UTMap: Triplet Neural Network for Uncertainty Medical Image Analysis

Rafael Souza e Silva[1], **André C. P. L. F. de Carvalho**[1]

[1] Institute of Mathematics and Computer Sciences – University of São Paulo (USP)
São Carlos – SP – Brazil

rafaelsoz@usp.br, andre@icmc.usp.br

Abstract. *This paper proposes a framework to enhance the reliability of deep classifiers in medical imaging by combining visualization and uncertainty quantification through meta-learning. The methodology employs a Triplet Neural Network (UTMap) to project instances into the Instance Uncertainty Space (IUS), which highlights patterns of confidence and uncertainty. Additionally, the Neighborhood Reliability Score (NRS) metric is introduced to estimate uncertainty based on the spatial relationships within the IUS. Experimental results show that the IUS effectively represents classifier behavior and that the NRS achieves competitive performance compared to traditional uncertainty estimation algorithms in distinguishing between correct and incorrect predictions.*

1. Introduction

There is a growing interest in the integration of medicine and artificial intelligence (AI) [Poon et al. 2025], driven by technological advances that enable increasingly sophisticated algorithms. In this context, models based on Deep Neural Networks (DNNs) have stood out for their ability to handle complex problems quickly and efficiently, becoming a promising tool to support the analysis and diagnosis of medical images [Mei et al. 2020, Gayathri et al. 2022].

Despite the potential of DNNs, especially Convolutional Neural Networks (CNNs) [LeCun et al. 2015], these models face significant challenges. These challenges include overconfidence even in the face of incorrect predictions [Nguyen et al. 2015, Goodfellow et al. 2015] and a tendency to overfit [Lee et al. 2017]. Although effective at identifying complex patterns, DNNs may also fit the noise present in training data [Algan and Ulusoy 2021].

In the healthcare domain, these limitations are exacerbated by the recurring scarcity and low quality of data [Bajwa et al. 2021, Ahmed et al. 2023, Hassan et al. 2024]. Preprocessing and data augmentation techniques are frequently applied to mitigate these limitations. [Islam et al. 2024, Goceri 2023]. Although such strategies improve model performance and data quality, they do not eliminate the biases associated with small datasets [Shorten and Khoshgoftaar 2019], compromising the reliability of deep learning-based solutions in clinical applications.

Uncertainty quantification methods have emerged as alternatives to increase the reliability of DNNs. In the medical domain, approaches such as Monte Carlo Dropout [Gal and Ghahramani 2016] and techniques based on DNN ensembles [Lakshminarayanan et al. 2017] have been largely explored in the literature

[Ling Huang et al. 2024, Benjamin Lambert et al. 2024]. However, the practical adoption of these techniques is limited by their high computational cost and the need for technical expertise in AI, which remains scarce among healthcare professionals [Ahmed et al. 2023, Hoffman et al. 2025].

In this work, we propose a novel methodology that integrates visualization and uncertainty quantification for the task of medical image classification, grounded in principles of Deep Meta-Learning [Huisman et al. 2021]. For such, we introduce the Uncertainty Triplet Network Mapping (UTMap), a projection meta-model responsible for constructing a two-dimensional metric space called the Instance Uncertainty Space (IUS). This space aims to model the behavior of deep classifiers with respect to instances, grouping those that are correctly classified and have low entropy, while placing more uncertain or incorrectly classified samples separately. Additionally, we propose a new uncertainty metric, named the Neighborhood Reliability Score (NRS), which is calculated based on distances in the IUS and is inspired particularly by the instance hardness measure N_2I [Smith et al. 2014].

According to experiments conducted on public medical image datasets, the IUS effectively represents the classifier's behavior in relation to the data. Moreover, the NRS metric demonstrated competitive and, in some aspects, superior performance compared to traditional uncertainty estimation methods in distinguishing between correct and incorrect predictions.

