Handling Ambiguity in Emotion: From Out-of-Domain Detection to Distribution Estimation

Wen Wu[1], **Bo Li**[2], **Chao Zhang**[3], **Chung-Cheng Chiu**[2], **Qiujia Li**[2], **Junwen Bai**[2], **Tara N. Sainath**[2], **Philip C. Woodland**[1]

[1] University of Cambridge, UK, [2] Google, LLC, USA, [3] Tsinghua University, China

[1]{ww368, pw117}@cam.ac.uk, [2]{boboli, chungchengc}@google.com, [3]cz277@tsinghua.edu.cn

Abstract

The subjective perception of emotion leads to inconsistent labels from human annotators. Typically, utterances lacking majority-agreed labels are excluded when training an emotion classifier, which cause problems when encountering ambiguous emotional expressions during testing. This paper investigates three methods to handle ambiguous emotion. First, we show that incorporating utterances without majority-agreed labels as an additional class in the classifier reduces the classification performance of the other emotion classes. Then, we propose detecting utterances with ambiguous emotions as out-of-domain samples by quantifying the uncertainty in emotion classification using evidential deep learning. This approach retains the classification accuracy while effectively detects ambiguous emotion expressions. Furthermore, to obtain fine-grained distinctions among ambiguous emotions, we propose representing emotion as a distribution instead of a single class label. The task is thus re-framed from classification to distribution estimation where every individual annotation is taken into account, not just the majority opinion. The evidential uncertainty measure is extended to quantify the uncertainty in emotion distribution estimation. Experimental results on the IEMOCAP and CREMA-D datasets demonstrate the superior capability of the proposed method in terms of majority class prediction, emotion distribution estimation, and uncertainty estimation.

1 Introduction

The inherent subjectivity of human emotion perception introduces complexity in annotating emotion datasets. Multiple annotators are often involved in labelling each utterance and the majority-agreed (MA) class is usually used as the ground truth (Busso et al., 2008; Cao et al., 2014). Utterances that have no majority-agreed (NMA) labels (*i.e.*, with tied votes) are typically excluded during emotion classifier training (Kim et al., 2013; Poria et al., 2017; Zou et al., 2022; Wu et al., 2021), which may cause issues when the system encounters such utterances in practical applications.

This paper investigates three approaches to handling ambiguous emotion data. First, a naive method is tested which aggregates NMA utterances into an additional class when training an emotion classifier. This approach proves problematic as NMA utterances contain a blend of emotions, thereby confusing the classifier and undermining the classification performance.

Then we explore if an emotion classifier can appropriately respond with "I don't know" for ambiguous emotion data that does not fit into any predefined emotion class. This is realised by quantifying the uncertainty in emotion classification using evidential deep learning (EDL) (Sensoy et al., 2018). When a classifier trained on MA data encounters an NMA utterance during the test, the model should identify it as an out-of-domain (OOD) sample by providing a high uncertainty score, indicating its uncertainty regarding the specific emotion class to which the NMA utterance belongs.

Moreover, to obtain fine-grained distinctions between ambiguous emotional data, we re-frame the task from classification to distribution estimation. Consider the example shown in Figure 1 with the annotations assigned to three utterances. Since the majority emotion classes are "angry" for both utterances (a) and (b), they will be assigned the same ground-truth label "angry" in the aforementioned classification system, which implies that they convey the same emotion content and is evidently unsuitable. On the contrary, utterance (c), though being an NMA utterance, is more likely to share similar emotional content with utterance (b). Therefore, in order to obtain more comprehensive representations of emotion content, we further propose representing emotion as a distribution rather

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2078–2093
August 11-16, 2024 ©2024 Association for Computational Linguistics

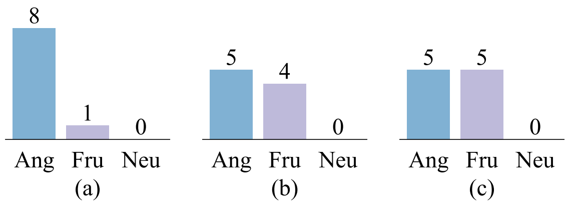


Figure 1: The bar chart shows the number of labels assigned by annotators to the emotion class "angry" (Ang), "frustrated" (Fru), and "neutral" (Neu) in an example. In utterance (a), eight annotators interpret the emotion as angry while one interprets it as frustrated.

than a single class label and re-framing emotion recognition as a distribution estimation problem rather than a classification problem. A novel algorithm is proposed which extends EDL to estimate the underlying emotion distribution given observed human annotations and quantify the uncertainty in emotion distribution estimation. The proposed approach considers all human annotations rather than relying solely on the majority vote class. Multiple evaluation metrics are adopted to evaluate the performance in terms of majority class prediction, uncertainty measure, and distribution estimation. Rather than simply saying "I don't know", the proposed system demonstrates the ability to estimate the emotion distributions of the NMA utterances and also offer a reliable uncertainty measure for the distribution estimation.

Our contributions are summarised as follows. **(i)** To the best of our knowledge, this paper is the first work that treats ambiguous emotion as OOD and detects it by uncertainty estimation; **(ii)** This is the first work that applies EDL to quantify uncertainty in emotion classification; **(iii)** Imposing a single ground truth through majority voting leads to under-representation of minority views. We instead estimate the distribution over emotion classes which provides a more comprehensive representation of emotion content as well as a more inclusive representation of human opinions; **(iv)** A novel algorithm is proposed that extends EDL to quantify uncertainty in emotion distribution estimation.

2 Related work

Human annotators often interpret the emotion of the same utterance differently due to their personal experiences and cultural backgrounds (Busso et al., 2008; Cowen and Keltner, 2017; Sethu et al., 2019). Instead of using the MA annotation as the ground truth label, some research suggests treating emotion classification as a multi-label task (Mower et al., 2010; Zadeh et al., 2018; Chochlakis et al., 2023) where all emotion classes assigned by any annotator are considered as correct classes and the ground truth label is presented as a multi-hot vector. The model is trained to predict the presence of each emotion class for each utterance. An issue with this approach is that it ignores the differences in strengths of different emotion classes.

An alternative approach uses "soft labels" as the proxy of ground truth, which is defined as the relative frequency of occurrence of each emotion class (Fayek et al., 2016; Han et al., 2017; Kim and Kim, 2018). The Kullback–Leibler (KL) divergence or distance metrics between the soft labels and model predictions are used to train the model. However, soft labels, being maximum likelihood estimates (MLE) of the underlying distribution based on observed samples, might not provide an accurate approximation to the unknown distribution when the number of observations (annotations) is limited. Also, although adopting soft labels, those methods still focus on obtaining a "correct" label (*i.e.,* pursuing improved classification accuracy).

So far, the calibration of emotion models has not been extensively studied. In this study, we introduce a novel approach which provides not only better emotion content estimation but also a reliable measure of the model's prediction confidence.

3 Detecting NMA as OOD by quantifying emotion classification uncertainty

As explained in the introduction and confirmed experimentally in Section 7.1, training an emotion classifier with NMA utterances grouped into an additional class degrades the classification performance. This section studies an alternative method. The emotion classifier is trained on MA utterances and NMA utterances are treated as OOD samples. By quantifying uncertainty in emotion classification, the model is expected to output a high uncertainty score when encountering ambiguous emotions, indicating that the utterance doesn't belong to any predefined MA class.

3.1 Limitation of modelling class probabilities with the softmax activation function

A neural network model classifier transforms the continuous logits at the output layer into class probabilities by a softmax function. The model prediction can thus be interpreted as a categorical distribution with the discrete class probabilities associated with the model outputs. The model is then optimised by maximising the categorical likelihood of the correct class, known as the cross-entropy loss.

However, the softmax activation function is known to have a tendency to inflate the probability of the predicted class due to the exponentiation applied to transform the logits, resulting in unreliable uncertainty estimations (Gal and Ghahramani, 2016; Guo et al., 2017). Furthermore, cross-entropy is essentially MLE, a frequentist technique lacking the capability to infer the variance of the predictive distribution.

In the following section, we estimate the model uncertainty using evidential deep learning (EDL) (Sensoy et al., 2018) which places a second-order probability over the categorical distribution.

