Beyond Confidence: Evidential Learning for Robust Classification and Uncertainty Quantification

Amir Aghdam

Abstract

This study investigates the limitations of traditional cross-entropy loss in classification tasks and presents evidential deep learning as an effective alternative. Cross-entropy models often yield overconfident predictions, particularly for out-of-distribution samples, as they lack mechanisms to quantify uncertainty. Evidential deep learning addresses this issue by parameterizing a Dirichlet distribution to model class probabilities, enabling the estimation of both aleatoric and epistemic uncertainties.

Our experiments, conducted on a fine-grained classification task, demonstrate that the evidential model consistently outperforms the cross-entropy model in terms of accuracy, precision, recall, and F1 score. GradCAM visualizations reveal that the evidential model focuses on more meaningful and discriminative features, while t-SNE representations highlight its superior class boundaries to capture uncertainity. Furthermore, uncertainty analysis shows that the evidential model exhibits higher uncertainty in incorrect predictions, reflecting its ability to capture ambiguity and insufficient evidence, whereas the cross-entropy model remains overconfident.

The findings demonstrate the advantages of evidential deep learning for both classification performance and uncertainty quantification. All code and implementations are made publicly available at `https://github.com/aghdamamir/evidential-classification`, encouraging further research in this domain.

1 Introduction

Deep learning models have achieved remarkable success across various classification tasks. Traditionally, these models are trained using the cross-entropy loss function, which encourages the network to maximize the probability of the correct class. While effective in many scenarios, this approach often leads to overconfident predictions, especially when the model encounters out-of-distribution (OOD) samples or ambiguous data points [2]. Such overconfidence poses significant challenges in applications where understanding the model's uncertainty is crucial.

Evidential Deep Learning (EDL) offers a compelling alternative by integrating principles from evidence theory into the learning process. Instead of directly predicting class probabilities, EDL models estimate the parameters of a Dirichlet distribution, representing a distribution over possible categorical distributions [5]. This formulation allows the model to quantify both aleatoric uncertainty (inherent data uncertainty) and epistemic uncertainty (model uncertainty) in a unified framework.

By interpreting network outputs as evidence, EDL accumulates support for different classes, forming beliefs based on the input data. This approach not only provides a measure of confidence in predictions but also facilitates uncertainty estimation through the properties of the Dirichlet distribution. Furthermore, visualizations such as Gradient-weighted Class Activation Mapping (Grad-CAM) can demonstrate that EDL models focus on more meaningful features during decision-making, leading to more interpretable and reliable predictions [1].

In this study, we compare the performance of traditional cross-entropy-based models with evidential deep learning models in terms of predictive accuracy and uncertainty estimation. We also

utilize Grad-CAM to visualize and analyze the learned representations, highlighting the advantages of EDL in capturing relevant features and providing calibrated uncertainty measures.

2 Methodology

2.1 Cross-Entropy Loss in Classification

In traditional classification tasks, models are trained to minimize the cross-entropy loss, which measures the dissimilarity between the true label distribution and the predicted probability distribution. For a given input $\mathbf{x}$ with true label y and predicted probabilities $\hat{\mathbf{p}} = [\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_C]$, where C is the number of classes, the cross-entropy loss is defined as:

$$\mathcal{L}_{\text{CE}} = -\log \hat{p}_y, \tag{1}$$

where $\hat{p}_y$ is the predicted probability for the true class y, obtained from the softmax function:

$$\hat{p}_c = \frac{\exp(z_c)}{\sum_{j=1}^{C} \exp(z_j)}, \quad \forall c \in \{1, \ldots, C\}. \tag{2}$$

Here, z_c represents the raw logits (pre-softmax activations) of the model for class c. The cross-entropy loss encourages the model to assign high probabilities to the correct class while penalizing incorrect predictions. However, this approach often results in overconfident predictions, particularly for out-of-distribution (OOD) samples [2].

2.2 Evidential Deep Learning

Evidence theory, also known as Dempster-Shafer theory, provides a mathematical framework for reasoning under uncertainty. It extends traditional probability theory by allowing the representation of partial knowledge. In this work, we leverage evidence theory to estimate classification probabilities and quantify uncertainty through the parameters of a Dirichlet distribution.

