Information Fusion 101 (2024) 101978

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

Evaluation of uncertainty quantification methods in multi-label classification: A case study with automatic diagnosis of electrocardiogram

Marília Barandas [a,b,*], Lorenzo Famiglini [c], Andrea Campagner [c,d], Duarte Folgado [a,b], Raquel Simão [b], Federico Cabitza [c,d], Hugo Gamboa [a,b]

[a] Associação Fraunhofer Portugal Research, Rua Alfredo Allen 455/461, Porto, 4200-135, Portugal
[b] LIBPhys (Laboratory for Instrumentation, Biomedical Engineering and Radiation Physics), NOVA School of Science and Technology, Campus de Caparica, 2829-516, Portugal
[c] Department of Informatics, Systemics and Communication, University of Milano-Bicocca, Viale Sarca 336, Milan, 20126, Italy
[d] IRCCS Istituto Ortopedico Galeazzi, Via Riccardo Galeazzi, 4, Milan, 20161, Italy

ARTICLE INFO

Keywords:
Artificial Intelligence
Uncertainty quantification
Multi-label classification
Cardiology

ABSTRACT

Artificial Intelligence (AI) use in automated Electrocardiogram (ECG) classification has continuously attracted the research community's interest, motivated by their promising results. Despite their great promise, limited attention has been paid to the robustness of their results, which is a key element for their implementation in clinical practice. Uncertainty Quantification (UQ) is a critical for trustworthy and reliable AI, particularly in safety-critical domains such as medicine. Estimating uncertainty in Machine Learning (ML) model predictions has been extensively used for Out-of-Distribution (OOD) detection under single-label tasks. However, the use of UQ methods in multi-label classification remains underexplored.

This study goes beyond developing highly accurate models comparing five uncertainty quantification methods using the same Deep Neural Network (DNN) architecture across various validation scenarios, including internal and external validation as well as OOD detection, taking multi-label ECG classification as the example domain. We show the importance of external validation and its impact on classification performance, uncertainty estimates quality, and calibration. Ensemble-based methods yield more robust uncertainty estimations than single network or stochastic methods. Although current methods still have limitations in accurately quantifying uncertainty, particularly in the case of dataset shift, incorporating uncertainty estimates with a classification with a rejection option improves the ability to detect such changes. Moreover, we show that using uncertainty estimates as a criterion for sample selection in active learning setting results in greater improvements in classification performance compared to random sampling.

1. Introduction

Machine learning has made significant progress in a variety of decision-critical domains, including medicine. However, as these advancements are applied in real-world safety-critical applications, it is crucial to consider the inherent uncertainty present in the ML process as a path toward trustworthy AI [1]. While AI research has achieved promising results across various domains, the adoption of AI in the medical field remains a challenge [2]. This can be attributed to various factors, including the lack of trust in AI decisions. In medical AI, it is essential to have the ability to abstain from providing a decision when there is a high level of uncertainty associated with it. This mirrors the clinical practice of seeking a second opinion in unusual or complex cases. However, the quantification and communication of uncertainty are not routinely addressed in the current literature, yet they are crucial in healthcare applications [3].

Another important topic is multi-label classification, where multiple nonexclusive labels may be assigned to each instance, as opposed to multi-class or binary classification where a single label is assigned to each instance. The applications of multi-label include many real world problems such as text classification, music information retrieval, image classification, and time series analysis problems (e.g. ECG classification). Although the applications are vast, the multi-label studies that include UQ in their analysis are mainly associated with image recognition [4] or text classification [5]. Still, even in the mentioned

* Corresponding author at: LIBPhys (Laboratory for Instrumentation, Biomedical Engineering and Radiation Physics), NOVA School of Science and Technology, Campus de Caparica, 2829-516, Portugal.
E-mail address: m.barandas@campus.fct.unl.pt (M. Barandas).

https://doi.org/10.1016/j.inffus.2023.101978
Received 11 January 2023; Received in revised form 8 August 2023; Accepted 16 August 2023
Available online 22 August 2023

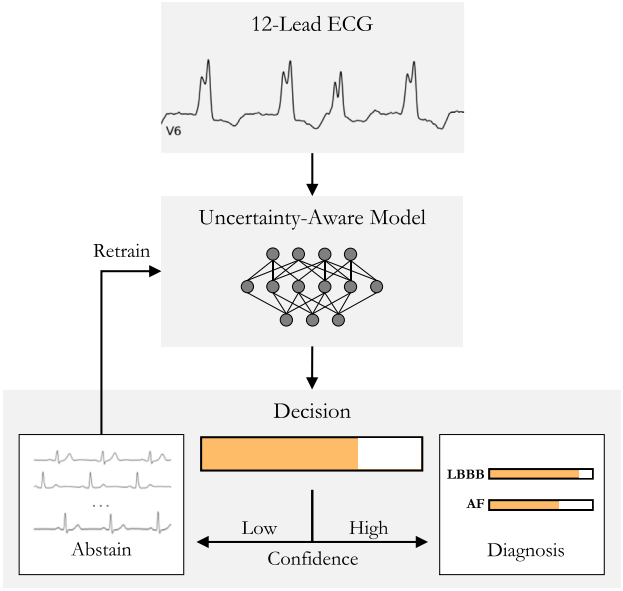

Fig. 1. Workflow for a clinical decision support system (CDSS) that includes uncertainty quantification techniques. The CDSS generates a prediction or diagnosis and estimates the associated uncertainty. The incorporation of UQ allows clinicians to make more informed decisions. Rejected predictions are used to retrain the model using an active learning workflow, improving the overall accuracy and reliability of the CDSS.

applications, UQ for multi-label classification remains underexplored and uses rudimentary techniques [6].

Previous research in the field of UQ has introduced numerous techniques, each tailored to specific evaluation tasks. However, the absence of a standardized approach for assessing uncertainty estimates has created challenges in comparing and selecting the most suitable techniques for different applications. With this work, we aim to fill this gap for automatic diagnosis of ECG. In particular, we establish a comparison of various UQ methods within a multi-label classification setting, using ECG analysis as our chosen domain. The motivation for focusing on ECG classification stems from the availability of large publicly accessible multi-label datasets, which enables us to address both UQ and multi-label challenges. To assess the robustness of uncertainty quantification measures, we evaluate the UQ methods on internal, external, and OOD validation sets. Additionally, we consider the calibration of uncertainty estimates, which is crucial in evaluating the reliability of uncertainty estimations. The calibration analysis allows us to define reliable threshold-based approaches to reject samples with high uncertainty and facilitating the integration of AI based methods into clinical practice.

In this sense, besides the comparison of UQ methods, we include in our work a clinical simulation scenario to assess the benefit of integrating AI uncertainty estimation methods into the practice of cardiology, as illustrated in Fig. 1. The system is based on an uncertainty aware AI model, trained to detect cardiac pathologies based on 12-lead ECG signals. In addition to the classification of cardiac pathologies, the model provides its overall confidence in predicting a given sample which is used to abstain from providing a diagnosis when there is a large amount of uncertainty. In the case of a prediction with low uncertainty, an independent confidence score is provided for each predicted diagnosis. With this ability, additional human expertise can be sought on those rejected samples that later can be used to retrain the model, improving its performance capabilities. Continuous training after a model is deployed is highly important since the environment continuously changes, and concept drifts are likely to occur. In this scenario, and due to the cost associated with data labeling, uncertainty estimation plays an important role in selecting the most informative samples to be labeled.

Contributions. We present a comprehensive comparison of UQ methods in a multi-label setting, focusing on ECG classification scenarios. Our evaluation of UQ methods across various validation scenarios highlights the importance of external validation and its influence on performance, the quality of uncertainty estimates, and calibration. Furthermore, we provide empirical evidence that incorporating UQ throughout the machine learning pipeline brings advantages in classification with a rejection option, dataset shift detection, and active learning. These contributions resulted from a research path that covered the following research questions:

- RQ1: Is the performance of internal validation consistently reproduced on external validation?
- RQ2: How does external validation affect the calibration of models' predictions?
- RQ3: How reliable are uncertainty methods in a multi-label setting under different validation strategies?
- RQ4: What is the impact of using sample rejection on ECG classification performance?
- RQ5: Are uncertainty measures suitable as selection criteria for active learning?

The rest of this paper is organized as follows: Section 2.1 introduces the background for ECG classification, and Section 2.2 presents the background of uncertainty estimation methods along with relevant related work. Section 4 provides a description of the methods used, while Section 5 presents the experimental results. Sections 6 and 7 discuss our findings and present the final remarks and conclusions.

2. Related work

2.1. ECG classification

Over the past decade, the automatic interpretation of ECG records has been widely investigated [7]. Automated classification pipelines have been proposed for classifying individual heartbeats [8–10] and longer intervals containing multiple heartbeats [11–13]. While traditional ML models have been successful in classifying some medical conditions [14], Deep Learning (DL) methods have gained increasing attention in recent years, motivated by their superior performance without requiring significant effort in feature engineering [15].

Hannun et al. [15] proposed a DNN model using the MIT-BIH Arrhythmia Database [16], consisting of single-lead ECG signals, for diagnosing 12 rhythm classes. Their model demonstrated superior performance compared to cardiologists. Ullah et al. [17] introduced a different approach that transforms 1D ECG signals into spectral images through Fourier transformation to classify cardiac pathologies using the same database (MIT-BIH). Although promising results were obtained using single-lead ECG, in realistic clinical settings, the standard technique is 12-lead ECG, which provides more valuable information than a single lead. In this context, He et al. [18] proposed a DL model based on a residual Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) layers to classify 9 classes using a 12-lead ECG and the CPSC dataset [19]. Chen et al. [13] also utilized the CPSC dataset and proposed an artificial neural network that combined CNN, Recurrent Neural Network (RNN), and attention mechanism layers, winning first place in the 2018 China Physiological Signal Challenge. Similarly, using the CPSC dataset, Zhang et al. [20] proposed a CNN model for classifying the 9 cardiac arrhythmias and compared their approach with baseline models employing different architectures such as LSTM, Time Incremental CNN, and Inception. The authors demonstrated that their approach achieved better results than the tested baseline architectures. Strodthoff et al. [21], evaluated multiple algorithms (CNN with feed-forward, ResNet, and Inception architectures, as well as RNN with LSTM and Gated Recurrent Units (GRUs)) using the PTB-XL dataset [22] to classify 9 ECG classes. The authors found that

ResNet and Inception-based architectures achieved the best results. In a recent study, Duong et al. [23] proposed a practical solution based on graph neural networks and showed that their approach had advantages in both performance and timing efficiency over other state-of-the-art baselines when compared using the MIT-BIH and PTB-XL datasets.