3.2 Quantify uncertainty in emotion classification by evidential deep learning

Consider an emotion class label as a one-hot vector $\boldsymbol{y}$ where y_k is one if the emotion belongs to class k else zero. $\boldsymbol{y}$ is sampled from a categorical distribution $\boldsymbol{\eta}$ where each component η_k corresponds to the probability of sampling a label from class k:

$$\boldsymbol{y} \sim \mathrm{P}(\boldsymbol{y}|\boldsymbol{\eta}) = \mathrm{Cat}(\boldsymbol{\eta}) = \eta_k^{y_k}. \tag{1}$$

To model the probability of the predictive distribution, we assume the categorical distribution is sampled from a Dirichlet distribution:

$$\boldsymbol{\eta} \sim \mathrm{p}(\boldsymbol{\eta}|\boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\eta}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \eta_k^{\alpha_k - 1} \tag{2}$$

where $B(\cdot)$ is the Beta function, α_k is the hyperparameter of the Dirichlet distribution. $\alpha_0 = \sum_{k=1}^{K} \alpha_k$ is the Dirichlet strength. The output of a standard neural network classifier is a probability assignment over the possible classes and the Dirichlet distribution represents the probability of each such probability assignment, hence modelling second-order probabilities and uncertainty.

Subjective logic (Jsang, 2018) establishes a connection between the Dirichlet distribution and the belief representation in Dempster–Shafer belief theory (Dempster, 1968), also known as evidence theory. Consider K classes each associated with a belief mass b_k and an overall uncertainty mass u, which satisfies $u + \sum_{k=1}^{K} b_k = 1$. The belief mass assignment corresponds to the Dirichlet hyperparameter α_k: $b_k = (\alpha_k - 1)/\alpha_0$, where $e_k = \alpha_k - 1$ is usually termed evidence. The overall uncertainty can then be computed as:

$$u = \frac{K}{\alpha_0}. \tag{3}$$

A neural network $\boldsymbol{f}_{\boldsymbol{\Lambda}}$ is trained to predict $\mathrm{Dir}(\boldsymbol{\eta}^{(i)}|\boldsymbol{\alpha}^{(i)})$ for a given sample $\boldsymbol{x}^{(i)}$ where $\boldsymbol{\Lambda}$ is the model parameters. The network is similar to standard neural networks for classification except that the softmax output layer is replaced with a ReLU activation layer to assure non-negative outputs, which is taken as the evidence vector for the predicted Dirichlet distribution: $\boldsymbol{f}_{\boldsymbol{\Lambda}}(\boldsymbol{x}^{(i)}) = \boldsymbol{e}^{(i)}$. The concentration parameter of the Dirichlet distribution can be calculated as $\boldsymbol{\alpha}^{(i)} = \boldsymbol{f}_{\boldsymbol{\Lambda}}(\boldsymbol{x}^{(i)}) + 1$. Given $\mathrm{Dir}(\boldsymbol{\eta}^{(i)}|\boldsymbol{\alpha}^{(i)})$, the estimated probability of class k can be calculated by:

$$\mathbb{E}[\eta_k^{(i)}] = \frac{\alpha_k^{(i)}}{\alpha_0{}^{(i)}}. \tag{4}$$

3.2.1 Training

For brevity, superscript i is omitted in this section. Given one-hot label $\boldsymbol{y}$ and predicted Dirichlet $\mathrm{Dir}(\boldsymbol{\eta}|\boldsymbol{\alpha})$, the network can be trained by maximising the marginal likelihood of sampling $\boldsymbol{y}$ given the Dirichlet prior. Since the Dirichlet distribution is the conjugate prior of the categorical distribution, the marginal likelihood is tractable:

$$\begin{aligned} \mathrm{P}(\boldsymbol{y}|\boldsymbol{\alpha}) &= \int \mathrm{P}(\boldsymbol{y}|\boldsymbol{\eta})\mathrm{p}(\boldsymbol{\eta}|\boldsymbol{\alpha})\mathrm{d}\boldsymbol{\eta} \\ &= \int \prod_k \eta_k^{y_k} \frac{1}{B(\boldsymbol{\alpha})} \prod_k \eta_k^{\alpha_k - 1} \\ &= \frac{B(\boldsymbol{\alpha} + \boldsymbol{y})}{B(\boldsymbol{\alpha})} = \frac{\prod_{k=1}^{K} \alpha_k^{y_k}}{\alpha_0^{\sum_{k=1}^{K} y_k}}. \end{aligned} \tag{5}$$

It is equivalent to training the model by minimising the negative log marginal likelihood:

$$\mathcal{L}^{\mathrm{NLL}} = \sum_{k=1}^{K} y_k(\log(\alpha_0) - \log(\alpha_k)). \tag{6}$$

Following (Sensoy et al., 2018), a regularisation term is added to penalise the misleading evidence:

$$\mathcal{L}^{\mathrm{R}} = \mathcal{KL}(\mathrm{Dir}(\boldsymbol{\eta}|\tilde{\boldsymbol{\alpha}}) || \mathrm{Dir}(\boldsymbol{\eta}|\mathbf{1})), \tag{7}$$

where $\mathrm{Dir}(\boldsymbol{\eta}|\mathbf{1})$ denotes a Dirichlet distribution with zero total evidence and $\tilde{\boldsymbol{\alpha}} = \boldsymbol{y} + (1 - \boldsymbol{y}) \odot \boldsymbol{\alpha}$ is the Dirichlet parameters after removal of the non-misleading evidence from predicted $\boldsymbol{\alpha}$. This penalty explicitly enforces the total evidence to shrink to zero for a sample if it cannot be correctly classified. The overall loss is $\mathcal{L} = \mathcal{L}^{\mathrm{NLL}} + \lambda\mathcal{L}^{\mathrm{R}}$ where λ is the regularisation coefficient.

4 Emotion distribution estimation

As illustrated in Figure 1, the majority vote class is not sufficient for fine-grained emotion representations. In this section, we describe emotion by a distribution instead of a single class label.

Consider an input utterance $\boldsymbol{x}^{(i)}$ associated with M_i labels from human annotators $\{\boldsymbol{y}_m^{(i)}\}_{m=1}^{M_i}$ where $\boldsymbol{y}_m = [y_{m1}, \ldots, y_{mK}]$ is a one-hot vector. Instead of representing the emotion content by the majority vote class, we propose estimating the underlying emotion distribution $\boldsymbol{\eta}$ based on the observations $\{\boldsymbol{y}_m^{(i)}\}_{m=1}^{M_i}$. The emotion classification problem is thus re-framed as a distribution estimation problem. In contrast to the "soft label" method in Section 2 which approximates the emotion distribution of each $\boldsymbol{x}^{(i)}$ solely based on $\mathcal{D}^{(i)} = \{\boldsymbol{y}_m^{(i)}\}_{m=1}^{M_i}$ by MLE and trains the model to learn this proxy in a supervised manner, the proposed approach meta-learns a distribution estimator $\boldsymbol{f}_{\boldsymbol{\Lambda}}$ across all data points $\mathcal{D}_{\text{meta}} = \{\mathcal{D}^{(i)}\}_{i=1}^{N}$ where N is the number of utterances in training. This uses the knowledge about the emotion expression and annotation variability across different utterances.

For brevity, superscript i is omitted thereafter. Assume $\{\boldsymbol{y}_m\}_{m=1}^{M}$ are samples drawn from a multinomial distribution. Let $\hat{\boldsymbol{y}} = \sum_{m=1}^{M} \boldsymbol{y}_m$ represent the counts of each emotion class:

$$\{\boldsymbol{y}_m\}_{m=1}^{M} \sim \mathrm{P}(\boldsymbol{y}|\boldsymbol{\eta}) = \mathrm{Mult}(\boldsymbol{\eta}, M) \tag{8}$$

$$\mathrm{Mult}(\boldsymbol{\eta}, M) = \frac{\Gamma(M+1)}{\prod_{k=1}^{K}\Gamma(\hat{y}_k+1)}\eta_k^{\hat{y}_k}. \tag{9}$$

The categorical distribution in Eqn. (1) is the special case when $M = 1$.

