Thesis Report

The measurement of uncertainty in deep learning prediction

by

Farhana Reza
16301025

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
January 2025

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Farhana Reza

16301025

Approval

The thesis/project titled "The measurement of uncertainty in deep learning prediction" submitted by

1. Farhana Reza (16301025)

Of Fall, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on Summer, 2025.

Examining Committee:

Supervisor:
(Member)

Moin Mostakim
Senior Lecturer
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Md. Golam Rabiul Alam, PhD
Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Ethics Statement (Optional)

This document has undergone language and grammar refinement using ChatGPT. No technical content has been generated or altered by AI.

Abstract

Uncertainty estimation is crucial for improving the reliability and robustness of image classification systems, particularly in safety-critical domains such as healthcare and autonomous driving. In this work, we propose a novel Enhanced Multi-Head Evidential Fusion (MCEF) framework that leverages multi-head evidential networks, meta-calibration, and attention-based fusion to provide comprehensive uncertainty quantification while improving predictive accuracy. Our method integrates diverse evidence heads with adaptive calibration and uncertainty-guided attention to capture both aleatoric and epistemic uncertainties, as well as head-level disagreement, enabling more reliable confidence estimation.

We validate the proposed method on standard benchmark datasets, including CIFAR-10, MNIST, and Fashion-MNIST, demonstrating that it consistently outperforms existing relevant methods in terms of classification accuracy while providing rich uncertainty decomposition. The proposed approach not only advances the state of uncertainty-aware image classification but also provides a robust foundation for reliable deployment in real-world applications where accurate predictions and uncertainty awareness are critical.

Keywords: Uncertainty Estimation, Evidential Deep Learning, Multi-Head Evidential Fusion (MCEF), Meta-Calibration, Attention-Based Fusion, Aleatoric Uncertainty, Epistemic Uncertainty

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Uncertainty estimation has emerged as a cornerstone in the development of reliable and interpretable machine learning systems. In image classification tasks, deep learning models have achieved remarkable accuracy on benchmark datasets; however, their ability to know when they are uncertain remains limited. This limitation is particularly critical in domains where mistakes carry significant consequences. For instance, in healthcare, a misclassified medical image could lead to a wrong diagnosis or treatment recommendation, while in autonomous driving, misidentifying an object could result in accidents or loss of life. Therefore, quantifying uncertainty is not merely an academic exercise but a practical necessity for ensuring the safe deployment of machine learning systems in the real world.

Traditional deep learning models typically produce point estimates, i.e., single deterministic predictions, without providing a measure of confidence. While techniques such as Monte Carlo (MC) dropout and deep ensembles have been proposed to estimate uncertainty, they often capture only epistemic uncertainty—uncertainty due to model parameters—and overlook aleatoric uncertainty, which arises from noise or ambiguity inherent in the data itself. Moreover, most of these approaches treat uncertainty estimation as a post-hoc analysis rather than integrating it directly into model training. This disconnection limits the model's ability to adaptively improve predictions in regions of high uncertainty and reduces overall reliability.

To address these limitations, we propose a novel Enhanced Multi-Head Evidential Fusion (MCEF) framework that simultaneously improves predictive accuracy and provides comprehensive uncertainty quantification. By leveraging multiple evidence heads, a meta-calibration network, and an attention-based fusion mechanism, the MCEF framework allows the model to capture diverse predictive signals, calibrate confidence scores, and measure both aleatoric and epistemic uncertainties effectively. This approach provides not only more accurate predictions but also richer insights into model confidence, enabling safer and more reliable decision-making in real-world applications.

1.2 Motivation

The growing adoption of machine learning in high-stakes domains underscores the need for models that are both accurate and uncertainty-aware. Existing uncer-

tainty estimation methods present several gaps that motivate the development of the MCEF framework:
Incomplete Uncertainty Coverage: Traditional approaches primarily estimate epistemic uncertainty through model ensembles or dropout sampling. Aleatoric uncertainty, which captures inherent variability in data (e.g., noisy or ambiguous images), is often neglected. Without modeling both types of uncertainty, the model's confidence estimates can be misleading.
Limited Training Integration: Most methods treat uncertainty as an afterthought, applying it only during evaluation. As a result, models cannot leverage uncertainty to guide learning, for example, by focusing more on uncertain samples or regularizing overconfident predictions.
Lack of Diversity Awareness: Single-headed models or naive ensembles often fail to capture disagreements between alternative predictive hypotheses. Measuring head-level disagreement can provide valuable insight into the reliability of predictions, yet few frameworks exploit this information.
Dataset Variability: Image datasets vary significantly in complexity, from simple grayscale digits (MNIST) to more challenging natural images (CIFAR-10) or structured fashion items (Fashion-MNIST). A robust uncertainty estimation framework must generalize across these diverse domains without sacrificing accuracy.
The MCEF framework addresses these challenges by combining multi-head evidential networks with meta-calibration and attention-based fusion, integrating uncertainty estimation directly into training, and ensuring that both aleatoric and epistemic uncertainties are captured comprehensively.

1.3 Problem Statement

Despite the progress in uncertainty estimation, current approaches suffer from several limitations that hinder their applicability in real-world image classification:
Partial uncertainty modeling: Many models focus on either epistemic or aleatoric uncertainty but do not provide a unified mechanism to quantify both. This can result in miscalibrated confidence scores and unreliable predictions.
Ineffective use of uncertainty during training: Without explicitly incorporating uncertainty into the learning process, models cannot adaptively prioritize difficult or ambiguous samples. This reduces the potential gains from uncertainty-guided optimization.
Limited interpretability and decomposition: Conventional methods often provide a single uncertainty score without explaining the underlying sources of uncertainty, making it harder to diagnose model failures.
Generalization across datasets: Many uncertainty estimation techniques are designed for specific datasets and fail to generalize across tasks with varying complexity and data distributions.
To address these issues, it is necessary to develop a framework that can:
Provide fine-grained uncertainty decomposition, including total, aleatoric, and epistemic components.
Leverage uncertainty during training to improve learning from difficult or ambiguous samples.
Ensure robust performance across diverse image classification datasets, such as MNIST, Fashion-MNIST, and CIFAR-10.

