# Uncertainty-Aware Ensemble Combination Method for Quality Monitoring Fault Diagnosis in Safety-Related Products

Jefkine Kafunah, Muhammad Intizar Ali, and John G. Breslin, *Senior Member, IEEE*

***Abstract*— With the advent of Industry 4.0 (I4.0) leading to the proliferation of industrial process data, deep learning (DL) techniques have become instrumental in developing intelligent fault diagnosis (FD) applications. However, despite their potentially superior process monitoring capabilities, DL-based FD models are poorly calibrated and generate point estimate predictions without the associated uncertainty estimates. For DL-based FD models, accurate predictive uncertainty estimates from well-calibrated models are essential in ensuring industrial process safety and reliability. This paper proposes Ensemble-to-Distribution (E2D), an uncertainty-aware combination method for quality monitoring FD based on an ensemble of deep neural networks (DNNs). First, E2D addresses safety by providing accurate uncertainty estimates on model predictions, enabling informed decision-making to minimize operational risks. Second, E2D improves model performance on out-of-distribution (OOD) detection tasks to facilitate deployments in the real world. Third, E2D is a post hoc application, implementable at inference time, and compatible with diverse pre-trained models. Finally, to demonstrate the effectiveness of E2D, we explore the problem of monitoring the stability of industrial processes and product quality using case studies on the steel plates faults and APS failure at Scania trucks datasets.**

***Index Terms*— Ensemble methods, deep learning, uncertainty estimation, calibration, fault diagnosis, process monitoring, safety-critical.**

## I. Introduction

INDUSTRY 4.0 (I4.0) has enabled dynamic modern-day industrial environments through rapid automation and improved access to real-time data from complex industrial operations [1]–[5]. Additionally, the systematic integration of the physical and virtual worlds through the cyber-physical system (CPS), a core concept of I4.0, enables the construction of expansive factories with high flexibility, adaptability, and even self-awareness [6]–[8]. These factories are physically interconnected large-scale industrial plants requiring a higher level of process and quality management strategies to improve overall production safety and efficiency. The rapidly increasing high-dimensional and nonlinear historical process data from large-scale industrial plants pose significant challenges to traditional process monitoring approaches. As a result, deep learning (DL) techniques have become the dominant approach to building intelligent fault diagnosis (FD) applications from large-scale industrial process data [9]–[11].

Nonetheless, for DL-based FD models deployed in the real world, accurate predictive uncertainty estimation and out-of-distribution (OOD) detection from well-calibrated models are essential in ensuring overall system safety and reliability. Accordingly, the ensemble models are well suited for developing robust FD applications due to the ability to obtain improved predictive performance [12]–[18], protect against adversarial attacks [19], [20], and decompose the total predictive uncertainty into epistemic and aleatoric uncertainty [21]–[25].

However, existing ensemble combination methods apply deterministic point estimation techniques that are ineffective in capturing the underlying ensemble member diversity and robustly represent the combined model predictive uncertainty [21], [26]–[28]. This paper proposes Ensemble-to-Distribution (E2D), a classifier combination method for an ensemble of DL-based FD models that applies parameter estimation techniques to fit a probability distribution over the combined model output. In particular, E2D is a distribution-based approach that enhances the ensemble model output capacity by generating a continuous multivariate probability distribution as the combined model output.

Following a growing demand for higher-quality products, reduced rejection rates, and compliance with safety and environmental regulations, modern-day industrial plants require effective monitoring and control strategies through intelligent fault management. FD systems improve production processes by identifying defect patterns that can lead to rejected products downstream, unintended downtimes, and insufferable economic losses. Nonetheless, for DL-based FD applications, exposure to highly dynamic industrial environments introduces uncertainties to the FD system. In particular, the sources of uncertainty for DL-based FD applications include: (i) model fit: uncertainty from the inherent limitations of a learned model as characterized by the degree of errors in outcomes, (ii) data quality: uncertainty from training data-related issues such as; noise

This publication has emanated from research supported in part by a grant from Science Foundation Ireland under Grant Number SFI/16/RC/3918 (Confirm), and also by a grant from SFI under Grant Number SFI 12/RC/2289_P2 (Insight). For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Jefkine Kafunah is with the Data Science Institute, University of Galway, H91 TK33 Galway, Ireland (e-mail: jefkine.kafunah@insight-centre.org).

Muhammad Intizar Ali is with the School of Electronic Engineering, Dublin City University, 9 Dublin, Ireland (e-mail: ali.intizar@dcu.ie).

John G. Breslin is with the Data Science Institute, University of Galway, H91 TK33 Galway, Ireland (e-mail: john.breslin@insight-centre.org).

in the data, varying quality of data features owing to sensor limitations, and imbalanced datasets lacking adequate samples from the uncommon defects [29], and; (iii) scope compliance: uncertainty resulting from the mismatch between the model scope of training and application scope (e.g., plant dynamics varying over time, and the model has to extrapolate beyond its specialization) [30]. Therefore, DL-based FD applications in industrial systems primarily depend on insights from model predictive uncertainty estimates to identify and avoid harmful unintended behavior during system operation, especially for safety-critical application domains.

In this paper, we seek to develop a product quality monitoring FD application based on an ensemble of DNN models constructed entirely from process data to help monitor the stability of industrial processes and product quality. We focus on addressing the problems of uncertainty estimation and OOD detection for the DL-based FD applications. In particular, our approach involves designing an uncertainty-aware ensemble combination method to generate FD model predictions accompanied by accurate predictive uncertainty estimates. Based on our approach, our main contributions can be summarized as follows:

- We propose E2D, a method that generates a continuous multivariate probability distribution as the combined DL-based FD ensemble model output, replacing deterministic point estimation techniques that are ineffective in capturing the underlying ensemble model uncertainties.
- We apply differential entropy as a measure of uncertainty for the output probability distribution, obtaining improved scores over standard entropy measures for in-distribution (ID) and OOD sample detection tasks.
- We propose a standard algorithm for E2D, as a post hoc application implementable at inference time and compatible with diverse pre-trained models.

The remainder of this paper is structured as follows. In Section II, we present a review of the literature and related work on methods used for ensemble classifier combinations. In Section III, we present the proposed approach. First, we describe an associated concept of Multinomial and Dirichlet conjugacy upon which we build our proposed method. We then detail the data transformations and numerical methods applied to the distribution estimation problem, including pseudocode for the E2D algorithm. In Section IV, we present the case study, outlining the datasets, experiments, and evaluation metrics upon which we analyze the effectiveness of our proposed method. In section V, we present the results obtained from the experiments and a detailed analysis of our findings. Finally, Section VI concludes the paper with an overview of our main contributions and potential future work.

## II. Related Works

In this section, we present the related works on ensemble combination methods and DL-based fault diagnosis.

### A. Ensemble Combination Methods

Classifier selection (CS) and classifier fusion (CF) are the two fundamental approaches used for the combiner module of an ensemble [31]. In this work, we focus on CF approaches where the aim is to aggregate predicted posterior probabilities from the multiple base classifiers through some efficient combiner module.

Simple averaging [32] is one of the fundamental combination approaches applied to ensembles by averaging the base classifier posterior probabilities to obtain a final mean estimate. Nonetheless, error reduction in simple averaging method holds the assumption that the errors from the individual classifiers are uncorrelated, even though for ensembles, training of the base classifiers on the same problem suggests that the errors are typically highly correlated [33, p. 69]. Another widely used approach, the weighted averaging method [34], extends the idea of simple averaging to incorporate weights that cater to the varied performance or implied importance of the individual base classifiers. Averaging based approaches implement point estimation techniques that do not preserve the diversity of the ensemble, while approaches such as weighted averaging are prone to overfitting [33, p. 72]. Kittler et al. [35] propose a collection of algebraic methods derived from the probabilistic framework upon which the maximum, minimum, and median combination rules apply to individual classifiers predicted outputs as the combined output. Voting [33, pp. 71–77] is a common combination strategy explored in ensembles where the base classifiers predict either crisp labels or class probabilities. Each output prediction from a base classifier is regarded as a vote, indicating some preferred choice in the final ensemble decision-making process. However, voting approaches apply a winner-take-all strategy, thus hindering cooperation amongst ensemble member classifiers. Kuncheva et al. [36] propose decision templates combination method based on a decision profile compiled as a matrix of all the predicted outputs from ensemble classifiers. Wolpert [37] proposes stacking, a combination method where predicted outputs from separately trained base classifiers are aggregated and used as input to another classifier known as a meta-classifier. However, the meta-classifier is usually a deterministic classifier that generates point estimates of individual class probabilities.

### B. DL-Based Fault Diagnosis

Fault diagnosis methods broadly categorize into model-based, signal-based, knowledge-based, and hybrid methods [38], [39]. Recently, FD applications have adopted the knowledge-based approach relying on historical process data to extract the underlying relationship between faults and process variables. In particular, the quantitative knowledge-based methods essentially formulate the diagnostic problem as a pattern recognition problem, applying techniques such as DL to handle the complexities of high dimensional nonlinear historical process data otherwise hard to establish through explicit system models or expert systems based on human reasoning [11], [39].

Jiang et al. [7] observe that one of the emerging challenges of industrial CPS (ICPS) monitoring and safety control is the development of artificial intelligence-based autonomous decision units able to achieve plant-wide monitoring and control through transparent process management. From the perspective of a plant-wide process monitoring system implemented as

a distributed framework, the sequential and temporal dependencies among sub-processes bring about interdependencies to the individual sub-process level FD models. Notably, a malfunctioning sub-process FD model can lead to cascading effects that cause performance degradation of the overall industrial process [6], [40]. Therefore, this research raises the following open questions: (i) how to leverage insights originating from uncertainty estimation to facilitate sub-process level autonomy and self-awareness, and; (ii) how to integrate and propagate sub-process level uncertainty estimates through the distributed modeling framework up to the final plant-wide monitoring result.