The main contributions of this work are:

- A novel framework integrating visualization and quantification for the analysis of uncertainty for medical image classification models;
- Development of a projection meta-model capable of mapping extracted representations into a structured two-dimensional space (IUS);
- Definition of NRS, a metric based on distance relationships in the IUS, empirically validated for accurately detecting classification failures;
- Practical validation of the approach on public medical image datasets, with source code made available [1] for experiment reproducibility.

2. Background and Related Work

2.1. Metric-Based Meta-Learning

Metric-based meta-learning methods aim to learn a latent space with meaningful representations, where the proximity between samples reflects their similarity [Tian et al. 2022]. From this space, new tasks are solved by directly comparing unseen inputs to the training examples, using the similarity relationships learned during the meta-learning process.

The main approaches found in the literature include: Siamese Networks [Koch et al. 2015], which learn a feature space based on comparisons between pairs of samples; Matching Networks [Vinyals et al. 2016], which perform pairwise comparisons between the support set and the new query inputs to construct a representational space; and Prototypical Networks [Snell et al. 2017], which organize latent representations around class prototypes.

[1] Public repository with the source code: `https://github.com/Rafaelsoz/UTMap`

2.2. Triplet Neural Network

Triplet Neural Networks (TNNs) [Hoffer and Ailon 2015], grounded in the principles of Metric Learning [Kaya and Bilge 2019], are deep learning models capable of mapping similarities through distance comparisons in the feature space. Their goal is to bring instances of the same class closer together while pushing apart instances with different labels, respecting a predefined margin.

Unlike Siamese Networks [Koch et al. 2015], which are trained on pairs of samples with either identical or different labels, TNNs operate on triplets of samples (x_i^a, x_i^p, x_i^n), consisting of an anchor x_i^a, a positive example x_i^p that shares the same label as the anchor, and a negative example x_i^n with a different label. The same network processes all three samples, and the loss is computed based on the distances between their representations.

However, the effectiveness of training heavily depends on selecting informative triplets to learn discriminative representations [Kaya and Bilge 2019], which remains a challenging task. Mining strategies are often employed to address this issue [Schroff et al. 2015, Simo-Serra et al. 2015], although they lead to higher computational costs and increased training complexity.

2.3. Confidence Estimation

An intuitive approach to estimating the uncertainty associated with a classification is to directly use the predicted probabilities, analyzing the Maximum Class Probability (MCP), extracted from the model's *softmax* layer. Despite its reasonable performance [Hendrycks and Gimpel 2016], this approach has important limitations, as deep learning models tend to be overconfident, even in incorrect predictions [Nguyen et al. 2015, Goodfellow et al. 2015].

Monte Carlo sampling-based methods have been explored to mitigate this issue, with Monte Carlo Dropout (MCDropout) [Gal and Ghahramani 2016] being one of the most widely used due to its ease of implementation. During inference, this technique applies dropout and performs multiple forward passes to estimate uncertainty. The average of the softmax outputs is used as the final prediction, while metrics such as entropy provide a measure of uncertainty. Although effective, the method can suffer from model overfitting [Nguyen et al. 2015], which compromises the reliability of the estimates. Additionally, interpreting the results requires appropriate statistical knowledge.

Another line of investigation involves neural network ensembles [Lakshminarayanan et al. 2017], which quantify uncertainty based on the variation in predictions from multiple independent models. This approach does not require modifications to individual architectures and tends to produce more robust estimates, but it comes with a high computational cost due to the need to train several networks.

2.4. Failure Prediction

Failure prediction aims to identify incorrect predictions made by a classification model based on the analysis of confidence scores associated with its outputs. For this purpose, a confidence scoring function $\kappa_f : \mathcal{X} \rightarrow R^+$ is defined, such as the MCP, which, in the context of deep learning, is associated with the model f to represent the level of

confidence in its predictions. An appropriate confidence function should assign lower scores to incorrectly classified instances compared to those that are correctly classified.

