EM-SEC: Efficient Multi-head Set-valued Evidential Classification

Grigor Bezirganyan[1] (✉), Sana Sellami[1], Laure Berti-Équille[2], and Sébastien Fournier[1]

[1] Aix-Marseille University, CNRS, LIS, Marseille, France
{grigor.bezirganyan,sana.sellami,sebasiten.fournier}@univ-amu.fr
[2] IRD, ESPACE-DEV, Montpellier, France laure.berti@ird.fr

Abstract. In machine learning and deep learning, uncertainty quantification helps to accurately assess a model's confidence in its predictions, enabling the rejection of uncertain outcomes in safety-critical applications. However, in scenarios involving AI-assisted decision-making, proposing multiple plausible decisions can be more beneficial than either not making any decisions or risking incorrect ones. Set-valued classification is a relaxation of standard multiclass classification where, in cases of uncertainty, the classifier returns a set of potential labels instead of a single label. Current methods for set-valued classification often suffer from high computational complexity or fail to adequately quantify uncertainty. In this paper, we introduce a novel, computationally efficient approach to set-valued classification leveraging evidential deep learning and subjective logic, explicitly providing a measure of classification uncertainty. Our method employs a dual-head architecture: one head conducts multiclass evidential classification, while the other suggests candidate label sets when uncertainty is high. The proposed approach has linear worst-case computational complexity with respect to the number of classes. Extensive evaluation on several benchmark datasets demonstrates that our method showcases comparable performance to baseline set-valued methods, while being up to 23 times faster at inference on the benchmark datasets.

Keywords: set-valued classification · evidential deep learning · subjective logic · utility maximization · uncertainty quantification.

1 Introduction

Artificial Intelligence (AI) has experienced a rapid surge in recent years, driven by advancements in deep learning, large-scale data availability, and increased computational power. From healthcare diagnostics to financial forecasting and autonomous systems, AI is increasingly used for both autonomous and assisted decision-making. Nevertheless, it has been shown that deep neural networks can produce incorrect output with high confidence [11], which can be catastrophic in safety-critical areas. To alleviate this issue, different uncertainty quantification

(UQ) techniques have been suggested [1] to understand the true confidence of the models and reject the decision of the model in case of high uncertainty.

Most of the UQ techniques for classification tasks operate in *precise classification* setting[3], where the model can either make a single prediction, or refuse to make a prediction under high uncertainty. However, in many applications of deep learning models for assisted decision-making, where a human expert makes the final decision, it would be more beneficial to suggest a reduced set of decision options with lower uncertainty rather than completely rejecting uncertain predictions. *Set-valued* or *imprecise* classification [5] tries to address this issue by suggesting a set of possible outputs in uncertain situations. [7] has recently demonstrated through trials that set-valued predictions enhance human decision-making and increase accuracy compared to no assistance or top-k predictions (suggesting the most probable k options).

Recent work [24,29,15] concentrates on using Dempster-Shafer (DS) theory of belief functions [8,22] to model uncertainty and imprecision in deep learning models. DS theory is a mathematical framework for reasoning under uncertainty, generalizing probability theory by allowing belief assignment to sets of possibilities rather than single events. However, considering all possible subsets of the power set of the label space can be computationally expensive and infeasible for large label spaces [19]. [19] proposes efficient set-valued classification algorithm, with $K \log K$ time complexity, where K is the number of classes. However, the method takes conditional class probabilities as a measure of uncertainty, which may not be accurate in practice [11].

In this work, we propose a novel set-valued evidential classification method, EM-SEC, which uses a multi-head architecture to efficiently model set-valued predictions. The first multiclass head is used to perform evidential multiclass classification, where it uses subjective logic (SL) [14] to get beliefs for each class as well as the total uncertainty. The second head is used to suggest one set-valued candidate set, which is obtained by evidential multi-label classification. Finally, subjective logic is used to allocate some belief mass from the uncertainty mass of the first head to the candidate set. At inference time, the model takes the prediction with highest belief mass, which can be either a singleton class or a set of classes. Our contributions hence are the following:

- We propose EM-SEC (**E**fficient **M**ulti-head **S**et-valued **E**vidential **C**lassification), a novel set-valued classification approach that scales linearly with the number of classes, hence being up to 23 times faster in our experiments compared with baseline methods.
- EM-SEC provides the uncertainty of the decision for both: single class and set-valued predictions. The uncertainty value can be used to avoid making uncertain decisions in safety-critical areas.
- We show that EM-SEC achieves comparable results with the baseline models on CIFAR-10 and CIFAR-100 datasets [17].

[3] *Precise* classification refers to returning only one class, in contrast to imprecise classification, which returns multiple classes, often referred to as *set-valued or imprecise* classification

– The code of EM-SEC is open and experiments are reproducible at https://github.com/bezirganyan/em_sec.

2 Related Work

Set-valued classification aims to minimize both the error rate and the expected size of the prediction set. A common approach is Top-K classification, which outputs a fixed number of labels per sample. However, choosing an optimal K is difficult, as different samples require different set sizes: high-confidence samples need fewer labels, while ambiguous ones benefit from larger sets.

Another approach is to use thresholding on the output probabilities of a multiclass classifier, where the classes with probabilities above a certain threshold are selected. However, a fixed threshold may not be optimal for all samples, and the set size can vary considerable by slight changes in the threshold [19]. Similar to EM-SEC, [10] propose a two-headed architecture, where the second (multi-label) head is trained on pseudo-labels derived from the softmax scores of the first head. These pseudo-positives are selected to maintain a batch-wise average set size of K. At test time, only the output of the multi-label head is used. However, this approach does not account for model uncertainty. In contrast, EM-SEC performs per-sample evidential fusion: if the Dirichlet evidence from the single-label head is sufficiently confident, it outputs a single class; otherwise, it defers to the evidential multi-label head's candidate set, while also providing a measure of uncertainty.

Conformal Prediction (CP) [26,23] is another popular framework for set-valued classification. CP is a distribution-free framework that provides a prediction set (or region in the case of regression) that is guaranteed to contain the true label with a certain probability. The theoretical guarantees and the simplicity of the framework make it a popular choice for set-valued classification. However, while these guarantees hold on the error rate, the expected set size is not guaranteed to be minimized, and the set size can be quite large for some datasets [19,27]. Conformal Prediction with strong coverage guarantees (full conformal prediction) can also be quite expensive and infeasible for big datasets, while optimized variants (e.g., split conformal prediction) have weaker coverage guarantees [4].

