Project Report
LSP-265

Artificial Intelligence for Decision Emulation (Medic-AIDE): FY19 Biomedical Sciences and Technologies Line-Supported Program

T. Tsiligkaridis

4 November 2019

Lincoln Laboratory
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LEXINGTON, MASSACHUSETTS

This material is based upon work supported under Air Force Contract No. FA8702-15-D-0001.

This report is the result of studies performed at Lincoln Laboratory, a federally funded research and development center operated by Massachusetts Institute of Technology. This material is based upon work supported under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Air Force.

Massachusetts Institute of Technology
Lincoln Laboratory

Artificial Intelligence for Decision Emulation (Medic-AIDE): FY19 Biomedical Sciences and Technologies Line-Supported Program

T. Tsiligkaridis

Group 45

Project Report LSP-265

4 November 2019

Lexington Massachusetts

This page intentionally left blank.

ABSTRACT

Precise estimation of uncertainty in predictions for AI systems is a critical factor in ensuring trust and safety. Replicating and enhancing experts' decisions while quantifying uncertainty in predictions is a challenging problem. Uncertainty-aware AI for safety-critical domains such as healthcare, autonomous navigation and cybersecurity is a requirement. In particular, when AI is used to emulate decisions of medical experts in the field, AI confidence needs to be measured and plays a key role in making effective triage decisions and choosing appropriate treatment options. While various aspects of deep learning, such as achieving high accuracy and optimizing architectures are maturing, precise predictive uncertainty estimation remains a subject of on-going research efforts. Conventional neural networks tend to be overconfident as they do not account for uncertainty during training. In contrast to Bayesian neural networks that learn approximate distributions on weights to infer prediction confidence, we propose a novel method, Information Robust Dirichlet networks, that provides accurate uncertainty estimates while maintaining high prediction accuracy. Properties of the new cost function are derived to indicate how improved uncertainty estimation is achieved. Experiments using real medical datasets on heart arrhythmia diagnosis and AI-assisted pre-hospital triage show that our technique outperforms state-of-the-art neural networks, by a large margin, for estimating predictive uncertainty.

This page intentionally left blank.

TABLE OF CONTENTS

This page intentionally left blank.

LIST OF FIGURES

LIST OF FIGURES
(Continued)

Figure No. **Page**

1. INTRODUCTION

1.1 PROBLEM AND PREVIOUS WORK

A key role of artificial intelligence (AI) has been to learn from challenging data patterns and make complex inferences that humans may not be very good at or as efficient at performing. Deep learning systems have achieved state-of-the-art performance in various domains (LeCun et al., 2015). The first successful applications of deep learning include large-scale object recognition (Krizhevsky et al., 2012) and machine translation (Sutskever et al., 2014; Wu et al., 2016). While further advances have achieved strong performance and often surpass human-level ability in computer vision (Ciresan et al., 2012; Geirhos et al., 2018; He et al., 2015), speech recognition (Hinton et al., 2012; Xiong et al., 2017), medicine (Wang et al., 2016), bioinformatics (Alipanahi et al., 2015), other aspects of deep learning are less well understood. Conventional neural networks (NNs) are overconfident in their predictions (Guo et al., 2017) and provide inaccurate predictive uncertainty (Louizos and Welling, 2017). Intepretability, robustness, and safety are becoming increasingly important as deep learning is being deployed across various industries including healthcare, autonomous driving and cybersecurity.

Motivated by the application of AI in medical decision making problems, including field-forward trauma care, characterizing predictive uncertainty is a crucial goal while maintaining high accuracy. This requires more capable AI in order to address more complex goals, including calibration of complex models, anomaly detection, learning complex personalized models, learning from limited noisy and missing data, active learning and experimental design, etc.

In field-forward trauma, about 1/4 of combat deaths in Iraq and Afghanistan were potentially survivable. Uncontrolled blood loss was the leading cause of death in 90% of potentially survival battlefield injuries (Eastridge et al., 2012). Specifically, the site of lethal hemorrhage was truncal (67%), followed by junctional (19%) and peropheral-extremity (14%) hemorrhage. Nosocomial infections account for the leading cause of late death after traumatic injury (Dente et al., 2017). Approximately 2/3 of deaths survived the initial injury but died before getting to a field hospital. Medics are frequently the only medically trained personnel available at the point of injury, and future force paradigm extends this isolated period to as long as 1.5 hours. It is important to note that medics often lack the years of experience in trauma and field-forward personnel could benefit from real-time decision aides or just-in-time training. In the trauma-care domain, we envision AI to be a tool for helping medics in the field treating patients, performing effective triage, training novices, predicting with confidence courses of action and characterizing uncertainty where prior experience might be limited and other factors come into the picture, including human fatigue, time restrictions, stress, etc. Furthermore, we envision AI to also be used to perform automated patient stabilization by predicting interventions.

To intuitively understand why conventional neural networks exhibit poor uncertainty estimates, it suffices to notice that the training objective does not account for uncertainty. The typical cross-entropy objective only maximizes the correct-class probability. Furthermore the softmax layer tends to inflate probabilities since it is based on exponentiation and normalization of activation values. Extensions include adding an extra class to capture unknown data (Bendale and Boult, 2016). However, this class augmentation approach has several problems; the data selected to train the unknown class induces bias, the approach does not address the calibration question, and the methods fail to detect adversarial attacks.

Bayesian neural networks (BNNs) (Blundell et al., 2015; Gal and Ghahramani, 2016; Kingma et al., 2015; Molchanov et al., 2017) treat the weights in neural networks as probabilistic parameters and aim to learn the weight distribution using Bayesian inference. The high-dimensionality of weights and the presence of nonlinearities make an exact Bayesian treatment intractable. Therefore, approximations to weight distributions are made, often making assumptions about these distributions that may not hold in practice. Posterior predictive distributions are obtained using approximate integration. Thus, during inference time BNNs have high computational complexity and potential bias. They take more effort to implement and are harder to train than conventional NNs. They obtain better predictive uncertainty estimates than conventional NNs but still leave a lot of room for improvement in terms of how accurate these estimates are. Also, they struggle detecting adversarial attacks.

Emerging architectures for uncertainty-aware deep learning that circumvent the weight inference of BNNs have been recently developed, referred here as Dirichlet neural networks (Malinin and Gales, 2019; Sensoy et al., 2018). Dirichlet NNs use deterministic neural networks to estimate the parameters of a predictive Dirichlet distribution that governs the distribution of class categorical distributions on the probability simplex. While these nascent methods improve upon BNNs, they have several drawbacks. The current training processes depend on the average case prediction error (Sensoy et al., 2018) which assumes that all data are easy to classify (expecting sharp Dirichlet predictive distributions) (Malinin and Gales, 2019), resulting in deteriorating uncertainty estimates for correct and misclassified data examples. Furthermore, a form of their training (Malinin and Gales, 2019) depends on learning unknown data, in similar spirit to (Bendale and Boult, 2016) which induces bias in classification. While they do well in terms of detecting simple adversarial attacks, they fail to reliably detect strong adversarial attacks as will be shown in this report. Connections between the current training criteria and the Dirichlet distribution have not been well established.

In our approach, we develop a new training method for Dirichlet neural networks. The new objective used for training fits predictive distributions to data by minimizing a calibration loss (expected L_p norm of prediction error) and an information divergence loss that penalizes information flow towards incorrect classes (sharpening or flattening distributions on the probability simplex based on how difficult each example is to classify) and maximizing the uncertainty (measured by differential entropy) for small adversarial perturbations. The end result yields an information-robust Dirichlet (IRD) neural network that is tightly coupled to the training data and is able to reliably measure predictive uncertainty within the data distribution (e.g. predict when an error will likely be made), detect anomalous examples outside the data distribution and even detect adversarial attacks (designed to fool the network with perfect knowledge of the network and its weights). We demonstrate the superiority of IRD on several applications through numerical experiments and dervie several properties of the training objective that show how improved uncertainty estimation is achieved.

1.2 BRIEF EXECUTIVE SUMMARY

We present a very brief executive summary of our work. In our probabilistic framework, a deep Dirichlet neural network is trained in a supervised fashion with a novel training objective that accounts for uncertainty in a non-trivial way. Our novel method, Information Robust Dirichlet networks (IRD), learns the Dirichlet distribution on prediction probabilities by minimizing the expected L_p norm of the prediction error and an information divergence loss that penalizes information flow towards incorrect classes, while simultaneously maximizing differential entropy of small adversarial perturbations to provide accurate uncertainty estimates. After obtaining a closed-form expression for the novel training objective, several properties are derived that support how improved uncertainty estimation is achieved. Experiments using real datasets show that our technique outperforms state-of-the-art neural networks, by a large margin, for estimating in-distribution and out-of-distribution uncertainty, and detecting adversarial examples.

First, we apply our method to a benchmark image classification task of handwritten digits. Our method achieves a competitive prediction accuracy in comparison to other state-of-the-art deep learning methods, while it achieves superior predictive uncertainty estimation for in-distribution, out-of-distribution and adversarial examples. IRD is able to predict when the AI system will likely make an error, it detects anomalous digital images with high confidence unlike the ones used for training, and succeeds in detecting adversarial attacks designed to fool the classifier that are very hard to visually detect as anomalous. Simple and more sophisticated adversarial attacks are both detectable with high confidence by our method, while other methods struggle to detect them. These attacks are generated using knowledge of the network structure and classification loss. Our results show unmatched performance against other conventional and state-of-the-art uncertainty-aware neural networks.

Second, our method is applied to an electrocardiogram (ECG)-based heart arrhythmia diagnosis task. ECG signals are often noisy due to electrode contact noise, motion artifacts, muscle contractions, etc., making it a challenging AI task to classify short ECG single-lead recordings into normal rhythm and atrial fibrillation. Atrial fibrillation is the most common sustained cardiac arrhythmia, and is associated with high mortality and morbidity rates (Clifford et al., 2017). Results show that IRD achieves high prediction accuracy on par with other state-of-the-art deep learning methods while outperforming them in terms of uncertainty estimation. This methodology may be extended further to augment clinical decision systems equipped with ECG devices. Furthermore, our method successfully detects anomalous ECG signals (e.g., too noisy or not indicative of either type of normal or atrial fibrillation rhythms).

Third, we apply our method to a trauma care decision making task developing an AI-supported pre-hospital triage tool that accurately identifies shock and predicts surgical and transfusion requirement. Through our collaboration with the Massachusetts General Hospital, Division of Trauma, Emergency Surgery and Surgical Critical Care, we used the Trauma Quality Improvement Program (TQIP) dataset over a two-year span of $2015-2016$ and $2016-2017$ hospital admissions with a specific focus on truncal gunshot wounds. Our technology gives certainty values of the predictions rather than risking overconfident predictions that could harm in-field triage decision making. The certainty values correspond very well to misclassified and correct predictions. Further development and implementation of this tool has the potential to optimize triage in the field, both in civilian and military settings.

1.3 SECTION CONTENTS

The sections in the remaining part of the report are as follows. Section 2 contains the uncertainty-aware AI algorithm development and presents theoretical properties of the new method. It serves as the basis for the remaining sections. This new AI method is applied to a benchmark image classification task in Section 3 and to ECG-based atrial fibrillation diagnosis in Section 4. Several comparisons with other state-of-the-art methods are included. The material up to this point was submitted as a technology disclosure, a provisional patent has been filed, and is under review at an AI conference. Section 5 contains applications of the AI method to trauma care decision problems. This material has been submitted as an abstract to a trauma competition (Committee on Trauma for the American College of Surgeons). The conclusions and directions for future work are presented in Section 6.

2. INFORMATION ROBUST DIRICHLET NETWORKS FOR PREDICTIVE UNCERTAINTY ESTIMATION

In this section, a novel deep learning method is presented that predicts uncertainty accurately while maintaining high prediction accuracy. Conventional neural networks tend to be overconfident as they do not account for uncertainty during training. In contrast to Bayesian neural networks that learn approximate distributions on weights to infer prediction confidence, we propose a novel method, Information Robust Dirichlet networks, that learns the Dirichlet distribution on prediction probabilities by minimizing the expected L_p norm of the prediction error and an information divergence loss that penalizes information flow towards incorrect classes, while simultaneously maximizing differential entropy of small adversarial perturbations to provide accurate uncertainty estimates. Properties of the new cost function are derived to indicate how improved uncertainty estimation is achieved. Experiments using real datasets show that our technique outperforms state-of-the-art neural networks, by a large margin, for estimating in-distribution and out-of-distribution uncertainty, and detecting adversarial examples.

2.1 INTRODUCTION AND PRIOR WORK

Uncertainty modeling in deep learning is a crucial aspect that has been the topic of various Bayesian neural network (BNN) research studies (Blundell et al., 2015; Gal and Ghahramani, 2016; Kingma et al., 2015; Molchanov et al., 2017). BNNs capture parameter uncertainty of the network by learning distributions on weights and estimate a posterior predictive distribution by approximate integration over these parameters. The non-linearities embedded in deep neural networks make the weight posterior intractable and several tractable approximations have been proposed and trained using variational inference (Blundell et al., 2015; Gal and Ghahramani, 2016; Kingma et al., 2015; Li and Gal, 2017; Molchanov et al., 2017), the Laplace approximation (MacKay, 1992; Ritter et al., 2018), expectation propagation (Hernandez-Lobato and Adams, 2015; Sun et al., 2017), and Hamiltonian Monte Carlo (Chen et al., 2014). The success of approximate BNN methods depends on how well the approximate weight distributions match their true counterparts, and their computational complexity is determined by the degree of approximation. Most BNNs take more effort to implement and are harder to train in comparison to conventional NNs. Furthermore, approximate integration over the parameter uncertainties increases the test time due to posterior sampling, and yields an approximate predictive distribution that is subject to bias, due to stochastic averaging. Thus, it is of interest to develop methods that provide good uncertainty estimates while reusing the training pipeline and maintaining scalability. To this end, a simple approach was proposed by (Lakshminarayanan et al., 2017) that combines NN ensembles with adversarial training to improve predictive uncertainty estimates in a non-Bayesian manner. It is known that deterministic NNs are brittle to adversarial attacks (Goodfellow et al., 2014; Kurakin et al., 2017) and various defenses have been proposed to increase accuracy for low levels of noise (Madry et al., 2018). Recently, a study (Lee et al., 2018) used generative adversarial networks to generate boundary samples and trained the classifier to be uncertain on those as a means to improve detection of out-of-distribution samples. While adversarial defense has been explored, the idea of maximizing uncertainty on low-noise adversarial examples to improve predictive uncertainty estimates has not been investigated to the best of our knowledge.

Recently, in (Malinin and Gales, 2019; Sensoy et al., 2018) the Dirichlet distribution was used to model distributions of class compositions and its parameters were learned by training deterministic neural networks. This approach for Bayesian classification yields closed-form predictive distributions and outperforms BNNs in uncertainty quantification for out-of-distribution and adversarial queries. However, uncertainty estimation performance for in-distribution queries was not studied, and out-of-distribution and adversarial queries performance can be significantly improved.

In this report, we propose Information Robust Dirichlet networks that deliver more accurate predictive uncertainty than other state-of-the-art methods. Our method modifies the output layer of neural networks and the training loss, therefore maintaining computational efficiency and ease of implementation. The contributions are as follows. First, a new training loss based on minimizing the expected L_p norm of the prediction error is proposed under which the prediction probabilities follow a Dirichlet distribution. A closed-form approximation to this loss is derived, under which a deterministic neural network is trained to infer the parameters of a Dirichlet distribution, effectively teaching neural networks to learn distributions over class probability vectors. Second, an information divergence is used to regularize the estimated Dirichlet distribution and a maximum entropy penalty on adversarial examples is used to maximize uncertainty near the edge of the data distribution. Third, an analysis is provided that shows how properties of the new loss improve uncertainty estimation. Finally, we demonstrate on real datasets that our technique obtains unmatched success in terms of uncertainty estimation for correct and incorrect predictions, detection of out-of-distribution queries and adversarial attacks.

2.2 PITFALLS OF CONVENTIONAL SOFTMAX NETWORKS

The conventional approach for the classification layer includes the softmax operator which takes continuous-valued activations of the output layer and converts them into probabilities. Typically the cross-entropy loss is used for training which does not account for uncertainty. As noted in Gal and Ghahramani (2016); Louizos and Welling (2017); Sensoy et al. (2018), the exponentiation involved to form a point estimate of the class probabilities tends to inflate them deteriorating uncertainty estimates derived from the softmax probabilities. As a result, uncertainty estimation suffers due to the parametrization and the fact that the training loss does not account for uncertainty. This is illustrated in Fig. 1 in which an image of a digit 6 is correctly classified initially, but as it rotates the softmax output incorrectly classifies it with high probability as an 8.[1] In contrast, our approach yields a near-uniform distribution during the rotation stage and thus provides a reasonable uncertainty estimate using the entropy of the predictive distribution.

[1] We remark that the networks are trained on MNIST without rotated data augmentations.