Despite the promising performance achieved by DL models, they are prone to a generalization gap, which refers to the difference between a model's performance on training data and its performance on unseen data drawn from the same distribution. DL models often fail to perform adequately on external test sets sampled from different distributions (cf. the concept of external validation [24] or out-of-distribution validation [25]). In this context, Zhu et al. [26] conducted a more extensive study in which they trained a CNN model capable of diagnosing 20 cardiac abnormalities on a private dataset and tested it on a public external dataset. The authors demonstrated the generalization capabilities of their model and showed that it outperformed trained physicians in ECG interpretation. In the study of Kent et al. [27], instead of testing the generalization capabilities using different datasets, the authors explored the generalization of DL models by testing different sampling frequencies and durations from the PTB-XL dataset. They concluded that the models were robust to changes in sampling frequency but not in duration. In a recent study, Rawi et al. [28] evaluated the effectiveness of different DL architectures using three independent datasets. Their results showed that MobileNet and AlexNet models outperformed the other models (Inception, LeNet, VGG16, and ResNet50). However, they did not perform external validation.

2.2. Uncertainty estimation methods

Uncertainty is classified in different ways by different research communities. However, in the ML and statistics literature, one usually distinguishes between two fundamental sources and types of uncertainty, namely *aleatoric* and *epistemic* uncertainty [1]. *Aleatoric uncertainty* refers to the notion of randomness arising from the data's complexity, multi-modality, and noise. *Aleatoric uncertainty*, also known as *data uncertainty*, cannot be reduced or entirely eliminated because it is a property of the underlying distribution that generated the data rather than a property of the model. On the other hand, *epistemic uncertainty* represents the uncertainty caused by a lack of knowledge of the underlying process being modeled, either due to the uncertainty associated with the model or the lack of data. In principle, this uncertainty can be reduced by providing more knowledge, i.e., extending the training data, better modeling, or better data analysis.

While different types of uncertainty should be measured differently, this distinction in ML has only recently gained attention [1]. For instance, the ability to separately quantify uncertainty has been utilized in active learning as a selection criterion for uncertainty sampling [29–31]. In the medical domain, this distinction was emphasized in the work of Senge et al. [32], where the authors demonstrated the usefulness of their approach in the context of medical decision-making. Various approaches for uncertainty quantification have been applied in different medical applications, such as skin cancer detection [33,34], COVID-19 detection [35,36], cancer image detection [37,38], and others.

In recent years, numerous approaches have been developed to equip DNN with the ability to incorporate uncertainty, due to neural networks' limited awareness of their own confidence [39–42]. Bayesian Neural Network (BNN) have been extensively used and depending on how the posterior is inferred, they can be classified as *Variational Inference* (VI), *sampling approaches* or *Laplace approximation* [43]. The Bayes-by-Backprop [44] is an example of a widely used algorithm in the variational inference literature. Another important example is Monte Carlo (MC) Dropout, introduced by Gal et al. [45], which approximates the posterior with a product of Bernoulli distributions. This method has been applied in various studies in literature [46–48], and different extensions have been built on top of it, such as the drop connect method [49], which has been found to be more robust in uncertainty representation [43,50]. On the other hand, *sampling approaches* has the advantage of not being restricted by the type of distribution and have been studied in the literature, including popular algorithms such as particle filtering, rejection sampling, importance sampling, and Markov Chain Monte Carlo sampling (MCMC) [51,52]. *Laplace approximation* was first proposed by Denker, and LeCun [53] and can be applied as a post-hoc method to already trained neural networks. In the literature, recent studies include the work from Kristiadi et al. [54] or Deng et al. [55].

Instead of stochastic approximations, another common approach to approximate Bayesian methods is through ensembles [56,57]. In particular, one popular approach was introduced by Lakshminarayanan et al. [56] where the same network is trained M independently times using different parameter initialization on the whole dataset. Osband et al. [58] and He et al. [59] proposed to train an ensemble by perturbing the loss function of each model with a random but fixed additive prior function. Dwaracherla et al. [60]took advantage of both prior functions and bootstrapping to improve uncertainty estimations.

Single deterministic methods are also quite common in the literature of deep learning [61–64], where a single forward pass generates uncertainty estimation either derived by using additional (external) methods or directly predicted by the network. In the works from Malinin et al. [61] or Sensoy et al. [65], the proposed neural networks were explicitly modeled and trained to quantify both aleatoric and epistemic uncertainties. In these approaches, along with the uncertainty quantification, the training procedure and network's predictions are affected. On the other hand, some studies argue that uncertainty quantification and prediction tasks should be two separate tasks for uncertainty quantification to be unbiased [66]. In this context, Raghu et al. [66] and Ramalho et al. [67], trained one neural network for the prediction task and another for the uncertainty estimation. Other approaches include the use of gradient metrics for uncertainty quantification for OOD detection [68]. Additionally, some more popular approaches in this area include isolation forests [69], auto-encoders [70], and local outlier factor [71]. More in the realm of anomaly or outlier detection, we highlight a few representatives works, such as the Maximum Logit score [72], Mahalanobis distance-based confidence score [73] and energy [74] or joint energy for multi-label setting [6]. Although these approaches are not developed to quantify uncertainty explicitly, they can be seen as a measure of knowledge uncertainty.

Recent works have provided comprehensive reviews on the topic of uncertainty quantification in the context of deep learning, such as the review of Abdar et al. [75] and the Gawlikowski et al. [43].

2.3. ECG classification under uncertainty quantification

Prior studies in ECG classification have often overlooked the evaluation and management of uncertainty associated with their estimations, focusing primarily on classification performance without considering practical implementation in real-world applications. Hong et al. conducted a systematic review of the PhysioNet/CinC Challenge 2020 [76], highlighting the importance of handling unknown classes and interpretability for real-world implementation. Surprisingly, none of the top 10 methods in the Challenge 2020 addressed these critical topics.

While research on UQ for ECG classification remains limited, some recent works have addressed this area, and are summarized in Table 1. Belen et al. [77] employed a variational encoder network to classify atrial fibrillation using the MITBIH Atrial Fibrillation database. Their method used KL Divergence as a loss function and estimated uncertainty by running the input through the network multiple times and computing the standard deviation of softmax probabilities. Vranken et al. [78] explored various uncertainty estimation methods, including Monte Carlo dropout, variational inference, ensemble, and snapshot ensemble. They evaluated the quality of uncertainty estimations using rank-based metrics, calibration evaluation, and OOD detection. Their

Table 1
A summary of related studies on ECG classification using uncertainty quantification measures.

Study	Data	Labels	External Validation	OOD	Calibration
Belen et al. [77] (2020)	MITBIH AF	Single	No	No	No
Vranken et al. [78] (2021)	UMCU-Triage UMCU-Diagnose CPSC2018	Single	No	Yes	Yes
Asseri et al. [79] (2021)	MITBIH ARR INCART BIDMC	Single	No	No	Yes
Elul et al. [80] (2021)	MITBIH NSR Long-Term AF MITBIH ARR MITBIH AF THEW CinC 2017	Multi	Yes	Yes	No
Zhang et al. [81] (2022)	CPSC2018	Single	No	No	No
Jahmunah et al. [82] (2023)	PTB-XL	Single	No	No	No
Park et al. [83] (2023)	MITBIH ARR CinC 2017 INCART	Single	No	No	No

results showed that variational inference with Bayesian decomposition and ensemble with auxiliary output outperformed other methods in terms of ranking and calibration across datasets and in both in-distribution and OOD settings. Aseeri et al. [79] developed a gated recurrent neural network trained using three types of datasets and estimated uncertainty using Monte Carlo dropout and deep ensemble methods. They also evaluated the uncertainty calibration of these methods and demonstrated that their proposed network achieved comparable results with state-of-the-art methods while having a strong capability of rejecting low-confidence examples. Elul et al. [80] presented a comprehensive study on integrating AI into clinical practice, emphasizing the importance of uncertainty estimation for handling OOD examples or multilabel diagnosis. They developed a DL model consisting of 10 binary classifiers for each trained ECG pathology, enabling the model to output any combination of known rhythms and handle unknown classes when the model outputs a negative prediction for every binary class. They employed the Monte Carlo dropout method to assess the confidence in predictions. Zhang et al. [81] employed a Bayesian neural network with Monte Carlo dropout for arrhythmia classification with a rejection option. They computed total uncertainty using an entropy-based decomposition of data and model uncertainty and explored different uncertainty thresholds to improve classification performance by rejecting high uncertainty samples. Jahmunah et al. [82] trained a Dirichlet DenseNet with reverse KL divergence to compute predictive entropy for model uncertainty in a multi-class classification task. The authors argue that their approach is faster and computationally lightweight compared to previous uncertainty quantification methods. Additionally, they included noisy ECG in their analysis. Recently, Park et al. [83] proposed a self-attention-based LSTM-FCN deep learning architecture using a deep ensemble approach to quantify uncertainty. Their results achieved state-of-the-art performance, showing that epistemic uncertainty is reliable for classifying the six arrhythmia types.

Even though some multi-label datasets were used in the previously presented studies, all of them employed a single-label classification approach, except for Elul et al. [80]. To the best of our knowledge, Elul et al.'s work [80] is the only one that applied an UQ method under the multi-label approach in ECG classification. While this study offers a comprehensive interpretation of the importance of handling a mixture of classes and demonstrates that their model is prepared to deal with the multi-label setting, no performance evaluation was conducted on multi-label datasets, making it difficult to thoroughly assess the performance of their model in such settings. Additionally, in this study, only the Monte Carlo Dropout method was used as the uncertainty quantification method.

Additionally, some studies focus on calibration metrics, others on OOD detection, and a few on external validation. However, we argue that a good uncertainty quantification measure should comply with all three validation procedures. In this sense, we focus our work on multi-label datasets, evaluating not only internal validation sets but also external sets, OOD, and calibration.

3. Background

As standard notation to introduce uncertainty estimation methods throughout this section, let us consider a standard setting of supervised learning with a finite training dataset, $D = \{(x_i, y_i)\}_i^N \subset \mathcal{X} \times \mathcal{Y}$, with N samples, composed of pairs of input instances x and outcomes y, where $\mathcal{X}$ is an instance space and $\mathcal{Y}$ the set of outcomes that can be associated with an instance. Suppose a hypothesis space $\mathcal{H}$ of probabilistic predictors, where a hypothesis h maps instances x to probability distributions on outcomes y. Note that we use the notation of hypothesis to make the interpretation more general. Nonetheless, in the case of neural networks, a hypothesis h can be interpreted by a weight vector w.

