



Universität Augsburg  
Fakultät für Angewandte  
Informatik

# Calibrated and Uncertainty-aware Multimodal Emotion Recognition

Nicolas Kolbensschlag

Lukas Stappen, M.Sc.

Bachelor thesis, 2021

# Agenda

---

- 1 Uncertainty in Deep Learning (Background)
- 2 Predictive Uncertainty, Subjectivity and Emotion Recognition (Research Goals)
- 3 Obtaining Estimates of Predictive Uncertainty (Approach and Results)
- 4 Evaluation and Discussion (Conclusion)



# Agenda

---

- 1** Uncertainty in Deep Learning (Background)
- 2** Predictive Uncertainty, Subjectivity and Emotion Recognition (Research Goals)
- 3** Obtaining Estimates of Predictive Uncertainty (Approach and Results)
- 4** Evaluation and Discussion (Conclusion)

# Uncertainty in Deep Learning

## Introduction to Uncertainty

- Imagine a model (e.g. a neural network), which makes a **prediction**  $\hat{y}$ .
- But **how confident** (or the opposite uncertain) is the model about this prediction?
- Fundamental intuition:
  1. Define true uncertainty.
  2. Measure/predict uncertainty, alongside with the actual prediction  $\hat{y}$ .
  3. Model is referred to as **well-calibrated**, if true uncertainty matches predicted uncertainty.
- Usually true uncertainty is defined as the prediction error, so **we want the models uncertainty to be high if its error (in terms of  $\hat{y}$ ) is high** and vice-versa.

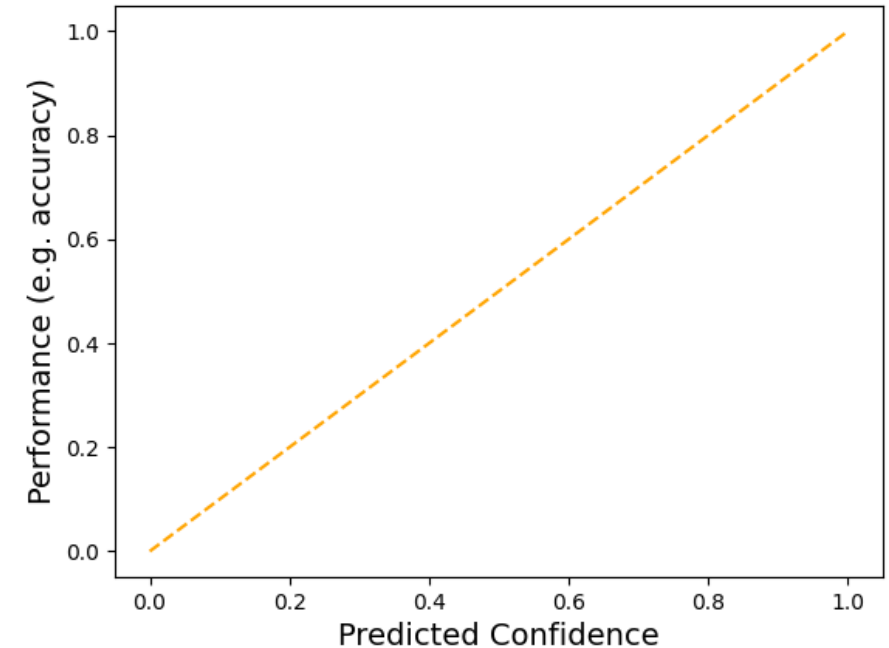


Fig.: *Well-calibrated* uncertainty measurements: the predicted confidence exactly matches the prediction error (here, represented by accuracy).

# Uncertainty in Deep Learning

## Types and Sources of Uncertainty

- **Aleatory** (*Data*) Uncertainty:
  - Caused by noise in the sample, as it leads to non-determination of an estimation problem.
  - Imagine the landing point of an arrow.
- **Epistemic** (*Model/Knowledge*) Uncertainty:
  - Appears if the model lacks in knowledge on a sample.
  - E.g., the image of a ship, fed into a cat-vs.-dog-classifier.