2.2.1 Dirichlet Distribution Parameterization

For a classification task with C classes, the Dirichlet distribution is parameterized by concentration parameters $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_C]$, where $\alpha_c > 0$ for all c. Its probability density function is given by:

$$\text{Dir}(\mathbf{p} \mid \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{c=1}^{C} p_c^{\alpha_c - 1}, \tag{3}$$

where $\mathbf{p} = [p_1, p_2, \ldots, p_C]$ is the vector of class probabilities such that $\sum_{c=1}^{C} p_c = 1$. The normalization term $B(\boldsymbol{\alpha})$ is the multivariate Beta function, defined as:

$$B(\boldsymbol{\alpha}) = \frac{\prod_{c=1}^{C} \Gamma(\alpha_c)}{\Gamma\left(\sum_{c=1}^{C} \alpha_c\right)}, \tag{4}$$

with $\Gamma(\cdot)$ denoting the Gamma function. In EDL, the network predicts evidence values $\mathbf{e} = [e_1, e_2, \ldots, e_C]$, which are then used to compute the Dirichlet parameters as:

$$\alpha_c = e_c + 1, \quad \forall c \in \{1, \ldots, C\}. \tag{5}$$

2.2.2 Class Probability Prediction

Using the obtained Dirichlet distribution, the predicted class probabilities $\hat{\mathbf{p}}$ are given by the mean of the distribution:

$$\hat{p}_c = \frac{\alpha_c}{\sum_{j=1}^{C} \alpha_j}, \quad \forall c \in \{1, \ldots, C\}. \tag{6}$$

This formulation allows the model to output probabilities that reflect the accumulated evidence for each class, as well as the uncertainty associated with the prediction.

2.2.3 Uncertainty Quantification

The Dirichlet distribution provides a framework for quantifying both aleatoric and epistemic uncertainties:

- **Total Uncertainty**: Measured by the entropy of the predictive distribution:

$$\text{Entropy} = -\sum_{c=1}^{C} \hat{p}_c \log \hat{p}_c. \tag{7}$$

- **Epistemic Uncertainty**: Related to the precision of the Dirichlet parameters:

$$\text{Epistemic Uncertainty} = \frac{C}{\sum_{c=1}^{C} \alpha_c}. \tag{8}$$

- **Aleatoric Uncertainty**: Captures inherent noise in the data, calculated as:

$$\text{Aleatoric Uncertainty} = \text{Entropy} - \text{Epistemic Uncertainty}. \tag{9}$$

2.2.4 Loss Function

To train the EDL model, a loss function is employed that balances the fit to the data with the quantification of uncertainty. The loss function comprises two components:

$$\mathcal{L}_{\text{EDL}} = \mathcal{L}_{\text{NLL}} + \lambda \mathcal{L}_{\text{KL}}, \tag{10}$$

where:

- $\mathcal{L}_{\text{NLL}}$ is the negative log-likelihood loss, encouraging the model to fit the data.
- $\mathcal{L}_{\text{KL}}$ is a Kullback-Leibler divergence term, regularizing the predicted Dirichlet distribution towards a prior distribution, promoting reasonable uncertainty estimates.
- λ is a hyperparameter controlling the trade-off between the two components.

This loss function enables the model to learn both accurate predictions and meaningful uncertainty estimates.

2.3 Gradient-weighted Class Activation Mapping (Grad-CAM)

Grad-CAM provides visual explanations of deep learning models by highlighting the regions of an input that are most influential in the model's prediction [4]. Grad-CAM computes a localization map $\mathbf{L}_{\text{Grad-CAM}}$ for a target class c as follows:

$$L^c_{\text{Grad-CAM}}(i,j) = \text{ReLU}\left(\sum_k \alpha^c_k A_k(i,j)\right), \tag{11}$$

where:

- $A_k(i,j)$ represents the activation of the k-th feature map at spatial location (i,j).
- α^c_k is the importance weight of the k-th feature map for class c, computed as:

$$\alpha^c_k = \frac{1}{Z}\sum_{i,j}\frac{\partial y_c}{\partial A_k(i,j)}, \tag{12}$$

  where y_c is the score for class c (logit output), and Z is a normalization constant (total number of spatial locations).

- The ReLU ensures that only positive contributions are considered in the localization map.

Grad-CAM provides insight into the decision-making process of the model by visualizing the spatial regions that contribute most to the model's predictions.

2.4 DINO v2 Vision Transformer

DINO v2 (Distillation with No Labels) is a foundational self-supervised Vision Transformer (ViT) that excels in learning visual representations without the need for labeled data. Its architecture is based on the standard Transformer model, which processes input data as a sequence of tokens.

