Using Explainable AI to Measure Feature Contribution to Uncertainty

Katherine E. Brown and **Douglas A. Talbert**
Department of Computer Science
Tennessee Technological University
1 William L. Jones Drive
Cookeville, TN 38505

Abstract

The application of artificial intelligence techniques in safety-critical domains such as medicine and self-driving vehicles has raised questions regarding its trustworthiness and reliability. One well-researched avenue for improving trust in and reliability of deep learning is uncertainty quantification. *Uncertainty* measures the algorithm's lack of trust in its predictions, and this information is important for practitioners using machine learning-based decision support. A variety of techniques exist that produce uncertainty estimations for machine learning predictions; however, very few techniques attempt to explain why that uncertainty exists in the prediction. Explainable Artificial Intelligence (XAI) is an umbrella term that encompasses techniques that provide some level of transparency to machine learning predictions. This can include information on which inputs contributed to or detracted from the algorithm's prediction. This work focuses on applying existing XAI techniques to deep neural networks to understand how features contribute to epistemic uncertainty. Epistemic uncertainty is a measure of confidence in a prediction given the training data distribution upon which the neural network was trained. In this work, we apply common feature attribution XAI techniques to efficiently deduce explanations of epistemic uncertainty in deep neural networks.

Introduction

Neural networks have produced state-of-the-art results in a variety of domains; however, their use in safety-critical domains has been limited due to their black-box nature and complex architectures (Begoli, Bhattacharya, and Kusnezov 2019). This has paved the way for research in uncertainty quantification and explainable artificial intelligence (XAI). Both of which strive to improve trust between human practitioners and their artificial intelligence counterparts.

Epistemic uncertainty provides a numerical measure of how similar an input instance is compared to the data on which the neural network has been trained and can be used to effectively indicate to human practitioners which instances should be referred to domain experts (Brown and Talbert 2019; Leibig et al. 2017). Only a very few techniques attempt to explain why that uncertainty exists in the prediction in a computationally efficient manner (Chai 2018).

Distilling a black-box model into a more easily interpretable model or algorithm has been extensively researched as a prediction explanation technique in the XAI literature (Frosst and Hinton 2017). In this work, we apply a similar approach to explaining epistemic uncertainty in deep neural networks. This is done by distilling Bayesian dropout uncertainty estimations into a regression-based machine learning model to which we can apply XAI techniques.

Motivation

There are several factors that motivate the explanation of epistemic uncertainty. We hypothesize that explaining epistemic uncertainty should identify which input features, or combinations thereof, result in increased epistemic uncertainty. The first benefit is this can yield usable information about how deep neural networks learn in complex feature spaces. This information could possibly be applied to improve deep neural networks in practice. Second, by identifying regions in the feature space that contribute to uncertainty, an active learning approach, for example, can be applied to sample additional training data that can be used to reduce uncertainty in that region of the feature space. Third, explaining model uncertainty can have practical benefits for the utilization of deep learning in safety-critical scenarios, where an improved understanding of the model's capabilities can improve its utility and reliability. For example, such techniques can provide more information to guide users and domain experts to critically assess the model's predictions.

Contributions

This work presents several contributions to the uncertainty quantification and explainable artificial intelligence literature. First, we provide an application of XAI techniques to epistemic uncertainty quantification in deep learning. Second, we provide an assessment of these explanations of epistemic uncertainty by determining if explanations of uncertainty reflect the true rationale of the uncertainty. Finally, we determine if the explained uncertainty is related to negative relevance from classification explanations.

Background

In this section, we present the background literature needed for this work.

Uncertainty Quantification

Uncertainty quantification is a growing field of study that provides information regarding the reliability of neural networks, important information in safety-critical tasks, (Begoli, Bhattacharya, and Kusnezov 2019) and can provide insight into the robustness of the model given its training distribution or the data input to the model for inference (Gal and Ghahramani 2016; Kendall and Gal 2017). In deep learning, two types of uncertainty are considered epistemic and aleatoric. Aleatoric uncertainty is beyond the scope of this work.

