university of groningen / faculty of science and engineering / artificial intelligence

Human-Centered Epistemic Uncertainties for Handwritten Digits

Bachelor's Project Thesis

Daniel Gentile, s4273389, d.gentile@student.rug.nl,
Supervisors: Ivo de Jong, MSc (supervisor) & dr. Matias Valdenegro-Toro (co-supervisor)

Abstract

Uncertainty Quantification (UQ) helps build trustworthy and interpretable neural networks. Most prior work uses synthetic or out-of-distribution data, which fails to reflect real-world challenges. In an attempt to bridge this gap, this thesis proposes a novel human-generated dataset of handwritten digits to analyze the effects of different UQ methods on user perception of model confidence and trust.

A web-based platform was built for users to view model predictions, assess confidence scores, and give Likert-scale feedback. We compared three different models: a simple convolutional neural network (CNN) with softmax for confidence estimation, a Monte Carlo Dropout model, and a Deep Ensemble model. Each model was compared in terms of accuracy, confidence calibration, and their effects on user trust.

The results show that even though the three models were equally accurate, the baseline model had a much higher level of confidence—usually bordering on overconfidence—when compared with the UQ models. Bayesian analysis also supported the fact that the baseline model was poorly calibrated, with strong evidence of being overconfident. User feedback on Likert-scale prompts did not show statistically significant differences, though the patterns indicated that the users perceived the confidence of the baseline model as being more extreme.

This work provides a dataset, process, and findings that support future studies of human-centered AI in real-world tasks.

1 Introduction

Machine learning (ML) models, especially neural networks (NNs), have made various applications, from autonomous vehicles to disease diagnosis, a reality. However, despite their capabilities, these models are often difficult for non-experts to understand and trust. One of the primary challenges is that NNs do not always get everything right, and although performance keeps improving, perfect accuracy remains an unrealistic goal. Researchers are looking to Uncertainty Quantification (UQ) methods to bridge the gap. Such methods enable models to convey to users their belief in predictions and offer an additional layer of information to form trust in AI systems.

1.1 Types of Uncertainty

To better tackle the issue of ML interpretability and trust, UQ techniques distinguish between two categories of uncertainty: epistemic and aleatoric. Both categories represent two different sources of uncertainty within ML systems and must be handled differently.

"Aleatoric uncertainty captures noise inherent in the observations. This could be, for example, sensor noise or motion noise, resulting in uncertainty which cannot be reduced even if more data were to be collected." (Kendall & Gal, 2017)

Aleatoric uncertainty—also called data uncertainty—is not eliminable and arises from randomness in the input data. It can be caused by factors such as faulty sensors, blurry images, or ambiguous class labels. No additional data can eliminate such

uncertainty as it reflects randomness or noise in the observation process.

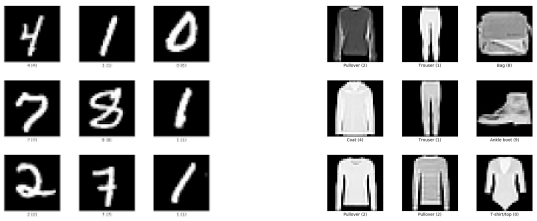
"Epistemic uncertainty accounts for uncertainty in the model parameters – uncertainty which captures our ignorance about which model generated our collected data. This uncertainty can be explained away given enough data, and is often referred to as model uncertainty." (Kendall & Gal, 2017)

Epistemic uncertainty arises from not knowing the data-generating process. It represents uncertainty in the model's predictions and parameters, especially in regions of the input space where the model was poorly trained. It is a form of reducible uncertainty and can be mitigated by getting more data, improving model architecture, or improving training methods. For example, suppose a neural network sees input data that is very different from its training data. In that case, the epistemic uncertainty will typically be higher, indicating less confidence in the model's predictions.

It is important to recognize and differentiate between these two kinds of uncertainty in order to construct ML systems that not only work well but are also interpretable. Accurate communication of model uncertainty in high-stakes fields like autonomous vehicles and medicine, where errors can be disastrous, enables users to make informed decisions and have faith in AI systems.

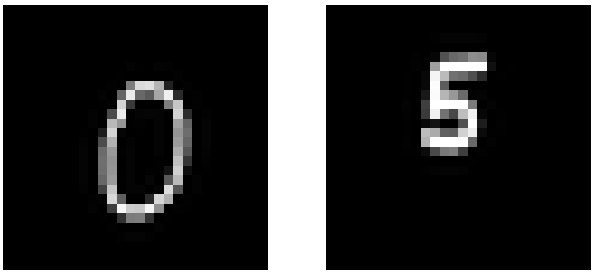
1.2 Limitations of Existing Approaches

Despite the progress made in UQ, significant limitations remain in how these methods are evaluated. A common practice in the literature is introducing epistemic uncertainty using artificial scenarios, such as out-of-distribution (OOD) datasets that differ substantially from the original. For example, many studies evaluate models trained on MNIST—a dataset of 70,000 grayscale images of handwritten digits (0–9)— (LeCun et al., 1998) by testing them on Fashion-MNIST, which contains 70,000 grayscale images of clothing items like shirts, shoes, and dresses (Xiao et al., 2017). While these techniques offer valuable insights, they fail to capture the complexity of uncertainty encountered in real-world applications. As a result, the uncertainties generated in these artificial setups often do not align with naturally occurring ones, limiting the generalizability and robustness of the conclusions. This highlights the need for evaluation methods and datasets that more closely reflect the uncertainty in real-world scenarios (Guth et al., 2024; Gawlikowski et al., 2023).



(a) MNIST **(b) Fashion MNIST**

Figure 1.1: Examples of synthetic datasets used for UQ.



(a) Drawing of digit 0 by a participant **(b) Drawing of digit 5 by a participant**

Figure 1.2: Examples from dataset created for this thesis.

1.3 User Feedback and a Realistic Dataset

Apart from measuring performance through quantitative means, this thesis also analyzes how users perceive and react to uncertainty in predictions from models. To quantify this, we receive participant feedback through Likert-scale questions. Allowing us to analyze how different modes of visualization of uncertainty affect user trust, confidence, and interpretability. It is important to know how people respond to uncertainty when developing ML systems that are not only accurate but also transparent and trustworthy from the user's perspective.

This thesis introduces a novel dataset of handwritten digits directly obtained from human users. In comparison to artificial OOD benchmarks, this dataset naturally contains epistemic uncertainty through the variability of writing style, clarity, and quality—attributes that better reflect the variability and unpredictability of real-world user input. These user samples present a realistic challenge to machine learning models, particularly in parts of the input space not properly represented during training.

1.4 Research Question and Contributions

This thesis investigates the following research question:

> Does visualizing epistemic uncertainty in machine learning models affect user perception of predictions in a handwritten digit classification task?

To answer this, we make the following contributions:

- Develop a web application to collect a realistic, human-generated dataset by allowing participants to draw handwritten digits.
- Implement and compare three uncertainty quantification methods: a base CNN with softmax confidence, Monte Carlo Dropout (MC Dropout), and Deep Ensembles.
- Evaluate user perception of uncertainty using subjective feedback gathered through Likert-scale questions.
- Analyze the relationship between quantitative metrics (e.g., accuracy, Calibration Error) and user-reported trust levels.

2 Related Work

This section outlines the key developments in Uncertainty Quantification (UQ) and related concepts, providing context for this thesis's contributions.

2.1 Uncertainty Quantification in Neural Networks

Firstly, the distinction between epistemic uncertainty (caused by limited knowledge, reducible with more data) and aleatoric uncertainty (caused by inherent data noise and irreducible) was clearly defined thanks to the paper by Kendall & Gal (2017). Additionally, the paper by Valdenegro-Toro & Mori (2022) provided further insight into disentangling these two types of uncertainty. Bayesian neural networks (BNNs) were proposed to better measure and interpret both types of uncertainty, but they were initially too computationally expensive for deep learning tasks. This changed when methods like Monte Carlo Dropout made scalable Bayesian approximations feasible for larger neural networks. Gal & Ghahramani (2016) introduced MC Dropout, which uses stochastic dropout at inference to estimate epistemic uncertainty and requires multiple forward passes—hence the high computational cost. Gal & Ghahramani (2016) tested MC Dropout's uncertainty estimates using the MNIST dataset (LeCun et al., 1998). Later, Lakshminarayanan et al. (2017) introduced Deep Ensembles, a method that involves training multiple independent models to capture diverse representations. This strategy often outperforms MC Dropout, particularly for out-of-distribution (OOD) detection, though at the cost of even greater computational expense, which can limit their use in real-time applications.

2.2 Real-World Applications

UQ is increasingly used in real-world AI applications, particularly in high-stakes domains. Filos et al. (2019) demonstrates that UQ can improve trust in AI-driven medical diagnostics. Some challenges remain not fully solved. Gustafsson et al. (2020) analyzed MC Dropout's performance under distribution shifts and revealed its weaknesses in detecting out-of-distribution data, particularly when models trained on synthetic datasets like Virtual KITTI and Synscapes were evaluated on real-world datasets such as KITTI and Cityscapes. Guth et al. (2024) highlighted gaps in OOD benchmarks, emphasizing the need for datasets reflecting naturally occurring diversity. Many studies rely on synthetic OOD scenarios, for example, training on MNIST

and testing on various variations of MNIST like Fashion-MNIST (Xiao et al., 2017), Dirty-MNIST (Mukhoti et al., 2021a), or Ambiguous-MNIST (Mukhoti et al., 2021b). While this can be very informative, it does not fully replicate real-world uncertainty that emerges naturally from user input variations. This shortcoming of synthetic OOD testing motivates the evaluation of UQ methods on datasets gathered in realistic conditions, where uncertainty emerges naturally from user variation. As Gawlikowski et al. (2023) argued, AI with UQ should be tested on realistic human interactions and datasets.

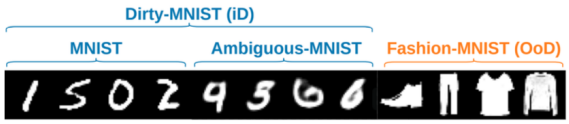


Figure 2.1: Dirty and Ambiguous MNIST. (Mukhoti et al., 2021b)

2.3 The Role of Uncertainty Visualization in Trust and Interpretability

Researchers have developed effective techniques for UQ, but communicating the resulting uncertainty is equally important. Accurate quantification without clear communication can lead to outcomes as ineffective as poor quantification itself. To this point, Kendall & Gal (2017) introduced uncertainty visualization techniques like confidence bars and heatmaps. Other works highlighted the need for suitable indicators of uncertainty to foster users' trust. In one case, Filos et al. (2019) employed predictive entropy with MC Dropout as an uncertainty indicator alerting human reviewers to uncertain examples in screening for diabetic retinopathy. Their system referred images with increased uncertainty to human reviewers, and they found this type of threshold-based indicator increased diagnostic accuracy and matched instances when the model would be incorrect — providing confidence in AI-informed decisions. Their work highlights the importance of evaluating not only how well models quantify uncertainty but also how effectively this information is conveyed to users.

