A subjective-logic-based model uncertainty estimation mechanism for out-of-domain detection

Jiarui Xie
LIMOS, UMR6158, CNRS, F-63000
Clermont Auvergne University
Clermont-Ferrand, France
jiarui.xie@uca.fr

Thierry Chateau
Logiroad
Institut Pascal, UMR6602, CNRS
Clermont Auvergne University
Clermont-Ferrand, France
thierry.chateau@logiroad.com

Violaine Antoine
LIMOS, UMR6158, CNRS, F-63000
Clermont Auvergne University
Clermont-Ferrand, France
violaine.antoine@uca.fr

***Abstract*—Deep neural networks are important for a wide range of scientific and industrial processes. However, a classical discriminative model always makes a classification with respect to the probabilities allocated to the training labels, even when the sample is out of the domain. Thus, it is of interest to assign uncertainty to a model prediction to avoid such a situation. Fortunately, there are many existing methods for dealing with this kind of problem, one branch of which involves combining neural networks with subjective logic (SL). Based on previous works, we propose a new method called subjective-logic-based uncertainty estimation (SLUE) that can take the base rate distribution explicitly into account to refine the Dirichlet distribution parameters and guide the model training. Experiments were performed on several public datasets and additional adversarial datasets. Compared with existed methods, SLUE reached better uncertainty assessment performance (15% improvement in terms of % max entropy) as well as comparable prediction accuracy performance.**

***Index Terms*—Base Rate Distribution, Uncertainty, Out-of-domain Detection, Subjective Logic**

I. Introduction

Currently, deep neural networks have a pivotal role in various applications of human endeavor. It possesses an admirable prediction accuracy but less prediction confidence. As an example, if we feed an image of car into a cat-dog neural network, this image will be classified as either being a cat or a dog rather than as being inappropriate. Obviously, to a human being, this is lacking in intelligence. To address this problem, we propose a new method called subjective-logic-based uncertainty estimation (SLUE). It is derived from evidential deep learning (EDL) [1] based on subjective logic (SL) by processing model outputs while giving uncertainty to each prediction. The impetus behind this is that the subjective logic model boosts the traditional evidence theory (belief function) in the sense that opinions take base rates into account, whereas evidence theory ignores base rates. With base rates, we can make good use of prior knowledge. At the same time, it also makes it possible to define a bijective mapping between subjective opinions and Dirichlet probability density functions (PDFs) [2]. With a bijective mapping, the uncertainty and probability expectation formula can be easily derived.

In EDL, there is no explicit use of base rates, because that it just used the default base rates neither analyze the potential initialization methods. In addition, the base rates is not updated during the training process. The Dirichlet distribution parameters appeared in EDL ($\boldsymbol{\alpha} = \mathbf{e} + 1$) are composed of evidence and a weight of one that is allocated to all the training classes. In comparison, the SLUE used the base rates to refine the Dirichlet distribution parameters ($\boldsymbol{\alpha} = \mathbf{e} + \mathcal{C}\mathbf{a}$) to guide the training process. Within the SLUE method, the base rates is updated after each batch leading to a more flexible and precise classification.

In addition to the underutilization of base rates within the EDL method, the sum of all the weights equals the number of training classes. Intuitively, the sum of all the weights could be a hyperparameter. Consequently, in SLUE, the sum of all the weights is represented by the prior constant $\mathcal{C}$; the optimum can be explored through experiments.

Compared with the existing methods, this work makes the following contributions:

(1) Taking the base rates explicitly into account, the initial choice is evaluated. Since the update of base rates is after each batch, comprehensive analysis with experiments under the batch size setting was carried out.
(2) Extract the hyperparameter $\mathcal{C}$ to generate a more uniform representation, the function of $\mathcal{C}$ is revealed by experiments.

The rest of the paper is organized as follows. Section II recalls the necessary background about SL (e.g., Dirichlet distribution and uncertainty calculation formula). Section III introduces the new SLUE method and presents its loss function. Several experiments (e.g., out-of-domain detection, adversarial samples detection, and batch size effect estimation) are described in Section IV. Section V lists the related work. Finally, Section VI makes a conclusion about the work.

II. Subjective Logic Basics

Before going deeper insight into the SLUE method, we must introduce some prerequisite definitions.

