Resilient Wireless Communications with Selective Deep Neural Network Classification

A. Q. M. Sazzad Sayyed*, Jonathan Ashdown‡, Michael De Lucia†,
Ananthram Swami†, Nathaniel D. Bastian° and Francesco Restuccia*

* Northeastern University, United States; ‡Air Force Research Laboratory, United States
†DEVCOM Army Research Laboratory, United States °United States Military Academy, United States

***Abstract*—The pervasive presence of deep neural networks (DNNs) has made artificial intelligence (AI) an integral part of tactical communication and networking systems. However, the overconfidence of modern DNNs poses significant risks in mission-critical applications. Selective Classification (SC) is a promising paradigm to mitigate this issue by enabling DNNs to abstain from unreliable predictions. Although there is a plethora of work on SC in the computer vision (CV) domain, there is no evaluation of the utility of these approaches in tactical communication scenarios. This work evaluates State-of-the-art (SOTA) SC methods within the tactical domain, focusing on their applicability to Automatic Modulation Classification (AMC) and Network Intrusion Detection System (NIDS) tasks using the RadioML and ACI IoT datasets, respectively. Our key findings show a 55% risk reduction for AMC with a 20% loss in coverage and near-zero risk for NIDS with minimal coverage loss. Furthermore, we analyze SC performance under distribution shifts, revealing limitations of traditional methods in handling covariate and semantic shifts. Finally, we explore training strategies to enhance SC performance.**

I. Introduction

The rapid advancements in AI have revolutionized numerous fields, from healthcare [1] to industrial automation [2]. Military communication and tactical systems are no exception to this trend [3], [4]. In particular, the integration of DNNs in areas such as automated modulation classification (AMC) and network intrusion detection systems (NIDS) has significantly enhanced capabilities for intelligence gathering and securing network infrastructures. However, these advancements come with challenges related to the reliability of DNN inference. In mission-critical applications, it is essential for DNNs to abstain from making predictions when confidence is low. Unfortunately, modern DNNs often display overconfidence in their predictions, even when incorrect [5]. This overconfidence undermines their trustworthiness and raises concerns about their deployment in high-stakes tactical scenarios.

Various approaches have been developed under the umbrella term of *SC*. The main approaches attempt to train DNNs to recognize and act upon their uncertainty. SC improves the resilience and reliability of the DNN by abstaining from giving uncertain or under-confident inferences. This reduces the error rate but comes at the cost of *coverage*, defined as the percent of inputs the DNN provides inference for. Approaches in this field has been developed with CV domain in focus. Prior work studied an abstention option into the DNN, either by modifying the training algorithm or the DNN architecture [6]–[9]. Another line of research addresses overconfidence through *post hoc* adjustments [10]–[13]. *To date, no study has explored the limitations of SC methods within this critical area.*

Approved for Public Release: Distribution Unlimited: AFRL-2025-0195.

In this work, we evaluate SOTA SC approaches in AMC and NIDS. Specifically, we seek to answer the following questions: (i) How much risk can SC reduce within a given coverage budget? (ii) Is SC effective under distribution shifts? (iii) Is there a particular training mechanism that can make DNNs selective by nature? To answer the questions posed above, we conduct a systematic study on a radio fingerprinting dataset *RadioML* [14] and a NIDS dataset *ACI IoT* [15]. In summary, we make the following contributions:

- We propose a framework for a tactical communication system enhanced by human feedback to improve reliability. The framework incorporates a *selective classification* module, which determines whether to defer decision-making to a human expert or proceed with the automated decision. We show that irrespective of the effectiveness of the human decision, SC can reduce risk for RadioML by 55% for a loss of coverage of only 20%. For ACI IoT dataset, the risk can be brought down from 2.5% to 0.05% for as little as 5% loss in coverage;
- We show that SC cannot handle distribution shifts. With traditional SC, only about 30% samples can be detected while about 80% can be detected with approaches focusing on detecting distribution shifts. This indicates that SC cannot address resiliency;
- We investigate different training strategies which have proved to be successful in vision domain for improving SC performance. We find that training with entropy regularization and stochastic weight averaging is conducive to SC and it provides 1% improvement in SC for the RadioML dataset and VGG8 architecture.