2.4 Real-World Uncertainty in Handwritten Digit Recognition

Most UQ research evaluates methods on controlled benchmarks which may not always generalize to real-world settings. Guth et al. (2024) called for more diverse datasets that naturally introduce epistemic uncertainty.

This thesis addresses this gap by:

- Introducing a human-generated dataset for handwritten digit classification.
- Evaluating UQ methods in a realistic user-driven setting rather than artificial OOD benchmarks.
- Incorporating user feedback to examine how uncertainty affects trust and interpretability.

By focusing on human-generated uncertainty, this thesis aims to enhance the transparency, interpretability, and trustworthiness of AI systems in real-world scenarios.

3 Methods

As discussed in Section 2, a clear lack of real-world data used to compare UQ in various models exists (Guth et al., 2024). This work addresses this gap by developing a dataset of handwritten digits that naturally exhibit epistemic uncertainty rather than relying on artificially induced variations such as added noise. Participants used a digital canvas to draw the digits by hand on a tablet, attempting to emulate their handwritten digits on paper. This can be seen in Figure 3.1.

The drawings were then preprocessed, which involved placing the drawn digit on a black background and making it white, converting to grayscale, resizing the image to 28×28 pixels, applying thresholding so pixels are either entirely white or black, and lastly, normalizing pixel values to the range [0, 1]. The *preprocess_image()* method outputs the image as a NumPy array for the models to take as input and a PIL image, to be shown later as a reference. A simple and minimal preprocessing pipeline ensured that the dataset created from this experiment maintained the handwritten digits as close to how they were drawn, without

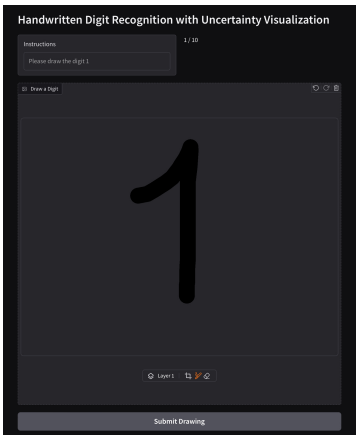


Figure 3.1: Web App Canvas.

any rotations, centering, or additional resizing. The pseudocode for this method can be seen below in Listing 1.

Listing 1: Pseudocode for Image Preprocessing

```
preprocess_image(img):
    Create a white background
    Composite image onto background
    Convert image to grayscale
    Invert image so the digit appears
        white on black
    Resize image to (28x28)
    Convert image to a NumPy array
    Apply thresholding:
        - Pixels > 20 → 255 (white)
        - Pixels ≤ 20 → 0 (black)
    Normalize pixels to range [0,1]
    Return processed image array and
        resized image
```

3.1 Participants and Data Collection

For this research, participants were selected with no specific criteria. People around various buildings of the Zernike Campus (Groningen, NL) were asked to participate if they wanted to. Everybody's data was used with no omissions. Each participant was allowed two practice runs of the experiment to familiarize themselves with the interface. We allowed them to ask questions while we observed and provided feedback. Next, they were told to draw each digit once for a total of 10 digits per participant. There was no support during this part of the experiment, as we wanted their drawings to be as genuine as possible without any outside influence.

Twenty participants participated in this experiment, resulting in a final dataset of 200 drawings (each digit 20 times). Preprocessing was minimal, ensuring no modifications to the drawing were made. The models only used this data to provide the output, which was used during the experiment without using it for training or anything else.

Figure 3.2 shows an example of a digit drawn by a participant.

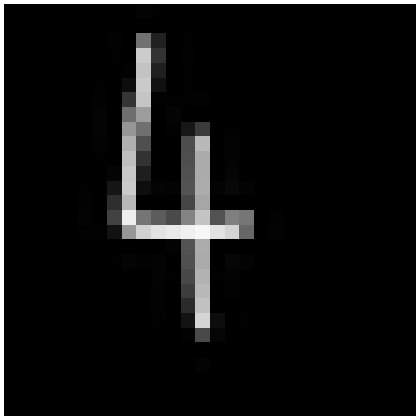


Figure 3.2: Drawing of Digit 4 by Participant 1.

3.2 Model Architectures and Training

For this experiment, three models were trained: a base CNN, a MC Dropout CNN, and a Deep Ensemble of five CNNs (each identical to the base CNN). The architectures shown in Figure 3.3 and Figure 3.4 show how each was structured.

The Base Model contains three convolution/max-pooling blocks followed by one fully connected layer of 64 neurons and one dense layer of size 10, corresponding to the softmax output of 10 classes. Early Stopping was utilized on the validation set to control overfitting. The overall model has approximately 280k parameters, of which approximately 93k are trainable.

Next, the baseline architecture was extended to have dropout layers that remain active during inference for the MC Dropout model. In addition to the three conv/max-pooling blocks, dropout layers were added after each pool and fully connected operation. By making many stochastic forward passes

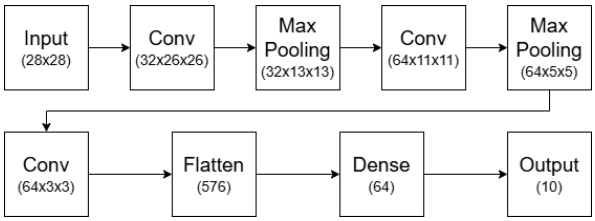


Figure 3.3: Base Model & Deep Ensemble Member

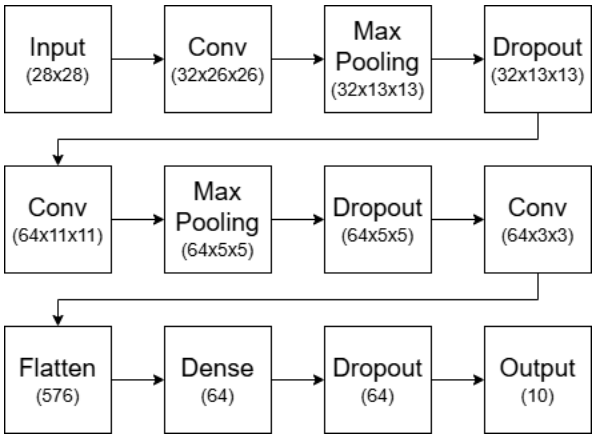


Figure 3.4: MC Dropout

(e.g., 100), a mean prediction and an uncertainty approximation are created (Gal & Ghahramani, 2016). The dropout additions do not change the overall number of parameters of the model (still approximately 280k), and Early Stopping was used in the training of this model, too.

Lastly, we utilized a Deep Ensemble of CNNs to further improve uncertainty estimation (Lakshminarayanan et al., 2017). Five separately initialized instances of the base CNN architecture were trained. Their estimated probabilities are averaged out during test time, and confidence is gathered from this, too. Though the process is more computationally expensive, it can yield better-calibrated estimates than one network alone. A single member of the Deep Ensemble's architecture is identical to that of the base CNN, as shown by Figure 3.3.

All models were implemented using TensorFlow/Keras. Critical hyperparameters (e.g., learning rate, batch size, dropout levels, size of the ensemble, optimizer parameters) are shown in Table 3.1 for reference.

Table 3.2 summarizes the model's final validation and test performance (on standard MNIST).

3.3 Confidence Calculation

Each model in this study computes a confidence score for every class based on the probability distribution output by its softmax layer. The confidence for class i is defined as the probability assigned to that class, expressed as a percentage. The final predicted class is determined as the one with the highest confidence.

3.3.1 Base CNN Confidence Calculation

The base CNN produces a single forward pass for each input image, outputting a softmax probability distribution:

$$p_i = f_\theta(x) \tag{3.1}$$

where p_i represents the probability assigned to class i for a given input x. The confidence for each class is then computed as follows:

$$\text{Confidence}_i = p_i \times 100 \tag{3.2}$$

The final predicted class $\hat{y}$ is selected as:

$$\hat{y} = \arg\max_i p_i \tag{3.3}$$

3.3.2 MC Dropout Confidence Calculation

MC Dropout models apply dropout during inference, allowing for multiple stochastic forward passes. Given T forward passes, each pass produces a probability distribution $p_i^{(t)}$ for class i. The confidence for each class is computed using the *mean softmax probability* across all runs:

$$\mathbb{E}[p_i] = \frac{1}{T} \sum_{t=1}^{T} p_i^{(t)} \tag{3.4}$$

The confidence score for each class is then:

$$\text{Confidence}_i = \mathbb{E}[p_i] \times 100 \tag{3.5}$$

The final predicted class is selected as:

$$\hat{y} = \arg\max_i \mathbb{E}[p_i] \tag{3.6}$$

Table 3.1: Key Hyperparameters for Each Model.

Model	Hyperparameter	Value
All	Learning Rate	0.001 (Adam default)
All	Batch Size	64
All	Early Stopping Patience	3
MC Dropout	Dropout Rate	0.50
MC Dropout	Number of Forward Passes	100
Deep Ensemble	Random Initializations	5 distinct seeds

Table 3.2: Summary of Key Training and Evaluation Results on MNIST.

Model	Final Epoch	Val Accuracy (Best)	Test Loss	Test Accuracy
Base CNN	12	99.12%	0.0267	99.21%
MC Dropout	13	99.13%	0.0280	99.09%
Ensemble (1–5)	8[a]	98.97%[b]	0.0306[c]	99.04%[d]

[a] Mean of the final epoch across the five Deep Ensemble members (range: 5–11).
[b] Mean of the best validation accuracy (range: 98.82–99.10%).
[c] Mean test loss (range: 0.0280–0.0325).
[d] Mean test accuracy (range: 98.98–99.09%).

3.3.3 Deep Ensemble Confidence Calculation

Deep Ensembles combine predictions from N independently trained models. Each model produces a probability distribution $p_i^{(n)}$ for class i, and the final confidence for each class is based on the *mean softmax probability* across all Ensemble members:

$$\mathbb{E}[p_i] = \frac{1}{N} \sum_{n=1}^{N} p_i^{(n)} \tag{3.7}$$

The confidence score for each class is then:

$$\text{Confidence}_i = \mathbb{E}[p_i] \times 100 \tag{3.8}$$

The final predicted class is selected as:

$$\hat{y} = \arg\max_i \mathbb{E}[p_i] \tag{3.9}$$

3.4 User Feedback Analysis

A web app was developed to create a dataset of handwritten digits and collect user perception data. The development of the app was done in Python with the use of Gradio (Abid et al., 2019) for the front-end user interface. The app was designed to allow me, the researcher, some customization before passing the tablet to the participants. Tablets allowed participants to draw digits that most closely resembled their handwritten digits on paper. On the tablet, the first page for the participant was simple, minimizing distractions. Participants were given instructions on which digit to draw, a way to measure their progress through the experiment, and a large canvas on which to draw their digit. A screenshot of this page can be seen in Figure B.2 in Appendix B with two examples of what the canvas looked like with digits drawn in Figure B.3 and Figure B.4.

