Introduction

Medical imaging, particularly chest radiographs, plays a crucial role in diagnosing various respiratory diseases, including pneumonia and COVID-19. Traditional deep learning methods, while effective in classifying known conditions, often struggle when confronted with unseen or out-of-distribution (OOD) data, such as new disease variants or images with subtle transformations. This is where Bayesian Neural Networks (BNNs) offer a promising advantage, as they allow for the explicit modeling of uncertainty in predictions.

Bayesian Neural Networks are a class of models that incorporate uncertainty by treating the network weights as distributions rather than fixed values. This approach provides two types of uncertainty: epistemic uncertainty, which reflects uncertainty in the model parameters due to insufficient or limited training data, and aleatoric uncertainty, which accounts for the inherent noise in the data itself. The ability to quantify epistemic uncertainty is particularly valuable in medical image analysis, where unseen data or novel disease variants may pose challenges for accurate classification.

This study investigates the role of uncertainty in the classification of lung radiographs using a Bayesian ResNet model, replacing the last layer with a Bayesian layer. ResNet was chosen for its strong performance in medical imaging tasks due to its ability to learn deep hierarchical features effectively while mitigating vanishing gradient issues through residual connections (He et al., 2015). The radiographs used in this study were downloaded from the COVID-QU-Ex dataset (Tahir et al., 2022) which contains 10,701 normal, 11,263 non-COVID and 11,956 COVID-19 chest X-ray images. Example images are shown in Fig. 1.

The Bayesian ResNet model was trained to distinguish between normal chest radiographs and those affected by non-COVID infections. COVID-19 radiographs were not used in model training. The goal of the study was to investigate whether uncertainty analysis, derived from variational inference, could effectively identify COVID-19 cases as well as normal radiographs subjected to out-of-distribution (OOD) transformations like rotations, cropping and color jitter. By analyzing epistemic uncertainty and entropy, the study aimed to assess the model's ability to flag data that deviated from the training distribution, providing insights into its reliability in detecting novel or altered inputs.

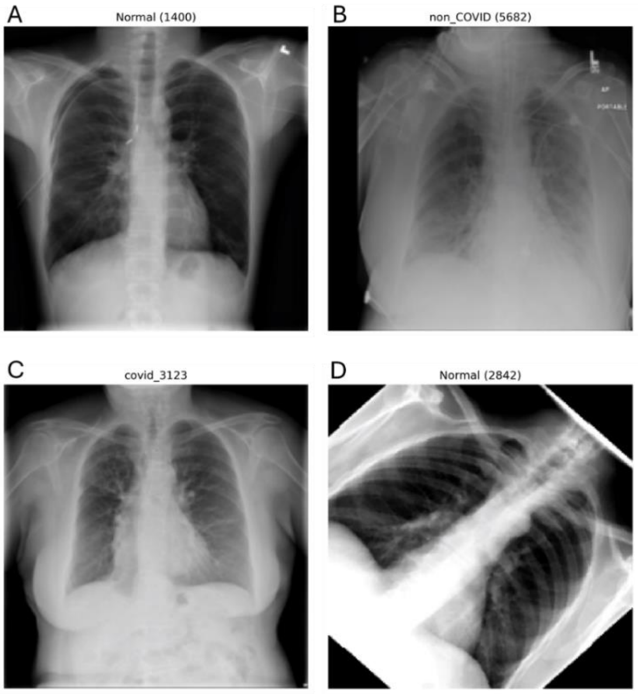

Fig. 1: Example chest X-ray images from the COVID-QU-Ex dataset. A (normal) and B (non-COVID) are images from classes used in model training. C (COVID-19) and D (normal images subjected to OOD transformations) were treated as OOD samples for uncertainty analysis.

Variational inference in Bayesian neural networks

The goal of Bayesian inference in the context of neural networks is to find the posterior distribution of weights θ given the data $D = \{(x_i, y_i)\}_{i=1}^{N}$:

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

Where:

- $p(D|\theta)$: likelihood of the data given the parameters (how well the model explains the data).
- $p(\theta)$: prior distribution of parameters.
- $p(D)$: evidence or marginal likelihood, representing the probability of observing the data under the entire model, calculated as an integral over the entire parameter space.