Incorporate head-level disagreement and attention-based fusion to maximize predictive reliability.

1.4 Research Objectives

The goal of this study is to develop and validate a novel Enhanced Multi-Head Evidential Fusion (MCEF) framework that integrates uncertainty estimation directly into the image classification pipeline while improving predictive accuracy. Specifically, this work aims to:

Design a multi-head evidential network: Implement multiple evidence heads with diverse activation functions to capture alternative hypotheses, enhancing the model's ability to quantify uncertainty and disagreement across predictions.

Incorporate meta-calibration and attention mechanisms: Use a meta-calibration network to adaptively scale and shift evidence, and attention-based fusion to weigh contributions from each head according to their uncertainty, improving both predictive accuracy and confidence estimation.

Develop uncertainty-aware training strategies: Introduce a composite loss function that integrates evidential loss, adaptive calibration loss, consistency loss, and diversity loss, ensuring that the model learns robust representations while leveraging uncertainty information during optimization.

Evaluate on benchmark datasets: Assess the framework on MNIST, Fashion-MNIST, and CIFAR-10, comparing performance with existing methods in terms of classification accuracy, uncertainty decomposition, confidence calibration, and head-level disagreement.

Demonstrate practical applicability: Provide a framework that can be readily deployed in high-stakes applications, offering both high predictive performance and reliable uncertainty estimates, laying the groundwork for safer, uncertainty-aware decision-making in real-world scenarios.

By achieving these objectives, the MCEF framework aims to advance the state-of-the-art in uncertainty-aware image classification, providing both enhanced accuracy and comprehensive uncertainty quantification, which are critical for reliable deployment in complex, real-world environments.

Chapter 2

Related Work

Machine learning (ML) models, particularly deep learning models, are notoriously difficult to interpret due to their inherent complexity and opacity. This lack of interpretability poses significant concerns and risks to trustworthiness, such as safety [11], security [3], and ethics [22]. Using ML models without adequate risk assessment can exacerbate these challenges. To address these issues, various risk assessment techniques have been developed for ML models [23], [18], [17], [20]. These techniques can be broadly categorized into two main approaches: (1) data distribution analysis and (2) uncertainty estimation. The data-centric nature of ML models necessitates careful consideration of data, particularly when there is a potential distributional discrepancy between training and test datasets. Techniques such as detecting distribution shifts [17] and identifying out-of-distribution samples [21] are commonly employed to mitigate risks associated with unseen data. There are various methods to measure uncertainty in Deep learning.
These methods calculate uncertainty using a deterministic neural network through a single forward pass. **Softmax-based Confidence:** The softmax probabilities, $\sigma(\mathbf{z})$, are computed as:

$$\sigma(\mathbf{z})_i = \frac{\exp(z_i)}{\sum_{j=1}^{C} \exp(z_j)}$$

where $\mathbf{z}$ represents the logits and C is the number of classes. High entropy or low maximum softmax probability, $H(\sigma(\mathbf{z}))$, indicates higher uncertainty:

$$H(\sigma(\mathbf{z})) = -\sum_{i=1}^{C} \sigma(\mathbf{z})_i \log \sigma(\mathbf{z})_i$$

Logits Variance: Uncertainty is derived from the variance of logits $\mathrm{Var}(\mathbf{z})$.
Bayesian methods place a distribution over network weights $\mathbf{w}$ to model uncertainty [6]. **Bayesian Neural Networks (BNNs):** Represent weights as $\mathbf{w} \sim p(\mathbf{w})$. The predictive distribution is given by:

$$p(\mathbf{y}|\mathbf{x}) = \int p(\mathbf{y}|\mathbf{x}, \mathbf{w}) p(\mathbf{w}) \, d\mathbf{w}$$

Monte Carlo (MC) Dropout: During inference, dropout is applied multiple times to sample T outputs $\{\hat{\mathbf{y}}_t\}_{t=1}^{T}$. Uncertainty is measured using variance:

$$\mathrm{Var}(\mathbf{y}) = \frac{1}{T} \sum_{t=1}^{T} (\hat{\mathbf{y}}_t - \bar{\mathbf{y}})^2$$

where $\bar{\mathbf{y}} = \frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{y}}_t$.
These methods use multiple models to estimate uncertainty [9]. **Deep Ensembles:** Train M models $\{f_m\}_{m=1}^{M}$ with different initializations. The mean prediction is:

$$\bar{\mathbf{y}} = \frac{1}{M}\sum_{m=1}^{M} f_m(\mathbf{x})$$

Uncertainty is derived from the prediction variance:

$$\mathrm{Var}(\mathbf{y}) = \frac{1}{M}\sum_{m=1}^{M}(f_m(\mathbf{x}) - \bar{\mathbf{y}})^2$$

Bagging: Train on bootstrap samples and aggregate predictions.
Test-Time Augmentation (TTA) is an approach for estimating uncertainty by introducing variations to the input data during inference [12]. Specifically, the input data $\mathbf{x}$ is augmented with a set of transformations $\{T_t\}_{t=1}^{T}$, such as flips, rotations, or other distortions. Each transformed input produces a corresponding prediction $\{\hat{\mathbf{y}}_t\}_{t=1}^{T}$. The variability across these predictions reflects the model's uncertainty. The variance of predictions is given by:

$$\mathrm{Var}(\mathbf{y}) = \frac{1}{T}\sum_{t=1}^{T}(\hat{\mathbf{y}}_t - \bar{\mathbf{y}})^2,$$

where the mean prediction is:

$$\bar{\mathbf{y}} = \frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{y}}_t.$$