Epistemic uncertainty is a measure of how confident the model is inferring for a data point given the model's training data (Gal and Ghahramani 2016). This encompasses the suitability of the training data at representing the decision space and the suitability of the model at determining classification boundaries within the training data (i.e., is the model complex enough?).

A common and accessible technique to measure epistemic uncertainty is Bayesian dropout (Gal and Ghahramani 2016). Dropout is a regularization technique that prevents overfitting in a deep neural network by zero-ing out, or removing, weights. When a neural network is sampled several times, dropout creates an ensemble effect by which epistemic uncertainty can be measured. Bayesian dropout has been applied in medical applications (Leibig et al. 2017; Brown and Talbert 2019) as well as in other critical fields (Michelmore, Kwiatkowska, and Gal 2018).

One shortcoming of Bayesian dropout (and other "variational inference" techniques) is the computational overhead. To ascertain uncertainty, multiple inferences are typically required (Gal and Ghahramani 2016). This has led to rising study of Direct Uncertainty Methods (DUM). DUM techniques attempt to ascertain uncertainty using less computational complexity, typically through a single inference of a proxy algorithm. A survey of popular DUM techniques is given in (Postels et al. 2021).

Explainable Artificial Intelligence

Explainable Artificial Intelligence (XAI) has exploded in recent years. This is due, in part, to growing concerns of bias in artificially intelligent decision support systems (Ahmad et al. 2020). Feature attribution is one popular subset of XAI techniques in machine learning where predictions are explained by how much each input feature contributed to the model's prediction. Two very popular feature attribution techniques that we utilize in this work are Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro, Singh, and Guestrin 2016) and Layer-Wise Relevance Propagation (LRP) (Bach et al. 2015).

Local Interpretable Model-Agnostic Explanations Local Interpretable Model-Agnostic Explanations is a model-agnostic XAI technique that explores the decision space of the data point in question to construct an explanation for a prediction (Ribeiro, Singh, and Guestrin 2016). LIME seeks to balance explanation understandability with a locally faithful representation of the classifier by minimizing the sum of the loss (lack of fidelity) and complexity (lack of interpretability). To do this, LIME samples data near the data point in question. These data points are used to build a linear model that serves as an interpretable surrogate for the more complex classifier. The coefficients of the linear model are then used to assess how much each feature contributed to the overall prediction.

Layer-Wise Relevance Propagation (LRP) Layer-wise Relevance Propagation (LRP) is an XAI technique for deep neural networks that backpropagates relevance using a series of constraints (Bach et al. 2015). After LRP is executed, the input variables are assigned a "relevance" value that is a measure of the magnitude and direction of the influence that each specific input pixel exerts on the classification. A higher magnitude of the relevance value, either positive or negative, indicates that the pixel had more influence on the final decision. LRP can assign positive and negative relevance, where positive relevance can be seen as "votes" *for* a decision and negative relevance can be seen as "votes" *against* a decision. During this backpropagation procedure, the relevance of a specific neuron is decomposed as a summation of relevance values of neurons that feed into the neuron in question. Further, relevance is *conserved* between layers such that the amount of relevance between any two adjacent layers is equivalent. Then, transitivity guarantees that the relevance of the input pixels is a decomposition of the output. Further treatment of this aspect of LRP is given in (Bach et al. 2015).

The determination of how much relevance, positive or negative, is assigned to a previous layer is set by constraints. A common constraint is the $\alpha\beta$-rule. The value of α determines the emphasis placed on positive relevance, and the value of β determines the emphasis placed on negative relevance as long as $\alpha - \beta = 1$ (Montavon, Samek, and Müller 2018). In this work, classification explanations are generated with $\alpha = 2, \beta = 1$. When LRP is used to generate uncertainty estimations, the values are $\alpha = 1, \beta = 0$ to focus on which inputs actively contribute to epistemic uncertainty. We incorporate negative relevance into the classification explanations in order to study the relationship between negative relevance and epistemic uncertainty.