Another direction, to which this paper belongs to, stems from decision theory, and tries to combine the error-rate minimization and set-size minimization objectives into a single utility function. [19] proposed an efficient set-valued classification, that relies on the conditional class probabilities and tries to find the subset of classes that maximizes the expected utility. This method is computationally efficient, but assumes that the uncertainty is quantified by class-conditional probabilities. It was shown [11] that standard deep learning models can be overconfident and poorly calibrated, which also motivates our use of evidential neural networks in this work. The efficient set-valued classifier proposed by [19] also sorts the classes by their conditional probabilities at inference time,

which, while can be negligible for small label spaces, can be computationally expensive for large label spaces.

Approaches based on Dempster-Shafer (DS) theory [8,22] offer an alternative framework for set-valued classification by directly addressing the inherent uncertainty in the data through belief assignments. In these methods, classifiers generate belief masses over subsets of the label space, following the principles laid out in DS theory. By assigning non-zero mass to composite hypotheses, DS-based methods enable the classifier to express uncertainty with a finer granularity than traditional probability estimates. [9] introduced an approach to neural network classification that integrates evidential reasoning. In this method, the similarity between an input pattern and a limited set of prototypes is evaluated, with each prototype contributing evidence about class membership in the form of belief functions. These individual pieces of evidence are then combined using Dempster's rule of combination to reach a final decision. Initially, the method could make only singleton predictions, or refuse to make a prediction under high uncertainty (classification with rejection). Later, [18] extended this approach to set-valued classification by allowing the classifier to output multiple classes. [24] integrated the set-valued classification framework for deep convolutional neural networks, showcasing their capabilities on complex data. Nevertheless, both [18,24] require computation of the extended utility matrices with shapes $(2^k - 1 \times k)$. Although this matrix can be precomputed, in our experiments we observed that for $k > 20$, the matrices already did not fit in the memory. Moreover, besides the memory limitation, the approach also requires $k(2^k - 1 - k)$ computations, which is infeasible for large label spaces. To solve that issue, [18] suggested to only consider 2-element sets, and the full set. This, however, limits the expressiveness of the model and still requires $\mathcal{O}(k^3)$ computations. In [24], the authors propose identifying, for each class, the two (or more) most similar classes and restricting the utility computation to these class pairs. While this approach can help reduce computational complexity, its performance heavily depends on the dataset's inherent structure (e.g one class can have many similar classes), and it may still limit the overall expressiveness of the model.

[13] proposed an imprecise re-labeling procedure that revises the training data by replacing precise class labels with subsets of candidate classes for samples located in overlapping or isolated regions, and then uses DS theory for learning and reasoning. However, the reasoning step of this approach can still be computationally expensive for large label spaces. The re-labeling step also depends on prediction methods to provide reliable posterior probabilities; if these estimates are poor, the subsequent imprecise labels may not accurately capture the underlying uncertainty. The work of [15] is also notable for combining evidential neural networks with conformal prediction to provide set-valued predictions with guaranteed error rates.

In this paper, however, our work mostly falls in the domains of evidential set-valued classifiers [9,18,24], and utility function maximization [19,18,24]. In our approach, we reduce the exponential computational complexity required by evidential classifiers, by considering only one additional set, suggested by a sec-

ond candidate proposal head. This allows us to scale linearly with the number of classes. The use of evidential neural network also enables us more accurate uncertainty quantification compared to standard deep learning models.

Similar to the work discussed in the previous paragraphs [9,18,24], we propose an evidential classification algorithm. However, instead of learning prototypes and fusing them using Dempster's rule, our architecture directly learns the parameters of the Dirichlet distribution, as introduced in [21]. The Dirichlet distribution models a distribution over class probabilities while also capturing the uncertainty in predictions. Following [21], we leverage subjective logic to reason about uncertainty. The next section provides a detailed discussion of subjective logic and evidential deep learning as used in this work.

3 Preliminaries

In this section, we provide background on Subjective Logic and precise evidential deep learning algorithms, as they form the foundation of our methodology.

3.1 Subjective Logic

Subjective Logic (SL) [14] is an extension of DS theory that provides an intuitive framework for modeling uncertain and imprecise information. SL defines the *domain* $\mathbb{X}$ as the set of all possible states (or classes), analogous to the frame of discernment in DS theory. Additionally, SL defines the *hyperdomain* $\mathscr{R}(\mathbb{X})$ as the reduced superset of $\mathbb{X}$, which is the set of all non-empty proper subsets of $\mathbb{X}$, excluding the full set $\mathbb{X}$ itself.

In SL, beliefs about states in domain $\mathbb{X}$ or hyperdomain $\mathscr{R}(\mathbb{X})$ are represented using *belief masses*. The belief mass distribution $\mathbf{b}_X$ assigns belief masses to the possible values of random variable X, where X can be a state in the domain or a set of states in the hyperdomain. The belief mass on the whole domain is denoted as the *uncertainty mass* u. The additivity property of belief masses is defined as $\sum_{X \in \mathscr{D}} b_X(X) + u_X = 1$, where $\mathscr{D} \in \{\mathbb{X}, \mathscr{R}(\mathbb{X})\}$.

A *subjective opinion* in SL is defined as a triplet $\omega_X = (\mathbf{b}_X, u_X, \mathbf{a}_X)$, where $\mathbf{b}_X$ is the belief mass vector, u_X is the uncertainty mass, and $\mathbf{a}_X$ is the vector of base rates. Base rates are the prior probabilities of the states in the domain $\mathbb{X}$ or hyperdomain $\mathscr{R}(\mathbb{X})$. The opinions are called *multinomial opinions* if $X \in \mathbb{X}$ and *hypernomial opinions* if $X \in \mathscr{R}(\mathbb{X})$. In other words, a multinomial opinion asigns belief masses to precise singleton states, while a hypernomial opinion assigns belief masses to sets of states. Belief mass assigned to a composite value represents vagueness and is referred as *vague belief mass*. A special case of multinomial opinion is when the opinion is about a binary state, which is called a *binomial opinion* and is represented as $\omega_X = (b_X, d_X, a_X, u_X)$, where b_X is the belief mass, d_X is the disbelief mass, a_X is the base rate, and u_X is the uncertainty mass.