High variability in predictions indicates greater uncertainty, signaling regions where the model may struggle due to ambiguity or limited training data. TTA is computationally efficient and flexible since it relies only on augmenting existing data and aggregating predictions. It is particularly effective in vision tasks, where transformations like cropping, resizing, and flipping are common. This method provides insights into model robustness by showing how sensitive predictions are to input perturbations. By averaging the predictions, TTA improves the reliability of predictions while offering a straightforward mechanism to quantify epistemic uncertainty without requiring major architectural modifications.
Stochastic Gradient-Based Approaches estimate uncertainty by incorporating stochasticity during training or inference [16]. A notable method in this category is Stochastic Weight Averaging-Gaussian (SWAG). SWAG models the posterior distribution of weights in a neural network as a Gaussian distribution:

$$\mathbf{w} \sim \mathcal{N}(\bar{\mathbf{w}}, \Sigma),$$

where $\bar{\mathbf{w}}$ represents the mean of the model weights, and Σ is the covariance matrix derived from stochastic samples during training.
To compute these, SWAG performs weight sampling at different points during training, capturing the variability in weight updates. This variability reflects epistemic uncertainty, which arises due to limited data or model ambiguity. During inference, multiple weight samples are drawn from the Gaussian distribution, and predictions are averaged to produce both a point estimate and uncertainty intervals.

SWAG combines the benefits of stochastic weight averaging (which improves generalization) with uncertainty estimation. It is especially useful in scenarios requiring probabilistic predictions, such as medical diagnostics or autonomous systems. Unlike traditional Bayesian methods, SWAG does not require a full probabilistic model of weights, making it computationally efficient while preserving interpretability.

Evidential Deep Learning explicitly models uncertainty by learning the evidence associated with each class in a classification task [15]. This approach relies on the Dirichlet distribution, parameterized as:

$$\mathbf{p} \sim \text{Dirichlet}(\boldsymbol{\alpha}), \quad \boldsymbol{\alpha} = \mathbf{1} + \mathbf{e}.$$

Here, $\mathbf{e}$ represents evidence learned by the neural network, and $\boldsymbol{\alpha}$ is the Dirichlet concentration parameter. The Dirichlet distribution captures the belief over possible class probabilities, inherently quantifying both epistemic and aleatoric uncertainty. Aleatoric uncertainty is derived from the amount of evidence collected for a given input, while epistemic uncertainty arises from limited or ambiguous training data. A low $\boldsymbol{\alpha}$ reflects high uncertainty, signaling insufficient evidence for confident predictions.

This method naturally combines predictive accuracy with uncertainty estimation. During training, the loss function encourages the model to fit data points confidently while penalizing overconfident predictions for uncertain or noisy data. Evidential Deep Learning finds applications in high-stakes fields such as healthcare or finance, where reliable uncertainty quantification is essential.

Confidence Calibration ensures that the predicted probabilities align closely with the actual likelihood of correctness [7]. This method typically involves adjusting the logits $\mathbf{z}$ output by a model using temperature scaling:

$$\mathbf{p} = \text{softmax}(\mathbf{z}/T),$$

where T is a temperature parameter optimized on a validation set to minimize calibration error. Larger values of T reduce confidence in predictions, while smaller values increase it.

A well-calibrated model provides probabilities that reflect true likelihoods. For example, if a model predicts a probability of 0.7 for a class, the class should be correct approximately 70% of the time. Calibration is crucial in domains where decisions are guided by predicted probabilities, such as weather forecasting or risk assessment.

Miscalibrated models, which are often overconfident, can lead to poor decision-making. Confidence Calibration corrects this by aligning probabilities, and enhancing model reliability and interpretability. The method is computationally efficient and widely applicable, improving trustworthiness in probabilistic outputs.

Variational Methods approximate the Bayesian posterior distribution $p(\mathbf{w}|\mathcal{D})$ over model weights by optimizing a variational distribution $q(\mathbf{w})$ [5]. The key objective is to minimize the Kullback-Leibler (KL) divergence between $q(\mathbf{w})$ and the true posterior:

$$\mathcal{L}_{\text{VI}} = \text{KL}(q(\mathbf{w}) \parallel p(\mathbf{w}|\mathcal{D})).$$

Bayes by Backprop is a prominent implementation of variational inference for neural networks. It uses backpropagation to update the parameters of $q(\mathbf{w})$, enabling the

model to capture uncertainty in weights. During inference, predictions are made by sampling weights from $q(\mathbf{w})$ and averaging the results.
Variational Methods are computationally efficient compared to exact Bayesian inference and provide uncertainty estimates critical in tasks like reinforcement learning or active learning. By modeling the posterior over weights, these methods enable robust decision-making under uncertainty.
Randomized prior functions combine random prior components with learned functions to model uncertainty [14]. The model output is represented as:

$$f(\mathbf{x}) = f_{\text{prior}}(\mathbf{x}) + f_{\text{learned}}(\mathbf{x}),$$

where $f_{\text{prior}}(\mathbf{x})$ introduces variability from a predefined random prior, and $f_{\text{learned}}(\mathbf{x})$ represents the learned model. The uncertainty is derived from the variability in the prior, enabling the model to account for epistemic uncertainty effectively.
Uncertainty-aware loss functions incorporate aleatoric uncertainty directly into the training process [8]. The modified loss function is expressed as:

$$\mathcal{L} = \frac{1}{2\sigma^2}||\mathbf{y} - \hat{\mathbf{y}}||^2 + \log \sigma,$$