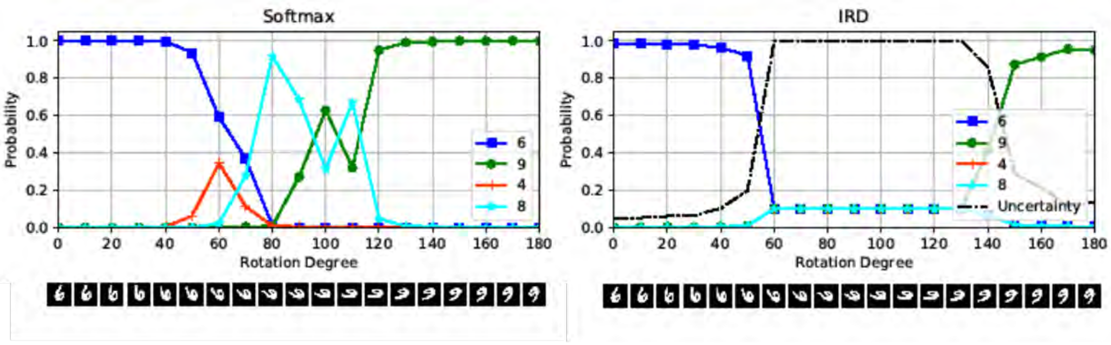


Figure 1. Classification of rotated digit 6 *spanning a 180-degree rotation for standard neural network with softmax output (left) and our proposed approach (right). Our approach tracks the uncertainty throughout the rotation and accurately predicts the correct class at both ends.*

2.3 LEARNING DISTRIBUTIONS ON THE PROBABILITY SIMPLEX

2.3.1 Dirichlet Distribution

Outputs of standard neural networks for classification tasks are probability vectors over classes. The basis of our approach lies in modeling the distribution of such probability vectors for each example using the Dirichlet distribution (Mauldon, 1959; Mosimann, 1962). Given the probability simplex as $\mathcal{S} = \{(p_1, \dots, p_K) : p_i \geq 0, \sum_i p_i = 1\}$, the Dirichlet distribution is a probability density function on vectors $\mathbf{p} \in \mathcal{S}$ given by

$$f_{\boldsymbol{\alpha}}(\mathbf{p}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{j=1}^{K} p_j^{\alpha_j - 1}$$

where $B(\boldsymbol{\alpha}) = \prod_{j=1}^{K} \Gamma(\alpha_j)/\Gamma(\alpha_0)$ is the multivariate Beta function. It is characterized by K parameters $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)$ here assumed to be larger than unity. [2] In the special case of the all-ones $\boldsymbol{\alpha}$ vector, the distribution becomes uniform over the probability simplex. The mean of the proportions is given by $\hat{p}_j = \alpha_j/\alpha_0$, where $\alpha_0 = \sum_j \alpha_j$ is the Dirichlet strength. The Dirichlet distribution is conjugate to the multinomial distribution, with posterior parameters updated as $\alpha'_j = \alpha_j + y_j$ for a multinomial sample $\mathbf{y} = (y_1, \dots, y_K)$. For a single sample, $y_j = I_{\{j=c\}}$, where c is the index of the correct class.

The marginal distributions of the Dirichlet distribution are Beta random variables, specifically, $p_j \sim \text{Beta}(\alpha_j, \alpha_0 - \alpha_j)$ with support on $[0, 1]$. The q-th moment of the Beta distribution $\text{Beta}(\alpha', \beta')$ is given by

$$\mathbb{E}[p^q] = \int_0^1 p^q \frac{p^{\alpha'-1}(1-p)^{\beta'-1}}{B_u(\alpha', \beta')} dp = \frac{B_u(\alpha' + q, \beta')}{B_u(\alpha', \beta')} \tag{1}$$

where $B_u(\alpha', \beta') = \Gamma(\alpha')\Gamma(\beta')/\Gamma(\alpha' + \beta')$ is the univariate Beta function.

[2] The reason for this constraint is that the Dirichlet distribution becomes inverted for $\alpha_j < 1$ concentrating in the corners of the simplex and along its boundaries.

2.3.2 Classification Loss

Consider given data $\{\mathbf{x}_i\}$ and associated labels $\{\mathbf{y}_i\}$ drawn from a set of K classes. We model the class probability vectors for sample i given by $\mathbf{p}_i$ as random vectors drawn from a Dirichlet distribution conditioned on the input $\mathbf{x}_i$. A neural network with input $\mathbf{x}_i$ is trained to learn this Dirichlet distribution, $f_{\boldsymbol{\alpha}_i}(\mathbf{p}_i)$, with output $\boldsymbol{\alpha}_i$. While the layers of the Dirichlet neural network can be similar to classical NNs, the softmax classification layer is replaced by a softplus activation layer that outputs non-negative continuous values, e.g., $g_\alpha(\mathbf{x}_i; w) \in \mathbb{R}_+^K$ where w are the network parameters, from which we obtain $\boldsymbol{\alpha}_i = g_\alpha(\mathbf{x}_i; w) + 1$.

Given one-hot encoded labels $\mathbf{y}_i$ of examples $\mathbf{x}_i$ with correct class c_i, the Bayes risk of the L_p prediction error for $p \geq 1$ is approximated using Jensen's inequality as

$$\begin{aligned}
\mathbb{E}\|\mathbf{y}_i - \mathbf{p}_i\|_p &\leq \left(\mathbb{E}[\|\mathbf{y}_i - \mathbf{p}_i\|_p^p]\right)^{1/p} \\
&= \left(\sum_{j=1}^{K} \mathbb{E}|y_{ij} - p_{ij}|^p\right)^{1/p} \\
&= \left(\mathbb{E}[(1 - p_{i,c_i})^p] + \sum_{j \neq c_i} \mathbb{E}[p_{ij}^p]\right)^{1/p} \qquad =: \mathcal{F}_i(w)
\end{aligned}$$

The max-norm can be approximated by using a large p. To calculate each term, we note $1 - p_{i,c_i}$ has a distribution $\text{Beta}(\alpha_{i,0} - \alpha_{i,c_i}, \alpha_{i,c_i})$ due to mirror symmetry, and p_{ij} has distribution $\text{Beta}(\alpha_{i,j}, \alpha_{i,0} - \alpha_{i,j})$. Using the moment expression (1) for Beta random variables:

$$\begin{aligned}
\mathcal{F}_i(w) &= \left(\frac{B_u(\alpha_{i,0} - \alpha_{i,c_i} + p, \alpha_{i,c_i})}{B_u(\alpha_{i,0} - \alpha_{i,c_i}, \alpha_{i,c_i})} + \sum_{j \neq c_i} \frac{B_u(\alpha_{i,j} + p, \alpha_{i,0} - \alpha_{i,j})}{B_u(\alpha_{i,j}, \alpha_{i,0} - \alpha_{i,j})}\right)^{\frac{1}{p}} \\
&= \left(\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + p)}\right)^{\frac{1}{p}} \left(\frac{\Gamma\left(\sum\limits_{k \neq c} \alpha_k + p\right)}{\Gamma\left(\sum\limits_{k \neq c} \alpha_k\right)} + \sum_{k \neq c} \frac{\Gamma(\alpha_k + p)}{\Gamma(\alpha_k)}\right)^{\frac{1}{p}}
\end{aligned}$$

The following theorem shows that the loss function $\mathcal{F}_i$ has the correct behavior as the information flow increases towards the correct class which is consistent when an image sample of that class is observed in a Bayesian Dirichlet experiment and hyperparameters are incremented (see Sec. 2.3.1).

Theorem 1. *For a given sample $\mathbf{x}_i$ with correct label c, the loss function $\mathcal{F}_i$ is strictly convex and decreasing in α_c increases (and increases when α_c decreases).*

Theorem 1 shows that our objective function encourages the learned distribution of probability vectors to concentrate towards the correct class. While increasing information flow towards the correct class reduces the loss, it is also important for the loss to capture elements of incorrect classes. It is expected that increasing information flow towards incorrect classes increases uncertainty. The next result shows that through our loss function the model avoids assigning high concentration parameters to incorrect classes as the model cannot explain observations that are assigned incorrect outcomes.

Theorem 2. *For a given sample $\mathbf{x}_i$ with correct label c, the loss function $\mathcal{F}_i$ is increasing in α_j for any $j \neq c$ as α_j grows.*

Theorem 2 implies that our loss function leads the model to push the distribution of class probability vectors away from incorrect classes. The proofs are included in Section 2.4.

2.3.3 Information Divergence Regularization Loss

The classification loss can discover interesting patterns in the data to achieve high classification accuracy. However, the network may learn that certain patterns lead to strong information flow towards incorrect classes, e.g., circular pattern of digit 6 might contribute to a large α associated with digit 8.

We regularize the Dirichlet distribution $f_{\boldsymbol{\alpha}}$ to concentrate away from incorrect classes. Given the auxiliary vector $\boldsymbol{\alpha}'_i = (1 - \mathbf{y}_i) + \mathbf{y}_i \odot \boldsymbol{\alpha}_i$, we minimize the Rényi information divergence (Erven and Harremos, 2014; Rényi, 1961) of the Dirichlet distribution $f_{\boldsymbol{\alpha}}$ from $f_{\boldsymbol{\alpha}'}$:

$$\begin{aligned} D_u^R(f_{\boldsymbol{\alpha}} \parallel f_{\boldsymbol{\alpha}'}) &= \frac{1}{u-1} \log \int_{\mathcal{S}} f_{\boldsymbol{\alpha}}(\mathbf{p})^u f_{\boldsymbol{\alpha}'}(\mathbf{p})^{1-u} d\mathbf{p} \\ &= \frac{1}{u-1} \log \left[\frac{B(u\boldsymbol{\alpha} + (1-u)\boldsymbol{\alpha}')}{B(\boldsymbol{\alpha})^u B(\boldsymbol{\alpha}')^{1-u}} \right] \\ &= \log \left[\frac{B(\boldsymbol{\alpha}')}{B(\boldsymbol{\alpha})} \right] + \frac{1}{u-1} \log \left[\frac{B(u\boldsymbol{\alpha} + (1-u)\boldsymbol{\alpha}')}{B(\boldsymbol{\alpha})} \right] \end{aligned} \tag{2}$$

The order $u > 0$ controls the influence of the likelihood ratio $f_{\boldsymbol{\alpha}}/f_{\boldsymbol{\alpha}'}$ on the divergence. This divergence is minimized if and only if $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}'_i$, in other words when $\alpha_{ij} = 1$ for $j \neq c_i$. The extended order $u = 1$ yields the Kullback-Leibler divergence.

The next theorem presents a local approximation of the divergence (2) in terms of Fisher information matrix $J(\boldsymbol{\alpha}) = \mathbb{E}[\nabla \log f_{\boldsymbol{\alpha}}(\mathbf{p}) \nabla \log f_{\boldsymbol{\alpha}}(\mathbf{p})^T] = -\mathbb{E}[\nabla^2 \log f_{\boldsymbol{\alpha}}(\mathbf{p})]$.

Theorem 3. *As $\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_2^2 = \sum_{j \neq c} (\alpha_j - 1)^2 \to 0$, the Rényi divergence can be locally approximated as:*

$$\begin{aligned} D_u^R(f_{\boldsymbol{\alpha}} \parallel f_{\boldsymbol{\alpha}'}) &\cong \frac{u}{2} (\boldsymbol{\alpha} - \boldsymbol{\alpha}')^T J(\boldsymbol{\alpha}) (\boldsymbol{\alpha} - \boldsymbol{\alpha}') \\ &= \frac{u}{2} \left[\sum_{i \neq c} (\alpha_i - 1)^2 \psi^{(1)}(\alpha_i) - (\sum_{i \neq c} (\alpha_i - 1))^2 \psi^{(1)}(\alpha_0) \right] \end{aligned}$$

where $\psi^{(1)}(z) = \frac{d}{dz}\psi(z)$ is the polygamma function of order 1.

Theorem 3 shows that as $\{\alpha_j\}_{j\neq c} \to 1$ during the training process, the regularization term becomes proportional to the order u that controls the local curvature of the divergence function. The proof is contained in Section 2.4.

Furthermore, the asymptotic approximation has an interesting behavior for various confidence levels α_c. Since the polygamma function is monotonically decreasing, it satisfies $\psi^{(1)}(\alpha_c + \sum_{i\neq c}\alpha_i) > \psi^{(1)}(\alpha'_c + \sum_{i\neq c}\alpha_i)$ for $\alpha_c < \alpha'_c$. Theorem 3 implies that during training, examples that exhibit larger confidence for the correct class c have a higher Rényi divergence associated with them compared to ones with a lower confidence α_c. This is numerically illustrated in Fig. 2 as a function of α_i for some $i \neq c$, when all concentration parameters are held fixed close to 1 and α_c has a low or high value. This implies that the model tends to learn to yield sharper Dirichlet distributions when the correct class confidence is higher since the Rényi divergence is minimized by concentrating away from incorrect classes through $\{\alpha_j\}_{j\neq c} \to 1$.

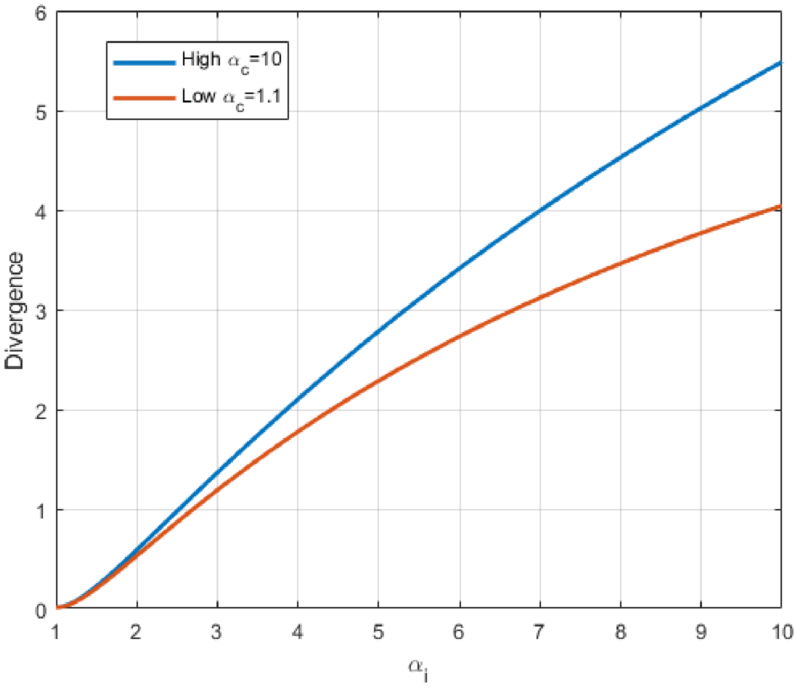


Figure 2. Rényi divergence illustration as $\alpha_i, i \neq c$ varies for the regime $\{\alpha_j\}_{j\neq c} \to 1$ with two different values for the correct class concentration parameter α_c. Here, $u = 2$ and $K = 10$.

2.3.4 Maximum Adversarial Entropy Regularization Loss

To further improve the network robustness, we first generate low-noise adversarial examples using the fast gradient sign method (FGSM) (Goodfellow et al., 2014),

$$\mathbf{x}_{adv} = \mathbf{x} + \epsilon \cdot \operatorname{sgn}(\nabla_{\mathbf{x}} \mathcal{F}(\mathbf{x}, y, w)).$$

Then the Dirichlet network generates $\boldsymbol{\alpha}_{adv}$ that parametrize a distribution on the simplex $f(\mathbf{p}|\mathbf{x}_{adv}, w) = f_{\boldsymbol{\alpha}_{adv}}(\mathbf{p})$, and we maximize the differential entropy of this Dirichlet distribution:

$$\begin{aligned}\mathcal{H}(f_{\boldsymbol{\alpha}_{adv}}(\mathbf{p})) &= -\int_{\mathcal{S}} f_{\boldsymbol{\alpha}_{adv}}(\mathbf{p}) \log f_{\boldsymbol{\alpha}_{adv}}(\mathbf{p}) d\mathbf{p} \\ &= \log B(\boldsymbol{\alpha}_{adv}) + (\alpha_{0,adv} - K)\psi(\alpha_{0,adv}) - \sum_{j=1}^{K}(\alpha_{j,adv} - 1)\psi(\alpha_{j,adv})\end{aligned}$$

This differential entropy captures distributional uncertainty and is maximized when all probability vectors have the same likelihood (pushing α_{adv} towards the all-ones vector). This penalty has the effect of robustifying the predictive Dirichlet distributions inferred by the network so that small adversarial perturbations of the inputs yield high distributional uncertainty. In our experiments we find that this improves the out-of-distribution uncertainty estimation performance as well.

