An Uncertainty Quantification Method Based on Evidence theory and Conformal Prediction

Rouaa HOBLOS[1], Noura DRIDI[1], Noureddine ZERHOUNI[1], Zeina AL MASRY[1]
[1] *SUPMICROTECH, CNRS, institut FEMTO-ST, 24 rue Alain Savary, 25000 Besançon, France*
{rouaa.hoblos, noura.dridi, noureddine.zerhouni, zeina.almasry}@femto-st.fr

***Abstract*—Conformal prediction (CP) provides a robust framework for quantifying the uncertainty of predictions for machine and deep learning. By leveraging the concept of prediction sets, it guarantees marginal coverage of the prediction to the true labels. In this paper, we develop a new method for quantifying uncertainty based on CP and the theory of evidence. The contribution includes a novel conformity score with a class probability modeled by Dirichlet distribution. Predictions of the neural network are interpreted as subjective opinions, and the network aggregates the evidence underlying these opinions. Moreover, using CP, the algorithm provides a set of predictions that can be empty if the model's confidence is lower than a fixed threshold. This would prevent the model from making uncertain predictions, thereby enhancing its overall reliability. Results on tabular and image datasets attest the superior performance of the proposed method compared to the state-of-the-art (SoA) in terms of robustness and coverage metrics as well as its effectiveness in detecting out-of-domain (OOD) data.**

***Index Terms*—uncertainty quantification, conformal prediction, conformity score, evidence, prediction set.**

I. Introduction

Recent advancements in Machine and Deep Learning (ML and DL) have improved the overall performance of various AI-driven systems. Therefore, applications covered large domains ranging from medical [1], [2] to renewable energy [3] to finance [4] and other fields. Moreover, to ensure efficient decision making, it is important to quantify the uncertainty associated to ML and DL predictions, especially when dealing with critical applications such as medical ones or autonomous vehicle control. Uncertainty Quantification (UQ) provides insights into the reliability of models' predictions. Several techniques have been developed to measure this uncertainty, including Bayesian Neural Networks (BNNs) [5], single-deterministic, and ensemble methods [6]. BNNs infer prediction uncertainty by modeling uncertainties in the Neural Network (NN) weights; the latter are represented using probability distributions instead of deterministic values. In [7], dropout is applied to estimate (NN) uncertainty based on the result that optimizing the loss function of a NN with dropout is equivalent to a Bayesian variational approximation of a Gaussian process. In ensemble method, the algorithm is trained multiple times, and the final prediction is obtained by averaging the predictions of the individual members [6].On the other hand, with single deterministic methods, the neural network's parameters are considered as deterministic. The prediction, as well as the uncertainty, are computed using one single forward pass on the network. In this category, several approaches for classification are based on Dirichlet distribution as a probability density function for class probabilities as in [8]. Moreover, in [9], the authors propose a model based on the theory of subjective logic [10]. The predictions of the NN are interpreted as subjective opinions. To implement this, they introduce a new loss function that aggregates the evidence supporting these opinions. A detailed review of the main approaches for UQ is given in [11]. The limit with the above cited approaches is the lack of guaranteed coverage, i.e the probability that the provided prediction covers the ground truth. Moreover, additional hypothesis related to the distribution is necessary. For example, in a BNN, the NN's weights are modelled as a Gaussian distribution. As an alternative, Conformal Prediction (CP), a distribution-free UQ framework, generates sets of possible labels for each input. An interesting property of CP is the guarantee of marginal coverage [12]. CP is largely applied to classification tasks [12]–[14] and is model-agnostic. In this paper, we propose a new method for classification with UQ based on CP. The idea is to integrate evidence in the chosen conformity score. The class probabilities are assumed to follow a Dirichlet distribution whose parameters are related to the evidence and the uncertainty. Two classifiers are then proposed based on LeNet [15] and ConvNet [16], respectively. The classifiers are integrated in a conformal prediction scheme to obtain the set of predictions. This framework provides a more robust measure of uncertainty by accounting the coverage probability.