Based on this principle, during inference, a predefined threshold $\delta \in R^+$ is applied to the confidence scores, allowing the rejection of potentially incorrect predictions using the following decision function:

$$g(\mathbf{x}) = \begin{cases} 1, & \text{if } \kappa_f(\mathbf{x}) \geq \delta, \\ 0, & \text{otherwise.} \end{cases}$$

The performance of failure prediction methods is commonly evaluated using standardized metrics in the literature [Hendrycks and Gimpel 2016], such as: Area Under the ROC Curve (AUROC); Area Under the Precision-Recall Curve (AUPR), considering both error (AUPR-Error) and success (AUPR-Success) as the positive class; False Positive Rate at 95% True Positive Rate (FPR@95%TPR); among others.

3. Uncertainty Triplet Network Mapping

The proposed framework, illustrated in Figure 1, is grounded in the principles of Deep Meta-Learning and leverages the ability of Triplet Neural Networks (TNNs) to model similarities in a metric space. To this end, we introduce the meta-model called Uncertainty Triplet Network Mapping (UTMap), which maps the classifier's behavior regarding the instances and their corresponding predictions into a two-dimensional space, referred to as the Instance Uncertainty Space (IUS). Based on this space, we also propose a neighborhood-based confidence metric, called the Neighborhood Reliability Score (NRS).

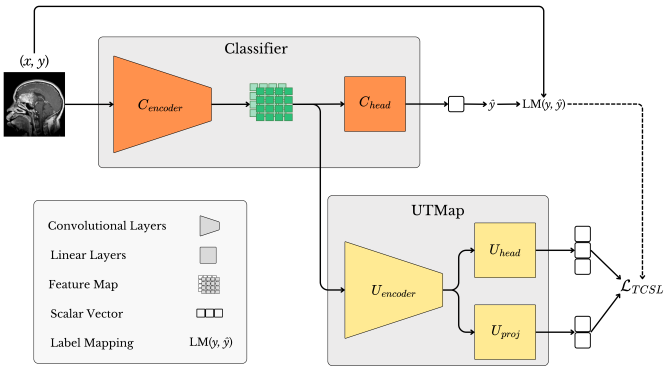

Figure 1. Uncertainty Triplet Network Mapping Framework

3.1. Instance Uncertainty Space

To construct the proposed space, we consider a binary classification problem over a dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$, composed of N samples, where $x_i \in \mathcal{X}$ represents the inputs and $y_i \in \mathcal{Y}$ their respective labels. During training, a set of stochastic transformations is applied to the samples, denoted by $\tilde{x_i}$, with the goal of introducing random perturbations. This strategy allows for the capture of instances that, although initially correctly classified, begin to exhibit increased prediction entropy or, in some cases, become misclassified. Such an approach supports the construction of a latent space that is more robust to stochastic variations in the inputs.

Although data augmentation techniques are widely used in classification tasks, in the medical context, overly aggressive transformations may compromise critical image regions, such as tumor areas. To mitigate this risk, we adopt three transformations, based on the study by [Goceri 2023], which analyzed common augmentation techniques and their impacts on medical image classification. The applied transformations were: (1) Translation, limited to the range $[-15°, 15°]$; (2) Shearing, within the same range; and (3) Rotation, limited to the range $[-25°, 25°]$.

From the transformed images $\tilde{x}_i$, feature maps are extracted using the classification neural network C, specifically from $C_{encoder}$, and the associated predictions $\hat{y}_i$ are produced by C_{head} over the dataset $\mathcal{D}$, as illustrated in Figure 1. In this step, the weights of C remain frozen, and no gradient is computed.

In this context, the architecture of C can be replaced, as long as the embedding extraction is performed from the last convolutional layer, since the initial layers capture generic visual patterns, while the final layers preserve spatial features specific to the image [Zeiler and Fergus 2013, Yosinski et al. 2014]. These representations contribute to the construction of a more robust and discriminative latent space, supporting the proposed reliability analysis.