The total loss is $\mathcal{G}_i = \mathcal{F}_i + \lambda D_u^R(f_{\boldsymbol{\alpha}_i} \parallel f_{\boldsymbol{\alpha}'_i}) - \gamma \mathcal{H}(f_{\boldsymbol{\alpha}_{i,adv}})$ where λ, γ are nonnegative parameters controlling the tradeoff between minimizing the approximate Bayes risk and the information regularization penalties. The total loss is summed over a batch of training samples $\mathcal{G}(w) = \sum_{i=1}^{N} \mathcal{G}_i(w)$. Training is performed using minibatches and the adversarial FGSM examples are generated for every minibatch as training progresses with λ, γ increasing using an annealing schedule, e.g., $\lambda_t = \lambda(1 - e^{-0.05t}), \gamma_t = \gamma \min(1, t/40)$.

2.3.5 Uncertainty Metrics

Dirichlet networks generate $\boldsymbol{\alpha} = g_\alpha(\mathbf{x}^*; w) + 1$ that correspond to a Dirichlet distribution on the simplex $f(\mathbf{p}|\mathbf{x}^*, w) = f_{\boldsymbol{\alpha}}(\mathbf{p})$. The predictive distribution is given by

$$P(y = j|\mathbf{x}^*; w) = \mathbb{E}_{f_{\boldsymbol{\alpha}}(\mathbf{p})}[P(y = j|\mathbf{p})] = \frac{\alpha_j}{\alpha_0}$$

Predictive entropy measures total uncertainty and can be decomposed into knowledge uncertainty (arises due to model's difficulty in understanding inputs) and data uncertainty (arises due to class-overlap and noise) (Malinin and Gales, 2019). This uncertainty metric is given by:

$$H(\mathbb{E}_{f_{\boldsymbol{\alpha}}(\mathbf{p})}[P(y|\mathbf{p})]) = -\sum_j \frac{\alpha_j}{\alpha_0} \log \frac{\alpha_j}{\alpha_0}$$

The mutual information between the labels y and the class probability vector $\mathbf{p}$, $I(y, \mathbf{p}|\mathbf{x}^*; w)$, captures knowledge uncertainty, and can be calculated by subtracting the expected data uncertainty from the total uncertainty:

$$\begin{aligned}I(y, \mathbf{p}|\mathbf{x}^*; w) &= H(\mathbb{E}_{f_{\boldsymbol{\alpha}}(\mathbf{p})}[P(y|\mathbf{p})]) - \mathbb{E}_{f_{\boldsymbol{\alpha}}(\mathbf{p})}[H(P(y|\mathbf{p}))] \\ &= -\sum_j \frac{\alpha_j}{\alpha_0}\left(\log \frac{\alpha_j}{\alpha_0} - \psi(\alpha_j + 1) + \psi(\alpha_0 + 1)\right)\end{aligned}$$

This metric is useful when measuring uncertainty for out-of-distribution or adversarial examples, and a variation of it was used in the context of active learning (Houlsby et al., 2011).

2.4 TECHNICAL PROOFS

We make use of the following lemmas in the proofs.

Lemma 1. *Consider the digamma function ψ. Assuming $x_1 > x_2 > 1$ and $p > 0$, the following inequality strictly holds:*

$$0 < \psi(x_1 + p) - \psi(x_2 + p) < \psi(x_1) - \psi(x_2)$$

Furthermore, we have $\lim_{x\to\infty} \psi(x + p) - \psi(x) = 0$.

Proof. Since $x_1 > x_2 > 1$, we can write $x_1 = s_1 + 1$ and $x_2 = s_2 + 1$ for some $s_1 > s_2$. Upon substitution of the Gauss integral representation $\psi(z + 1) = -\gamma + \int_0^1 \left(\frac{1-t^z}{1-t}\right) dt$ (here γ is the Euler-Mascheroni constant), we have:

$$\psi(x_1) - \psi(x_2) = \int_0^1 \left(\frac{t^{s_2} - t^{s_1}}{1 - t}\right) dt$$

which is strictly positive since the integrand is positive for $t \in (0, 1)$. Using the integral representation again, the inequality $\psi(x_1 + p) - \psi(x_2 + p) < \psi(x_1) - \psi(x_2)$ is equivalent to:

$$\int_0^1 \left(\frac{(1 - t^p)(t^{s_2} - t^{s_1})}{1 - t}\right) > 0$$

which holds since the integrand is positive due to $t^p < 1$ an $t^{s_1} < t^{s_2}$. The limit of $\psi(x+p)-\psi(x)$ follows from the asymptotic expansion $\psi(x) = \log(x) - \frac{1}{2x} + O\left(\frac{1}{x^2}\right)$, which yields $\psi(x+p)-\psi(x) \sim \log(1 + p/x) - \frac{1}{2(x+p)} + \frac{1}{2x} \to 0$ as $x \to \infty$. This concludes the proof. □

Lemma 2. *Consider the polygamma function of order 1 $\psi^{(1)}(z) = \frac{d}{dz}\psi(z)$. Assuming $x_1 > x_2 > 1$ and $p > 0$, the following inequality strictly holds:*

$$\psi^{(1)}(x_1) - \psi^{(1)}(x_2) < \psi^{(1)}(x_1 + p) - \psi^{(1)}(x_2 + p) < 0$$

Proof. Proceeding similarly as in the Proof of Lemma 1, we write $x_1 = s_1 + 1$ and $x_2 = s_2 + 1$ for some $s_1 > s_2$. Upon substitution of the integral representation $\psi^{(1)}(z+1) = \int_0^1 \left(\frac{t^z}{1-t} \ln\left(\frac{1}{t}\right)\right) dt$, we have:

$$\psi^{(1)}(x_1) - \psi^{(1)}(x_2) = \int_0^1 \left(\frac{t^{s_1} - t^{s_2}}{1 - t} \ln\left(\frac{1}{t}\right)\right) dt$$

which is strictly negative since the integrand is negative for $t \in (0, 1)$. Using the integral representation again, the inequality $\psi^{(1)}(x_1) - \psi^{(1)}(x_2) < \psi^{(1)}(x_1 + p) - \psi^{(1)}(x_2 + p)$ is equivalent to:

$$\int_0^1 \left(\frac{(1 - t^p)(t^{s_1} - t^{s_2})}{1 - t} \ln\left(\frac{1}{t}\right)\right) < 0$$

which holds true since $\ln(1/t) > 0$ for $t \in (0, 1)$. This concludes the proof. □

2.4.1 Proof of Theorem 1

Proof. Taking the logarithm of $\mathcal{F}_i$, we have:

$$\begin{aligned}\log \mathcal{F}_i &= \frac{1}{p}\log\left(\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0+p)}\right)\\ &+\frac{1}{p}\log\left(\frac{\Gamma(\sum_{k\neq c}\alpha_k+p)}{\Gamma(\sum_{k\neq c}\alpha_k)}+\sum_{j\neq c}\frac{\Gamma(\alpha_j+p)}{\Gamma(\alpha_j)}\right)\end{aligned}$$

where the second term is independent of α_c. Letting the first term be denoted as $g(\alpha_c) := \frac{1}{p}\log\left(\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0+p)}\right)$, it suffices to show $f(\alpha_c) := \exp(g(\alpha_c))$ is strictly convex and decreasing in α_c.

Differentiating $g(\alpha_c)$ twice we obtain:

$$\begin{aligned}g'(\alpha_c) &= \frac{1}{p}\left(\psi(\alpha_0)-\psi(\alpha_0+p)\right)\\ g''(\alpha_c) &= \frac{1}{p}\left(\psi^{(1)}(\alpha_0)-\psi^{(1)}(\alpha_0+p)\right)\end{aligned}$$

Lemmas 1 and 2 then yield that $g'(\alpha_c) < 0$ and $g''(\alpha_c) > 0$ respectively. Differentiating $f(\alpha_c)$ twice, we have:

$$f'(\alpha_c) = e^{g(\alpha_c)}g'(\alpha_c) \tag{3}$$

$$f''(\alpha_c) = e^{g(\alpha_c)}\left(g''(\alpha_c)+(g'(\alpha_c))^2\right) \tag{4}$$

Using the inequalities above and the positivity of $e^{g(\alpha_c)}$, it follows that $f'(\alpha_c) < 0$ and $f''(\alpha_c) > 0$. Thus, $f(\alpha_c)$ is a strictly convex decreasing function in α_c. This concludes the proof. □

2.4.2 Proof of Theorem 2

Proof. Consider a concentration parameter α_j corresponding to an incorrect class, i.e., $j \neq c$. Define the ratio of Gamma functions as:

$$\mu(\alpha) \overset{\text{def}}{=} \frac{\Gamma(\alpha+p)}{\Gamma(\alpha)}$$

This function is positive, increasing and convex with derivative given by:

$$\begin{aligned}\mu'(\alpha) &= -\frac{\Gamma(\alpha+p)\Gamma'(\alpha)}{\Gamma(\alpha)^2}+\frac{\Gamma'(\alpha+p)}{\Gamma(\alpha)}\\ &= -\frac{\Gamma(\alpha+p)\psi(\alpha)}{\Gamma(\alpha)}+\frac{\Gamma(\alpha+p)\psi(\alpha+p)}{\Gamma(\alpha)}\\ &= \mu(\alpha)\left(\psi(\alpha+p)-\psi(\alpha)\right)\\ &= \mu(\alpha)\nu(\alpha) \qquad (5)\end{aligned}$$

where we used the relation $\Gamma'(z) = \Gamma(z)\psi(z)$ and defined

$$\nu(\alpha) \overset{\text{def}}{=} \psi(\alpha + p) - \psi(\alpha).$$

From Lemma 1, it follows that $\nu(\alpha) > 0$ which implies $\mu(\alpha)$ is increasing.

Since $(\cdot)^{1/p}$ is a continuous increasing function, it suffices to show the objective $\mathcal{G} = \mathcal{F}_i^p$ is increasing, given by

$$\mathcal{G}(\alpha_j) = \frac{\mu\left(\sum_{l \neq c} \alpha_l\right) + \sum_{l \neq c} \mu(\alpha_l)}{\mu(\alpha_0)}$$

The derivative is then calculated as:

$$\mathcal{G}'(\alpha_j) = \frac{\mu'\left(\sum_{l \neq c} \alpha_l\right) + \mu'(\alpha_j)}{\mu(\alpha_0)} - \frac{\mu'(\alpha_0) \cdot \left[\mu\left(\sum_{l \neq c} \alpha_l\right) + \sum_{l \neq c} \mu(\alpha_l)\right]}{\mu(\alpha_0)}$$

The condition $\mathcal{G}'(\alpha_j) > 0$ is equivalent to:

$$\frac{\mu'\left(\sum_{l \neq c} \alpha_l\right) + \mu'(\alpha_j)}{\mu'(\alpha_0)} > \frac{\mu\left(\sum_{l \neq c} \alpha_l\right) + \sum_{l \neq c} \mu(\alpha_l)}{\mu(\alpha_0)} = \mathcal{G}$$

Upon substituting the expression (5), this condition becomes:

$$\mu\left(\sum_{l \neq c} \alpha_l\right) \nu\left(\sum_{l \neq c} \alpha_l\right) + \mu(\alpha_j)\nu(\alpha_j) > \left[\mu\left(\sum_{l \neq c} \alpha_l\right) + \sum_{l \neq c} \mu(\alpha_l)\right] \nu(\alpha_0) \tag{6}$$

From Lemma 1, it follows that $\nu\left(\sum_{l \neq c} \alpha_l\right) > \nu(\alpha_0)$ and $\nu(\alpha_j) > \nu(\alpha_0)$. In addition, the functions $\mu\left(\sum_{l \neq c} \alpha_l\right) \nu\left(\sum_{l \neq c} \alpha_l\right)$ and $\mu(\alpha_j)\nu(\alpha_j)$ are both increasing as α_j grows. Using these results and the fact that $\left[\sum_{l \neq c,j} \mu(\alpha_l)\right] \nu(\alpha_0) \to 0$ as α_j grows (due to Lemma 1), it follows that the inequality (6) holds true for large α_j. Thus, we conclude that the loss function is increasing as α_j gets large. The proof is complete. □

An illustration of Theorem 2 is shown in Fig. 3 below. An approximate loss function is also shown due to $\lim_{\alpha \to \infty} \frac{\Gamma(\alpha+p)}{\Gamma(\alpha)\alpha^p} = 1$, from which we obtain the approximation $\mu(\alpha) \sim \alpha^p$. This approximation to the loss behaves similarly. Despite the initial dip, the loss is increasing as α_j increases. We remark that the loss is neither convex nor concave in α_j.

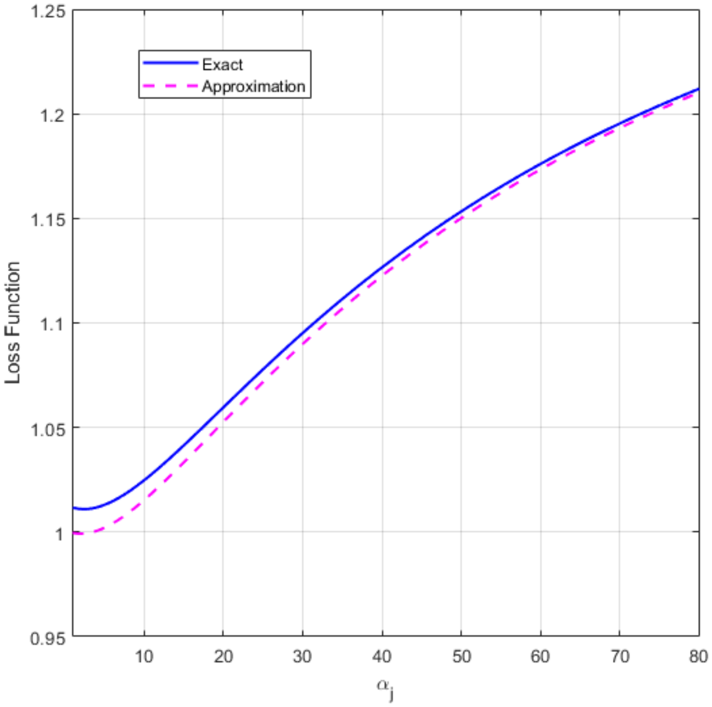


Figure 3. Illustrative example for Theorem 2. Here, the loss function $\mathcal{F}_i$ is plotted as a function of α_j, $j \neq c$. Parameters $p = 2$ and a random $\boldsymbol{\alpha}$ vector were used for $K = 10$ classes with α_c small relative to other concentration parameters. As Theorem 2 shows, the loss is increasing for large α_j.

2.4.3 Proof of Theorem 3

Proof. From Haussler and Opper (1997) (p. 2472), we have

$$\frac{\partial}{\partial \alpha'_i} D_u^R(f_{\boldsymbol{\alpha}} \parallel f_{\boldsymbol{\alpha}'})|_{\boldsymbol{\alpha}'=\boldsymbol{\alpha}} = 0$$
$$\frac{\partial^2}{\partial \alpha'_i \partial \alpha'_j} D_u^R(f_{\boldsymbol{\alpha}} \parallel f_{\boldsymbol{\alpha}'})|_{\boldsymbol{\alpha}'=\boldsymbol{\alpha}} = uJ_{ij}(\boldsymbol{\alpha})$$

and using Taylor's expansion to second order:

$$D_u^R(f_{\boldsymbol{\alpha}} \parallel f_{\boldsymbol{\alpha}'}) = \frac{u}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}')^T J(\boldsymbol{\alpha})(\boldsymbol{\alpha} - \boldsymbol{\alpha}') + O(\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_2^3) \tag{7}$$

where $J(\boldsymbol{\alpha})$ is the Fisher information matrix corresponding to the Dirichlet distribution. Taking the logarithm of the density, we have $\log f_{\boldsymbol{\alpha}}(\mathbf{p}) = \log \Gamma(\alpha_0) - \sum_j \log \Gamma(\alpha_j) + \sum_j (\alpha_j - 1) \log p_j$, and differentiating twice we obtain:

$$\frac{\partial}{\partial \alpha_i} \log f_{\boldsymbol{\alpha}}(\mathbf{p}) = \psi(\alpha_0) - \psi(\alpha_i) + \log p_i$$
$$\frac{\partial^2}{\partial \alpha_i \partial \alpha_j} \log f_{\boldsymbol{\alpha}}(\mathbf{p}) = \psi^{(1)}(\alpha_0) - \psi^{(1)}(\alpha_i) I_{\{i=j\}}$$

Since $J_{ij}(\boldsymbol{\alpha}) = -\mathbb{E}[\frac{\partial^2}{\partial \alpha_i \partial \alpha_j} \log f_{\boldsymbol{\alpha}}(\mathbf{p})]$, we obtain:

$$J(\boldsymbol{\alpha}) = \mathrm{diag}(\{\psi^{(1)}(\alpha_i)\}_{i=1}^{K}) - \psi^{(1)}(\alpha_0) 1_{K \times K}$$

Here $\psi^{(1)}$ is the first order polygamma function. Substituting this into (7) and simplifying, we obtain:

$$\begin{aligned} D_u^R(f_{\boldsymbol{\alpha}} \parallel f_{\boldsymbol{\alpha}'}) &\cong \frac{u}{2} \sum_{i \neq c} \sum_{j \neq c} (\alpha_i - 1)(\alpha_j - 1) J_{ij}(\boldsymbol{\alpha}) \\ &= \frac{u}{2} \left(\sum_{i \neq c} (\alpha_i - 1)^2 \psi^{(1)}(\alpha_i) - (\sum_{i \neq c} (\alpha_i - 1))^2 \psi^{(1)}(\alpha_0) \right) \end{aligned}$$

□

3. EXPERIMENTAL RESULTS ON MNIST DATASET

We follow the same experimental setup as in (Louizos and Welling, 2017) and (Sensoy et al., 2018). The MNIST dataset (LeCun et al.) is a popular benchmark for image classification. The task is to classify grayscale images of handwritten digits (size 28×28) into 10 classes. The training set contains $60,000$ images and the testing set contains $10,000$ images. Image pixel intensities were normalized to $[0, 1]$ range. It is well known that convolutional neural networks (CNNs) achieve the highest accuracies for this dataset.