Intersection of Explainable Artificial Intelligence and Uncertainty Quantification in Deep Learning

Limited work exists in the intersection of uncertainty quantification in deep learning and XAI. First, Bykov et al. utilize classification uncertainty to enhance explanations. Earlier work noted that dropout produces perturbations in explanations when left active as part of a Bayesian neural network (Bykov et al. 2020). These perturbations represent the "uncertainty in the explanation." Explanations are augmented with this uncertainty information by highlighting which pixels (or input features) contribute most and least frequently (Bykov et al. 2021). Thus, this work does little to explain *why* the uncertainty is associated with specific inputs but

presents that information in a human interpretable manner.

The work by Chai attempted to explain which input features contributed to uncertainty in classification tasks. This work required the neural network to be continuously retrained with specific features removed to ascertain the overall effect on prediction uncertainty. For image data, this required removing specific patches of images rather than specific pixels to reduce the computational complexity of the technique. The technique was able to identify uncertain pixels but, as stated, did so at a potentially exponential cost (Chai 2018). We hypothesize recent developments in XAI and DUM techniques will help us garner similar information without the computational overhead

Application of XAI to Explain Uncertainty

To explain epistemic uncertainty in a deep neural network, our technique involves distilling Bayesian dropout uncertainty to a proxy model that can either be more easily interpreted via XAI or is an interpretable model itself. While several direct and deterministic uncertainty measures have been proposed in the literature (Van Amersfoort et al. 2020; Jain et al. 2021; Postels et al. 2021), we opt to use a simple regression technique proposed in (Brown and Talbert 2022). This technique distills Bayesian dropout-based epistemic uncertainty into a machine learning-based regression model using an augmentation of the training and validation sets of the classification algorithm. The use of machine learning regression allows us to apply common and verified XAI techniques such as LIME or LRP on the uncertainty estimation model. Moreover, we are attempting to explain Bayesian dropout uncertainty as defined by Gal and Ghahramani that specifically measures epistemic uncertainty (Gal and Ghahramani 2016). This should minimize the amount of aleatoric uncertainty present in the explanation; although it is not guaranteed that aleatoric (e.g., data uncertainty/noise) will not present in the explanation.

Experimental Methodology

In this section, we present the experimental methodology of the described techniques using a synthetic dataset and the MNIST image classification dataset.

Simulation with Synthetic Data

Dataset Description We first use a synthetic data so we can evaluate the techniques in an easily visualizable environment. A total of 500 points are generated with the two moons dataset in Scikit-Learn (Pedregosa and others 2011). The noise parameter is set to 0.3 and a random state of 0 is used. Each datapoint is associated with one of two possible classes, yielding a binary classification problem. A point (x, y) in the dataset is constrained as follows: $-1.4 \leq x_1 \leq 2.5$ and $-1.4 \leq x_2 \leq 1.7$. Training data is defined geometrically to be within the circle of radius 0.4 centered at $(0.5, 0.25)$. Figure 1 provides a visualization of the synthetic dataset.

Neural Networks Table 1 provides the architecture and training epoch information for neural networks used for the synthetic dataset. For the classification network, hidden layers use the Rectified Linear Unit (ReLU) as their activation functions, and the output layer uses the softmax function as its activation function. Dropout is applied before each hidden layer to calculate epistemic uncertainty (Gal and Ghahramani 2016). Binary cross-entropy is optimized using the ADAM optimization algorithm (Kingma and Ba 2014).

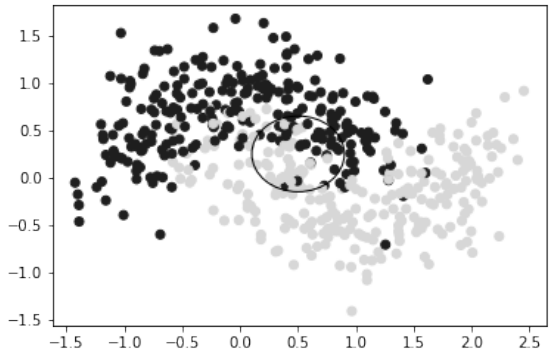

Figure 1: Visualization of the synthetic dataset generated. Two moons form a binary classification problem. The class of a point is indicated its color. Points within the circle serve as the training set and points outside as an evaluation set.