From the predictions generated by C, we define a label mapping function (LM) responsible for assigning pseudo-labels used in the optimization of U. This function organizes the samples into three distinct groups, as follows:

$$LM(y, \hat{y}) = \begin{cases} 0, & \text{if } y = 0 \text{ and } \hat{y} = 0 \quad \text{(True Positive)} \\ 1, & \text{if } y = 1 \text{ and } \hat{y} = 1 \quad \text{(True Negative)} \\ 2, & \text{otherwise} \qquad \text{(Type 1 and Type 2 Error)} \end{cases} \tag{1}$$

Thus, U is trained to group correctly classified instances in the latent space according to their true labels, while aiming to separate those that were misclassified.

The architecture of U follows a simple structure, consisting of an encoder with two convolutional layers [LeCun et al. 2015], followed by two fully connected (FC) layers, all with batch normalization [Ioffe and Szegedy 2015] (BN) and ReLU activation functions [Nair and Hinton 2010]. In addition to the encoder, the architecture includes two additional sub-networks, both implemented as single linear layers: U_{head}, responsible for predicting the pseudo-labels; and U_{proj}, which projects the instances into the latent space.

The optimization of U is performed using the *Triplet Center Softmax Loss* [He et al. 2018], an alternative to the traditional Triplet Loss, which incorporates class centers in the latent space, a concept introduced by [Wen et al. 2016]. Instead of directly comparing instance pairs, this technique employs the distance between a sample and the class centers, which are updated iteratively with each mini-batch during training. Thus, U receives as input the feature maps extracted by $C_{encoder}$ to generate the embeddings (e_i), as well as the logits corresponding to the assigned pseudo-labels:

$$\mathcal{L}_{tcl} = \sum_{i=1}^{N} \max(0, m + D(e_i, c_{y_i}) - \min_{y_j \neq y_i} D(e_i, c_{y_j})) \tag{2}$$

$$\mathcal{L}_{tcsl} = \lambda \mathcal{L}_{tcl} + \mathcal{L}_{softmax} \tag{3}$$

where c_{y_i} denotes the center of class y_i, which is used as the positive sample, and the negative sample is the nearest center from a different class. This approach removes the need for triplet mining, reducing computational cost and training complexity. The term $D(e_i, c_{y_i})$ represents the squared Euclidean distance between the sample and center. The hyperparameter λ controls the contribution of the *Triplet Center Loss* in the combined loss function. Finally, the *softmax* function guides the association of samples to pseudo-labels, while the *Triplet Center Loss* directly shapes the embedding space.

3.2. Neighborhood Reliability Score

After constructing the IUS metric space, it becomes possible to compute the *Neighborhood Reliability Score (NRS)*, inspired by the instance hardness measure known as the *Ratio of Intra-Extra Class Distances at Instance Level* (N_2I) [Smith et al. 2014]. This measure assesses how well-positioned an instance is in the feature space by computing the ratio between the distance to its nearest neighbor of the same class and the distance to its nearest neighbor from a different class.

The proposed metric follows the same principle as N_2I but with two key differences: (1) distances are computed based on the projections in the IUS space; and (2) the labels used to determine the neighbors are, respectively, the true labels of the training instances and the predictions provided by C.

Thus, NRS allows us to infer the reliability of a prediction based on the relative position of the instance within the IUS space, reflecting the similarities learned by the meta-model U. The metric is defined as follows:

$$NRS(x_i, \hat{y}_i) = \frac{1}{IntraInter(x_i, \hat{y}_i) + 1} \qquad IntraInter(x_i, \hat{y}_i) = \frac{d(x_i, NN(x_i) \wedge y_j = \hat{y}_i)}{d(x_i, NN(x_i) \wedge y_j \neq \hat{y}_i)} \tag{4}$$

Values close to zero indicate that the instance lies near examples from other classes, that is, in regions of the IUS space composed of high-entropy or misclassified samples, reflecting greater uncertainty in the prediction. On the other hand, values close to one indicate that the instance is located within well-defined clusters, signaling higher reliability in the classification provided by C.