In the ViT framework, an input image is divided into patches, each linearly embedded into a vector. A special learnable token, known as the `[CLS]` (classification) token, is prepended to the sequence of patch embeddings. This `[CLS]` token serves to aggregate global information from the image, as it attends to all other tokens during the Transformer's self-attention operations.

For our classification task, we utilize the pretrained DINO v2 Small model with a patch size of 14. This choice is advantageous in scenarios with limited data, as the model has already learned rich visual features from extensive pretraining. To adapt the model to our specific task, we modify its architecture as follows:

- **Head Removal**: We remove the original classification head from the pretrained DINO v2 model.
- **Addition of Fully Connected Layer**: We introduce a new fully connected layer that takes the output of the `[CLS]` token and produces either class probabilities or the parameters of a Dirichlet distribution, depending on the desired output.

During fine-tuning, **we freeze the weights of the pretrained network except for the last four layers**. This strategy allows the model to adapt to the specific nuances of our dataset while retaining the robust features learned during pretraining.

3 Experiments

3.1 Dataset

For our experiments, we use the **Flowers Dataset** from Kaggle [3], which consists of 3,670 images of variable sizes across 5 classes. The classes include: Daisy, Dandelion, Roses, Sunflowers, and Tulips. This fine-grained dataset includes a diverse collection of flower images characterized by high intra-class variability (e.g., variations in color, angle, and lighting) and low inter-class variability, making it a challenging benchmark for classification.

Figure 1 illustrates representative samples from the dataset, showcasing the visual complexity.

Figure 1: Sample images from the Flowers dataset

3.2 Baseline

To evaluate the impact of evidential deep learning, we establish a comparison between two models for a direct comparison:

- A model trained with the traditional **cross-entropy loss** as baseline.
- A model trained with the **Dirichlet-based evidential loss**, allowing for uncertainty quantification.

Both models share the same architecture but differ in their loss functions.

3.3 Metrics

We assess the models using standard evaluation metrics:

- **Accuracy**: The fraction of correctly classified samples.
- **Precision**: The proportion of true positive predictions among all positive predictions.
- **Recall**: The proportion of true positive predictions among all actual positives.
- **F1 Score**: The harmonic mean of precision and recall.

These metrics provide a comprehensive view of model performance in terms of both classification accuracy and robustness to imbalances.

3.4 Uncertainty Evaluation

To validate our hypothesis regarding the overconfidence of cross-entropy models, we compare the **mean total uncertainty** for correct and incorrect predictions. The total uncertainty for a single prediction is calculated using:

$$\text{Total Uncertainty} = -\sum_{c=1}^{C} \hat{p}_c \log \hat{p}_c, \tag{13}$$

where $\hat{p}_c$ represents the predicted probability for class c. By comparing uncertainty levels, we aim to highlight the superior calibration of evidential models in reflecting prediction confidence.

3.5 Experimental Setting

Both models are fine-tuned for 10 epochs using the Adam-Weighted optimizer with a learning rate of 1×10^{-4}. We resize all images to a dimension of 224×224 and normalize them as pre-processing steps. The model architecture features a hidden dimension of 384. During fine-tuning, we observe the evolution of the training and validation performance metrics over epochs for both models. The dataset is split into 10% of validation and 10% of test data representing a similar distribution of classes similar to the original dataset. For final evaluations and analysis, **we select the best model over the validation set** for both cases.

Figure 2 and Figure 3 depict the fine-tuning progress for the cross-entropy and evidential models, respectively.