Kamal et al. [9] propose a fault detection and classification (FDC) system for a nuclear power plant based on wavelet transform and NNs, resulting in model with enhanced speed of fault recognition, accuracy and robustness. Wen and Gao in [10] propose a deep transfer learning (DTL) method for fault diagnosis using a three-layer sparse auto-encoder (SAE) for feature extraction and a maximum mean discrepancy (MMD) term to minimize penalty between source and target features. Zhao et al. [41] develop a new DL method, deep residual shrinkage networks (DRSNs) for FD tasks with highly noised vibration signals. Jia et al. [42] present a DNN-based intelligent method for diagnosing the faults of rotating machinery. The proposed DNN models trained on massive datasets are less dependent on human labor or prior knowledge about signal processing techniques and diagnostic expertise. We note that the NN-based FD implementations mentioned above are deep networks with softmax layers as the network output, resulting in overconfident model predictions for both ID and OOD samples. Lu et al. [43] propose a DNN-based model for fault diagnosis referred to as DAFD to address cross-domain learning problems in FD. DAFD models trained in a particular source domain are adoptable in a different but related target domain. Our method seeks to address, among others, the problem of OOD detection where samples emerge from unrelated target domains.

Wang et al. [44] propose a data-driven approach based on a deep belief network (DBN) optimized through a particle swarm optimization (PSO) algorithm to predict material removal rate (MRR) during wafer polishing. Zhang et al. [45] enhance the MRR prediction model through another data-driven approach based on random forests and residual CNN (ResCNN). For the industrial hydrocracking process, Yuan et al. [46] propose a dynamic CNN to learn hierarchical local nonlinear dynamic features for soft sensor modeling. Further, to enhance soft sensor modeling capabilities to time series process data, Yuan et al. [47] propose a spatiotemporal attention-based LSTM network. Loy-Benitez et al. [48] present a memory-gated recurrent neural networks-based autoencoders (MG-RNN-AE) to perform FD on measurements of the multivariate indoor air quality data in subway stations. Le et al. [49] propose FDC-CNN, a convolutional neural network (CNN) model for FDC. In the proposed FDC-CNN model, specially designed receptive fields in the convolutional layer operate as fault feature extractors, capturing the structural characteristics of the multivariate fault data. We observe that FD models depending on robust feature extraction can be application-specific, requiring explicit knowledge of relations between process variables. Furthermore, the proposed DNN-based FD models are deterministic, generating point estimates with no representation of uncertainty. Based on our approach, we seek to enhance the deterministic models by leveraging an ensemble of models coupled with an uncertainty-aware combination method that generates a distribution over distributions as the combined model output.

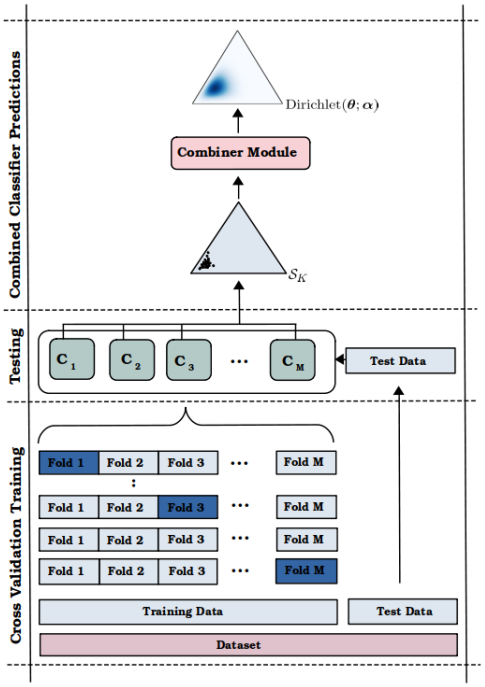


Fig. 1: Schematic description of our approach Ensemble-to-Distribution. From a diverse set of $M$ base classifiers, the collection of $K$-dimensional transformed logit output forms a dataset upon which the combiner module fits a probability distribution $\mathrm{Dirichlet}(\boldsymbol{\theta}; \boldsymbol{\alpha})$, establishing the final output of the ensemble.

## III. Ensemble to Distribution

In this section, we present our proposed method, E2D. First, we begin by outlining a related concept of conjugacy upon which we build our proposed method. We then outline our method in the proposed approach section, detailing the data transformation techniques, numerical methods applied to the distribution estimation problem, and parameter initialization routines. Finally, we present the pseudo-code for the algorithm and a brief discussion of the implementation.

### A. Multinomial and Dirichlet Conjugacy

The multinomial distribution of $N$-independent trials and $K$ categories has the probability mass function given by:

$$\mathrm{Mult}(N; \boldsymbol{\theta}) = \frac{N!}{\prod_{k=1}^{K} x_k!} \prod_{k=1}^{K} \theta_k^{x_k} \tag{1}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ and $\sum_{k=1}^{K} \theta_k = 1$, represents the distribution parameter vector, a set probabilities that the $K$

categories occur, while $\boldsymbol{x} = (x_1, \ldots, x_K)$ is a nonnegative integer count vector of observations where $\sum_{k=1}^{K} x_k = N$.

The Dirichlet distribution is the conjugate prior for the multinomial [50]. Therefore, for $\boldsymbol{x}$ that follows a multinomial distribution with a Dirichlet prior, the joint distribution given by:

$$\operatorname{Mult}(N;\boldsymbol{\theta}) \cdot \operatorname{Dir}(\boldsymbol{\theta};\boldsymbol{\alpha}) = \frac{\Gamma\left(N + \sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma\left(x_k + \alpha_k\right)} \prod_{k=1}^{K} \theta_k^{x_k + \alpha_k - 1} \tag{2}$$

The posterior distribution of the Dirichlet-multinomial 2 is itself a Dirichlet distribution with the parameters $\boldsymbol{\theta} \sim \operatorname{Dir}(\alpha_1 + x_1, \ldots, \alpha_K + x_K)$ where, the Dirichlet parameters $\boldsymbol{\alpha}$ act as pseudo-counts that initially allocate some weight on each of the $K$ categories before the actual data emerges to reveal the true underlying distribution.

### B. Proposed Approach

For an ensemble of $M$ members, let $\tilde{\boldsymbol{\theta}}_i \in \mathbb{R}^K$ be the logit vector of the $i^{th}$ member output, such that $\tilde{\boldsymbol{\Theta}} = \{\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_M\}$ represents the collection of all member outputs. In this work, we seek to estimate a probability distribution over all possible ensemble member outputs through maximum likelihood estimate (MLE). Therefore, we apply data transformation to the aggregated ensemble logit space $\tilde{\boldsymbol{\Theta}}$ to obtain datasets suitable for the Dirichlet distribution and its compound variant, the Dirichlet-multinomial [51], [52]:

*1) Dirichlet dataset:* For the Dirichlet distribution, we transform $\tilde{\boldsymbol{\Theta}}$ into a set of multinomial distribution parameter vectors which we denote $\boldsymbol{\Theta}_{\mathrm{Dir}}$. Initially, we convert all the logits into positive by applying the exponential function to the logits. We then apply a combination of L1 normalization and label smoothing [53] to obtain zero-avoiding probability vectors as confidence scores for each of the ensemble member output, each representing an element on the probability simplex $\mathcal{S}_K$ (see Fig. 1).

The Dirichlet distribution, a multivariate probability distribution over the probability simplex is capable of modeling the categorical data from the ensemble member outputs. Appropriately, a $K$-dimensional Dirichlet distribution can be thought of as a distribution over a $(K-1)$ simplex $\mathcal{S}_K$, representing the space of all $K$-dimensional categorical data $\boldsymbol{\Theta}_{\mathrm{Dir}}$. The Dirichlet distribution parameterized by a vector $\boldsymbol{\alpha}$ has a probability density function given by:

$$\begin{aligned} \operatorname{Dir}(\boldsymbol{\theta};\boldsymbol{\alpha}) &= \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma\left(\alpha_k\right)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \\ \text{where } & \sum_{k=1}^{K} \theta_k = 1, \quad \theta_k \geq 0 \end{aligned} \tag{3}$$

Therefore, given the observed set of categorical data $\boldsymbol{\Theta}_{\mathrm{Dir}}$, we seek to obtain the parameters of a Dirichlet distribution which maximize the likelihood of that data. Estimation of the $\boldsymbol{\alpha}$ parameters for this Dirichlet distribution is obtained by maximizing the following log-likelihood function:

$$\log p(\boldsymbol{\Theta}_{\mathrm{Dir}}|\boldsymbol{\alpha}) \propto M\left(\log\Gamma\left(\sum_k \alpha_k\right) - \sum_k \log\Gamma\left(\alpha_k\right) + \sum_k \left(\alpha_k - 1\right)\log\bar{\theta}_k\right) \tag{4}$$

where $\log\bar{\theta}_k = \frac{1}{M}\sum_i \log\theta_{ik}$.