# Uncertainty in Deep Learning

## Measuring Predictive Uncertainty

- There are many approaches to obtain measurements of uncertainty from an neural network.
- For classification: most common and easiest approach is to apply softmax activation function on output layer.
- But what about regression?
  - Bayesian modelling:
    - Obtain multiple forecasts for  $\hat{y}$  (with possibly different outcomes).
    - Use the mean of them as final prediction and the **variance** among them **as measurement for uncertainty**.
    - Approaches:
      - Monte Carlo Dropout: Enable dropout during inference.
      - Ensemble Averaging: Train multiple model from varying starting point (seeds).
      - ...
  - Distributional parameter estimation:
    - Let model output parameters for probability distribution over  $\hat{y}$ , instead of  $\hat{y}$ .
- Note: approaches, mentioned for regression are possible for classification, too.

# Uncertainty in Deep Learning

## Calibration

- Often predicted uncertainty is not well-calibrated initially.
- *(Re-) calibration* describes the procedure of (attempting to) make non well-calibrated uncertainties well-calibrated.
- Therefore:
  1. Obtain non-well calibrated (*uncalibrated*) measurements of uncertainty  $\hat{U}$  from a hold-out validation set.
  2. Calculate true uncertainty  $U$  of the validation set.
  3. Train an auxiliary model, the *calibrator*, to map  $\hat{U}$  to  $U$ .
  4. Now, the calibrator can be applied on uncalibrated uncertainty quantification obtained from the test set.
- “Features  $\rightarrow$  neural network  $\rightarrow$  uncalibrated uncertainty  $\rightarrow$  calibrator  $\rightarrow$  calibrated uncertainty”
- Often used as calibrator, *Isotonic Regression*: fit non-decreasing function through validation points and interpolate to cover complete target space.

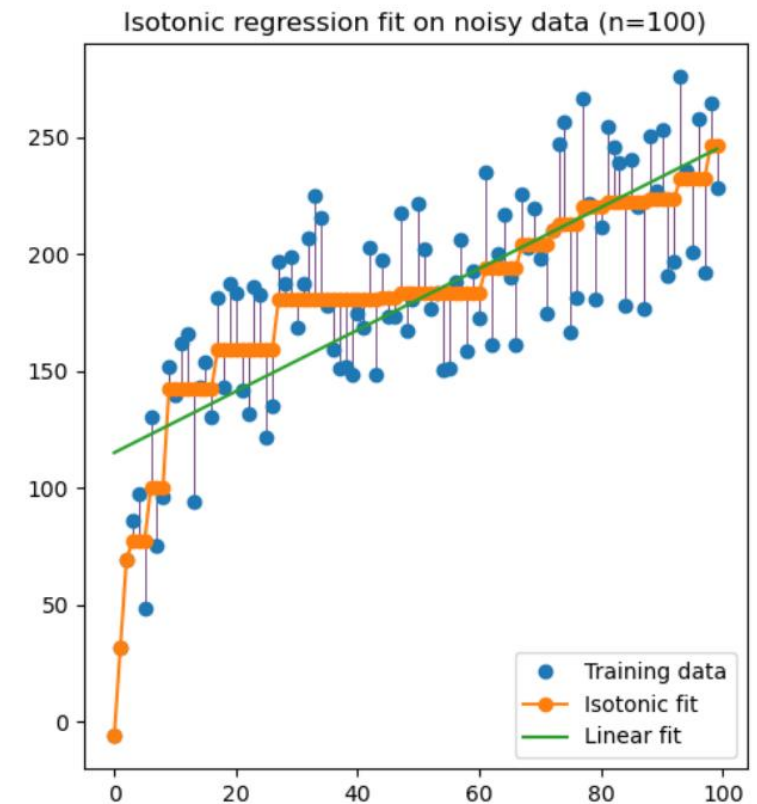


Fig.: Isotonic Regression. Image taken from: [https://scikit-learn.org/stable/auto\\_examples/miscellaneous/plot\\_isotonic\\_regression.html](https://scikit-learn.org/stable/auto_examples/miscellaneous/plot_isotonic_regression.html)

# Agenda

---

- 1 Uncertainty in Deep Learning (Background)
- 2 Predictive Uncertainty, Subjectivity and Emotion Recognition (Research Goals)
- 3 Obtaining Estimates of Predictive Uncertainty (Approach and Results)
- 4 Evaluation and Discussion (Conclusion)