Subjective logic also provides a convenient bijective mapping between multinomial opinions and Dirichlet distributions [14]. Dirichlet distribution with parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ is a probability distribution over the K-simplex, and

is defined as:

$$\mathrm{Dir}(\mathbf{p} \mid \boldsymbol{\alpha}) = \begin{cases} \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} p_k^{\alpha_k - 1}, & \text{for } \mathbf{p} \in \mathcal{S}_K \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $B(\boldsymbol{\alpha})$ is the multivariate beta function, and $\mathcal{S}_K$ is the K-simplex. The Dirichlet distribution is convenient for modeling multinomial opinions, as the parameters $\boldsymbol{\alpha}$ can be interpreted as the number of observations of each class. The bijective mapping between multinomial opinions and Dirichlet distributions is given by:

$$b_k = \frac{e_k}{S} = \frac{\alpha_k - 1}{S}, u = \frac{K}{S}, \tag{2}$$

where $S = \sum_{k=1}^{K} \alpha_k$ is called the Dirichlet strength, and $e_k = \alpha_k - 1$ is the supporting evidence for class k. To be able to represent hypernomial opinions using Dirichlet distributions, [14] suggests to use the hyper-Dirichlet distribution, which is a generalization of the Dirichlet distribution to the hyperdomain. The hyper-Dirichlet distribution is defined similarly to the Dirichlet distribution, but with an additional artificial assumption that the states in hyperdomain are mutually exclusive (i.e. $\forall x, y \in \mathscr{R}(\mathbb{X}), x \cap y = \emptyset$).

3.2 Evidential Deep Learning using Subjective Logic

Subjective logic has been used in deep learning to model uncertainty and imprecision in classification tasks. [21] proposed an evidential deep learning framework that uses subjective logic to model uncertainty in deep neural networks. In this framework, the output logits of the neural network are transformed into evidences with some monotonically increasing and non-negative activation function (e.g., softplus, exponent, ReLU, etc.). Then, the evidences are transformed into the parameters of a Dirichlet distribution, with $\alpha_k = e_k + 1$ for each class k. The network is trained using the adapted cross-entropy loss, which is defined as:

$$\begin{aligned} L_{ace}\left(\boldsymbol{\alpha}_n\right) &= \int \left[\sum_{k=1}^{K} -y_{nk} \log p_{nk} \right] \frac{\prod_{k=1}^{K} p_{nk}^{\alpha_{nk}-1}}{B\left(\boldsymbol{\alpha}_n\right)} d\mathbf{p}_n \\ &= \sum_{k=1}^{K} y_{nk} \left(\psi\left(S_n\right) - \psi\left(\alpha_{nk}\right) \right), \end{aligned} \tag{3}$$

where ψ is the digamma function, $S_n = \sum_{k=1}^{K} \alpha_{nk}$ is the Dirichlet strength, and y_{nk} is the one-hot encoded target label for sample n. An additional Kullback-Leibler divergence term is added between incorrect class probabilities and the uniform distribution to encourage the network to output high uncertainty for

incorrect predictions. The KL divergence term is defined as:

$$L_{KL}(\boldsymbol{\alpha}_n) = KL\left[D(\boldsymbol{p}_n \mid \tilde{\boldsymbol{\alpha}}_n) \| D(\boldsymbol{p}_n \mid \mathbf{1})\right] = \log\left(\frac{\Gamma\left(\sum_{k=1}^{K} \tilde{\alpha}_{nk}\right)}{\Gamma(K)\prod_{k=1}^{K}\Gamma(\tilde{\alpha}_{nk})}\right) + \sum_{k=1}^{K}(\tilde{\alpha}_{nk}-1)\left[\psi(\tilde{\alpha}_{nk}) - \psi\sum_{k=1}^{K}\tilde{\alpha}_{nk}\right], \tag{4}$$

where $\tilde{\alpha}_n = \mathbf{y}_n + (\mathbf{1} - \mathbf{y}_n) \odot \boldsymbol{\alpha}_n$ are the Dirichlet parameters after removing the misleading evidence, and the Γ is the gamma function. The total loss is defined as $L = L_{ace} + \sigma_t L_{KL}$, where $\sigma_t[0, 1]$ is an annealing coefficient.

4 Our Methodology

In this paper, we propose a novel approach for Evidential Set-Valued Classification based on Subjective Logic. As discussed in previous section, set-valued classification, especially the ones utilizing DS theory or utility functions often face the problem of handling the exponentially growing number of possible subsets of the frame of discernment. Similarly, in subjective logic, the cardinality of the hyperdomain grows exponentially with the number of states and is equal to $2^K - 2$. Modeling these sets with a neural network becomes infeasible quickly, especially for large K. To address this issue, we propose a novel approach, which suggests assigned belief masses on a reduced hyperset consisting of $K + 1$ elements. We propose to use a 2-head approach, where the first head, called the **multinomial head**, is responsible for providing precise multinomial evidences $e_1, e_2, \ldots, e_K$, and the second head, called the **candidate proposal head**, is responsible for proposing candidate sets of classes and providing the evidence $e_{\mathcal{C}}$. More formally, instead of modeling the evidence of the hyperset $e_H = \{e_1, e_2, \ldots, e_K, e_{K+1}, \ldots, e_{(2^K-2)}\}$, we model the evidence of $e_H = \{e_1, e_2, \ldots, e_K, e_{\mathcal{C}}\}$, where $e_k,\ k \in \{1, \ldots, K\}$ are the evidences for the singleton classes predicted from the multinomial head, and $e_{\mathcal{C}}$ is the evidence on a set of classes $\mathcal{C} \subset \mathcal{X}$ predicted from the candidate proposal head. Finally, the outputs of the two heads are combined using subjective logic to obtain the set-valued predictions together with the uncertainty estimates. The combination strategy ensures that the model provides set-valued predictions when the uncertainty is high, and singleton predictions when the uncertainty is low.

4.1 Multinomial Head

This head is a standard evidential multi-class classifier, which outputs a multinomial opinion for each class. The output of this head are evidences $e_1, e_2, \ldots, e_K$, which are obtained by applying an activation function with non-negative outputs, such as exponential function or softplus function, to the logits of the model. The evidences then can be converted either to multinomial opinions, or to the