4. Experiments

The experiments were conducted with two main objectives: (1) to evaluate UTMap's ability to structure the latent space in a way that reflects the classifier's behavior toward the instances; and (2) to compare the proposed metric, NRS, with well-established uncertainty estimation methods in inference scenarios.

For the first objective, a visual analysis of the IUS space was performed, examining the organization of the samples in terms of class labels and predictive entropy. For the second, a quantitative evaluation was conducted by directly comparing NRS to traditional uncertainty estimation approaches. All metrics were computed using stratified 10-fold cross-validation, with 80% of the data used for training and 20% for testing, with 10% of the training set reserved for validation.

4.1. Experimental Setup

To evaluate the proposed method, three public binary classification datasets with medical images exhibiting distinct visual characteristics were used. The selection of these datasets aims to cover various clinical scenarios and imaging modalities, thereby ensuring the greater robustness of the results obtained.

The datasets include cases of Brain Cancer [Sartaj Bhuvaji 2020] using Magnetic Resonance Imaging (MRI), Breast Cancer [Al-Dhabyani et al. 2020] using Ultrasound (US), and SARS-CoV-2 [Soares et al.] using Computed Tomography (CT). A summary of the datasets, as well as the hyperparameters used for training the classifier, is presented in Table 1.

Table 1. Datasets used and base classifier hyperparameters

ID	Dataset	Modality	Instances	Epochs	Batch Size	Learning Rate
1	Brain Cancer	MRI	3.264	10	256	1×10^{-4}
2	Breast Cancer	US	780	5	128	1×10^{-4}
3	SARS-CoV-2	CT	2.279	10	256	1×10^{-4}

In this study, we adopted ResNet-18 [He et al. 2016] as the architecture for the base classifier. The model was trained individually for each dataset using transfer learning through fine-tuning on weights previously trained on ImageNet [Russakovsky et al. 2015]. Additionally, data augmentation techniques were applied during training to improve the model's robustness. The transformations used follow the same ones described in Section 3.1.

The algorithms adopted for comparison with the proposed method are: Maximum Class Probability (MCP) [Hendrycks and Gimpel 2016]; Monte Carlo Dropout (MCDropout) [Gal and Ghahramani 2016], using the mean predictive entropy as the uncertainty measure with 50 samples; and Deep Ensemble (DE) [Lakshminarayanan et al. 2017], configured with 5 independent models. For UTMap, the λ value in the loss function was empirically set to 0.1. All models were trained using the Adam optimizer [Kingma and Ba 2015], and the respective hyperparameters used in each approach are described in Table 2.

Table 2. Hyperparameters used for training the uncertainty estimation algorithms, organized by dataset.

Dataset	MCDropout				DE	UTMap			
	Ep	Batch	Lr	P	Models	Ep	Batch	Lr	Center Lr
Brain Cancer	30	128	1×10^{-4}	0.1	5	20	16	5×10^{-5}	5×10^{-3}
Breast Cancer	30	128	1×10^{-4}	0.1	5	20	16	5×10^{-5}	5×10^{-3}
SARS-CoV-2	30	128	1×10^{-4}	0.1	5	20	16	5×10^{-5}	5×10^{-3}

To demonstrate the effectiveness of the proposed method, we evaluated the algorithms' ability to detect prediction failures using well-established metrics from the literature [Hendrycks and Gimpel 2016], as detailed in Section 2.4: AUROC; AUPR-Error, due to its direct relevance to the task of failure detection; and finally, FPR@95%TPR.

4.2. Experimental Results

4.2.1. Analysis of Latent Projections

Based on the visualization presented in Figure 2, it can be observed that the meta-model is capable of grouping samples according to the generated pseudo-labels, separating those that are correctly classified from those associated with prediction errors. However, in some instances, although correctly classified, are projected closer to regions associated with errors than to regions corresponding to their true label.