Once a digit was drawn, participants submitted their drawing, and this is when, in the background, all the preprocessing was done, and the randomly selected model processed the drawn digit. Participants were then shown the results page, which included their drawn digit for reference, the model's top prediction with confidence as a percentage, and a bar plot showing the confidence percentages for all ten digits. An example of this can be seen in Figure B.5 in Appendix B. Below this, participants were asked five questions with possible answers:

Strongly disagree, *Disagree*, *Neutral*, *Agree*, *Strongly agree*, and *Can't answer* for edge cases. A screenshot of this page can be seen in Figure B.6 in Appendix B. The five questions are explained below:

1. **Is the top prediction appropriate?**
   This question asked if the model's top prediction (labeled on the bar plot in a different color) was appropriate. It aimed to measure whether the user agreed with the top prediction, especially when incorrect since the model predicting a 7 instead of a 1 makes more sense than predicting an 8 instead of a 1.
2. **Is the top prediction's confidence appropriate?**
   This question asked if the confidence value attributed to the top prediction made sense, aiming to measure user confidence in the model's prediction.
3. **Are the alternative predictions appropriate?**
   This question asked if all other predictions (all but the top) felt appropriate. For example, the model correctly predicting a 1, but having 7 be the next most confident guess should be seen as more appropriate than the next guess being a 0.
4. **Are the alternative predictions' confidences appropriate?**
   Same idea as question 2 but refers to the alternative predictions instead of the top one.
5. **In relation to how clear the drawing is, is the prediction too confident?**
   This question aimed to measure how appropriate the model's confidence felt to the participants relative to their satisfaction with their drawing.

This app contributed to this research by providing a dataset of handwritten digits and a measure of user perception on a number of outputs from the various models. The results section will analyze this user perception to gather any possible conclusions regarding the differences between the three models.

3.5 Ethical Considerations

Regarding data and ethics, participants were informed that their digit drawings and questionnaire responses would be collected anonymously for use within this study. No identifiable data were collected, and students could opt out whenever they wanted. Two documents were provided to each participant:

1. **Consent Form:**
   A consent form for each participant to sign was collected and stored. See Figure A.1 in Appendix A.
2. **Participant Handout:**
   A handout containing an explanation of the experiment and how the data will be used. It contained contact information if participants wanted to withdraw their data or ask further questions. See Figure A.2 and Figure A.3 in Appendix A.

4 Results

Before conducting any statistical testing, the first step was determining the distribution of all the data points used in this section for analysis. Table C.2 in Appendix C shows the results of these checks.

Furthermore, **QQ plots** and **histograms** were generated to provide a more intuitive way to interpret the various distributions. The following conclusions were gathered from the **Shapiro-Wilk tests** and the various plots: confidence scores were all non-normally distributed with a strong negative skew, especially in the Base Model. A substantial deviation from normality was also observed for accuracy. The answers to the five questions in the Likert scale also showed a non-normal distribution when *"Cannot answer"* answers were ignored. Lastly, the calibration error (difference between accuracy and confidence) showed a deviation from normality for the Base Model, weak evidence for normality for MC Dropout, and a normal distribution for the Deep Ensemble. These were all treated as non-normal for consistency and ease of comparison during statistical analysis.

4.1 Accuracy Analysis

The mean accuracy (per participant) across the three models was compared using a **Friedman test**.

The test results showed no statistically significant difference between the models, $\chi^2(2) = 0.125$, $p = .939$. Given this high p-value, no post-hoc tests were conducted.

These results show that no significant difference was found between the models' accuracies, which can be seen visually in Figure 4.1. The box plot shows a wide range of values for each model, with the only notable difference being a larger range for the Base Model compared to the other two.

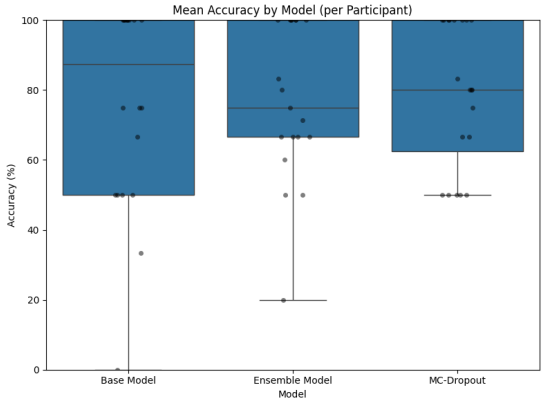


Figure 4.1: Distribution of mean accuracy for each model.

4.2 Confidence Analysis

The mean confidence (per participant) across the three models was also compared using a **Friedman test**. The results revealed a statistically significant difference, $\chi^2(2) = 8.000$, $p = .018$.

Following this, **post-hoc Wilcoxon signed-rank tests** with **Bonferroni correction** were used to examine pairwise comparisons:

- **Base Model vs Deep Ensemble**: $p_{\text{corrected}} = .042$
- **Base Model vs MC-Dropout**: $p_{\text{corrected}} = .007$
- **Deep Ensemble vs MC-Dropout**: $p_{\text{corrected}} = 1.000$

These results indicate that confidence in the Base Model was significantly higher than in both the Deep Ensemble and MC Dropout. At the same time, no significant difference was found between the Ensemble Model and MC dropout after correction.

We can confirm these findings visually by looking at Figure 4.2. The Base Model had the highest confidence, both in terms of higher median and tighter range. On the other hand, confidence for the other two models was lower and with higher variability.

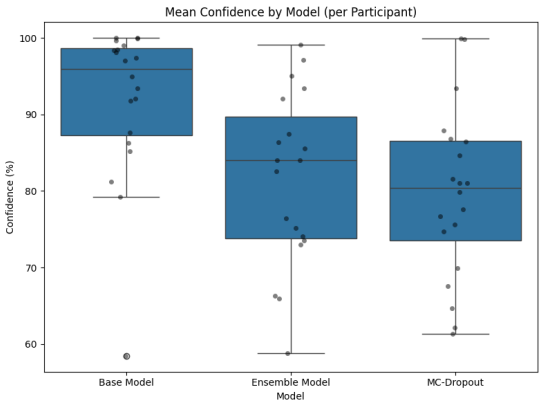


Figure 4.2: Distribution of mean confidence for each model.

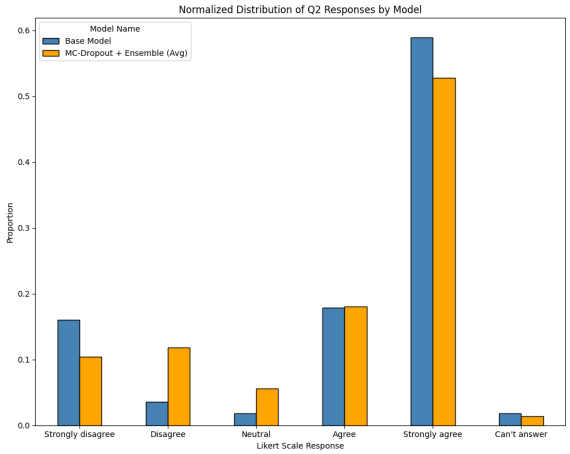
4.3 User Feedback Analysis

The answers to the five questions were stored on a Likert scale, and, as discussed before, the data deviates from normality. Therefore, a **Friedman test** was conducted here, too. The results are below:

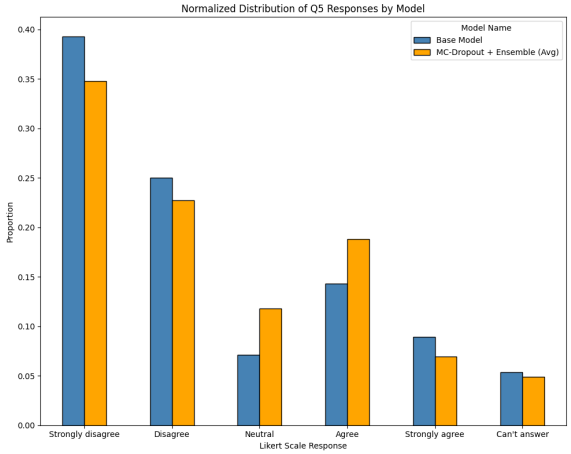
- **Q1**: $\chi^2(2) = 1.156$, $p = .561$
- **Q2**: $\chi^2(2) = 1.583$, $p = .453$
- **Q3**: $\chi^2(2) = 2.459$, $p = .292$
- **Q4**: $\chi^2(2) = 0.704$, $p = .703$
- **Q5**: $\chi^2(2) = 5.365$, $p = .068$

Although Q5 approached significance, it did not meet the conventional threshold of $p < .05$, and therefore no follow-up tests were performed.

The following bar plots (Figure 4.3) were constructed to analyze the answers to the five questions visually. They show the frequency of each answer for the base model compared to an average of the MC Dropout and Deep Ensemble models.



(a) Comparison of proportions of answers between Base Model and average between MC Dropout and Deep Ensemble for Q2.
Q2: Is the top prediction's confidence appropriate?



(b) Comparison of answers proportions between Base Model and average between MC Dropout and Deep Ensemble for Q5.
Q5: In relation to how clear the drawing is, is the prediction too confident?

Figure 4.3: Normalized Likert response distributions for Q2 and Q5.

These figures also show *"Cannot answer"*. However, this answer was excluded from the statistical tests not to disrupt the statistical significance of the test.

Notably, the bar plot for Q2 and Q5 shows the base model over-performing over the other two models in the *Strongly disagree* and *Strongly agree* answer options while under-performing on the more neutral answers. These bar plots show both the answers from participants when the model predicted correctly and when it predicted incorrectly. This could be a sign of overconfidence from the base model, which is interpreted as appropriate when correct, as shown by the high number of *Strongly agree* answers, and the overconfidence interpreted as incorrect when the model was wrong, as shown by the high number of *Strongly disagree* answers. On the other hand, the MC Dropout model and Deep Ensemble model still have a high number of *Strongly agree* and *Strongly disagree* answers but also more *Disagree*, *Neutral*, and *Agree*. This could show that these models predict with more appropriate levels of confidence in the eyes of the participants.

4.4 Calibration Error Frequentist Analysis

The difference between participants' mean confidence and actual accuracy was computed to measure the calibration error for the three models. This provides insight into how well the confidences align with the accuracy of the models.

For each model, a **Wilcoxon signed-rank test** was conducted to test whether the difference between confidence and accuracy was significantly different from zero.