The main contributions of this study are as follows:

- Propose an evidential conformity score.
- Use reliable metrics including robustness and coverage to attest the proposed method performance.
- Generate an empty set when the algorithm is unsure of its prediction.

A similar work based on Evidential Deep Learning and CP was done in [17]. The authors evaluate their method using only coverage evaluation metrics without considering the point-prediction performance.

The rest of the paper is organised as follows. In II, the proposed method is detailed. Section III is dedicated to the evaluation metrics used to attest the performance of the UQ. Results are provided in section IV. Finally, a conclusion with some perspectives are provided.

II. Evidential Conformal Prediction for Classification with Uncertainty Quantification

We propose a new method for uncertainty quantification based on conformal prediction called Evidential Conformal Prediction for Classification (ECPC). Before presenting the ECPC, we recall the idea of CP.

A. Preliminaries: Conformal Prediction

Consider n data samples $\{(X_i, Y_i)\}_{i=1}^{n}$, where X_i is the input and Y_i is the output. Y_i belongs to a set of discrete labels $Y_i \in \mathcal{Y} = \{1, 2, ..., K\}$. We use a neural network parameterised by θ to predict the output Y_i, given by $Y_i = f_\theta(X_i)$. The algorithm of CP is resumed as follows:

1) Split the data into three sets: a training set, a calibration set of size n_{cal} unused for training, and a test set to evaluate the performance of size n_{test}.
2) Train a classifier $\hat{f}$ on the training set.
3) Evaluate the model on the calibration set to get the predictions on the calibration set using f_θ.
4) Fix a conformity score for data noted cs_i, with i being the input index. The conformity score measures how familiar the data is to the model. We choose our conformity score to be $cs_i = 1 - \hat{f}(X_i)_{Y_i}$ then compute it for all data in the calibration set: $S_{cal} = \{cs_i, i \in \{1, ..., n_{cal}\}\}$
5) Compute the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile noted $\hat{q}$ where α is a user-fixed error rate.
6) Construct prediction sets for data in the test set. For a new test data point $(X_i, Y_i), i \in \{1, ..., n_{test}\}$, the prediction set equals:

$$C(X_\text{i}) = \{y : cs(X_\text{i}, y) \leq \hat{q}\}. \tag{1}$$

An important result of CP is to guarantee the marginal coverage [18]. The theorem is given below.

Theorem 1. *(Conformal Coverage Guarantee [18]) Suppose $(X_i, Y_i)_{i=1,...,n_{cal}}$ and (X_{test}, Y_{test}) are exchangeable (i.e. samples' order doesn't effect results). For a new data in the test set (X_t, Y_t), the following holds :*

$$P(Y_t \in C(X_t)) \geq 1 - \alpha \tag{2}$$

where α is a user-fixed error rate.

B. Evidential Conformal Prediction for Classification (ECPC)

The main idea of ECPC is to integrate evidence in the conformal score to generate "evidential" prediction sets. To do so, we propose using Deep Evidential Classification (DEC) algorithm [9] as the classifier to CP. As shown in Figure 1, the input data are used to learn the parameters of the NN. Unlike [9], different architectures are chosen to test the model (more details are provided in the Results section). Moreover, instead of using a softmax function in the NN, an evidence function is applied to the logits z_k so that the NN outputs positive evidence e_k for K classes instead of probabilities, and predictions are represented as subjective opinions. The evidence function is given by: $e_k = \exp(\min(\max(z_k, -c), c))$

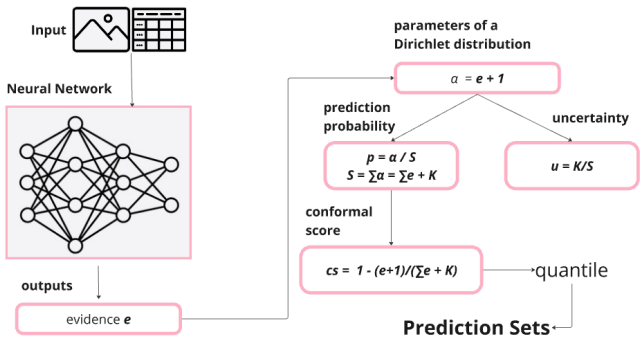

Fig. 1: Different steps of the ECPC method.

where c is a fixed positive constant. Note that other functions can be used like ReLU...