Comparisons are made with the following methods: (a) L2 corresponds to deterministic neural network with softmax output and weight decay, (b) Dropout is the uncertainty estimation method of (Gal and Ghahramani, 2016), (c) Deep Ensemble is the non-Bayesian approach of (Lakshminarayanan et al., 2017), (d) FFG is the BNN used in (Blundell et al., 2015), (e) FFLU is the BNN used in (Kingma et al., 2015) with the additive parameterization (Molchanov et al., 2017), (f) MNFG is the multiplicative normalizing flow VI inference method in (Louizos and Welling, 2017), (g) PN is the reverse KL divergence-based prior network method of (Malinin and Gales, 2019), (h) EDL is the evidential approach of (Sensoy et al., 2018) and (i) IRD is our proposed technique.

3.1 NETWORK ARCHITECTURE

The LeNet CNN architecture with 20 and 50 filters of size 5×5 is used for the MNIST dataset with 500 hidden units at the dense layer. In our implementation of PN and IRD, FGSM adversarial examples were generated using $\epsilon = 0.1$ noise. Hyperparameter values $u = 2.0, \lambda = 0.5, \gamma = 0.1$ were used to generate these results with $p = 15$.

3.2 ACCURACY

Table 1 shows the test accuracy on MNIST for these methods; IRD is shown to be competitive assigning low uncertainty to correct predictions and high uncertainty to misclassifications.

3.3 UNCERTAINTY ESTIMATION

Fig. 4 shows the distribution of entropies of predictive distributions for correct and misclassified examples across competing methods. The overconfidence of softmax NNs is evident since both correct and wrong entropy distributions are concentrated on lower uncertainties. The Dirichlet-based methods, EDL and PN, are better calibrated offering a good balance between correct and misclassified entropies. IRD offers a drastic improvement over all methods with 90% of the misclassified samples falling within 95% of the max-entropy ($\log 10 \approx 2.3$), as opposed to 58% and 5% of the misclassified samples of the PN and EDL methods respectively. Predictive uncertainty can also be measured in terms of the inverse Dirichlet strength K/α_0 which captures the spread of the Dirichlet distribution. Fig. 5 shows the resulting performance of various algorithms.

IRD is tested on notMNIST (Bulatov, 2011) which contains only letters serving as out-of-distribution data. The uncertainty is expected to be high for all such images as letters do not fit into

TABLE 1

MNIST Dataset: Test accuracy (%), median % max-entropy for correct and misclassified examples for various deep learning methods

Method	Accuracy	Median %Max-Entropy - Correct	Median %Max-Entropy - Misclassified
L2	99.4	-	-
Dropout	99.5	-	-
Deep Ensemble	99.3	-	-
FFG	99.1	-	-
FFLU	99.1	-	-
MNFG	99.3	-	-
PN	99.3	19.5	56.7
EDL	99.2	24.9	99.6
IRD	98.2	6.4	100.0

any digit category. Fig. 6 shows the empirical CDF of the predictive entropy for all models. CDF curves close to the bottom right are more desirable as higher entropy is desired for all predictions. IRD is much more tightly concentrated towards higher entropy values with an impressive 96% of letter images having entropy larger than 95% of the max-entropy, while EDL and PN have 61% and 63% approximately. Fig. 7 compares all Dirichlet neural network using the mutual information metric that measures distributional uncertainty on the notMNIST anomaly detection task. As expected from Fig. 6 that shows IRD has the largest total uncertainty, the same relative trend continues.

Fig. 8 shows the adversarial performance when each model is evaluated using adversarial examples generated with the Fast Gradient Sign method (FGSM) (Goodfellow et al., 2014) for different noise values ϵ, i.e.,

$$\mathbf{x}_{adv} = \mathbf{x} + \epsilon \text{sgn}(\nabla_{\mathbf{x}} \mathcal{F}(\mathbf{x}, y, w)).$$

We observe that IRD achieves higher entropy on adversarial examples as ϵ increases. Dropout outperforms other BNN methods at the expense of overconfident predictions. While PN asymptotically achieves very high uncertainty as well to the same level as IRD, we remark that IRD achieves a lower average predictive entropy for $\epsilon = 0$ due to the higher confidence of correct predictions and assigns a large entropy to misclassified samples as Fig. 4 also supports. Fig. 9 compares all Dirichlet distribution-based uncertainty estimation methods using the mutual information metric that measures distributional uncertainty on FGSM adversarial examples detection task. Figs. 10 and 11 show sample digits generated using the FGSM method once the IRD and L2 networks are trained. From left to right, the noise level ϵ is increasing from 0 to 1. We note slightly different things are happening when adversarial noise is added to the images since the adversarial noise is dependent on what features the network has learned to extract and the classification cost function.

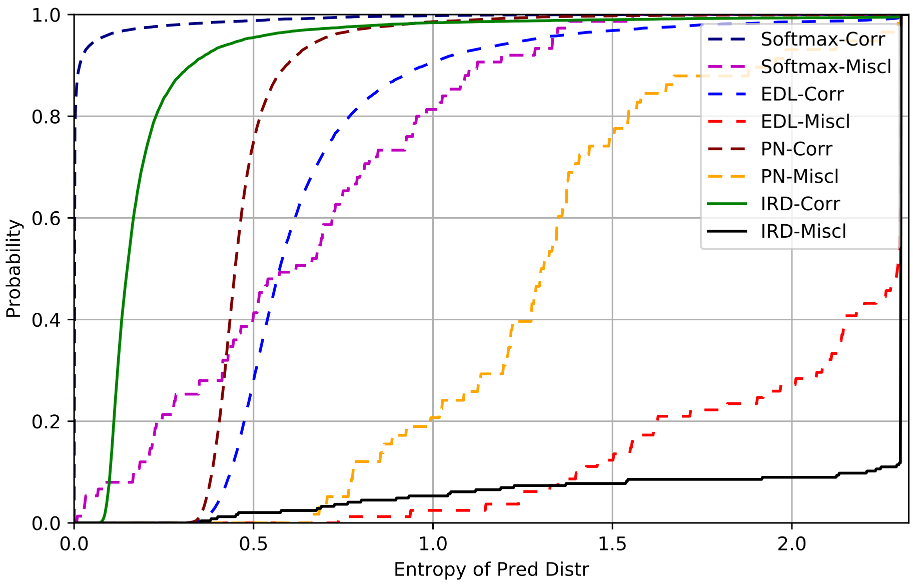


Figure 4. Empirical CDF of predictive distribution entropy on MNIST dataset.

Fig. 12 shows the adversarial performance of the Dirichlet-based methods (the most competitive ones) on examples generated with the projected gradient descent (PGD) method (Kurakin et al., 2017) for different noise levels ϵ, i.e.,

$$\begin{aligned}\mathbf{x}_{adv}^{t+1} &= \Pi_{\mathbf{x}+\epsilon\ell_\infty}(\mathbf{x}_{adv}^{t} + \alpha \mathrm{sgn}(\nabla_{\mathbf{x}}\mathcal{F}(\mathbf{x}_{adv}^{t}, y, w)))\\ \mathbf{x}_{adv}^{0} &= \mathbf{x}.\end{aligned}$$

Here, $\Pi_{\mathbf{x}+\epsilon\ell_\infty}(\cdot)$ is the projection onto the ℓ_∞ ball of size ϵ centered at $\mathbf{x}$. This multi-step variant of FGSM uses a small step size $\alpha = 0.01$ over $T = 40$ steps. We observe that IRD achieves the highest uncertainty on PGD adversarial examples as the noise level increases while PN asymptotically achieves a mid-range uncertainty, EDL is inconsistent and Softmax NNs cannot reliably detect these stronger attacks. We further remark that IRD has lower predictive entropy for $\epsilon = 0$ due to the higher confidence of correct predictions as Fig. 4 also shows.

Figs. 13 and 14 show sample digits generated using the PGD method once the IRD and L2 networks are trained. From left to right, the noise level ϵ is increasing from 0 to 0.5. We note that the PGD attacks are harder to detect than the simpler one-step FGSM attacks. IRD reliably detects those. Also, we note that the adversarial attacks based on the standard L2 network can be visually detected well for larger enough ϵ, but they are harder to detect for the IRD network, again highlighting the fact that the IRD network has learned different patterns and has a more robust classifier.

The adversarial experiment with PGD was repeated for a larger step size of 0.03 with 40 steps to examine trends as larger adversarial steps are taken. Fig. 15 shows the accuracy, predictive entropy and mutual information metrics as the noise level ϵ increases from 0 to 0.5. IRD continues

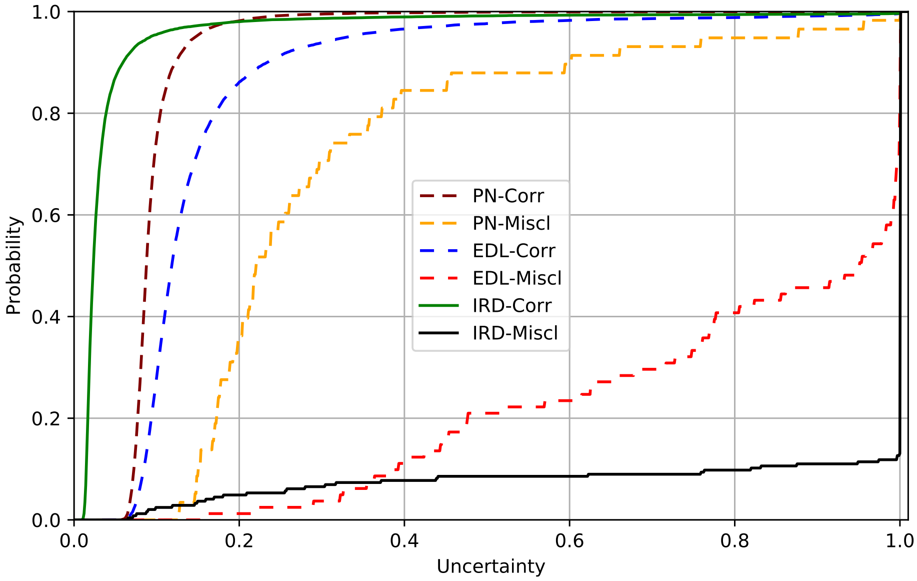


Figure 5. Empirical CDF of inverse Dirichlet strength for correct and incorrect predictions on MNIST dataset.

to be the most confident method in terms of detecting these PGD adversarial examples, while PN and EDL lag behind despite the fact that the accuracy of all methods decays as noise increases. Figs. 16 and 17 show sample digits generated using the PGD method once the IRD and L2 networks are trained. From left to right, the noise level ϵ is increasing from 0 to 0.5. We note that the PGD attacks with a larger step size yield more aggressive noise patterns on the digits, still being much more subtle when compared to FGSM. IRD reliably detects those. Furthermore, the adversarial attacks based on the standard L2 network can be visually detected well for larger enough ϵ and shaping of digits into patterns of other digits are apparent, but they are much harder to detect for the IRD network.

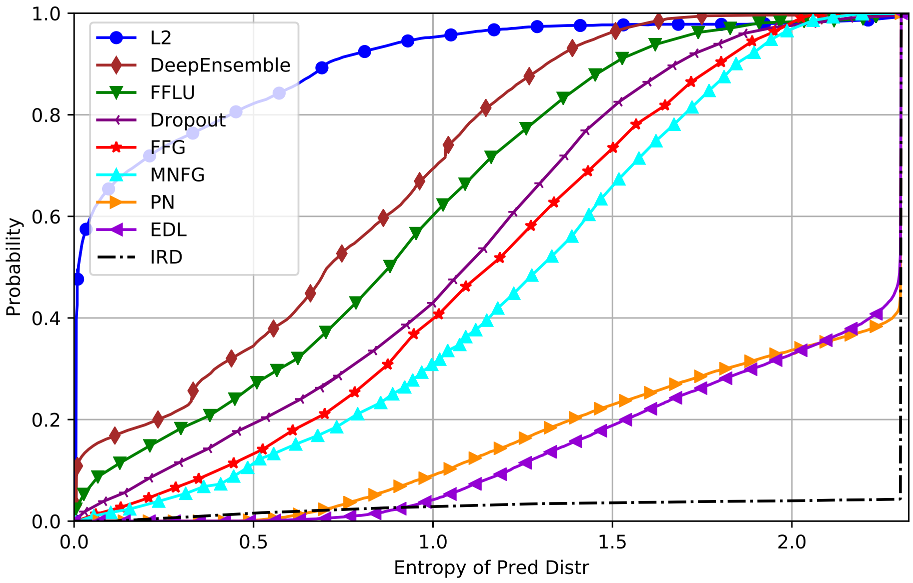


Figure 6. Empirical CDF of predictive distribution entropy on notMNIST dataset.

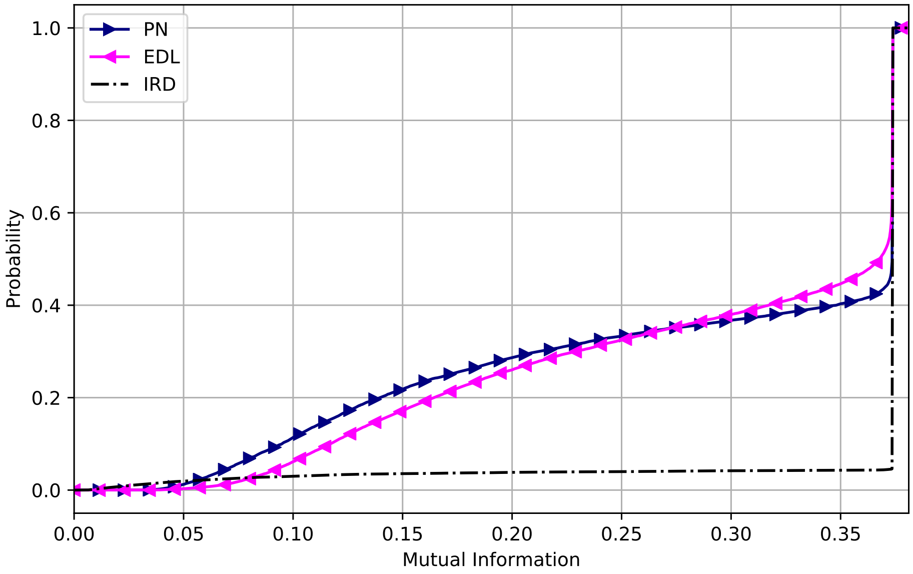


Figure 7. Empirical CDF of mutual information between labels $\mathbf{y}$ *and categorical distribution* $\boldsymbol{p}$ *out-of-distribution notMNIST dataset.*

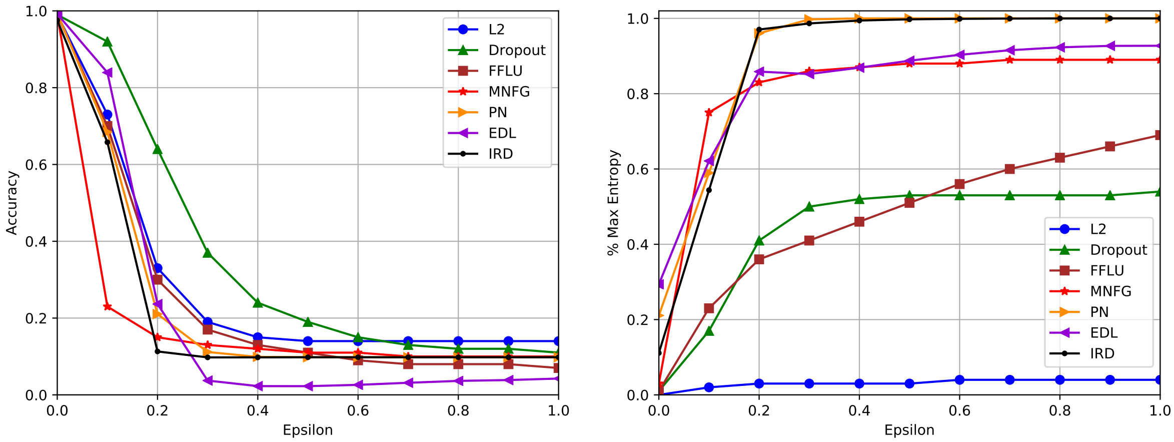


Figure 8. *Test accuracy (left) and predictive entropy (right) for FGSM adversarial examples as a function of adversarial noise ϵ on MNIST dataset.*