II. Background and Related Work

A. The Problem of Selective Classification

Let X represent the feature space and Y the label space, where X could denote the distribution of input I/Q samples or network traffic features, and Y the corresponding class labels. The aim is to learn the conditional distribution $P(Y \mid X)$, with a prediction model $f(\cdot;\theta) : X \rightarrow Y$,

parameterized by θ (we will omit the θ in future references to f unless its explicit mention is necessary). The risk of the task, evaluated with a loss function $\ell(\cdot)$, is expressed as:

$$\mathbb{E}_{P(X,Y)}[\ell(f(x;\theta), y)] \ .$$

A prediction model with a rejection option is defined by two functions (f, g), where $g_\tau : X \to \mathbb{R}$ acts as a selection function, serving as a binary qualifier for f:

$$(f,g)(x) = \begin{cases} f(x) & \text{if } g_\tau(x) \geq \tau, \\ \text{not sure} & \text{otherwise.} \end{cases}$$

In this framework, the model refrains from making a prediction when the value of the selection function $g(x)$, also known as the confidence score, falls below a predefined threshold τ. Different methods use varying forms of $g(x)$ to quantify uncertainty. The *covered* dataset is defined as:

$$\{x \mid g_\tau(x) \geq \tau\},$$

and the coverage is the ratio of the size of the covered dataset to the total dataset. This setup allows for a trade-off between coverage and risk, which motivates rejection option methods. These methods provide a way for the model to abstain from making predictions when the uncertainty is too high, thus improving reliability.

B. Existing Work on Selective Classification

Softmax Probability Based Approaches. The work in [16] first proposed to use softmax probability to quantify the confidence of DNNs and detect misclassified samples. Geifman et al. [17] developed a procedure for determining the threshold for attaining a target risk with a theoretical guarantee. Later work [18], [19] has further investigated ways to improve the softmax response of the DNNs via robust training. Feng et al. [18] proposed to use the SC mechanisms suggested in [6], [8], [9] but discards the modifications to the DNN architecture after training. After that, the softmax probability would be used for quantifying the confidence of the DNN. Zhu et al. [19] argue that the overconfidence issue is connected to the convergence of the DNN. They propose to use stochastic weight averaging [20] and sharpness-aware minimization [21] to achieve a flatter minimum.

Confidence Calibration Based Approaches. One influential baseline for modern DNN calibration is temperature scaling [22]. This post hoc method multiplies logits by a scalar temperature parameter to adjust confidence scores without altering model accuracy. Temperature scaling demonstrates significant improvements in calibration across multiple datasets and architectures, serving as a lightweight yet effective solution for classification tasks. Kull et al. [23] extended temperature scaling by modeling logits as a Dirichlet distribution, enabling better calibration for imbalanced datasets. Gupta et al. [24] fit a piecewise spline function to adjust confidence scores, allowing more flexibility than linear methods like temperature scaling. Cattelan et al. [13] introduced logit normalization with tunable parameters to improve calibration across diverse architectures.

Training Approaches. Prior work modifies the DNN architecture and/or the training mechanism. Geifman et al. [6] proposed a three-headed DNN consisting of a selection head, prediction head, and an auxiliary prediction head. The work in [8] and [9] proposed a $C + 1$ way classifier where the C is the number of classes and the additional logit is used to decide on the abstention decision. These two approaches differ in the training mechanism. While [8] proposes to change the weights of the samples dynamically during training, [9] proposes to use portfolio theory [25]. The work in [7] estimates model confidence by training a separate confidence estimator which is then used for abstention decision.

Bayesian Uncertainty Based Approaches. Bayesian deep learning provides the foundation for uncertainty modeling in DNNs, but it faces challenges due to the large number of parameters, making traditional Bayesian inference computationally intractable. Two prominent methods for estimating uncertainty are Monte Carlo Dropout (MC Dropout) [26] and Deep Ensembles [27]. MC Dropout adds dropout layers to the network, using multiple forward passes at inference to approximate uncertainty. Deep Ensembles train multiple models with different initializations or data splits, aggregating their predictions for uncertainty estimation. While both approaches are effective, they struggle with real-time performance and scalability. A recently growing line of work focuses on using Dirichlet distributions in training DNNs [28]–[30].