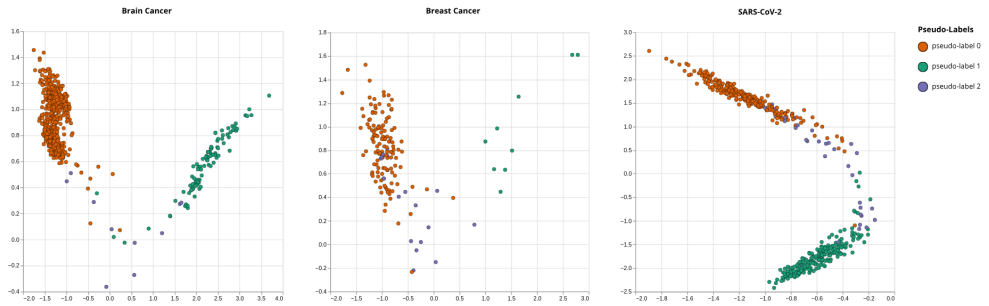

Figure 2. Latent space projections with generated pseudo-labels.

This behavior can be attributed to the application of stochastic perturbations to the images, which significantly affect the probability distribution generated by the classifier [Goodfellow et al. 2015]. In such cases, instances originally associated with low predictive entropy may begin to exhibit higher uncertainty or even be misclassified after perturbations, and are thus displaced to regions of the latent space near misclassified samples, as illustrated in Figure 3.

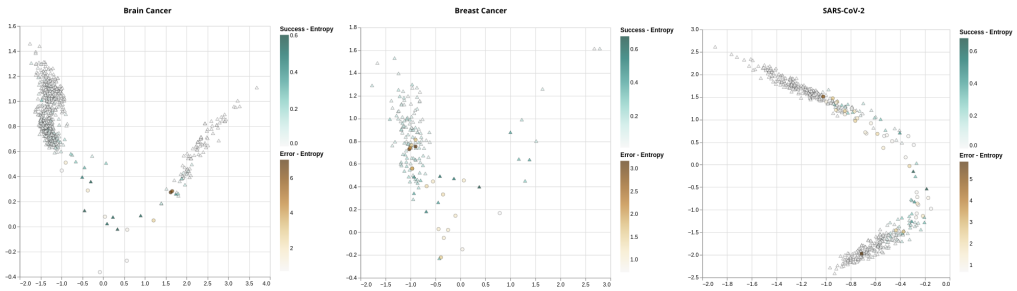

Figure 3. Latent space projections with the classifier's predictive entropy.

Furthermore, Figure 3 shows that the meta-model, when mapping instances based on pseudo-labels, also incorporates information about the classifier's predictive entropy. It is evident that samples with high entropy tend to cluster near regions associated with classification errors, even when correctly classified. Conversely, samples located farther away tend to exhibit lower entropy, revealing a gradual transition between regions of high confidence and high uncertainty as one approaches the error zones.

Lastly, Figure 2 also shows cases where misclassified samples are projected into densely populated regions of correctly classified instances. This scenario may indicate the presence of outliers, mislabeled samples, or limitations in the discriminative capacity

of the feature space extracted by the classifier, as the misclassified instances in question share latent representations that are very similar to those of correctly classified samples. This similarity hinders their separation in the projection step, which depends on both the pseudo-labels and the representations generated by the classification model.

4.2.2. Uncertainty Estimation Methods

The comparative results presented in Table 3 indicate that the proposed method demonstrates competitive or superior performance compared to established approaches from the literature in two out of the three evaluated metrics across different datasets, with particular emphasis on AUPR and FPR@95%TPR.

Table 3. Comparison for Failure prediction performance.