Definition 1 (State Space): A state space, also known as a frame of discernment, is an exhaustive set of mutually exclusive atomic states.

Definition 2 (Evidence): Let $\Omega = \{\omega_i \mid i = 1, \ldots, K\}$ be a state space, and $\mathbf{e} = \{e_i \mid i = 1, \ldots, K\}$ representing evidence according to each element in Ω that satisfies $e_i \geq 0$.

Definition 3 (Subjective Logic Opinion): Let $\Omega = \{\omega_i \mid i = 1, \ldots, K\}$ be a state space. A SL opinion is an ordered triple $(\mathbf{b}, u, \mathbf{a})$, with

$$\sum_{i=1}^{K} a_i = 1, \quad 0 \leq a_i \leq 1. \tag{1}$$

$$u + \sum_{i=1}^{K} b_i = 1, \quad 0 \leq b_i \leq 1, \tag{2}$$

where $\mathbf{b}$ delegating the belief mass distribution over Ω, u is the uncertainty mass, and $\mathbf{a}$ is the base rate distribution representing the prior knowledge over Ω.

Given a state space of cardinality K, the default base rates for each element in the state space is $\frac{1}{K}$, but it is possible to define the other base rates for all the mutually exclusive elements of the state space, as long as the additivity constraint (1) is satisfied. Base rates can also be dynamically updated as a function of observed evidence. For example, in a box that contains red and black balls of unknown proportion, the initial base rates of the balls can be set to 0.5. After having picked (with return) several balls the base rates can be updated according to the observed balls proportions [2].

Definition 4 (Dirichlet Distribution): Let Ω be a state space of K mutually disjoint values, $\mathbf{e}$ be the evidence for outcome $\omega_i \in \Omega$, $\mathbf{a}$ a base rates over Ω, and $\mathbf{p}$ the probability distribution of $\omega_i \in \Omega$. Then, the probability density function

$$Dir(\mathbf{p}, \mathbf{e}, \mathbf{a}) = \frac{\Gamma\left(\sum_{i=1}^{K} (e_i + \mathcal{C}a_i)\right)}{\prod_{i=1}^{K} \Gamma\left(e_i + \mathcal{C}a_i\right)} \prod_{i=1}^{K} p_i^{(e_i + \mathcal{C}a_i - 1)}, \tag{3}$$

where $\mathcal{C}$ is a prior constant, $\Gamma(\cdot)$ represents a gamma function.

The probability expectation of the K possible outcomes can now be written as

$$\mathrm{E}\left(\mathbf{p} \mid \mathbf{e}, \mathbf{a}\right) = \frac{\mathbf{e} + \mathcal{C}\mathbf{a}}{\mathcal{C} + \sum_{i=1}^{K} e_i}. \tag{4}$$

Meanwhile, the uncertainty can be derived from the the parameters of the Dirichlet distribution [1]. Then, the following equivalence holds:

$$u = \frac{\mathcal{C}}{\mathcal{C} + \sum_{i=1}^{K} e_i}. \tag{5}$$

III. The SLUE Method

Suppose the state space is composed by sample predictions; we define the evidence $\mathbf{e}_t$, the probability $\mathbf{p}_t$, and base rates $\mathbf{a}_t$ for the current batch t. Consequently, the probability $\mathbf{p}_{t-1}$ for past batch $t-1$. We feed the model outputs into the rectified linear unit (ReLU) and take the outputs as $\mathbf{e}_t$, meanwhile, $\mathbf{p}_t$ can be calculated with (4).

In the beginning, we do not know the class proportion. However, once made a batch prediction, we can take the previous probability $\mathbf{p}_{t-1}$ to update the base rates $\mathbf{a}_t$. Exactly as we do in the "ball game", as we do not know the class proportion, consequently, after having made predictions, we can take relative proportions of observed classes probability as the base rates. Then, $\boldsymbol{\alpha} = \mathbf{e} + \mathcal{C}\mathbf{a}$ are used as the parameters of a Dirichlet distribution. Sample uncertainty can be calculated with (5) to determine whether to accept or reject the current prediction. The SLUE method uses (5) to quantify uncertainty directly and probability calculated from (4) to make prediction decisions. The choosing strategy for initial base rates and prior constant $\mathcal{C}$ are discussed in Section IV-B.