In a classification task, the most straightforward way of quantifying uncertainty is by using the output of the classification task that represents the class probabilities. Therefore, a simple uncertainty measure given by the confidence in a prediction x can be obtained by the probability of the predicted class, or maximum probability, by Eq. (1), as used in studies such as [56,84,85].

$$p(\hat{y}|x) = \max_{y \in \mathcal{Y}} p(y|x) \tag{1}$$

Additionally, the entropy of the predictive posterior modeled by the (Shannon) entropy, is the most well-known measure of uncertainty of a single probability distribution [1]. For discrete class labels is given by Eq. (2):

$$H[p(y|x)] = -\sum_{y \in \mathcal{Y}} p(y|x) \log_2 p(y|x) \tag{2}$$

Both maximum probability and entropy of the predictive posterior distribution can be seen as measures of the total uncertainty in predictions [61]. These measures of uncertainty for probability distributions

Table 2
Overview of multi-label datasets statistics.

Class	CPSC	G12EC	PTB-XL	Total
AF	1,220	568	1,514	3,302
I-AVB	722	766	795	2,283
LBBB	235	231	536	1,002
NSR	918	1,735	18,058	20,711
PAC	614	636	398	1,648
RBBB	1,857	554	542	2,953
STD	868	38	1,009	1,915
STE	220	134	28	382
VEB	699	41	1,153	1,893
# Labels	7,353	4,703	24,033	36,089
# Recordings	**6,871**	**4,301**	**20,214**	**31,386**

primarily capture the shape of the distribution and, hence, are mostly concerned with the aleatoric part of the overall uncertainty.

For Bayesian approaches, the predictive posterior distribution is approximated by a finite set of Monte Carlo samples or by the individual ensemble members' predictions. In both methods, the predictive variance of the M predictions is a measure of epistemic uncertainty used in various studies [77,78,80] and given by Eq. (3).

$$\sigma[p(y|x)]^2 = \frac{1}{M}\sum_{i=1}^{M}\left(p(y|h_i,x) - \bar{p}\right)^2 \tag{3}$$

where $\bar{p}$ is defined as $\bar{p} = \frac{1}{M}\sum_{i=1}^{M} p(y|h_i,x)$

Additionally, instead of considering the probability variance, one can consider the variation ratios that measure the variability of predictions by computing the fraction of samples with the correct output [86,87]. This heuristic is a measure of the dispersion of the predictions around its mode. For a given instance x, with M output predictions, the variation ratios is calculated as follows,

$$vr(x) = 1 - \frac{\sum_{i=1}^{M}[\![\hat{y}_i = \hat{y}]\!]}{M} \tag{4}$$

where $\hat{y}$ corresponds to the sampled majority class obtained and $[\![\hat{y}_i = \hat{y}]\!]$ is an indicator function that takes the value 1 if the expression is true, and to 0 otherwise.

Furthermore, an explicit attempt at measuring and separating aleatoric and epistemic uncertainty was made by Depeweg et al. [88] who proposed an approach to quantify and separate uncertainties with classical information-theoretic measures of entropy.

In more detail, the *total uncertainty* is measured in terms of the entropy of the predictive posterior distribution approximated by:

$$u_t(x) := -\sum_{y\in\mathcal{Y}}\left(\frac{1}{M}\sum_{i=1}^{M} p(y|h_i,x)\right)\log_2\left(\frac{1}{M}\sum_{i=1}^{M} p(y|h_i,x)\right) \tag{5}$$

The *aleatoric uncertainty* is measured considering the average entropy of each individual prediction in terms of the expectation over the entropies of distributions. The idea is that by fixing a hypothesis h, the *epistemic uncertainty* is essentially removed. Its approximation is given by Eq. (6):

$$u_a(x) := -\frac{1}{M}\sum_{i=1}^{M}\sum_{y\in\mathcal{Y}} p(y|h_i,x)\log_2 p(y|h_i,x) \tag{6}$$

Then, *epistemic uncertainty* is measured in terms of mutual information between hypotheses and outcomes and can be expressed as the difference between the *total uncertainty*, captured by the entropy of expected distribution, and the expected data uncertainty, captured by expected entropy of each individual prediction [61].

$$u_e(x) := u_t(x) - u_a(x) \tag{7}$$

Thus, *epistemic uncertainty* is high if the distribution $p(y|h)$ varies a lot for different hypotheses h with high probability but leading to quite different predictions. This approach was used in different studies, such as [81,89,90].

4. Methods

We conducted an analysis of various uncertainty quantification methods, following the steps illustrated in Fig. 2 and dividing this section accordingly. We begin by discussing the datasets employed and the considerations for data preprocessing. Subsequently, we provide details on the neural network architecture and its variations for uncertainty estimation. The section concludes with an explanation of the validation, which involved three distinct sets (internal, external, and OOD) to assess the methods, the implemented evaluation measures, and particular applications of uncertainty methods (classification with rejection option, dataset shift, and active learning).

4.1. Datasets and preprocessing

For dataset selection, our primary criterion was to choose datasets that included 12-lead ECG data. The PhysioNet/CinC Challenge 2020 provided 12-lead multi-label ECG datasets from four different data sources. However, due to our validation procedure, which involved internal and external validation using different data sources, we could not use the standard 27 classes (out of 111 classes) selected by PhysioNet/CinC Challenge 2020, as not all classes were present in every dataset.

As a result, we decided to utilize only the classes that were common among the datasets. This approach yielded nine classes (NSR, AF, I-AVB, LBBB, RBBB, PAC, VEB, STD, and STE) that were represented across the entire CPSC dataset, enabling us to conduct consistent validation across the different datasets. Furthermore, these nine classes are available in three different data sources. The first source is the China Physiological Signal Challenge 2018 (CPSC) [19], the second is the Physikalisch Technische Bundesanstalt XL (PTB-XL) [22] from Brunswick, Germany, and the third is the Georgia 12-lead ECG Challenge (G12EC) [91] Database, Emory University, Atlanta, Georgia, USA. The three datasets contain data from the 12-leads ECG signals, demographic information (age and gender), and multi-label annotations. The annotations between databases were previously standardized by PhysioNet/CinC Challenge 2020. However, following the evaluation procedure of PhysioNet/CinC Challenge 2020, we relabeled the class CRBBB in G12EC and PTB-XL dataset to RBBB.

For the ECG signals preprocessing, due to the different characteristics of each dataset, the preprocessing included a resampling mechanism to 250 Hz and a truncation of 10 s long. For ECG signals with more than 10 s, the 10 s in the center of the window were selected. The choice of using the 10 s in the window center was done due to the poor signal quality at the beginning and end of some ECG signals. Additionally, each ECG signal was filtered using a 2nd order bandpass Butterworth filter between 1 and 40 Hz and normalized through a z-normalization over the complete dataset.

Table 2 presents a summary of ECG data used per class and dataset. The Table contains information about the number of labels and recordings per dataset.

4.2. Uncertainty methods

As previously mentioned, the main objective of this work is not to explore better model architectures or improve the accuracy of already developed methods. Instead, we aim to understand the potential use of uncertainty measures as a safety mechanism in a practical ECG classification domain. Therefore, as baseline architecture, we decided to use the proposed neural network architecture, which was ranked first in the China Physiological Signal Challenge [13]. The model is a combined architecture of five CNN blocks, followed by a bidirectional gated recurrent unit (GRU), an attention layer, and a finally dense layer. For more details, please refer to Chen et al. [13]. The training was done using the Adam optimizer with a learning rate of 0.001. To counteract class imbalance in the data, the binary focal loss was used

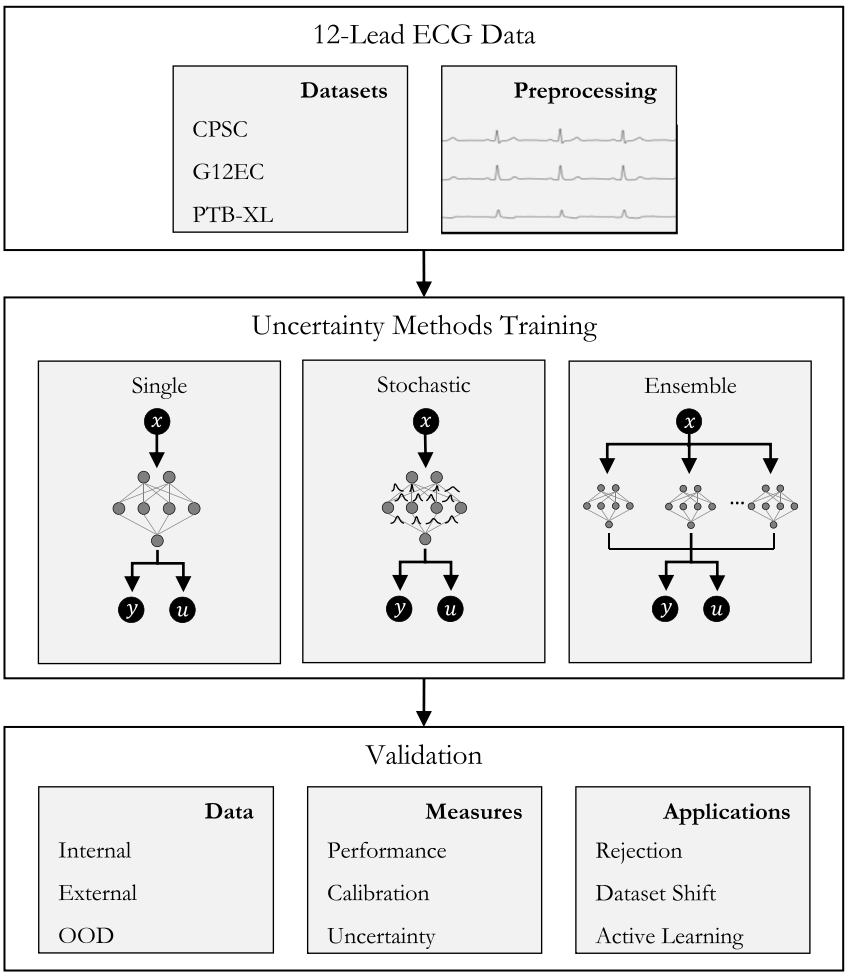

Fig. 2. Overview of the methodology used for uncertainty methods evaluation. Data is based on 12-lead ECG signals, and uncertainty methods are divided into three main categories: Single, Stochastic, and Ensemble. Validation is done using three test sets (internal, external, and OOD) evaluated in terms of performance, calibration, and uncertainty measures, with application on classification with rejection option, dataset shift, and active learning.

as the loss function with the focusing parameter set to 1. The training was performed for 100 epochs using mini-batches of size 64. The best model, which was the one with the smallest loss on the validation set, was selected as the baseline for the uncertainty methods.