This work. In this paper, we focus on *post hoc* confidence calibration and training-based approaches as these two have proved to be the most successful in the CV domain. Uncertainty based approaches are too computation-expensive to be practical in dynamic tactical scenarios.

III. Methodology and Experimental results

Figure 1 shows the framework we consider in this work. The input data traffic being received is passing through a DNN trained to extract specific intelligence from the raw input. To improve the reliability, instead of relying directly on the output of the DNN, we place a selective classification module and a human authority in between. If the SC determines the output of the DNN be reliable enough, it passes the output for necessary action. Otherwise the output is passed to the human authority. The authority, based on the output of the SC and additional contextual information not available to the classifier, decides whether to ignore the prediction or to pass to a more capable DNN or a human expert. This way, we create a human-in-the-loop system striking a balance between the automation from the DNN and the expertise provided by a human. The

most crucial part of this framework is the *SC*. We analyze different SOTA designs of the SC to find the answers to the questions posed in Section I.

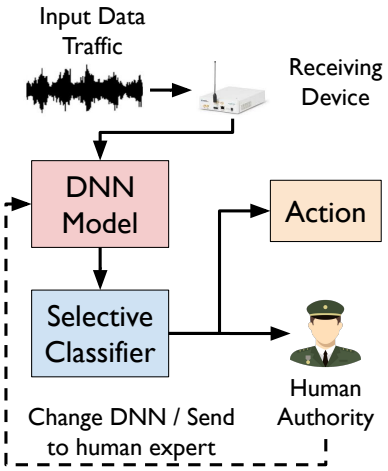


Fig. 1: Framework for reliable tactical communication.

A. Experimental Setup

Datasets. To analyze the SC methods, we consider two different tasks of military importance - AMC and NIDS. For the AMC task, we utilized the RadioML 2018.01A open-source dataset [14], which contains labeled data for 24 different analog and digital modulation schemes across various propagation scenarios. The dataset includes over-the-air I/Q samples transmitted via the B210 universal software radio peripheral (USRP), employing the Analog Devices AD9361 as the RF front-end. Modulations were captured under varying signal-to-noise ratios (SNRs), ranging from -20 dB to 30 dB. For the NIDS task, we employ the ACI IoT [15] dataset. It consists of 12 different classes of attacks on IoT network and 84 distinct features. This dataset simulates a realistic scenario reflecting the dynamic behavior of the connected devices.

DNN architectures: For each dataset, we investigate two different network architectures to study the relation between the network complexity and the SC. We use a custom 5 layered Convolutional Neural Network (CNN) (for both datasets, we refer to it as CNN5), a 1D variant of the VGG architecture consisting of 8 layers (for RadioML dataset), and a three layered Multi-layer Perceptron (MLP) (for ACI IoT datsaet).

Training. We train the DNNs for 30 epochs with Adam optimizer using learning rate of 0.01. We decrease the learning rate to $\frac{1}{10}$-th at 15-th and 25-th epochs. We set the batch size to 256. We refer to this as the vanilla training approach. As SNRs below 0 dB are of limited practical utility, we only train with data samples from SNRs above or equal to 0 dB for the RadioML dataset. For the ACI IoT dataset, we drop the classes *ARP Spoofing* and *UDP Flood* as these classes of attack have negligible number of samples (5 and 791 respectively) as compared to the other classes (rest of the classes have minimum of 6,000 samples).

SC approaches. To analyze the utility of the SC, we investigate a wide array of approaches - *trust score* [31], *temperature scaling* [22], *confidnet* [7], *p-Norm* [13], *latent heteroscedastic classifiers (HET)* [32], *adaptive temperature scaling (ATS)* [33]. Apart from these approaches, we also investigate the utility of *energy* [34] and *generalized entropy (GEN)* [35] for out-of-distribution detection.