The evidence is the degree of belief (or mass) assigned to different possible outcomes. The higher it is, the more certain the model is. This evidence is used in the calculation of the parameters α_k of a Dirichlet distribution: $\alpha_k = e_k + 1$. By doing so, the prediction probability would be taking into consideration the uncertainty u_k as it is based on evidence: $u_k = \frac{K}{S}$ and $S = \sum_{k=1}^{K} \alpha_k = K + \sum_{k=1}^{K} e_k$. The lack of evidence would allow ECPC to give empty prediction sets when it is unsure. This is very important as it improves ECPC's reliability with respect to other methods where they always give predictions even when unsure. The prediction probability is given by the mean of the Dirichlet distribution:

$$\hat{p}_k = \frac{\alpha_k}{S} \tag{3}$$

where k is the label index. This probability can be written in terms of the evidence, we get:

$$\hat{p}_k = \frac{e_k + 1}{K + \sum_{j=1}^{K} e_k}. \tag{4}$$

An important step in CP approaches is the choice of conformity score (Step 4 in Section II-A). It is critical as with bad scores, prediction sets may be useless [12]; a prediction set may be large enough to not offer real help in decision making. We propose an evidential conformity score given by:

$$1 - \hat{p}(X_i)_{Y_i}, \tag{5}$$

where $\hat{p}(X_i) = (\hat{p_1}, ..., \hat{p_K})$ is the vector of the predicted probabilities for data sample i. $\hat{p}(X_i)_{Y_i}$ corresponds to the probability of the true class. This score is also called the Hinge Loss [19]. Substituting equation (4) in (5), we write the conformal score in terms of evidence making it an evidential conformity score:

$$cs_i = \frac{K - 1 - e_i + \sum_{j=1}^{K} e_j}{K + \sum_{j=1}^{K} e_j}. \tag{6}$$

This score provides a measure of conformity or "strangeness" of a data-point to the model. The higher the evidence is, the more conformal a data-point would be to the model. The validity of coverage guarantees in conformal prediction is

distribution-free and holds irrespective of the specific conformity score used, thereby justifying the validity of our evidential conformity score.

III. Evaluation Metrics

To assess the performance of ECPC, we employ metrics for single predictions and prediction sets for classification. Two categories of metrics are considered: robustness and coverage.

A. Robustness metrics

We used two metrics to assess the robustness of the method: Shannon Entropy and Brier Score.

Shannon Entropy measures the amount of randomness or uncertainty in the outcomes. The closer the entropy is to zero, the more certain the outcome is. It is defined as:

$$H(P) = -\sum_{k=1}^{K} p_k \log(p_k). \tag{7}$$

where p_k is the predicted probability for the k^{th} label. The entropy is calculated per data-point (to simplify the notation, the index of the data-point is omitted). We propose to relate the class probabilities given in Eq(3) with the entropy. The latter can easily be written as

$$H(P) = -\frac{1}{S}[\sum_{k} \alpha_k log(\alpha_k) - SlogS]. \tag{8}$$

The derivative is given by:

$$\frac{dH}{d\alpha_k} = -\frac{1}{S}[log\alpha_k + 1] \leq 0. \tag{9}$$

Hence, entropy is a decreasing function of α_k. Moreover, α_k is related to the evidence by $\alpha_k = e_k + 1$. Therefore when the evidence increases, α_k increases and entropy decreases leading to a more confident prediction.