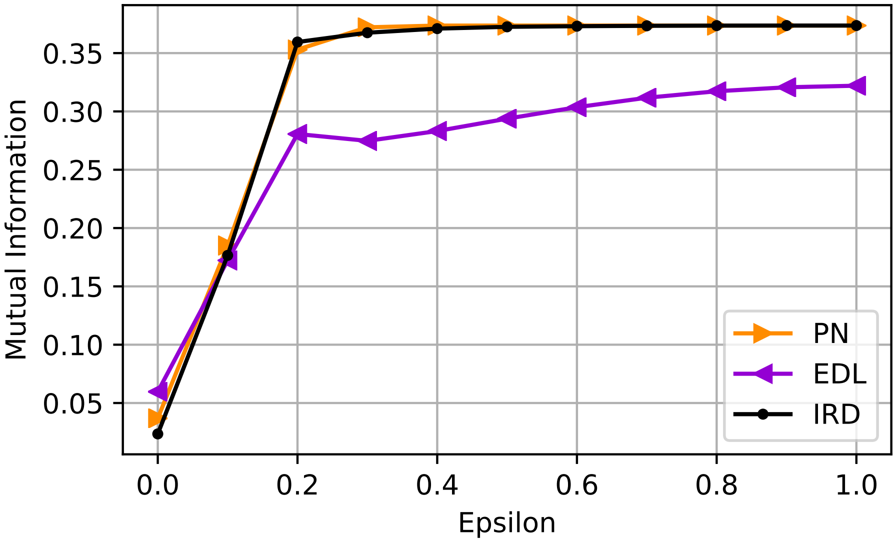


Figure 9. *Empirical CDF of mutual information between labels* $\mathbf{y}$ *and categorical distribution* $\boldsymbol{p}$ *out-of-distribution notMNIST dataset.*

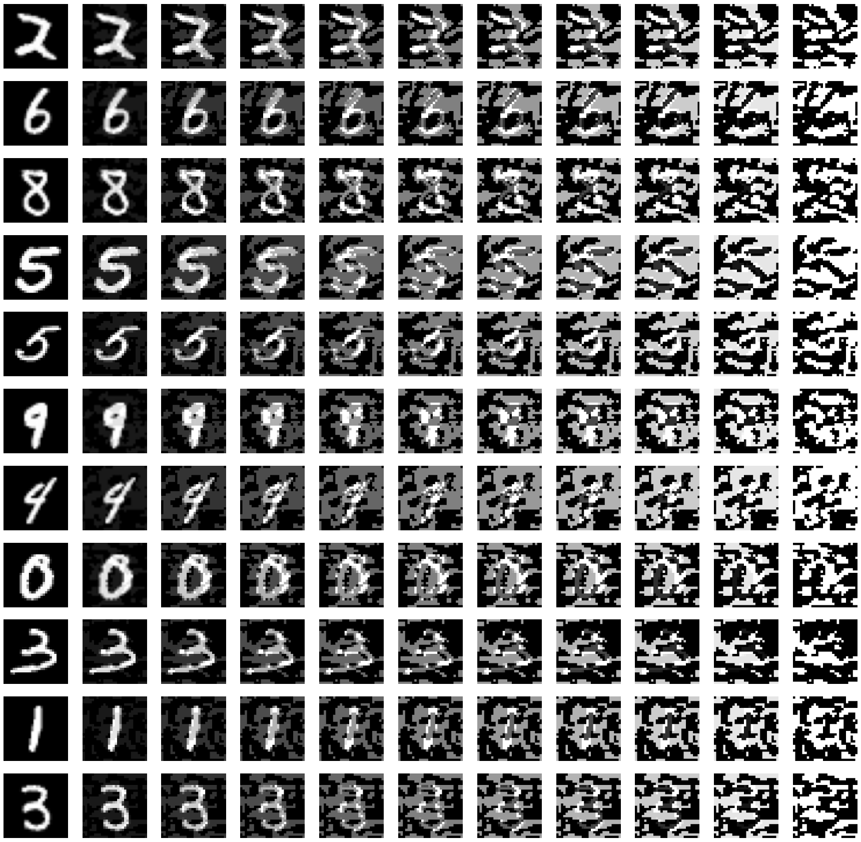


Figure 10. Sample MNIST adversarial digits generated with the FGSM method for the IRD network.

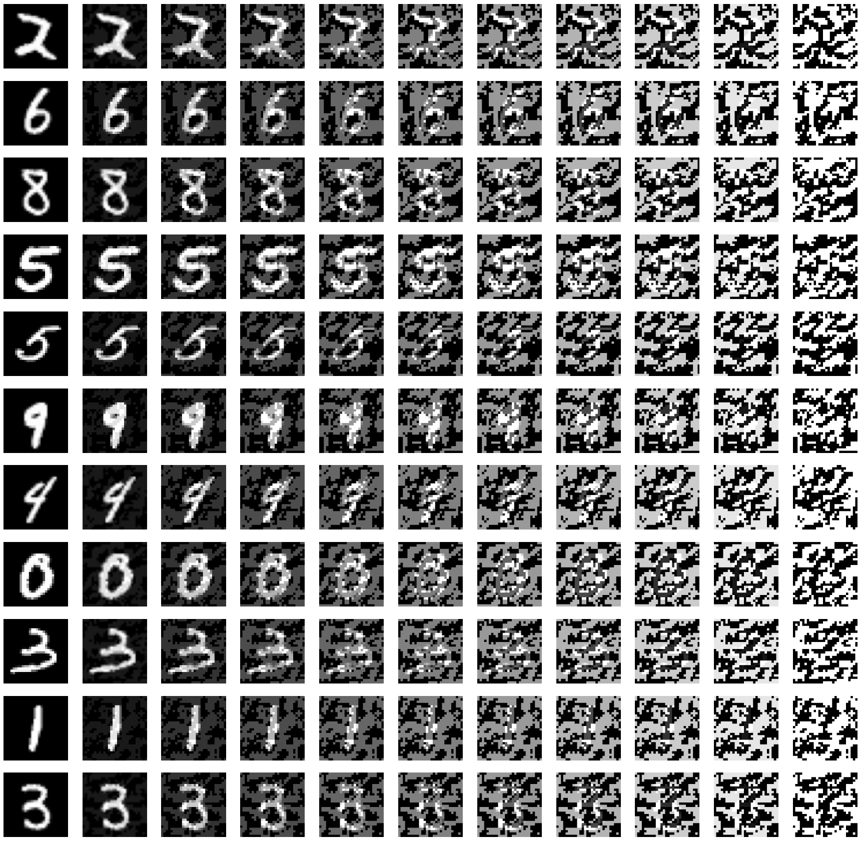


Figure 11. Sample MNIST adversarial digits generated with the FGSM method for L2 network.

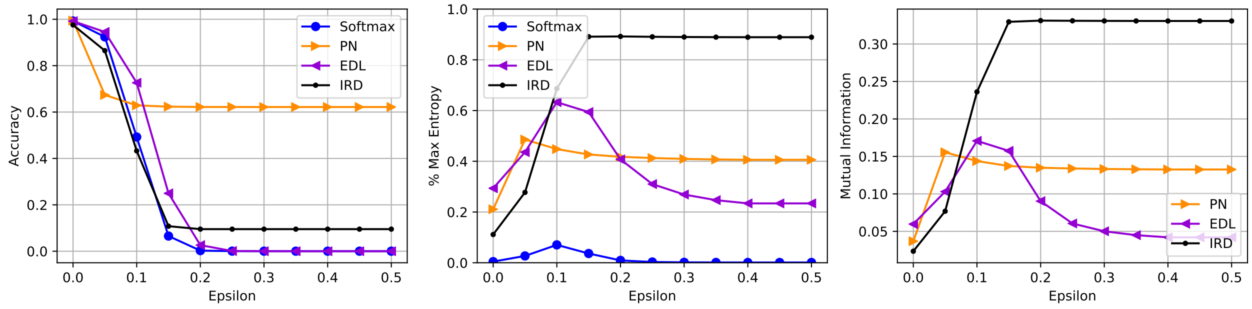


Figure 12. Test accuracy (left), predictive entropy (middle), and mutual information (right) for PGD adversarial examples as a function of adversarial noise ϵ on MNIST dataset.

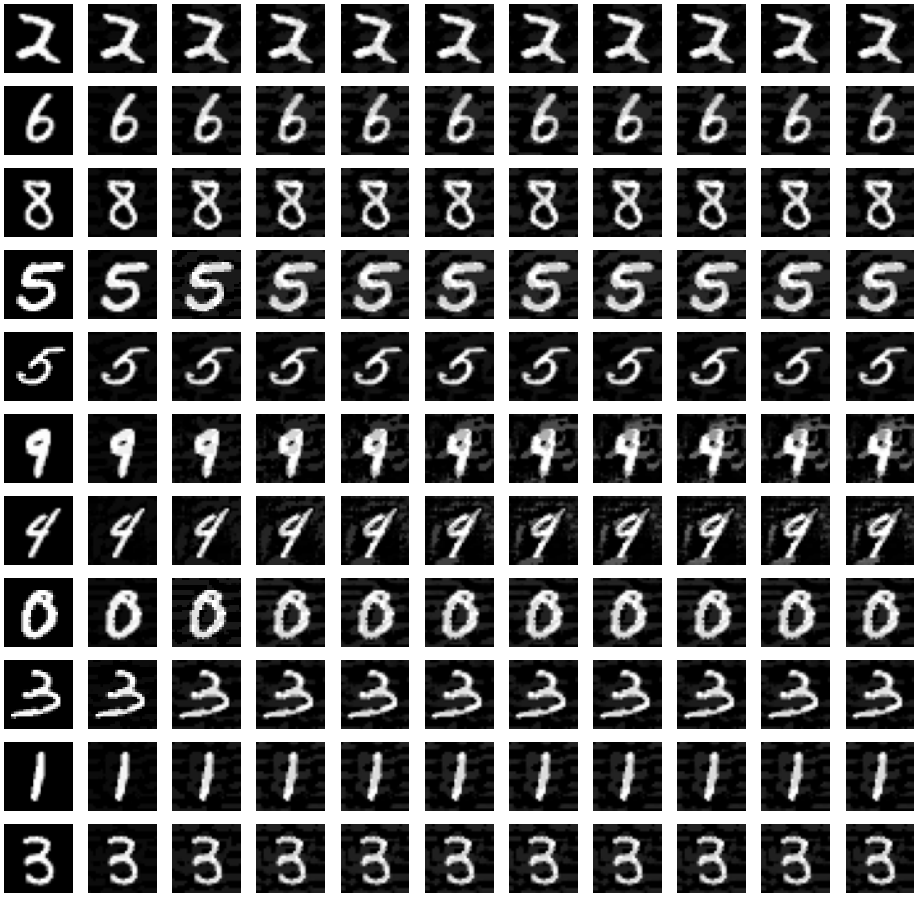


Figure 13. Sample MNIST adversarial digits generated with the PGD method for the IRD network. Here 40 *steps were taken with a step size of* 0.01*.*

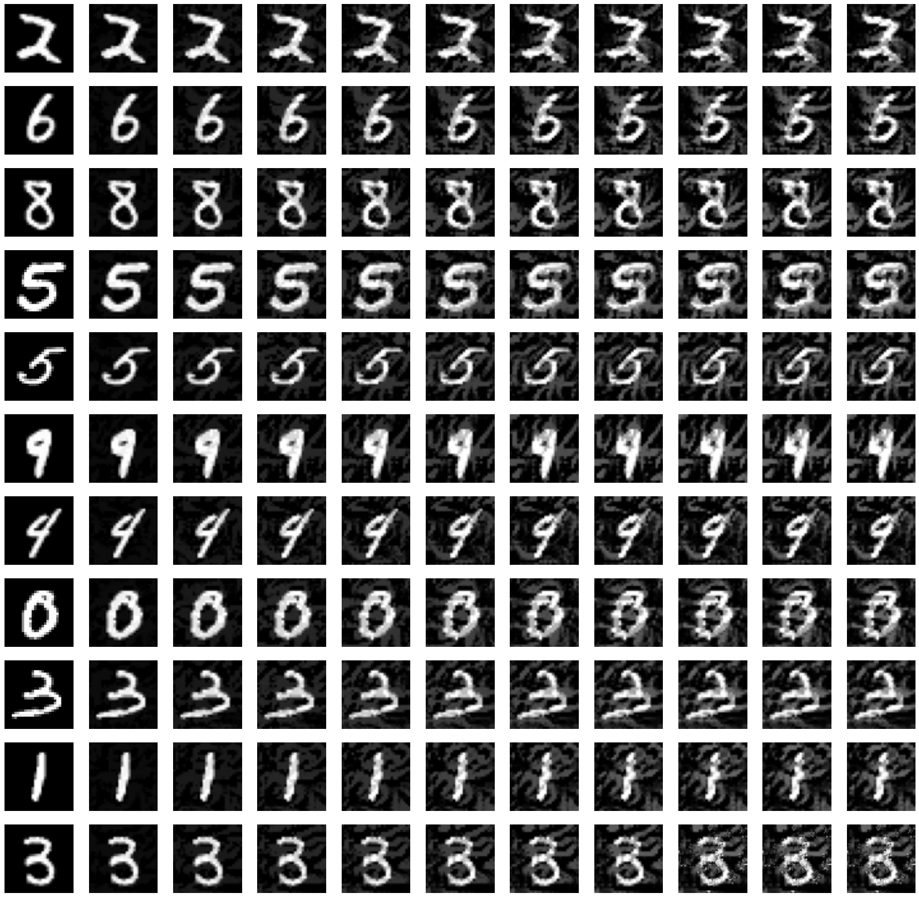


Figure 14. Sample MNIST adversarial digits generated with the PGD method for L2 network. Here 40 *steps were taken with a step size of* 0.01*.*

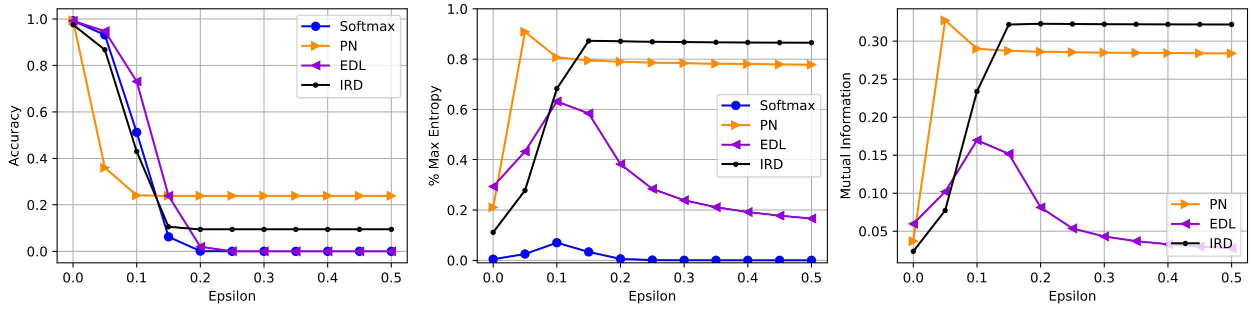


Figure 15. Test accuracy (left), predictive entropy (middle), and mutual information (right) for PGD adversarial examples as a function of adversarial noise ϵ on MNIST dataset. Here, 40 *steps were taken with step size* 0.03*.*

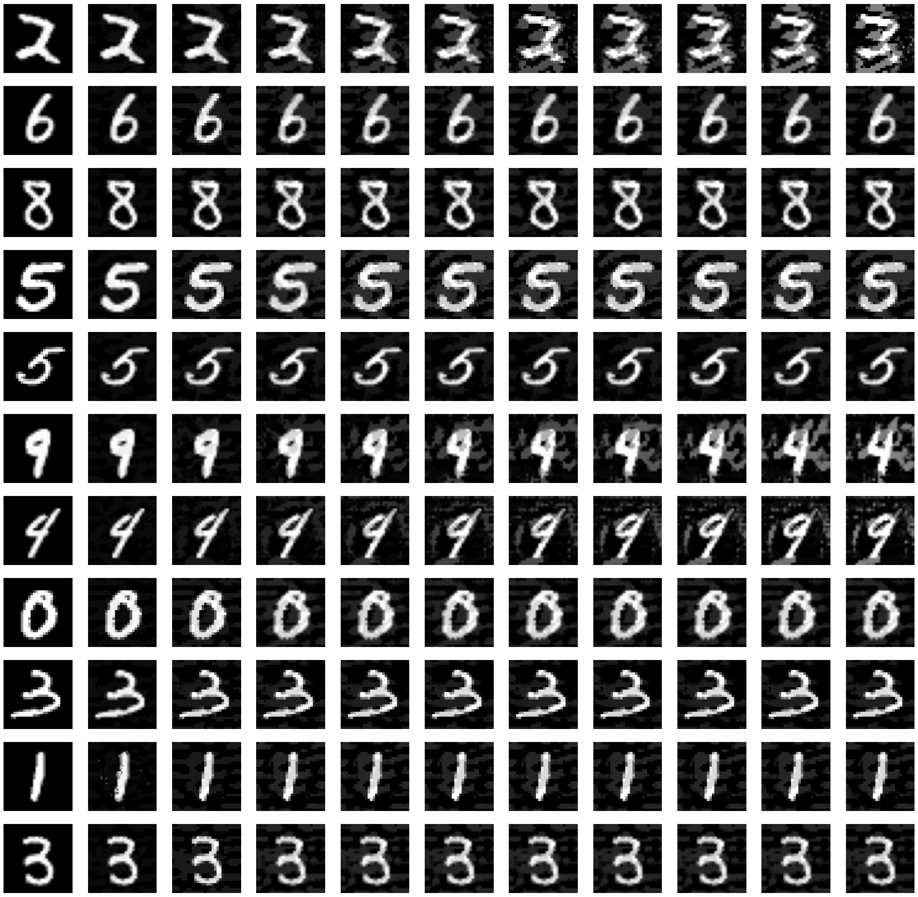


Figure 16. Sample MNIST adversarial digits generated with the PGD method for the IRD network. Here 40 *steps were taken with a step size of* 0.03.