For stochastic methods, we implemented both MC Dropout and Laplace approximation to their easy implementation with slight changes in training logic. For MC Dropout, the same trained network was used without retraining since the baseline architecture contains dropout layers. In the testing, dropout layers were kept active, and 15 MC samples were used. For the Laplace approximation, the same trained network was also used since this method can be applied post-hoc to trained neural networks that use an exponential family loss function and piece-wise linear activation functions [4]. Therefore, to approximate the intractable posterior distribution over the parameters of neural networks, we used the implementation of Rewicki et al. [4] developed under the multi-label scenario and publicly available.[1] Similar to MC Dropout, 15 samples were used for testing.

For ensemble methods, the popular approach introduced by Lakshminarayanan et al. [56] where the same network is trained M independently times using different parameter initialization was selected. We will refer to this approach as DeepEnsemble. Additionally, an ensemble based on bootstrapping approach was also trained. Both approaches are composed of 15 individual ensemble members.

Regarding the employed measures to quantify aleatoric and/or epistemic uncertainty, we used different measures depending on whether a single network or a Bayesian approximation was used. For single methods, aleatoric uncertainty estimation was calculated based on both maximum probability and (Shannon) entropy. For epistemic uncertainty, we selected baseline measures developed to improve OOD uncertainty estimation, namely Joint Energy [6], Maximum Logit [72], Isolation Forest [69], Local Outlier Factor [71], and Mahalanobis distance-based confidence score [73]. For Bayesian approximations, we employed maximum probability, predictive variance, variation ratios, and the decomposition of entropy-based measures into aleatoric and epistemic uncertainty. Fig. 3 presents a summary of the uncertainty estimation methods and corresponding uncertainty measures applied on top of it.

It is important to note that we considered independence between labels to calculate uncertainty measures that are directly dependent on class probabilities. For instance, entropy measures are applied in a binary setting scenario for each label, which results in an uncertainty measure per label. To consider the joint uncertainty across labels, we summed the measure of label uncertainties.

[1] https://github.com/ferewi/tf-laplace

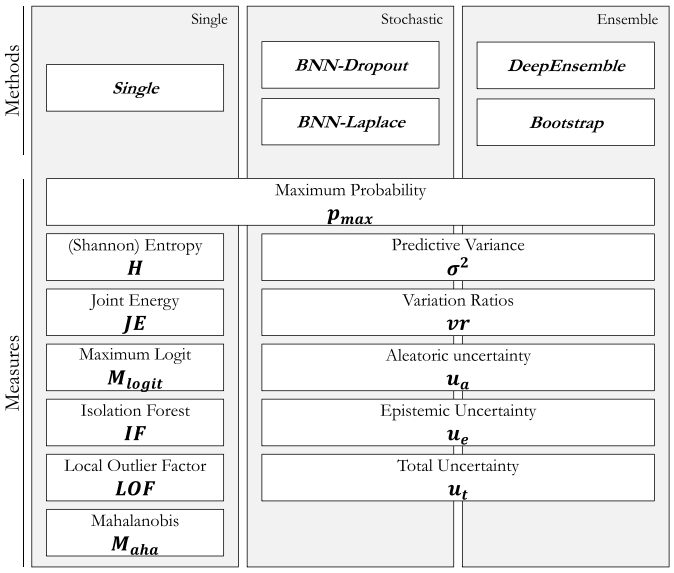

Fig. 3. Uncertainty methods and corresponding uncertainty measures selected for this analysis. The acronyms used throughout this work are represented in bold.

4.3. Validation approach

4.3.1. Training, validation, and test sets

To evaluate the generalization capabilities of the trained models, we conducted assessments on three different test sets, focusing on classification performance, calibration, and uncertainty measures' quality. For model training and internal validation, we used the CPSC dataset, employing an 80–10%–10% train–validation–test split. To ensure an equal distribution of class labels, gender, and age information in each set, we used these criteria as splitting factors. For external validation sets, we used G12EC and PTB-XL datasets.

Additionally, two OOD datasets were also considered for uncertainty quantification evaluation. Since we are not using all datasets' available classes for models' training, we selected a group of unknown classes as OOD. For this purpose, the hierarchical organization in terms of coarse superclasses and subclasses for the diagnostic labels provided by the PTB-XL dataset [22] was used. To reduce the similarity between the diagnostic labels used, we selected the Myocardial Infraction (MI) superclass and the Hypertrophy (HYP) superclass as OOD datasets. As the heterogeneous mixture of known and unknown classes can be presented in this set of labels, we removed all records that contain known classes mixed with these sets of unknown classes to ensure that OOD dataset contained only unknown classes.

Thus, the following test sets were used for evaluation purposes:

- IN (CPSC): Test set used for internal validation, i.e., an independent test from the same data source as the training set. This set contains a total of 687 recordings with the same proportion of class labels identified in Table 2;
- EXT (G12EC): The entire dataset from the G12EC dataset was used for external validation, containing a total of 4301 recordings;
- EXT (PTB-XL): The entire dataset from PTB-XL dataset was used for external validation, containing a total of 20,214 recordings;
- OOD-MI: OOD dataset containing IMI, AMI, LMI, and PMI diagnostic labels from PTB-XL dataset, totaling 2214 records.
- OOD-HYP: OOD dataset containing LVH, LAO/LAE, RVH, RAO/RAE, and SEHYP diagnostic labels from the PTB-XL dataset, totaling 1553 records.

The provided abbreviations will refer to each test set during the experimental analysis.

4.3.2. Evaluation measures

The empirical evaluation of methods for quantifying uncertainty is a non-trivial problem due to the lack of ground truth uncertainty information. A common approach for indirectly evaluating the predicted uncertainty measures is by accessing their usefulness to improve classification performance. In this sense, ranking-based methods can be used to evaluate the uncertainty measures' capability of ordering predictions based on their own uncertainty estimation. The idea is to evaluate how the classification performance varies as a function of the percentage of rejections. If a measure is able to quantify its own uncertainty well, the classification performance should improve with an increasing percentage of rejections. This approach can only be directly applied to compare different uncertainty measures using the same predictive model since the classification performance curves depend not only on the uncertainty ordering but also on the predictive model performance. Although the applied uncertainty methods are based on the same model architecture, due to the specific details of each approach, the classification performance slightly varies between methods. Thus, for a fair comparison between uncertainty measures, we will use the Area Under the Confidence-Oracle (AUCO) error [92] that computes the area between the theoretically perfect ordering and the ordering made by each uncertainty measure.

The oracle confidence curve represents the best possible ordering of predictions by their confidence, with the true error imposing the ordering. The AUCO value is calculated as the area under the curve representing the difference between the given uncertainty estimation and the oracle confidence curve. Smaller values of AUCO indicate that the given uncertainty estimation is closer to the oracle confidence curve and therefore is a better predictor of uncertainty. The formula of AUCO is as follows:

$$AUCO = \int_0^1 \left(conf_r^u - conf_r^o\right) dr \tag{8}$$

where $conf^u$ is the confidence curve for a given uncertainty estimation, $conf^o$ is the oracle confidence curve and r is the fraction of rejections. Thus, the integration is performed over the range of confidence values.

For the special case of OOD datasets, we used the Area Under the Receiver Operating Characteristic (AUROC) curve metric, which is commonly applied in most recent studies and is a threshold-independent performance method for evaluating OOD detection methods. The AUROC can be interpreted as the probability that a positive example is assigned a higher detection score than a negative example. Consequently, a random positive example detector corresponds to a 50% AUROC, and a perfect detector corresponds to an AUROC score of 100% [84].

In our study, we also employed threshold-dependent measures, adopting the approach used in recent studies for evaluating uncertainty estimations based on the concept of binary confusion matrix [35,93]. In this context, predictions are classified as correct or incorrect, and, depending on a threshold, predictions are also classified as certain or uncertain. As a result, four combinations are identified: (i) True Certainty (TC): correct and certain; (ii) True Uncertainty (TU): incorrect and uncertain; (iii) False Uncertainty (FU): correct and uncertain; and (iv) False Certainty (FC): incorrect and certain. Based on these combinations, we calculated Uncertainty Accuracy (UAcc), Uncertainty Sensitivity (USens), Uncertainty Specificity (USpec), and Uncertainty Precision (UPrec) using the following formulas:

$$UAcc = \frac{TU + TC}{TU + TC + FU + FC} \tag{9}$$

$$USen = \frac{TU}{TU + FC} \tag{10}$$

$$USpec = \frac{TC}{TC + FU} \tag{11}$$

$$UPrec = \frac{TU}{TU + FU} \tag{12}$$

Table 3
Global performance of uncertainty methods in internal (IN) and external (EXT) validation sets. The highest scores are represented in bold.

Model	IN (CPSC)		EXT (G12EC)		EXT (PTB-XL)	
	AUROC	F1-Score	AUROC	F1-Score	AUROC	F1-Score
Single Network	0.896	0.826	0.830	0.715	**0.734**	**0.567**
BNN-Dropout	0.890	0.833	0.811	0.699	0.700	0.516
BNN-Laplace	0.896	0.830	0.830	0.715	**0.735**	**0.568**
DeepEnsemble	**0.903**	**0.856**	**0.831**	**0.736**	0.724	0.559
Bootstrap	**0.903**	0.851	0.821	0.718	0.717	0.548

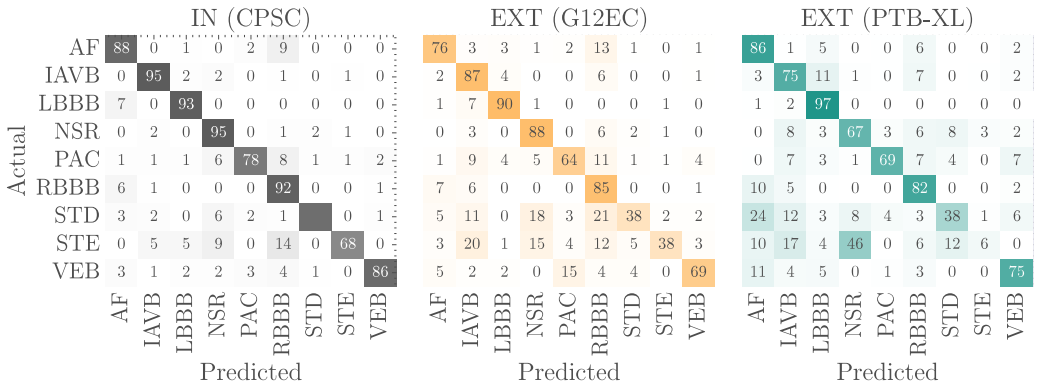

Fig. 4. Binary multi-class multi-label confusion matrices for DeepEnsemble method in internal (IN) and external (EXT) validation sets.