Since the Dirichlet distribution belongs to a larger class of distributions, the exponential family, the objective $\log p(\boldsymbol{\Theta}_{\mathrm{Dir}}|\boldsymbol{\alpha})$ is concave in $\boldsymbol{\alpha}$ and thus converges to the global optimum [54], [55]. To obtain the likelihood estimates, we apply the Newton-Raphson algorithm. The Newton's method is an efficient numerical technique for estimating the Dirichlet $\boldsymbol{\alpha}$ parameters given the log likelihood. In (4), the gradient of the log-likelihood with respect to $\alpha_k$ is obtained as follows:

$$g_k = \frac{\partial \log p(\boldsymbol{\Theta}_{\mathrm{Dir}}|\boldsymbol{\alpha})}{\partial \alpha_k} = M\left(\Psi\left(\textstyle\sum_k \alpha_k\right) - \Psi\left(\alpha_k\right) + \log\bar{\theta}_k\right) \tag{5}$$

where $\Psi(\cdot)$ is the diagamma function. The update step is as follows:

$$\alpha^{\text{new}} = \alpha^{\text{old}} - \mathbf{H}^{-1}\mathbf{g} \tag{6}$$

$\mathbf{H}$ is the Hessian matrix derived from the second-derivatives of the log-likelihood, and in matrix form is represented as:

$$\begin{aligned} \mathbf{H} &= \mathbf{Q} + \mathbf{1}\mathbf{1}^{\top} z \\ q_{jk} &= -M\Psi'\left(\alpha_k\right)\delta(j-k) \\ z &= M\Psi'\left(\sum_k \alpha_k\right) \end{aligned} \tag{7}$$

$\Psi'(\cdot)$ is the triagamma function, while $\delta(\cdot)$ is the Dirac function. Notably, the Newton's method in this context is suitable for high dimension data as the Hessian matrix can be computed in linear time and does not require storing or explicitly inverting the matrix [56].

*2) Dirichlet-multinomial dataset:* For second and alternative case, the Dirichlet-multinomial distribution, we transform $\tilde{\boldsymbol{\Theta}}$ into a set of the multinomial count vectors which we denote $\boldsymbol{\Theta}_{\mathrm{DirMult}}$. Initially, we convert all the logits into positive by applying the exponential function to the logits. We then apply a combination of L1 normalization and label smoothing [53] to obtain zero-avoiding confidence scores and convert them to percentages, generating the categorical nonnegative integer vector counts for each of the ensemble member output.

Given the observed set of categorical data $\boldsymbol{\Theta}_{\mathrm{DirMult}}$ we seek to obtain the parameters of a Dirichlet-multinomial distribution which maximize the likelihood of that data. Estimation of the $\boldsymbol{\alpha}$ parameters for this compound distribution is obtained by maximizing the following log-likelihood function [57, p. 213]:

$$\begin{aligned} \log p(\boldsymbol{\Theta}_{\mathrm{DirMult}}|\boldsymbol{\alpha}) \propto\ & M\left\{\log\Gamma\left(\sum_k \alpha_k\right) - \sum_{k=1}^{K}\log\Gamma\left(\alpha_k\right)\right\} \\ & + \sum_{m=1}^{M}\sum_{k=1}^{K}\log\Gamma\left(x_{mk} + \alpha_k\right) \\ & - \sum_{m=1}^{M}\log\Gamma\left(N_m + \sum_k \alpha_k\right) \end{aligned} \tag{8}$$

In (8), given $\boldsymbol{\Theta}_{\mathrm{DirMult}}$, the Newton-Raphson method accurately computes MLE for the $\boldsymbol{\alpha}$ parameters [51], [52].

For efficient convergence of the proposed algorithm, proper initialization is deemed vital. Improper initialization causes the Newton-Raphson algorithm to experience slow convergence while the final parameter estimates can be outside the permissible range. In our implementation, we observe that initializing with a vector of $K$-ones obtains the best results. Further, in order to avoid overfitting, we perform additive smoothing [58] on both $\boldsymbol{\Theta}_{\mathrm{Dir}}$ and $\boldsymbol{\Theta}_{\mathrm{DirMult}}$ before the numerical optimization step.

### C. E2D Learning Algorithm

We outline our proposed algorithmic approach for the E2D in Algorithm 1. In our implementation, we seek to train a diverse set of classifiers that will be combined using the E2D technique. Operations in step one apply the cross-validation (CV) sampling technique where the choice of the number of folds is equal to the total number of ensemble base classifiers, therefore obtaining a unique classifier per fold. Operations in step two, begin by generating the datasets $\boldsymbol{\Theta}_{\mathrm{Dir}}$ or $\boldsymbol{\Theta}_{\mathrm{DirMult}}$ from the combined set of ensemble base classifiers and subsequently fit a Dirichlet distribution. The final ensemble output is either the distribution mean, distribution mode, or the mean of random samples drawn from the generated Dirichlet distribution. A schematic representation of the E2D Algorithm is presented if Fig. 1.

## IV. Case Study

In this paper, we analyze the effectiveness of our proposed method on two real-world industrial datasets; APS Failure at Scania Trucks dataset [59] and the Steel Plates Faults dataset [60].

### A. Steel Plates Faults dataset

In the steel industry, intelligent fault diagnosis during steel plate production is essential for the timely identification of defects that directly influence the final product safety and performance. Notably, fault diagnosis in steel plate production is challenging due to the complex nature of defects owing to the dynamic production process and the quality of raw materials [61]–[63]. The steel-plate surface defect inspection system involves capturing video images of the steel plates on the rolling equipment, followed by image processing and analysis, detecting the area of the defect, extracting features from the defect area, and finally, defect classification [64].

The steel plates faults dataset consists of a total of 1941 instances meant for the classification of surface defects in stainless steel plates during industrial production. This is a labelled dataset where instances are classified into either of the seven distinct typologies of faults: Pastry, Z Scratch, K Scatch, Stains, Dirtiness, Bumps, and Other Faults. Each recorded instance consists of 27 attributes representing the geometric shape of the fault and its contour. For this dataset, we apply FD to diagnose the source of the fault from among the seven commonly occurring faults of the steel plates. The target class distribution reveals an imbalanced dataset. For OOD detection evaluation, we generate a set of OOD data based on the Steel Plates Faults ID dataset using the Gaussian Hyperspheric Offset method [65] with a mean of 2 and standard deviation 0.3.

**Algorithm 1:** E2D Learning Algorithm

**Input:**
$M$, number of weak base classifiers to be combined.
$S$, draw sample size for the estimated distribution.
**Data:**
$\mathcal{D}^{\mathrm{train}} = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{N^{\mathrm{train}}}$, set of $N^{\mathrm{train}}$ i.i.d. labeled samples from the training dataset.
$\mathcal{D}^{\mathrm{test}} = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{N^{\mathrm{test}}}$, set of $N^{\mathrm{test}}$ i.i.d. labeled samples from the test dataset.
**Training**
  Step 1: Train $\boldsymbol{C} = \{C_m\}_{m=1}^{M}$ base learners using $\boldsymbol{\mathcal{F}} = \{\mathcal{F}_m\}_{m=1}^{M}$ classifiers on $\mathcal{D}^{\mathrm{train}}$
  Split $\mathcal{D}^{\mathrm{train}}$ into $M$ groups $\left\{\left(\mathcal{D}_1^{\mathrm{train}}, \mathcal{D}_1^{\mathrm{valid}}\right), \ldots, \left(\mathcal{D}_M^{\mathrm{train}}, \mathcal{D}_M^{\mathrm{valid}}\right)\right\}$
  **for** $m = 1$ **to** $M$ **do**
    **for** $i = 1$ **to** $N^{\mathrm{train}}$ **do**
      $C_m =$ $\left\{\text{train: } \mathcal{F}_m(\mathcal{D}_{m,i}^{\mathrm{train}}),\ \text{validate: } \mathcal{F}_m(\mathcal{D}_{m,i}^{\mathrm{valid}})\right\}$
    **end**
  **end**
**end**
**Testing**
  Step 2: Test and combine base learners $\boldsymbol{C} = \{C_m\}_{m=1}^{M}$ on $\mathcal{D}^{\mathrm{test}}$
  **for** $i = 1$ **to** $N^{\mathrm{test}}$ **do**
    **for** $m = 1$ **to** $M$ **do**
      $p_m(\tilde{\boldsymbol{y}}_i|\boldsymbol{x}_i) = \{\text{evaluate: } C_m(\boldsymbol{x}_i)\}$
    **end**
    Stack all the logits from the $M$ classifiers into $\boldsymbol{\mathcal{P}} = \{p_1(\tilde{\boldsymbol{y}}_i|\boldsymbol{x}_i), \ldots, p_M(\tilde{\boldsymbol{y}}_i|\boldsymbol{x}_i)\}$
    Generate $\boldsymbol{\mathcal{P}}_{\mathrm{D}} = \boldsymbol{\Theta}_{\mathrm{Dir}}(\boldsymbol{\mathcal{P}})$ or $\boldsymbol{\Theta}_{\mathrm{DirMult}}(\boldsymbol{\mathcal{P}})$
    Initialize the Dirichlet parameters $(\alpha_{i,1}, \ldots, \alpha_{i,K})$
    Estimate the Dirichlet parameters from the ensemble prediction matrix $\boldsymbol{\mathcal{P}}_{\mathrm{D}}$
      $\mathrm{Dir}(\alpha_{i,1}, \ldots, \alpha_{i,K}) = \mathrm{MLE}\ \boldsymbol{\mathcal{P}}_{\mathrm{D}}$
    Calculate $p_{\mathrm{ens}}(\boldsymbol{y}_i|\boldsymbol{x}_i)$ as either (i) mode, (ii) mean, or (iii) mean of $S$ random samples
      (i) $p_{\mathrm{ens}}(\boldsymbol{y}_i|\boldsymbol{x}_i) =$ distribution mode of $\{\mathrm{Dir}(\alpha_{i,1}, \ldots, \alpha_{i,K})\}$
      (ii) $p_{\mathrm{ens}}(\boldsymbol{y}_i|\boldsymbol{x}_i) =$ distribution mean of $\{\mathrm{Dir}(\alpha_{i,1}, \ldots, \alpha_{i,K})\}$
      (iii) $p_{\mathrm{ens}}(\boldsymbol{y}_i|\boldsymbol{x}_i) =$ mean of $\{p_s(\boldsymbol{y}_i|\boldsymbol{x}_i)\}_{i=1}^{S} \sim \{\mathrm{Dir}(\alpha_{i,1}, \ldots, \alpha_{i,K})\}$
  **end**
**end**
**Output:**
Class Label Prediction: $\hat{\boldsymbol{y}}(\boldsymbol{x}_i) = p_{\mathrm{ens}}(\boldsymbol{y}_i|\boldsymbol{x}_i)$

### B. APS Failure at Scania Trucks dataset

The Air Pressure System (APS) is an essential part of the vehicle's pneumatic system consisting of pressurized compressed air for power distribution [66]. In particular, the APS control unit intelligently manages pressurized air, engaging and disengaging the compressor to regulate energy during braking and gear changes. APS is useful in compressed air brake systems of large commercial and passenger vehicles such as trucks, buses, trailers, and railroad trains. As a result, FD systems involving APS are safety-critical, owing to the catastrophic consequences of road accidents resulting from brake system failures.