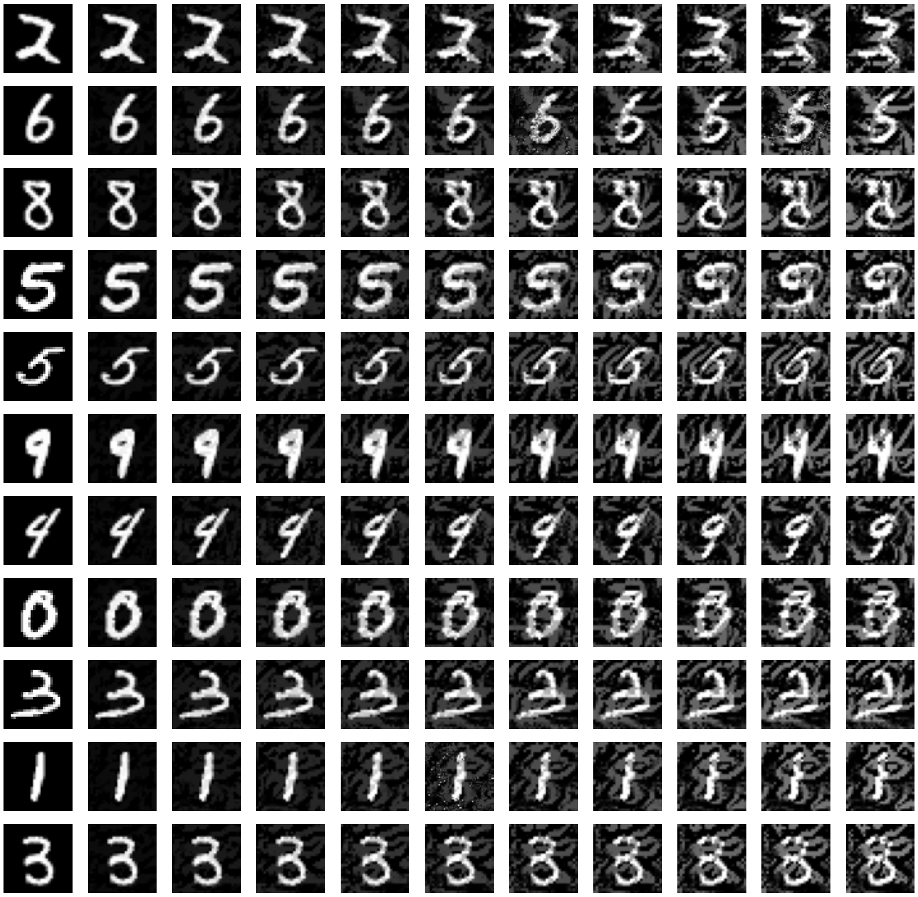


Figure 17. Sample MNIST adversarial digits generated with the PGD method for L2 network. Here 40 *steps were taken with a step size of* 0.03.

4. EXPERIMENTAL RESULTS ON PHYSIONET 17 CHALLENGE

The PhysioNet17 challenge dataset (Clifford et al., 2017) contains $5,707$ electrocardiogram (ECG) signals of length $9,000$ sampled at 300 samples/sec. The task is to classify a single short ECG lead recording into a normal sinus rhythm or atrial fibrillation (Afib). Atrial fibrillation is the most common sustained cardiac arrhythmia occurring when the heart's upper chambers beats out of synchronization with the lower chambers, and is hard to detect due to its episodic presence. The raw ECG signals were bandpass filtered for baseline wander removal, and then normalized to zero mean and unit variance over the 30s duration.

About 13% of the recordings correspond to Afib, and oversampling was used to account for class imbalance. A train/test split of 90/10% was used. As EDL and PN were shown to be most competitive with our method based on the benchmark image dataset shown above, we compare IRD with the L2, Dropout, PN and EDL methods.

4.1 NETWORK ARCHITECTURE

The CNN architecture consists of six 1D Conv layers with stride-2 max-pooling, with 8, 16, 32, 64, 128, 128 filters of sizes 9, 9, 7, 7, 5, 5 respectively, followed by a filter-wise sum-pooling layer, 100 hidden units with dropout and a binary classification layer.The IRD parameters used were $u = 0.95, \lambda = 2.3, \gamma = 0.07, \epsilon = 0.02$ with $p = 15$.

4.2 ACCURACY

The accuracies for various deep learning methods are shown in Table 2, and IRD achieves a high prediction accuracy on par with other methods. Furthermore, the last two columns of the table show that on average what % of the max-entropy correct and misclassified ECG signals are assigned by the algorithms. IRD obtains a great tradeoff between correct and misclassified prediction entropy.

TABLE 2

PhysioNet ECG Dataset: Test accuracy (%), median % max-entropy for correct and misclassified examples for various deep learning methods

Method	Accuracy	Median %Max-Entropy - Correct	Median %Max-Entropy - Misclassified
L2	94	1.7	81.4
Dropout	94	4.2	70.4
PN	96	15.1	65.0
EDL	95	23.4	59.5
IRD	95	10.2	100.0

4.3 UNCERTAINTY ESTIMATION

Fig. 18 shows the cumulative density function of the predictive entropy for correct and misclassified examples. The median entropy normalized by the maximum entropy is shown in the last two columns of Table 2, which reflects that IRD assigns very low uncertainty for correct classifications and large uncertainty to misclassifications. The tail of the entropy distribution of misclassified samples shows that IRD assigns entropy values larger than 90% of the max-entropy to 69% of the misclassified samples while L2, Dropout, PN and EDL methods assign that to only 44%, 27%, 3% and 37% of their misclassified examples respectively. Fig. 19 shows the empirical CDF of the inverse Dirichlet strength K/α_0 for various deep learning methods. This metric also captures uncertainty in predictions. The IRD method assigns large uncertainty to incorrect predictions, outperforming PN and EDL by a large margin, and low uncertainty to correct ones.

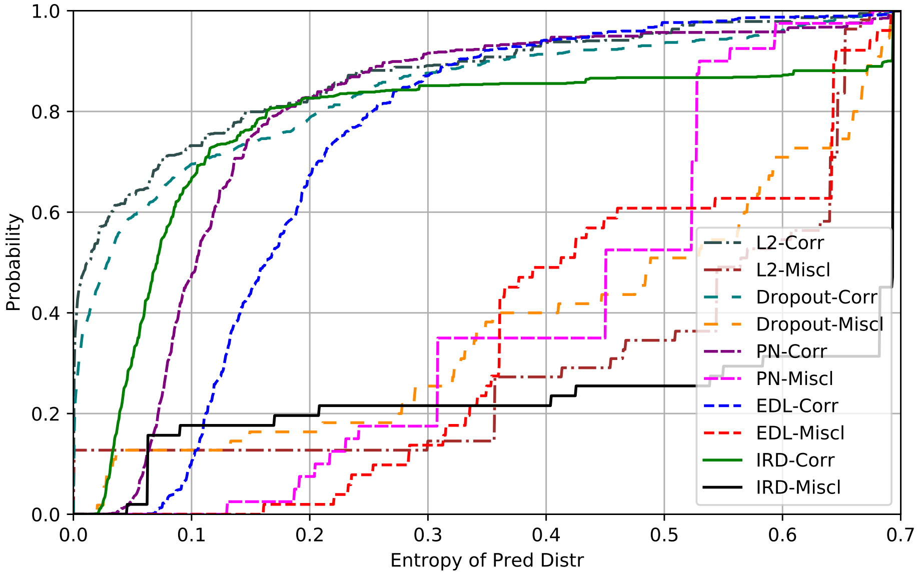


Figure 18. Empirical CDF of predictive entropy on correct and misclassified ECG signals for various deep learning methods.

Fig. 20 shows correct and misclassified ECG signals from the test set; the top plots show correctly classified normal rhythms (top two) and AFib (next two) signals with low prediction entropy, and the bottom two plots show incorrectly classified AFib signals characterized by high prediction entropy. It is evident that the algorithm correctly forms high-confidence opinions about signals that exhibit strong characteristics of normal heartbeat (e.g., regular occurrence with identifiable P wave, QRS complex and T wave) and AFib (e.g., irregular spacing of pulses with often a lack of a P wave). Visual inspection of the high-entropy misclassified signals show that although local peaks tend to be irregular hinting at AFib, but there is too much noise in the intermediate waves and transient irregularity to reliably classify them.

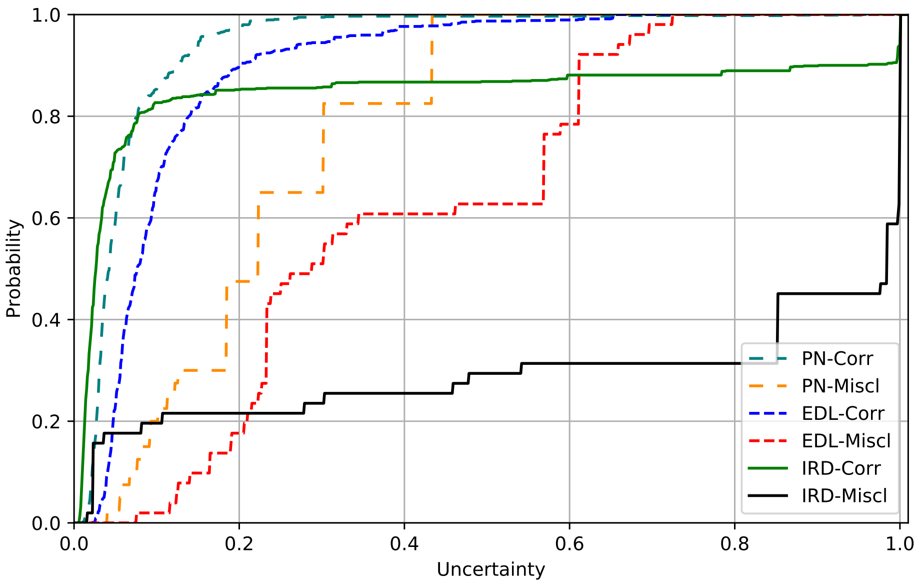
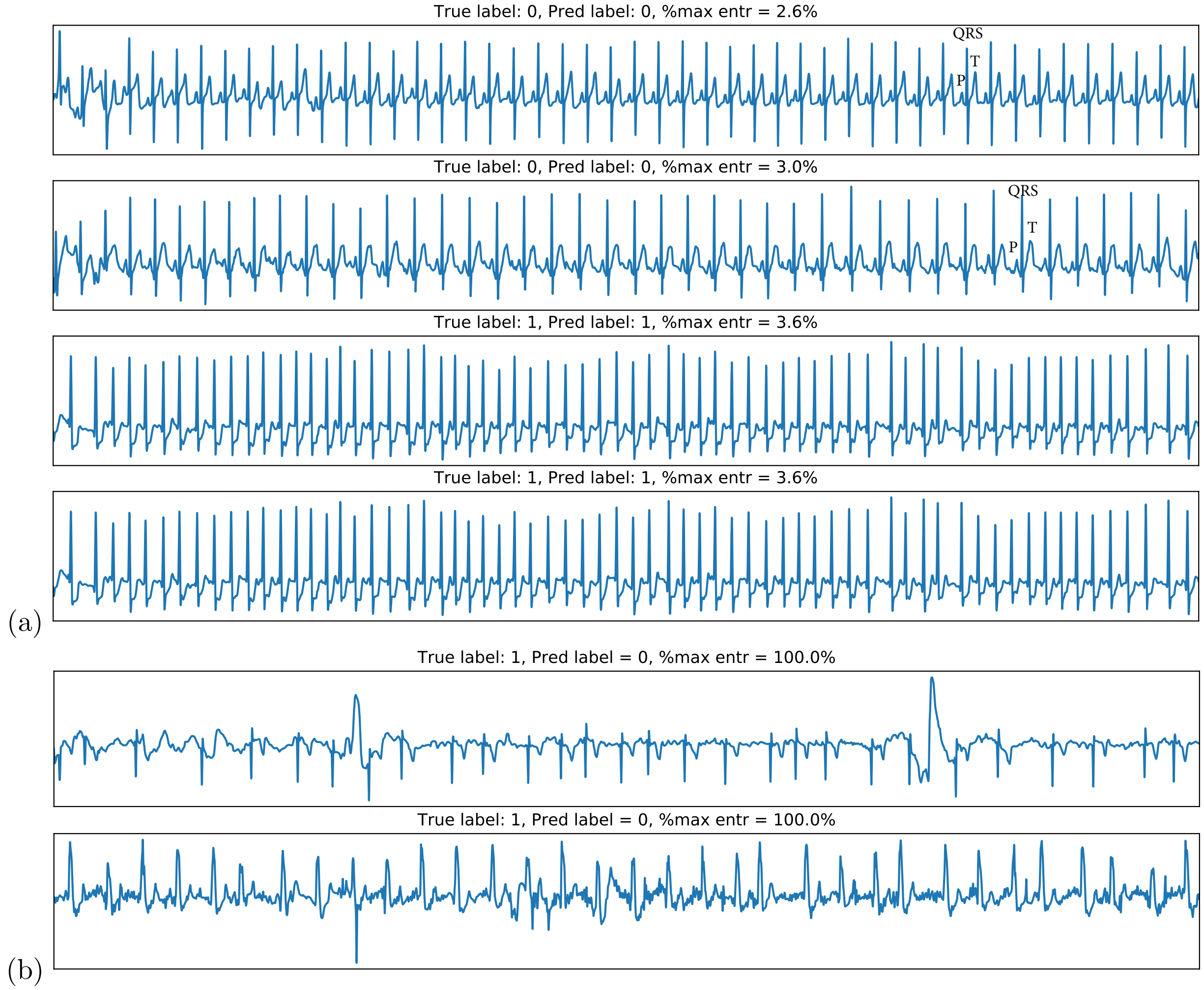


Figure 19. Empirical CDF of inverse Dirichlet strength for correct and incorrect predictions on PhysioNet ECG dataset.

To test detection of out-of-distribution signals, we constructed a modified dataset from the test set by adding sparse random noise (zero-mean Gaussian with $\sigma = 5$ at 5% of total time locations uniformly at random) followed by temporally smoothing the whole waveform with a 1D Gaussian filter of $\sigma = 15$. Fig. 21 contains several anomalous generated waveforms. Empirical CDFs of predictive entropy and mutual information are shown in Fig. 22, in which IRD outperforms other methods by a large margin. Specifically, IRD assigns a predictive entropy of 90% max-entropy or higher to 81% of the anomalous signals as opposed to $17\%, 27\%, 6\%, 20\%$ for L2, Dropout, PN and EDL methods respectively.

To further test the detection of anomalous signals, we used a subset of the full PhysioNet 17 challenge data labeled as “too noisy to be classified” by experts. A sample of these noisy ECG waveforms is shown in Fig. 23. It is evident that these waveforms are very hard to classify due to transients, low signal to noise ratios and inconsistent temporal statistical behavior. Fig. 24 shows the empirical CDFs of predictive entropy and mutual information. We observe that IRD assigns higher uncertainty to these noisy ECG recordings than other competing methods implying that the empirical entropy and mutual information distributions are concentrated towards higher uncertainties.