where σ represents the model's estimate of data uncertainty. This approach allows the model to adapt to varying levels of noise in the data, leading to robust predictions in uncertain environments.
Functional priors integrate Gaussian processes (GPs) with neural networks, providing a probabilistic framework for uncertainty estimation [4]. A GP prior is defined as:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')),$$

where $m(\mathbf{x})$ is the mean function, and $k(\mathbf{x}, \mathbf{x}')$ is the kernel function. This combination allows the model to leverage the strengths of GPs for uncertainty-aware predictions while benefiting from the scalability of neural networks.
Gaussian Processes (GPs) at the output layer model predictions as probabilistic functions, allowing for uncertainty quantification [13]. A GP is defined as a collection of random variables where any finite subset follows a joint Gaussian distribution. The output $\mathbf{y}$ is assumed to follow a GP prior:

$$\mathbf{y} \sim \mathcal{GP}(\mathbf{f}, K),$$

where $\mathbf{f}$ is the mean function, and K is the kernel matrix, which captures the similarity between input points. GPs are particularly effective for regression, providing predictive means and variances to balance accuracy and uncertainty estimation.
Quantile regression estimates prediction intervals by modeling conditional quantiles of the output distribution [1]. For a given quantile $\tau \in [0, 1]$, the model minimizes the pinball loss function $L_\tau(y, \hat{y})$, defined as:

$$L_\tau(y, \hat{y}) = \max(\tau(y - \hat{y}), (1 - \tau)(\hat{y} - y)).$$

The predicted quantiles $\hat{y}_\tau$ represent the range within which the true output is expected to lie, providing uncertainty estimates. This approach is particularly useful for interval predictions in regression problems.

Conformal prediction constructs prediction sets with guaranteed statistical coverage [2]. For a given input $\mathbf{x}$, the prediction set is defined as:

$$\mathcal{C}(\mathbf{x}) = \{\mathbf{y} : S(\mathbf{x}, \mathbf{y}) \leq \epsilon\},$$

where S is a scoring function, and ϵ is a threshold calibrated to achieve the desired confidence level. Conformal prediction is model-agnostic, leveraging past observations to quantify uncertainty. It ensures reliable prediction intervals for both regression and classification tasks.

Evidential deep learning for regression replaces Bayesian neural networks' weight priors with priors directly over the likelihood function, allowing a model to represent uncertainty without sampling [19]. Specifically, a neural network is trained to output the parameters of a *Normal-Inverse-Gamma (NIG)* distribution, $p(\mu, \sigma^2 \mid \gamma, \nu, \alpha, \beta)$, which serves as a conjugate prior over Gaussian likelihoods. From this distribution, the predictive mean is $\hat{y} = E[\mu] = \gamma$, the **aleatoric uncertainty** is $E[\sigma^2] = \frac{\beta}{\alpha-1}$, and the **epistemic uncertainty** is $\mathrm{Var}[\mu] = \frac{\beta}{\nu(\alpha-1)}$. Training is guided by maximizing the marginal likelihood of the data, which corresponds to minimizing the negative log-likelihood (NLL) under a Student-t distribution. To prevent the model from assigning strong evidence when predictions are wrong, a regularization penalty is introduced, $L_R = |y - \gamma| \cdot (2\nu + \alpha)$, which increases when large errors occur with high confidence. The overall loss thus combines data fit and regularization: $L = L_{\mathrm{NLL}} + \lambda L_R$. This framework produces meaningful uncertainty that grows in regions without training data, while maintaining competitive predictive accuracy. Empirical results show that it yields sharper and better-calibrated uncertainty estimates in regression tasks, improves robustness to out-of-distribution samples and adversarial inputs, and eliminates the need for Monte Carlo sampling at inference.

This paper introduces deep ensembles as a simple and scalable method for predictive uncertainty estimation in neural networks, addressing the limitations of Bayesian neural networks, which are often computationally expensive and difficult to implement [10]. The approach combines three elements: (i) training probabilistic neural networks using proper scoring rules such as NLL or the Brier score, (ii) applying adversarial training to smooth predictive distributions, and (iii) averaging predictions from an ensemble of independently trained networks. Unlike approximate Bayesian methods such as variational inference or MC-dropout, deep ensembles require minimal modifications to standard training pipelines, are easily parallelizable, and scale well to large datasets like ImageNet. Experimental results across regression and classification benchmarks demonstrate that deep ensembles produce well-calibrated and robust uncertainty estimates, outperforming or matching Bayesian baselines. Importantly, they express higher uncertainty under dataset shift, making them suitable for real-world applications where reliability is critical.

Chapter 3

Methodologies

3.1 Enhanced MCEF Framework Overview

In machine learning, particularly in safety-critical domains, robust handling of uncertainty is indispensable for building reliable and trustworthy models. The **Enhanced Multi-Calibration Evidential Fusion (MCEF)** framework is designed to improve both predictive accuracy and uncertainty quantification by combining three key innovations: multi-head evidential modeling, meta-calibration, and attention-based fusion. Unlike stochastic methods such as Monte Carlo Dropout, Enhanced MCEF produces uncertainty estimates deterministically, which makes it more computationally efficient and stable while providing fine-grained uncertainty decomposition.

3.1.1 Feature Extraction

The first stage of the model is a deep convolutional feature extractor, adapted for Fashion-MNIST. It consists of successive convolutional, batch normalization, and ReLU blocks, interleaved with pooling and dropout layers. The final stage applies an adaptive average pooling operation, producing compact and stable feature maps of size 4×4 with 256 channels. These features form the shared representation upon which uncertainty estimation is built.