Evaluation Metrics. When evaluating the performance of these approaches, we use the following metrics:

- *Area Under Receiver Operating Characteristic (AUROC)*: Measures the ability of the SC to distinguish between correct and incorrect predictions across all possible decision thresholds. It reflects the trade-off between correctly detecting incorrect inference and incorrectly identifying correct predictions as incorrect ones. A higher value of AUROC denotes high capability of separating correct and incorrect predictions.
- *False Positive rate at x% True Positive rate (FPR)*: Represents the rate of false positives (correct inferences attributed as incorrect) when the DNN achieves a true positive rate (correctly detecting incorrect samples) of x%. Lower FPR indicates better DNN calibration and high capability of separating correct and incorrect predictions leading to enhanced reliability, particularly at high sensitivity thresholds.
- *Risk at x% Coverage (Risk)*: Quantifies the expected risk (e.g., error rate or loss) when the DNN makes predictions on x% of the data (i.e., abstains from predicting for the most uncertain (1-x)%). Lower risk at a given coverage reflects better selective prediction performance.
- *Area Under the Risk Coverage Curve (AURC)*: Summarizes the trade-off between risk and coverage across all possible coverage levels. A lower AURC indicates a DNN that minimizes risk more effectively while maintaining high coverage. This acts as an average of the risk across different coverages.

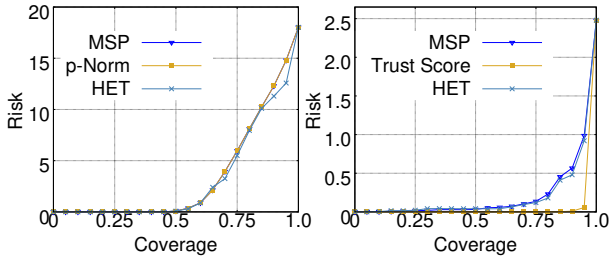
We report FPR at 95% TPR and Risk at 95% coverage.

B. Performance Evaluation

Tables I, II, III and IV present the performance of the SC approaches across different architectures and datasets. For the ACI IoT dataset, the results demonstrate exceptional reliability, with the MLP architecture achieving a risk as low as 0.01% and the CNN architecture achieving 0.05%—both utilizing the trust score. In addition to the minimal risk, the trust score achieves a very low FPR, highlighting its effectiveness in distinguishing between correct and incorrect predictions for this dataset. Conversely, the trust score performs poorly on the RadioML dataset. Here, the best-performing methods are HET (for the CNN architecture), with a risk of 12.58%, and p-norm (for the VGG8 architecture), with a risk of 5.14%. Notably, the p-norm

method demonstrates that when an appropriate SC approach is applied, substantial improvements in reliability can be achieved. For example, even at a coverage of 95%, p-norm attains a remarkably low risk.

This conclusion is further supported by the risk-coverage plots shown in Figure 2. These plots illustrate that for the RadioML dataset and CNN architecture, the risk decreases from 18% at 100% coverage to 8% at 80% coverage—a 55% reduction in risk for a 20% loss in coverage. Similarly, for the ACI IoT dataset, the risk effectively drops to zero when the coverage falls below 95%. From the results presented in Table I - IV and Figure 2, it emerges that utilizing *appropriate* SC provides substantial gain. At the same time, these results also highlight that the choice of the *appropriate* method is not straightforward and currently there is no principled way for the task even in the CV domain. This opens up a new research direction for *post hoc* selective classification approaches.



(a) Risk-Coverage curve for RadioML dataset (b)Risk-Coverage curve for ACI IoT dataset

Fig. 2: We show the RC curve for the top two SC methods on CNN architecture. The MSP is shown as the default baseline.

TABLE I: Performance comparison of different SC approaches for the RadioML dataset and CNN architecture.