Brier score (BS) measures how accurate the predictions are by calculating the error between the predicted probabilities and the actual outcomes. It is given by

$$\text{BS} = \frac{1}{K}\sum_{k=1}^{K}(p_k - o_k)^2 \tag{10}$$

where p_k is the predicted probability of label k and o_k is the true outcome. The lower the Brier score is, the more certain the predictions are.

B. Coverage metrics

Unlike UQ methods with single output predictions, conformal predictors provide a set of predictions. The proposed method is based on CP, therefore, coverage metrics [19] are used to attest the performance.

Mean Prediction Set Width (MPSW). The width of a prediction set for classification is given by the number of classes in the set as follows

$$MPSW = \frac{1}{n}\sum_{i=1}^{n} W_i \tag{11}$$

where n is the total number of data-points and W_i is the width of the prediction set of data-point i. The tighter the MPSW, the higher the confidence in the prediction. **OneC** represents the percentage of singletons in the prediction sets:

$$\text{OneC} = \frac{1}{n}\sum_{i=1}^{n} sing_i, \quad sing_i = \begin{cases} 1, & |\mathcal{C}(X_i)| = 1 \\ 0, & |\mathcal{C}(X_i)| \neq 1 \end{cases}$$

where $C(X_i)$ is the prediction set corresponding to data-point i. $|C(X_i)|$ denote the cardinal of the set $C(X_i)$. High OneC corresponds to a more certain model.

p-value measures how different a data-point is from the data sample in the calibration set based on the conformity score.

$$p_i = \frac{|cs_j \geq cs_i| + 1}{n_{cal} + 1} \tag{12}$$

where cs_i and cs_j represent respectively the conformity score of the data-point i in test set and j in calibration one, $cs_j \in S_{cal}$. A p-value close to 0 means that the data-point is highly unusual compared to calibration data.

Prediction Set Coverage Probability (PSCP) represents the percentage of times the true value is included in the prediction set.

$$\text{PSCP} = \frac{1}{n}\sum_{i=1}^{n} c_i, \quad c_i = \begin{cases} 1, & y_i \in C(X_i) \\ 0, & y_i \notin C(X_i) \end{cases}$$

For single output prediction methods, the PSCP is equivalent to the accuracy of classification.

IV. Results

Two datasets are used: MNIST and Titanic. Details about the neural networks' hyperparameters and the noise are provided in table I. LeNet and ConvNet architectures are used [15], [16]. The dataset split is fixed across runs, resulting in negligible variation between repeated experiments; therefore, results are reported for a single representative run.

ECPC performance. Table II shows the different coverage metrics used to evaluate the ECPC method. As expected, MPSW increases on noisy and OOD data of Titanic dataset using both NN architectures. The highest value is for the OOD dataset, illustrating the ability of the method to detect the distribution shift. However, it is not the case for MNIST. MPSW decreases as the data becomes stranger to the model. This is due to the fact that on MNIST data, ECPC generates empty sets when images are ambiguous. This actually shows the efficiency of the method by returning "I don't know" as a response instead of wrong classification. For example, using the LeNet architecture, the percentage of empty sets for training is 0.17% and for testing is 0.23 %. This increases on noisy and OOD data to 27.83 and 23.76 % respectively. This means the stranger the data is to the model, the harder it gets for it to make predictions. Indeed, the percentage of empty sets on noisy data is higher than that of OOD as the added noise is large enough to make images so blurry, they become stranger to model than the OOD data. Conversely to MPSW, PSCP decreases as the data gets unfamiliar to the model. The PSCP

Dataset	Format	Total Size	Train	Calibration	Test	OOD Dataset	Epochs
MNIST	Images	70,000	48,000	12,000	10,000	FMNIST	100
Titanic	Tabular	712	455	114	143	Forestfires	400
Noisy Data	Gaussian noise applied on test data						
MNIST	Mean = 2 & Std = 4						
Titanic	Different means and stds on 3 out of 7 features						
CP Coverage Parameter	MNIST: $\alpha = 0.01$			Titanic: $\alpha = 0.1$			

TABLE I: Dataset splits, noisy data settings, CP coverage parameters, and number of epochs for MNIST and Titanic datasets.