Table 1: Hyperparameter information for synthetic and MNIST datasets

Dataset	Task	Hidden Layer Architecture	No. Training Epochs
Synthetic	Classification	512	15
Synthetic	Uncertainty Regression	2000,1000,50	50
MNIST	Classification	128	15
MNIST	Uncertainty Regression	2000,1000,50	50

For the uncertainty regression network, mean squared error is optimized using the ADAM optimization algorithm (Kingma and Ba 2014). Hidden layers are activated using the Rectified Linear Unit (ReLU) activation function. Dropout is strictly used as a regularization technique (Srivastava et al. 2014).

Experiment Setup We evaluate our proposed technique as follows: First, we define the x and y boundaries of the training distribution. We define the lower and upper x-axis boundary as α_1 and α_2, respectively, and the upper and lower y-axis boundary as β_1 and β_2, respectively. Thus, for our defined synthetic scenario, $\alpha_1 = 0.1$, $\alpha_2 = 0.9$, $\beta_1 = -0.15$, $\beta_2 = 0.65$. This allows us to define 3 specific scenarios to evaluate. Given a point x_1, x_2:

- Scenario 1: $x_1 \notin [\alpha_1, \alpha_2]$ and $x_2 \in [\beta_1, \beta_2]$. Thus, x_2 is the coordinate that causes the point not to be within the training distribution.
- Scenario 2: $x_1 \in [\alpha_1, \alpha_2]$ and $x_2 \notin [\beta_1, \beta_2]$. Thus, x_1 is the coordinate that causes the point not to be within the training distribution.
- Scenario 3: $x_1 \notin [\alpha_1, \alpha_2]$ and $x_2 \notin [\beta_1, \beta_2]$. Thus, neither

coordinate is within the training distribution.

We use LIME (Ribeiro, Singh, and Guestrin 2016) as the explanation algorithm. We generate explanations of uncertainty with two different techniques. The first is explanation of uncertainty using LIME in which we use a regression model to estimate Bayesian dropout uncertainty and explaining its outputs using the LIME algorithm.

Evaluation Using MNIST

Dataset Description The second dataset is the MNIST dataset (Deng 2012), which contains images of size 28×28 that depict handwritten numerical digits between 0 and 9. The multi-class classification task is to determine which digit is depicted per image. The features used by the classification neural network are the pixels of the 28×28 image that are converted into a one-dimensional vector of length 784. The pixel values are normalized by dividing each by 255 (the maximum pixel value). The hidden layer architecture for the classification task consists of 2 hidden layers of width 75 neurons each and is trained for 100 epochs.

Neural Networks Table 1 provides the architecture and training epoch information for neural networks used for the MNIST dataset. For the classification network, hidden layers use the Rectified Linear Unit (ReLU) as their activation functions, and the output layer uses the softmax function as its activation function. Dropout is applied before each hidden layer to calculate epistemic uncertainty (Gal and Ghahramani 2016). Binary cross-entropy is optimized using the ADAM optimization algorithm (Kingma and Ba 2014).

For the uncertainty regression network, mean squared error is optimized using the ADAM optimization algorithm (Kingma and Ba 2014). Hidden layers are activated using the Rectified Linear Unit (ReLU) activation function. Dropout is strictly used as a regularization technique (Srivastava et al. 2014).

Experiment Setup We use LRP (Bach et al. 2015) as the explanation algorithm. We generate explanations of uncertainty by using a regression model to estimate Bayesian dropout uncertainty and explain its outputs using the LRP algorithm.