The APS Failure at Scania Trucks dataset consists of sensor data collected from the Air Pressure System (APS) equipment in heavy Scania trucks mapping out their everyday usage. The task is to generate a model that predicts whether or not a vehicle faces imminent failure specific to the APS component or not. The APS Failure at Scania Trucks dataset is an imbalanced dataset consisting of 60000 training set instances, where 59000 belong to the positive and 1000 to the negative. The positive class consists of truck failures linked to a specific component failure in the APS. On the other hand, the negative class consists of records of truck failures emerging from failures in components unrelated to APS. Also included is a test set consisting of 16000 instances. In total, the dataset has 171 attributes anonymized for proprietary reasons. We apply FD to diagnose the source of the fault as either from the APS or not. For OOD detection evaluation, we generate a set of OOD data based on the APS Failure at Scania Trucks ID dataset using the Gaussian Hyperspheric Offset method [65] with a mean of $4$ and standard deviation $0.7$.

### C. Experimental Setup

For experiments on the Steel Plates Faults dataset, we utilize an ensemble of deep feedforward neural networks (DFNNs). The architecture for each base network consists of four fully-connected layers (216, 108, 54, and 13 output features), with each layer followed by a rectified linear unit (ReLU) [67], a batch normalization layer [68] and a dropout layer [69]. All the base networks use the cross-entropy (CE) loss function during training. We apply $M$-fold CV sampling to the training data, obtaining a classifier per fold to result in a diverse ensemble of $M$ base classifiers. In our case, we evaluate for $M = (5, 10, 20, 30, 50, 100)$; the assortment of base classifiers per ensemble. We train each base classifier for 100 epochs using the Adam optimizer [70] and a base learning rate of 0.1. Through a learning rate scheduler, the base learning rate adaptively changed to 0.01 at epoch 75 and 0.001 at epoch 90 during training. For the optimizer tuning, we ultimately settle on Adam with $\epsilon$ values of $10^{-4}$. We use a large batch size of 128 for all experiments on the Steel Plates Faults dataset, an imbalanced dataset, hence increasing the chances of samples from the minority classes included in each batch during training. We conduct a thorough empirical evaluation of our proposed ensemble combination method, E2D, on a real-world industrial task of defects classification under uncertainty. We compare E2D against the combination methods; averaging and decision templates, as baselines. Additionally, we include evaluations on the same datasets based on evidential deep learning (EDL) DNNs [71] and Deep belief networks (DBNs) [72]

For experiments on the APS Failure at Scania Trucks dataset, we utilize an ensemble of DFNNs. The architecture for each base network consists of three fully-connected layers (492, 328, and 82 output features), with each layer followed by a rectified linear unit (ReLU) [67], and a dropout layer [69]. All the base networks use the cross-entropy (CE) loss function during training. We train the base networks for 10 epochs each, using the stochastic gradient descent (SGD) [73] optimizer and a base learning rate of 0.1. Similar to the previous experiment, the learning rate is adaptively changed through a learning rate scheduler while the selected batch size is 256, informed by the imbalanced nature of the dataset. We apply $M$-fold CV sampling to the training data and obtain a classifier per fold, resulting in a diverse ensemble with $M = (5, 10, 20, 30, 50, 100)$ base classifiers.

Notably, ensemble sizes (5, 10) and (50 and 100) produce similar results, therefore, we report results for $M = 5, 100$.

### D. Evaluation Metrics

For the evaluation of models on predictive uncertainty and OOD detection, we choose the following metrics[1]:

*1) Accuracy (Acc.):* $\uparrow$ measures the model performance as a percentage of correct predictions out of the sum total predictions made. $\text{acc} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\left(y_n \neq \hat{y}_n\right)$, evaluates the model's generalization performance on a hold-out test set. The higher the accuracy score, the more accurate the model's prediction.

*2) Expected Calibration Error (ECE):* $\downarrow$ measures the consensus between classifiers predicted probabilities (confidence) and empirical accuracy.
$\text{ECE} = \sum_{j=1}^{J} \frac{|B_j|}{n} \left|\text{acc}\left(B_j\right) - \text{conf}\left(B_j\right)\right|$, where $n$ represents the number of samples and $B_j$ is the bin $j$ [74].

*3) Maximum Calibration Error (MCE):* $\downarrow$ measures the maximum discrepancy between classifiers predicted probabilities (confidence) and empirical accuracy.
$\text{MCE} = \max_j \left|\text{acc}\left(B_j\right) - \text{conf}\left(B_j\right)\right|$, where $B_j$ represents the bin $j$ [75].

*4) Brier Score (BS):* $\downarrow$ measures the accuracy of predicted probabilities. $\text{BS} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\boldsymbol{p}}_i - \boldsymbol{y}_i\right)^2$, computed as the mean squared error of predicted probabilities and true classes where $\hat{\boldsymbol{p}}$ is a vector of predicted probabilities and $\boldsymbol{y}$ is the one-hot encoded ground truth [76].

*5) Confidence Calibration:* measures the correlation between confidence and correctness of model predictions. For a selected threshold, the metric is provided by the area under the precision-recall curve (AUPRC) [77] as follows:

- Aleatoric Confidence (Alea. Conf.) $\uparrow$ obtained using maximum class probability $\max_k \hat{\boldsymbol{p}}_k$ as the threshold and a binary set of labels where 1 corresponds to correct predictions while 0 to incorrect predictions.
- Epistemic Confidence (Epist. Conf.) $\uparrow$ For E2D: Dirichlet, we use $\max_k \hat{\boldsymbol{\alpha}}_k$ as the threshold, while for averaging

[1]Arrows next to the evaluation metric indicate which direction is better

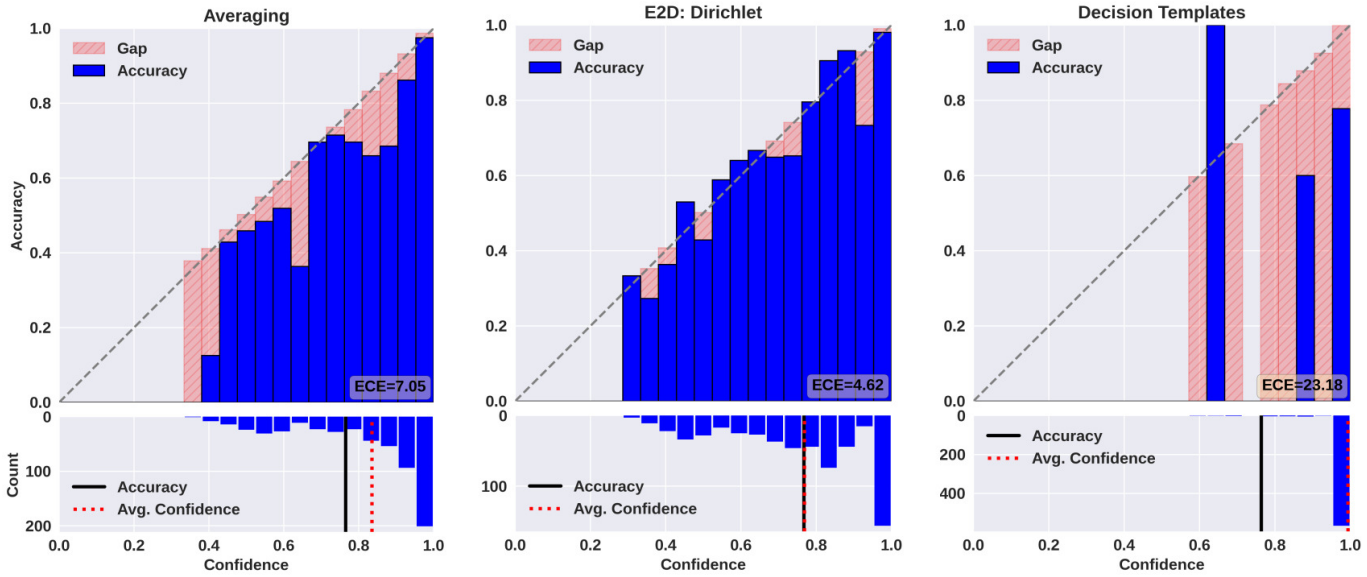


Fig. 2: Reliability diagrams ($B = 21$ bins) and confidence histograms for ensemble size $M = 100$ evaluated on the Steel Plates Faults dataset. Reliability diagrams (top) visualize model calibration generated through accuracy as a function of confidence plots. Confidence histograms (bottom) represent the number of test samples per bin, including two vertical lines indicating the overall accuracy and average confidence. E2D: Dirichlet performs best with the lowest ECE score.