True label: 0, Pred label: 0, %max entr = 2.6%

True label: 0, Pred label: 0, %max entr = 3.0%

True label: 1, Pred label: 1, %max entr = 3.6%

True label: 1, Pred label: 1, %max entr = 3.6%

True label: 1, Pred label = 0, %max entr = 100.0%

True label: 1, Pred label = 0, %max entr = 100.0%

Figure 20. (a) Correctly classified ECG signals with low uncertainty. (b) Misclassified ECG signals with high uncertainty.

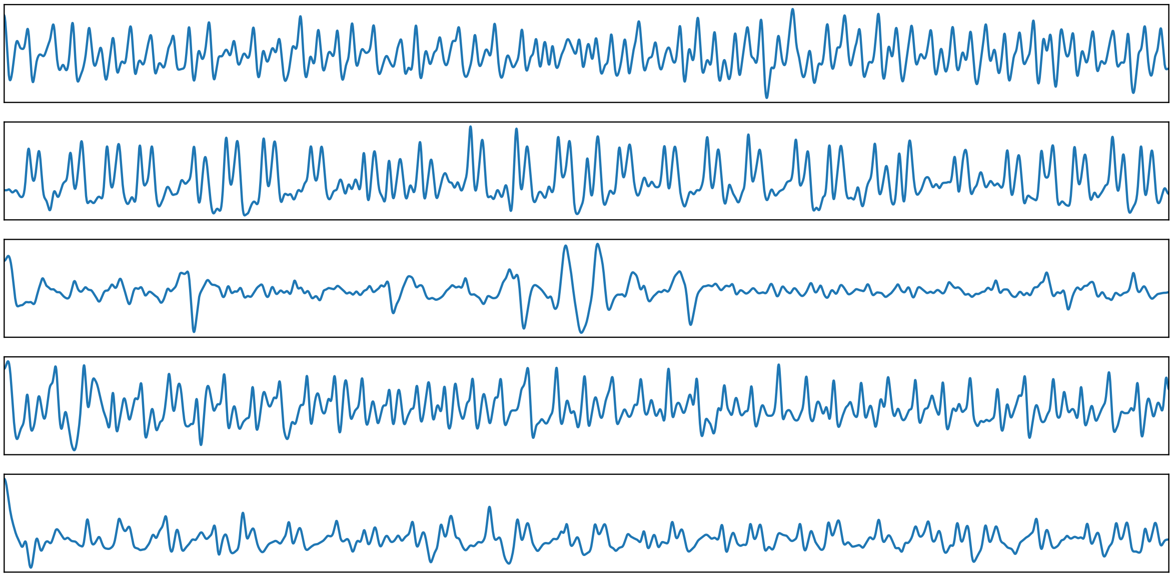


Figure 21. Sample out-of-distribution signals for PhysioNet ECG dataset.

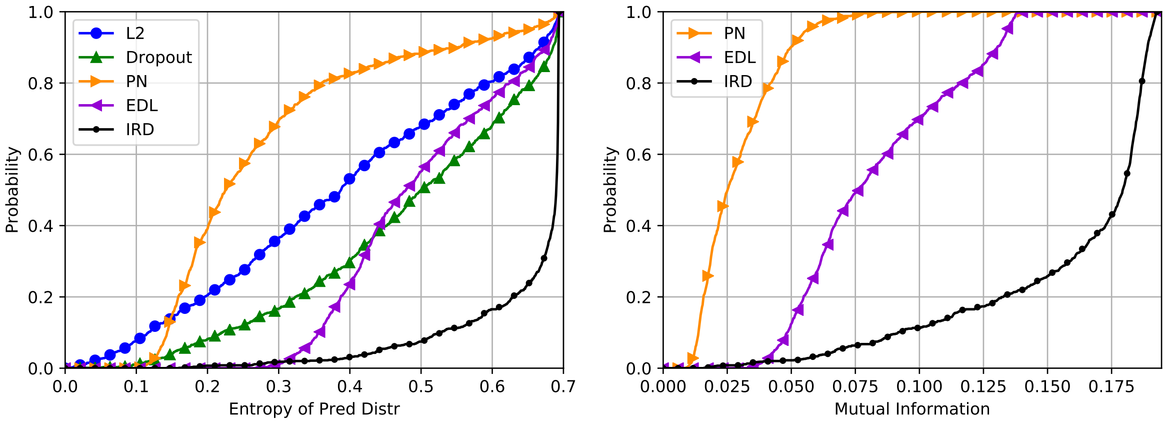


Figure 22. Empirical CDF of predictive entropy and mutual information on out-of-distribution signals for various deep learning methods.

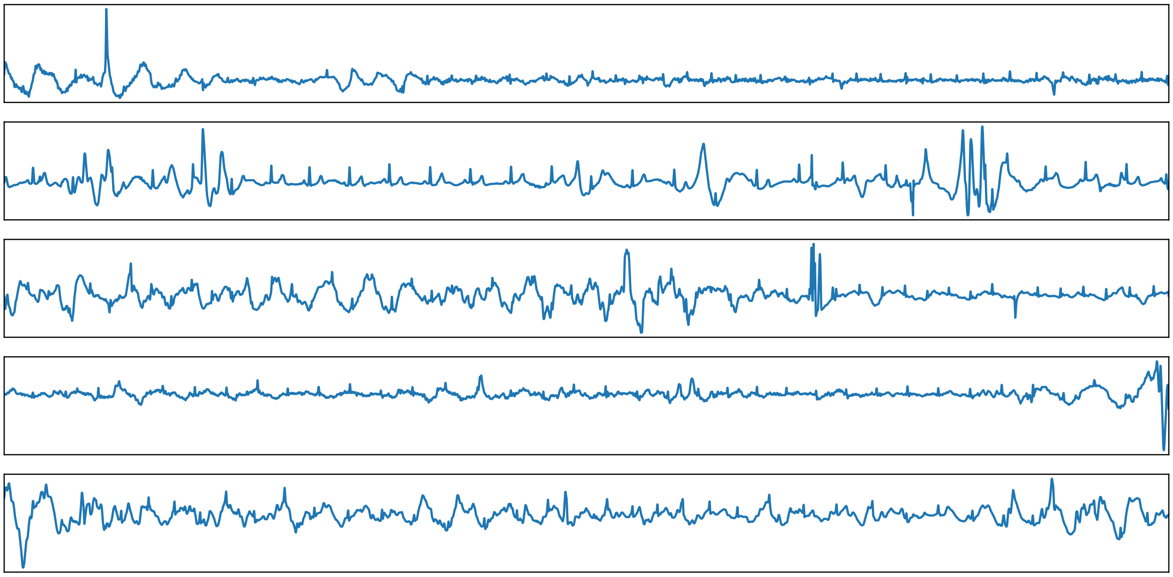


Figure 23. Sample noisy ECG signals from PhysioNet ECG dataset.

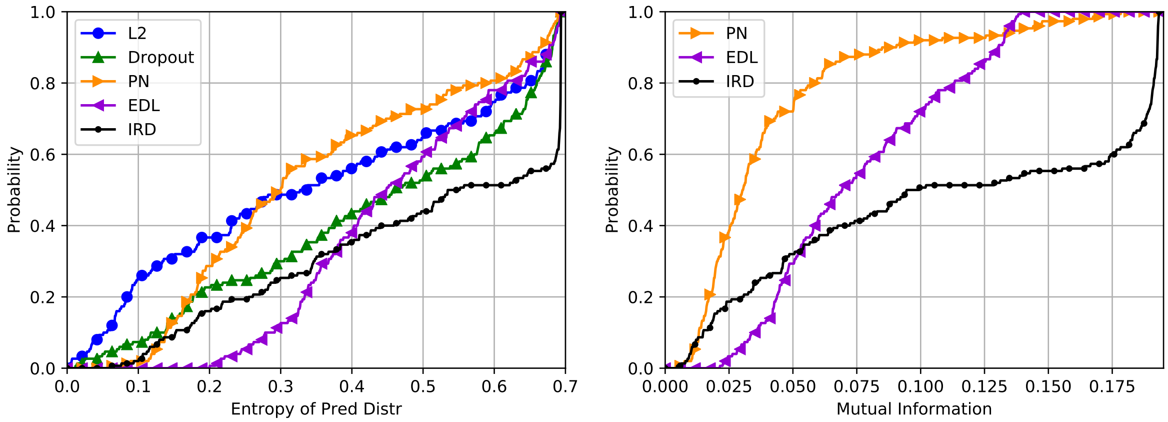


Figure 24. Empirical CDF of predictive entropy and mutual information on noisy ECG signals for various deep learning methods.

5. EXPERIMENTAL RESULTS ON TQIP 16-17 DATASET

Effective field triage of severely injured patients relies on accurate, rapid assessment and (experienced) clinical judgment. Due to the unavailability of more advanced diagnostic techniques in the field, triage decisions are made based on limited information, often by less experienced providers. We aim to design and test an AI-supported tool to accurately identify and triage high-risk military age patients in shock, needing transfusion of major operative intervention who sustained truncal gunshot wounds (GSW), based on the information that is available in the field.

The trauma quality improvement program (TQIP) database (2015-2017) was used to identify all military age (16-60) patients with truncal gunshot wounds (GSW). Information available in the field was identified: vital signs, age, sex, race, body mass index, visible wounds to neck, shoulder/axilla, thorax, abdomen, hip/thigh region (further specified using ICD-10 codes). An information-robust deep Dirichlet neural network (IRD), a form of Artificial Intelligence (AI), was designed to provide expected data uncertainty and total uncertainty estimation results along with its predictions. Contrary to conventional AI neural networks (Softmax), that are often overconfident in their predictions, our approach accounts for uncertainty in the training process and has the ability to measure uncertainty.

With a focus on penetrating gunshot wound injuries, our dataset is comprised of $24,428$ patients with a mean age of 29.5 (SD 10.1), median ISS of 16 (IQR 10-24, for 2015-16). To extract this data from the TQIP database, constraints were placed including age limit 16-60, identified bullet wounds on the trunk/junctional areas, and no missing vitals (except for temperature). Data available per patient include vital signs, age, weight, demographics, prior comorbidities (for patient context) and injury pattern encoding all injury locations visible from the outside. Outcomes of interest to predict using AI include shock, need for massive transfusion and need for major surgery.

Shock was defined as a combination of pulse $>$ 100, systolic blood pressure $<$ 100, pRBC transfusion of $>$ 5 units within 4 hours, requiring urgent hemorrhage control within 4 hours, or diagnosis of shock as per ICD9 /10. Major hemorrhage control surgery is defined as undergoing laparotomy, thoracotomy, sternotomy, pericardiotomy or open vascular repair procedure within 2 hours. Massive transfusion was defined as requiring $>$ 10 units of pRBC within 4 hours. Visible GSW to the thorax and abdomen were identified in 50.2% and 26.7% of patients respectively. Shock was identified in 11% of the included patients, 20% of patients underwent major operative intervention within 2 hours, and 5% received early massive transfusion.

A 90/10% train-test split was used in combination with oversampling to account for class imbalance. Results were averaged over twenty random train/test splits to account for training/testing distribution variability. For uncertainty, the metrics of interest are test prediction accuracy, max-probability (total uncertainty, abbreviated as Prob) and expected data uncertainty. Max-probability of 0.5 corresponds to low certainty, whereas a score of 1.0 corresponds to very high certainty.

5.1 NETWORK ARCHITECTURE

The 43 context variables and 73 injury pattern variables make up a total of 116 input variables per patient. A deep neural network consisting of an input embedding layer of dimension 60, three hidden layers with ReLU activations of size 100, 80, 30 respectively, and a binary classification layer, was used. The form of the classification layer depends on the method used. For L2 the conventional softmax layer is used while for IRD the softplus layers is used that outputs $\boldsymbol{\alpha}$ that parametrize the predictive Dirichlet distribution. The parameters used in training the IRD method were $p = 6, u = 1.5, \lambda = 1.5, \gamma = 0.1, \epsilon = 0.1$.

5.2 ACCURACY

The accuracy, true positive rate, true negative rate, false positive rate and false negative rates for each outcome prediction task are shown in Table 3. The first set of results correspond to conventional neural networks (L2) and the second set of results is based on the IRD method. The IRD method achieves competitive performance with conventional networks, identifying shock with test accuracy $81.3 \pm 1.7\%$, predicting major operative intervention with test accuracy $76.2 \pm 1.4\%$ and transfusion requirements with test accuracy $80.3 \pm 2.3\%$.

TABLE 3

TQIP 16-17 Dataset: Test accuracy, true positive rate, true negative rate, false positive rate and false negative rates (%) for each trauma care prediction task

Prediction Task	Accuracy	TP Rate	TN Rate	FP Rate	FN Rate
L2/Shock	82.7 ± 0.9	85.1	80.3	19.7	14.9
L2/Mass. Transf.	81.1 ± 1.8	82.8	79.4	20.6	17.2
L2/Maj. Surg.	77.7 ± 1.0	81.2	74.2	25.8	18.8
IRD/Shock	81.3 ± 1.7	84.4	78.1	21.9	15.6
IRD/Mass. Transf.	80.3 ± 2.3	82.7	77.9	22.1	17.3
IRD/Maj. Surg.	76.2 ± 1.4	87.3	65.0	35.0	12.7

5.3 UNCERTAINTY ESTIMATION

Max-probability of the predictive distribution $\max_c\{P(Y = c; \mathbf{x}^*, w)\} = \max\{P_0, P_1\}$ is another measure of total uncertainty in predictions. It is particularly interesting and interpretable for binary classification problems with values near 0.5 imply high uncertainty and values near unity imply low uncertainty.

For uncertainty values of shock predictions (Fig. 25), the median max-probability of misclassified/correct predictions is 0.51/0.92 for our method as opposed to 0.75/0.89 for conventional neural networks). For major operative intervention (Fig. 26), median max-probability for misclassified/correct examples is 0.52/0.81 for our method as opposed to 0.72/0.83). For early massive transfusion (Fig. 27), median max-probability for misclassified/correct examples is 0.52/0.92 for our method as opposed to 0.73/0.87).

The uncertainty metrics show that the IRD method on average yields very uncertain scores (prob ≈ 0.5) when errors are made and more certain score (prob ≈ 1.0) when correct predictions are made, in addition to measuring expected data uncertainty that shows the difficulty of a decision problem (e.g., predicting major surgery need is harder than predicting shock based on our results). The same trends are supported by the predictive entropy uncertainty measures.

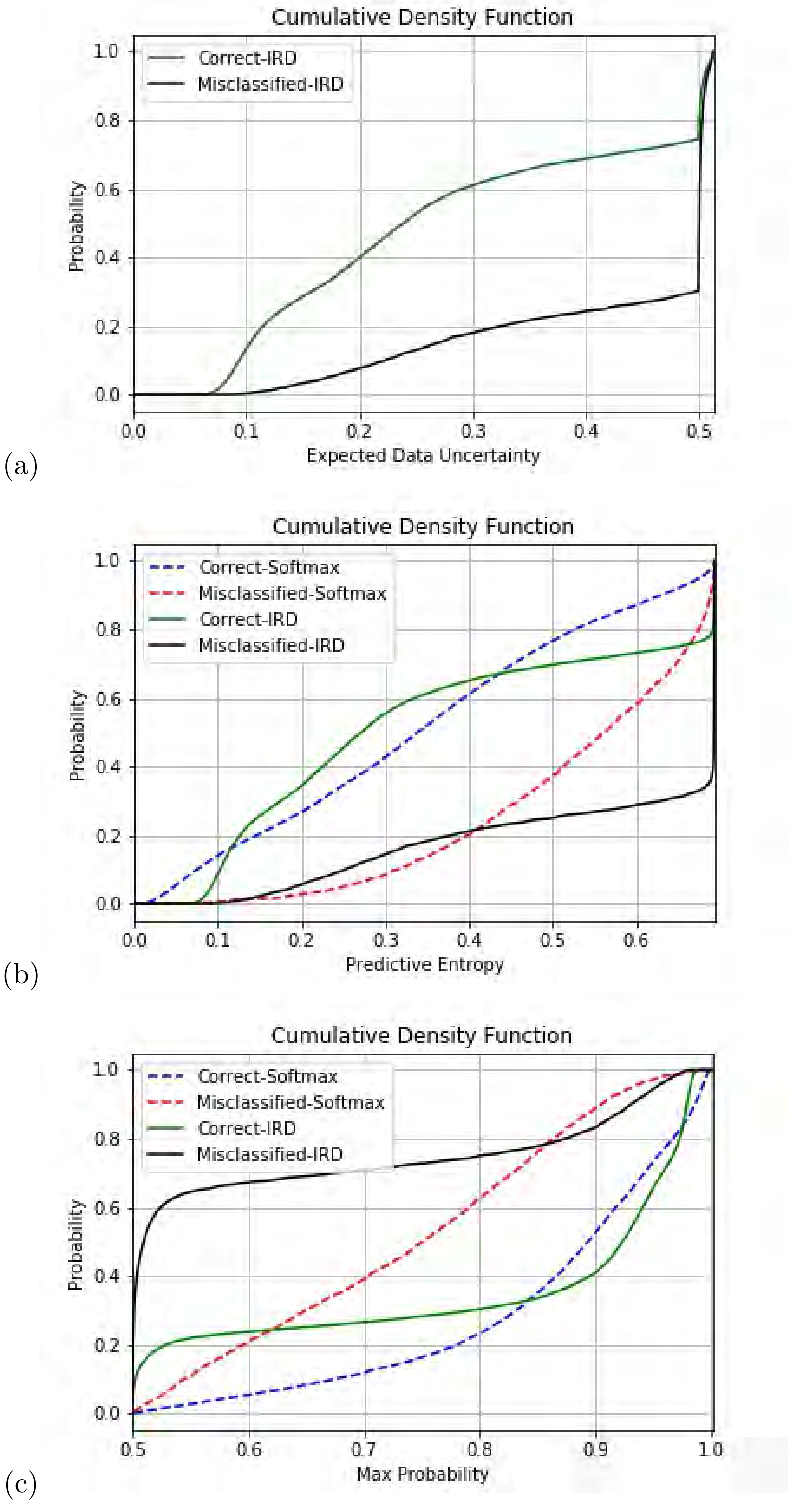


Figure 25. Shock identification task. Empirical CDF curves for (a) expected data uncertainty, (b) predictive entropy, and (c) max-probability.