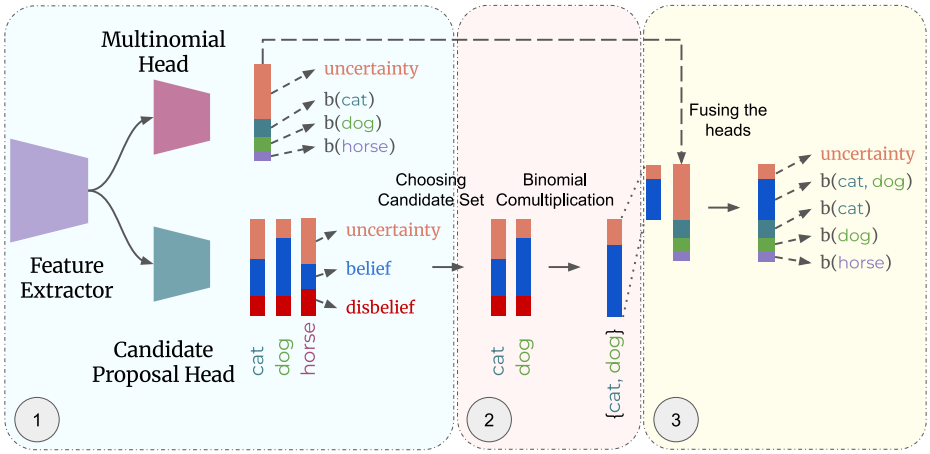

Fig. 1. Pipeline of the proposed approach. (1) Extracted features are fed into multinomial and candidate proposal heads. (2) The candidate proposal head's outputs determine the candidate set using belief and disbelief masses, followed by binomial co-multiplication to compute belief and uncertainty masses. The disbelief masses are merged with the uncertainty mass. (3) The final hypernomial output is obtained by scaling and fitting these masses into the multinomial head's uncertainty. In the figure b(·) represents the belief mass on the specific class.

parameters of Dirichlet distribution. For notation convenience, we note the subjective multinomial opinion representation as: $\boldsymbol{\omega_M}(\mathbf{b}_M, \mathbf{a}_M, u_M)$, where

$$b_{Mk} = \frac{e_k}{\sum_{i=1}^{K} e_i + 1}; \quad a_{Mk} \quad = \frac{1}{K}; \quad u_M = \frac{1}{\sum_{i=1}^{K} e_i + 1}. \tag{5}$$

As we can see, the multinomial head does not provide any set-valued predictions, but unlike standard softmax based classifiers, it provides the uncertainty in the form of uncertainty mass. Traditionally, the uncertainty mass is used to reject the prediction of the model in case of high uncertainty. However, instead of rejecting predictions due to high uncertainty, our approach proposes making multiple plausible predictions.

4.2 Candidate Proposal Head

The candidate proposal head is responsible for proposing candidate sets of classes. This task can be framed similarly to multi-label classification, where the model is trained to predict the likelihood of each class being associated with the input sample, allowing for the possibility of multiple classes being selected simultaneously. However, unlike multilabel classification, the ground truth includes only one positive label. Training with single positive and $k-1$ negative classes penalizes the model for predicting any class other than the ground truth, restricting or preventing multiple predictions. To address this issue, we follow

the approach proposed by [6], where the authors propose to reduce the penalty of false positives by down-weighting the terms in the loss function corresponding to the negative classes. To achieve this they propose the *weak assume negative (WAN)* loss defined as:

$$\begin{aligned}\mathcal{L}_{\text{WAN}}\left(\mathbf{p}_n, \mathbf{y}_n\right) = -\frac{1}{K}\sum_{k=1}^{K}\big[&\mathbb{1}_{[y_{nk}=1]}\log\left(p_{nk}\right) \\ &+\mathbb{1}_{[y_{nk}\neq 1]}\gamma\log\left(1-p_{nk}\right)\big],\end{aligned} \tag{6}$$

where K is the number of classes, $\mathbf{p}_n$ are the conditional class probabilities of the probabilistic classifier, $\mathbf{y}_n$ is the ground truth, $\mathbb{1}_{[\cdot]}$ is the indicator function, and $\gamma = \frac{1}{K-1}$ is the down-weighting factor.

This approach, while suiting to our candidate proposing task, does not quantify the uncertainties in the candidate set, and will also not allow us to obtain a joint belief mass for the proposed set. To address this issue, we will make use of evidential multi-label classification [30], which is a generalization of evidential multi-class classification to multi-label setting. The evidential multi-label classifier puts a prior $\text{Beta}\left(\alpha_{nk}, \beta_{nk}\right)$ on the presence of each class k in the candidate set. The parameters of the Beta distribution are obtained by a neural network predicting two evidence parameters e^{+}_{nk} and e^{-}_{nk} for each class. The evidences are then converted to the parameters of the Beta distribution α_{nk} and β_{nk} by adding one to each evidence. The model is learned by optimizing the Beta loss defined as:

$$\begin{aligned}\mathcal{L}_n(\boldsymbol{\theta}) &= \sum_{k=1}^{K}\int \mathbf{BCE}\left(y_{nk}, p_{nk}\right)\text{Beta}\left(p_{nk}; \alpha_{nk}, \beta_{nk}\right)dp_{nk} \\ &= \sum_{k=1}^{K}\big[y_{nk}\left(\psi\left(\alpha_{nk}+\beta_{nk}\right)-\psi\left(\alpha_{nk}\right)\right) \\ &+ \left(1-y_{nk}\right)\left(\psi\left(\alpha_{nk}+\beta_{nk}\right)-\psi\left(\beta_{nk}\right)\right)\big],\end{aligned} \tag{7}$$

where K is the number of classes, $\mathbf{BCE}\left(\cdot\right)$ is the binary cross entropy loss, $\psi(\cdot)$ is the digamma function. The parameters α_{nk} and β_{nk} can also be converted to subjective binomial opinions $\omega(b_{nk}, d_{nk}, a_{nk}, u_{nk})$, which we will use later for joint belief mass calculation. We can integrate the weak assume negative loss and the Beta loss to obtain the *weak assume negative evidential (WANE)* loss:

$$\begin{aligned}\mathcal{L}_{n\text{WANE}}\left(\alpha_n, \beta_n, \mathbf{y}_n\right) = \sum_{k=1}^{K}\big[&y_{nk}\left(\psi\left(\alpha_{nk}+\beta_{nk}\right)-\psi\left(\alpha_{nk}\right)\right) \\ &+\gamma\left(1-y_{nk}\right)\left(\psi\left(\alpha_{nk}+\beta_{nk}\right)-\psi\left(\beta_{nk}\right)\right)\big].\end{aligned} \tag{8}$$

To control how generous or strict the models shall be in terms of set-sizes, we propose two strategies. First, we introduce penalties prediction set-sizes, to directly control how big the prediction is allowed to be. Second, we employ learnable down-weighting factors that allow the model to infer class-specific preferences for set size strictness directly from the data.