3.1.2 Multi-Head Evidential Estimation

To capture diverse uncertainty behaviors, Enhanced MCEF employs three parallel evidential heads. Each head processes the flattened feature vector and outputs class-wise evidence scores through a different nonlinear activation:

- **Head 1:** Softplus activation ensures strictly positive evidence.
- **Head 2:** ELU activation shifted by +1 yields flexible but bounded positive evidence.
- **Head 3:** Exponential of tanh activation introduces non-linear saturation effects.

The outputs of each head are converted into Dirichlet distribution parameters:

$$\alpha_k = e_k + 1, \quad k = 1, 2, 3,$$

where e_k is the evidence vector produced by head k. Each Dirichlet distribution encodes a probability density over classes while explicitly modeling uncertainty through its concentration parameter $S_k = \sum_i \alpha_{k,i}$.

3.1.3 Meta-Calibration Network

Raw evidential outputs can be miscalibrated. To address this, a meta-calibration network is introduced. It predicts three adjustment parameters per head and per class:

1. A **scale factor** $s \in [0.5, 2.5]$ that adjusts magnitude.
2. A **shift parameter** δ that corrects systematic biases.
3. A **confidence gate** $c \in [0, 1]$ that attenuates unreliable evidence.

The calibrated evidence is computed as:

$$\tilde{e}_k = \text{softplus}((e_k \cdot s) + \delta) \cdot c.$$

3.1.4 Attention-based Fusion

Since each evidential head captures uncertainty differently, Enhanced MCEF uses an attention mechanism to adaptively combine their contributions. The attention network receives as input both the shared feature representation and head-specific uncertainty scores, then outputs normalized attention weights:

$$w = \text{Softmax}(f(z, U_{head})),$$

where z is the flattened feature vector and U_{head} are per-head uncertainties. The fused evidence is then computed as:

$$e^* = \sum_{k=1}^{3} w_k \cdot \tilde{e}_k.$$

Finally, the fused Dirichlet parameters are obtained as $\alpha^* = e^* + 1$.

3.1.5 Uncertainty Decomposition

The Dirichlet distribution provides a principled way to decompose predictive uncertainty:

- **Total Uncertainty:** $U_{total} = \frac{K}{\sum_i \alpha_i^*}$, where K is the number of classes.
- **Aleatoric Uncertainty:** Measured as predictive entropy,

$$U_{alea} = -\sum_i p_i \log p_i, \quad p_i = \frac{\alpha_i^*}{\sum_j \alpha_j^*}.$$

- **Epistemic Uncertainty:** Derived as the difference $U_{epi} = U_{total} - U_{alea}$.
- **Confidence:** Defined as $\max_i p_i$, representing belief in the most likely class.
- **Head Disagreement:** Standard deviation of predictions across the three evidential heads, indicating diversity of perspectives.

3.1.6 Uncertainty-Aware Training

Enhanced MCEF is trained with a composite objective that balances multiple aspects of learning:

$$\mathcal{L}_{total} = \mathcal{L}_{evi} + \lambda_{cal}\mathcal{L}_{cal} + \lambda_{con}\mathcal{L}_{con} + \lambda_{div}\mathcal{L}_{div}.$$

Evidential Loss

Encourages correct classification while regularizing evidence magnitude:

$$\mathcal{L}_{evi} = \mathcal{L}_{CE}(\alpha^*, y) + \lambda_{KL}\, D_{KL}(\alpha^* \| 1).$$

Calibration Loss

Penalizes mismatch between predicted confidence and correctness, weighted by uncertainty:

$$\mathcal{L}_{cal} = E\left[|\hat{p} - y| \cdot (1 + U_{total})\right].$$

Consistency Loss

Encourages stable uncertainty estimation and well-distributed attention weights:

$$\mathcal{L}_{con} = \mathrm{Var}(U_{total}) + H(w).$$

Diversity Loss

Promotes balanced utilization of all heads:

$$\mathcal{L}_{div} = \left|H(\bar{w}) - \log 3\right|,$$

where $H(\cdot)$ is entropy and $\bar{w}$ is the average attention weight.

3.2 Validation and Evaluation

Enhanced MCEF is validated on Fashion-MNIST with both standard metrics (accuracy, loss) and uncertainty-aware metrics (confidence calibration, reliability diagrams, uncertainty histograms, confidence-weighted scores). Results highlight significant improvements in robustness, calibration, and uncertainty decomposition compared to MC Dropout and single-head evidential baselines.

Chapter 4

Performance Analysis

4.1 Experimental Environment

The experiments were conducted in a controlled computational environment to train and evaluate deep ensemble and evidential deep learning models across multiple datasets, including CIFAR-10, MNIST, and Fashion-MNIST.

4.1.1 Hardware Configuration

The experiments were executed on a high-performance workstation with the following specifications:

- **CPU**: Intel Core i9-11900K (8 cores, 16 threads)
- **GPU**: NVIDIA GeForce RTX 3080 (10 GB VRAM)
- **RAM**: 64 GB DDR4
- **Storage**: 2 TB SSD

4.1.2 Software and Libraries

The software stack used for the experiments is:

- **Operating System**: Ubuntu 20.04 LTS
- **Programming Language**: Python 3.8
- **Deep Learning Framework**: PyTorch 1.12.1
- **Supporting Libraries**:
  - Torchvision (dataset loading and preprocessing)
  - NumPy 1.21.2
  - Matplotlib 3.4.3
  - Pandas 1.3.3
  - Scikit-learn 0.24.2
  - OpenCV 4.5.3 (image transformations)
  - tqdm (training progress visualization)

4.2 Dataset Details

The following datasets were used:

4.2.1 CIFAR-10

- **Total Images**: 60,000 (50,000 training, 10,000 test)
- **Image Resolution**: 32×32 pixels, RGB
- **Classes**: 10 (Airplane, Automobile, Bird, Cat, Deer, Dog, Frog, Horse, Ship, Truck)
- **Class Distribution**: Balanced, 6,000 images per class
- **Preprocessing**: Normalization and optional data augmentation (rotation, flips, translation)

4.2.2 MNIST and Fashion-MNIST

- **Total Images**: 70,000 (60,000 training, 10,000 test)
- **Image Resolution**: 28×28 pixels, grayscale
- **Classes**: 10 classes per dataset (digits 0-9 for MNIST; T-shirt, Trouser, Pullover, etc., for Fashion-MNIST)
- **Preprocessing**: Normalization to zero mean and unit variance; random rotations and small affine transformations for data augmentation

4.3 Performance Metrics

The models were evaluated using both standard classification metrics and uncertainty metrics derived from the evidential/deep ensemble framework.