The format loss function, which comes from [1], is adopted as follows:

$$\mathcal{L}(\Theta) = \sum_{i=1}^{N} \mathcal{L}_i(\Theta) = \sum_{i=1}^{N} \sum_{j=1}^{K} \left(y_{ij} - p_{ij}\right)^2 + \frac{p_{ij}\left(1 - p_{ij}\right)}{\left(S_i + 1\right)}, \tag{6}$$

where N is the number of samples, K is the number of classes, $\mathbf{y}_i = \{y_{ij} \mid j = 1, \ldots, K\}$ is a one-hot vector that encodes the ground-truth class of sample $\mathbf{x}_i$ with $y_{ij} = 1$ and $y_{ik} = 0$, for all $k \neq j$, $\mathbf{p}_i = \{p_{ij} \mid j = 1, \ldots, K\}$ is a vector representing class assignment probabilities, and $S_i = \mathcal{C} + \sum_{i=1}^{K} e_i$.

A Kullback-Leibler (KL) term is used to estimate the divergence caused by unknown states. In order to alleviate the error brought by misclassified samples, a KL term is incorporated into the loss function as follows:

$$\begin{aligned} \mathcal{L}(\Theta) = & \sum_{i=1}^{N} \mathcal{L}_i(\Theta) \\ & + \lambda_s \sum_{i=1}^{N} KL\left[D\left(\mathbf{p}_i \mid \mathbf{e}_i, \mathbf{a}_i\right) \| D\left(\mathbf{p}_i \mid \langle 1, \ldots, 1 \rangle\right)\right], \end{aligned} \tag{7}$$

where the annealing coefficient $\lambda_s = \min(1.0, \frac{s}{10}) \in [0, 1]$, s representing the current training epoch index, and $D\left(\mathbf{p}_i \mid \langle 1, \ldots, 1 \rangle\right)$ is the uniform Dirichlet distribution.

IV. Experiment

A. Experimental protocol

We use the standard convolutional neural networks (CNNs) with ReLU as the neural network architecture, all experiments are implemented in Pytorch[1]. For the MNIST dataset, a standard LeNet was trained. Following the suggestion of [3], an augmented LeNet version that contained 192 filters at each convolutional layer and had 1000 hidden units for the fully connected layers was trained for CIFAR10 and CIFAR100 datasets. The characteristics of the datasets are shown in Table. I.

The criteria were as follows:

(1) *Real accuracy* is the number of correctly classified in-domain samples plus the rejected out-of-domain samples divided by the total number of samples. This is different from test accuracy, which only takes correctly classified samples into account; with uncertainty, the rejected out-of-domain samples should also be regarded

[1]The SLUE result visualization demo is available under https://github.com/Soplia/SLUE-demo

TABLE I
A QUICK VIEW OF DATASETS INVOLVED

Name	# Classes	# Training samples	# Testing samples
MNIST [4]	10	55000	10000
CIFAR10 [5]	10	50000	10000
CIFAR100 [5]	10	5000	1000
LSUN [6]	10	-	3000
TEXTURE [7]	10	-	1300
PLACES365 [8]	10	-	4000
MNIST5	5	25000	5000
CIFAR5	5	28000	4800

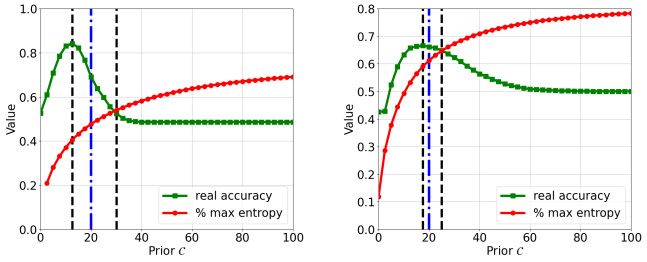


Fig. 1. Estimation results based on different prior constant $\mathcal{C}$ values for MNIST (left) and CIFAR10 (right) datasets

as correctly classified samples and be taken into account. A higher value is better.

(2) *Entropy* is used to evaluate the prediction uncertainty as described in [3]. The increase in prediction uncertainty leading to an increase in entropy. Consequently, % max entropy which means the ratio of prediction entropy to the maximum prediction entropy and cumulative distribution function (CDF) is used for uncertainty estimation. For out-of-domain samples, a higher % max entropy value is better.