4.3 Constraint-based WANE Optimization

An important aspect of the candidate proposal head is that it should propose a set of classes, where the cardinality of the set is higher than one: $|\mathcal{C}| > 1$. To achieve this, we can define a penalty term in the loss function, which penalizes the model for $|\mathcal{C}| \leq 1$. The penalty term can be defined as:

$$\mathcal{L}_{\text{p2}}(\mathbf{p}) \;=\; \lambda \cdot \min\left(0, \sum_{k=1}^{K} \mathbb{1}_{[p_k > 0.5]} - 2\right), \tag{9}$$

where λ is the penalty coefficient. However, the indicator function is not a differentiable function, hence we perform a smooth relaxation of the penalty term by using an estimated soft cardinality. For predicted conditional probabilities $\mathbf{p} = [p_1, \ldots, p_L]$, we define the estimated (soft) cardinality s as:

$$s = \sum_{k=1}^{K} \sigma\Big(\eta\,(p_k - 0.5)\Big), \tag{10}$$

where $\sigma(\cdot)$ is the sigmoid function and η is the steepness parameter. Then, the relaxed penalty term will become:

$$\mathcal{L}^{r}_{\text{p2}}(\mathbf{p}) \;=\; \lambda \cdot \min\Big((s-2)\,, 0\Big). \tag{11}$$

The resulting loss function in its current form constrains the candidate set's cardinality from below, but it does not regulate its average size. Following [6], the original formulation sets $\gamma = \frac{1}{K-1}$, which reduces the penalty on all negative classes in the loss. Although this allows the model to propose multiple classes, it inadvertently encourages high-cardinality outputs. Instead, our objective is to design a loss function that imposes higher penalties for false positives from classes not relevant to the input, while applying lower penalties for false negatives in the relevant classes. To address this issue, we propose a cardinality-based loss term that directly controls the average size of the candidate set. Our approach leverages a differentiable approximation of the set's cardinality from Equation 10. We then augment the original loss (e.g. the WANE loss defined in Eq. 8) with a penalty term that directly discourages excessively large candidate sets:

$$\mathcal{L}_{\text{card}} = \gamma\, h(s), \tag{12}$$

where $\gamma > 0$ is a hyperparameter, and the penalty function $h(s)$ is chosen to be monotonically increasing in s. Inspired by decision-theoretic utility functions, a reasonable choice is

$$h(s) = 1 - \frac{1+\beta^2}{s+\beta^2}, \tag{13}$$

with *beta* (distinct from parameters of the Beta distribution) controlling the tolerance for larger candidate sets. Lower β values enforce stricter cardinality control, while higher β values allow for larger sets.

In our proposed training regime, the overall loss is defined as

$$\mathcal{L}_{\text{const}} = \mathcal{L}_{\text{WANE}} + \mathcal{L}_{\text{p2}}^{r} + \mathcal{L}_{\text{card}}, \tag{14}$$

which ensures that the model is penalized not only for misclassifications but also for deviating from the desired candidate set size. We train the model with $\mathcal{L}_{\text{WANE}}$ and $\mathcal{L}_{\text{p2}}^{r}$ for a number of epochs to obtain reliable conditional class probabilities, and the cardinality penalty is introduced gradually (via an annealing coefficient $\tau \in [0, 1]$) so that the loss function adapts smoothly to the new objective.

4.4 Alternative Learnable-Factor Based WANE Optimization

In the previous sub-section, we introduced two penalty terms to enforce a minimum cardinality for the predicted set and to penalize excessively large sets. However, this soft-thresholding approach introduces several challenges. First, the soft estimate of the set size, computed using sigmoid activations, is only an approximation of the true cardinality. When the predicted probabilities are close to the 0.5, the estimated size can deviate significantly from the actual number of selected elements. Second, using steep sigmoid functions to approximate hard thresholds can lead to optimization issues such as vanishing or unstable gradients. Third, the gradients induced by the WANE loss and the cardinality-based penalties may conflict, potentially pulling the model in opposing directions and hindering effective learning. Finally, the penalties are not class-dependent, meaning the model applies the same set-size constraints uniformly across all classes. This prevents the model from adapting its prediction behavior based on class-specific characteristics, such as semantic similarity or varying levels of ambiguity between classes. To address this issue we propose an alternative optimization idea based on learnable, class-specific down-weighting factors.

As discussed before, the down-weighting factor γ in equation 8 is controlling how strong the penalization for false positives shall be. While the authors [6] choose $\gamma = \frac{1}{K-1}$, it puts a uniform penalty for false positives on all classes. Nevertheless, based on the semantics of classes and the uncertainties, it may be more acceptable to have false positives for some classes than the others. For example, for an image of a dog, it is more acceptable to have the class `wolf` as a false positive, than the class `car`. To achieve this, we propose to have a matrix of down-weighting factors $\Gamma = [\gamma_{jk}] \in [0, 1]^{K \times K}$, where each γ_{jk} represents the down-weighting factor for penalization of a sample belonging to class j with a false positive in class k. The WANE loss wit learnable factors then will be:

$$\begin{aligned} \mathcal{L}_{n\text{WANE}-\text{LF}}\left(\alpha_n, \beta_n, \mathbf{y}_n\right) = \sum_{k=1}^{K} \big[& y_{nk}\left(\psi\left(\alpha_{nk} + \beta_{nk}\right) - \psi\left(\alpha_{nk}\right)\right) \\ & + \gamma_{ji}\left(1 - y_{nk}\right)\left(\psi\left(\alpha_{nk} + \beta_{nk}\right) - \psi\left(\beta_{nk}\right)\right)\big], \end{aligned} \tag{15}$$

where j is the index of the correct ground truth class, such that $y_{nj} = 1$.

To obtain the matrix $\Gamma = [\gamma_{jk}]$, we introduce a learnable real-valued parameter matrix $Z = [z_{jk}] \in \mathbb{R}^{K \times K}$. Each element γ_{jk} is then computed as

$\gamma_{jk} = \sigma(z_{jk})$, where $\sigma(\cdot)$ denotes the sigmoid function. This ensures that all down-weighting factors γ_{jk} lie in the range $[0, 1]$, while allowing the model to learn them in a fully differentiable manner.

However, since the down-weighting parameters are now learnable, the optimal values for minimizing the WANE-LF loss could trivially become $\gamma_{jk} = 0$, effectively ignoring false positive penalties. To prevent Γ from collapsing to zero, we introduce a regularization term to the loss function:

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_{n\text{WANE}-\text{LF}} + \lambda e^{-\bar{z}}, \tag{16}$$

where $\bar{z} = \frac{1}{K^2} \sum_{i,j=1}^{K} z_{ij}$ is the average of the elements of the Z matrix, and $\lambda \geq 0$ is a hyperparameter that controls the strength of the regularization. By adjusting the value of λ, we can effectively control how tolerant the model is to larger predicted label sets. Lower values of λ encourage the model to assign lower down-weighting factors γ_{jk}, thereby reducing the penalty for false positives and promoting broader predictions. Conversely, higher values of λ result in stricter penalization, favoring more conservative and precise label sets.