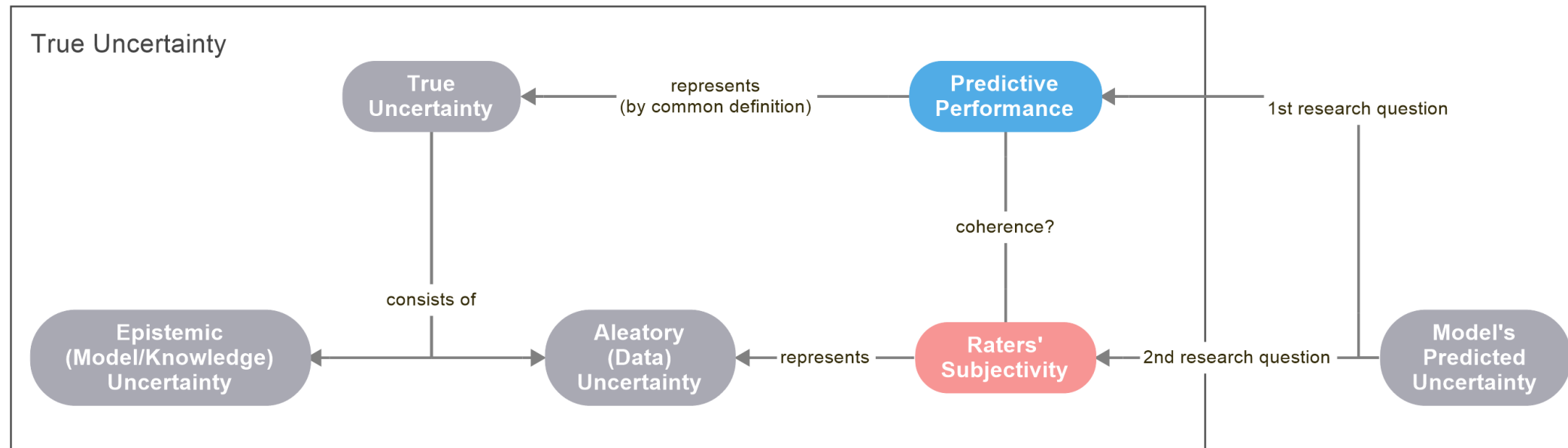
# Predictive Uncertainty, Subjectivity and Emotion Recognition

## Emotion Recognition and the MuSe-CaR dataset

- **Emotion Recognition:** prediction of the emotional state, here described by valence (sentiment) and arousal (excitement).
- The **MuSe-CaR** dataset:
  - Contains YouTube videos of car reviews contiguously (one label per time step) annotated by valence and arousal.
  - For each sample, there are available **5 annotations by different human annotators**.
  - Further, there exists a *ground-truth* calculated, from the multiple annotations, by fusion techniques.
- Aims of this work:
  1. Measure **predictive uncertainty** for the task of emotion recognition with MuSe-CaR.
  2. Investigate **subjectivity among annotations** as source of predictive uncertainty.

# Obtaining Estimates of Predictive Uncertainty

## Research Questions



# Predictive Uncertainty, Subjectivity and Emotion Recognition

## Before we start: Replacing Variance by Correlation

- We want to use approach like Monte Carlo Dropout to measure uncertainty, but such techniques use variance to obtain the final measurement of confidence, as mentioned before.
- We predict steps in time of a **contiguous** emotional state.
- Each point in time  $t$  is rated based on the evaluation of the signal at  $t$  and the annotation of the latest time step  $t - 1$ .
- Therefore, we replace variances by correlations:
  - We obtain multiple (differing) predictions, as usual.
  - We compute the **average correlation** (PCC or CCC) **among each pair** of them.
  - As higher the correlations, as higher the model's confidence and vice-versa.