B. Initial stage choice

This section describes and compares several strategies to choose the prior constant $\mathcal{C}$ and initial base rates. The first five classes of the MNIST and CIFART10 datasets were extracted to generate MNIST5 and CIFAR5 datasets. The model with the SLUE method was trained based on these two datasets; then it was tested with MNIST and CIFAR10 test datasets that contained all ten classes. During this process, the first five class samples played the role of in-domain samples, while the last five class samples acted as out-of-domain samples.

To select the optimum prior constant $\mathcal{C}$, we examined various settings from zero to 100 at an interval of five (the number of training classes). As we can see from Fig. 1, the % max entropy kept increasing; meanwhile, the real accuracy reached an optimum. To balance these two criteria and take the integer multiples value of the number of training classes, a value four times the number of training classes was used in the experiments. The optimal real accuracy and intersection are indicated by two dashed black lines, and the chosen prior constant $\mathcal{C}$ is indicated by dash-dot blue lines.

TABLE II
ESTIMATION RESULTS BASED ON DIFFERENT BASE RATES INITIAL STRATEGIES

Strategy	Real Accuracy	% Max Entropy
	MNIST / CIFAR10	MNIST / CIFAR10
Uniform Prior	**0.733 / 0.645**	**0.448 / 0.608**
Frequency Prior	0.733 / 0.645	0.448 / 0.608
Highest Frequency Prior	0.733 / 0.645	0.446 / 0.607
Lowest Frequency Prior	0.737 / 0.645	0.447 / 0.607

TABLE III
TEST ACCURACIES FOR MNIST AND CIFAR5 DATASETS

Method	MNIST	CIFAR5
CNN	0.994	0.764
EDL	0.993	0.843
SLUE	**0.997**	**0.843**

Events that can be repeated many times are typically frequentist in nature, meaning that base rates for such events typically can be derived from statistical observations [2]. Thus, for the initial base rates, there were four candidates:

(1) Use the uniform base rates (hereafter called uniform prior).
(2) Use each training class frequency as the initial base rates (hereafter called frequency prior).
(3) Assign the whole base rates to the training class that has the highest frequency (hereafter called highest frequency prior).
(4) Assign the whole base rates to the training class that has the lowest frequency (hereafter called lowest frequency prior).

Prior constant $\mathcal{C}$ is set to equal four times the number of training classes, then each strategy is verified in turn. The training and testing period operations are the same as those for choosing optimum prior constant $\mathcal{C}$. As can be seen from Tables II, there was not an obvious difference between the four; because the initial base rates is just initialization, the final classification is determined by all the base rates in every iteration step, and the initial base rates does not dominate. As a result, the easiest achieved strategy (uniform prior) was chosen for determining the initial base rates.

C. Synthetic dataset

First and foremost, SLUE accuracy performance was evaluated. Following the suggestion of [3], the CIFAR10 dataset was reduced by selecting the first five classes; it is referred to as CIFAR5. Since with CNNs method the model cannot calculate uncertainty for out-of-domain samples, to be fair, classical test accuracy was adopted instead of real accuracy. Table III demonstrates the main characteristics for comparing the SLUE method and the existing methods. The SLUE method achieved a match and better performance in test accuracy. Hence, the uncertainty estimation extensions of SLUE do not decline the model performance on in-domain sample classification.

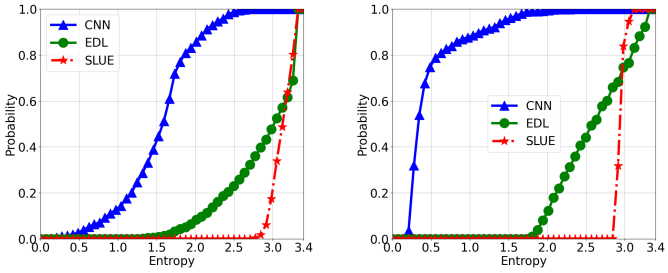
TABLE IV
COMPARISON BETWEEN DIFFERENT ESTIMATION METHODS. ↑ INDICATES LARGER VALUES ARE BETTER, BOLD NUMBERS ARE SUPERIOR RESULTS. ALL VALUES ARE PERCENTAGES AND ARE AVERAGED OVER THE FOUR OUT-OF-DOMAIN DATASETS DESCRIBED IN SECTION IV-C