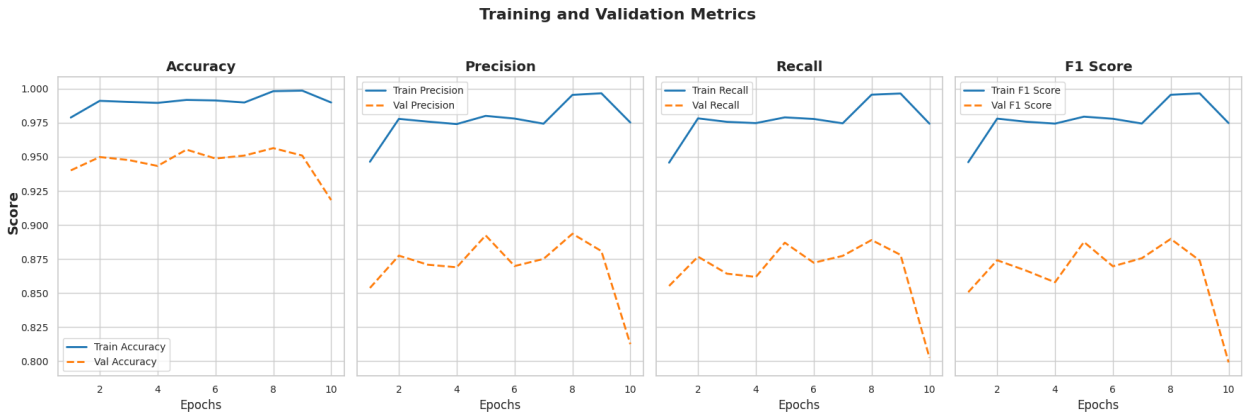


Figure 2: Training and validation performance of the cross-entropy model over 10 epochs.

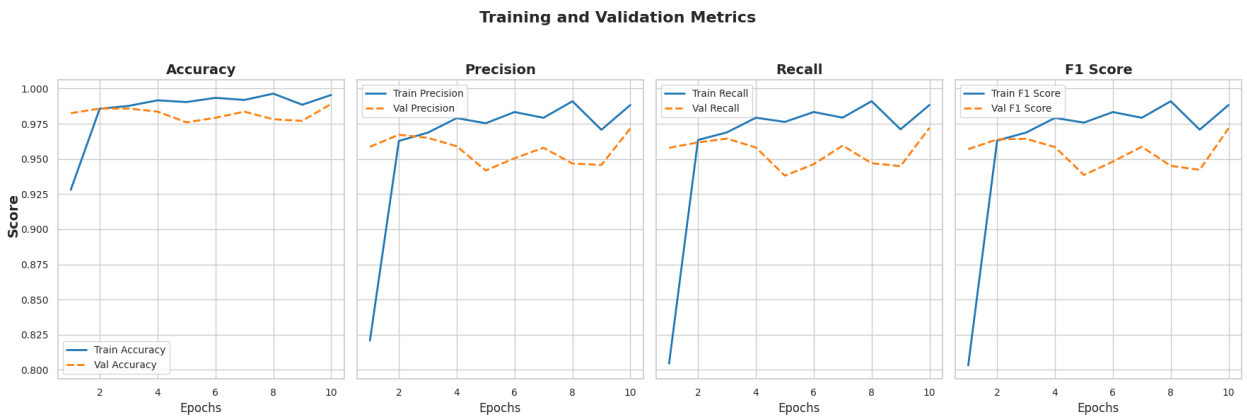


Figure 3: Training and validation performance of the evidential model over 10 epochs.

4 Results and Discussion

4.1 Quantitative Results

4.1.1 Classification Results

Table 1 presents the classification performance of both the cross-entropy and evidential models on the test dataset. Evidential deep learning demonstrates superior performance across all metrics, including accuracy, precision, recall, and F1 score. This improvement highlights the model's ability to make more informed predictions by leveraging the Dirichlet-based uncertainty framework.

Metric	Cross-Entropy Model	Evidential Model
Accuracy	94.55%	98.69%
Precision	86.32%	96.55%
Recall	86.07%	96.61%
F1 Score	86.14%	96.55%

Table 1: Classification results of cross-entropy and evidential models on the test dataset. Evidential deep learning consistently outperforms cross-entropy in all metrics.

4.1.2 Uncertainty Measures

Table 2 compares the mean total uncertainty of correct and incorrect predictions for both models. Evidential deep learning exhibits higher uncertainty in wrong predictions, which demonstrates its ability to identify cases with insufficient supporting evidence. In contrast, the cross-entropy model is overconfident, even for incorrect predictions.

Model	Correct	Incorrect
Evidential	0.445	0.630
Cross-Entropy	0.070	0.429

Table 2: Mean uncertainty comparison between models. Evidential deep learning better calibrates uncertainty, showing higher values for incorrect predictions.

4.2 Qualitative Results

4.2.1 GradCAM Visualization

To analyze the differences in feature attention between the models, we use GradCAM visualizations.

Figure 4 shows samples where both models predict correctly. The evidential model focuses on meaningful and discriminative features of the input, whereas the cross-entropy model exhibits scattered attention.