Table 4.1 shows the results of the tests.

These results show that the Base Model was the most overconfident, though this difference was not statistically significant. The difference for the Deep Ensemble model was minor, and MC Dropout was the most calibrated.

4.5 Calibration Error Bayesian Analysis

Since no statistically significant results were found regarding calibration error using Frequentist statistics, a Bayesian approach was also used to gather more insight into the significance of these differences.

Table 4.1: Summary of Wilcoxon signed-rank test results for calibration error per model.

Model	n	W	p	Median Accuracy	Median Confidence	Mean Difference
Base Model	20	58.0	.083	87.5	95.99	−15.66
Deep Ensemble Model	19	67.0	.275	75.0	83.99	−4.92
MC Dropout	20	105.0	1.000	80.0	80.42	−0.55

Models were randomly picked for each trial hence the number of samples for Deep Ensemble Model being 19 instead of 20 (One participant never saw this model's outputs)

Table 4.2: Bayesian one-sample t-test results for calibration error per model.

Model	Posterior Mean	94% HDI[a]	BF_{10}
Base Model	−14.70	[−23.56, −6.10]	26.18
Deep Ensemble	−4.82	[−13.66, 3.52]	0.39
MC Dropout	−0.43	[−7.33, 6.59]	0.18

[a] Highest Density Interval (HDI).

A Bayesian one-sample t-test was used to estimate the posterior of μ, the average calibration error, and compute the Bayes Factor (BF_{10}) against the null hypothesis that $\mu = 0$.

The results displayed in Table 4.2 show **moderate evidence in favor of a negative calibration error** (overconfidence) in the Base Model, as indicated by a Bayes Factor well above 10 and an HDI that excludes zero. The evidence was inconclusive for the Deep Ensemble and MC-Dropout, with HDIs that included zero and Bayes Factors indicating support for the null.

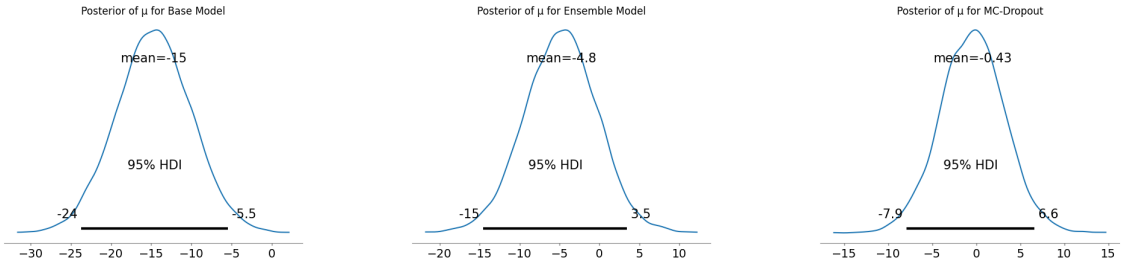
The plots in Figure 4.4 show the calibration error distributions. In the case of the Base Model, the posterior is shifted from zero, while for the other models, this is not the case.

5 Discussion

This thesis examined how different uncertainty quantification (UQ) techniques affect user perception of model predictions in a handwritten digit classification task, focusing on trust and the perceived appropriateness of model confidence. The dataset of handwritten digits we created was ideal for this, as it allowed testing with real-world data instead of the usual synthetic alternatives. The results show interaction with realistic uncertainties for the participants, as will be explored below.

5.1 Accuracy

Predicting handwritten digits is quite simple for modern model structures, like the one employed in the base model and, in turn, the other two. When tested on the MNIST dataset, all three models had test accuracies over 99% (see Table 3.2). This explains why no statistically significant difference was found when comparing the accuracies of the three models during this experiment. The Base, MC Dropout, and Deep Ensemble Models had accuracies of 76.25%, 79.08%, and 76.65%, respectively, when observing all the samples gathered from the experiment. This shows that the models struggled more with the human-drawn digits than the MNIST dataset digits but still held relatively high accuracies. These findings align with prior work showing that UQ methods such as Dropout



(a) Base Model **(b) Deep Ensemble** **(c) MC-Dropout**

Figure 4.4: Posterior distributions for the calibration error (μ) per model. The 95% HDI is marked along the x-axis.

and Ensembles degrade in calibration and accuracy under dataset shift (Ovadia et al., 2019).

5.2 Confidence

A statistically significant difference was found between the three models when comparing confidence, especially between the Base Model and the other two. The Base Model had a higher average confidence than both other models, while there was little to no difference between MC Dropout and the Deep Ensemble. This is in line with earlier predictions and the theory behind this research. The Base Model does not quantify epistemic uncertainty in any particular way and, therefore, predicts with consistent overconfidence. This overconfidence is consistent with findings by Guo et al. (2017), who showed that modern deep networks, including those using Softmax outputs, tend to be overconfident without post-hoc calibration techniques.

5.3 User Feedback

This experiment collected user feedback through five questions with answers on the Likert scale. Analysis of the responses showed no statistically significant difference in answers between the three models. The reasons are discussed later in Section 5.7. That said, when analyzing the bar plots showing the normalized distributions of answers per question (see Figure C.1-C.5 in Appendix C), some trends indicative to a difference between the models could be seen. As discussed in the results section, in the plots for Questions 2 and 5, which both ask about the appropriateness of the model's confidence for the top prediction, the Base Model showed more responses on answers that begin with "Strongly" than on the more neutral ones, compared to the other two models. This clearly shows that participants found the Base Model's confidence predictions to be more extreme, either in the disagree or agree direction. These bar plots show answers picked by the participants regardless of whether the model predicted correctly or not. Therefore, it is possible to see how the Base Model's overconfidence could appear appropriate in cases where the prediction was correct and very distant from that when it was incorrect. That said, no statistically significant results were found, so there are few conclusions we can draw from this data.

5.4 Calibration Error

Two kinds of statistical analysis were employed to better interpret this data: Frequentist and Bayesian. No statistically significant results were found from the former, showing that the difference in calibration error was not significant enough to be statistically relevant. That said, by looking at the mean differences, it was clear that the Base Model was the most overconfident. This was further demonstrated by the Bayesian analysis conducted on the same data, which showed that a Bayes Factor of 26.18 for the Base Model was notably higher than 0.39 and 0.18 for the Deep Ensemble and MC Dropout, respectively. This also aligns with the earlier predictions and assumptions, stating that the Base Model, with no UQ implemented, will predict with excessive overconfidence compared to models incorporating various UQ techniques.

5.5 Cannot answer

We also provided participants with a "Cannot answer" option for all Likert-type questions to account for cases where they could not make sense of the question or felt that none of the other responses were applicable. We included this option to avoid forcing participants to give a response that did not accurately represent their opinion.

For the Likert answer analysis, we did not consider responses labeled "Can't answer." We omitted these because the Likert answers were numerically encoded from 1 (Strongly Disagree) to 5 (Strongly Agree), while "Can't answer" would be encoded as 6. Including them would have artificially skewed averages, increased standard deviations, and ultimately made the results uninterpretable.

Below is a summary of participants who used the "Cannot answer" option:

Participant	Cannot Answers (CA)
06	15
08	1
11	3
12	11
13	6
15	9
Total	**45**

Table 5.1: Summary of "Cannot answer" responses across participants.

5.6 Key Takeaways

The Base Model is quite overconfident, especially compared to the other two. We showed that some trends hint at users perceiving this overconfidence, but there is no statistically significant evidence to back this up. Lastly, we showed that the models performed worse in the handwritten digits dataset created by the participants in this experiment than in the MNIST one, proving the importance of a real-world dataset for testing UQ instead of synthetic alterations of existing datasets.

5.7 Limitations

As with most experiments, some limitations were discovered after the data collection was completed. The key ones are listed below:

- **Phrasing of the questions:** We carefully crafted the five questions, but some participants criticized them as confusing and somewhat too similar. The scale on which they could answer was also relatively small. Implementing a question asking, "On a scale of 1-10, how much do you agree with the above statement?" would have provided finer outputs to be used in analyzing the three models and more straightforward interpretations for the participants.
- **Problems with the web app:** The web app received much attention as it was one of the central elements developed for this experiment, but due to time constraints, the final version was not fully polished. Some bugs arose during the experiment, causing participants to be kicked out and lose progress—though no collected data was ever lost. Using two tablets simultaneously caused problems, as the app—running on a single computer—initially could not handle multiple inputs and required modifications. A few participants also inadvertently relaunched the browser page, which wiped their progress. Overall, these issues made the task more cumbersome for participants and the experimenter, possibly impacting the experience and the results.
- **Sample size:** The sample size of 20 participants ensured that the experiment was completed quickly but made it difficult to gather meaningful conclusions in the analysis of the data gathered.
- **Generalizability of the task:** We discuss this in more detail below in Section 5.8, but handwritten digit classification is a straightforward task that does not generalize well to other fields or offer many interesting comparisons.

5.8 Future Work

This thesis accomplished nearly everything it set out to do, and the approach employed was a solid foundation for investigating how people perceive uncertainty. However, there are several different

ways that this work could be improved or altered to be more informative.

While classifying handwritten digits is a relatively simple task, it served as a valuable starting point for exploring how users perceive uncertainty in model predictions. An improvement in the future is to apply the same techniques - mixing uncertainty methods with real human feedback - to more complex and significant domains. For instance, experiments classifying medical images, such as detecting problems in X-rays or detecting early signs of disease in MRI scans, could benefit from learning how users (particularly doctors or technicians) perceive and respond to various forms of uncertainty. The stakes are much higher in these domains, so explaining and understanding uncertainty is even more critical.

This research can also be applied to natural language processing models, particularly chatbots and large language models such as ChatGPT, DeepSeek, and more. Recent surveys like Shorinwa et al. (2024) show a growing number of methods for quantifying uncertainty in transformer-based models, from confidence intervals to decoding strategies. These developments open many paths to study how trust and uncertainty interact in NLP applications. These models generate responses based on context but typically do not indicate to users how certain or uncertain they are. An obvious next step would be investigating ways of clearly indicating uncertainty in text and how this influences trust. Even minor actions, such as indicating when a model is uncertain about its responses or offering alternative options, would help users better understand and enable them to make wiser decisions.

Another idea is to show uncertainty differently. This work considered confidence shown in percentages and bar charts, but future studies could look at how various images or methods of displaying uncertainty influence user trust, satisfaction, and comprehension, particularly for various user groups (such as experts and non-experts).

This thesis examined a basic task, but the larger concepts of uncertainty quantification, trust, and human feedback can be applied to many fields. By investigating broader uses and new forms of interaction, future studies can make AI systems in daily life more understandable and trustworthy.