The network is trained by maximising the marginal likelihood of sampling $\{\boldsymbol{y}_m\}_{m=1}^{M}$ given the predicted Dirichlet prior $\mathrm{Dir}(\boldsymbol{\eta}|\boldsymbol{\alpha})$:

$$\mathrm{P}(\{\boldsymbol{y}_m\}_{m=1}^{M}|\boldsymbol{\alpha}) = \int \mathrm{P}(\{\boldsymbol{y}_m\}_{m=1}^{M}|\boldsymbol{\eta})\mathrm{p}(\boldsymbol{\eta}|\boldsymbol{\alpha})\mathrm{d}\boldsymbol{\eta}$$

$$= \frac{\Gamma(M+1)}{\prod_{k=1}^{K}\Gamma(\hat{y}_k+1)}\frac{\prod_{k=1}^{K}\alpha_k^{\hat{y}_k}}{\alpha_0{}^{\sum_{k=1}^{K}\hat{y}_k}}. \tag{10}$$

The multinomial coefficient is independent of $\boldsymbol{\alpha}$, we thus verify that $\mathcal{L}^{\text{NLL}}$ in Eqn. (6) can be generalised to the distribution estimation framework by replacing one-hot majority label $\boldsymbol{y}$ with $\hat{\boldsymbol{y}}$:

$$\mathcal{L}^{\text{NLL}^*} = \sum_{k=1}^{K}\hat{y}_k(\log(\alpha_0) - \log(\alpha_k)). \tag{11}$$

The regulariser in Eqn. (7) is then replaced with:

$$\mathcal{L}^{\text{R1}} = \mathcal{KL}(\mathrm{Dir}(\boldsymbol{\eta}|\hat{\boldsymbol{\alpha}})||\,\mathrm{Dir}(\boldsymbol{\eta}|\mathbf{1})) \tag{12}$$

where $\hat{\boldsymbol{\alpha}} = \bar{\boldsymbol{y}} + (1-\bar{\boldsymbol{y}})\odot\boldsymbol{\alpha}$ and $\bar{\boldsymbol{y}} = \frac{1}{M}\sum_{m=1}^{M}\boldsymbol{y}_m$ is the soft label. An alternative regulariser is proposed in order to explicitly regularise the predicted multinomial distribution:

$$\mathcal{L}^{\text{R2}} = \mathcal{KL}(\bar{\boldsymbol{y}}||\mathbb{E}[\boldsymbol{\eta}]). \tag{13}$$

Hence, we have extend the EDL method described in Section 3.2 for classification to quantify the uncertainty in distribution estimation, with the original method (Sensoy et al., 2018) being a special case when $M = 1$ and $\hat{\boldsymbol{y}}$ becomes the one-hot majority label $\boldsymbol{y}$. In addition, it's worth noting that the proposed approach does not require a fixed number of annotators for every utterance and can easily generalise to a large number of annotators (*i.e.*, for crowd-sourced datasets). There has been previous work on emotion distribution estimation using another variant of Dirichlet prior network (DPN) (Wu et al., 2022). Note that although both EDL* and DPN involve a Dirichlet distribution, EDL* has a different problem formulation to DPN. The key differences between EDL* and DPN are explained in detail in Appendix H.

5 Evaluation metrics

The proposed method is evaluated in terms of majority prediction, uncertainty estimation, OOD detection, and distribution estimation.

Majority prediction. Majority prediction for MA utterances is evaluated by classification accuracy (ACC) and unweighted average recall (UAR) which is the sum of class-wise accuracy divided by the number of classes.

Uncertainty estimation. Model calibration is evaluated by expected calibration error (ECE) (Naeini et al., 2015) and maximum calibration error (MCE) (Naeini et al., 2015). ECE measures model calibration by computing the difference in expectation between confidence and accuracy. Predictions are partitioned into Q bins equally spaced in the [0,1] range and ECE can be computed as follows:

$$\mathrm{ECE} = \sum_{q=1}^{Q}\frac{|B_q|}{n}\left|\mathrm{Acc}(B_q) - \mathrm{Conf}(B_q)\right|. \tag{14}$$

MCE is a variation of ECE which measures the largest calibration gap:

$$\mathrm{MCE} = \max_{q\in\{1,\ldots,Q\}}\left|\mathrm{Acc}(B_q) - \mathrm{Conf}(B_q)\right|. \tag{15}$$

OOD detection. The area under the receiver operating characteristic (AUROC) and the area under the precision-recall curve (AUPRC) are used to evaluate the performance of OOD detection. The estimated uncertainty is used as a decision threshold for both AUROC and AUPRC. The baseline is 50% for AUROC and is the fraction of positives for AUPRC. NMA utterances are set as the positive class to detect.

Distribution estimation. Emotion distribution estimation performance is measured by the negative log-likelihood (NLL) of sampling human annotations from the predicted multinomial distribution.

6 Experimental setup

6.1 Baselines

The proposed methods were compared to the following baselines:

- MLE: a deterministic classification network with softmax activation trained by the cross-entropy loss between the majority vote label and model predictions;

- MCDP: a Monte-Carlo dropout (Gal and Ghahramani, 2016) model with a dropout rate of 0.5 which is forwarded 100 times to obtain 100 samples during testing;

- Ensemble: an ensemble (Lakshminarayanan et al., 2017) of 10 MLE models with the same structure trained by bagging;

- MLE+: a MLE model with NMA as an extra class.

An additional baseline for distribution estimation:

- MLE*: the "soft label" approach mentioned in Section 2 which is trained by minimising KL divergence between the soft label $\bar{y}$ and predictions. It is an extension of MLE from one-hot majority vote labels to soft labels.

The system described in Section 3.2 is denoted as **"EDL"**. **"EDL*(R1)"** and **"EDL*(R2)"** refers to the systems proposed in in Section 4 using regularisation terms defined in Eqn. (12) and Eqn. (13) respectively. Uncertainty estimation of EDL models are computed by Eqn. (3) while max probability is used as confidence measure for other methods.

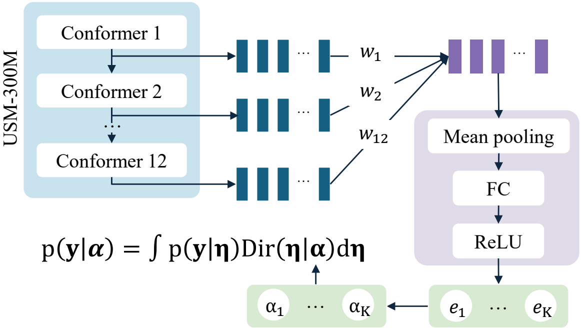

USM-300M
Conformer 1
Conformer 2
Conformer 12
w_1
w_2
w_{12}
Mean pooling
FC
ReLU
$\mathrm{p}(\mathbf{y}|\boldsymbol{\alpha}) = \int \mathrm{p}(\mathbf{y}|\boldsymbol{\eta})\mathrm{Dir}(\boldsymbol{\eta}|\boldsymbol{\alpha})\mathrm{d}\boldsymbol{\eta}$
α_1 ··· α_K
e_1 ··· e_K

Figure 2: Illustration of the model structure.

6.2 Datasets

Two publicly available datasets were used in the experiments. The IEMOCAP corpus (Busso et al., 2008) is one of the most widely used emotion datasets. It consists of 10,039 English utterances from 5 dyadic conversational sessions. Each utterance was evaluated by at least three human annotators. Only 16.1% of utterances have an all-annotators-agreed emotion label. The emotion distribution is represented using a five-dimensional categorical distribution, including happy (merged with excited), sad, neutral, angry, and others. The "others" category includes all emotions not covered in the previous four categories which is dominated by frustration (92%). 14.2% of the utterances don't have a majority agreed emotion class label.

The CREMA-D corpus (Cao et al., 2014) contains 7,442 English utterances from 91 actors. Actors spoke from a selection of 12 sentences using one of six different emotions (anger, disgust, fear, happy, neutral and sad). The dataset was annotated by crowd-sourcing. Ratings based on audio alone were used in this work. Utterances have 9.21 ratings on average. 5.1% of utterances have an all-annotators-agreed emotion label and 8.7% don't have a majority agreed emotion class label.

Both datasets were divided into an MA subset and an NMA subset. All methods were trained only on MA data except for MLE+ where 25% of NMA utterances were reserved for testing and the rest were included in training. For IEMOCAP, Session 5 was reserved for testing, and Sessions 1-4 were split into training and validation with a ratio of 4:1. For the CREMA-D dataset, the MA subset was split into train, validation, test in the ratio 70 : 15 : 15 following Ristea and Ionescu (2021).

	Classify MA				Detect NMA (all)		Detect NMA (test)	
	ACC ↑	UAR ↑	ECE ↓	MCE ↓	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑
MLE+	0.447	0.438	0.303	0.383	/	/	0.461	0.139
MLE	0.582	0.577	0.206	0.239	0.550	0.471	0.549	0.177
MCDP	0.584	0.572	0.128	0.184	0.566	0.491	0.568	0.203
Ensemble	0.593	0.595	0.439	0.594	0.567	0.491	0.563	0.192
EDL	**0.611**	**0.596**	**0.103**	**0.145**	**0.610**	**0.530**	**0.620**	**0.227**

Table 1: Results of quantifying uncertainty in emotion classification on the IEMOCAP dataset. The baseline for AUPRC is 0.433 for the entire NMA set and 0.160 for the NMA test subset. The best value in each column is indicated in bold, and the second-best value is underlined.