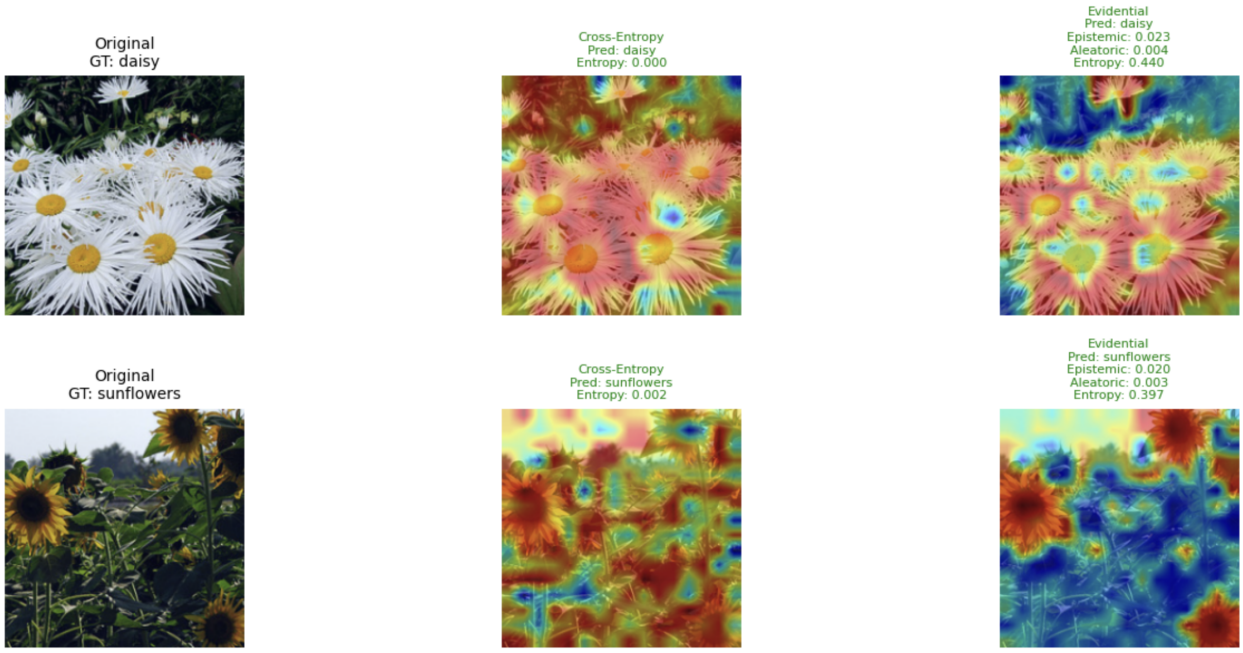


Figure 4: GradCAM visualizations for samples correctly classified by both models. The evidential model shows focused and meaningful feature activations.

Figure 5 highlights samples correctly classified by the evidential model but misclassified by the cross-entropy model. The evidential model demonstrates the ability to "know" when there is insufficient evidence or the presence of OOD samples, as indicated by its Dirichlet-based uncertainty.