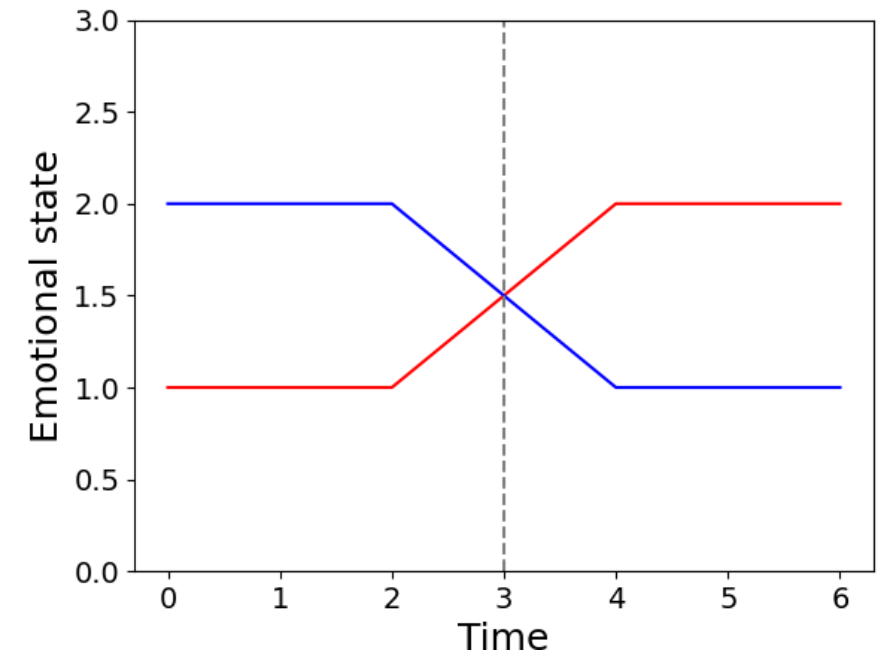


Fig.: At  $x = 3$  both lines have the same value. But when it comes to measuring the disagreement of them, the correlation at this point is much more expressive, especially because the shown values at each step in time come from one contiguous measurement. Each point in time is not a standalone prediction.

# Predictive Uncertainty, Subjectivity and Emotion Recognition

## Correlation Metrics

- Pearson Correlation Coefficient (PCC):

$$PCC(x, y) = \frac{Cov(x, y)}{\sqrt{\sigma^2(x)}\sqrt{\sigma^2(y)}}$$

- Concordance Correlation Coefficient (CCC):

$$CCC(x, y) = \frac{2Cov(x, y)}{\sigma^2(x) + \sigma^2(y) + (\mu(x) - \mu(y))^2}$$

Notation: the mean  $\mu$ , variance  $\sigma^2$  and  $Cov(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \mu(x))(y_i - \mu(y))$

Both are limited to  $[-1, 1]$  with -1 indicating total negative correlation and +1 perfect correlation.

# Agenda

---

- 1 Uncertainty in Deep Learning (Background)
- 2 Predictive Uncertainty, Subjectivity and Emotion Recognition (Research Goals)
- 3 Obtaining Estimates of Predictive Uncertainty (Approach and Results)
- 4 Evaluation and Discussion (Conclusion)

# Obtaining Estimates of Predictive Uncertainty

## Overview over Concepts

- Why investigate subjectivity among annotations?
  - We **want uncertainty** to appear, where the **prediction error** is large (common definition).
  - But why should it indeed appear there? Or, why actually is there prediction error, if it is?
  - Whether model nor anybody know the prediction error during inference. So, there must be underlying sources that cause it (refers to *data* and *knowledge* uncertainty).
  - We investigate **subjectivity among annotations** as indicator for where we **expect uncertainty**.
  - So, we assume, it is a source of measured uncertainty.
- Two main conceptional approaches:
  - Local uncertainty quantification: for now, anything happens **per time step**, or *locally*.
  - Global uncertainty quantification: measurements are done for **multiple time steps** (sub-samples, respectively) **together**, or *globally*.

# Obtaining Estimates of Predictive Uncertainty

## Local Uncertainty Quantification: Formalization

### True Uncertainty (at time step j)

- **Predictive Performance** (left figure)

$$P_j = PCC(y_{j-t} \dots y_j, \hat{y}_{j-t} \dots \hat{y}_j) \text{ with } t = 2$$

- With the prediction for the emotion  $\hat{y}_i$  and the label  $y$ .

- **Subjectivity among raters** (right figure)

$$S_j = \frac{1}{10} \sum_{i=1}^5 \sum_{k=i+1}^5 PCC(x_{i,j-t} \dots x_{i,j}, x_{k,j-t} \dots x_{k,j-t}) \text{ with } t = 2$$

- With  $x_i$ , the annotation of the i-th rater.