6 Conclusions

This work aimed to understand better how various UQ methods impact user perceptions of a model's outputs. For this, a web app was developed, allowing easy customization of how the experiment can be run, making it easily adaptable to new model structures, and offering a way to collect user input. Furthermore, we created a methodology for experimentation in this field. This set a series of questions to model user perception, numerical and graphical visualizations of confidence, and a data analysis process aimed at better understanding the impact of various model architectures on the participants. This provides the base for future work in the field, giving an idea of how the methodology can be implemented and what results can be gathered from it.

With the analysis of all the collected data, various conclusions were drawn. Firstly, a statistically significant difference in confidence between the Base Model and the other two models was found, showing the expected overconfidence of a model that does not implement any UQ. Then, we showed that no significant difference in the accuracies of the three models existed, meaning that any other differences discovered cannot be attributed to drastically different performances, allowing more conclusions to be drawn. Next, no statistically significant difference was found when analyzing the responses to the five questions. However, we showed trends discovered through various graphs for users perceiving the Base Model as more extreme in its predictions. Lastly, with the analysis of calibration error, we found that the Base Model was visibly overconfident, especially when comparing the Bayes Factors between the three models.

In summary, this thesis laid the groundwork for future research in the field by supplying the following contributions:

1. **Human-generated dataset for handwritten digit classification:** Creating this dataset was an essential step in achieving the key goal of this work. Synthetic datasets used to test epistemic uncertainty do not provide as much ecological validity as a human-generated one can.

2. **Evaluation of three UQ methods (Base Model, MC Dropout, Deep Ensemble)**

in a real-world setting: There are many ways of measuring epistemic uncertainty in a machine learning model. For this thesis, MC Dropout and Deep Ensemble provided a look into how UQ can work and allowed for easy comparison between them to find any differences in user perception.

3. **Web-based experimental interface for collecting and evaluating user trust in model predictions:** Data collection is a key step in any experiment, and the web app developed here provides a great starting point for anybody attempting to measure similar phenomena between various models.

4. **Quantitative and qualitative analysis of how different UQ methods impact user perception:** Measuring user perception is not a trivial task. However, the analysis found in this thesis provides just that—a comparison of how users perceived the three models by comparing their answers to the five questions.

References

Abid, A., Abdalla, A., Abid, A., Khan, D., Alfozan, A., & Zou, J. (2019). Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*.

Filos, A., Farquhar, S., Gomez, A. N., Rudner, T. G., Kenton, Z., Smith, L., ... Gal, Y. (2019). A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks. *arXiv preprint arXiv:1912.10481*.

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

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., ... others (2023). A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, *56*(Suppl 1), 1513–1589.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International conference on machine learning* (pp. 1321–1330).

Gustafsson, F. K., Danelljan, M., & Schon, T. B. (2020). Evaluating scalable bayesian deep learning methods for robust computer vision. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops* (pp. 318–319).

Guth, S., Mojahed, A., & Sapsis, T. P. (2024). Quality measures for the evaluation of machine learning architectures on the quantification of epistemic and aleatoric uncertainties in complex dynamical systems. *Computer Methods in Applied Mechanics and Engineering*, *420*, 116760.

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems* (Vol. 30).

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, *30*.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Mukhoti, J., Kirsch, A., van Amersfoort, J., Torr, P., & Gal, Y. (2021b). Deterministic neural networks with inductive biases capture epistemic and aleatoric uncertainty. *arXiv preprint arXiv:2102.11582*, *2*.

Mukhoti, J., Kirsch, A., van Amersfoort, J., Torr, P. H., & Gal, Y. (2021a). Deterministic neural networks with inductive biases capture epistemic and aleatoric uncertainty. *arXiv preprint arXiv:2102.11582*.

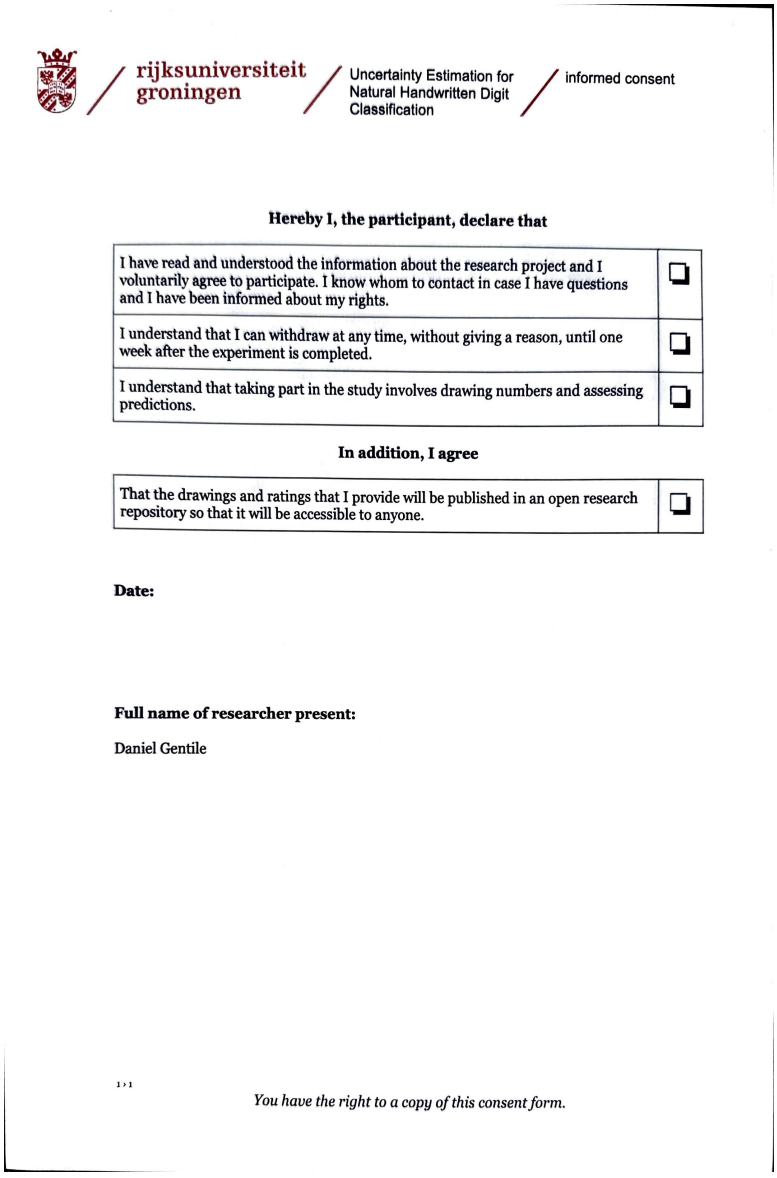
Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., ... Snoek, J. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, *32*.

Shorinwa, O., Mei, Z., Lidard, J., Ren, A. Z., & Majumdar, A. (2024). A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *arXiv preprint arXiv:2412.05563*.

Valdenegro-Toro, M., & Mori, D. S. (2022). A deeper look into aleatoric and epistemic uncertainty disentanglement. In *2022 ieee/cvf conference on computer vision and pattern recognition workshops (cvprw)* (p. 1508-1516).

Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

A Participant Handouts



rijksuniversiteit groningen / Uncertainty Estimation for Natural Handwritten Digit Classification / informed consent

Hereby I, the participant, declare that

I have read and understood the information about the research project and I voluntarily agree to participate. I know whom to contact in case I have questions and I have been informed about my rights.	❑
I understand that I can withdraw at any time, without giving a reason, until one week after the experiment is completed.	❑
I understand that taking part in the study involves drawing numbers and assessing predictions.	❑

In addition, I agree

That the drawings and ratings that I provide will be published in an open research repository so that it will be accessible to anyone.	❑

Date:

Full name of researcher present:

Daniel Gentile

1 / 1

You have the right to a copy of this consent form.

Figure A.1: Consent form for participants.

Why this research?

- We are interested in whether different Machine Learning (Artificial Intelligence) methods give perceivably better predictions and uncertainty estimations. That is: do people find the predictions to be correct, and do they find the predictions are to be appropriately confident. We use handwritten numbers as a simple task, but future research may extend to other Machine Learning systems.

What do we ask of you during the research?

- We will ask you to draw handwritten numbers with the laptop. It is okay if those drawings are not perfect.
- Then we will ask the Machine Learning system to try to predict which number you drew.
- After this you will be asked to rate the prediction of the Machine Learning system on
  a) How correct is the prediction?
  b) How appropriate is the confidence of the prediction?

How will we treat your data?

- Your handwritten numbers and the ratings that you give will be stored with a randomly generated anonymous subject number. We will keep a record of your name and which subject number belongs to you for one week. During this week you may request your data to be removed. After this week, the record will be destroyed, the data will be fully anonymous and you will not be able to request your data to be deleted anymore.
- We will publish the anonymous handwritten numbers and ratings in a public online database and encourage other researchers to build safe Machine Learning systems. Additionally, your numbers and ratings will be stored on a secure server owned by the University.
- Your data will be analyzed for a BSc thesis project. This involves temporarily storing your handwritten digits and ratings on the student's personal computer until the analysis is completed.
- The signed consent form will be archived securely at the University and will only be made available for auditing. These will be stored for ten years.

As a research participant, you have the right to a copy of this research information.

Figure A.2: Page 1 of the information handout.

Do you have to participate in this study?

Participation in this study is voluntary. If you decide not to participate, you do not need to explain why, and there will be no negative consequences for you. You have the right to withdraw from this study at any time until one week after the experiment.

What else do you need to know?

We encourage you to share any questions or comments you may have about the study. You can ask your questions now, during, or after the study. You can do so by speaking with one of the researchers present or by emailing the research team Daniel Gentile (d.gentile@rug.nl), drs. Ivo de Jong (ivo.de.jong@rug.nl), dr. Matias Valdenegro Toro (m.a.valdenegro.toro@rug.nl) or dr. Andreea Sburlea (a.i.sburlea@rug.nl).

Do you have questions/concerns about your rights as a research participant or about the conduct of the research team? In that case, please contact the Research Ethical Review Committee (CETO) of the Faculties of Arts, Philosophy, and Science & Engineering of the University of Groningen: ceto@rug.nl.

Do you have questions or concerns regarding the handling of your personal data? You may also contact the University of Groningen's Data Protection Officer: privacy@rug.nl.

As a research participant, you have the right to a copy of this research information.

Figure A.3: Page 2 of the information handout.