Metric	MNIST								Titanic							
	Train		Test		Noisy		OOD		Train		Test		Noisy		OOD	
	LeNet	ConvNet	LeNet	ConvNet	LeNet	ConvNet	LeNet	ConvNet	LeNet	ConvNet	LeNet	ConvNet	LeNet	ConvNet	LeNet	ConvNet
MPSW	0.998	0.79	0.999	0.79	0.73	0	0.83	0.31	1.43	1.6	1.5	1.65	1.48	1.66	1.73	2
OneC	99.82	78.93	99.57	78.95	71.64	0	69.39	25.16	56.92	39.34	49.65	34.96	52.45	33.57	26.69	0
p-value	0.81	0.53	0.82	0.53	0.0069	0.22	0.0063	0.17	0.53	0.55	0.5	0.48	0.5	0.46	0.33	0.25
PSCP	99.78	88.73	99.29	88.58	10.48	10.53	–	–	95.82	98.46	94.4	93.71	93.71	91.6	–	–

TABLE II: Performance metrics for ECPC for MNIST and Titanic datasets using LeNet and ConvNet architectures.

of the OOD data can't be calculated because the ground truth is unknown. As for the percentage of singletons, represented by OneC, it is clear that ECPC is predicting wider sets as data is getting different from the training set in the Titanic dataset. This means, it is more certain of its prediction on in-distribution data. Besides, for the OOD data, ECPC is less certain so it is lost between more choices to predict. As for the MNIST dataset, OneC indeed decreases. This is coherent with the increase in the number of empty sets. The p-value given in table II is the mean computed over all the data-points.

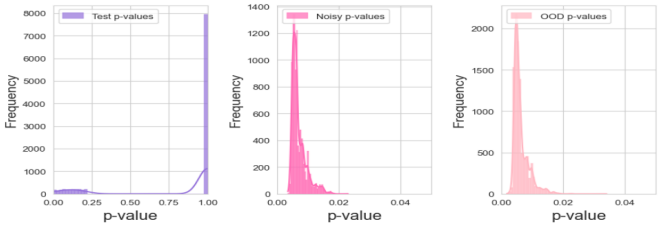

Fig. 2: Histogram of the p-values on MNIST test data.

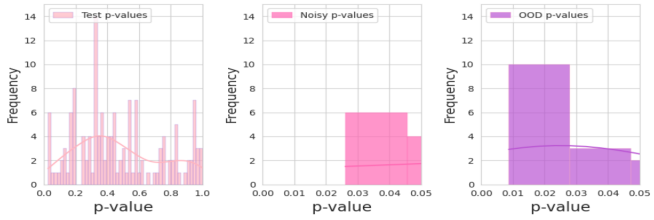

Fig. 3: Histogram of the p-values on Titanic test data

p-value for Noisy and OOD. To illustrate the behaviour on individual data-points, we propose the histograms of the p-values in Figures 2 and 3 respectivelly for MNIST and Titanic datasets. In Figure 2, we observe that both for noisy and OOD data, the p-values are close to zero. This reveals a high degree of non-conformity with the training data. This result illustrates that the ECPC method successfully recognizes when the input data varies significantly from what it has learned, leading to higher uncertainty on the predictions. On the other hand, for MNIST test data, the p-values approach 1, reflecting that the model is well-acquainted with this data and permitting it to provide more certain forecasts.

Figure 3 shows that for the test dataset, p-values range between 0 and 1. Meanwhile, for noisy and OOD datasets, p-values are between 0 and 0.05. This reflects the uncertainty of the model on strange data.

Comparison with the SoA. In Table III, a comparative analysis of the performance of ECPC on MNIST data using the LeNet architecture with the following approaches is proposed: 1) Dropout applied to Convolutional Neural Network (CNN-Dropout) proposed in [7], 2) Conformal Prediction with Logistic Regression as a classifier (LR-CP), 3) Bayesian Neural Networks (BNN) proposed in [20]. Results are compared for MNIST test data-with and without noise-and FMNIST as an OOD data. The BS can't be calculated on the latter dataset since the ground truth is unknown. Likewise, coverage metrics, in table II, cannot be calculated for CNN-Dropout and BNN as they are single-prediction classifiers unlike ECPC and LR-CP which generate prediction sets.