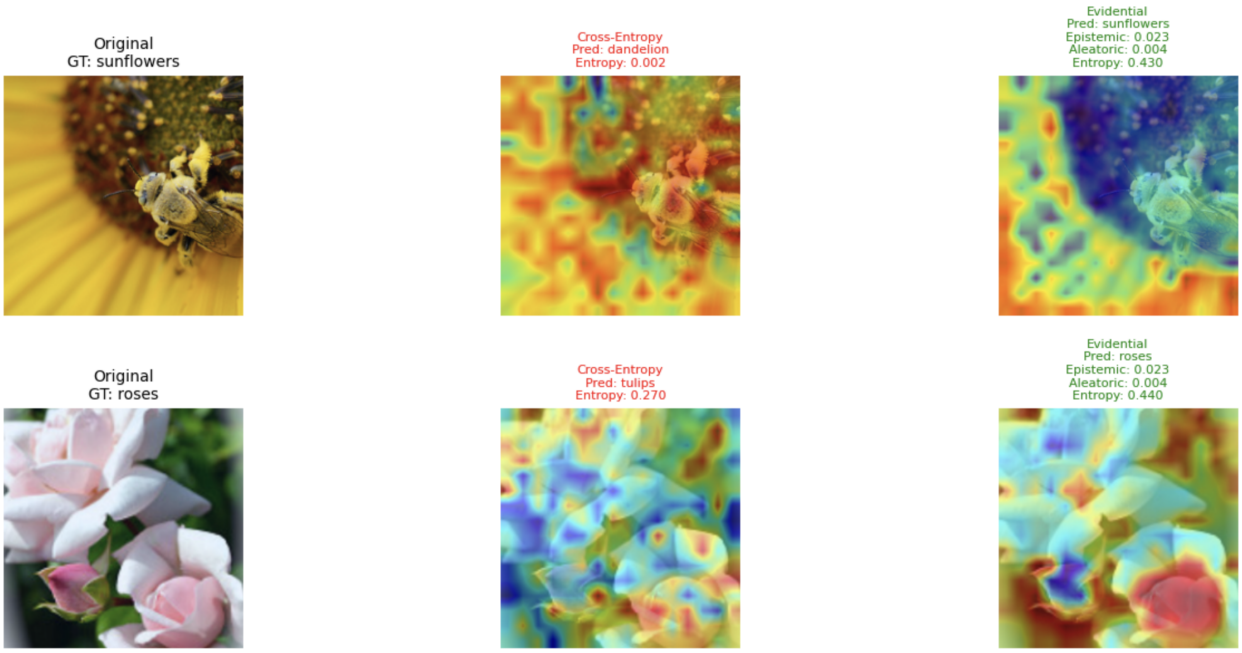
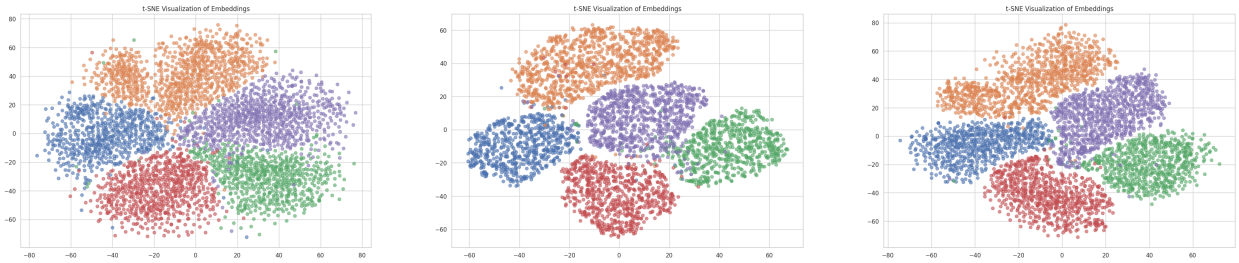


Figure 5: GradCAM visualizations for samples correctly classified by the evidential model but misclassified by the cross-entropy model. The evidential model focuses on key regions and reflects uncertainty for ambiguous inputs.

4.2.2 t-SNE Visualization

We use t-SNE to visualize the feature space representations learned by the models. Figure 6 illustrates three plots: the initial representations (pre-finetuning) and the final representations after fine-tuning for both models.

The evidential model's embeddings show better representation on uncertain regions compared to the cross-entropy model, suggesting its improved ability to learn meaningful representations.



Initial embeddings | Cross-Entropy Model | Evidential Model

Figure 6: t-SNE visualization of feature representations. (a) Initial embeddings before fine-tuning. (b) Final embeddings of the cross-entropy model. (c) Final embeddings of the evidential model. Evidential learning models uncertain regions more effectively.

4.2.3 Uncertainty Analysis

Figure 7 presents a scatter plot comparing the total uncertainty values between the evidential and cross-entropy models. Evidential models display higher uncertainty for incorrect predictions, which

indicates their ability to reflect ambiguity and insufficient evidence. In contrast, cross-entropy models demonstrate overconfidence in both correct and incorrect predictions.

The distinct separation in uncertainty distributions highlights the superior calibration of evidential models. This allows them to better distinguish between confident and ambiguous predictions, especially in out-of-distribution scenarios.