In contrast to the optimization strategy proposed in the previous section, the learnable discounting factors introduce a quadratic number of parameters with respect to the number of classes. However, these factors are required only during training and are not used at inference time, which ensures that the inference complexity remains linear.

4.5 Combining the Heads

The outputs of the two heads are combined using subjective logic. The output of the multinomial head is a set of multinomial evidences, which can be converted to subjective multinomial opinions. The output of the candidate proposal head is a set of positive and negative evidences, which can be converted to subjective binomial opinions about each class. More formally, the output of the multinomial head can be represented as $\boldsymbol{\omega}_M(\mathbf{b}, a, \mathbf{u})$, and the output of the candidate proposal head can be represented as the set $\{\omega_k\}, k \in \{1, \ldots, K\}$, where $\omega_k = \omega(b_k, d_k, a_k, u_k)$. We want to combine the evidences in such a way, that the model provides set-valued prediction only on very hard samples, where providing a precise prediction would have high risk of being incorrect. Conveniently, in most evidential deep learning approaches, the loss function is designed to assign higher uncertainty masses to incorrect classifications. This is achieved by decreasing the Kullback-Leibler (KL) divergence between the incorrect predictions and the uniform distribution [21]. Hence, the uncertainty value of the multinomial head is a good measure to assess the complexity of the sample. It then follows that to reduce the classification error, we want to make set-valued predictions when the uncertainty mass is high. To achieve this, we will move some of the uncertainty mass from the multinomial head as a belief mass to the candidate proposal head.

First, let us recall that in the candidate proposal head, we have belief, disbelief and uncertainty masses for each of the classes, but not for the selected candidate set. To obtain these masses for the candidate set, we will make use of the binomial co-multiplication operation from subjective logic.

Definition 1 (Subjective Binomial Co-multiplication [14]). *Let* $\mathbb{X} = \{x, \bar{x}\}$ *and* $\mathbb{Y} = \{y, \bar{y}\}$ *be two domains. Let* $\omega_x = \omega(b_x, d_x, a_x, u_x)$ *and* $\omega_y = \omega(b_y, d_y, a_y, u_y)$ *be two independent subjective binomial opinions on* x *and* y *respectively. The binomial co-multiplication* $\omega_x \sqcup \omega_y$ *provides the subjective binomial opinion on disjunction* $x \vee y = \{(xy), (x\bar{y}), (\bar{x}y)\}$ *and is defined as:*

$$\omega_{x\vee y} : \begin{cases} b_{x\vee y} = b_x + b_y - b_x b_y, \\ d_{x\vee y} = d_x d_y + \frac{a_x(1-a_y)d_x u_y + (1-a_x)a_y u_x d_y}{a_x + a_y - a_x a_y}, \\ u_{x\vee y} = u_x u_y + \frac{a_y d_x u_y + a_x u_x d_y}{a_x + a_y - a_x a_y}, \\ a_{x\vee y} = a_x + a_y - a_x a_y. \end{cases} \tag{17}$$

Since our candidate proposal head is designed similar to multi-label classification, we can follow the binary relevance approach [28] of assuming independence between the candidates, and apply the binomial co-multiplication operation to obtain the subjective binomial opinion on the candidate set. To achieve this, first we form the candidate set by selecting all the classes where the belief mass is greater than the disbelief mass: $\mathcal{C} = \{i | b_i > d_i\}$. Then we apply the binomial co-multiplication operation to obtain the joint belief mass $b_\mathcal{C}$ with:

$$b_\mathcal{C} = 1 - \prod_{i \in \mathcal{C}} (1 - b_i). \tag{18}$$

The proof of Equation 18 can be found in the our GitHub repository. To simplify the computations we will join the disbelief mass and the uncertainty mass into a single uncertainty mass, which can be obtained with: $u_\mathcal{C} = 1 - b_\mathcal{C}$, due to the additivity property of belief masses. Having the subjective binomial opinion on the candidate set, we can now combine it with the subjective multinomial opinion from the multinomial head. To do that, we need to scale the belief and uncertainty masses of the candidate to fit into the uncertainty mass of the multinomial opinion. The scaled belief and uncertainty masses of the candidate set can be obtained with:

$$b'_\mathcal{C} = b_\mathcal{C} \cdot u_M \quad u'_\mathcal{C} = u_\mathcal{C} \cdot u_M \quad a_\mathcal{C} = \sum_{i \in \mathcal{C}} a_i, \tag{19}$$

where u_M is the uncertainty mass from the multinomial opinion. Finally, the combined hyper-opinion $\boldsymbol{\omega}_H(\mathbf{b}_H, \mathbf{a}_H, u_H)$ can be formed, where:

$$\mathbf{b}_H = \{b_{M1}, b_{M2}, \ldots, b_{MK}, b'_\mathcal{C}\}, \quad \mathbf{a}_H = \{a_{M1}, a_{M2}, \ldots, a_{MK}, a_\mathcal{C}\}, \quad u_H = u'_\mathcal{C}. \tag{20}$$

During the decision making stage, a singleton class can be selected if the belief b_{Mi} is greater than the other belief masses, and the proposed candidate set can be selected if the belief $b'_\mathcal{C}$ is greater than the other belief masses.

All operations described here have worst case $\mathcal{O}(k)$ time complexity, which means our approach scales linearly with the number of classes.

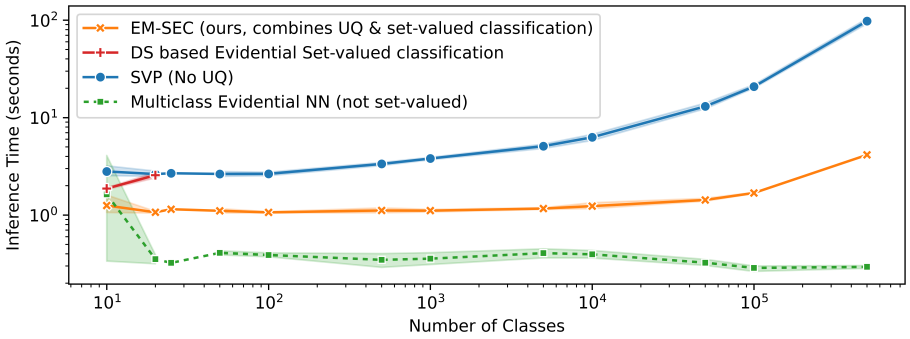

Fig. 2. Inference time (log-log scale) on 10,000 samples for varying class counts. DS is only reported for 10 and 20 classes due to scalability issues. EM-SEC is 2–23× faster than SVP and DS. Multiclass evidential runtime is shown for reference, but not as a baseline, since t does not perform set-valued classification.