The ranking-based methods are an essential measure to compare different uncertainty estimations. However, they do not consider the actual values expressed by uncertainty. In this sense, calibration measures can be used to assess if observed empirical frequencies are consistent with outputting probability distributions. Thus, to measure calibration, a reliability diagram and the Expected Calibration Error (ECE) were used as calibration measures. Reliability diagrams depict accuracy on the y-axis and average confidence on the x-axis. A perfectly calibrated model outputs probabilities that match up with the accuracy, yielding a diagonal line, where confidence is equal to accuracy. Additionally, the ECE was computed to measure the difference in expectation between confidence and accuracy.

4.3.3. Applications

For the classification with rejection option, the uncertainty measures were used as a measure for rejection. The rejection threshold was obtained using the training data, where a given uncertainty training percentile is selected to reject samples on test data. Thus, for each test sample, the uncertainty is computed and compared with the defined threshold. If the obtained uncertainty value is greater than the threshold the sample is rejected and no prediction is made. On the other hand, if the uncertainty is lower than the threshold the model accepts the prediction, and a confidence level is also returned.

For dataset shift validation, the statistical divergence measure, Wasserstein distance [94], was applied to measure dataset similarity between internal and external datasets. The Wasserstein-1 version of Wasserstein distance [94] was used and is given by:

$$W_1(X,Y) = \inf_{\pi \in \Gamma(X,Y)} \int_{\mathbb{R}\times\mathbb{R}} |x-y| \mathrm{d}\pi(x,y), \tag{13}$$

where $\Gamma(X,Y)$ is the set of distributions whose marginals are X and Y on the first and second factors, respectively. The variables x and y are samples from each distribution $\pi(x,y)$ from the set. Intuitively, the distance is given by the optimal cost of moving a distribution until it overlaps with the other. In our experiments, x and y are the feature representations of subsets of the train and test data; thus, W_1 represents the cost of mapping the distribution of x into the distribution of y (or vice versa). The similarity measure was computed on the latent feature space, i.e., the embeddings extracted from the neural network, between the training set and each of the test sets.

For active learning validation, the samples are sorted based on their uncertainty values, and the highest n uncertain samples are used for retraining the model. This process is performed using different uncertainty sources and compared to random sampling. The evaluation is based on the improvement of classification performance metrics.

5. Experimental results

The experimental results are organized to address the five research questions previously introduced.

5.1. External validation

RQ1: Is the performance of internal validation consistently reproduced on external validation?

Although all the uncertainty methods share the same deep learning architecture, differences in training or testing procedures between them might affect not only the uncertainty estimation but also the predictive performance. To properly assess these uncertainty methods, we first present the classification performance for each method in internal and external validation.

Table 3 compares the AUROC and F1-score for each method during internal and external validation. The comparison indicates that the DeepEnsemble method performs slightly better than the other methods. However, the performance achieved within the same test set is similar across all methods. Table 3 also reveals a significant drop in performance during external validation, particularly in the PTB-XL dataset.

To analyze the class level performance between datasets, a binary multi-class, multi-label confusion matrix for each dataset was computed using the implementation provided by the PhysioNet/CinC Challenge

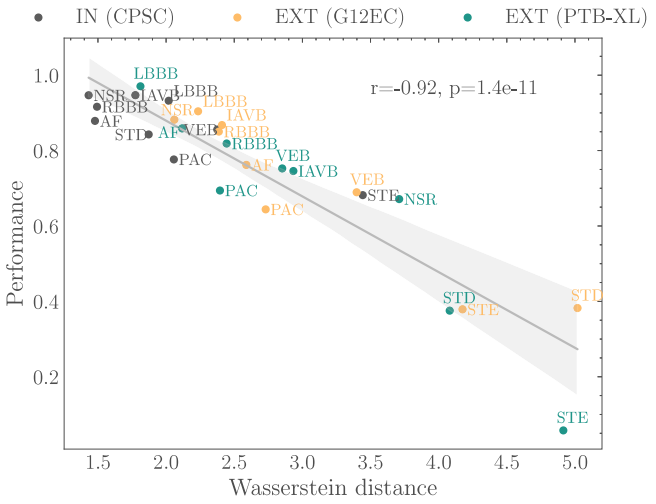

Fig. 5. Class performance drop as a function of Wasserstein distance between training and each represented test set. Each point is annotated with the class name abbreviation and the color represents the dataset. The linear regression is obtained with all datasets and represented in gray. The Pearson correlation coefficient (r) and p-value (p) for testing non-correlation are annotated in the graph area.

2020. As all methods demonstrated comparable performance measures, we only present the confusion matrices for the DeepEnsemble method in Fig. 4. These confusion matrices reveal that in both external datasets, STE and STD diagnoses are accurately recognized. In contrast, the Bundle Branch Blocks (LBBB and RBBB) maintain consistent performance across both internal and external datasets.

The correlation between data similarity and generalization properties across datasets has been previously identified as a strong indicator that the datasets originate from different distributions. Consequently, information about similarity can offer valuable insights into understanding why a machine learning model exhibits poor performance on an external dataset [95].

Fig. 5 illustrates the correlation between the performance drop and Wasserstein distance using the three datasets. The worst class performances observed in confusion matrices (Fig. 4) also correspond to those with higher Wasserstein distances. The calculated Pearson correlation coefficient ($r = -0.92$) suggests that there is a potential shift (label-dependent) in external datasets, and the Wasserstein distance proves to be useful in detecting it. In addition to the STE and STD classes, the NSR (Normal Sinus Rhythm) class from the PTB-XL dataset also exhibits a higher distance and a significant drop when compared to the same class in the CPSC and G12EC datasets.

Based on this observation, we carried out a thorough examination of the NSR class label and discovered a significant difference in NSR annotations across the three datasets. To align with the annotation of the training dataset, only a subset of the NSR class from the PTB-XL dataset will be utilized for the remainder of the analysis. We refer to this subset as PTB-XL*. A comprehensive explanation and the results obtained can be found in Appendix A.

Table 4 presents the performance results for various combinations of internal and external sets and Fig. 6 the correlation between Wasserstein distance and global model performance of different combinations. All models followed the same training procedure, as detailed in Section 4.2. Independent validation sets were utilized for internal validation, either using the publicly available data partition or an 80–10%–10% train–val–test split, with class labels, gender, and age serving as splitting criteria. Regardless of the combination, internal validation sets consistently achieved a performance higher than 0.80,

Table 4
Performance comparison of different combinations of in internal (IN) and external (EXT) validation sets.

CPSC		G12EC		PTB-XL*	
Validation	F1-score	Validation	F1-score	Validation	F1-score
IN	0.856	EXT	0.736	EXT	0.699
IN	0.807	EXT	0.741	IN	0.891
IN	0.849	IN	0.818	EXT	0.728
EXT	0.722	IN	0.832	IN	0.885
IN	0.826	IN	0.815	IN	0.884

Table 5
Expected Calibration Error (ECE) for internal (IN) and external (EXT) validation sets. The lowest errors are represented in bold.

Model	IN (CPSC)	EXT (G12EC)	EXT (PTB-XL*)
Single Network	0.047	0.121	0.115
BNN-Dropout	0.057	0.043	**0.040**
BNN-Laplace	0.048	0.120	0.115
DeepEnsemble	**0.026**	**0.034**	0.045
Bootstrap	0.045	0.048	0.062

*Subset of PTB-XL with only Normal class.

while external validation sets showed a performance below 0.75. Nevertheless, incorporating additional datasets for training led to enhanced performance on external datasets.

5.2. Calibration

RQ2: How does external validation affect the calibration of models' predictions?

Table 5 shows the ECEs for all uncertainty methods and datasets. Reliability diagrams are shown in Fig. 7. For both measures, 10 bins were used. All uncertainty methods achieved equal or lower ECEs compared to the Single Network, with BNN-Dropout on the internal validation being the only exception. The DeepEnsemble model obtained the lowest ECE in CPSC and G12EC datasets, while BNN-Dropout obtained the lowest ECE in PTB-XL. BNN-Laplace was the least effective uncertainty method, exhibiting similar results to the Single Network.

From the reliability diagrams, we can observe that the Single Network and BNN-Laplace exhibit similar behavior, with their estimates being overconfident across all datasets. Both ensemble methods display similar behavior in all datasets, with the DeepEnsemble appearing to be more robust across the various datasets.

5.3. Uncertainty evaluation

RQ3: How reliable are uncertainty methods in a multi-label setting under different validation strategies?

As an initial illustrative visualization, we present the overall uncertainty of internal, external, and OOD datasets using the DeepEnsemble method in Fig. 8. The uncertainty values were normalized to their maximum theoretical values, ensuring that all uncertainty measures are bounded within the range of 0 and 1. Noticeably, all measures consistently increase the overall uncertainty, regardless of the uncertainty measure employed. Ideally, we would like the increase in uncertainty values to coincide with the reduction in classification performance observed earlier. In other words, we would expect the uncertainty values to remain as consistent as possible under different external validations (as well as on OOD data) to indicate that models are uncertain when predicting a given input.