In practice, the exact computation of the complex posterior is intractable due to the normalizing factor $p(D)$. Variational inference offers an approach to approximate the posterior through optimization during network training.

Instead of $p(\theta|D)$, a simple parametrized distribution $q(\theta)$ is initialized, e.g. a Gaussian with mean μ and variance σ^2 referred to as variational parameters. The aim is to make $q(\theta)$ as close as possible to $p(\theta|D)$ by maximizing the evidence lower bound (ELBO):

$$ELBO = E_{q(\theta)}[\log p(D|\theta)] - KL(q(\theta)||p(\theta))$$

Where:

- $E_{q(\theta)}[\log p(D|\theta)]$: expected log likelihood under the approximate posterior where $p(D|\theta) = \sum_{i=1}^{N} p(y_i|x_i, \theta)$.
- $KL(q(\theta)||p(\theta))$: Kullback-Leibler divergence between the approximate posterior and prior of the variational parameters. As a weighted term it can be used as a regularizer to encourage the variational parameters to stay close to the prior distribution.

For a classification task the expected log likelihood is the negative of cross-entropy loss:

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N} logp(y_i|x_i, \theta) = -E_{q(\theta)}[\log p(D|\theta)]$$

Thus, the variational parameters can be found by minimizing the negative of ELBO, combining CE loss and KL divergence (weighted by hyperparameter α) in a loss function during network training on the classification task:

$$L = -ELBO = L_{CE} + \alpha KL(q(\theta)||p(\theta))$$

Each weight in the Bayesian component of the network is assigned variational parameters (such as mean and variance). During training, the Bayesian weights are sampled for each forward pass (for each minibatch) and backpropagation updates the variational parameters of the approximate posterior.

The goal of inference is to compute the posterior predictive distribution over classes for new output y^* given new data x^* by integrating the product of likelihood of new output (softmax model output given a set of parameters) and posterior distribution of parameters over all possible values of parameters:

$$p(y^*|x^*, D) = \int p(y^*|x^*, \theta)p(\theta|D)d\theta$$

Given that the integral is intractable due to the high dimensionality of θ and the true posterior $p(\theta|D)$ is unknown, the predictive distribution can be found by Monte Carlo sampling $\theta_1, \ldots, \theta_T$ from the approximate posterior $q(\theta)$ (for which parameters were determined during training) and performing a forward pass of the model for each θ_t. Then the posterior predictive distribution can be approximated as:

$$p(y^*|x^*, D) \approx \frac{1}{T}\sum_{t=1}^{T} p(y^*|x^*, \theta_t)$$

$$\theta_t \sim q(\theta)$$

Uncertainty in Bayesian neural networks

Prediction uncertainty can be broken down into epistemic and aleatoric components. Epistemic uncertainty refers to the lack of knowledge about the best model parameters and can be caused by insufficient training data or lack of diversity in the training set. On the other hand, aleatoric uncertainty is inherent in the data itself, caused by stochasticity or noise that cannot be explained by the model and it cannot be reduced with more data.

Total uncertainty can be quantified with entropy which measures the uncertainty in the model's average prediction. For a BNN an average prediction is the mean of the softmax outputs over all forward passes for a given image while the predicted class is the one with the highest predicted probability. A high entropy indicates a high spread across multiple classes (uncertain predictions). The metric can be calculated as follows:

$$H[p(y|x)] = -\sum_{c=1}^{C} p(y = c|x) log p(y = c|x)$$

Where:

- $p(y = c|x)$ is the mean softmax probability for class c across forward passes.
- C is the number of classes.