D^{train}	Method	% max entropy ↑
Cifar10	EDL	0.78
	SLUE	**0.93**
Cifar100	EDL	0.76
	SLUE	**0.92**

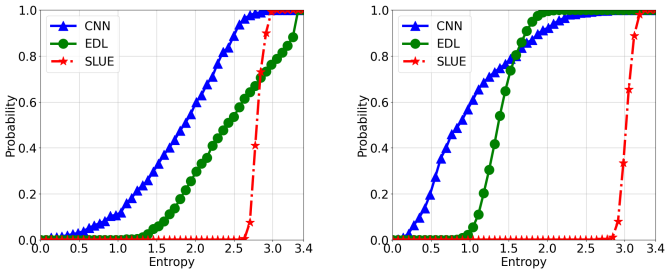
There is one more point that should be touched upon, the model uncertainty performance. The models were trained with the CIFAR10 and CIFAR100 datasets separately. The trained models were tested with the out-of-domain datasets (e.g., TEXTURE, PLACES365, LSUN, CIFAR10, and CIFAR100 datasets). To be homogeneous between testing datasets and training datasets, for TEXTURE, PLACES365, and CIFAR100 datasets, the first ten classes were extracted and then used for the experiments. The results of the prediction entropy CDF on the out-of-domain datasets are shown in Fig. 2. Since the predictions for these samples are almost wrong, predictions with maximum entropy are expected. A high start entropy value will be observed, and after this beginning point, there will be a dramatic increase in probability. Regarding the figure, the curves closer to the bottom right corner of the plot are wanted, which demonstrates maximum entropy in all predictions [3]. What is striking about the curves in these figures is that the SLUE method associates much more uncertainty with its predictions than other methods. Compared to the decentralized entropy distribution on the EDL method, for the SLUE method, it is more concentrated. This attribute makes it easier to distinguish out-of-domain samples. As Table. IV demonstrated that SLUE reached better uncertainty assessment performance 15% improvement in terms of % max entropy. It is apparent that the uncertainty estimates of the SLUE method are better than those of the baseline methods.

D. Adversarial dataset

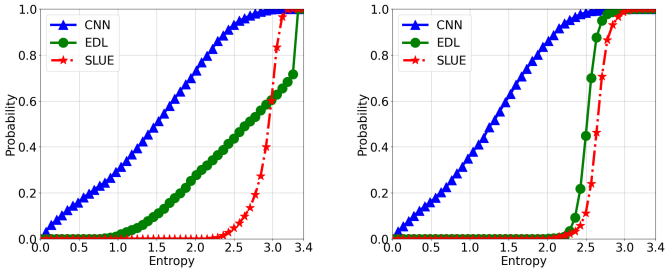
Last but not least, the different methods were also evaluated against adversarial samples [1], [3], [9]. Using the fast gradient sign method, adversarial MNIST and CIFAR10 datasets are generated. The feature is that the bigger is, the generated datasets were closer to the out-of-domain datasets. Because it becomes harder to make correct predictions, bigger % max entropy would be observed. Fig. 3 presents an overview of the performance of real accuracy and % max entropy for the SLUE method against adversarial datasets. These figures are quite revealing in several ways. First, the figure indicates that the SLUE method has the highest real accuracy for the adversarial datasets as shown in the left column of the figure. Second, with the comparable % max entropy on all of its predictions as indicated by the right column of the figure, the SLUE method can be used for the identification of out-of-domain samples. The SLUE method represents a good balance between prediction uncertainty and real accuracy criteria. It associates high uncertainty with the wrong predictions, which can be used to reject out-of-domain samples, improving model robustness.