	Classify MA				Detect NMA (all)		Detect NMA (test)	
	ACC ↑	UAR ↑	ECE ↓	MCE ↓	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑
MLE+	0.568	0.540	0.216	0.476	/	/	0.552	0.156
MLE	0.714	0.672	0.150	0.156	0.578	0.467	0.571	0.179
MCDP	0.717	0.687	0.102	0.109	0.619	0.481	0.614	0.201
Ensemble	**0.731**	0.674	0.362	0.496	0.598	0.481	0.605	0.198
EDL	0.711	**0.714**	**0.057**	**0.080**	**0.645**	**0.506**	**0.657**	**0.234**

Table 2: Results of quantifying uncertainty in emotion classification on the CREMA-D dataset. The baseline for AUPRC is 0.387 for the entire NMA set and 0.097 for the NMA test subset.

6.3 Model structure

The backbone structure used in this paper is illustrated in Figure 2 which follows an upstream-downstream paradigm (Bommasani et al., 2021). The upstream model uses the universal speech model (USM) (Zhang et al., 2023) with 300M parameters which contains a CNN-based feature extractor and 12 Conformer (Gulati et al., 2020) encoder blocks of dimension 1024 with 8 attention heads. The structure of the downstream model follows SUPERB (Yang et al., 2021), a benchmark for evaluating pre-trained upstream models, which performs utterance-level mean-pooling followed by a fully-connected layer. The pre-trained upstream USM model is frozen. The downstream model computes the weighted sum of the hidden states extracted from each layer of the upstream model. The backbone structure has been shown to outperform state-of-the-art methods for emotion classification (see Table 5 in Appendix C). The implementation details for model training can be found in Appendix A.

7 Results

This section presents experimental results of the three approaches for handling ambiguous emotion: incorporating NMA as an extra class (Section 7.1), detecting NMA as OOD (Section 7.2), and representing emotion as distributions (Section 7.3). The average of three runs are reported for all results.

7.1 Including NMA as an additional category degrades the performance

The first approach, which trains an emotion classifier with NMA as an extra class, is denoted "MLE+" in Table 1 and 2. Some of the NMA utterances are included in MLE+ training while the remainder are used for testing. Therefore, OOD detection is evaluated only on NMA (test) data for MLE+. The results reveal that the addition of the NMA class has a detrimental impact on the classification performance of the original MA emotion classes. Comparing to MLE, MLE+ observes a ∼23% relative decrease in both ACC and UAR on IEMOCAP and a ∼20% relative decrease in ACC and UAR on CREMA-D. The confusion matrices of the MLE+ model can be found in Appendix D, which shows that NMA itself is challenging to predict and it also confuses the model when predicting classes such as neutral, sad, frustrated, and disgust.

7.2 Detecting NMA as OOD

The proposed EDL-based method is compared to baselines in Tables 1 and 2 on the IEMOCAP and CREMA-D datasets respectively. First, as shown by the values of ACC and UAR, the proposed method demonstrates comparable classification performance to the baselines, suggesting that the extension for uncertainty estimation does not undermine the model's capabilities. Although the Ensemble achieves the highest accuracy on CREMA-D, it involves training 10 individual systems. The proposed method achieves overall the best classification performance with only a tenth of the computational cost of Ensemble during both training and testing. In addition, the proposed method offers superior model calibration, as shown by the lowest values of ECE and MCE. It also outperforms the baselines in effectively identifying NMA as OOD samples, as shown by the highest AUROC and AUPRC values.

Figure 3 shows the change of accuracy when samples with uncertainty larger than a threshold are excluded. The model tends to provide less accurate predictions when it is less confident about its prediction, shown by the decrease of classification accuracy when the uncertainty threshold increases, which demonstrates the effectiveness of uncertainty prediction.

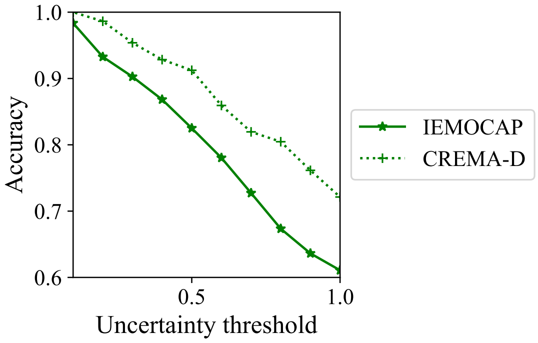


Figure 3: The change of accuracy with respect to the uncertainty threshold for EDL-based methods on IEMOCAP and CREMA-D.

7.3 Estimating emotion distribution

The proposed EDL* methods were first evaluated in terms of majority class prediction. The results of distribution-based methods on classification of MA data are shown in Table 3. Compared to the classification-based methods in Table 1 and Table 2, it can be seen that EDL* does not reduce the performance of emotion classification (in terms of ACC and UAR) and model calibration (in terms of ECE and NCE) on MA data. This indicates that the information of the majority class is retained when representing emotion as a distribution. Note that when representing emotion as a distribution, it is no longer appropriate to consider NMA utterances as OOD samples, as illustrated by case (b) and (c) in Figure 1. Although still trained only on MA data, the proposed distribution-based system shows good generalisation ability in predicting the emotion distribution of NMA data, which we will see shortly. When encountering NMA data during testing, instead of simply returning "I don't know", the proposed system can provide reliable estimation of its emotional content, which is a key benefit.

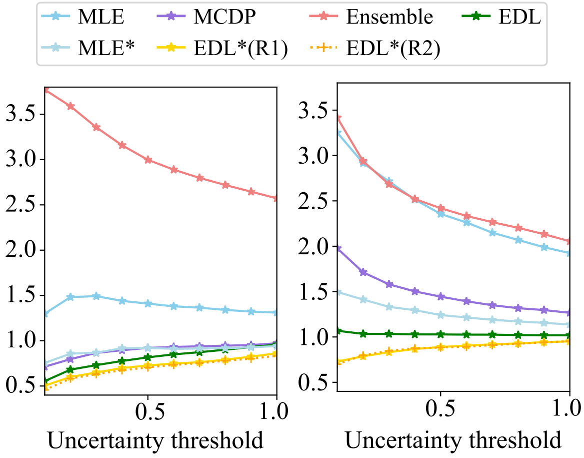
IEMOCAP	ACC	UAR	ECE	MCE
MLE*	0.564	0.562	0.151	0.279
EDL*(R1)	0.623	0.612	0.081	0.208
EDL*(R2)	**0.624**	**0.616**	**0.025**	**0.201**
CREMA-D	**ACC**	**UAR**	**ECE**	**MCE**
MLE*	0.693	0.621	0.109	0.115
EDL*(R1)	**0.740**	0.694	**0.029**	**0.095**
EDL*(R2)	0.718	**0.722**	0.084	0.107

Table 3: Classification and calibration performance of distribution-based methods on MA data. The best value in each column is indicated in bold.

NLL	IEMOCAP		CREMA-D	
	MA	NMA	MA	NMA
MLE	1.310	1.924	1.532	2.054
MCDP	0.972	1.266	0.965	1.292
Ensemble	2.572	2.055	2.285	2.089
EDL	0.958	1.019	0.757	1.021
MLE*	0.941	1.137	0.648	0.774
EDL*(R1)	0.861	**0.951**	0.614	0.722
EDL*(R2)	**0.833**	0.953	**0.606**	**0.698**

Table 4: Distribution estimation results. NMA stands for NMA(all).

The proposed EDL* methods were then evaluated regarding distribution estimation. Table 4 compares EDL* to the baselines in terms of the negative log likelihood of sampling target labels from the predicted emotion distribution. As can be seen from the table, EDL* produce improved distribution estimation, achieving the smallest NLL values on both MA and NMA data. Among the two EDL* methods employing different regularisation terms, EDL* with R2 (defined in Eqn. (13)), which directly applies regularisation to the predicted distribution, exhibits better distribution estimation with-



(a) MA data (b) NMA data

Figure 4: Reject option for NLL on IEMOCAP. Trends on CREMA-D are similar, shown in Appendix G.