A standard (non-Bayesian) neural network has fixed parameters and cannot represent uncertainty about the model's knowledge, referred to as epistemic uncertainty. In contrast, a Bayesian Neural Network (BNN) uses a distribution over its weights, allowing it to quantify and express uncertainty about the data. Epistemic uncertainty in a BNN can be assessed by calculating the variance of the predicted class probabilities across multiple forward passes. This variance indicates how much the model's predictions for each class fluctuate due to the uncertainty in the model's parameters.

Neural network design and training

The Bayesian ResNet model featured a Bayesian layer as its final layer, allowing for the quantification of uncertainty in its predictions. This layer employed trainable parameters for the mean μ and variance σ of the weights and biases, allowing stochastic sampling during forward passes. The prior distributions for weights and biases were defined as normal distributions with a mean μ of 0 and a standard deviation σ of 0.2. These choices reflect typical assumptions of zero-mean priors with moderate variance to balance regularization and model flexibility (Blundell et al., 2015).

The reparameterization trick was employed to enable backpropagation through stochastic layers. By expressing the weights and biases as deterministic transformations of sampled noise, gradients could flow through the network effectively during optimization. Specifically, parameters were reparameterized as $w = \mu + \sigma + \epsilon$ where ϵ is sampled from a standard normal distribution. This approach is critical for variational inference in Bayesian neural networks, allowing the optimization of posterior distributions via gradient descent (Kingma & Welling, 2022).

The model underwent hyperparameter tuning to optimize learning rates and batch sizes. The dataset was split into a test set (20% of the total data), with the remaining 80% further divided into training (80% of the remaining data) and validation sets (20% of the remaining data). Validation loss guided the selection of the best configuration, which was subsequently used for further training. Loss was calculated as a combination of cross-entropy loss and KL divergence, weighted by a KL coefficient (10^{-3}), to balance task performance with regularization. KL weight tuning was excluded from hyperparameter tuning due to its sensitivity in stabilizing loss during training (Graves, 2011).

Variational inference was performed on a balanced sample of 2,000 test images, evenly divided between in-distribution classes and out-of-distribution (OOD) COVID-19 images. Each image underwent 100 stochastic forward passes, generating predictive distributions for uncertainty analysis. Equal class representation ensured robust uncertainty estimates by mitigating bias in sample composition.

Results and discussion

Model performance was assessed using accuracy, precision, recall, and F1 scores on in-distribution classes. A confusion matrix (Fig 2.) highlighted classification performance, while a separate report evaluated predictions for OOD samples, including COVID-19 images and normal images subjected to random transformations.

The Bayesian ResNet model achieved an overall accuracy of 93.3% in classifying normal versus non-COVID infection radiographs, demonstrating the model's capability to distinguish between these two classes effectively, even under the influence of uncertainty in the Bayesian framework. Precision, recall, and F1-score metrics, all approximately 93.3%, further emphasize the model’s balanced performance across the classes.

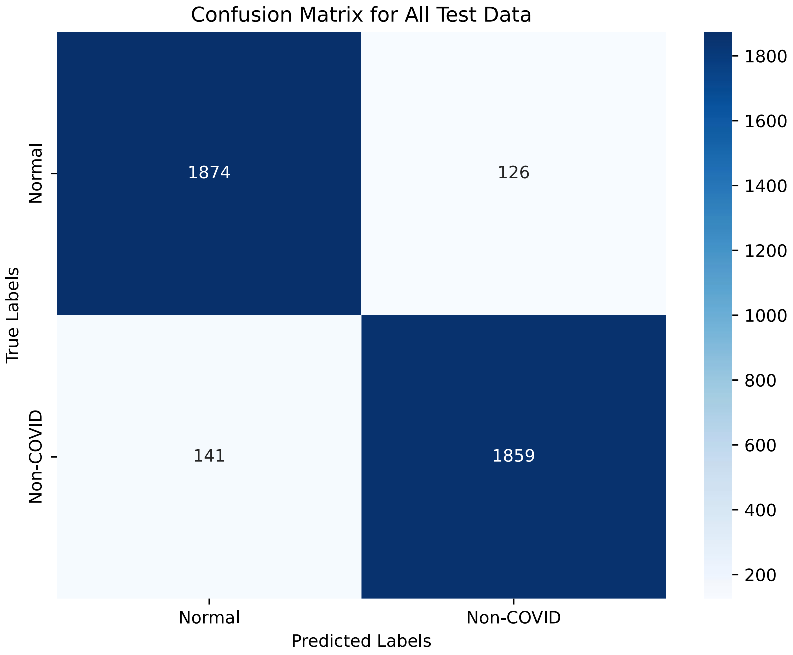

Fig. 2: Confusion matrix for test data (normal and non-covid radiographs).