In addition to assessing the uncertainty between datasets, it is also possible to statistically evaluate the relationship between the distributions of uncertainty values for correctly and incorrectly classified samples. In a multi-label setting, we can consider two scenarios: (1) a label

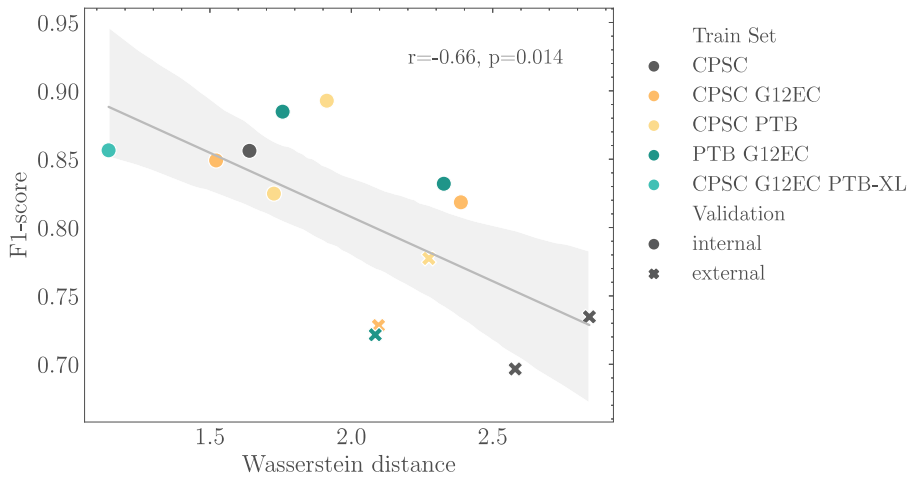

Fig. 6. Correlation between Wasserstein distance and F1-score using different datasets combinations for internal and external datasets. The Pearson correlation coefficient (r) and p-value (p) for testing non-correlation are annotated in the graph area.

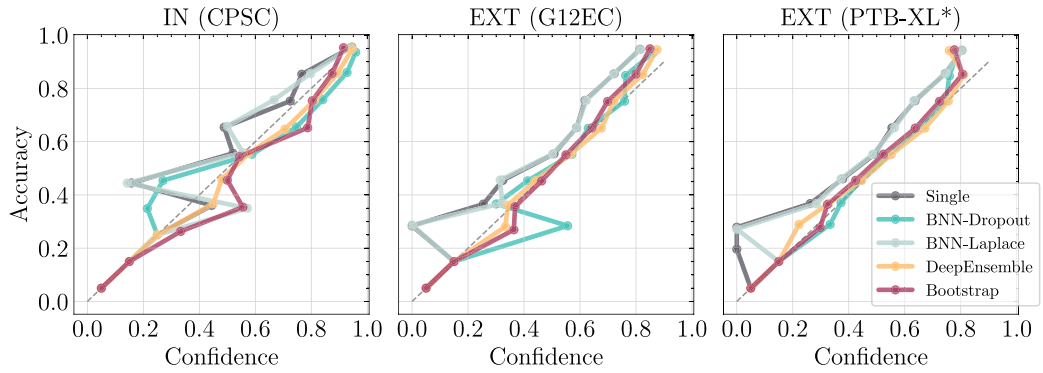

Fig. 7. Reliability diagrams for internal (IN) and external (EXT) validation sets. The diagonal dashed line represents the perfect calibration.

dependence scenario, in which the entire label combination is treated as either correct or incorrect, and (2) a label independence scenario, in which each class is addressed as a separate binary classification problem. Fig. 9 illustrates the distributions of these two scenarios, using DeepEnsemble as an example. The uncertainty distributions for the label independence scenario display a more pronounced distinction between correctly and incorrectly classified samples compared to the label dependence scenario. When applying the non-parametric Mann–Whitney U statistical test [96] for unpaired groups, both approaches were found to be statistically significant at a .05 significance level (even with Benjamini–Hochberg p-values correction [97]). In addition to evaluating the practical significance using Cohen's d effect size [98], we also computed the effect sizes for each method. The label independence approach was found to yield larger effect sizes. The complete results can be found in the Appendix B.

To enable a fair comparison among all uncertainty methods and their corresponding measures, the AUCO metric was calculated for both internal and external test sets. Smaller AUCO values indicate better performance. Fig. 10 presents the results, with the same color representing the same uncertainty measure across different methods. In addition to the differences in performance between internal and external validation, Fig. 10 also clearly illustrates that uncertainty estimation measures are affected in external validation. In general, ensemble-based uncertainty measures appear more robust in preserving the correct uncertainty ordering compared to other methods, and epistemic uncertainty measures outperform aleatoric uncertainty measures in external validation. In internal validation, maximum probability (p_{max}) achieved the lowest AUCO across all methods. This is somewhat expected, as the internal dataset exhibits low epistemic uncertainty, unlike the external validation datasets. It is also worth noting that OOD detection measures (JE, M_{Logit}, IF, LOF, M_{aha}) do not perform as well as other methods in this rank-based analysis. Although there is a close relationship between uncertainty estimation and OOD detection, ordering uncertainty values and detecting OOD are not the same problem, which might explain the lower performance of these methods.

In regard to OOD detection, the two superclass sets (MI and HYP) from the PTB-XL dataset, consisting solely of unknown classes, were employed as OOD samples. The AUROC was calculated with the OOD samples as positive instances and the internal CPSC test samples as negative instances. The obtained results are presented in Table 6. In line with the previous analysis, ensemble methods surpassed other approaches in terms of AUROC. For the MI set, total uncertainty u_t (or entropy H for Single methods) achieved the highest AUROC across all methods. In contrast, for the HYP set, epistemic uncertainty u_e yielded higher AUROC values for ensemble methods and BNN-Laplace. As for the methods specifically designed for OOD detection (JE, IF, LOF), their performance in distinguishing OOD samples was surprisingly poor.

While the OOD problem typically refers to anomaly and/or outlier detection, where OOD samples come from entirely different distributions, in our setting, the OOD samples consist of classes from the same datasets and are thus more related to novelty detection associated with the Open Set Recognition (OSR) scenario. Although OSR is similar to OOD detection, it is likely more challenging to address, as the statistics of the new classes often resemble those of existing classes within the dataset [99].

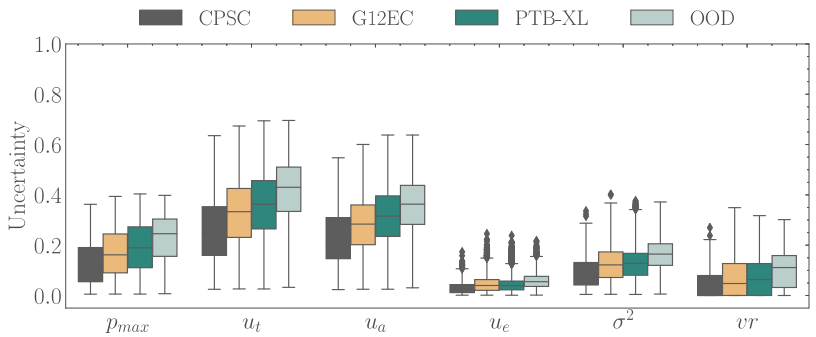

Fig. 8. DeepEnsemble uncertainty measures distributions for internal (CPSC), external (G12EC and PTB-XL), and Out-of-Distribution (OOD) datasets. Uncertainty measures are normalized with their maximum theoretical value, where 1 represents the maximum possible uncertainty. The OOD label contains data from both OOD-MI and OOD-HYP sets.

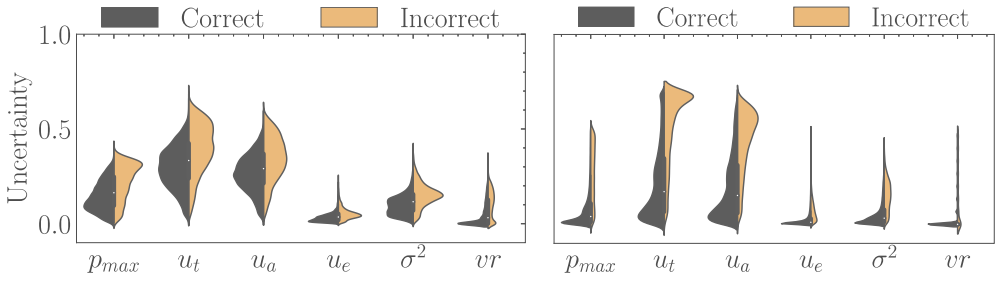

Fig. 9. Comparison of uncertainty value distributions for correctly and incorrectly classified samples using DeepEnsemble model. (a) label dependence approach, where the entire label combination is considered as either correct or incorrect, and (b) label independence approach, where each class is treated as a separate binary classification problem.

Table 6
OOD detection performance comparison using all uncertainty methods and measures. OOD datasets are composed of two superclasses sets (MI and HYP) with only unknown classes from the PTB-XL dataset.

Model	Uncertainty	AUROC	
		OOD-MI	OOD-HYP
Single Network	p_{max}	0.758	0.702
	H	0.767	0.703
	JE	0.749	0.715
	M_{Logit}	0.761	0.716
	IF	0.524	0.617
	LOF	0.502	0.633
	M_{aha}	0.614	0.569
BNN-Dropout	p_{max}	0.763	0.717
	vr	0.671	0.647
	σ^2	0.645	0.648
	u_a	0.773	0.719
	u_e	0.450	0.529
	u_t	0.767	0.717
BNN-Laplace	p_{max}	0.759	0.704
	vr	0.574	0.575
	σ^2	0.742	0.727
	u_a	0.767	0.704
	u_e	0.710	0.730
	u_t	0.767	0.704
DeepEnsemble	p_{max}	0.781	0.752
	vr	0.721	0.736
	σ^2	0.778	0.790
	u_a	0.787	0.740
	u_e	0.751	0.787
	u_t	**0.794**	0.764
Bootstrap	p_{max}	0.775	0.757
	vr	0.735	0.747
	σ^2	0.783	0.793
	u_a	0.776	0.730
	u_e	0.764	**0.794**
	u_t	0.791	0.767

5.4. Classification with rejection option

RQ4: What is the impact of using sample rejection on ECG classification performance?

In the previous sections, we compared various uncertainty estimation methods using threshold-independent measures. However, to evaluate the benefits of integrating AI uncertainty estimation methods in supporting medical decision-making within cardiology, a confidence threshold must be established. This threshold enables the classifier to abstain in situations with high uncertainty. The selection of a threshold restricts the comparison among methods, as each method may have a varying optimal threshold.

Figs. 11 and 12 depict the predictive uncertainty performance evaluation metrics for the three datasets, using the uncertainty estimation methods while varying the uncertainty threshold. Fig. 11 employs epistemic uncertainty (σ^2) as an uncertainty measure, while Fig. 12 uses aleatoric uncertainty (maximum probability p_{max}) as an uncertainty measure. Although these differences are more pronounced when using epistemic uncertainty, aleatoric uncertainty also presents some disparities between methods. Regardless of the chosen threshold, we can observe from the figures that there is a degradation of performance metrics in the external datasets. For instance, the uncertainty accuracy in internal validation reaches a maximum of over 0.8, while in external datasets, the maximum is approximately 0.70.