5 Experiments

We conduct extensive experiments to evaluate the performance of our proposed method. As baselines, we use the evidential classifier by [24], referred to as DS (Dempster-Shafer-based Classifier), and the efficient set-valued classifier SVP by [19]. We denote with EM-SEC the model with WANE loss (Eq. 8), while EM-SEC-LF is the model with WANE-LF loss (Eq. 15). These baselines demonstrate that EM-SEC(-LF) is both faster than other evidential set-valued classifiers and competitive with the highly efficient SVP.

For evaluation, we use CIFAR-10 and CIFAR-100 [17] due to their widespread use in image classification benchmarks. Since the original DS implementation was in TensorFlow, we adapted it to PyTorch for compatibility. The SVP implementation was in C++, which posed an unfair advantage. To ensure a fair comparison, we used a Large Language Model (ChatGPT o3-mini-high) to translate it into Python and verified that key components fully utilized PyTorch vectorized operations. We used ResNet-18 [12] architecture as encoders for all baselines.

Scalability Analysis First, we will try to understand how the proposed approach scales to classification tasks with higher number of classes. To have a controlled experimental setting, we will take random 32 x 32 images, and try to classify them into 10 - 500,000 classes, and log the inference time for 10000 samples. For the DS approach, the extended utility matrix did not fit in GPU memory for $K > 20$. Hence, we will not provide the results for DS for classes higher than 20. **As we can see in Figure 2, the proposed EM-SEC[4] approach is consistently faster (from 2 up to 23 times) than the other baseline set-valued classifiers. Specifically, for 500,000 classes the inference takes around 4 seconds for EM-SEC and 97 seconds for SVP.**

[4] EM-SEC and EM-SEC-LF have identical inference times due to shared architecture.

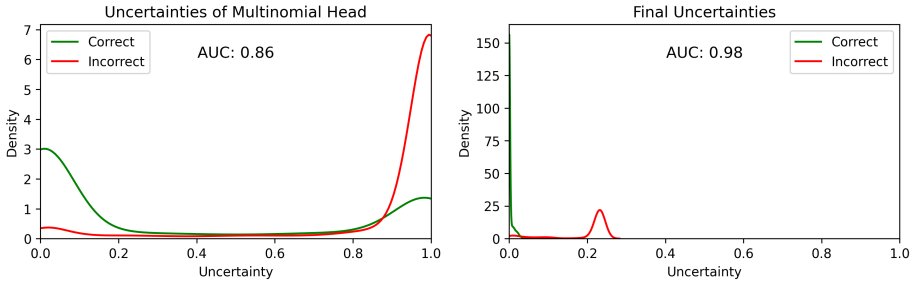

Fig. 3. Uncertainties of the multinomial head (left) and after fusing with candidate proposal head (right) on CIFAR-100 dataset. The AUC score of misclassification detection based on uncertainties is also provided. The average set size is 1.69. In the multinomial head incorrectly classified samples have very high uncertainty and correctly classified ones have low uncertainty. After fusing with candidate proposal head, most of the incorrect classification uncertainty mass is moved to correct classification.

Uncertainty Analysis Here, we empirically motivate our approach by analyzing uncertainties before and after EM-SEC set-valued classification. As shown in Figure 3, in standard evidential multiclass classification (multinomial head), incorrectly classified samples exhibit high uncertainty, with an area-under-the-curve (AUC) value of 0.86. With EM-SEC, we identify these high-uncertainty points and instead predict multiple possible labels. As illustrated in Figure 3, EM-SEC redistributes most of the uncertainty mass from misclassified samples to correctly classified points with low uncertainty. We also observe a secondary spike in uncertainty after fusion, which can serve as a reject option, enabling the model to further reduce incorrect classifications. **Notably, with a misclassification detection AUC score of 0.98, our approach effectively eliminates the risk of incorrect predictions on this dataset while maintaining an average of just 1.69 predictions per set.**

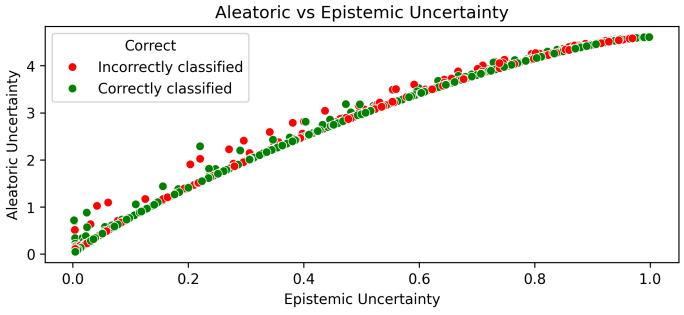

Fig. 4. Aleatoric vs Epistemic uncertainties of multinominal head on CIFAR-100 dataset. The disentanglement is performed following the formulas from [25].

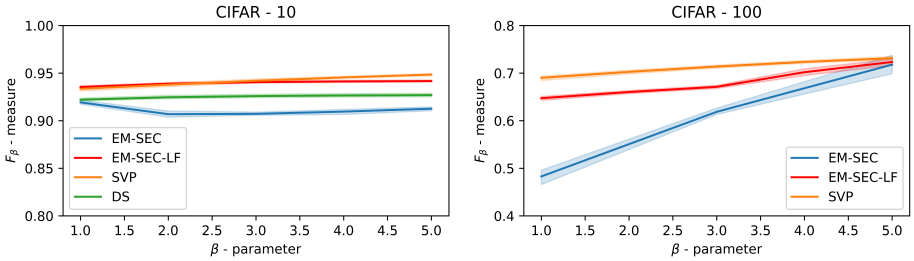

Fig. 5. F_β Utility for various beta values. EM-SEC represents the the model with WANE loss (Eq. 8), while EM-SEC-LF is the model with WANE-LF loss (Eq. 15).

Another important consideration is the type of uncertainty used to determine when to provide set-valued decisions. In the uncertainty quantification (UQ) literature, two main types of uncertainty are typically defined: epistemic and aleatoric [16]. Epistemic uncertainty arises from a lack of knowledge in the model, whereas aleatoric uncertainty reflects inherent noise in the data. In subjective logic, the uncertainty mass is more closely related to epistemic uncertainty. **Compared to the baseline methods, EM-SEC efficiently estimates the epistemic, and if needed aleatoric uncertainties.** Ideally, set-valued predictions are suited for high aleatoric uncertainty, while no decision shall be made under high epistemic uncertainty. However, as shown by [20], these two often correlate strongly. Our results in Figure 4 confirm this, so we do not distinguish between them in our approach.