B Web App Images

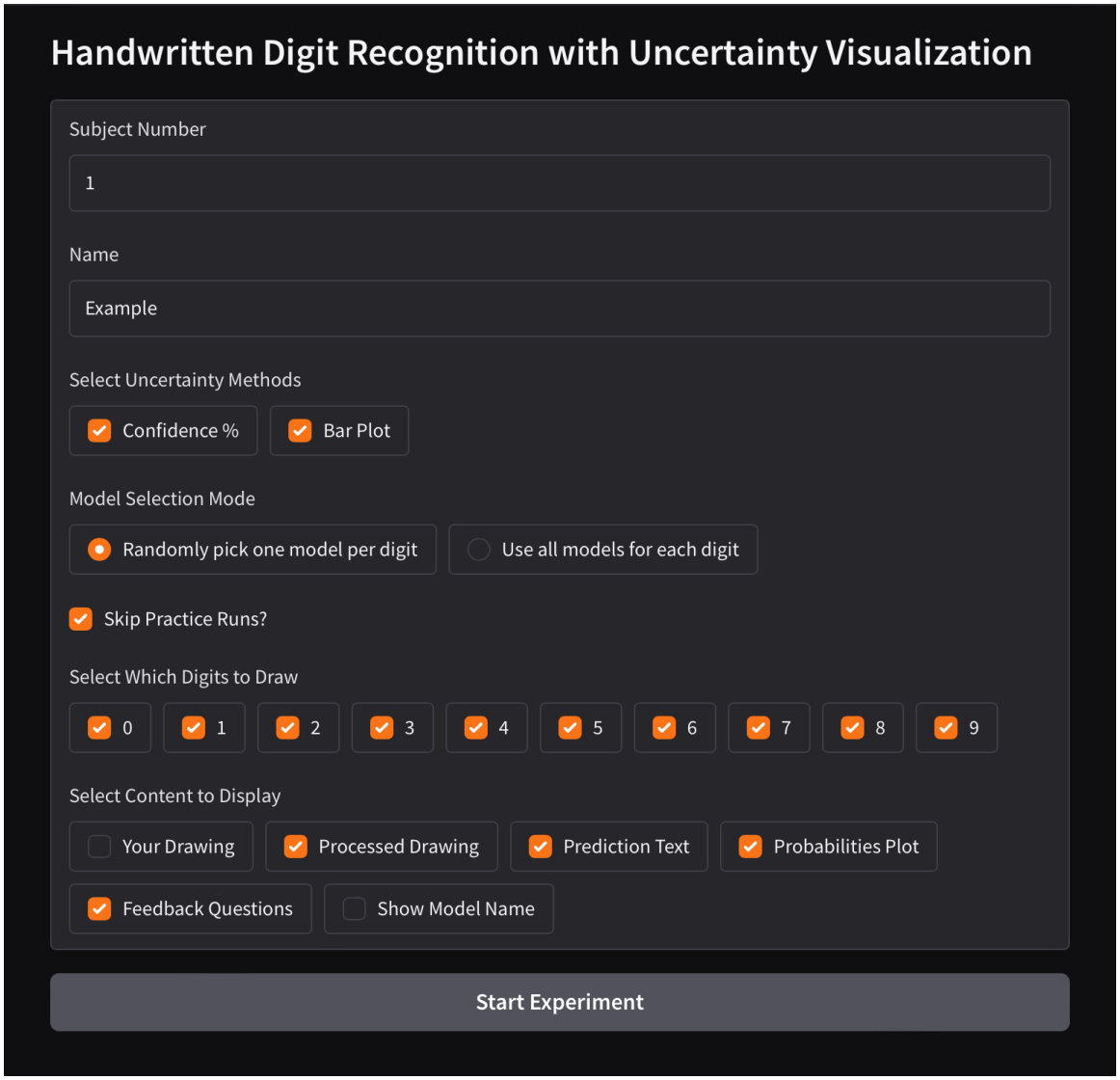


Figure B.1: Web App: Initial page for setup (Not shown to participants).

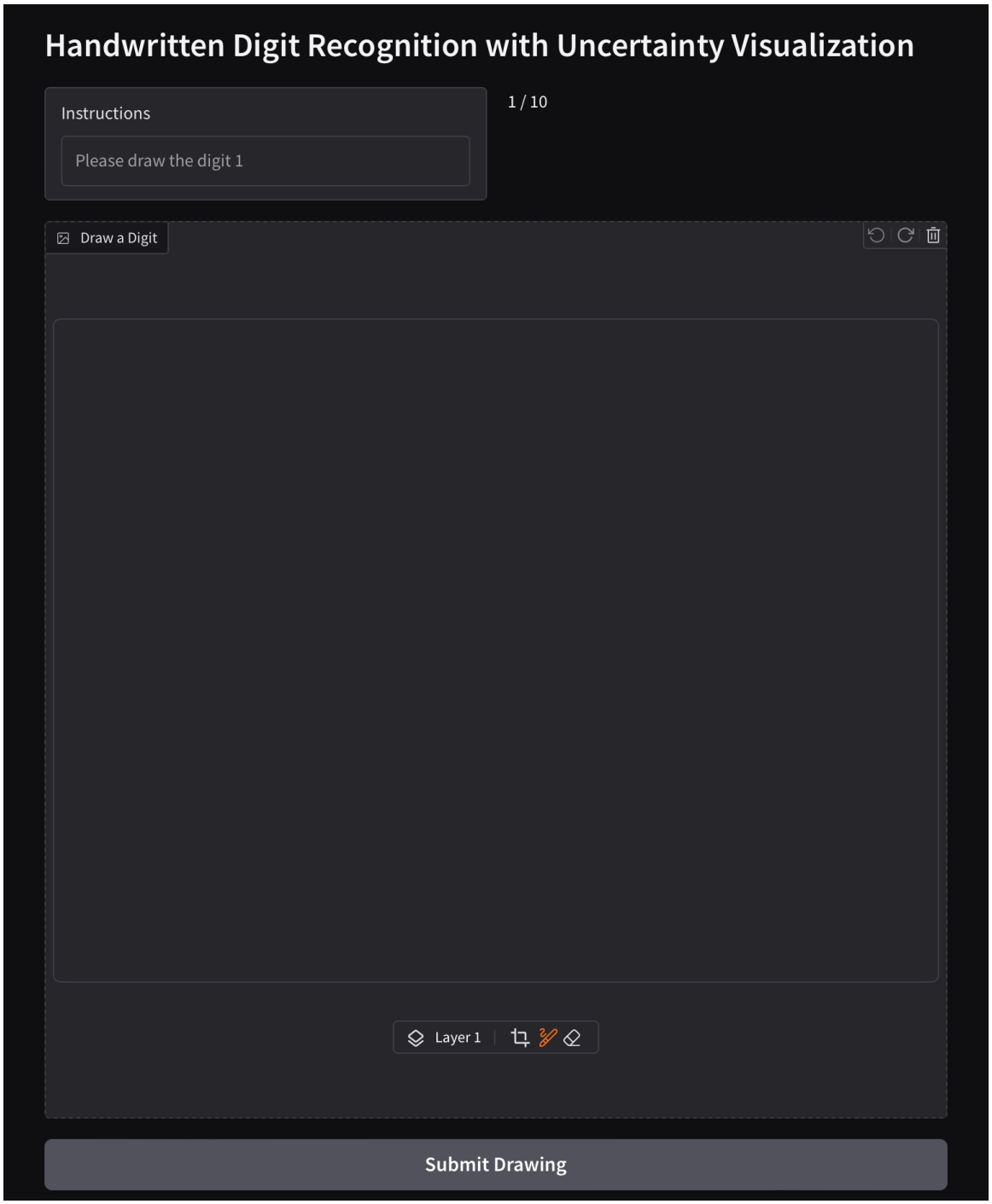


Figure B.2: Web App: First Page for Participants with Empty Canvas.

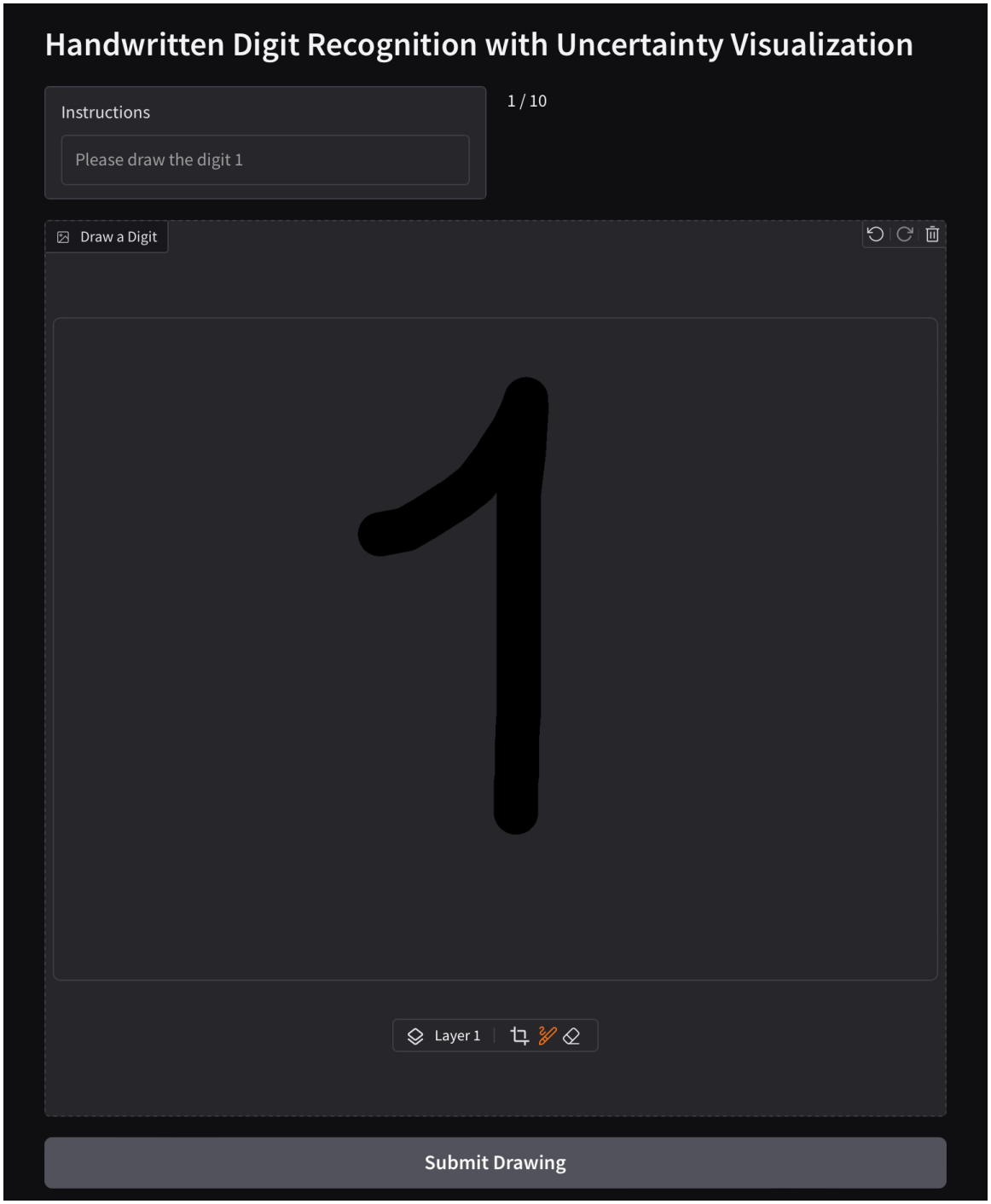


Figure B.3: Web App: Example of drawn digit on canvas.