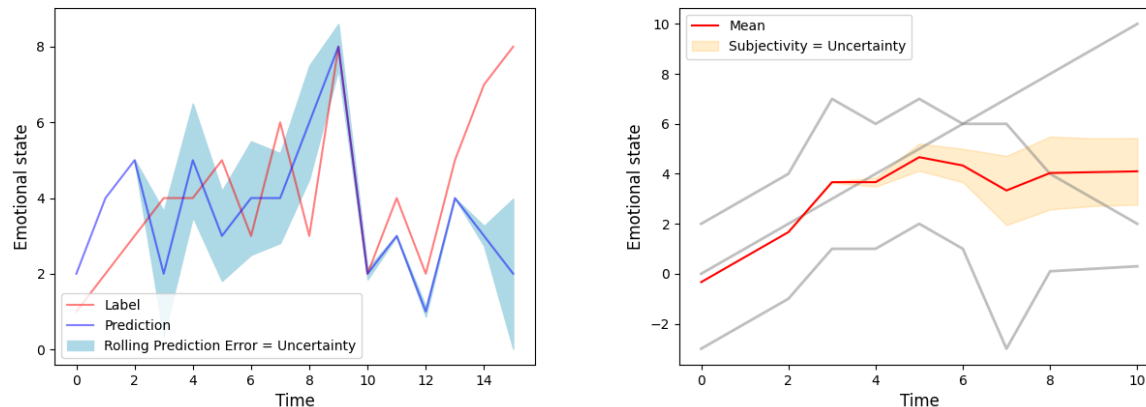


Fig.: Predictive Performance (left) and subjectivity among raters (right).

### Predicted Uncertainty (at time step j)

- Monte Carlo Dropout

$$MCD_j = \frac{1}{10} \sum_{i=1}^5 \sum_{k=i+1}^5 PCC(\hat{y}_{i,j-t} \dots \hat{y}_{i,j}, \hat{y}_{i,k-t} \dots \hat{y}_{k,j}) \text{ with } t = 2$$

- With the model's i-th prediction for the emotional state  $\hat{y}_i$ .

- **tilted CCC (tCCC)**

- Network has three output nodes:

- Mid node: usual CCC loss criterion (used for inference of emotional state).
- Outer nodes:  $RCE_3$  and  $RCE_{10}$  as criterion ( $n$ : sample length).

$$RCE_w(\hat{y}) = \frac{1}{n} \sum_{j=w}^n (1 - PCC(\hat{y}_{j-w+1} \dots \hat{y}_j, y_{j-w+1} \dots y_j))$$

- Creates mini-ensemble with 2 nodes, focusing on different errors (rolling correlation error over the 3 and 10 latest time steps, respectively).
- Used to obtain uncertainty quantification, by computing correlation between them (similar to  $P_j$ ).

# Obtaining Estimates of Predictive Uncertainty

## Local Uncertainty Quantification: Results

- Quantified uncertainty did not match true uncertainty, whether subjectivity nor predictive performance.
- But the reason, that it did not work, was not the unfeasibility of the uncertainty quantification approaches themselves.
- We found a conceptional issue with the emotion recognition itself:
  - With Monte Carlo Dropout, the model is optimized using CCC loss, which evaluates the forecast per complete sample, not per time step.
  - Further, we observed, that during training with tCCC, the outer nodes did not converge significantly.
  - Indicates, that whole emotion recognition (here) works more globally. So, the model is capable to predict longer trends of the emotion state, but not locally (over a few points in time).
  - **If the model's (local) forecasts are prone to be wrong, so meaningless, they naturally lead to non-reliable confidence estimates.**
  - Therefore, we decided to measure uncertainty globally, for multiple time steps together, respectively.



# Obtaining Estimates of Predictive Uncertainty

## Global Uncertainty Quantification: Formalization

- Instead of measuring one value for uncertainty per time step, we now define uncertainty for complete samples.
- But, as we expect uncertainty to vary over time, we split up samples into non-overlapping sub-samples of length 20 (5 seconds).

### True Uncertainty (for one sub-sample)

- **Predictive Performance**

$$P = CCC(y, \hat{y})$$

- With label  $y$  and prediction  $\hat{y}$ .

- **Subjectivity among raters**

$$S = \frac{1}{10} \sum_{i=1}^5 \sum_{k=i+1}^5 CCC(x_i, x_k)$$

- With  $x_i$ , the annotation of the  $i$ -th rater.

### Predicted Uncertainty (for one sub-sample)

- **Monte Carlo Dropout**

$$MCD = \frac{1}{10} \sum_{i=1}^5 \sum_{k=i+1}^5 CCC(\hat{y}_i, \hat{y}_k)$$

- With the model's  $i$ -th prediction for the emotional state  $\hat{y}_i$ .