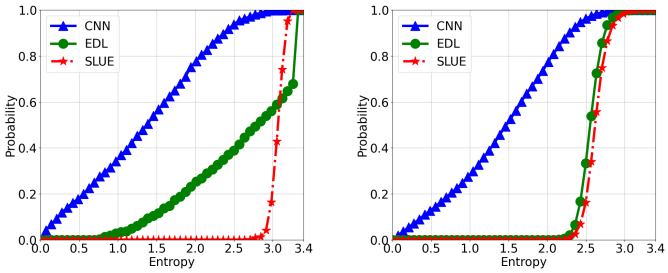
(a) TEXTURE dataset



(b) PLACES365 dataset



(c) LSUN dataset

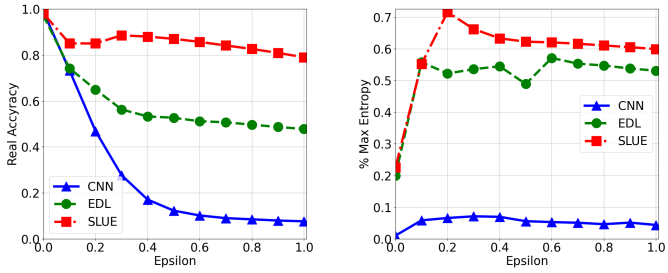


(d) CIFAR100 dataset (left) and CIFAR10 dataset (right)

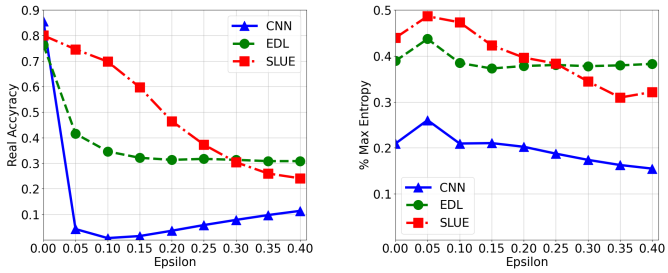
Fig. 2. Empirical CDF for the entropy of the predictive distributions on the out-of-domain datasets based on a model trained with CIFAR10 (left column) and CIFAR100 (right column) datasets

E. Batch size explanation

Since base rates update starts at the end of each batch, the effect of different batch sizes on uncertainty estimation was evaluated. We used different batch size settings (e.g., 20, 50, 100, 200, 500) to train models based on CIFAR10 and CIFAR100 separately, and used the SVHN and LSUN as out-of-domain datasets. As can be seen from Fig. 4 that the

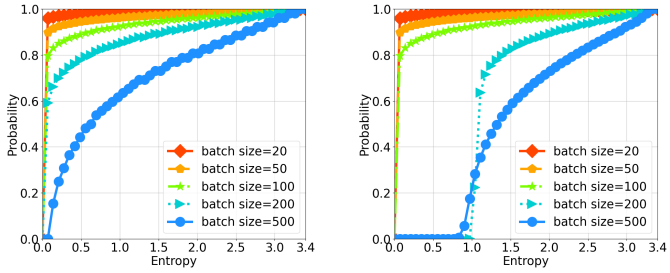


(a) MNIST dataset

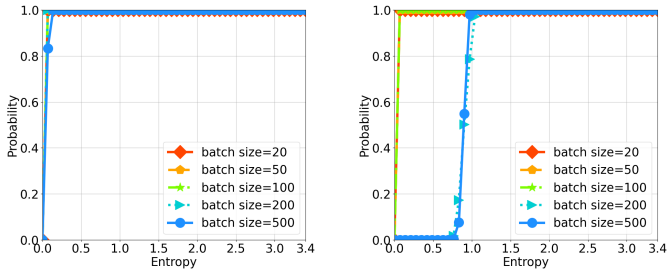


(b) CIFAR10 dataset

Fig. 3. Real accuracy and % max entropy as a function of the adversarial perturbation



(a) SVHN dataset



(b) LSUN dataset

Fig. 4. Empirical CDF for the entropy of different batch sizes on the out-of-domain datasets based on a model trained with CIFAR10 (left column) and CIFAR100 (right column) datasets

smaller the batch size is, the better the uncertainty estimation performance. It is reasonable because of the small batch size leading to a more precise update. On the other hand, the batch size does not need to be infinitely small. If it is very small, the training speed will be reduced, and it will also encounter overfitting. It is obvious when the batch size greater than 100, there is a significant performance gap. In comparison, there is no big difference when the batch size lower than 100. In conclusion, a batch size equal to 100 can meet the requirements.

V. Related Work

The problem that prediction lacks confidence has been examined by many researchers [1], [10], [11], and uncertainty has been used in the past to investigate out-of-domain detection in image segmentation [12]–[15] and image classification [16], [17].