For the out-of-distribution (OOD) transformation of normal radiographs, the model misclassified 636 images as normal and 1364 as non-COVID. This highlights the challenges of handling OOD data, particularly radiographs with subtle or complex transformations that deviate from the distribution of the training data. In the case of COVID-19 OOD data, the model predicted 990 images as normal and 1010 as non-COVID, yielding an approximately equal distribution of predictions between the two classes. As the model was not exposed to COVID-19 data during training, this indicates that the model does not strongly associate COVID-19 radiographs with normal or non-COVID classes present in training.

The Kruskal-Wallis tests for entropy and prediction variance equality showed significant differences between the various image classes. Dunn's post-hoc test revealed significant differences in entropy and prediction variance between each pairwise comparison, indicating that the model's uncertainty behaved differently for each category. Results are summarized in Fig. 1-2 and Table 1. A detailed report of p values for each category can be found in the attached data analysis notebook.

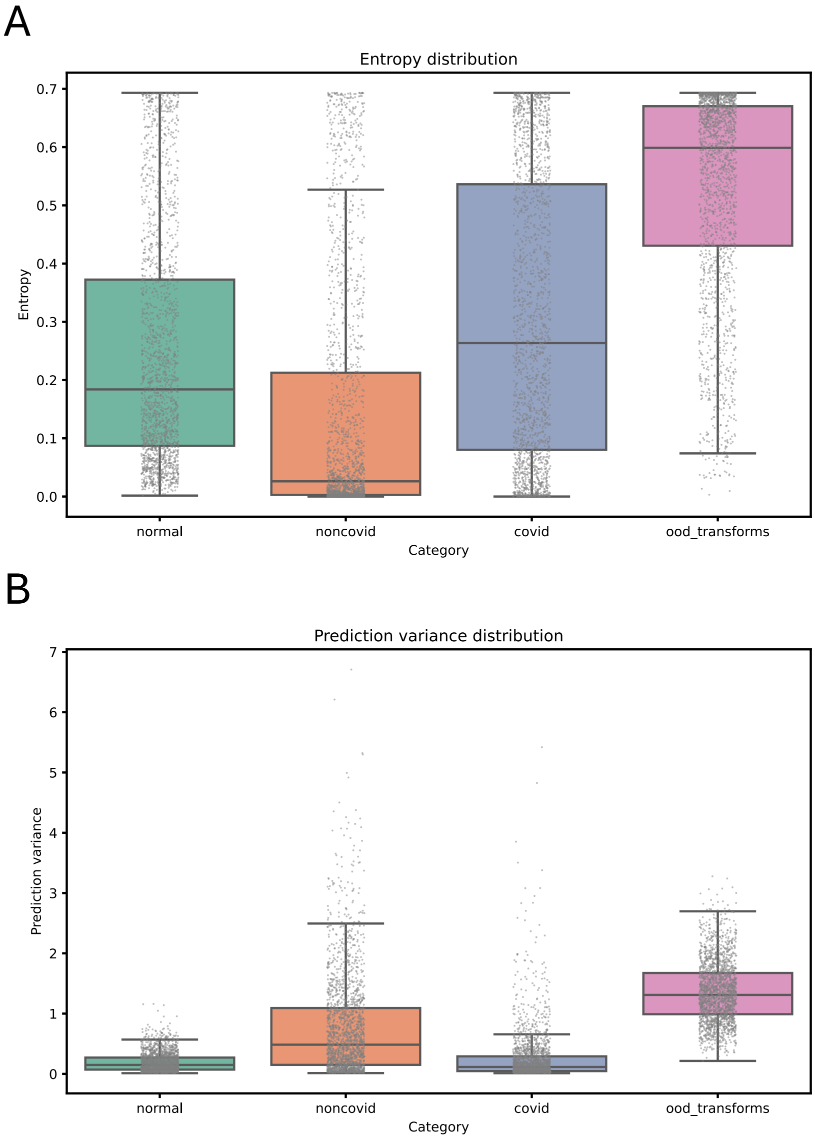

Fig. 3: Entropy (A) and prediction variance distribution (B) for test (normal and non-COVID) and OOD data (COVID-19 and normal data subjected to OOD transforms).

Table 1. Mean and variance of prediction variance and entropy.

	Mean Variance	Std Variance	Mean Entropy	Std Entropy
Normal	0.1945	0.1632	0.2481	0.2000
Non-COVID	0.7663	0.8385	0.1454	0.2140
COVID-19	0.2605	0.4193	0.3052	0.2381
Normal (OOD)	1.3619	0.5034	0.5271	0.1774

Non-COVID infection radiographs showed relatively low entropy and high epistemic uncertainty compared to the normal class. This underscores a key advantage of the Bayesian framework in this study. While classical models typically provide a single deterministic output, the Bayesian approach quantifies uncertainty by offering a distribution of possible outcomes, enabling it to express uncertainty in ways that traditional models cannot. The high epistemic uncertainty indicates that while the model is sure about the classification, it still has uncertainty about its learned parameters and how well they apply to future, potentially slightly different examples. To reduce epistemic uncertainty for non-COVID infections, increasing the amount of training data for this class could be beneficial. Additionally, techniques such as data augmentation, semi-supervised learning, or active learning could further help. These strategies, in conjunction with the Bayesian framework's ability to express uncertainty, could enhance the model's confidence in classifying non-COVID infections.