Metric	Dataset	CNN-Dropout	BNN	LR-CP	ECPC
PSCP / Accuracy	Test	86.4	95.9	99.1	**99.24**
Brier Score (BS)	Test	0.72	0.07	0.12	**0.0154**
	Noisy	0.9	0.09	**1.54**	0.9
Entropy	Test	2.25	0.11	0.28	**0.067**
	Noisy	**2.29**	0.232	0.58	2.26
	FMNIST	**2.27**	0.73	1.33	2.24
MPSW	Test	–	–	2.5	**1**
	Noisy	–	–	4.12	**0.73**

TABLE III: Comparison with SoA.

Table III shows that ECPC achieves a superior performance in quantifying uncertainty overall as it has the lowest entropy and Brier score for the test data. Notably, it correctly detects OOD data, as illustrated by the comparison of entropy between MNIST test data and FMNIST data: the difference of entropies for ECPC is 2.14, but it is only 0.02 for the CNN-Dropout. This means that CNN-dropout isn't able to detect well OOD data, meanwhile, ECPC exhibits enhanced sensitivity to distributional shifts. Additionally, it is noted that the differences between entropies of test and noisy data is significantly higher for ECPC (2.193) than LR-CP (0.3) and BNN (0.122). Furthermore, ECPC offers a distinct advantage: MPSW in ECPC is nearly 1, significantly shorter than that in LR-CP which is 2.5. This advantage indicates a more concise and informative prediction.

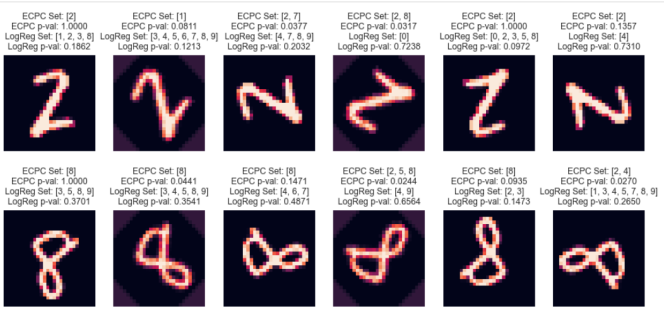

Fig. 4: Single predictions, prediction sets, and p-values of 2 random rotated images from the MNIST dataset

Data with Rotation Figure 4 shows the prediction sets by ECPC and LR-CP on two rotated images with rotation angles $\{0, 45, 90, 135, 180, 270\}$ from MNIST alongside their associated p-values. For the original image of digit 2, ECPC predicts a singleton of the true label 2 with a maximum p-value of 1, indicating high confidence. In contrast, LR-CP produces a set of length 4, accompanied by a low p-value, indicating less confidence. This behaviour is similarly observed with the second un-rotated image of digit 8.

For a rotation of 90°(third column), ECPC predicts a set containing the correct label (digit 2), but with a lower p-value, indicating increased uncertainty. On the other hand, LR-CP fails to cover the true label. For the image of digit 8, despite its horizontal orientation, ECPC predicts a singleton of the true label, while LR-CP produces a 5-elements set. Comparing p-values, for both digits, shows that LR-CP is more confident despite its wrong classification.With 135° rotation (fourth column) of the first image of digit 2, ECPC is uncertain but managed to generate a set of two labels, one of which is correct. The method's uncertainty is expected given the high angle of rotation. However, LR-CP confidently predicts a singleton of label 0. The same conclusion is drawn for the image of digit 8.