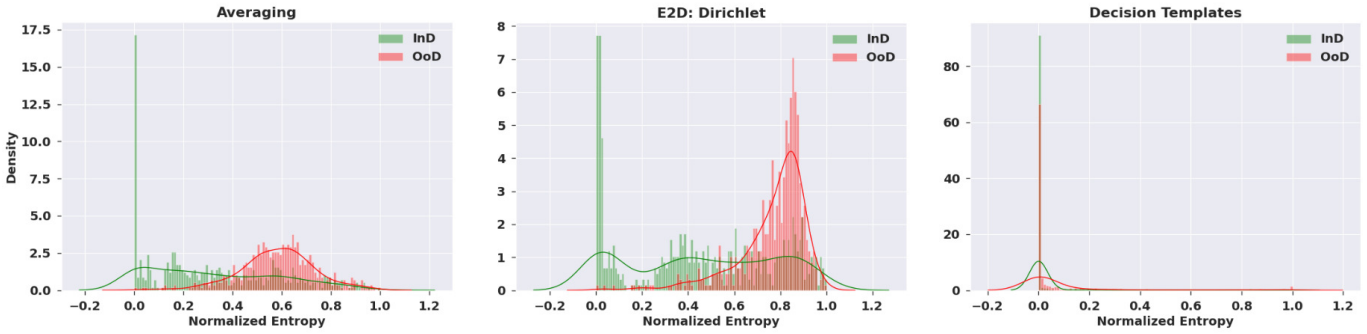


Fig. 3: Predictive entropy density plots of ID and OOD data for ensemble combination methods on the Steel Plates Faults dataset. Entropy scores have been normalized into the range $[0, 1]$. E2D: Dirichlet yields the best divergence between predictive entropies of ID and OOD samples.

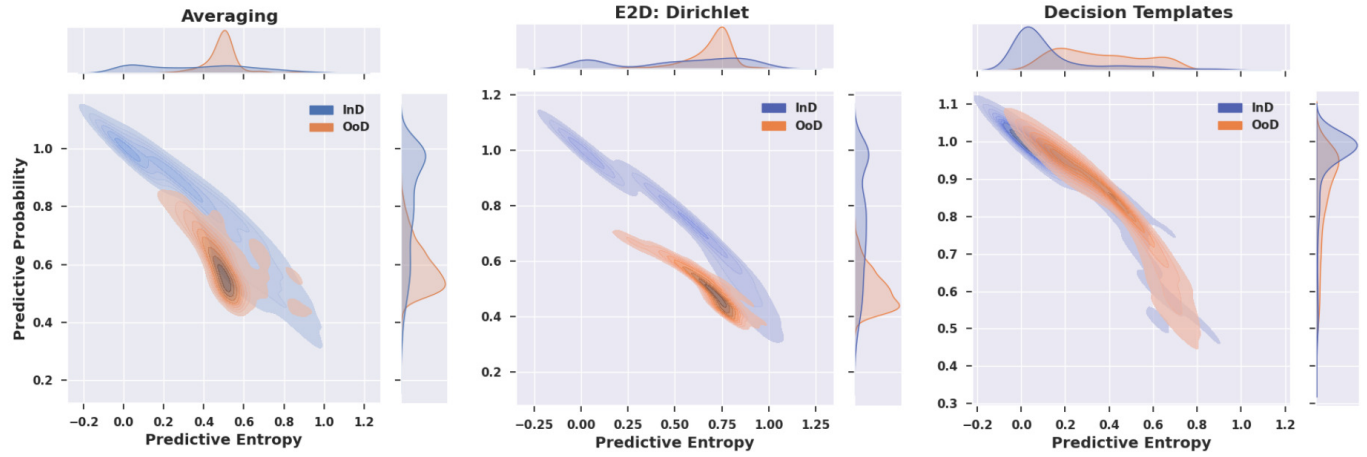


Fig. 4: Contour plots of predictive probabilities against predictive entropies of the combination methods on the Steel Plates Faults dataset. E2D: Dirichlet attains the best separation between ID and OOD data.

and decision templates combination methods, we use the inverse empirical variance of the predicted class $\hat{\boldsymbol{p}}_k$, estimated from $M$ base classifiers predictions as the threshold against a binary set of labels where 1 corresponds to correct predictions while 0 to incorrect predictions.

*6) OOD Detection:* measures the models' ability to detect OOD samples. For a selected threshold, the metric is provided by the area under the precision-recall curve (AUPRC) [77] as follows:

- Aleatoric OOD Detection (OOD Alea.) ↑ obtained using maximum class probability $\max_k \hat{\boldsymbol{p}}_k$ as the threshold and a binary set of labels where 1 corresponds to in-domain data while 0 to out-of-domain data.
- Epistemic OOD Detection (OOD Epist.) ↑ For E2D: Dirichlet, we use $\alpha_0 = \sum_k \hat{\alpha}_k$ as the threshold, while for averaging and decision templates combination methods, we use the inverse empirical variance of the predicted class $\hat{\boldsymbol{p}}_k$, estimated from $M$ base classifiers predictions as the threshold against a binary set of labels where 1 corresponds to in-domain data while 0 to out-of-domain data.

*7) Uncertainty Matrix:* implements the concept of a confusion matrix for $K$ classes using the dimensions of accuracy and uncertainty to yield a $K \times K$ matrix with accurate and inaccurate as rows, certain and uncertain as the columns [78], [79]. The combination of the rows and columns form four possible outcomes; accurate and certain $(\mathrm{AC})$, accurate and uncertain $(\mathrm{AU})$, inaccurate and certain $(\mathrm{IC})$ and inaccurate and uncertain $(\mathrm{IU})$. The following quantitative performance metrics objectively quantify predictive uncertainty estimates:

- $p(\text{accurate}|\text{certain})$ ↑ metric evaluates the quality of predictive uncertainty estimates through a conditional probability measure that the model is accurate on its predictions given it is confident on the same [79]. $p(\mathrm{A}|\mathrm{C}) = \mathrm{n}_{\mathrm{AC}}/(\mathrm{n}_{\mathrm{AC}} + \mathrm{n}_{\mathrm{IC}})$, where $\mathrm{n}$ represents the total number of samples in the given category.
- $p(\text{uncertain}|\text{inaccurate})$ ↑ metric evaluates the quality of predictive uncertainty estimates through a conditional probability measure that the model is uncertain on its predictions given it is inaccurate on the same [79]. $p(\mathrm{U}|\mathrm{I}) = \mathrm{n}_{\mathrm{IU}}/(\mathrm{n}_{\mathrm{IU}} + \mathrm{n}_{\mathrm{IC}})$, where $\mathrm{n}$ represents the total number of samples in the given category.

## V. Results and Discussions

We begin by assessing the calibration of models based on averaging, decision templates, and E2D classifier combination methods across ensemble sizes $M = 5, 100$. In Fig. 2 (top), reliability diagrams visualize model calibration for the three ensemble combination methods evaluated on the Steel Plates Faults dataset. We observe that E2D obtains the best calibration compared to the other methods. In particular, E2D exhibits the lowest deviation from the perfect diagonal, with an ECE score of 4.62 indicating better calibration. Additionally, in Fig. 2 (bottom), the confidence histograms present the distribution of test samples per bin, including model accuracy indicated by the black line and average confidence by the red dotted line. Based on the accuracy-confidence gap, we observe that model predictions from the averaging and decision templates combination methods are over-confident (confidence $>$ accuracy), similar to setting a low decision threshold that is likely to generate false positives. In contrast, the E2D method bridges this gap by attaining a more logical proportion where model confidence and accuracy are equal. Table I presents the results of ECE, MCE, and Brier scores as quantitative measures of model calibration. E2D achieves improved ECE and MCE scores, implying a combination method that generates well-calibrated predictions. For both Steel Plates Faults and APS Failure at Scania datasets, E2D consistently achieves the lowest Brier scores indicating model predictions that are both accurate and confident. We note that the E2D method generates the best model for the true-data distribution, evidenced by the lowest ECE scores across all ensemble sizes.

We then investigate the quality of predictive uncertainty estimates using conditional probabilities $p(\text{accurate}|\text{certain})$ and $p(\text{uncertain}|\text{inaccurate})$ evaluated against the predictive, mutual and differential entropies as uncertainty thresholds. Table II presents the conditional probabilities results from the experiments on the Steel Plates Faults dataset. E2D achieves improved $p(\mathrm{U}|\mathrm{I})$ and $p(\mathrm{A}|\mathrm{C})$ scores on all the thresholds. Crucially, E2D and EDL methods generates a Dirichlet distribution as the output upon which we apply differential entropy, a metric that captures elements of data uncertainty and is suitable for measuring distributional uncertainty [80]. We observe that both E2D and EDL achieves significant improvements on $p(\mathrm{U}|\mathrm{I})$ and $p(\mathrm{A}|\mathrm{C})$ scores across all ensemble sizes based on the differential entropy threshold. Nonetheless, compared to EDL, E2D has superior score based on differential entropy. Table III presents the results of these conditional probabilities on the APS Failure at Scania Trucks dataset. E2D and the averaging combination methods achieve comparable performance above the decision templates method on the predictive and mutual entropy thresholds. Nonetheless, for differential entropy threshold, E2D achieves the best scores for both $p(\mathrm{U}|\mathrm{I})$ and $p(\mathrm{A}|\mathrm{C})$. It is important to note that the differential entropy from EDL and E2D-based models provides a reliable measure of uncertainty, especially useful for tasks in the safety-critical domain.