Since the proper definition of an uncertainty threshold is beyond the scope of this work, we opted for an analysis based on a given rejection rate obtained in training. Consequently, we defined a 15% rejection rate on the training set and used the corresponding uncertainty value to reject samples on the testing sets. Each threshold is represented by a data point placed on top of each line plot in Fig. 11, emphasizing that the threshold value differs for each method and selecting the same threshold for all methods does not provide a fair comparison between methods.

To simplify the analysis of experimental results for classification with a rejection option, Table 7 presents a summary of the complete

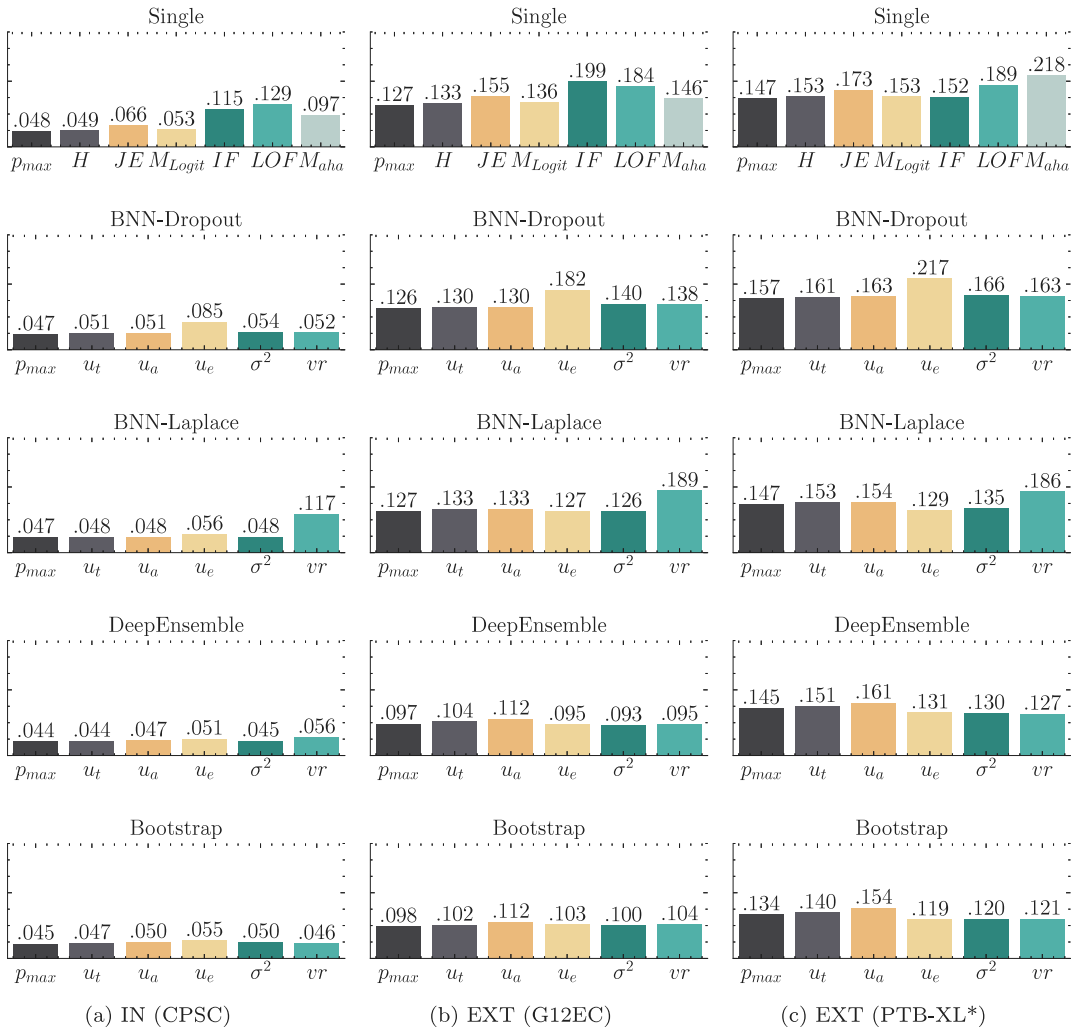

Fig. 10. Ranking performance evaluation using Area Under the Confidence-Oracle (AUCO) for all datasets and uncertainty methods.

Table 7
Performance evaluation metrics for classification with rejection option using DeepEnsemble method. F1-score is presented without rejection (baseline) and with rejection (Non-rejected F1-score). The rejection threshold was set to 15% rejection on the training set. Predictive uncertainty evaluation measures used the same threshold.

Metric	IN (CPSC)	EXT (G12EC)	EXT (PTB-XL*)
F1-score (Baseline)	0.856	0.736	0.699
Non-rejected F1-score	0.915	0.844	0.795
Rejection Rate	0.207	0.385	0.317
Uncertainty Accuracy	0.818	0.722	0.716
Uncertainty Sensibility	0.576	0.661	0.554
Uncertainty Specificity	0.880	0.755	0.823
Uncertainty Precision	0.548	0.595	0.675

performance results with rejection for the best uncertainty method, the DeepEnsemble. The uncertainty rejection measure used was the variance of ensemble members' probabilities, σ^2, as it achieved better performance measures in the previous analysis.

The first observation from Table 7 is that rejecting highly uncertain samples improves classification performance across all datasets. For the same threshold, the rejection rate varies considerably between datasets. As expected, the internal dataset CPSC exhibits a lower rejection rate (similar to the 15% applied in training), but for the external datasets, the rejection rate more than doubles, reaching 0.385 and 0.317 for G12EC and PTB-XL datasets, respectively. This observation aligns with the results obtained so far, in which the external test sets contain more uncertain samples. Applying the same threshold for the OOD datasets results in rejection rates of 0.634 and 0.666 for OOD-MI and OOD-HYP, respectively. While the performance of non-rejected samples can be considered acceptable (at least comparable to the internal validation), more than 30% of OOD samples were not rejected, which might be a substantial proportion of OOD samples. Naturally, lowering the uncertainty threshold would reject more OOD samples. However, this comes at the cost of rejecting more samples from known classes.

Table 7 also highlights acceptable uncertainty accuracy. However, with the selected threshold, all models exhibit higher specificity than sensitivity. This means that if we want to increase sensitivity (while decreasing specificity), the rejection rate will also increase.

5.5. Active learning

RQ5: Are uncertainty measures suitable as selection criteria for active learning?

Apart from classification with a rejection option, an essential procedure after deploying a model in clinical practice is continuous training to respond to changes in the data and prevent models from becoming unreliable and inaccurate. For model retraining, it is necessary to label data that requires expert knowledge. Obtaining large amounts of labeled data can be unfeasible during clinical practice. One possible approach to reduce this effort is to rely on active learning to select

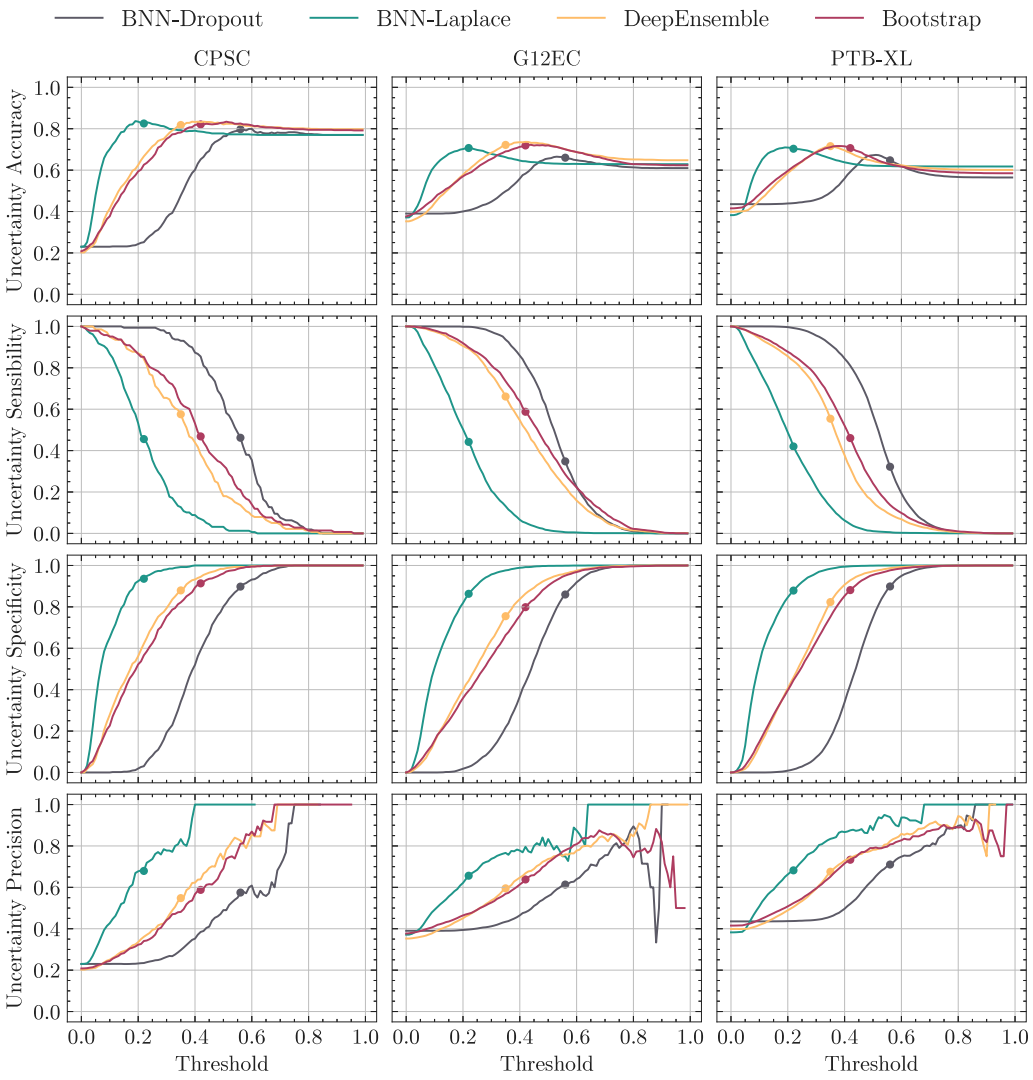

Fig. 11. Uncertainty performance measures for varying threshold values across different datasets using four uncertainty estimation methods with epistemic uncertainty. The chosen threshold for each method is denoted by a data point superimposed on each line plot.

what unlabeled data would be most informative to the model and ask an expert annotator for a label on only these selected samples.