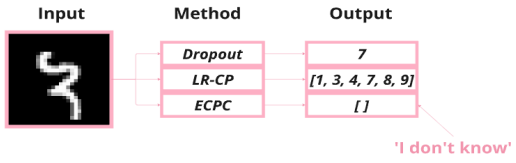

Fig. 5: Image of label 3 with an empty prediction set.

Empty Set. A final illustration of the performance of ECPC is presented in Figure 5. The image appears ambiguous or unclear. This unclarity likely complicates accurately identifying the correct labels. ECPC method has provided an empty prediction set. Instead of providing a potentially incorrect prediction, the algorithm indicates high uncertainty by outputting a response of 'I don't know'. On the other hand, LR-CP generates a prediction set of size 6 and CNN-Dropout predicts a wrong label of 7 .

V. Conclusion

In this paper, ECPC, a new method for quantifying uncertainty in classification tasks, is proposed. The method is based on conformal prediction to generate reliable uncertainty estimates. ECPC is compared to the SoA. Its relevance is highlighted in several experimental setups such as different data formats, architectures, and unfamiliarity. The possibility of predicting an empty set is attained, allowing the algorithm to say 'I don't know' when it lacks enough evidence. Future work will be focused on the coverage in conformal prediction and on dealing with real-world applications.

References

[1] A. Agha, M. I. Imran, F. de Castro, S. Tabrez, V. Vijayabhaskar, R. Ranju, and H. Harpreet, "A review of deep neural network-based uncertainty quantification methods for the classification of breast cancer," *NeuroQuantology*, vol. 20, no. 10, pp. 9702–9715, 2022.

[2] L. T. Pflieger, C. C. Mason, and J. C. Facelli, "Uncertainty quantification in breast cancer risk prediction models using self-reported family health history," *Journal of Clinical and Translational Science*, vol. 1, no. 1, pp. 53–59, 2017.

[3] T. Fu, C. Wang, and N. Cheng, "Deep-learning-based joint optimization of renewable energy storage and routing in vehicular energy network," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6229–6241, 2020.

[4] T. Blasco, J. S. Sánchez, and V. García, "A survey on uncertainty quantification in deep learning for financial time series prediction," *Neurocomputing*, vol. 576, pp. 127339, 2024.

[5] R. M. Neal, *Bayesian Learning for Neural Networks*, Springer-Verlag New York, Inc., 1996.

[6] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[7] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of the International Conference on Machine Learning*, 2016, pp. 1050–1059.

[8] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[9] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[10] A. Jsang, *Subjective Logic: A Formalism for Reasoning Under Uncertainty*, Springer Publishing Company, Incorporated, 2016.

[11] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, et al., "A survey of uncertainty in deep neural networks," *Artificial Intelligence Review*, vol. 56, no. Suppl 1, pp. 1513–1589, 2023.

[12] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," *arXiv preprint arXiv:2107.07511*, 2021.

[13] A. Podkopaev and A. Ramdas, "Distribution-free uncertainty quantification for classification under label shift," in *Uncertainty in Artificial Intelligence*. PMLR, 2021, pp. 844–853.

[14] J. Lei and L. Wasserman, "Distribution-free prediction bands for non-parametric regression," *Quality Control and Applied Statistics*, vol. 60, no. 1, pp. 109–110, 2015.

[15] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio, "Object recognition with gradient-based learning," in *Shape, Contour and Grouping in Computer Vision*, 1999.

[16] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," pp. 319–345, 1998.

[17] H. Karimi and R. Samavi, "Evidential uncertainty sets in deep classifiers using conformal prediction," 2024.

[18] V. Vovk, A. Gammerman, and C. Saunders, "Machine-learning applications of algorithmic randomness," 1999.

[19] V. Manokhin, *Practical Guide to Applied Conformal Prediction in Python: Learn and Apply the Best Uncertainty Frameworks to Your Industry Applications*, Packt Publishing Ltd., 2023.

[20] N. Dridi, L. Boffelli, and Z. Al Masry, "Uncertainty quantification using bayesian neural networks," in *Workshop on Frontiers of Uncertainty Quantification*, 2024.