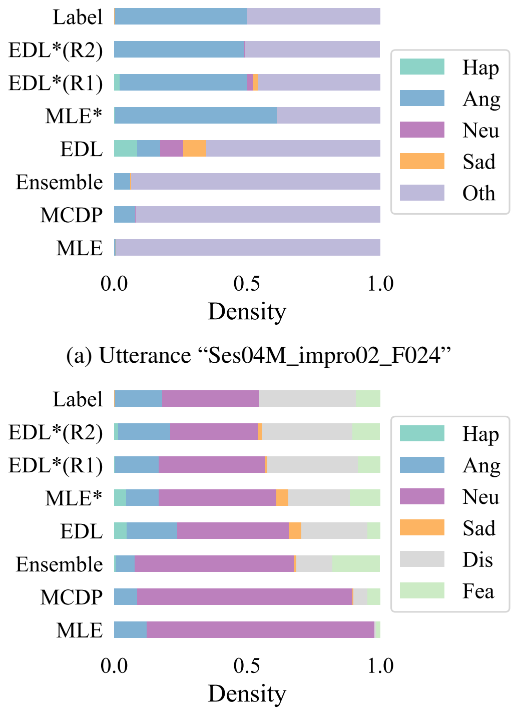
out sacrificing model calibration.

A reject option was then evaluated for NLL (instead of accuracy) to examine model calibration. For a well-calibrated model, an increase in the NLL value, which is associated with poorer distribution estimation, is expected when the model becomes less confident. Figure 4 visualises the change of NLL for MA data and NMA data when uncertainty increases. For MA data, *i.e.* the type of data that has been seen by the models during training, most methods can successfully reject uncertain samples except for MLE and Ensemble, as shown by an increase in NLL values when the uncertainty threshold increases. However, for NMA data which the model hasn't seen in training, only the EDL* methods exhibit the ability to demonstrate an increasing trend in NLL values.

The proficiency of the proposed EDL* methods for estimating the emotion distribution and providing reliable confidence predictions, demonstrates the method's capacity to estimate both aleatoric uncertainty (Matthies, 2007; Der Kiureghian and Ditlevsen, 2009; Hu et al., 2024), arising from data complexity (*i.e.*, the ambiguity of emotion expression), and epistemic uncertainty, corresponding to the amount of uncommitted belief in subjective logic.

7.4 Case study

Emotion distributions estimated by different methods are visualised against the label distributions for two representative examples in Figure 5. In general, distribution-based methods show superior performance for distribution estimation than



(a) Utterance “Ses04M_impro02_F024”

(b) Utterance “1084_TSI_ANG_XX”

Figure 5: Visualisation of emotion distribution for case study. Utterance (a) is selected from IEMOCAP. Utterance (b) is selected from CREMA-D.

classification-based methods. In the case of utterance (a) which receives two “angry” labels and two “frustrated” labels, the proposed EDL* methods stands out by effectively capturing the tie between the emotions, whereas the predictions of classification-based methods tend to be predominantly skewed towards “frustrated”. As for utterance (b), where both “disgust” and “neutral” receive four votes, along with two votes for “angry” and one for “fear”, the emotion distributions predicted by the EDL* methods also show a similar pattern. These examples show that the proposed method can not only provide a more comprehensive emotion representation but also better reflect the variability of human opinions. Additional examples can be found in Appendix J and Appendix K.

8 Conclusions

This paper re-examines the emotion classification problem, starting with an exploration of ways to handle data with ambiguous emotions. It is first shown that incorporating ambiguous emotions as an extra class reduces the classification performance of the original emotion classes. Then, evidence theory is adopted to quantify uncertainty in emotion classification which allows the classifier to output “I don’t know” when it encounters utterances with ambiguous emotion. The model is trained to predict the hyperparameters of a Dirichlet distribution, which models the second-order probability of the probability assignment over emotion classes. Furthermore, to capture finer-grained emotion differences, we transform the emotion classification problem into an emotion distribution estimation problem. All annotations are taken into account rather than only the majority opinion. A novel approach is proposed which extends standard EDL to quantify uncertainty in emotion distribution estimation. Experimental results show that given an utterance with ambiguous emotion the proposed approach is able to provide a comprehensive representation of its emotion content as a distribution with a reliable uncertainty measure.

Ethics Statement

In this work, all human annotations used for training were taken from existing publicly available corpora. No new human annotations were collected.

In subjective tasks like emotion recognition, there is usually no single “correct” answer. The conventional approach of imposing a single ground truth through majority voting may overlook valuable nuances within each annotator’s evaluation and the disagreements between them, potentially resulting in the under-representation of minority views. This study, instead of exclusively relying on the majority vote, integrates emotion annotations from all annotators for each utterance during model training. It is hoped that this work could contribute to a more inclusive representation of human opinions.

Limitations

The proposed approach requires the raw labels from different human annotators for each sentence to be provided by the datasets. Although validated only for emotion recognition, the proposed method could also be applied to other tasks with disagreements in subjective annotations, which will be investigated in future work.

Different people perceive emotion differently and hence the motivation of the paper is to handle such ambiguity. It may be that if annotators were to also provide confidence ratings during annotation, which is not the case in the emotion datasets we have used, then this information could be used to weight the observations when estimating the emotion distribution.

Acknowledgements

Work performed while W.W. was an intern at Google. W.W.’s PhD research is supported by the Cambridge Trust.

References

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *Proc. ICML*, Baltimore, USA.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proc. NeurIPS*, Vancouver, Canada.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. 2014. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Georgios Chochlakis, Gireesh Mahajan, Sabyasachee Baruah, Keith Burghardt, Kristina Lerman, and Shrikanth Narayanan. 2023. Leveraging label correlations in a multi-label setting: A case study in emotion. In *Proc. ICASSP*, Rhodes, Greece.

Alan S Cowen and Dacher Keltner. 2017. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the national academy of sciences*, 114(38):E7900–E7909.

Arthur P Dempster. 1968. A generalization of Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):205–232.

Armen Der Kiureghian and Ove Ditlevsen. 2009. Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105–112.

H.M. Fayek, M. Lech, and L. Cavedon. 2016. Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels. In *Proc. IJCNN*, Vancouver.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proc. ICML*, New York, USA.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for speech recognition. In *Proc. Interspeech*, Shanghai, China.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *Proc. ICML*, Sydney, Australia.

Jing Han, Zixing Zhang, Maximilian Schmitt, Maja Pantic, and Björn Schuller. 2017. From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty. In *Proc. ACM MM*, Mountain View, USA.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Yuchen Hu, Chen Chen, Chao-Han Huck Yang, Ruizhe Li, Chao Zhang, Pin-Yu Chen, and EnSiong Chng. 2024. Large language models are efficient learners of noise-robust speech recognition. *arXiv preprint arXiv:2401.10446*.

Audun Jsang. 2018. *Subjective Logic: A formalism for reasoning under uncertainty*. Springer Publishing Company, Incorporated.

Y. Kim, H. Lee, and E. M. Provost. 2013. Deep learning for robust feature generation in audiovisual emotion recognition. In *Proc. ICASSP*, Vancouver, Canada.

Yelin Kim and Jeesun Kim. 2018. Human-like emotion recognition: Multi-label learning from noisy labeled audio-visual expressive speech. In *Proc. ICASSP*, Calgary, Canada.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proc. NeurIPS*, Long Beach, USA.

Hermann G Matthies. 2007. Quantifying uncertainty: Modern computational representation of probability and applications. In *Extreme Man-made and Natural Hazards in Dynamics of Structures*, pages 105–135. Springer.

Emily Mower, Maja J Matarić, and Shrikanth Narayanan. 2010. A framework for automatic human emotion classification using emotion profiles. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5):1057–1070.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proc. AAAI*, Austin, USA.

S. Poria, E. Cambria, R. Bajpai, and A. Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125.

Nicolae-Cătălin Ristea and Radu Tudor Ionescu. 2021. Self-paced ensemble learning for speech and audio classification. In *Proc. Interspeech*, Brno, Czech.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential deep learning to quantify classification uncertainty. In *Proc. NeurIPS*, Montréal, Canada.

Vidhyasaharan Sethu, Emily Mower Provost, Julien Epps, Carlos Busso, Nicholas Cummins, and Shrikanth S. Narayanan. 2019. The ambiguous world of emotion representation. *arXiv preprint arXiv:1909.00360*.

Wen Wu, Chao Zhang, and Philip C. Woodland. 2021. Emotion recognition by fusing time synchronous and time asynchronous representations. In *Proc. ICASSP*, Toronto, Canada.

Wen Wu, Chao Zhang, Xixin Wu, and Philip C Woodland. 2022. Estimating the uncertainty in emotion class labels with utterance-specific Dirichlet priors. *IEEE Transactions on Affective Computing*.

Shu-Wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. In *Proc. Interspeech*, Brno, Czech.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proc. ACL*, Melbourne, Australia.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al. 2023. Google USM: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*.

Heqing Zou, Yuke Si, Chen Chen, Deepu Rajan, and Eng Siong Chng. 2022. Speech emotion recognition with co-attention based multi-level acoustic information. In *Proc. ICASSP*, Singapore.