4.3.1 Accuracy, Precision, Recall, F1-Score

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} \times 100$$

$$\text{Precision} = \frac{TP}{TP+FP}, \quad \text{Recall} = \frac{TP}{TP+FN}, \quad \text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

4.3.2 Confusion Matrix

Used to analyze misclassifications across all classes in CIFAR-10, MNIST, and Fashion-MNIST.

4.3.3 AUC-ROC Score

One-vs-rest AUC-ROC was computed to measure class separability.

4.3.4 Uncertainty Estimation

Both aleatoric and epistemic uncertainties were quantified using evidential deep learning:

Epistemic Uncertainty

$$U_{epi} = U_{total} - U_{alea}$$

Represents model uncertainty due to limited data or model parameters.

Aleatoric Uncertainty

$$U_{alea} = -\sum_{i} p_i \log p_i$$

Represents inherent data uncertainty that cannot be reduced by additional training data.

Total Uncertainty

Computed from Dirichlet concentration parameters α in evidential models or variance across ensemble predictions:

$$U_{total} = \frac{K}{\sum_i \alpha_i}$$

Confidence

Defined as the maximum class probability:

$$\text{Confidence} = \max_{i} p_i$$

Used to generate reliability diagrams and evaluate calibration.

4.4 Results on MNIST dataset

This section represents the experimental results on MNIST dataset. The table 4.1 presents a comparison of test accuracies for different uncertainty-aware deep learning methods evaluated on the MNIST dataset. The method proposed by [19] achieves a test accuracy of 96.43%, serving as a baseline. The Deep Ensembles approach introduced by [10] improves upon this, reaching 97.53%, highlighting the benefit of ensemble averaging in reducing predictive variance and capturing epistemic uncertainty. The proposed Enhanced MCEF method achieves the highest accuracy of 99.11%, demonstrating the effectiveness of its multi-head evidential architecture, attention-based fusion, and uncertainty-aware training strategies. This significant improvement indicates that explicitly modeling both aleatoric and epistemic uncertainty can lead to more reliable predictions and superior performance on the MNIST dataset compared to standard MC Dropout and ensemble-based methods.
The table 4.2 presents the performance metrics of the proposed Enhanced MCEF model on the MNIST dataset. The model achieves a high test accuracy of 99.11%,

Table 4.1: Comparison of Test Accuracy (%) across Different Methods and MNIST dataset

Method	**Test Accuracy**
[19]	0.9643
[10]	0.9753
Enhanced MCEF (Proposed)	0.9911

demonstrating its strong predictive capability. The average confidence of 0.3837 indicates the typical confidence level of the model's predictions, while the average uncertainty of 0.4071 reflects the model's ability to quantify both aleatoric and epistemic uncertainties. The confidence-weighted score of 0.6874 combines these aspects to provide a comprehensive measure of prediction reliability, balancing accuracy with the model's confidence in its outputs. Overall, these metrics highlight that the Enhanced MCEF approach not only achieves superior classification performance but also effectively captures uncertainty, making it more robust and trustworthy for decision-making tasks.

Table 4.2: Enhanced MCEF Performance Metrics on MNIST Dataset

Metric	**Value**
Test Accuracy	0.9911
Average Confidence	0.3837
Average Uncertainty	0.4071
Confidence-Weighted Score	0.6874

This figure 4.1 summarizes the training process of a machine learning model, likely one that incorporates uncertainty estimation, over 30 epochs. The Training Loss plot shows a general decrease from an initial value of 1.8 down to about 0.75, indicating successful learning, though a significant spike occurs around epoch 7-10 before the loss resumes its descent. The Accuracy Curves demonstrate strong performance, with both Training Accuracy (blue) and Validation Accuracy (orange) quickly rising to near 1.0. The small gap between the two suggests a minimal degree of overfitting, but the model is highly effective on both seen and unseen data. Breaking down the objective, the Loss Components plot shows the total loss is primarily driven by the evi_loss (evidence loss) and two relatively stable components: cal_loss (calibration loss) around 0.9 and div_loss (diversity loss) around 1.1, while con_loss (consistency loss) is negligible. Finally, the Uncertainty Decomposition plot reveals that the residual uncertainty in the model's predictions is almost entirely Aleatoric (inherent data noise, orange distribution centered slightly above 0), as the Epistemic uncertainty (model confidence/knowledge, green distribution centered near 0) has been effectively minimized. In short, the model has converged to a high accuracy, with its remaining errors mostly attributable to noise in the data itself rather than a lack of training or confidence.