Method	AUROC	FPR	Risk	AURC
MSP	91.23	56.12	14.85	3.38
Trust Score	87.56	69.49	15.82	4.16
P-Norm	91.23	55.88	14.8	3.37
Temperature Scaling	91.24	55.98	14.79	3.37
ATS	82.18	59.68	15.28	7.25
HET	**92.56**	**52.45**	**12.58**	**2.68**
energy	91.98	54.68	13.74	3.25
GEN	90.68	57.37	14.95	3.5

C. SC under Distribution Shift

To evaluate the utility of the SC approaches under distribution shifts, we analyze *covariate shift* and *semantic shift*.

1) *Covariate Shift*: For the RadioML dataset, covariate shift is introduced by training the DNNs only on samples with signal-to-noise ratio (SNR) values of 0 dB and above. The unseen samples, those with SNR values below 0 dB, represent covariate-shifted inputs, as they differ in noise power compared to the training set.

TABLE II: Performance comparison of different SC approaches for the RadioML dataset and VGG8 architecture.

Method	AUROC	FPR	Risk	AURC
MSP	93.06	46.03	12.25	2.42
Trust Score	88.66	48.59	13.62	3.16
P-Norm	93.06	46.03	**5.14**	**2.12**
Temperature Scaling	93.02	46.27	12.25	2.43
ATS	76.93	48.93	12.76	6.95
HET	93.46	45.45	**12.05**	2.35
energy	**93.98**	45.95	12.14	2.35
GEN	93.07	45.48	12.41	2.47

2) *Semantic Shift*: Semantic shift occurs when the DNN encounters semantically distinct or out-of-distribution (OOD) inputs. For the RadioML dataset, this is modeled by training the VGG architecture on all modulation types except FM and GMSK. These excluded modulation types are treated as OOD classes. During evaluation, we measure the percentage of the samples from FM and GMSK modulation for which the SC abstains.

TABLE III: Performance comparison of different SC approaches for the ACI IoT dataset and CNN architecture.

Method	AUROC	FPR	Risk	AURC
MSP	94.256	29.83	0.98	0.19
Trust Score	**99.8**	**0.78**	**0.05**	**0.03**
P-Norm	94.02	30.7	1.01	0.19
Temperature Scaling	69.4	56.7	2.01	2.36
ATS	46.9	80.24	1.01	0.19
HET	95.22	28.45	0.9	0.15
energy	94.50	29.6	0.96	0.2
GEN	94.98	28.2	0.94	0.18

TABLE IV: Performance comparison of different SC approaches for the ACI IoT dataset and MLP architecture.

Method	AUROC	FPR	Risk	AURC
MSP	95.23	23.51	0.35	0.085
Trust Score	**99.82**	**0.78**	**0.01**	**0.011**
P-Norm	95.17	22.84	0.36	0.084
Temperature Scaling	59.3	76.6	1.1	0.12
ATS	43.23	90.23	1.1	0.12
HET	95.53	22.4	0.29	0.076
energy	95.63	22.45	0.2	0.075
GEN	95.43	21.45	0.33	0.079

A similar approach is applied to the ACI IoT dataset. The MLP architecture is trained without the *Ping Sweep* and *Syn Flood* attack classes, as well as *ARP Spoofing* and *UDP Flood*. These four excluded attack types become OOD classes, and the robustness of the SC is measured by evaluating on what percentage it abstains from inferring on these OOD inputs.

Figure 3 shows the detection rate of OOD samples across various SC methods. Generalized approaches (*MSP, Trust Score, p-norm, HET*) perform poorly on both datasets having detection rate of about 50% for ACI IoT with

MLP3 and 30% for RadioML with VGG8. In contrast, specialized methods (*GEN, Energy*) perform significantly better—GEN detects 83% and Energy 63% on ACI IoT with MLP3, and 73% and 56% on RadioML, respectively. However, under covariate shift in RadioML, all methods detect less than 20% of shifted samples. This gap highlights that generalized SC methods lack robustness under domain shift, and specialized approaches are necessary.