A Implementation details

This section describes the implementation details. The model is implemented using Pax[1]. The batch size is set to 256, The coefficient λ is set to 0.8 for IEMOCAP and 0.2 for CREMA-D. The Adafactor optimiser and Noam learning rate scheduler are used with 200 warm up steps and a peak learning rate of 8.84×10^{-4}. Since the CREMA-D dataset is extremely imbalanced (*i.e.*, neutral accounts for over 50%), a balanced sampler is applied during training. The model is trained for 20k steps which takes $\sim$ 5 hours on 8 TPU v4s.

B Discussion of statistical significance

The results reported in Section 7 are the average of three runs with different seeds. Regarding ECE, MCE, AUROC and AUPRC, the improvements in the results in bold are consistent across all three runs. For ACC and NLL, the differences are statistically significant with $p < 0.05$.

C Comparing the backbone structure to SOTA models

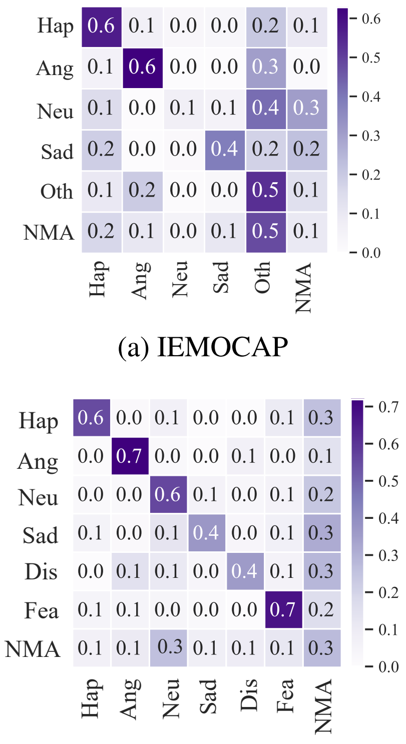
The USM-based backbone structure is evaluated following the setup of the emotion recognition task of the SUPERB benchmark (Yang et al., 2021): four-way emotion classification (happy, sad, angry, neutral) on IEMOCAP dataset with leave-one-session-out five-fold cross validation. The USM-300M model is compared to multiple state-of-the-art models of similar size. Results are shown in Table 5. Except for the USM-300M model used in the paper, all other results are quoted from the cited papers. As shown in the table, the USM-based backbone structure outperforms other state-of-the-art methods[2] and yields the highest accuracy.

D Confusion matrices of MLE+

As described in Section 7.1, including NMA as an extra class reduces the classification performance. This section analyses the confusion matrices of the MLE+ model, shown in Figure 6. It can be seen from the bottom right entry that NMA itself is challenging to predict, possibly because it essentially contains a mix of different emotion content. The last column demonstrates that grouping these utterances into one class can confuse the model, particularly for the classes neutral, sad, frustrated, and disgust.

Model	# Param	ACC (%)
Wav2vec 2.0 large (Baevski et al., 2020)	317M	65.64
Data2vec large (Baevski et al., 2022)	314M	66.31
HuBERT large (Hsu et al., 2021)	317M	67.62
WavLM large (Chen et al., 2022)	317M	70.62
USM-300M (Zhang et al., 2023)	290M	**71.06**

Table 5: Four-way classification results IEMOCAP following the SUPERB-ER benchmark setup.



(a) IEMOCAP

(b) CREMA-D

Figure 6: Confusion matrix of the MLE+ system on IEMOCAP and CREMA-D.

E Alternative activation functions

As described in Section 3.2, ReLU is used as the output activation function in EDL to make sure the evidence is non-negative. This section compares the use of different activation functions including ReLU, softplus and exponential functions. The three activation functions are plotted in Figure 7.

[1] https://github.com/google/paxml

[2] https://superbbenchmark.org/leaderboard

IEMOCAP	Classify MA ACC	 UAR	 ECE	 MCE	Detect NMA (all) AUROC	 AUPRC	Detect NMA (test) AUROC	 AUPRC
EDL (ReLU)	**0.611**	0.596	0.103	**0.145**	0.610	0.530	0.620	0.227
EDL (Softplus)	0.608	0.574	**0.035**	0.173	**0.617**	**0.534**	**0.639**	**0.251**
EDL (Exponential)	0.588	**0.601**	0.167	0.230	0.593	0.502	0.619	0.225

CREMA-D	Classify MA ACC	 UAR	 ECE	 MCE	Detect NMA (all) AUROC	 AUPRC	Detect NMA (test) AUROC	 AUPRC
EDL (ReLU)	0.701	**0.714**	**0.057**	**0.080**	**0.645**	**0.506**	**0.657**	**0.234**
EDL (Softplus)	0.692	0.696	0.113	0.309	0.640	**0.506**	0.633	0.230
EDL (Exponential)	**0.723**	0.602	0.277	0.277	0.623	0.495	0.626	0.197

Table 6: Comparison of EDL methods with different activation functions on IEMOCAP and CREMA-D.

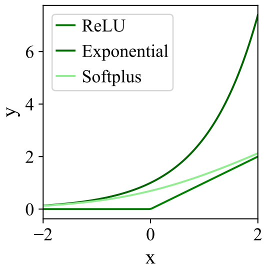
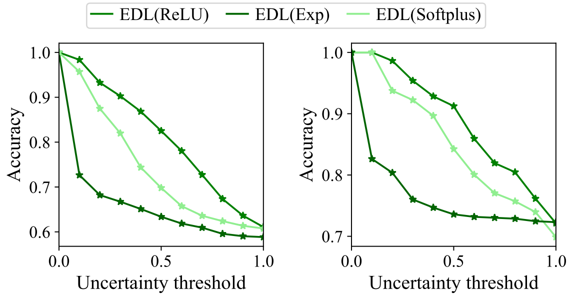


Figure 7: Illustration of the activation functions.

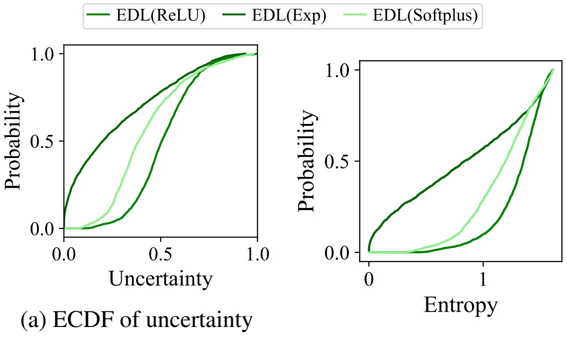
As shown in Table 6, using exponential function tends to result in less effective model calibration, shown by the largest ECE and MCE values. It also produces worse performance for NMA detection, shown by the smallest AUROC and AUPRC. Figure 8 shows the reject option for accuracy of EDL with different activation functions. A drop in accuracy when the uncertainty threshold increases from 0 to 0.1 is observed for model using exponential activation. This indicates that exponential activation tends to lead to smaller uncertainty.



(a) IEMOCAP (b) CREMA-D

Figure 8: Reject option for accuracy for EDL methods with different activation functions.

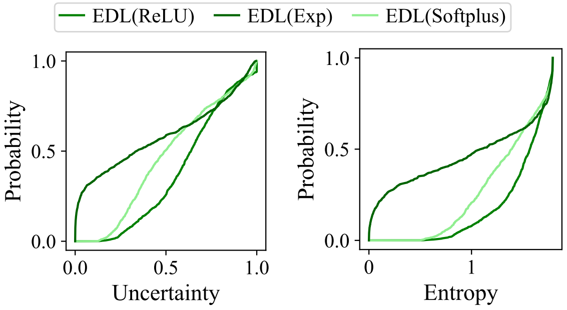
The empirical cumulative distribution function (ECDF) of uncertainty and entropy on IEMOCAP and CREMA-D are plotted in Figure 9 and Figure 10 respectively. It can be seen that exponential activation leads to smaller uncertainty and entropy, which echos the statement in Section 3.1 that exponential activation tends to inflate the probability of the correct class.



(a) ECDF of uncertainty

(b) ECDF of entropy

Figure 9: Empirical CDF of uncertainty (left) and entropy (right) on IEMOCAP for EDL method with different activation functions.

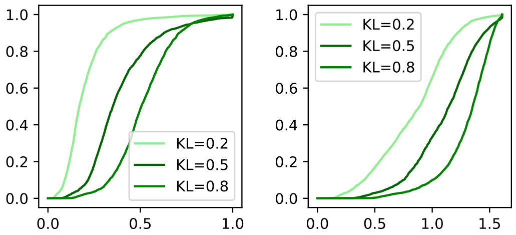


(a) ECDF of uncertainty (b) ECDF of entropy

Figure 10: Empirical CDF of uncertainty (left) and entropy (right) on CREMA-D for EDL method with different activation functions.