For each of the below classification tasks, we also generate explanations using LRP-$\alpha\beta$ with $\alpha = 2$ and $\beta = 1$. These parameters for LRP-$\alpha\beta$ are popular parameter values for LRP that allow empahsis on positive relevance while incorporating the effects of negative relevance (Sun 2021; Kohlbrenner et al. 2020). This will allow us to investigate if a potential relationship exists between input feature uncertainty and negative relevance of classification explanations. These explanations are denoted as $e_{\text{LRP}-\alpha\beta}(f(x_1, x_2))$

To evaluate the effectiveness of the proposed techniques on explaining out-of-distribution uncertainty, we train a classification neural network on the full MNIST dataset and multi-class classification task. We then generate uncertainty estimations and explanations of uncertainty having the model infer using 100 randomly selected images from the Fashion-MNIST dataset. The Fashion-MNIST (FMNIST) datset contains 28×28 grayscale images of various clothing items including handbags, pants, shirts, and shoes. Since the data type matches that of MNIST, it can easily serve as an unseen data source for MNIST-trained neural networks (Postels et al. 2021). This experiment will be referred to as the MNIST/FMNIST experiment in the Results and Discussion sections.

Finally, we consider the effects of uncertainty explanations for explaining in-distribution uncertainty. To do this, we construct a binary classification task by training the classification network to distinguish 0's and 6's. We then evaluate our technique with unseen data with results averaged from the following digits: 0, 1, 2, 3. We randomly select 100 testing images from each digit, resulting in 400 images total. This experiment will be referred to as the "Binary MNIST" experiment in the Results and Discussion sections.

Table 2: Out-of-Distribution Coordinate Identification Rate by Uncertainty Explanation Technique

Scenario	Exp. of Unc.	True Uncertainty Value
1 - x_1 outside	0.922	0.16
2 - x_2 outside	0.381	0.17
3 - x_1 and x_2 outside	0.547	0.22
Training Data	N/A	0.11

Results

In this section, we present quantitative results from the evaluation of the proposed techniques.

Synthetic Data

We first report results from the experiments on the synthetic data. Using the scenarios described above, we evaluate the rate at which the uncertainty explanation techniques can correctly identify which coordinate lies at a greater distance from the training distribution bounding box (defined by the values $\alpha_1, \alpha_2, \beta_1, \beta_2$). Table 2 depicts these results. Note that we do not consider the scenario that consists of points within the training distribution as this is an evaluation of out-of-distribution data. We also report the average dropout uncertainty for the training data.

MNIST

Our next experiment investigates the similarity between uncertainty explanations and negative relevance from classification explanations. We do this through measuring the size of the intersection of pixels that have negative classification relevance with the pixels that have a high magnitude in their uncertainty explanations. We measure this intersection in increments using the top $X\%$ of pixels with the highest magnitude in their uncertainty explanation, where $X \in [25, 50, 75, 90]$. This experiment will help us quantitatively determine if uncertainty explanations capture similar information to negative relevance in classification explanations. Table 3 provides these results for both the FMNIST experiment and the Binary MNIST experiment.

Our final experiment calculates the Pearson correlation coefficient when comparing the explanation of uncertainty

Table 3: Comparing the percentage of shared pixels in the intersection of negatively relevant classification pixels to the subset of the highest magnitude pixels in uncertainty explanations for the MNIST-based experiments experiment.

Experiment	25%	50%	75%	90%
MNIST/FMNIST	50.15	88.37	99.46	100
Binary MNIST	98.47	100	100	100

Table 4: Pearson Correlation Coefficient when comparing the explanation of dropout uncertainty with the classification explanations with only negative relevance values (all positive relevance is set to 0). The One-Tailed P-value is the result of a one-sided T-Test determining if the mean of the correlation coefficients is statistically significantly less than 0.

Experiment	Pearson's r	One-Tailed P-Value
MNIST/FMNIST	-0.13	2.2E-16
Binary MNIST	-0.28	8.1E-5

with the negative relevance of classification explanations. In this scenario, positive relevance is set to 0 so only negative relevance is considered. If negative relevance explains uncertainty or is some component of epistemic uncertainty, then there should be a negative correlation between the two as they would be "inverses". We also calculate the p-value for a two-tailed test indicating the probability that the given p-value could originate from 2 randomly-sampled distributions. Finally, we present the p-value from a one-tailed T-test to determine if the mean of the correlation coefficients is indeed negative (< 0). Table 4 contains the result of this experiment.

Discussion

In this section, we present a discussion of our results and possible options for operationalizing our technique.