Dataset	Metric	MCP	MCDropout	DE	NRS
Brain Cancer	AUROC (↑)	0.86 ±0.12	0.91 ±0.02	0.85 ±0.12	**0.96 ±0.03**
	AUPR (↑)	0.37 ±0.12	0.41 ±0.06	0.49 ±0.22	**0.71 ±0.20**
	FPR @95% TPR(↓)	0.22 ±0.14	0.34 ±0.07	0.25 ±0.19	**0.09 ±0.12**
Breast Cancer	AUROC (↑)	0.87 ±0.05	**0.89 ±0.06**	0.88 ±0.08	0.85 ±0.08
	AUPR (↑)	0.56 ±0.11	0.48 ±0.20	0.54 ±0.17	**0.60 ±0.19**
	FPR @95% TPR(↓)	**0.37 ±0.13**	0.43 ±0.20	0.44 ±0.22	0.42 ±0.19
SARS-CoV-2	AUROC (↑)	**0.90 ±0.03**	0.56 ±0.02	0.90 ±0.04	0.89 ±0.03
	AUPR (↑)	0.39 ±0.14	**0.51 ±0.04**	0.31 ±0.10	0.51 ±0.13
	FPR @95% TPR(↓)	0.48 ±0.17	0.92 ±0.03	0.44 ±0.09	**0.37 ±0.15**

These results show that, with respect to the AUROC metric, the proposed method stands out on the brain cancer dataset, shows competitive performance on the SARS-CoV-2 dataset, and underperforms on the breast cancer dataset. Regarding AUPR, arguably the metric most directly related to the task of failure detection, the proposed approach performs strongly across all three datasets, ranking among the top two in each. It achieved the best performance on the Brain Cancer and Breast Cancer datasets and the second-best on the SARS-CoV-2 dataset.

Concerning the FPR@95%TPR metric, the proposed method proved to be the most effective at identifying misclassified examples while maintaining a low false positive rate. Again, it ranked among the top two methods in all datasets, achieving the best results on Brain Cancer and SARS-CoV-2, and the second-best on the Breast Cancer dataset.

These results suggest that NRS is effective in identifying misclassified samples while maintaining a low false-positive rate in error detection. Moreover, as it is based on distances in the IUS space, the method provides a visual and intuitive representation for identifying low-confidence instances. In contrast, approaches such as MCDropout and Deep Ensemble entail higher computational costs, with MCDropout also requiring more advanced technical understanding for proper result interpretation.

5. Discussion and Conclusion

This work proposes a method that integrates visualization and uncertainty quantification in deep classification models applied to medical images. The approach consists of incor-

porating a projection meta-model into the classification network, enabling the construction of a two-dimensional metric space that represents the classifier's uncertainty with respect to the instances. From this space, a confidence metric is defined based on learned similarities and grounded in hardness measures.

The main objective is to evaluate the proposed method's ability to model a space that adequately reflects the classifier's behavior, as well as assess whether the uncertainty metric achieves performance comparable to well-established approaches in the literature. To this end, a visual analysis of the constructed space was performed, along with a quantitative comparison against three well-known methods, using three distinct binary datasets.

The experimental results show that the proposed method is capable of modeling a latent space in which instances with similar predictive entropy tend to cluster together, while separating high-entropy instances from those with low entropy. Furthermore, the proposed metric demonstrates competitive or superior performance in two of the three evaluated metrics across different datasets, with notable results in AUPR and FPR@95%TPR. These findings demonstrate the effectiveness of the metric in identifying misclassified samples while maintaining a low false positive rate in error detection.

Given the strong performance observed in modeling and quantifying uncertainty for binary classification tasks, future work will focus on generalizing the method to multiclass problems, a scenario that poses additional challenges, particularly in defining consistent pseudo-labels. Overcoming these limitations will allow the approach to be evaluated in more complex contexts, where multiple classes and diverse types of errors coexist.

References

Ahmed et al. (2023). A systematic review of the barriers to the implementation of artificial intelligence in healthcare. *Cureus*, 15(10).

Al-Dhabyani, W., Gomaa, et al. (2020). Dataset of breast ultrasound images. *Data in brief*, 28:104863.

Algan, G. and Ulusoy, I. (2021). Image classification with deep learning in the presence of noisy labels: A survey. *Knowledge-Based Systems*, 215:106771.

Bajwa, J., Munir, et al. (2021). Artificial intelligence in healthcare: transforming the practice of medicine. *Future healthcare journal*, 8(2):e188–e194.