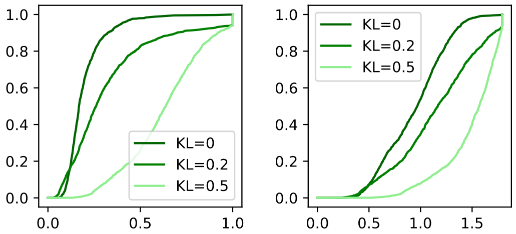
F Analysis of regularisation coefficient

This section analyses the effect of regularisation coefficient λ in Eqn. (7) for EDL. The empirical CDF of uncertainty and entropy when different regularisation coefficient was used is plotted in in Figure 11 and Figure 12 for IEMOCAP and CREMA-D respectively. We observed that larger lambda values lead to a larger entropy and uncertainty. This aligns with the definition of the regularisation term in Eqn. (7) which tends to enforce a flat prior with small evidence.



(a) ECDF of uncertainty (b) ECDF of entropy

Figure 11: Empirical CDF of uncertainty (left) and entropy (right) on IEMOCAP for EDL method with different regularisation coefficient λ.

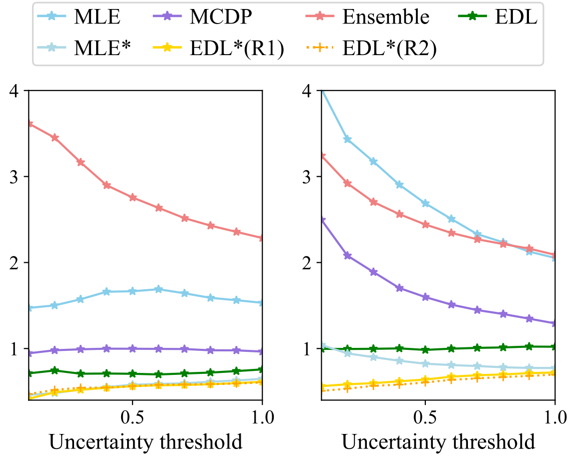


(a) ECDF of uncertainty (b) ECDF of entropy

Figure 12: Empirical CDF of uncertainty (left) and entropy (right) on CREMA-D for EDL method with different regularisation coefficient λ.

G Reject option for NLL on CREMA-D

This section shows the reject option for NLL on CREMA-D dataset. The change of NLL for MA data and NMA data when the uncertainty threshold increases are shown in Figure 13. For a well-calibrated model, an increase in the NLL value is expected when the model becomes less confident. Similar to the findings in Figure 4, most methods are effective for rejecting uncertain samples in the MA data, as shown by an increase in NLL values when the uncertainty threshold increases. However, only the EDL* methods are successful for NMA data.



(a) MA data (b) NMA data

Figure 13: Reject option for NLL on NMA data of CREMA-D.

H Differences between the proposed EDL* method and Dirichlet prior networks

The proposed EDL* method is similar to the Dirichlet prior network (DPN) used in Wu et al. (2022) but essentially different. Both methods involve a Dirichlet distribution, but the difference lies in the problem formulation and the connection of the Dirichlet distribution to evidence theory.

(i) **Problem formulation.** In DPN, individual emotion class labels provided by human annotators are treated as one-hot categorical distributions ($\{\boldsymbol{\eta}_m\}_{m=1}^M$) and the model is trained by maximising the likelihood of sampling those one-hot categorical distributions given the Dirichlet prior.

$$\mathcal{L}_{\text{DPN}}^{\text{NLL}} = \log \text{p}(\{\boldsymbol{\eta}_m\}_{m=1}^M|\boldsymbol{\alpha}) \tag{16}$$

However, the proposed EDL* method preserves the target emotion labels as discrete class labels ($\{\boldsymbol{y}_m\}_{m=1}^M$) which are drawn from an unknown categorical likelihood ($\boldsymbol{\eta}$). The model is then trained by maximising the likelihood of sampling discrete labels given the Dirichlet prior by marginalising out all possible categorical distributions.

$$\begin{aligned}\mathcal{L}_{\text{EDL*}}^{\text{NLL}} &= \log \text{P}(\{\boldsymbol{y}_m\}_{m=1}^M|\boldsymbol{\alpha}) \\ &= \log \int \text{P}(\{\boldsymbol{y}_m\}_{m=1}^M|\boldsymbol{\eta})\text{p}(\boldsymbol{\eta}|\boldsymbol{\alpha})\text{d}\boldsymbol{\eta}\end{aligned} \tag{17}$$

Therefore, the problem setting and NLL loss for two methods are essentially different. A drawback of DPN is that it greatly reduces the classification accuracy (as also pointed out by Wu et al. (2022)), while the proposed EDL* method improves over

IEMOCAP	ACC	UAR	ECE	MCE
DPN	0.545	0.477	0.214	0.285
EDL*(R1)	0.623	0.612	0.081	0.208
EDL*(R2)	**0.624**	**0.616**	**0.025**	**0.201**
CREMA-D	**ACC**	**UAR**	**ECE**	**MCE**
DPN	0.702	0.637	0.489	0.645
EDL*(R1)	**0.740**	0.694	**0.029**	**0.095**
EDL*(R2)	0.718	**0.722**	0.084	0.107

Table 7: Comparison of DPN and EDL* in terms of classification and calibration performance.

NLL	IEMOCAP		CREMA-D	
	MA	NMA	MA	NMA
DPN	1.081	0.954	0.958	0.845
EDL*(R1)	0.861	**0.951**	0.614	0.722
EDL*(R2)	**0.833**	0.953	**0.606**	**0.698**

Table 8: Comparison of DPN and EDL* in terms of distribution estimation results. NMA stands for NMA(all).

DPN by preserving or even boosting the classification performance.

(ii) **Connection to evidence theory.** By connecting to evidence theory, the proposed method provides an explicit uncertainty score (Eqn. (3)) while DPN doesn't. Therefore, DPN can only estimates aleatoric uncertainty (*i.e.*, data complexity) but it cannot predict epistemic uncertainty (*i.e.*, model confidence). In contrast, the proposed EDL* method provides superior performance in estimating both aleatoric and epistemic uncertainty, as explained in the last paragraph of Section 7.3.

The performance of DPN and EDL* are compared in the Table 7 and Table 8 (results taken for EDL* from Table 3 and Table 4 with the addition of DPN results). It can be seen that the proposed EDL* methods outperform DPN in all metrics on both datasets.

I Analysis of Instances When OOD Detection Fails

This section includes examples and analysis of particular utterances where OOD detection is problematic and compares these examples across the techniques discussed. It is shown that distribution-based methods improve over the classification-based systems in handling complex ambiguous emotions.

Consider the following false negative case where the OOD detection model fails to detect an NMA sample. Utterance "Ses04M_impro02_F024" from the IEMOCAP dataset has two "angry" labels and two "frustrated" labels as shown in Figure 14. The EDL system predicts this utterance as "frustrated" with a belief mass of 0.567 and an overall uncertainty score of 0.433, which reveals that the system fails to detect the utterance as NMA.

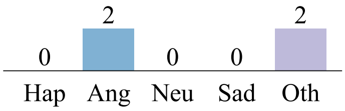


Figure 14: Human annotations for NMA utterance "Ses04M_impro02_F024".

A possible cause of this failure is that the model gets confused by MA utterances seen in the training that convey similar emotional content, such as "Ses05M_impro01_M014" whose annotations are shown in Figure 15 with an MA emotion class "frustrated". Although one annotator considered it as "angry", the MA ground-truth target was "frustrated" in a classification-based system.

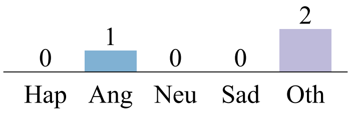


Figure 15: Human annotations for MA utterance "Ses05M_impro01_M014".

Both utterances occur within a dyadic situation where two people disagree, with the speaker being the one who compromises, feeling unhappy and frustrated. Such similar emotional content may confuse a classification-based system to also predict the NMA utterance as frustrated. It is worth noting that data with the same distribution as "Ses04M_impro02_F024", which has tied votes, is not included during the training of a classification-based model because there is no majority vote available to serve as ground truth.