- **Ensemble Averaging**

- Exact the same, as Monte Carlo Dropout, only with forecasts  $\hat{y}_i$  from five different models/seeds, instead of five forward runs.

# Obtaining Estimates of Predictive Uncertainty

## Global Uncertainty Quantification: Results

True Uncertainty	Emotional Dimension	Method	CCC devel / test [uncalibrated]	CCC devel / test [Isotonic Regression]
Predictive Performance	Valence	Ensemble Averaging	<b>0,18 / 0,27</b>	0,13 / 0,16
		Monte Carlo Dropout	0,18 / 0,24	0,16 / 0,17
	Arousal	Ensemble Averaging	<b>0,23 / 0,12</b>	0,22 / 0,11
		Monte Carlo Dropout	0,20 / 0,05	0,27 / 0,08
Raters' Subjectivity	Valence	Ensemble Averaging	0,03 / 0,04	0,02 / 0,03
		Monte Carlo Dropout	0,03 / 0,04	0,02 / 0,03
	Arousal	Ensemble Averaging	0,08 / 0,04	<b>0,13 / 0,06</b>
		Monte Carlo Dropout	0,04 / 0,01	0,06 / 0,02

# Obtaining Estimates of Predictive Uncertainty

## Global Uncertainty Quantification: Results

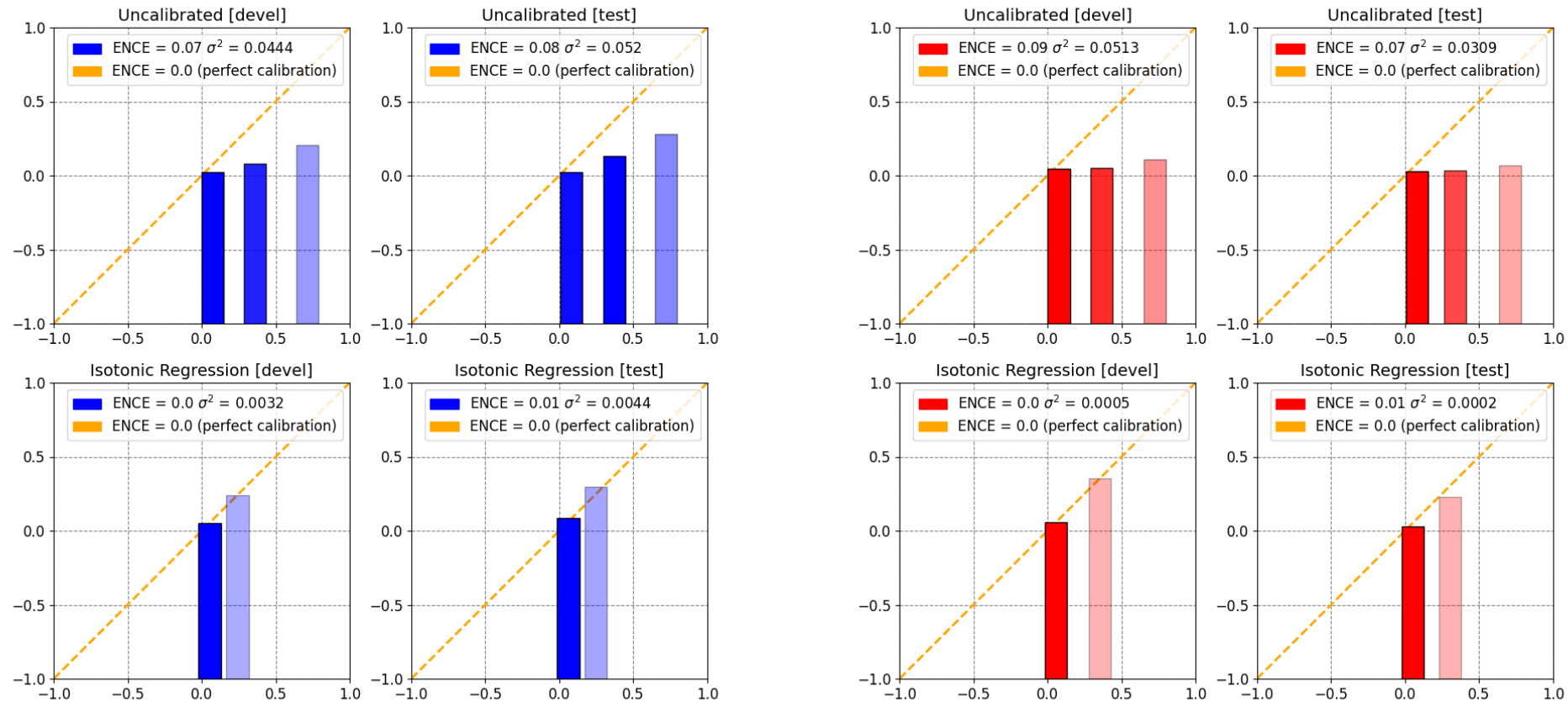


Fig.: Some results visualized with Reliability Diagrams. Predictive Performance as true confidence with Ensemble Averaging for Valence (left) and Subjectivity as true confidence with Ensemble Averaging for Arousal (right). As more transparent a bin, the less (sub-) samples fall into it.

# Agenda

---

- 1 Uncertainty in Deep Learning (Background)
- 2 Predictive Uncertainty, Subjectivity and Emotion Recognition (Research Goals)
- 3 Obtaining Estimates of Predictive Uncertainty (Approach and Results)
- 4 Evaluation and Discussion (Conclusion)

# Evaluation and Discussion

## Summary

- We proposed techniques to transfer the procedure of uncertainty quantification to contiguous forecasting, by replacing variance with correlations.
- We extended the measurement of true uncertainty, which is used to judge quality of predicted uncertainty, by another approach: we defined **subjectivity among multiple (human) raters** as additional comparative value for a model's confidence, next to the **predictive performance** of the model.
- Initial assumption: *Prediction Error*  $\sim^{(1)}$  *Predicted Uncertainty*  $\sim^{(2)}$  *Raters' subjectivity*
- As experiments in local uncertainty failed because of conceptional issues, we restrict further thoughts on global uncertainty measurements.

# Evaluation and Discussion

## Conclusion

- We observed, that there indeed seems to exist a coherence between **predictive performance** and predicted uncertainty.
- The correlation of about 0.2 could even be considered as weak, under the circumstance, that the predictive performance, as measurement of true confidence, unites each source of uncertainty, because no matter what underlying reasons (noise, lacking model knowledge, ...) cause the model's uncertainty, they should reflect on its predictive performance as well