Subjective logic Subjective logic [2] has been applied in a variety of ways, for example, in fraud detection [18], video segmentation [19], and image classification [20], [21]. Due to the unique nature of subjective logic, it can easily achieve bijection with Dirichlet distribution. The derived uncertainty score [1] has been proved to be effective in uncertainty estimation as well as out-of-domain detection. But its main problem is insufficient uses of subjective logic, especially the overlook of base rates. As a comparison, our work makes a good command of basic rates.

Uncertainty estimation Interest in assessing and quantifying uncertainty has increased in recent years, a reliable uncertainty estimation has been recognized as crucial for decision making. There are three main families that aim to provide meaningful uncertainty estimation. The first family are Bayesian Neural Networks [22]–[28], the disadvantage is that computation consumption. The second family is composed of Monte-Carlo drop-out based models [29]–[31] and ensembles [10]. Uncertainty estimation is achieved by computing statistics such as mean and variance. A disadvantage is that uncertainty estimation is fulfilled in the sacrifice of time consumption.

Out-of-domain detection Out-of-domain detection [17], [32]–[35] is a very trendy research field. There are various research methods, including generative-based and discriminative-based. The above methods have their own advantages and disadvantages, and detailed analysis can be obtained from these two surveys [36], [37]. Our research focuses on processing the model output to obtain a value that can distinguish the out-of-domain samples. The existing similar methods are softmax score [38], energy score [16], trust score [39], and response score [40]. They are all based on Bayesian probability theory, and the information that can be obtained is limited. The subjective logic used in this article is a generalization of the belief function, which can better extract more information from the input samples to make more accurate and reasonable predictions.

VI. Conclusion

The aim of the present research was to extend the existing EDL method by taking base rates into account. It verified the SLUE method uncertainty performance through out-of-domain and adversarial datasets. It introduced how to select the initial parameters and use real accuracy as one of the criteria. Meanwhile, we manifested the batch size effect on uncertainty estimation. From existed results, we can tell 100 is a sufficient

option for batch size setting. The most obvious finding to emerge from this study is that guided by base rates, the new SLUE method works better not only in terms of real accuracy but also on the uncertainty estimation performance. As it can reject out-of-domain samples, this approach will prove useful in improving model robustness. Although extensive research has been carried out, one issue is that the uncertainty delineation is still not complete. In this study, the uncertainty is calculated for the whole state space. Consequently, in the future, we are going to calculate uncertainty for all subsets of state space to provide more information for making predictions.

ACKNOWLEDGMENT

This work was funded by the Auvergne Rhône Alpes region: project AUDACE2018.

REFERENCES

[1] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 3183–3193.

[2] A. Jøsang, *Subjective Logic: A formalism for reasoning under uncertainty*. Springer Publishing Company, Incorporated, 2018.

[3] C. Louizos and M. Welling, "Multiplicative normalizing flows for variational bayesian neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 2218–2227.

[4] Y. LeCun, C. Cortes, and C. Burges, "Mnist handwritten digit database," *ATT Labs*, vol. 2, 2010.

[5] A. Krizhevsky, "Learning multiple layers of features from tiny images," 2009.

[6] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao, "LSUN: construction of a large-scale image dataset using deep learning with humans in the loop," *CoRR*, vol. abs/1506.03365, 2015.

[7] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2014, pp. 3606–3613.

[8] B. Zhou, A. Khosla, À. Lapedriza, A. Torralba, and A. Oliva, "Places: An image database for deep scene understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40(6), pp. 1452—1464, 2017.

[9] A. Malinin and M. Gales, "Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness," in *Advances in Neural Information Processing Systems*, 2019, pp. 14 547–14 558.

[10] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in neural information processing systems*, 2017, pp. 6402–6413.

[11] D. P. Kingma, T. Salimans, and M. Welling, "Variational dropout and the local reparameterization trick," *Advances in neural information processing systems*, vol. 28, pp. 2575–2583, 2015.

[12] P. Bogaert, F. Waldner, and P. Defourny, "An information-based criterion to measure pixel-level thematic uncertainty in land cover classifications," *Stochastic Environmental Research and Risk Assessment*, vol. 31, no. 9, pp. 2297–2312, 2017.

[13] K. M. Brown, G. M. Foody, and P. M. Atkinson, "Estimating per-pixel thematic uncertainty in remote sensing classifications," *International Journal of Remote Sensing*, vol. 30, no. 1, pp. 209–229, 2009.