This complex emotional expression can be better described by the distribution-based EDL* systems. The predicted distribution of the MA utterance is shown in Figure 16 and the predicted distribution of the NMA utterance can be found in Figure 5(a). It can be seen that the classification-based methods produce a similar distribution for the two utterances, with "frustrated" being dominant. However, the proposed EDL* methods can better match the label distribution and distinguish between these two cases. Although not been trained on NMA data, the EDL* methods are still capable of providing accurate predictions of its emotional content. This is a key benefit of the distribution-based approaches.

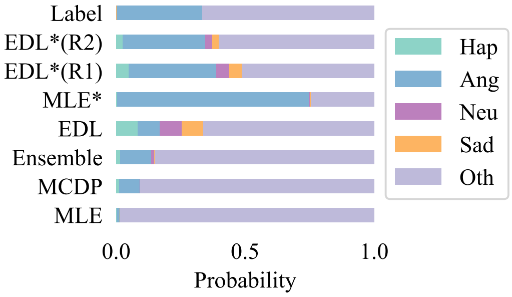


Figure 16: Predicted emotion distribution of MA utterance "Ses05M_impro01_M014".

Next, we provide a typical false positive instance where an MA utterance is mis-detected as OOD. The MA utterance "1087_IEO_FEA_LO" from the CREMA-D dataset has four "neutral", two "sad", and three "fear" human labels as in Figure 17. The NMA utterance "1052_ITH_FEA_XX" has four "neutral", two "sad", and four "fear" human labels as in Figure 18. The OOD system successfully predicts the NMA utterance as OOD with an overall uncertainty of 0.691 while also predicting the MA utterance as an OOD sample with an overall uncertainty of 0.623.[3] This failure is possible because the MA utterance "1087_IEO_FEA_LO" contains a complex mixture of emotions shown by the rather flat label distribution similar to "1052_ITH_FEA_XX", which confuses the OOD detection system. Note that the MA class "neutral" in Figure 17 comprises only $\frac{4}{4+2+3} \times 100\% = 44.4\%$ of the annotations and hence is not an absolute majority, which reduces the severity of this detection error.

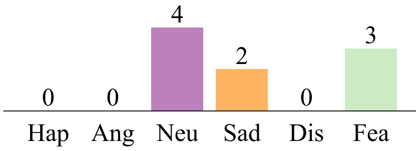


Figure 17: Human annotations for MA utterance "1087_IEO_FEA_LO".

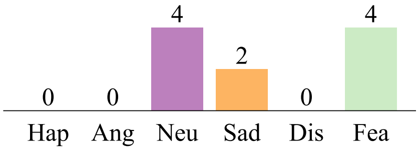


Figure 18: Human annotations for NMA utterance "1052_ITH_FEA_XX".

Again, the distribution-based EDL* methods show superior capability in handling such complex cases. The predicted distribution of the MA utterance is shown in Figure 19 and the predicted distribution of the NMA utterance can be found in Figure 21(b). For the MA utterance, although human opinions diverge, the classification-based methods only capture the majority prediction, with the predicted distribution being dominated by "neutral". However, the emotion distribution predicted by the proposed EDL* methods retains the probability for "sad" and "fear" which accounts for the minority human opinions. Therefore, we show that the proposed EDL* method improves over the OOD system by providing a more comprehensive representation of emotional content as well as a more inclusive representation of human opinions.

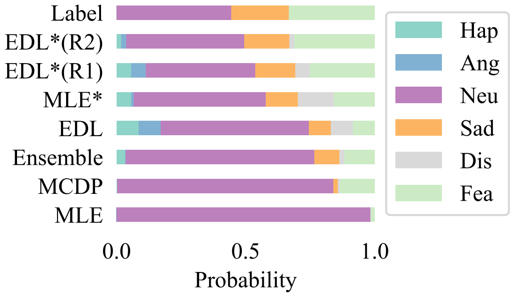
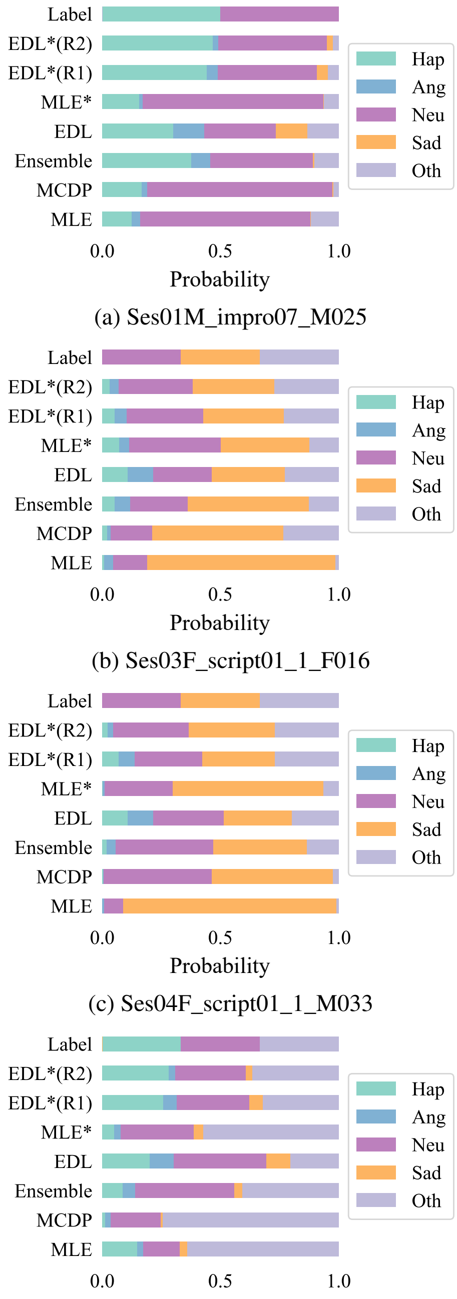


Figure 19: Predicted emotion distribution of MA utterance "1087_IEO_FEA_LO".

[3] Assume the OOD detection threshold is taken as 0.5.

J Further visualised examples: IEMOCAP

This section shows more examples on IEMOCAP. Aligning with the findings in Section 7.4, EDL* methods show better estimation of emotion distribution.



(a) Ses01M_impro07_M025

(b) Ses03F_script01_1_F016

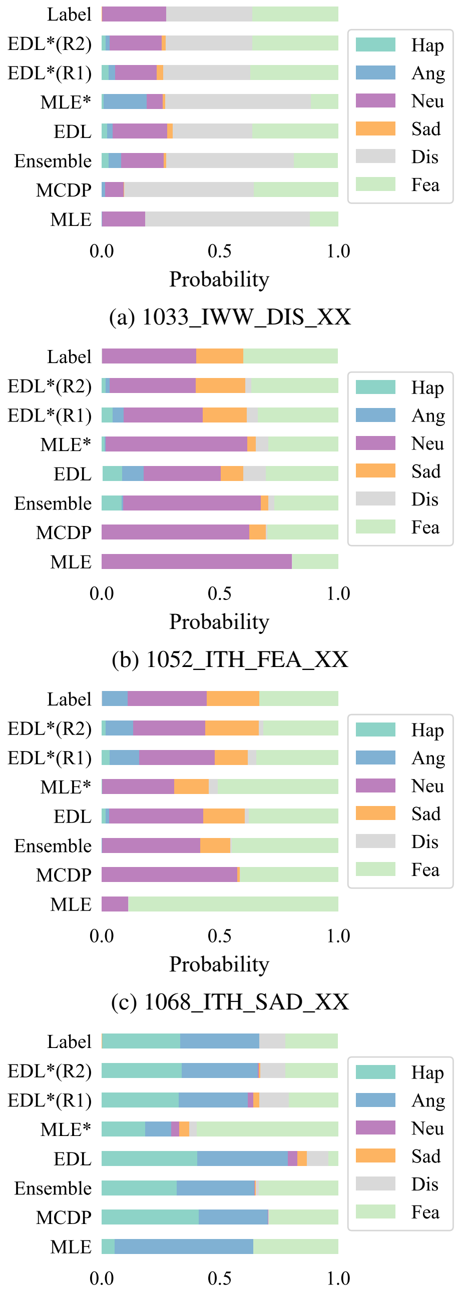
(c) Ses04F_script01_1_M033

(d) Ses04M_script01_3_M013

Figure 20: Case study on IEMOCAP.

K Further visualised examples: CREMA-D

This section shows more examples of CREMA-D. As can be seen, EDL* methods can better approximate the distribution of emotional content of an utterance.



(a) 1033_IWW_DIS_XX

(b) 1052_ITH_FEA_XX

(c) 1068_ITH_SAD_XX

(d) 1009_IWL_FEA_XX

Figure 21: Case study on CREMA-D.