Despite COVID-19 data being absent from the training set, the model did not exhibit high prediction variance when presented with this class. This result challenges the assumption that uncertainty metrics, such as epistemic uncertainty, can reliably detect novel classes during inference. Typically, high epistemic uncertainty is expected for truly out-of-distribution (OOD) data, reflecting the model's inability to generalize to unseen classes. The model's lack of increased uncertainty when encountering COVID data suggests it may not be effectively capturing uncertainty for novel classes. This could point to limitations in the training process or a failure of the model to generalize to new, unseen types of infections.

Interestingly, normal data subjected to OOD transformations exhibited high epistemic uncertainty, as expected for data that deviates from the training distribution. The increased uncertainty suggests the model struggled to generalize to these image variations. Additionally, entropy was high for these OOD-transformed images, indicating low confidence in the predictions, further highlighting the challenge of handling OOD data in a model trained on limited classes.

Overall, this study demonstrates the advantages of Bayesian models in medical imaging, particularly for capturing and quantifying epistemic uncertainty. However, it also highlights the challenges posed by OOD data, especially when the model has not been exposed to certain disease classes during training.

References

1. Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). *Weight Uncertainty in Neural Networks* (No. arXiv:1505.05424). arXiv. https://doi.org/10.48550/arXiv.1505.05424
2. Graves, A. (2011). *Practical Variational Inference for Neural Networks*.
3. He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition* (No. arXiv:1512.03385). arXiv. https://doi.org/10.48550/arXiv.1512.03385
4. Kingma, D. P., & Welling, M. (2022). *Auto-Encoding Variational Bayes* (No. arXiv:1312.6114). arXiv. https://doi.org/10.48550/arXiv.1312.6114
5. Tahir, A. M., Chowdhury, M. E. H., Qiblawey, Y., Khandakar, A., Rahman, T., Kiranyaz, S., Khurshid, U., Ibtehaz, N., Mahmud, S., & Ezeddin, M. (2022). *COVID-QU-ex dataset*. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/3122958