;44(8):1585–96.

[8] De Rosa Laura, Salvati Antonio, Martini Nicola, Chiappino Dante, Cappelli Simone, Mancini Marcello, et al. An ultrasound multiparametric method to quantify liver fat using magnetic resonance as standard reference. Liver international: official journal of the International Association for the Study of the Liver, 2024.

[9] Biswas Mainak, Kuppili Venkatanareshbabu, Edla D, Suri Harman S, Saba L, Marinho R, et al. Symtosis: a liver ultrasound tissue characterization and risk stratification in optimized deep learning paradigm. Comput Methods Programs Biomed 2018;155:165–77.

[10] Byra Michal, Styczynski G, Szmigielski C, Kalinowski P, Michalowski L, Paluszkiewicz R, et al. Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images. Int J Comput Assisted Radiol Surg 2018;13:1895–903.

[11] Cao Wen, An Xing, Cong L, Lyu Chaoyang, Zhou Qian, Guo R. Application of deep learning in quantitative analysis of 2-dimensional ultrasound imaging of nonalcoholic fatty liver disease. J Ultrasound Med 2019;39.

[12] Cowin G, Jonsson JR, Bauer Judith, Ash S, Ali A, Osland E, et al. Magnetic resonance imaging and spectroscopy for monitoring liver steatosis. J Magn Reson Imaging 2008;28.

[13] Colantonio Sara, Salvati Antonio, Caudai Claudia, Bonino Ferruccio, De Rosa Laura, Pascali Maria Antonietta, et al. A deep learning approach for hepatic steatosis estimation from ultrasound imaging. In: International conference on computational collective intelligence; 2021.

[14] Popa Stefan L, Ismaiel Abdulrahman, Cristina Pop, Cristina Mogosan, Chiarioni Giuseppe, David Liliana, et al. Non-alcoholic fatty liver disease: implementing complete automated diagnosis and staging. A systematic review. Diagnostics 2021;11(6).

[15] Reddy DS, Bharath R, Rajalakshmi P. A novel computer-aided diagnosis framework using deep learning for classification of fatty liver disease in ultrasound imaging. In: 2018 IEEE 20th international conference on e-health networking, applications and services (healthcom); 2018. p. 1–5.

[16] Jeon Sun Kyung, Lee Jeong Min, Joo Ijin, Yoon Jeong Hee, Lee Gunwoo. Two-dimensional convolutional neural network using quantitative us for noninvasive assessment of hepatic steatosis in nafld. Radiology April 2023;307(1).

[17] Larocque-Rigney Boustros P, Patry-Beaudoin C, Luo L, Aslan YH, Marinos E, Alamri J, et al. Comparison of radiologists and deep learning for us grading of hepatic steatosis. Radiology 2023;309.

[18] Jiang Heinrich, Kim Been, Guan Melody, Gupta Maya. To trust or not to trust a classifier. Adv Neural Inf Process Syst 2018;31.

[19] Varshney Kush R, Alemzadeh Homa. On the safety of machine learning: cyber-physical systems, decision sciences, and data products. Big Data 2017;5(3):246–55.

[20] Kuleshov Volodymyr, Liang Percy S. Calibrated structured prediction. Adv Neural Inf Process Syst 2015;28.

[21] Guo Chuan, Pleiss Geoff, Sun Yu, Weinberger Kilian Q. On calibration of modern neural networks. In: International conference on machine learning. PMLR; 2017. p. 1321–30.

[22] MacKay David JC. Bayesian neural networks and density networks. Nucl Instrum Methods Phys Res, Sect A, Accel Spectrom Detect Assoc Equip 1995;354(1):73–80.

[23] Lakshminarayanan Balaji, Pritzel Alexander, Blundell Charles. Simple and scalable predictive uncertainty estimation using deep ensembles. Adv Neural Inf Process Syst 2017;30.

[24] Hara Kazuyuki, Saitoh Daisuke, Shouno Hayaru. Analysis of dropout learning regarded as ensemble learning. In: Artificial neural networks and machine learning–ICANN 2016: 25th international conference on artificial neural networks. Proceedings, part II, vol. 25. Springer; 2016. p. 72–9.

[25] Jospin Laurent Valentin, Laga Hamid, Boussaid Farid, Buntine Wray, Bennamoun Mohammed. Hands-on Bayesian neural networks—a tutorial for deep learning users. IEEE Comput Intell Mag 2022;17(2):29–48.

[26] Magris Martin, Iosifidis Alexandros. Bayesian learning for neural networks: an algorithmic survey. Artif Intell Rev 2023;56(10):11773–823.

[27] Zou Ke, Chen Zhihao, Yuan Xuedong, Shen Xiaojing, Wang Meng, Fu Huazhu. A review of uncertainty estimation and its application in medical imaging. Meta-Radiology 2023;1(1):100003.

[28] Jeon Yeseul, Chang Won, Jeong Seonghyun, Han Sanghoon, Park Jaewoo. A Bayesian convolutional neural network-based generalized linear model. Biometrics 2022;80(2).

[29] Lee Sungyoon, Kim Hoki, Lee Jaewook. Graddiv: adversarial robustness of randomized neural networks via gradient diversity regularization. IEEE Trans Pattern Anal Mach Intell 2022.

[30] Reeder Scott B, Hu Houchun H, Sirlin Claude B. Proton density fat-fraction: a standardized mr-based biomarker of tissue fat concentration. J Magn Reson Imaging 2012;36.

[31] Hüllermeier Eyke, Waegeman Willem. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach Learn 2021;110:457–506.

[32] Stone Mervyn. Cross-validatory choice and assessment of statistical predictions. Journal of the royal statistical society: Series B (Methodological) 1974;36(2):111–33.

[33] Srivastava Nitish, Hinton Geoffrey, Krizhevsky Alex, Sutskever Ilya, Salakhutdinov Ruslan. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15(1):1929–58.

[34] Baldi Pierre, Sadowski Peter. The dropout learning algorithm. Artif Intell 2014;210:78–122.

[35] Gal Yarin, Ghahramani Zoubin. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: International conference on machine learning. PMLR; 2016. p. 1050–9.

[36] MacKay David JC. A practical Bayesian framework for backpropagation networks. Neural Comput 1992;4(3):448–72.

[37] Hoeting Jennifer A, Madigan David, Raftery Adrian E, Volinsky Chris T. Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and E.I. George, and a rejoinder by the authors). Stat Sci 1999;14(4):382–417.

[38] Gamerman Dani, Lopes Hedibert F. Markov chain Monte Carlo: stochastic simulation for Bayesian inference. CRC Press; 2006.

[39] Blei David M, Kucukelbir Alp, McAuliffe Jon D. Variational inference: a review for statisticians. J Am Stat Assoc 2017;112(518):859–77.

[40] Der Kiureghian Armen, Ditlevsen Ove. Aleatory or epistemic? Does it matter? Struct Saf 2009;31(2):105–12.

[41] Buongiorno Rossana, Del Corso Giulio, Germanese Danila, Colligiani Leonardo, Python Lorenzo, Romei Chiara, et al. Enhancing covid-19 ct image segmentation: a comparative study of attention and recurrence in unet models. J Imag 2023;9(12):283.

[42] Del Corso Giulio, De Rosa Laura, Pascali Maria Antonietta, Faita Francesco, Colantonio Sara. Echolocator: an open source python package for the standardisation of echographic images in multicentre analysis.