Results on Synthesis Data

The primary goal of this simple binary classification simulation was to ascertain if techniques for mapping epistemic uncertainty can identify 1) which feature caused a data point to be out-of-distribution and 2) which feature contributed most if both features caused a data point to be out-of-distribution. We can force uncertainty to increase by manipulating the values of x_1 and x_2.

Table 2 provides the accuracy rate at which input explanations can identify coordinate contribution to epistemic uncertainty. In Scenario 1, coordinate x_1 was fixed to be outside the bounding box of the training distribution. The explanation of uncertainty can identify x_1 as the contributing coordinate with 92.2% accuracy. A dramatic change in performance occurs when identifying x_2 as the contributing coordinate. Explanation vectors generated by explaining predicted uncertainty were 38.1% accurate.

When both x_1 and x_2 were outside the training distribution bounding box, we considered the accuracy at predicting which coordinate was at a farther distance from the bounding box. Explanations of uncertainty were 54.7% accuracy was noted for explanations of predicted uncertainty.

From the dropout uncertainty averages given in Table 2, we can see that manipulating the values of the x_1 and x_2 coordinates so input is out-of-distribution does result in an increased uncertainty compared to training data. The explanation of uncertainty using LIME is able to identify x_1 as the out-of-distribution coordinate with 92% accuracy; however, the technique does not hold for when x_2 is the out-of distribution coordinate (38% accuracy) nor when both coordinates are out-of-distribution (55% accuracy). Although this technique is accurate at approximating dropout uncertainty (Brown and Talbert 2022), it is possible that the distillation of uncertainty into this model creates a high-fidelity model that does not truly capture the rationale of the causes of epistemic uncertainty.

Further research in this direction could include developing and extending current XAI techniques to recent advances in direct uncertainty measures (Postels et al. 2021) as a technique to explaining uncertainty. The drawback to this approach is these techniques do not model uncertainty from variational inference techniques but generate their own uncertainty value.

MNIST Results

We now discuss aspects of the experiments performed on the MNIST datasets.

Uncertainty Explanation Techniques and Negative Relevance Table 3 contain the percent of shared pixels in the intersection of negatively relevant classification pixels to the X% of the highest magnitude pixels from the uncertainty explanation. For the Binary MNIST experiments, we see 98% of the pixels that had negative classification relevance were contained in 25% of the pixels that contributed most to epistemic uncertainty, according to our uncertainty explanation techniques. For the MNIST/FMNIST experiments, 85% - 88% of the most negatively relevant pixels are contained in 50% of the pixels that contributed most to epistemic uncertainty.

This implies there is a relationship between negative relevance in classification explanations and magnitude of contribution of a pixel to epistemic uncertainty. Particularly, negative relevance pixels are contained within the top 25-50% of the pixels in the explanation of uncertainty are associated with negatively relevant pixels in the classification explanation.

This does not tell the full story, however. It is not impossible that there are few negatively relevant pixels. Thus, we consider the correlation between explanation of uncertainty and the negatively relevant pixels in the classification explanation. If negative relevance is a component of epistemic uncertainty, then there should be a negative correlation between these vectors. As negative relevance of a pixel increases, so should its influence on making the decision uncertain, which should result in a negative correlation

between the vectors. Table 4 provides Pearson's r correlation coefficient for the MNIST/FMNIST and Binary MNIST experiments. For the Binary MNIST experiment, the correlation coefficient is -0.28, and for the MNIST/FMNIST experiment, it is -0.13 with these coefficients statistically significantly negative (p-values of 8.1E-5 and 2.2E-16, respectively). Both of these experiments, there exists a slight negative correlation between the explanation of uncertainty and negatively relevant pixels to classification. This further supports the existence of relationship between uncertainty and negative relevance.