We also investigate the proposed combination methods for the task of OOD samples detection. Fundamentally, the metrics aleatoric and epistemic confidence seek to establish the likelihood of correct predictions given high confidence [77]. Table IV presents results from experiments of confidence calibration and OOD detection evaluated on the Steel Plates Faults dataset. E2D achieves improved aleatoric and epistemic confidence scores, indicating that the generated confident predictions are likely to be accurate. Confidence calibration in safety-critical systems is essential in providing insight into the level of trust accorded to system-generated predictions. Additionally, E2D emerges as the best combination method, consistently outperforming the other two methods on both the aleatoric OOD and epistemic OOD scores. Table V presents results from evaluation on the APS Failure at Scania Trucks dataset. E2D is the top performing combination method, achieving perfect scores on OOD detection and confidence

TABLE I: Brier scores, ECE, and MCE test set results in ($B = 21$ bins) for ensemble ($M = 5,100$ base classifiers), DBN and Evidence DNN evaluated on Steel Plates Fault and APS Failure at Scania Trucks datasets. Best results are in bold.

| M | Method | Steel Plates Faults | | | APS Failure at Scania | | |
|---|---|---|---|---|---|---|---|
| | | Brier ↓ | ECE ↓ | MCE ↓ | Brier ↓ | ECE ↓ | MCE ↓ |
| 5 | Averaging | 0.0473 | 5.9445 | 0.6510 | 0.0213 | 2.3839 | 0.3687 |
| | D. Template | 0.0614 | 20.3154 | 0.6300 | 0.0603 | 4.3605 | 0.2096 |
| | E2D: Dirichlet | 0.0475 | 5.2809 | 0.3463 | 0.0213 | 2.2625 | 0.3438 |
| 100 | Averaging | 0.0471 | 7.0521 | 0.3782 | 0.0214 | 2.5671 | 0.4974 |
| | D. Template | 0.0658 | 23.1761 | 0.9255 | 0.0606 | 6.0056 | 0.4311 |
| | E2D: Dirichlet | 0.0464 | 4.6168 | 0.1958 | 0.0216 | 1.3018 | 0.7885 |
| – | DBN-Logistic | 0.0672 | 5.8215 | 0.2885 | 0.0559 | 15.6194 | 0.2080 |
| | DBN-SGD | 0.1104 | 3.3971 | 0.0340 | 0.0229 | 0.6690 | 0.0067 |
| | DBN-MLP | 0.0553 | 9.9783 | 0.3156 | 0.0154 | 0.5010 | 0.1991 |
| – | EDL-MSE | 0.0896 | 6.3977 | 0.8054 | 0.0241 | 3.4268 | 0.0343 |
| | EDL-Log | 0.0788 | 6.1832 | 0.5095 | 0.0234 | 2.2143 | 0.0221 |
| | EDL-Digamma | 0.0704 | 13.1888 | 0.6413 | 0.0235 | 2.5663 | 0.0257 |

TABLE II: Results of the quality of uncertainty estimates using conditional probabilities $p(\text{accurate}|\text{certain})$ and $p(\text{uncertain}|\text{inaccurate})$ evaluated against the predictive, mutual, and differential entropy thresholds across ensemble ($M = 5,100$ base classifiers), DBN, and Evidence DNN on the Steel Plates Faults dataset. Best results are in bold.

| M | Method | Predictive Ent. | | Mutual Ent. | | Differential Ent. | |
|---|---|---|---|---|---|---|---|
| | | $p(\text{U}\|\text{I})\uparrow$ | $p(\text{A}\|\text{C})\uparrow$ | $p(\text{U}\|\text{I})\uparrow$ | $p(\text{A}\|\text{C})\uparrow$ | $p(\text{U}\|\text{I})\uparrow$ | $p(\text{A}\|\text{C})\uparrow$ |
| 5 | Averaging | 0.58 ± 0.37 | 0.88 ± 0.09 | 0.58 ± 0.37 | 0.88 ± 0.09 | – | – |
| | D. Template | 0.24 ± 0.25 | 0.80 ± 0.06 | 0.24 ± 0.25 | 0.80 ± 0.06 | – | – |
| | E2D: Dirichlet | 0.68 ± 0.36 | 0.90 ± 0.08 | 0.74 ± 0.31 | 0.91 ± 0.07 | 0.91 ± 0.25 | 0.97 ± 0.06 |
| 100 | Averaging | 0.54 ± 0.36 | 0.87 ± 0.08 | 0.54 ± 0.36 | 0.87 ± 0.08 | – | – |
| | D. Template | 0.11 ± 0.19 | 0.78 ± 0.04 | 0.11 ± 0.20 | 0.79 ± 0.05 | – | – |
| | E2D: Dirichlet | 0.67 ± 0.34 | 0.90 ± 0.08 | 0.71 ± 0.30 | 0.91 ± 0.07 | 0.85 ± 0.29 | 0.94 ± 0.07 |
| – | DBN-Logistic | 0.62 ± 0.39 | 0.77 ± 0.21 | 0.62 ± 0.39 | 0.77 ± 0.21 | – | – |
| | DBN-SGD | 0.00 ± 0.00 | 0.37 ± 0.00 | 0.00 ± 0.00 | 0.37 ± 0.00 | – | – |
| | DBN-MLP | 0.52 ± 0.35 | 0.84 ± 0.09 | 0.52 ± 0.35 | 0.84 ± 0.09 | – | – |
| – | EDL-MSE | 0.92 ± 0.21 | 0.92 ± 0.10 | 0.93 ± 0.21 | 0.94 ± 0.11 | 0.93 ± 0.21 | 0.94 ± 0.11 |
| | EDL-Log | 0.88 ± 0.20 | 0.89 ± 0.10 | 0.91 ± 0.21 | 0.93 ± 0.11 | 0.92 ± 0.21 | 0.94 ± 0.11 |
| | EDL-Digamma | 0.80 ± 0.22 | 0.90 ± 0.08 | 0.88 ± 0.21 | 0.93 ± 0.08 | 0.88 ± 0.22 | 0.94 ± 0.08 |

TABLE III: Results of the quality of uncertainty estimates using conditional probabilities $p(\text{accurate}|\text{certain})$ and $p(\text{uncertain}|\text{inaccurate})$ evaluated against the predictive, mutual, and differential entropy thresholds across ensemble ($M = 5,100$ base classifiers), DBN, and Evidence DNN on the APS Failure at Scania Trucks dataset. Best results are in bold.

| M | Method | Predictive Ent. | | Mutual Ent. | | Differential Ent. | |
|---|---|---|---|---|---|---|---|
| | | $p(\text{U}\|\text{I})\uparrow$ | $p(\text{A}\|\text{C})\uparrow$ | $p(\text{U}\|\text{I})\uparrow$ | $p(\text{A}\|\text{C})\uparrow$ | $p(\text{U}\|\text{I})\uparrow$ | $p(\text{A}\|\text{C})\uparrow$ |
| 5 | Averaging | 0.90 ± 0.20 | 1.00 ± 0.00 | 0.90 ± 0.20 | 1.00 ± 0.00 | – | – |
| | D. Template | 0.29 ± 0.20 | 0.95 ± 0.01 | 0.29 ± 0.20 | 0.95 ± 0.01 | – | – |
| | E2D: Dirichlet | 0.89 ± 0.23 | 1.00 ± 0.01 | 0.87 ± 0.25 | 1.00 ± 0.01 | 0.93 ± 0.21 | 1.00 ± 0.00 |
| 100 | Averaging | 0.82 ± 0.27 | 1.00 ± 0.01 | 0.82 ± 0.27 | 1.00 ± 0.01 | – | – |
| | D. Template | 0.11 ± 0.20 | 0.94 ± 0.01 | 0.11 ± 0.20 | 0.94 ± 0.01 | – | – |
| | E2D: Dirichlet | 0.88 ± 0.25 | 1.00 ± 0.01 | 0.86 ± 0.27 | 1.00 ± 0.01 | 0.93 ± 0.21 | 1.00 ± 0.00 |
| – | DBN-Logistic | 0.53 ± 0.21 | 0.87 ± 0.14 | 0.53 ± 0.21 | 0.87 ± 0.14 | – | – |
| | DBN-SGD | 0.00 ± 0.00 | 0.98 ± 0.00 | 0.00 ± 0.00 | 0.98 ± 0.00 | – | – |
| | DBN-MLP | 0.62 ± 0.22 | 0.99 ± 0.00 | 0.62 ± 0.22 | 0.99 ± 0.00 | – | – |
| – | EDL-MSE | 0.56 ± 0.21 | 0.98 ± 0.24 | 0.56 ± 0.21 | 0.98 ± 0.21 | 0.91 ± 0.24 | 0.98 ± 0.21 |
| | EDL-Log | 0.57 ± 0.20 | 0.98 ± 0.22 | 0.57 ± 0.20 | 0.98 ± 0.22 | 0.92 ± 0.22 | 0.98 ± 0.22 |
| | EDL-Digamma | 0.53 ± 0.23 | 0.98 ± 0.21 | 0.53 ± 0.23 | 0.98 ± 0.21 | 0.89 ± 0.22 | 0.98 ± 0.21 |

TABLE IV: Results of accuracy, confidence calibration, and OOD detection for ensemble ($M = 5, 100$ base classifiers), DBN, and Evidence DNN evaluated on the Steel Plates Faults dataset. Best results are in bold.

| M | Method | Acc ↑ | Alea Conf ↑ | Epist Conf ↑ | OOD Alea ↑ | OOD Epist ↑ |
|---|---|---|---|---|---|---|
| 5 | Averaging | 76.16 | 93.40 | 62.33 | 34.50 | 64.10 |
| | D. Template | 75.13 | 92.52 | 59.78 | 82.80 | 39.50 |
| | E2D: Dirichlet | 76.16 | 92.98 | 92.54 | 87.40 | 84.10 |
| 100 | Averaging | 76.50 | 93.79 | 61.73 | 33.6 | 69.50 |
| | D. Template | 76.32 | 91.20 | 63.02 | 89.35 | 34.72 |
| | E2D: Dirichlet | 76.67 | 93.58 | 92.54 | 89.40 | 88.40 |
| – | DBN-Logistic | 63.98 | 85.51 | 50.20 | 50.0 | 50.0 |
| | DBN-SGD | 36.71 | 36.71 | 36.71 | 50.0 | 50.0 |
| | DBN-MLP | 73.58 | 90.92 | 60.70 | 50.0 | 50.0 |
| – | EDL-MSE | 38.25 | 79.86 | 79.86 | 66.88 | 66.88 |
| | EDL-Log | 51.97 | 92.02 | 92.00 | 67.69 | 67.70 |
| | EDL-Diagamma | 57.97 | 90.98 | 90.98 | 71.23 | 71.23 |