- We believe, that this weaker than expected correlation, is caused by multiple conceptional obstacles in our experiments, in contrast to basic single point regression.
- Particularly, we had to specify a fixed window length for the sub-samples. The choice of 20 was, although considered, nevertheless an arbitrary guess.
- Confidences do not vary in periodic intervals of any fixed length, so we must admit, that the shift from single point regression and variances to contiguous time series forecasting and correlation causes (at least for our approach) vagueness in the measurements.

# Evaluation and Discussion

## Conclusion

- As for the coherence between **subjectivity among annotations** and predicted uncertainty, we observed, that it is, if at all, barely available.
- This might indicate, that subjectivity among annotations is only one source of uncertainty among others.
- Subjectivity among raters indicates a noisy sample (as features did not transport an unambiguous signal), so aleatory uncertainty.
- So, what about epistemic uncertainty?
- Epistemic uncertainty is caused by lacks knowledge from the model about the fed in data, which is not contained in raters' subjectivity.
- The weaker correlations between subjectivity among annotations and predicted uncertainty, than those between predictive performance and predicted uncertainty, might be explained by the disregard of epistemic uncertainty in this definition of true uncertainty.

# Evaluation and Discussion

## Future Work

1. Vary the length of sub-samples (not just 20)
  - First experiments with lengths 10 and 50 do **not** indicate a significant improvement.
2. Incorporate epistemic uncertainty:
  - We could use simpler (actually less powerful) feature embedding techniques and try to observe, if this increases predicted uncertainty and decreases the proportion of aleatory uncertainty relative to the overall uncertainty.
  - Furthermore, we could think about manually quantifying the lack of knowledge for a sub-sample and use this as a definition for true uncertainty, as we did with subjectivity among annotations.
3. Use advanced fusion technique to obtain final prediction from Monte Carlo Dropout and Ensemble Averaging:
  - This would reflect on predictive performance and make it, as a definition of true confidence, more reliable.





*Thank you for your attention!*

*Any questions?*



Nicolas Kolbenschlager  
Universität Augsburg  
nicolas.kolbenschlager@student.uni-augsburg.de  
www.uni-augsburg.de

# Backlog

# Research Questions