Conclusion

In this work, we present an XAI-based technique to attempt to measure feature attribution to Bayesian dropout-measured epistemic uncertainty. We do this through distilling epistemic uncertainty into a machine learning algorithm and applying common feature attribution algorithms (LIME and LRP). We assessed this technique using both a 2-dimensional synthetic dataset and variations of the MNIST dataset as a proof of concept. We determine through the synthetic experiments that uncertainty explanations using our simplistic technique do not adequately capture the rationale behind epistemic uncertainty for out-of-distribution data detection. Through our MNIST experimentations, however, we uncover a relationship between explanations of uncertainty and negative classification relevance.

There are several opportunities for future work. First is determining development of other distillation techniques that could possibly explain the rationale of epistemic uncertainty more effectively. The second is adapting existing XAI techniques to current direct uncertainty measures to determine if these techniques capture the rationale behind epistemic uncertainty. Finally, further work is needed to completely capture the relationship between negative relevance and uncertainty explanations and create techniques that utilize this information for safety-critical domains.

References

Ahmad, M. A.; Patel, A.; Eckert, C.; Kumar, V.; and Teredesai, A. 2020. Fairness in machine learning for healthcare. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3529–3530.

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 10(7).

Begoli, E.; Bhattacharya, T.; and Kusnezov, D. 2019. The Need for Uncertainty Quantification in Machine-Assisted Medical Decision Making. *Nature Machine Intelligence* 1(1):20–23.

Brown, K. E., and Talbert, D. A. 2019. Estimating uncertainty in deep image classification. In *AMIA*.

Brown, K. E., and Talbert, D. A. 2022. Uncertainty quantification in multimodal ensembles of deep learners. In *Submitted to The Thirty-Fifth International Flairs Conference*.

Bykov, K.; Höhne, M. M.-C.; Müller, K.-R.; Nakajima, S.; and Kloft, M. 2020. How much can i trust you?–quantifying uncertainties in explaining neural networks. *arXiv preprint arXiv:2006.09000*.

Bykov, K.; Höhne, M. M.-C.; Creosteanu, A.; Müller, K.-R.; Klauschen, F.; Nakajima, S.; and Kloft, M. 2021. Explaining bayesian neural networks. *arXiv preprint arXiv:2108.10346*.

Chai, L. R. 2018. Uncertainty estimation in bayesian neural networks and links to interpretability.

Deng, L. 2012. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine* 29(6):141–142.

Frosst, N., and Hinton, G. 2017. Distilling a neural network into a soft decision tree. *arXiv preprint arXiv:1711.09784*.

Gal, Y., and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of The 33rd International Conference on Machine Learning*, 1050–1059.

Jain, M.; Lahlou, S.; Nekoei, H.; Butoi, V.; Bertin, P.; Rector-Brooks, J.; Korablyov, M.; and Bengio, Y. 2021. Deup: Direct epistemic uncertainty prediction. *arXiv preprint arXiv:2102.08501*.

Kendall, A., and Gal, Y. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, 5574–5584.

Kingma, D. P., and Ba, J. 2014. Adam: A Method for Stochastic Optimization.

Kohlbrenner, M.; Bauer, A.; Nakajima, S.; Binder, A.; Samek, W.; and Lapuschkin, S. 2020. Towards best practice in explaining neural network decisions with lrp. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–7. IEEE.

Leibig, C.; Allken, V.; Ayhan, M. S.; Berens, P.; and Wahl, S. 2017. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific reports* 7(1):1–14.

Michelmore, R.; Kwiatkowska, M.; and Gal, Y. 2018. Evaluating uncertainty quantification in end-to-end autonomous driving control. *arXiv preprint arXiv:1811.06817*.

Montavon, G.; Samek, W.; and Müller, K.-R. 2018. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* 73:1–15.

Pedregosa, F., et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Postels, J.; Segu, M.; Sun, T.; Van Gool, L.; Yu, F.; and Tombari, F. 2021. On the practicality of deterministic epistemic uncertainty. *arXiv preprint arXiv:2107.00649*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15(1):1929–1958.

Sun, L. 2021. The effects of α, β and patch size on the performance of layer wise relevance propagation. Master's thesis, Tennessee Technological University.

Van Amersfoort, J.; Smith, L.; Teh, Y. W.; and Gal, Y. 2020. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*, 9690–9700. PMLR.