Utility-based comparison In this section, we compare model performances using the F_β measure ($F_\beta(s) = \frac{1+\beta^2}{s+\beta^2}$). Higher β values are more tolerant to larger set sizes, whereas lower values impose stricter penalties. A high F_β score at low β values indicates accurate predictions with minimal expected set sizes. For SVP and EM-SEC, the β parameter was chosen corresponding to the evaluation F_β-score. For DS method, the γ parameter was chosen as 0.9, since it provided the best results in the paper. Finally, for EM-SEC-LF the λ parameter was tuned for each F_β measure (See the supplementary material in GitHub for more details).

As shown in Figure 5, on the CIFAR-10 dataset, EM-SEC achieves utility values comparable to the DS method for $\beta = 1$, but the performance worsens with higher values of β. This suggests that EM-SEC provides better predictions with lower set-size budget. In contrast, EM-SEC-LF, which incorporates learnable down-weighting factors, reaches performance on par with SVP, indicating its effectiveness in adapting to class-specific uncertainties. On the more challenging CIFAR-100 dataset, EM-SEC underperforms for lower β values but shows improved results as β increases, gradually approaching the performance of SVP. EM-SEC-LF significantly narrows the performance gap with SVP across all β values, with a similar upward trend as β increases. We were unable to evaluate the DS on CIFAR-100 due to to out-of-memory issues.

6 Conclusion

In this paper, we introduced the efficient evidential set-valued classification approach, EM-SEC, that leverages subjective logic and evidential deep learning to quantify prediction uncertainty and generate set-valued outputs when uncertainty is high. Our experiments demonstrate that EM-SEC scales efficiently to datasets with a large number of classes in terms of inference time and provides reliable uncertainty estimates that can filter out unreliable predictions.

Given the promising performance of EM-SEC in the unimodal case, we aim to extend it to a multimodal setting, where conflicting information across modalities introduces additional uncertainty and increased computational complexity. While [2] address modality conflict by reallocating conflict mass to uncertainty through evidential fusion, we propose an alternative approach: redirecting the conflict mass toward composite classes. We also plan to evaluate the effectiveness of this strategy using multimodal extensions of CIFAR, such as the LUMA dataset [3].

References

1. Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., et al.: A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information fusion **76**, 243–297 (2021)
2. Bezirganyan, G., Sellami, S., Berti-Equille, L., Fournier, S.: Multimodal learning with uncertainty quantification based on discounted belief fusion. In: Proceedings of The 28th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 258, pp. 3142–3150. PMLR (2025)
3. Bezirganyan, G., Sellami, S., Berti-Équille, L., Fournier, S.: Luma: A benchmark dataset for learning from uncertain and multimodal data. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (2025)
4. Cherubin, G., Chatzikokolakis, K., Jaggi, M.: Exact optimization of conformal predictors via incremental and decremental learning. In: International Conference on Machine Learning. pp. 1836–1845. PMLR (2021)
5. Chzhen, E., Denis, C., Hebiri, M., Lorieul, T.: Set-valued classification–overview via a unified framework. arXiv preprint arXiv:2102.12318 (2021)
6. Cole, E., Mac Aodha, O., Lorieul, T., Perona, P., Morris, D., Jojic, N.: Multi-label learning from single positive labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 933–942 (2021)
7. Cresswell, J.C., Sui, Y., Kumar, B., Vouitsis, N.: Conformal prediction sets improve human decision making. In: International Conference on Machine Learning. pp. 9439–9457. PMLR (2024)
8. Dempster, A.P.: A generalization of bayesian inference. Journal of the Royal Statistical Society: Series B (Methodological) **30**(2), 205–232 (1968)
9. Denoeux, T.: A neural network classifier based on dempster-shafer theory. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans **30**(2), 131–150 (2000)
10. Garcin, C., Servajean, M., Joly, A., Salmon, J.: A two-head loss function for deep average-k classification. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 7358–7367. IEEE (2025)

11. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Imoussaten, A., Jacquin, L.: Cautious classification based on belief functions theory and imprecise relabelling. International Journal of Approximate Reasoning **142**, 130–146 (2022)
14. Jøsang, A.: Subjective logic, vol. 3. Springer (2016)
15. Kempkes, M.C., Dunjko, V., van Nieuwenburg, E., Spiegelberg, J.: Reliable classifications with guaranteed confidence using the dempster-shafer theory of evidence. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 89–105. Springer (2024)
16. Kiureghian, A.D., Ditlevsen, O.: Aleatory or epistemic? Does it matter? Structural Safety **31**(2), 105–112 (Mar 2009)
17. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
18. Ma, L., Denoeux, T.: Partial classification in the belief function framework. Knowledge-Based Systems **214**, 106742 (2021)
19. Mortier, T., Wydmuch, M., Dembczyński, K., Hüllermeier, E., Waegeman, W.: Set-valued prediction in multi-class classification. In: 31st Benelux conference on Artificial Intelligence (BNAIC 2019); 28th Belgian Dutch conference on Machine Learning (Benelearn 2019). vol. 2491. CEUR (2019)
20. Mucsányi, B., Kirchhof, M., Oh, S.J.: Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. Advances in Neural Information Processing Systems **37**, 50972–51038 (2024)
21. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. Advances in neural information processing systems **31** (2018)
22. Shafer, G.: A Mathematical Theory of Evidence. Princeton university press (1976)
23. Shafer, G., Vovk, V.: A tutorial on conformal prediction. Journal of Machine Learning Research **9**(3) (2008)
24. Tong, Z., Xu, P., Denoeux, T.: An evidential classifier based on dempster-shafer theory and deep learning. Neurocomputing **450**, 275–293 (2021)
25. Ulmer, D., Hardmeier, C., Frellsen, J.: Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation. Trans. Mach. Learn. Res. **2023** (2023)
26. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic learning in a random world, vol. 29. Springer (2005)
27. Wang, Z., Qiao, X.: Set-valued classification with out-of-distribution detection for many classes. Journal of Machine Learning Research **24**(375), 1–39 (2023)
28. Zhang, M.L., Li, Y.K., Liu, X.Y., Geng, X.: Binary relevance for multi-label learning: an overview. Frontiers of Computer Science **12**, 191–202 (2018)
29. Zhang, Z., Liu, Z., Ning, L., Martin, A., Xiong, J.: Representation of imprecision in deep neural networks for image classification. IEEE Transactions on Neural Networks and Learning Systems (2023)
30. Zhao, C., Du, D., Hoogs, A., Funk, C.: Open set action recognition via multi-label evidential learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22982–22991 (2023)