TABLE V: Results of accuracy, confidence calibration, and OOD detection for ensemble ($M = 5, 100$ base classifiers), DBN, and Evidence DNN evaluated on the APS Failure at Scania Trucks dataset. Best results are in bold.

| M | Method | Acc ↑ | Alea Conf ↑ | Epist Conf ↑ | OOD Alea ↑ | OOD Epist ↑ |
|---|---|---|---|---|---|---|
| 5 | Averaging | 97.66 | 99.92 | 91.68 | 50.00 | 95.60 |
| | D. Template | 93.18 | 91.98 | 88.35 | 35.16 | 95.37 |
| | E2D: Dirichlet | 97.66 | 99.92 | 99.92 | 97.70 | 97.70 |
| 100 | Averaging | 97.66 | 99.97 | 91.19 | 50.0 | 100.0 |
| | D. Template | 93.75 | 97.80 | 86.95 | 99.06 | 32.29 |
| | E2D: Dirichlet | 97.66 | 99.97 | 99.97 | 100.0 | 100.0 |
| – | DBN-Logistic | 96.51 | 93.89 | 93.97 | 50.0 | 50.0 |
| | DBN-SGD | 97.66 | 97.66 | 97.66 | 50.0 | 50.0 |
| | DBN-MLP | 97.88 | 99.87 | 92.40 | 50.0 | 50.0 |
| – | EDL-MSE | 97.52 | 97.66 | 97.30 | 50.0 | 50.0 |
| | EDL-Log | 97.77 | 97.86 | 97.56 | 50.0 | 50.0 |
| | EDL-Diagamma | 97.78 | 97.56 | 97.50 | 50.0 | 50.0 |

calibration for all the ensemble sizes $M$. DBN and EDL perform poorly on the task of OOD detection, possibly because they lack the additional insight available to ensemble classifiers.

In Fig. 3, we illustrate the divergence between the predictive entropies of in-domain and OOD sample predictions from an ensemble ($M = 100$ base classifiers) using the three combination methods. E2D yields the best divergence in predictive entropies among the three methods, with OOD samples predominantly obtaining high entropies while ID obtaining low entropies. This divergence highlights the E2D model's ability to distinguish between ID and OOD samples through the measure of entropy. For an industrial use case scenario, we select a normalized entropy threshold of 0.4, the lowest, most feasible setting across the three combination methods, based on the plots in Fig. 3. E2D model is more robust as it can operate under higher threshold levels, increasing the accuracy of OOD detection while still minimizing the number of false negatives. In particular, setting the threshold at 0.6 for E2D ensures only the high entropy samples are classified as OOD, reducing the likelihood of incorrectly declaring ID data as OOD data. Further, we report additional visualizations of the predictive entropies divergence in Fig. 4. E2D model achieves the best separation between ID and OOD samples, evidenced by the centers of the contours appearing furthest apart. Therefore, E2D is the most effective method for distinguishing between ID and OOD samples, especially with OOD samples typically yielding high predictive entropies. We observe that E2D predictions are more reliable as the OOD samples rarely attain high probabilities. For safety-critical systems, it is much easier to introduce threshold strategies under which the system can distinguish between ID and OOD samples.

## VI. Conclusion

In this paper, we have presented E2D, an uncertainty-aware ensemble combination method for an ensemble of DL-based FD models to help monitor the stability of industrial processes and product quality. In particular, E2D generates a continuous multivariate probability distribution as the combined model output, replacing deterministic point estimation techniques that are ineffective in capturing the underlying model predictive uncertainties. Further, E2D is a post hoc application, implementable at inference time, and compatible with diverse pre-trained models. E2D enables robust uncertainty estimates through differential entropy, particularly useful in generating application-grounded interpretability of the model predictions, further enhancing the safety of the end task. From experiments on the steel plates faults and APS failure at Scania trucks datasets, we demonstrate that E2D achieves high-quality uncertainty predictions, improved model calibration, and OOD

detection. In future work, we aim to extend the combiner module to a more general implementation that includes different probability distributions. We also intend to evaluate the method on additional ensemble-type strategies such as bagging and boosting.

## REFERENCES


[1] K.-D. Thoben, S. Wiesner, and T. Wuest, ""industrie 4.0" and smart manufacturing – a review of research issues and application examples," *International Journal of Automation Technology*, vol. 11, pp. 4–19, 01 2017.

[2] P. O'Donovan, K. Bruton, and D. T. O'Sullivan, "Case study: the implementation of a data-driven industrial analytics methodology and platform for smart manufacturing," *International Journal of Prognostics and Health Management*, vol. 7, no. 3, 2016.

[3] J. Davis, T. Edgar, R. Graybill, P. Korambath, B. Schott, D. Swink, J. Wang, and J. Wetzel, "Smart manufacturing," *Annual Review of Chemical and Biomolecular Engineering*, vol. 6, no. 1, pp. 141–160, 2015, pMID: 25898070. [Online]. Available: https://doi.org/10.1146/annurev-chembioeng-061114-123255

[4] J. G. Koomey, H. Scott Matthews, and E. Williams, "Smart everything: Will intelligent systems reduce resource use?" *Annual Review of Environment and Resources*, vol. 38, no. 1, pp. 311–343, 2013. [Online]. Available: https://doi.org/10.1146/annurev-environ-021512-110549

[5] D. M. Tilbury, "Cyber-physical manufacturing systems," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 2, no. 1, pp. 427–443, 2019. [Online]. Available: https://doi.org/10.1146/annurev-control-053018-023652

[6] Y. Jiang, S. Yin, and O. Kaynak, "Performance supervised plant-wide process monitoring in industry 4.0: A roadmap," *IEEE Open Journal of the Industrial Electronics Society*, vol. 2, pp. 21–35, 2021.

[7] J. Yuchen, Y. Shen, and K. Okyay, "Data-driven monitoring and safety control of industrial cyber-physical systems: Basics and beyond," *IEEE Access*, vol. 6, pp. 47 374–47 384, 2018.

[8] Y. Jiang, K. Li, and S. Yin, "Cyber-physical system based factory monitoring and fault diagnosis framework with plant-wide performance optimization," *2018 IEEE Industrial Cyber-Physical Systems (ICPS)*, pp. 240–245, 2018.

[9] K. Hadad, M. Pourahmadi, and H. Majidi-Maraghi, "Fault diagnosis and classification based on wavelet transform and neural network," *Progress in nuclear energy*, vol. 53, no. 1, pp. 41–47, 2011.

[10] L. Wen, L. Gao, and X. Li, "A new deep transfer learning based on sparse auto-encoder for fault diagnosis," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 1, pp. 136–144, 2019.

[11] Y.-J. Park, S.-K. S. Fan, and C.-Y. Hsu, "A review on fault detection and process diagnostics in industrial processes," *Processes*, vol. 8, no. 9, p. 1123, 2020.

[12] B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski, and M. Woźniak, "Ensemble learning for data stream analysis: A survey," *Information Fusion*, vol. 37, pp. 132–156, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253516302329

[13] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*. Springer, 2000, pp. 1–15.

[14] R. Saini and S. Ghosh, "Ensemble classifiers in remote sensing: A review," in *2017 International Conference on Computing, Communication and Automation (ICCCA)*, 2017, pp. 1148–1152.

[15] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1249, 2018.

[16] Y. Liu, "How to find different neural networks by negative correlation learning," in *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, vol. 5. IEEE, 2005, pp. 3330–3333.

[17] Y. Liu and X. Yao, "Ensemble learning via negative correlation," *Neural networks*, vol. 12, no. 10, pp. 1399–1404, 1999.

[18] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE transactions on pattern analysis and machine intelligence*, vol. 12, no. 10, pp. 993–1001, 1990.

[19] A. Malinin and M. Gales, "Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[20] L. Smith and Y. Gal, "Understanding measures of uncertainty for adversarial example detection," *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, vol. 2, pp. 560–569, 2018.

[21] A. Malinin, B. Mlodozeniec, and M. J. F. Gales, "Ensemble distribution distillation," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=BygSP6Vtvr

[22] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6405–6416.

[23] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *Advances in neural information processing systems*, vol. 30, 2017.

[24] S. Depeweg, J. M. Hernández-Lobato, F. Doshi-Velez, and S. Udluft, "Decomposition of uncertainty for active learning and reliable reinforcement learning in stochastic systems," *stat*, vol. 1050, p. 11, 2017.

[25] S. Depeweg, J.-M. Hernandez-Lobato, F. Doshi-Velez, and S. Udluft, "Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1184–1193.

[26] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[27] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *CoRR*, vol. abs/1503.02531, 2015. [Online]. Available: http://arxiv.org/abs/1503.02531

[28] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine learning*, vol. 51, no. 2, p. 181, 2003.

[29] J. Kafunah, M. I. Ali, and J. G. Breslin, "Handling imbalanced datasets for robust deep neural network-based fault detection in manufacturing systems," *Applied Sciences*, vol. 11, no. 21, p. 9783, 2021.

[30] M. Kläs and A. M. Vollmer, "Uncertainty in machine learning applications: A practice-driven classification of uncertainty," in *Computer Safety, Reliability, and Security: SAFECOMP 2018 Workshops, ASSURE, DECSoS, SASSUR, STRIVE, and WAISE, Västerås, Sweden, September 18, 2018, Proceedings 37*. Springer, 2018, pp. 431–438.

[31] K. S. Woods, W. P. Kegelmeyer, and K. Bowyer, "Combination of multiple classifiers using local accuracy estimates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, pp. 405–410, 1997.