[14] L. Loosvelt, J. Peters, H. Skriver, H. Lievens, F. M. B. Van Coillie, B. De Baets, and N. E. C. Verhoest, "Random forests as a tool for estimating uncertainty at pixel-level in sar image classification," *International Journal of Applied Earth Observation and Geoinformation*, vol. 19, pp. 173–184, 2012.

[15] P. Stone, M. Sridharan, D. Stronger, G. Kuhlmann, N. Kohl, P. Fidelman, and N. K. Jong, "From pixels to multi-robot decision-making: A study in uncertainty," *Robotics and Autonomous Systems*, vol. 54, no. 11, pp. 933–943, 2006.

[16] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," *Advances in Neural Information Processing Systems*, 2018.

[17] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," *Proceedings of the International Conference on Learning Representations*, 2019.

[18] D. S. Xing and M. Girolami, "Employing latent dirichlet allocation for fraud detection in telecommunications," *Pattern Recognition Letters*, vol. 28, no. 13, pp. 1727–1734, 2007.

[19] X. G. Wang, X. X. Ma, and W. E. L. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models," *Ieee Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 3, pp. 539–555, 2009.

[20] C. Chen, A. Zare, H. N. Trinh, G. O. Omotara, J. T. Cobb, and T. A. Lagaunne, "Partial membership latent dirichlet allocation for soft image segmentation," *Ieee Transactions on Image Processing*, vol. 26, no. 12, pp. 5590–5602, 2017.

[21] N. Rasiwasia and N. Vasconcelos, "Latent dirichlet allocation models for image classification," *Ieee Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2665–2679, 2013.

[22] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 37, 2015, pp. 1613–1622.

[23] K. Osawa, S. Swaroop, M. E. E. Khan, A. Jain, R. Eschenhagen, R. E. Turner, and R. Yokota, "Practical deep learning with bayesian principles," in *Advances in neural information processing systems*, 2019, pp. 4287–4299.

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

[25] C. E. Rasmussen, "Gaussian processes in machine learning," in *Summer School on Machine Learning*, 2003, pp. 63–71.

[26] Y. Gal and Z. Ghahramani, "Bayesian convolutional neural networks with bernoulli approximate variational inference," *arXiv preprint arXiv:1506.02158*, 2015.

[27] D. P. Kingma, T. Salimans, and M. Welling, "Variational dropout and the local reparameterization trick," in *Advances in Neural Information Processing Systems*, 2015, pp. 2575–2583.

[28] D. Molchanov, A. Ashukha, and D. Vetrov, "Variational dropout sparsifies deep neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, 2017, p. 2498–2507.

[29] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," vol. 48, 2016, pp. 1050–1059.

[30] A. G. Roy, S. Conjeti, N. Navab, and C. Wachinger, "Inherent brain segmentation quality control from fully convnet monte carlo sampling," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2018, pp. 664–672.

[31] A. G. Roy, S. Conjeti, N. Navab, C. Wachinger, A. D. N. Initiative *et al.*, "Bayesian quicknat: model uncertainty in deep whole-brain segmentation for structure-wise quality control," *NeuroImage*, vol. 195, pp. 11–22, 2019.

[32] N. Papernot and P. D. McDaniel, "Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning," 2018.

[33] J. Tack, S. Mo, J. Jeong, and J. Shin, "Csi: Novelty detection via contrastive learning on distributionally shifted instances," in *Advances in Neural Information Processing Systems*, 2020.

[34] J. Ren, P. Liu, E. Fertig, J. Snoek, R. Poplin, M. DePristo, J. Dillon, and B. Lakshminarayanan, "Likelihood ratios for out-of-distribution detection," 06 2019.

[35] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," 2015.

[36] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.

[37] A. Zimek, E. Schubert, and H.-P. Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data," vol. 5, no. 5, p. 363–387, 2012.

[38] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *Proceedings of International Conference on Learning Representations*, 2017.

[39] H. Jiang, B. Kim, and M. Gupta, "To trust or not to trust a classifier," 05 2018.

[40] L. Gao and S. Wu, "Response score of deep learning for out-of-distribution sample detection of medical images," *Journal of Biomedical Informatics*, vol. 107, p. 103442, 2020.