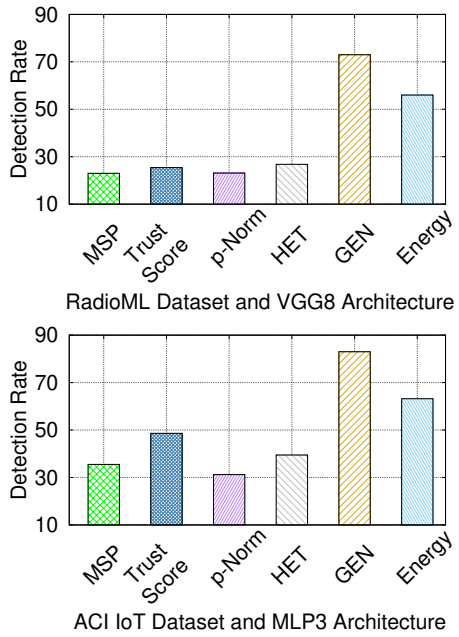


Fig. 3: The detection rate for OOD samples

D. Training DNNs to be Selective

In order to evaluate training strategies to make the DNN more selective, we investigate combination of four strategies: (i) *entropic regularization* [18], (ii) supervised contrastive learning [36], (iii) stochastic weight averaging (SWA) [20], and (iv) cosine similarity classifier (CSC). First, we minimize the loss function: $\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda * \mathcal{L}_{entropy}$ where $\mathcal{L}_{CE}$ is the cross-entropy loss, $\mathcal{L}_{entropy} = H(\text{softmax}(f(\mathbf{x};\theta)))$ is the Shannon-entropy of the output softmax probability, and $\mathcal{L}_{total}$ is the total loss. Second, we extract features $f_{L-1}(\mathbf{x})$ from $L-1$-th layer of a L layered DNN and employ supervised contrastive loss (supcon loss) together with standard cross-entropy loss to train the network. The is motivated by the theory of neural collapse [37] showing features (of a particular class) from a well-trained network cluster around the class mean. Conversely, with supcon loss, we constrain the feature vectors from a class to align in one direction. This, in turn, should place the incorrectly predicted samples in between the direction of two classes and thus improve SC performance. Third, we note that SWA is proven to be conducive to generalization [20], [38] and SC [19] in CV domain. In this approach, when the DNN is close to convergence, we checkpoint the DNN after each epoch and average the parameters of those checkpoints to obtain the final DNN. Fourth, we replace the logit layer with CSC after training the DNN. The CSC measures the similarity of the input feature to the weight vectors for each class. Formally, given a feature vector $f_{L-1}(\mathbf{x})$, and the weights matrix $\mathbf{W}$, the output of the cosine similarity classifier for i-th class is $\tilde{y}_i = \frac{f_{L-1}(\mathbf{x})\cdot\mathbf{W}_i}{||f_{L-1}(\mathbf{x})||_2||\mathbf{W}_i||_2}$, where $\mathbf{W}_i$ is the i-th column of the weight matrix corresponding to the i-th class and $||\cdot||_2$ denotes the L_2 norm. The motivation for this step comes from the effectiveness of CSC on few-shot classification [39], [40]. As the weight matrix, we use the weight matrix of the pre-trained logit layer.

TABLE V: Comparison among different training approaches for SC performance. We show for VGG8 architecture and RadioML dataset.

Training Approach	Risk	AURC
Vanilla training	5.14	2.42
Entropy Regularization	5.65	2.54
Entropy Regularization + SWA	**4.19**	**2.11**
Entropy + SWA + CSC	6.24	2.67
Entropy + SWA + Supcon	5.14	2.34
Entropy + SWA + Supcon + CSC	6.2	3.12

Table V reports the performance for various combinations of these training strategies for the RadioML dataset and VGG8 architecture. From Table V, we notice that the combination of entropy regularization and SWA is clearly the only one improving performance over the vanilla training. As SWA provides with wider and consequently less sharp minima, sharper minima is harmful for SC. As a result, research into techniques which encourage flatter minima might be more useful for SC.