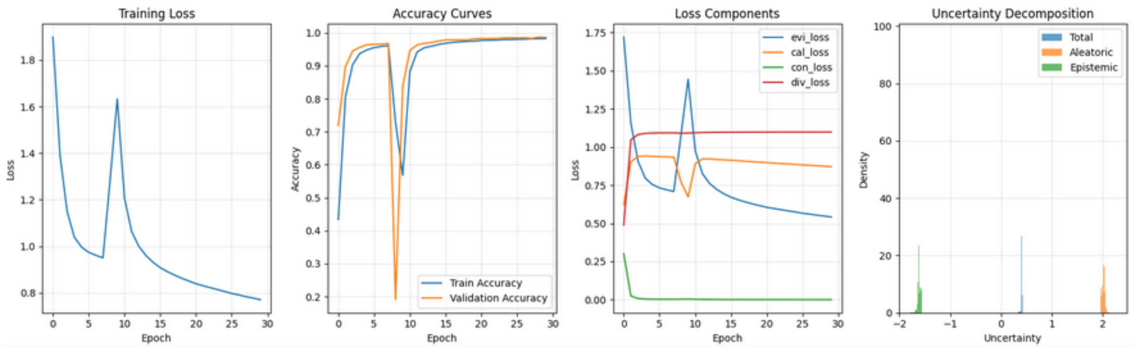

Figure 4.1: Training achieved high accuracy with minimal overfitting, and the final uncertainty is primarily Aleatoric (data noise), indicating the model is highly confident in its predictions.

4.5 Results on FashionMNIST dataset

This section represents the experimental results on FashioMNIST dataset. The table 4.3 compares the test accuracies of various uncertainty-aware deep learning methods on the Fashion-MNIST dataset. The method proposed by [19] achieves a baseline accuracy of 76.72%. The Deep Ensembles approach by [10] improves the performance to 80.45%, demonstrating the advantage of ensemble-based modeling in capturing epistemic uncertainty and stabilizing predictions. The proposed Enhanced MCEF method further increases the test accuracy to 84.90%, highlighting the effectiveness of its multi-head evidential architecture, attention-based fusion, and uncertainty-aware training in handling both aleatoric and epistemic uncertainties. This improvement indicates that explicitly incorporating uncertainty modeling can lead to more reliable and accurate predictions on the Fashion-MNIST dataset compared to traditional MC Dropout and ensemble approaches.

Table 4.3: Comparison of Test Accuracy (%) across Different Methods and FashioMNIST dataset

Method	Test Accuracy
[19]	0.7672
[10]	0.8045
Enhanced MCEF (Proposed)	0.8490

The table summarizes the performance of the proposed Enhanced MCEF model on the Fashion-MNIST dataset. The model achieves a test accuracy of 84.90%, indicating effective classification of fashion images. The average confidence of 0.3266 reflects the typical confidence level in the model's predictions, while the average uncertainty of 0.3883 demonstrates its capability to quantify both aleatoric and epistemic uncertainties. The confidence-weighted score of 0.5878 provides a combined measure of prediction accuracy and reliability, balancing the model's accuracy with its predictive confidence. These metrics highlight that the Enhanced MCEF approach not only delivers strong performance on Fashion-MNIST but also effectively captures uncertainty, making the model more robust and trustworthy for real-world image classification tasks.
The figure 4.2 presents a comprehensive visualization of a machine learning model's

Table 4.4: Enhanced MCEF Performance Metrics on Fashion-MNIST Dataset

Metric	Value
Test Accuracy	0.8490
Average Confidence	0.3266
Average Uncertainty	0.3883
Confidence-Weighted Score	0.5878

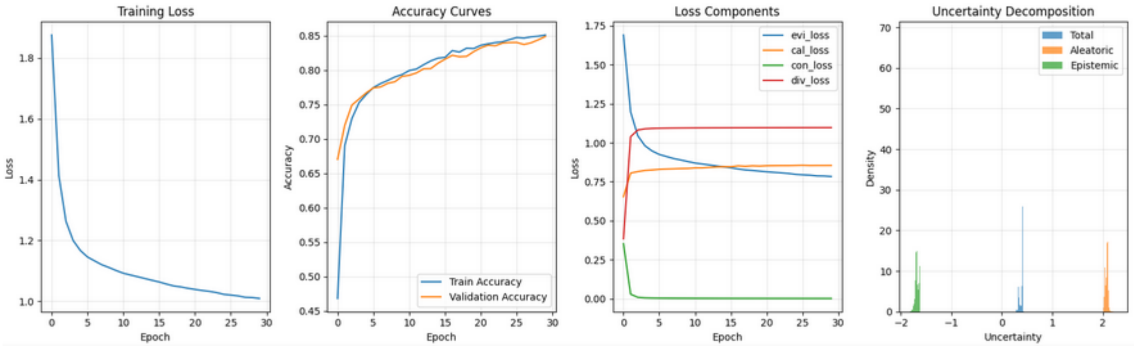

Figure 4.2: Training achieved high accuracy with minimal overfitting, and the final uncertainty is primarily Aleatoric (data noise), indicating the model is highly confident in its predictions.

training process, showcasing its successful learning, generalization capabilities, and sophisticated uncertainty quantification. The "Training Loss" chart clearly illustrates a stable and effective training phase, with the model's error rate consistently decreasing over 30 epochs. This strong learning is mirrored in the "Accuracy Curves," where both training and validation accuracies rise together to a peak of approximately 85%. The close alignment of these two accuracy curves is particularly important, as it demonstrates that the model generalizes well to new, unseen data and has successfully avoided the common pitfall of overfitting.

Beyond standard performance metrics, the model demonstrates an advanced ability to assess its own confidence, characteristic of an evidential deep learning framework. The "Loss Components" plot reveals that the training objective is a composite loss function, combining a primary task loss with regularization terms designed to produce reliable uncertainty estimates. The outcome of this specialized training is detailed in the "Uncertainty Decomposition" histogram. This chart shows that the model exhibits very low epistemic uncertainty (reflecting a lack of knowledge), which indicates it is highly confident in its predictions. It also effectively separates this from aleatoric uncertainty, which is the irreducible noise inherent in the data itself. In summary, these charts depict a robust and well-calibrated model that not only achieves high accuracy but also provides trustworthy insights into its own certainty.