Benjamin Lambert, F. F. et al. (2024). Trustworthy clinical ai solutions: A unified review of uncertainty quantification in deep learning models for medical image analysis. *Artificial Intelligence in Medicine*, 150:102830.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Gayathri, J., Abraham, B., Sujarani, M., and Nair, M. S. (2022). A computer-aided diagnosis system for the classification of covid-19 and non-covid-19 pneumonia on chest x-ray images by integrating cnn with sparse autoencoder and feed forward neural network. *Computers in biology and medicine*, 141:105134.

Goceri, E. (2023). Medical image data augmentation: techniques, comparisons and interpretations. *Artificial Intelligence Review*, 56(11):12561–12605.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples.

Hassan, M., Kushniruk, A., and Borycki, E. (2024). Barriers to and facilitators of artificial intelligence adoption in health care: scoping review. *JMIR Human Factors*, 11:e48633.

He, K. et al. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

He, X. et al. (2018). Triplet-center loss for multi-view 3d object retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1945–1954.

Hendrycks, D. and Gimpel, K. (2016). A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Hoffer, E. and Ailon, N. (2015). Deep metric learning using triplet network. In *Similarity-based pattern recognition: third international workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3*, pages 84–92. Springer.

Hoffman, J. et al. (2025). Overcoming barriers and enabling artificial intelligence adoption in allied health clinical practice: A qualitative study. *Digital Health*, 11:20552076241311144.

Huisman, M., Van Rijn, J. N., and Plaat, A. (2021). A survey of deep meta-learning. *Artificial Intelligence Review*, 54(6):4483–4541.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr.

Islam, T. et al. (2024). A systematic review of deep learning data augmentation in medical imaging: Recent advances and future research directions. *Healthcare Analytics*, 5:100340.

Kaya, M. and Bilge, H. Ş. (2019). Deep metric learning: A survey. *Symmetry*, 11(9):1066.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. international conference on learning representations (2015). *San Diego, California*.

Koch et al. (2015). Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, pages 1–30. Lille.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Lee, J.-G. et al. (2017). Deep learning in medical imaging: general overview. *Korean journal of radiology*, 18(4):570.

Ling Huang, S. R. et al. (2024). A review of uncertainty quantification in medical image analysis: Probabilistic and non-probabilistic methods. *Medical Image Analysis*, 97:103223.

Mei, X. et al. (2020). Artificial intelligence–enabled rapid diagnosis of patients with covid-19. *Nature medicine*, 26(8):1224–1228.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

Nguyen, A., Yosinski, J., and Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436.

Poon, E. G. et al. (2025). Adoption of artificial intelligence in healthcare: survey of health system priorities, successes, and challenges. *Journal of the American Medical Informatics Association*, 32(7):1093–1100.

Russakovsky, O., Deng, et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252.

Sartaj Bhuvaji, A. K. o. (2020). Brain tumor classification (mri).

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48.

Simo-Serra, E. et al. (2015). Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Smith, M. R., Martinez, T., and Giraud-Carrier, C. (2014). An instance level analysis of data complexity. *Machine learning*, 95:225–256.

Snell, J., Swersky, K., and Zemel, R. (2017). Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Soares et al. Sars-cov-2 ct-scan dataset: A large dataset of real patients ct scans for sars-cov-2 identification. medrxiv 2020. *Google Scholar*.

Tian, Y., Zhao, X., and Huang, W. (2022). Meta-learning approaches for learning-to-learn in deep learning: A survey. *Neurocomputing*, 494:203–223.

Vinyals et al. (2016). Matching networks for one shot learning. *Advances in neural information processing systems*, 29.

Wen, Y. et al. (2016). A discriminative feature learning approach for deep face recognition. In *Computer vision–ECCV 2016: 14th European conference, amsterdam, the netherlands, October 11–14, 2016, proceedings, part VII 14*, pages 499–515. Springer.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks?

Zeiler, M. D. and Fergus, R. (2013). Visualizing and understanding convolutional networks.