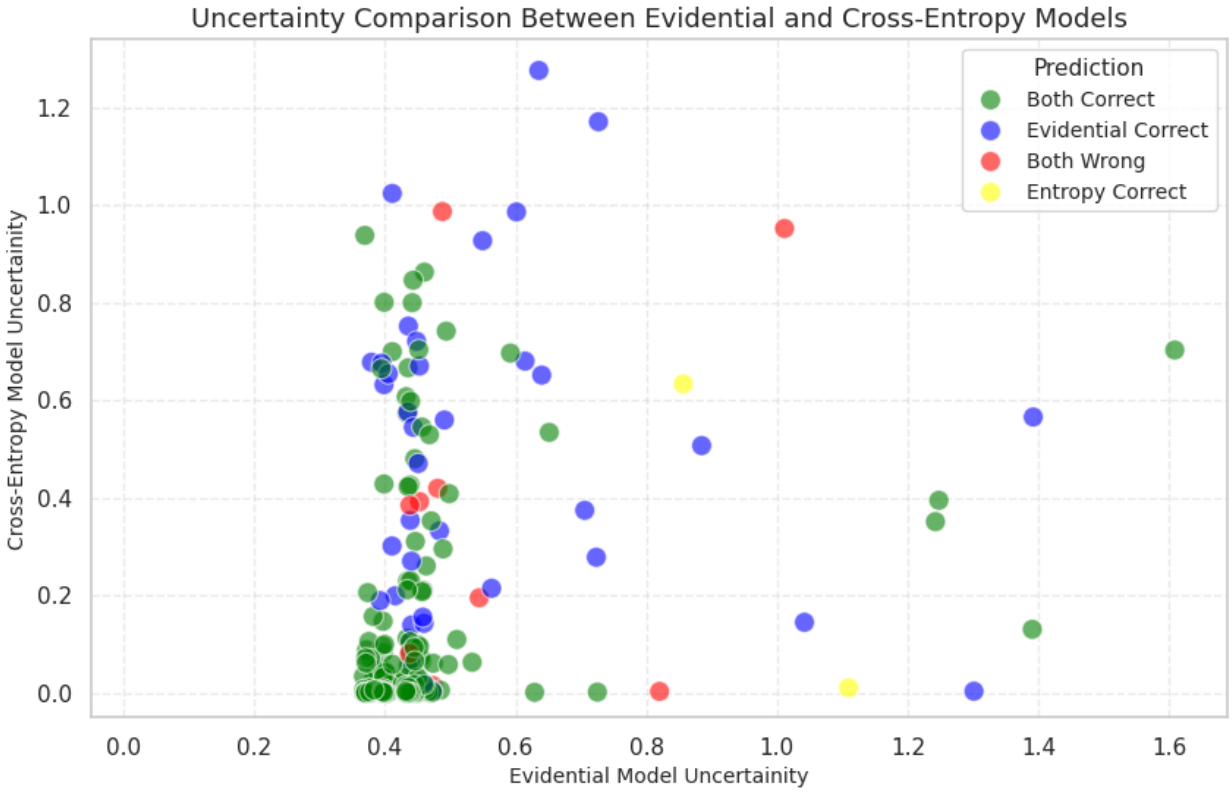


Figure 7: Uncertainty comparison between evidential and cross-entropy models. Evidential model exhibit higher uncertainty for incorrect predictions, while cross-entropy models remain overconfident overall.

5 Conclusion

In this study, we demonstrated the advantages of evidential deep learning over traditional cross-entropy loss for classification tasks. Our experiments revealed that the evidential model consistently outperformed the cross-entropy model across all key metrics, including accuracy, precision, recall, and F1 score.

Through GradCAM visualizations, we observed that the evidential model attends to more meaningful and discriminative features, while the cross-entropy model often shows scattered attention. Furthermore, t-SNE embeddings highlighted more meaninngful class boundaries achieved by the evidential model, emphasizing its capability to identify uncertainty and learn discriminative representations.

Uncertainty analysis confirmed that the evidential model better captures predictive uncertainty, exhibiting higher uncertainty in incorrect predictions. In contrast, the cross-entropy model displayed overconfidence, even in incorrect predictions. This ability to quantify uncertainty makes the evidential model more reliable, particularly in scenarios involving ambiguous or out-of-distribution

data.

These findings shows the potential of evidential deep learning to provide both improved performance and better uncertainty calibration in challenging classification tasks.

References

[1] Danruo Deng, Guangyong Chen, Yang Yu, Furui Liu, and Pheng-Ann Heng. Uncertainty estimation by fisher information-based evidential deep learning. *arXiv preprint arXiv:2303.02045*, 2023.

[2] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[3] Kaggle. Flowers recognition dataset, 2023. URL `https://www.kaggle.com/datasets/rahmasleam/flowers-dataset`.

[4] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[5] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, pages 3179–3189, 2018.