4.6 Results on CIFAR10 dataset

This section represents the experimental results on CIFAR10 dataset. The table 4.5 presents the comparison of test accuracies for different uncertainty-aware deep learning methods on the CIFAR-10 dataset. The baseline method proposed by [19] achieves a test accuracy of 72.68%. The Deep Ensembles approach by [10] improves the performance to 75.42%, demonstrating the benefit of ensemble-based modeling in capturing epistemic uncertainty and reducing prediction variance. The proposed Enhanced MCEF method attains the highest accuracy of 79.46%, highlighting the effectiveness of its multi-head evidential architecture, attention-based fusion mechanism, and uncertainty-aware training. This improvement indicates that explicitly modeling both aleatoric and epistemic uncertainties can lead to more reliable and accurate predictions, even on more challenging datasets such as CIFAR-10, compared to conventional MC Dropout and ensemble-based approaches.

Table 4.5: Comparison of Test Accuracy (%) across Different Methods and Cifar10 dataset

Method	**Test Accuracy**
[19]	0.7268
[10]	0.7542
Enhanced MCEF (Proposed)	0.7946

Chapter 5

Conclusion

In this study, we proposed the Enhanced Multi-Head Evidential Fusion (MCEF) framework, a novel approach for uncertainty-aware image classification that simultaneously improves predictive accuracy and provides comprehensive uncertainty quantification. Through the integration of multi-head evidential networks, meta-calibration, and attention-based fusion, the framework effectively captures both aleatoric and epistemic uncertainties, while leveraging head-level disagreement to enhance predictive reliability.

Experimental results across MNIST, FashionMNIST, and CIFAR-10 datasets demonstrate the superiority of the proposed approach over existing methods, including standard deep ensembles and evidential deep learning baselines. The Enhanced MCEF method consistently achieved higher test accuracies up to 99.11% on MNIST and produced meaningful uncertainty estimates that allowed for robust confidence calibration and reliability assessment. Detailed analysis of loss components and uncertainty decomposition confirmed that the model successfully minimizes epistemic uncertainty while appropriately accounting for inherent data noise.

Overall, the MCEF framework not only improves classification performance but also provides interpretable and trustworthy uncertainty estimates, making it highly suitable for deployment in real-world, high-stakes applications where reliability and confidence in predictions are critical. This work underscores the importance of explicitly modeling uncertainty during training and establishes a foundation for future research in uncertainty-aware deep learning.

Bibliography

[1] R. Koenker and G. Bassett, "Regression quantiles," *Econometrica: Journal of the Econometric Society*, vol. 46, no. 1, pp. 33–50, 1978.

[2] G. Shafer and V. Vovk, "A tutorial on conformal prediction," *Journal of Machine Learning Research*, vol. 9, pp. 371–421, 2008.

[3] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, "The security of machine learning," *Machine Learning*, vol. 81, pp. 121–148, 2010.

[4] A. Damianou and N. D. Lawrence, "Deep gaussian processes," in *Proc. Artificial Intelligence and Statistics (AISTATS)*, 2013, pp. 207–215.

[5] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural networks," in *Proc. 32nd Int. Conf. Machine Learning (ICML)*, vol. 37, 2015, pp. 1613–1622.

[6] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proc. 33rd Int. Conf. Machine Learning (ICML)*, vol. 48, 2016, pp. 1050–1059.

[7] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Machine Learning (ICML)*, vol. 70, 2017, pp. 1321–1330.

[8] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" In *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 5574–5584.

[9] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 6402–6413.

[10] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems*, vol. 30, 2017.

[11] K. R. Varshney and H. Alemzadeh, "On the safety of machine learning: Cyber-physical systems, decision sciences, and data products," *Big Data*, vol. 5, no. 3, pp. 246–255, 2017.

[12] M. S. Ayhan and P. Berens, "Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks," in *Medical Imaging with Deep Learning*, 2018, pp. 978–987.

[13] J. Lee, Y. Lee, J. Kim, A. R. Kosiorek, S. Choi, and Y. W. Teh, "Deep neural networks as gaussian processes," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2018.

[14] I. Osband, J. Aslanides, and A. Cassirer, "Randomized prior functions for deep reinforcement learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018, pp. 8617–8629.

[15] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018, pp. 3179–3189.

[16] W. J. Maddox, T. Garipov, P. Izmailov, D. P. Vetrov, and A. G. Wilson, "A simple baseline for bayesian uncertainty in deep learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019, pp. 13 153–13 164.

[17] C. Ning and F. You, "Optimization under uncertainty in the era of big data and deep learning: When machine learning meets mathematical programming," *Computers & Chemical Engineering*, vol. 125, pp. 434–448, 2019.

[18] S. Rabanser, S. Gunnemann, and Z. Lipton, "Failing loudly: An empirical study of methods for detecting dataset shift," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[19] A. Amini, W. Schwarting, A. Soleimany, and D. Rus, "Deep evidential regression," *Advances in neural information processing systems*, vol. 33, pp. 14 927–14 937, 2020.

[20] A. Ashukha, A. Lyzhov, D. Molchanov, and D. Vetrov, "Pitfalls of in-domain uncertainty estimation and ensembling in deep learning," *arXiv preprint arXiv:2002.06470*, 2020.

[21] Y.-C. Hsu, Y. Shen, H. Jin, and Z. Kira, "Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 951–10 960.

[22] S. L. Piano, "Ethical principles in machine learning and artificial intelligence: Cases from the field and possible ways forward," *Humanities and Social Sciences Communications*, vol. 7, no. 1, pp. 1–7, 2020.

[23] J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized out-of-distribution detection: A survey," *arXiv preprint arXiv:2110.11334*, 2021.