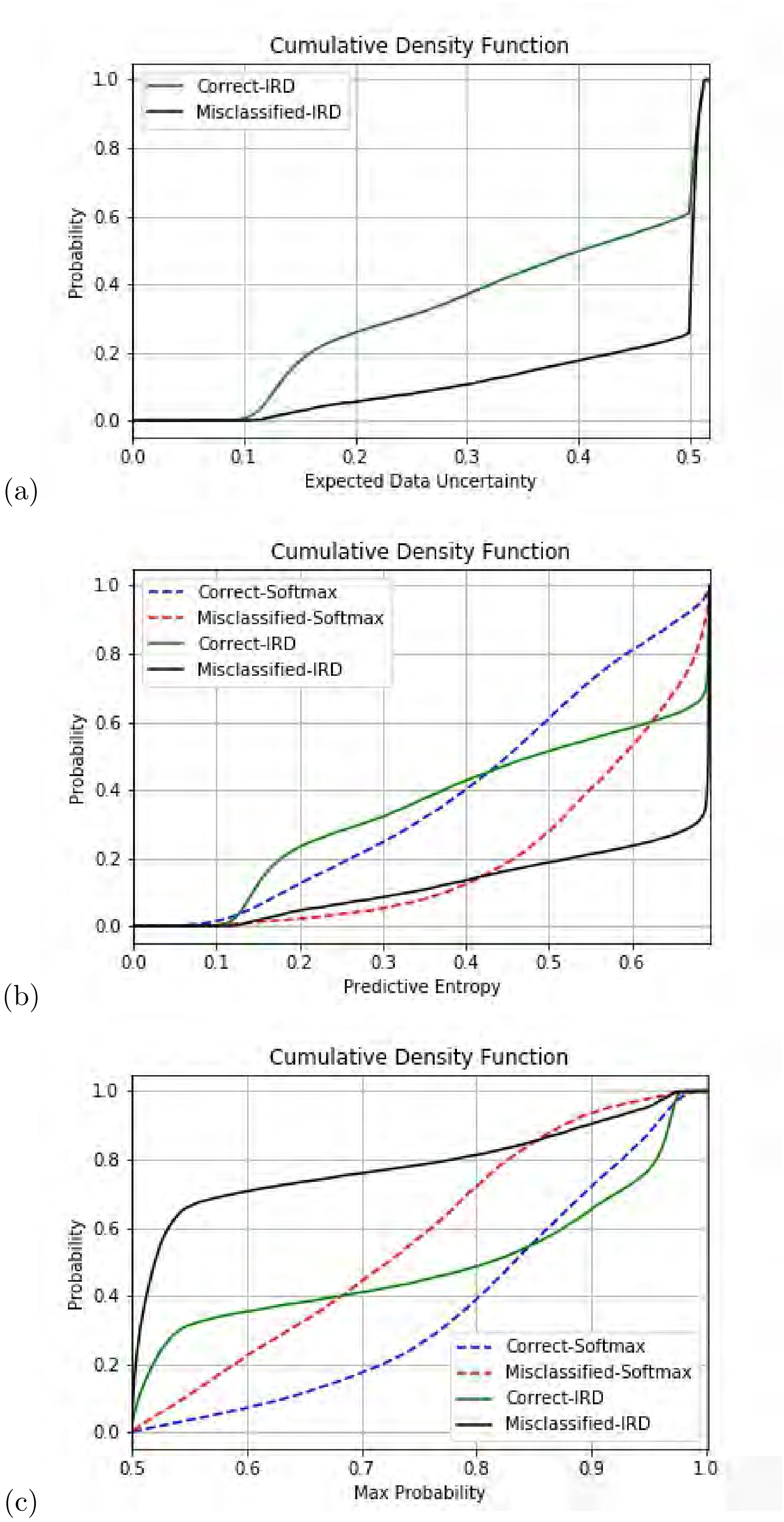


Figure 26. Major operative surgery prediction task. Empirical CDF curves for (a) expected data uncertainty, (b) predictive entropy, and (c) max-probability.

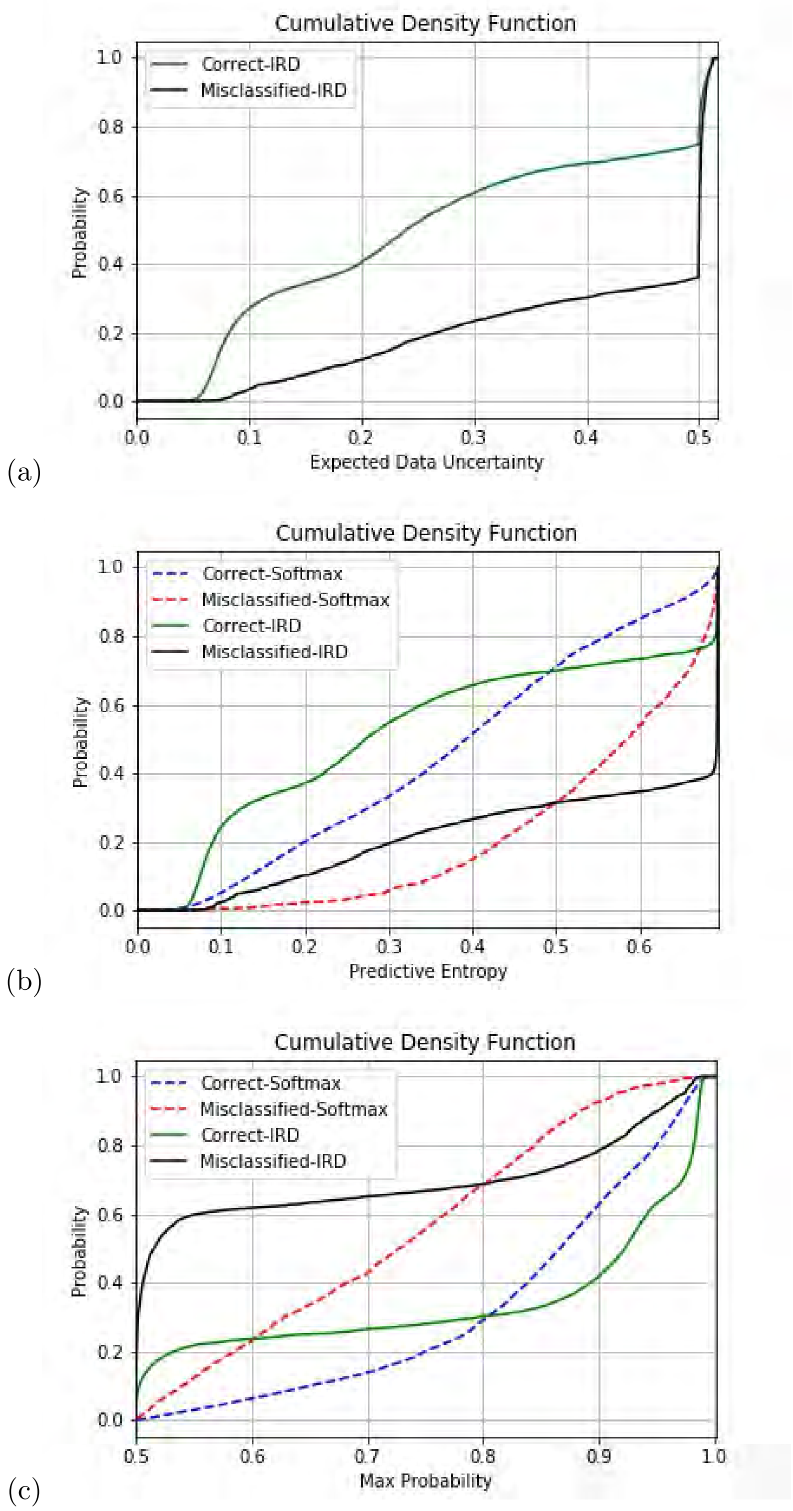


Figure 27. Massive transfusion prediction task. Empirical CDF curves for (a) expected data uncertainty, (b) predictive entropy, and (c) max-probability.

6. CONCLUSION AND FUTURE WORK

In this report we presented a new method for training Dirichlet neural networks that are aware of the uncertainty associated with predictions. Our training objective, which fits predictive distributions to data, consisted of three elements; a calibration loss that minimizes the expected L_p norm of the prediction error, an information divergence loss that penalizes information flow towards incorrect classes, and a maximum entropy loss that maximizes uncertainty for small adversarial perturbations. We derived closed-form expressions for our training loss and desirable properties on how improved uncertainty estimation is achieved. Experimental results highlighted the unmatched improvements in predictive uncertainty estimation made by our method over conventional softmax neural networks, Bayesian neural networks, and other recent Dirichlet networks trained with different criteria. Furthermore, due to the explicit modeling of the categorical distributions over classes, our approach does not require ensembling multiple predictions or performing multiple evaluations of the network at inference time (e.g., as BNNs do approximate integration over the parameter uncertainties to obtain approximate predictive distributions) to arrive at predictive distributions and compute uncertainty metrics.

The benefits of this novel AI uncertainty-aware method were demonstrated on an image classification task, a ECG-based heart condition diagnosis task, and a trauma care decision making task. In all these different domains, the results showed that information-robust neural networks can be tightly coupled to the training data and have the ability to yield accurate predictive uncertainty estimates that potentially allow the AI system to predict when errors are likely to be made (due to insufficient evidence for example) and detect anomalous data with high confidence, while maintaining a high prediction accuracy.

Future work will include extending the uncertainty modeling framework to temporal prediction tasks in order to accommodate more diverse data inputs (vitals time series, labs, etc), in addition to using this framework to learn which patterns of data (e.g., injury patterns) are easier to predict and associated with certain conditions, and which have higher predictive uncertainty in order to obtain data-driven insights into the trauma care decision problems under study. Further extensions of the methods will be explored including leveraging patients with missing vitals and using semi-supervised learning to improve accuracy and uncertainty estimation. Through our collaboration with Massachusetts General Hospital Trauma Division, we plan on applying our algorithms to real patient data both for trauma care decision support and automatic closed-loop monitoring and intervention prediction (e.g., ICU monitoring). Future work may also include using this framework for early sepsis detection and prediction of other nosocomial infections.

REFERENCES

B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–838, 2015.

A. Bendale and T. E. Boult. Towards Open Set Deep Networks. In *Computer Vision and Pattern Recognition*, 2016.

C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight Uncertainty in Neural Networks. In *International Conference on Machine Learning (ICML)*, 2015.

Y. Bulatov. notMNIST dataset, 2011. URL http://yaroslavvb.com/upload/notMNIST/.

T. Chen, E. Fox, and C. Guestrin. Stochastic Gradient Hamiltonian Monte Carlo. In *International Conference on Machine Learning*, 2014.

D. C. Ciresan, U. Meier, J. Masci, and J. Schmidhuber. Multi-column deep neural network for traffic sign classification. *Neural Networks*, 32:333–338, 2012.

G. D. Clifford, C. Liu, B. Moody, Li wei H. Lehman, I. Silva, Q. Li, A. E. Johnson, and R. G. Mark. Af Classification from a short single lead ECG recording: the PhysioNet/Computing in Cardiology Challenge 2017. In *Computing in Cardiology*, 2017. URL https://physionet.org/challenge/2017/.

C. Dente, M. Bradley, S. Schobel, B. Gaucher, T. Buchman, A. Kirk, and E. Elster. Towards precision medicine: Accurate predictive modeling of infectious complications in combat casualties. *Journal of Trauma and Acute Care Surgery*, 83(4):609–616, October 2017.

B. J. Eastridge et al. Death on the battlefield (2001-2011): implications for the future of combat casualty care. *Journal of Trauma and Acute Care Surgery*, 73(6):S431–S437, December 2012.

T. Van Erven and P. Harremos. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

Y. Gal and Z. Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *International Conference on Machine Learning (ICML)*, 2016.

R. Geirhos, C. R. M. Temme, J. Rauber, M. Bethge, and F. A. Wichmann. Generalization in humans and deep neural networks. In *Advances in Neural Information Processing Systems*, 2018.

I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and Harnessing Adversarial Examples. In *International Conference for Learning Representations*, 2014.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning*, 2017.

D. Haussler and M. Opper. Mutual Information, Metric Entropy and Cumulative Relative Entropy Risk. *The Annals of Statistics*, 25(6):2451–2492, 1997.

K. He, X. Zhang, S. Ren, and J. Sun. Delving Deep into Rectifiers: Surpassing Human-level Performance on ImageNet classification. In *IEEE International Conference on Computer Vision (ICCV)*, December 2015.

J. M. Hernandez-Lobato and R. P. Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International Conference on Machine Learning*, 2015.

G. Hinton, L. Deng, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

N. Houlsby, F. Huszar, Z. Ghahramani, and M. Lengyel. Bayesian Active Learning for Classification and Preference Learning. Technical report, 2011. arXiv:1112.5745.

D. P. Kingma, T. Salimans, and M. Welling. Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing (NIPS)*, 2015.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

A. Kurakin, I. J. Goodfellow, and S. Bengio. Adversarial Machine Learning at Scale. In *International Conference for Learning Representations*, 2017.

B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems*, 2017.

Y. LeCun, C. Cortes, and C.J.C. Burges. The MNIST Database. URL http://yann.lecun.com/exdb/mnist/index.html.

Y. LeCun, Y. Bengio, and G. Hinton. Deep Learning. *Nature*, 521(7533):436–444, 2015.

K. Lee, H. Lee, K. Lee, and J. Shin. Training Confidence-Calibrated Classifiers for Detecting Oout-of-Distribution Samples. In *International Conference for Learning Representations*, 2018.

Y. Li and Y. Gal. Dropout inference in Bayesian neural networks with alpha-divergences. In *International Conference on Machine Learning*, 2017.

C. Louizos and M. Welling. Multiplicative Normalizing Flows for Variational Bayesian Neural Networks. In *International Conference on Machine Learning (ICML)*, 2017.

David J.C. MacKay. A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference for Learning Representations*, 2018.

A. Malinin and M. Gales. Reverse KL-Divergence Training of Prior Networks: Improved Uncertainty and Adversarial Robustness. Technical report, 2019. arXiv:1905.13472.

J. G. Mauldon. A generalization of the Beta-distributions. *Annals of Mathematical Statistics*, 30: 502–520, 1959.

D. Molchanov, A. Ashukha, and D. Vetrov. Variational dropout sparsifies deep neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

J. E. Mosimann. On the compound multinomial distribution, the multivariate beta-distribution, and correlations among proportions. *Biometrika*, 49:65–82, 1962.

Alfred Rényi. On measures of entropy and information. In *Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 547–561, 1961.

H. Ritter, A. Botev, and D. Barber. A Scalable Laplace Approximation for Neural Networks. In *International Conference on Learning Representations*, 2018.

M. Sensoy, L. Kaplan, and M. Kandemir. Evidential Deep Learning to Quantify Classification Uncertainty. In *Advances in Neural Information Processing Systems (NIPS) 31*, 2018.

S. Sun, C. Chen, and L. Carin. Learning Structured Weight Uncertainty in Bayesian Neural Networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, 2014.

D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck. Deep Learning for Identifying Metastatic Breast Cancer. Technical report, June 2016. arXiv:1606.05718.

Y. Wu et al. Google‘s neural machine translation system: Bridging the gap between human and machine translation. Technical report, 2016. arXiv:1609.08144.

W. Xiong, J. Droppo, X. Huang, F. Seide, M. L. Seltzer, A. Stolcke, D. Yu, and G. Zweig. Toward Human Parity in Conversational Speech Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 25(12):2410–2423, December 2017.