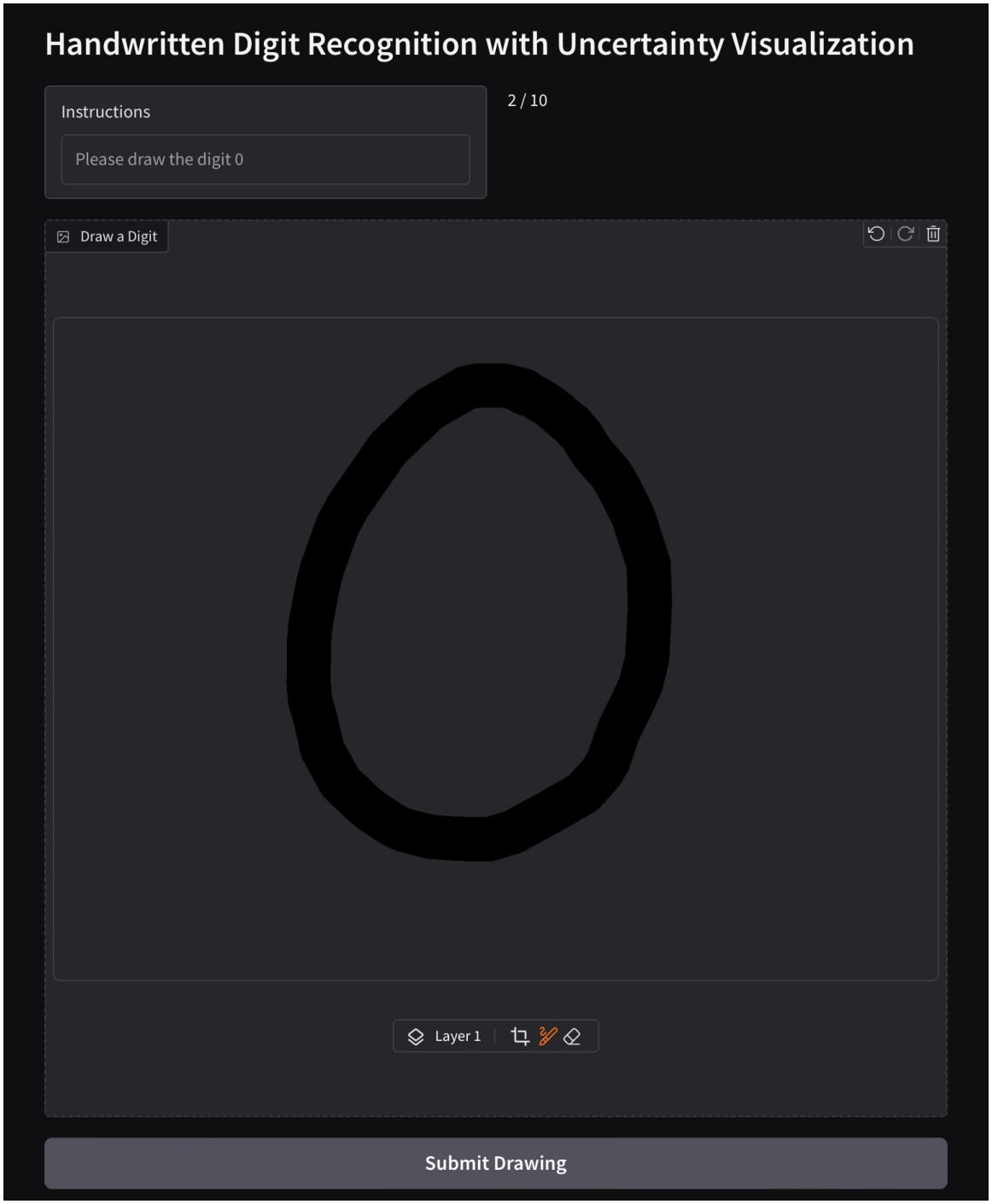


Figure B.4: Web App: Example of drawn digit on canvas 2.

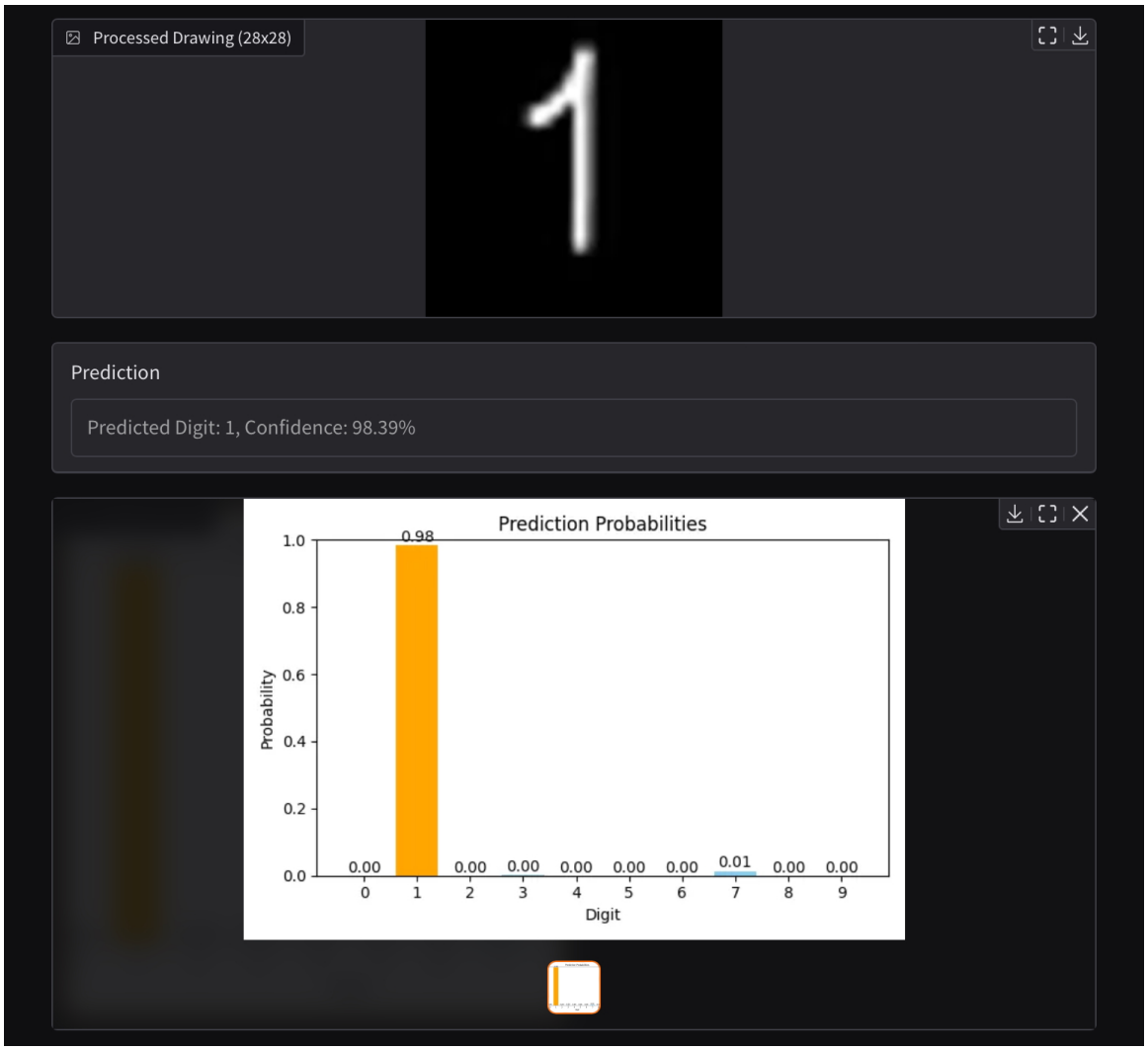


Figure B.5: Web App: Results page showing initial drawing, model prediction with confidence, and bar plot

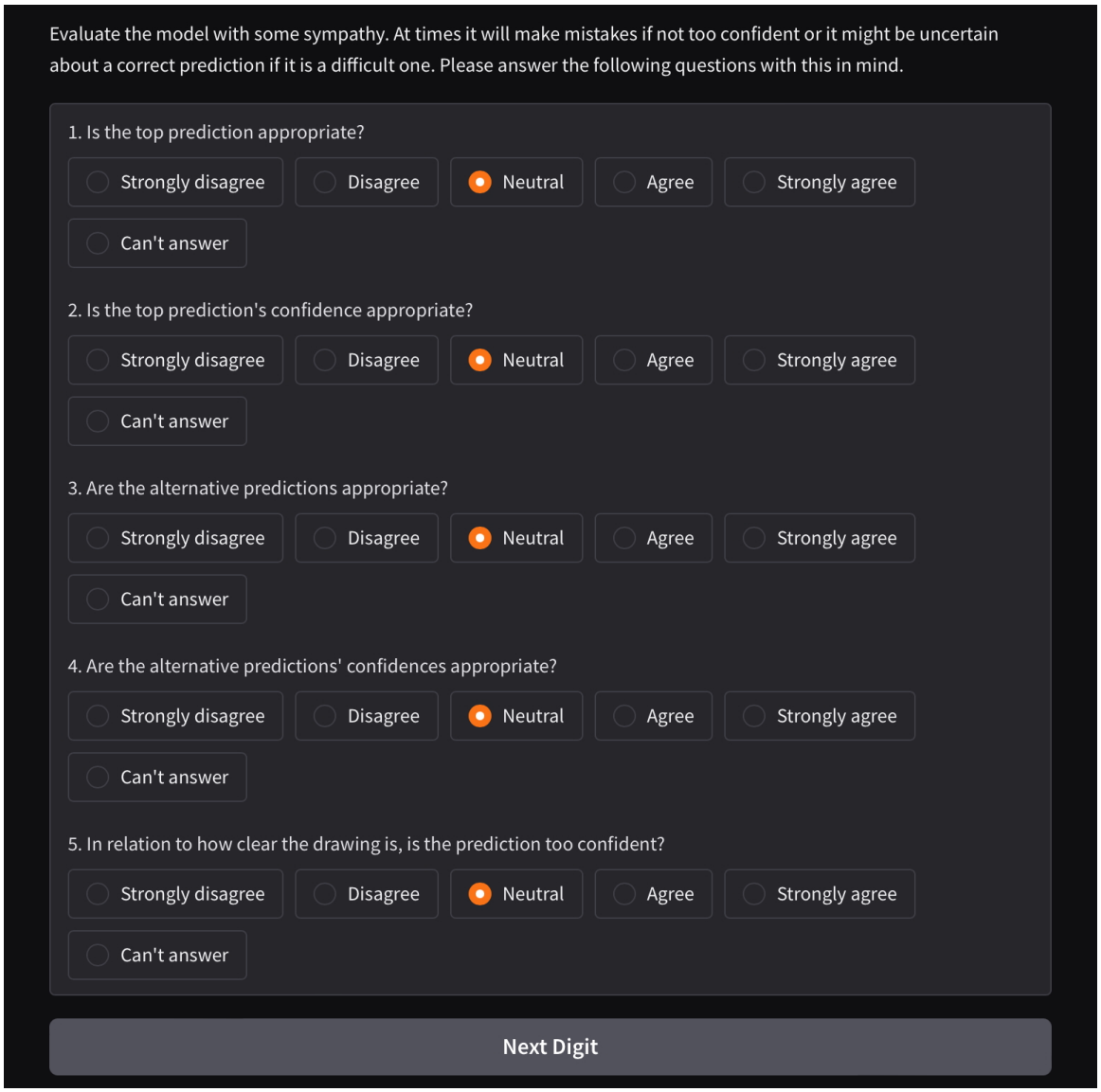


Figure B.6: Web App: Liker scale questions regarding results.