IV. Conclusion

In this paper, we have evaluated the state of the art approaches for Selective Classification (SC). We have cast the problem in the context of military settings and selected the problems Automatic Modulation Classification (AMC) and Network Intrusion Detection System (NIDS) tasks using the RadioML and ACI IoT datasets, respectively. We have shown that SC allows up to 55% risk reduction for AMC with a 20% loss in coverage and near-zero risk for NIDS with minimal coverage loss. We have also analyzed the performance of SC under distribution shifts. Our investigation has shown that existing art has limitations when covariate and semantic shifts are presented to the deep neural network (DNN). We have explored training strategies such as entropy regularization and stochastic weight averaging to enhance SC performance. We hope that this paper will spur further investigations.

Acknowledgments

This work has been funded in part by the National Science Foundation under grants ECCS-2229472 and CNS-2312875, by the Air Force Office of Scientific Research

under contract number FA9550-23-1-0261, by the Office of Naval Research under award number N00014-23-1-2221, and by the Defense Advanced Research Projects Agency (DARPA) under the Young Faculty Award program.

REFERENCES

[1] J. Bajwa, U. Munir, A. Nori, and B. Williams, "Artificial Intelligence in Healthcare: Transforming the Practice of Medicine," *Future healthcare journal*, vol. 8, no. 2, pp. e188–e194, 2021.

[2] N. Jazdi, B. A. Talkhestani, B. Maschler, and M. Weyrich, "Realization of AI-enhanced Industrial Automation Systems Using Intelligent Digital Twins," *Procedia CIRP*, vol. 97, pp. 396–400, 2021.

[3] M. Cummings, *Artificial Intelligence and the Future of Warfare*. Chatham House for the Royal Institute of International Affairs London, 2017.

[4] I. Szabadföldi, "Artificial Intelligence in Military Application–Opportunities and Challenges," *Land Forces Academy Review*, vol. 26, no. 2, pp. 157–165, 2021.

[5] A. Nguyen, J. Yosinski, and J. Clune, "Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[6] Y. Geifman and R. El-Yaniv, "Selectivenet: A Deep Neural Network with an Integrated Reject Option," in *International conference on machine learning*, pp. 2151–2159, PMLR, 2019.

[7] C. Corbière, N. THOME, A. Bar-Hen, M. Cord, and P. Pérez, "Addressing Failure Prediction by Learning Model Confidence," in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.

[8] L. Huang, C. Zhang, and H. Zhang, "Self-Adaptive Training: beyond Empirical Risk Minimization," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 19365–19376, Curran Associates, Inc., 2020.

[9] Z. Liu, Z. Wang, P. P. Liang, R. R. Salakhutdinov, L.-P. Morency, and M. Ueda, "Deep Gamblers: Learning to Abstain with Portfolio Theory," in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.

[10] C. Corbière, N. Thome, A. Saporta, T.-H. Vu, M. Cord, and P. Pérez, "Confidence Estimation via Auxiliary Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6043–6055, 2022.

[11] F. Granese, M. Romanelli, D. Gorla, C. Palamidessi, and P. Piantanida, "Doctor: A simple Method for Detecting Misclassification Errors," *Advances in Neural Information Processing Systems*, vol. 34, pp. 5669–5681, 2021.

[12] M. Shen, Y. Bu, P. Sattigeri, S. Ghosh, S. Das, and G. Wornell, "Post-hoc Uncertainty Learning Using a Dirichlet Meta-model," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 9772–9781, 2023.

[13] L. F. P. Cattelan and D. Silva, "How to Fix a Broken Confidence Estimator: Evaluating Post-hoc Methods for Selective Classification with Deep Neural Networks," in *The 40th Conference on Uncertainty in Artificial Intelligence*.

[14] T. J. O'Shea, T. Roy, and T. C. Clancy, "Over-the-Air Deep Learning Based Radio Signal Classification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 1, pp. 168–179, 2018.

[15] N. Bastian, D. Bierbrauer, M. McKenzie, and E. Nack, "Aci iot network traffic dataset 2023," 2023.

[16] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *arXiv preprint arXiv:1903.12261*, 2019.

[17] Y. Geifman and R. El-Yaniv, "Selective cClassification for Deep Neural Networks," *Advances in neural information processing systems*, vol. 30, 2017.