[32] D. M. Tax, M. Van Breukelen, R. P. Duin, and J. Kittler, "Combining multiple classifiers by averaging or by multiplying?" *Pattern recognition*, vol. 33, no. 9, pp. 1475–1485, 2000.

[33] Z.-H. Zhou, *Ensemble methods: foundations and algorithms*. CRC press, 2012.

[34] G. Fumera and F. Roli, "Performance analysis and comparison of linear combiners for classifier fusion," in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*. Springer, 2002, pp. 424–432.

[35] J. Kittler, "Combining classifiers: A theoretical framework," *Pattern Anal. Appl.*, vol. 1, no. 1, pp. 18–27, 1998. [Online]. Available: https://doi.org/10.1007/BF01238023

[36] L. I. Kuncheva, J. C. Bezdek, and R. P. Duin, "Decision templates for multiple classifier fusion: an experimental comparison," *Pattern recognition*, vol. 34, no. 2, pp. 299–314, 2001.

[37] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.

[38] Z. Gao, C. Cecati, and S. X. Ding, "A survey of fault diagnosis and fault-tolerant techniques—part i: Fault diagnosis with model-based and signal-based approaches," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 6, pp. 3757–3767, 2015.

[39] G. Zhiwei, C. Carlo, and D. S. X., "A survey of fault diagnosis and fault-tolerant techniques—part ii: Fault diagnosis with knowledge-based and hybrid/active approaches," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 6, pp. 3768–3774, 2015.

[40] Z. Ge and J. Chen, "Plant-wide industrial process monitoring: A distributed modeling framework," *IEEE Transactions on Industrial Informatics*, vol. 12, pp. 310–321, 2016.

[41] M. Zhao, S. Zhong, X. Fu, B. Tang, and M. Pecht, "Deep residual shrinkage networks for fault diagnosis," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 7, pp. 4681–4690, 2019.

[42] F. Jia, Y. Lei, J. Lin, X. Zhou, and N. Lu, "Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data," *Mechanical systems and signal processing*, vol. 72, pp. 303–315, 2016.

[43] W. Lu, B. Liang, Y. Cheng, D. Meng, J. Yang, and T. Zhang, "Deep model based domain adaptation for fault diagnosis," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 3, pp. 2296–2305, 2016.

[44] P. Wang, R. X. Gao, and R. Yan, "A deep learning-based approach to material removal rate prediction in polishing," *CIRP annals*, vol. 66, no. 1, pp. 429–432, 2017.

[45] J. Zhang, Y. Jiang, H. Luo, and S. Yin, "Prediction of material removal rate in chemical mechanical polishing via residual convolutional neural network," *Control Engineering Practice*, vol. 107, p. 104673, 2021.

[46] X. Yuan, S. Qi, Y. Wang, and H. Xia, "A dynamic cnn for nonlinear dynamic feature learning in soft sensor modeling of industrial process data," *Control Engineering Practice*, vol. 104, p. 104614, 2020.

[47] X. Yuan, L. Li, Y. A. Shardt, Y. Wang, and C. Yang, "Deep learning with spatiotemporal attention-based lstm for industrial soft sensor model development," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 5, pp. 4404–4414, 2020.

[48] J. Loy-Benitez, S. Heo, and C. Yoo, "Soft sensor validation for monitoring and resilient control of sequential subway indoor air quality through memory-gated recurrent neural networks-based autoencoders," *Control Engineering Practice*, vol. 97, p. 104330, 2020.

[49] K. B. Lee, S. Cheon, and C. O. Kim, "A convolutional neural network for fault classification and diagnosis in semiconductor manufacturing processes," *IEEE Transactions on Semiconductor Manufacturing*, vol. 30, no. 2, pp. 135–142, 2017.

[50] J. Lin, "On the dirichlet distribution," *Department of Mathematics and Statistics, Queens University*, pp. 10–11, 2016.

[51] T. Minka, "Estimating a dirichlet distribution," 2000.

[52] M. Sklar, "Fast mle computation for the dirichlet multinomial," *arXiv preprint arXiv:1405.0099*, 2014.

[53] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[54] G. Ronning, "Maximum likelihood estimation of dirichlet distributions," *Journal of Statistical Computation and Simulation*, vol. 32, pp. 215–221, 1989.

[55] N. Wicker, J. Muller, R. K. R. Kalathur, and O. Poch, "A maximum likelihood approximation method for dirichlet's parameter estimation," *Computational Statistics & Data Analysis*, vol. 52, no. 3, pp. 1315–1322, 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167947307002848

[56] M. Giordan and R. Wehrens, "A comparison of computational approaches for maximum likelihood estimation of the dirichlet parameters on high-dimensional data," *Sort-statistics and Operations Research Transactions*, vol. 39, pp. 109–126, 2013.

[57] K. W. Ng, G.-L. Tian, and M.-L. Tang, *Dirichlet and related distributions: Theory, methods and applications*. John Wiley & Sons, 2011.

[58] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language*, vol. 13, no. 4, pp. 359–394, 1999.

[59] D. Dua and C. Graff, "UCI Machine Learning Repository: APS Failure at Scania Trucks Data Set," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[60] D. Dheeru and G. Casey, "UCI Machine Learning Repository: Steel Plates Faults Data Set," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[61] E. C. ÖZKAT, "A method to classify steel plate faults based on ensemble learning," *Journal of Materials and Mechatronics: A*, vol. 3, no. 2, pp. 240–256.

[62] L. Yang, X. Huang, Y. Ren, and Y. Huang, "Steel plate surface defect detection based on dataset enhancement and lightweight convolution neural network," *Machines*, vol. 10, no. 7, p. 523, 2022.

[63] Z. Hao, Z. Wang, D. Bai, B. Tao, X. Tong, and B. Chen, "Intelligent detection of steel defects based on improved split attention networks," *Frontiers in Bioengineering and Biotechnology*, vol. 9, p. 1478, 2022.

[64] N. Chen, J. Sun, X. Wang, Y. Huang, Y. Li, and C. Guo, "Research on surface defect detection and grinding path planning of steel plate based on machine vision," in *2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA)*. IEEE, 2019, pp. 1748–1753.

[65] F. Möller, D. Botache, D. Huseljic, F. Heidecker, M. Bieshaar, and B. Sick, "Out-of-distribution detection and generation using soft brownian offset sampling and autoencoders," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021, pp. 46–55.

[66] B. Karanja and P. Broukhiyan, "Commercial vehicle air consumption: Simulation, validation and recommendation," 2017. [Online]. Available: http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-209657

[67] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML 2010), Haifa, Israel*, 2010.

[68] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 448–456. [Online]. Available: http://proceedings.mlr.press/v37/ioffe15.html

[69] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, p. 1929–1958, Jan. 2014.

[70] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pp. 1–15, 2015.

[71] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," *Advances in neural information processing systems*, vol. 31, 2018.

[72] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[73] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural networks: Tricks of the trade*. Springer, 2012, pp. 9–48.

[74] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*. PMLR, 2017, pp. 1321–1330.

[75] M. P. Naeini, G. F. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using bayesian binning," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15), Austin Texas, USA.*, vol. 2015. AAAI Press, 25–30 Jan 2015, pp. 2901–2907.

[76] G. W. Brier *et al.*, "Verification of forecasts expressed in terms of probability," *Monthly weather review*, vol. 78, no. 1, pp. 1–3, 1950.

[77] B. Charpentier, D. Zügner, and S. Günnemann, "Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1356–1367, 2020.

[78] H. Asgharnezhad, A. Shamsi, R. Alizadehsani, A. Khosravi, S. Nahavandi, Z. A. Sani, D. Srinivasan, and S. M. S. Islam, "Objective evaluation of deep uncertainty predictions for covid-19 detection," *Scientific Reports*, vol. 12, no. 1, pp. 1–11, 2022.

[79] J. Mukhoti and Y. Gal, "Evaluating bayesian deep learning methods for semantic segmentation," *ArXiv*, vol. abs/1811.12709, 2018.

[80] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," *Advances in neural information processing systems*, vol. 31, 2018.



**Jefkine Kafunah** received a B.Sc. degree in Actuarial Science from the Jomo Kenyatta University of Agriculture and Technology, Kenya in 2008. Since 2019, he has been working toward the Ph.D. degree in computer science from the College of Science and Engineering, University of Galway, Ireland. His research interests include data analytics, data-driven process monitoring, fault diagnosis and prognosis, industrial cyber-physical systems, and artificial intelligence. Jefkine's homepage: jefkine.com



**Muhammad Intizar Ali** is an Assistant Professor in the School of Electronic Engineering, Dublin City University. He received the Ph.D. (Hons) degree from the Vienna University of Technology, Austria, in 2011. His research interests include semantic Web, data analytics, Internet of Things (IoT), linked data, federated query processing, stream query processing, and optimal query processing over large scale distributed data sources. He is actively involved in various EU funded and industry-funded projects aimed at providing IoT enabled adaptive intelligence for smart applications. He serves as a PC Member of various journals, international conferences, and workshops. Ali's homepage: intizarali.org

**John Breslin** (M'94–SM'16) is a Professor in Electronic Engineering at the University of Galway, where he is Director of the TechInnovate and AgInnovate entrepreneurship programmes. Associated with three SFI Research Centres, he is a Co-Principal Investigator at Confirm (Smart Manufacturing) and Insight (Data Analytics), a Funded Investigator at VistaMilk (AgTech), and a Principal Investigator on the Horizon 2020 CSA OntoCommons. He has co-authored around 300 publications, including the books "The Social Semantic Web", "Social Semantic Web Mining", "Old Ireland in Colour" and "Old Ireland in Colour 2". He co-created the SIOC framework, implemented in hundreds of applications (by Yahoo, Boeing, Vodafone, etc.) on at least 65,000 websites with 35 million data instances. He is co-founder of the PorterShed, boards.ie and adverts.ie.