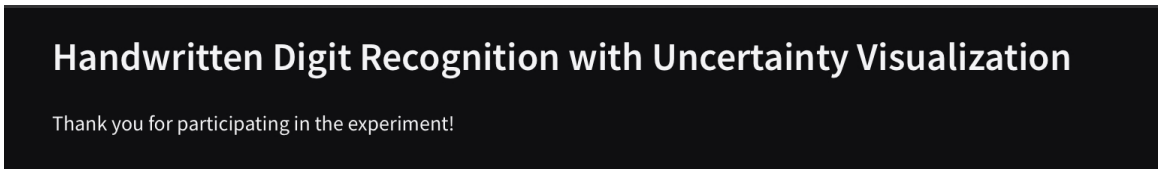
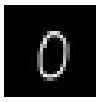
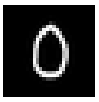


Figure B.7: Web App: Thank you page shown after completion of experiment.

C Tables and Graphs

C.1 Examples of Digits Drawn by Participants

Table C.1: Examples of handwritten digits.

Digit	Example 1	Example 2	Example 3	Example 4	Example 5
Digit 0					
Digit 1					
Digit 2					
Digit 3					
Digit 4					
Digit 5					
Digit 6					
Digit 7					
Digit 8					
Digit 9					

C.2 Normality Testing Results

Table C.2: Summary of Normality Test Results by Model and Variable.

Variable	Group	N	Mean	StdDev	Skew	Kurtosis	Shapiro_W	Shapiro_p
confidence	Base Model	56	92.561	14.953	-2.306	4.731	0.571	0.000
confidence	MC-Dropout	73	79.329	22.980	-0.853	-0.811	0.806	0.000
confidence	Ensemble Model	71	82.443	22.720	-1.251	0.720	0.786	0.000
mean_accuracy	Base Model	20	76.250	29.154	-1.006	0.243	0.800	0.001
mean_accuracy	Ensemble Model	19	76.654	22.618	-0.686	0.028	0.875	0.017
mean_accuracy	MC-Dropout	20	79.083	20.615	-0.330	-1.418	0.815	0.001
q1_answer	Base Model	55	4.273	1.484	-1.712	1.062	0.509	0.000
q1_answer	MC-Dropout	73	4.205	1.443	-1.452	0.350	0.578	0.000
q1_answer	Ensemble Model	69	4.188	1.458	-1.478	0.451	0.585	0.000
q2_answer	Base Model	55	4.018	1.509	-1.271	-0.048	0.653	0.000
q2_answer	MC-Dropout	73	3.877	1.433	-0.951	-0.594	0.751	0.000
q2_answer	Ensemble Model	69	3.971	1.424	-1.056	-0.413	0.719	0.000
q3_answer	Base Model	52	3.846	1.433	-0.938	-0.587	0.762	0.000
q3_answer	MC-Dropout	72	3.597	1.083	-0.487	-0.617	0.882	0.000
q3_answer	Ensemble Model	62	3.484	1.555	-0.418	-1.451	0.801	0.000
q4_answer	Base Model	50	3.800	1.400	-0.947	-0.471	0.781	0.000
q4_answer	MC-Dropout	72	3.514	1.210	-0.585	-0.713	0.865	0.000
q4_answer	Ensemble Model	63	3.413	1.531	-0.228	-1.589	0.804	0.000
q5_answer	Base Model	53	2.245	1.385	0.780	-0.776	0.805	0.000
q5_answer	MC-Dropout	71	2.254	1.216	0.754	-0.476	0.846	0.000
q5_answer	Ensemble Model	66	2.500	1.470	0.293	-1.468	0.818	0.000
acc_conf_diff	Base Model	20	-15.661	23.033	-0.506	-1.248	0.878	0.016
acc_conf_diff	Ensemble Model	19	-4.918	21.120	0.038	0.517	0.964	0.663
acc_conf_diff	MC-Dropout	20	-0.547	16.849	-0.301	-1.066	0.932	0.169

C.3 User Feedback Answer Distributions

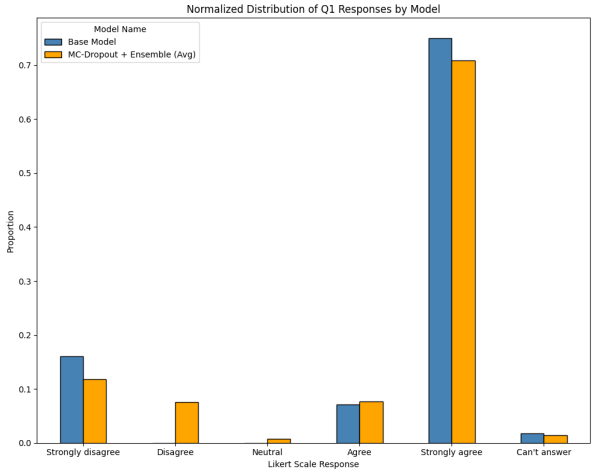


Figure C.1: Comparison of proportions of answers between Base Model and average between MC Dropout and Deep Ensemble for Q1.
Is the top prediction appropriate?

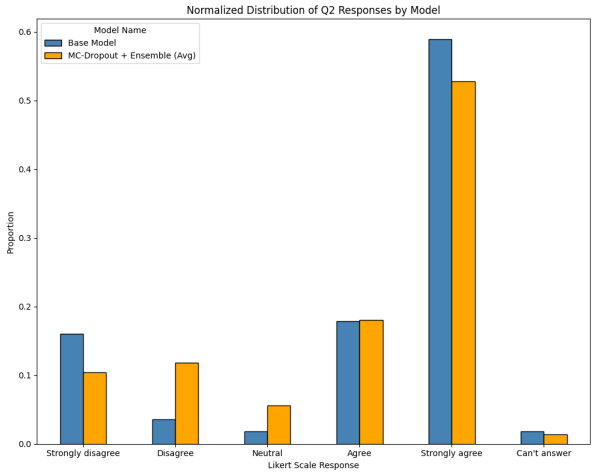


Figure C.2: Comparison of proportions of answers between Base Model and average between MC Dropout and Deep Ensemble for Q2.
Is the top prediction's confidence appropriate?

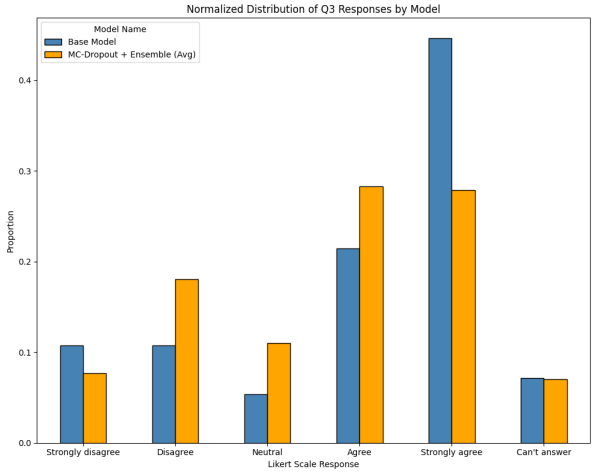


Figure C.3: Comparison of proportions of answers between Base Model and average between MC Dropout and Deep Ensemble for Q3.
Are the alternative predictions appropriate?

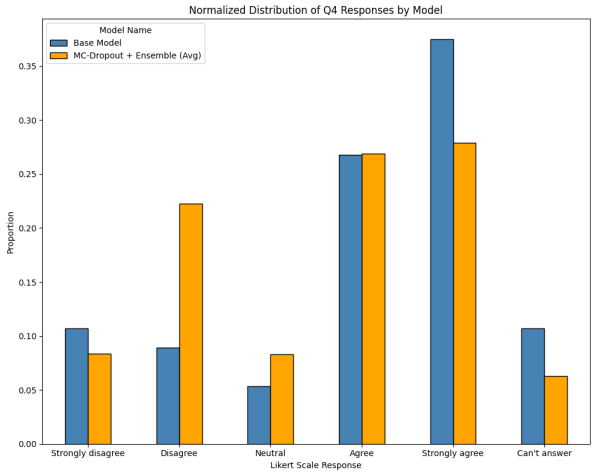


Figure C.4: Comparison of proportions of answers between Base Model and average between MC Dropout and Deep Ensemble for Q4.
Are the alternative predictions' confidences appropriate?

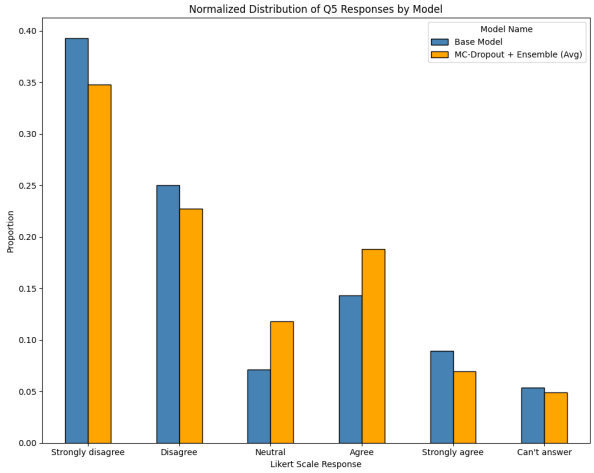


Figure C.5: Comparison of proportions of answers between Base Model and average between MC Dropout and Deep Ensemble for Q5.
In relation to how clear the drawing is, is the prediction too confident?