[18] L. Feng, M. O. Ahmed, H. Hajimirsadeghi, and A. Abdi, "Towards Better Selective Classification," *arXiv preprint arXiv:2206.09034*, 2022.

[19] F. Zhu, Z. Cheng, X.-Y. Zhang, and C.-L. Liu, "Rethinking Confidence Calibration for Failure Prediction," in *European Conference on Computer Vision*, pp. 518–536, Springer, 2022.

[20] P. Izmailov, A. Wilson, D. Podoprikhin, D. Vetrov, and T. Garipov, "Averaging Weights Leads to Wider Optima and Better Generalization," in *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, pp. 876–885, 2018.

[21] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware Minimization for Efficiently Improving Generalization," *arXiv preprint arXiv:2010.01412*, 2020.

[22] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in *International conference on machine learning*, pp. 1321–1330, PMLR, 2017.

[23] M. Kull, M. Perello Nieto, M. Kängsepp, T. Silva Filho, H. Song, and P. Flach, "Beyond Temperature Scaling: Obtaining Well-calibrated Multi-class Probabilities with Dirichlet Calibration," *Advances in neural information processing systems*, vol. 32, 2019.

[24] K. Gupta, A. Rahimi, T. Ajanthan, T. Mensink, C. Sminchisescu, and R. Hartley, "Calibration of Neural Networks Using Splines," *arXiv preprint arXiv:2006.12800*, 2020.

[25] H. Markowitz, "Portfolio selection," *The Journal of Finance*, vol. 7, no. 1, pp. 77–91, 1952.

[26] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," in *international conference on machine learning*, pp. 1050–1059, PMLR, 2016.

[27] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles," *Advances in neural information processing systems*, vol. 30, 2017.

[28] A. Malinin and M. Gales, "Predictive Uncertainty Estimation via Prior Networks," *Advances in neural information processing systems*, vol. 31, 2018.

[29] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential Deep Learning to Quantify Classification Uncertainty," *Advances in neural information processing systems*, vol. 31, 2018.

[30] J. Nandy, W. Hsu, and M. L. Lee, "Towards Maximizing the Representation Gap Between In-domain & Out-of-distribution Examples," *Advances in neural information processing systems*, vol. 33, pp. 9239–9250, 2020.

[31] H. Jiang, B. Kim, M. Guan, and M. Gupta, "To Trust or Not to Trust a Classifier," *Advances in neural information processing systems*, vol. 31, 2018.

[32] M. Collier, R. Jenatton, B. Mustafa, N. Houlsby, J. Berent, and E. Kokiopoulou, "Massively Scaling Heteroscedastic Classifiers," in *The Eleventh International Conference on Learning Representations*, 2023.

[33] S. A. Balanya, J. Maroñas, and D. Ramos, "Adaptive Temperature Scaling for Robust Calibration of Deep Neural Networks," *Neural Computing and Applications*, vol. 36, no. 14, pp. 8073–8095, 2024.

[34] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based Out-of-distribution Detection," *Advances in Neural Information Processing Systems*, 2020.

[35] X. Liu, Y. Lochman, and C. Zach, "GEN: Pushing the Limits of Softmax-Based Out-of-Distribution Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23946–23955, June 2023.

[36] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18661–18673, 2020.

[37] J. Jiang, J. Zhou, P. Wang, Q. Qu, D. Mixon, C. You, and Z. Zhu, "Generalized Neural Collapse for a Large Number of Classes," *arXiv preprint arXiv:2310.05351*, 2023.

[38] B. Athiwaratkun, M. Finzi, P. Izmailov, and A. G. Wilson, "There are Many Consistent Explanations of Unlabeled Data: Why You Should Average," *arXiv preprint arXiv:1806.05594*, 2018.

[39] S. Gidaris and N. Komodakis, "Dynamic Few-Shot Visual Learning without Forgetting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4367–4375, 2018.

[40] S. X. Hu, P. G. Moreno, Y. Xiao, X. Shen, G. Obozinski, N. Lawrence, and A. Damianou, "Empirical Bayes Transductive Meta-Learning with Synthetic Gradients," in *International Conference on Learning Representations*, 2020.