Following this reasoning, we retrained the DeepEnsemble method using data from PTB-XL and G12EC datasets. The retraining procedure consisted of selecting the rejected samples with higher uncertainty from one of the datasets and retraining the model with these new samples. For comparison purposes, we repeated this process 5 times using different uncertainty measures and random sampling. The process consisted of retraining the model using 400 new samples and repeating the process eight more times with a step of 400 new samples, totaling the usage of 3200 new samples at the end of the process. For this analysis, we split the external datasets into train and test sets and present the results always using the test set for a fair comparison. Since PTB-XL has an available 10-fold split provided by PhysioNet, we used the last fold, as proposed by PhysioNet, for the test set and the other folds for training. For the G12EC dataset, since there is no proposed split, we used a 90%–10% train-test split using classes, gender, and sex as group criteria for balanced data splitting. The obtained performance in these test sets was similar to the performance using the entire dataset and represented the first point (0 samples) in the plots of Fig. 13.

Fig. 13 shows the evolution of classification performance with the increased number of samples used to retrain the models. In the first row, data from the G12EC dataset was used to retrain the model, and the in the second row, PTB-XL data was used. The gray background represents the dataset used to retrain the models. Besides the performance evolution within the dataset used for retraining, we also show the classification performance in the other datasets to ensure that the increase in performance in one dataset does not represent a performance degradation in the other datasets. Observing Fig. 13 we note that adding new samples from external datasets does not affect the performance in the internal CPSC dataset. Contrary, adding new samples from one of the external datasets increased not only the performance on that dataset but also the performance in the other external dataset. Comparing the random sampling with the different uncertainty measures, we conclude that every uncertainty measure performs better than using random samples to retrain the model. Even though random sampling also increases the classification performance but at a slower rate. As for the uncertainty measure used to retrain the model, aleatoric, epistemic, and total uncertainty obtained similar results on the G12EC dataset. Otherwise, on the PTB-XL dataset, epistemic uncertainty obtained an higher improvement compared to aleatoric and total uncertainty, with the only exception on the first 400 samples of the PTB-XL dataset.

6. Discussion

This study addresses the importance of uncertainty quantification in multi-label ECG classification to develop a practical approach suitable

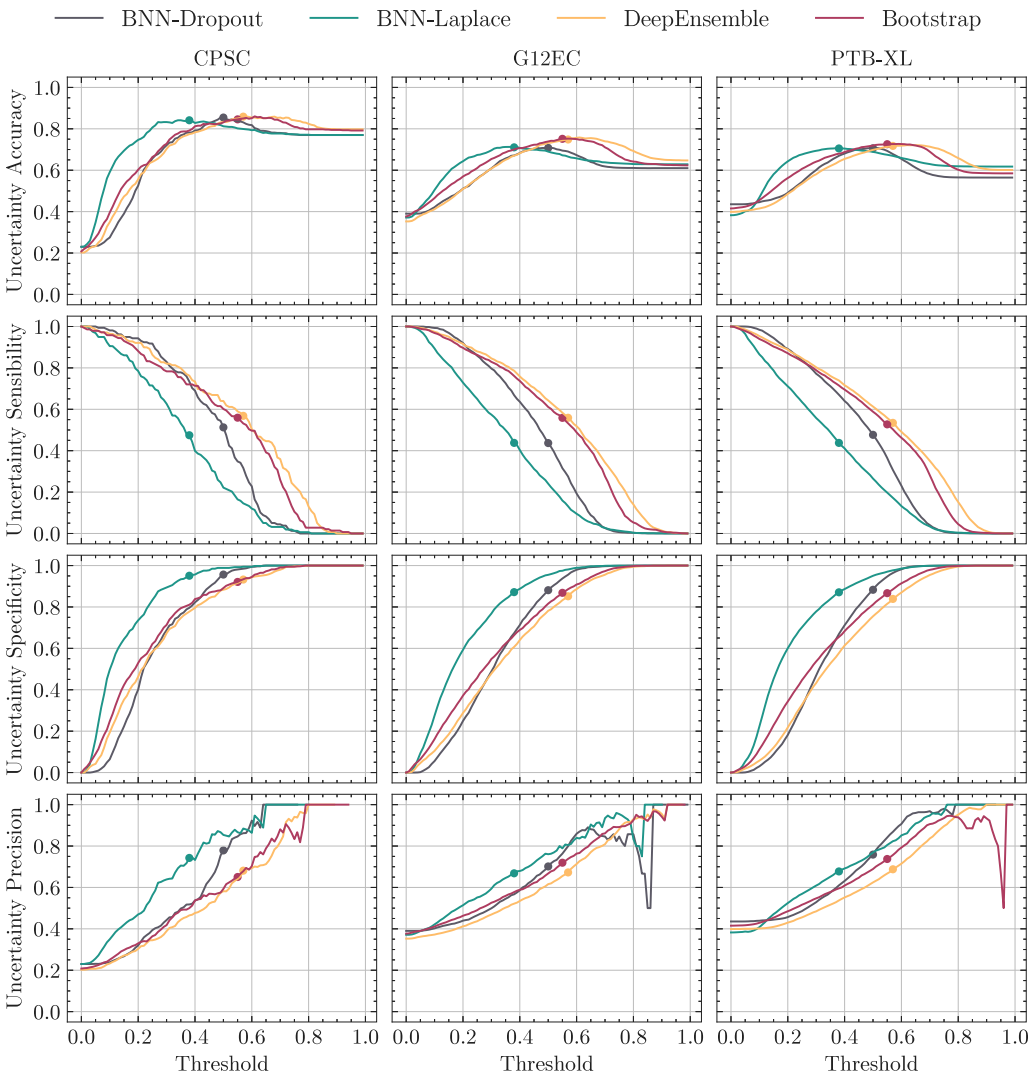

Fig. 12. Uncertainty performance measures for varying threshold values across different datasets using four uncertainty estimation methods with aleatoric uncertainty. The chosen threshold for each method is denoted by a data point superimposed on each line plot.

for implementation in clinical practice. The discussion is organized into four subsections for clarity. We begin by discussing the main conclusions concerning external validation and its connection to dataset shift. Next, we explore uncertainty estimation and calibration results, encompassing internal, external, and OOD validation sets. The subsection on the clinical scenario covers classification with rejection option and active learning experiments, focusing on their practical implementation after deployment.

Finally, we conclude the discussion with a reflection on the limitations of our work and outline potential directions for future research.

6.1. External validation and dataset shift

External validation of machine learning models is becoming increasingly important, particularly in the medical domain. Although it offers more reliable validation compared to internal validation, the results do not necessarily guarantee reliability on their own [95]. Our results revealed that a trained model, which performs well on internal validation (with comparable classification performance to similar studies in the literature [13]), may be significantly affected when validated on an external dataset. Specifically, our findings demonstrated a drop in F1-Score from 0.86 to a range between 0.74 and 0.70 on external validation, depending on the dataset used. Besides being from a completely different source, the external datasets included not only the known classes for the model but also a mixture with unknown classes, i.e., since a multi-label setting is being used, a sample can be labeled with a known and an unknown class. In fact, in the external datasets, 50% of samples include unknown classes, and out of the remaining 50%, only 20% of samples do not belong to the Normal class. As a result, the majority of cardiac pathologies in the external datasets represent a heterogeneous mixture of medical conditions, which can be a major contributing factor to the performance drop. In line with this, we demonstrated a strong correlation ($r = -0.92$) between the drop in class performance and the distance between the train and test sets using the Wasserstein distance.

These findings on external validation align with studies in the literature [95,100], where models trained in one setting (data from the same source) do not generalize well to other external data sources. Additionally, incorporating more data sources into the training scheme improves overall performance on both internal and external data sources. However, it still does not guarantee the same level of performance as with internal datasets.

6.2. Uncertainty and calibration

Although uncertainty quantification does not solve the problem on its own, it plays a crucial role in identifying and mitigating unreliable or inaccurate predictions when dealing with external factors [101].

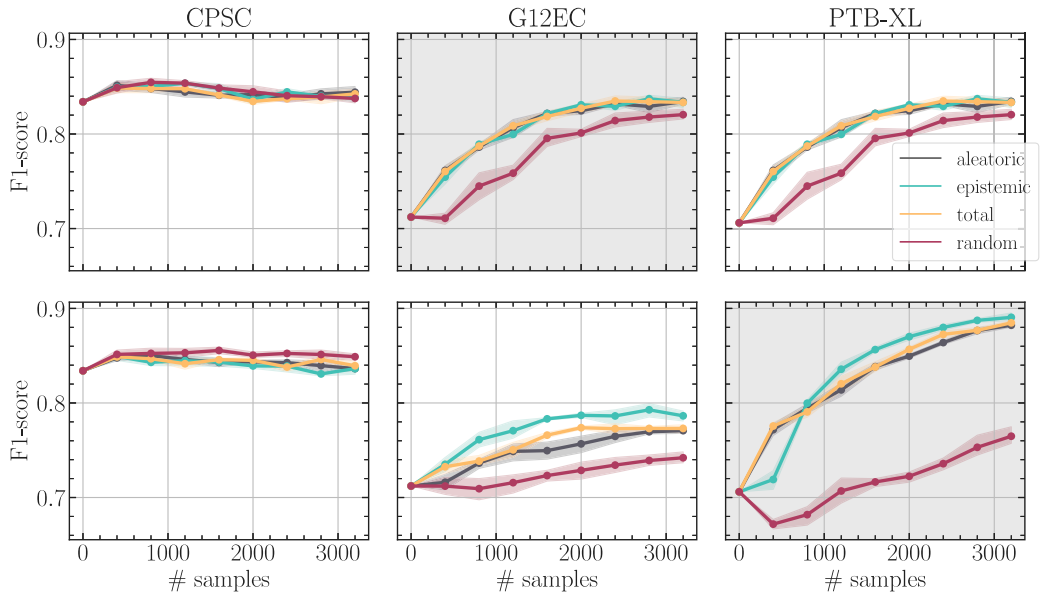

Fig. 13. Classification performance as a function of the number of samples used to retrain the DeepEnsemble. In the upper plots data from the G12EC dataset was used to retrain the model, and the in the lower plots, PTB-XL data was used. The gray background represents the dataset used to retrain the models.

Consequently, we investigate the feasibility of various UQ methods applied to multi-label ECG classification. Our results demonstrated that ensemble-based methods yielded more robust uncertainty estimations compared to single or Bayesian methods. In terms of calibration analysis, MC-Dropout and ensemble methods achieved lower ECE values than the baseline network. Therefore, the uncertainty measures not only provide an assessment of uncertainty but also offer an improved and better-calibrated probability measure.

To the best of our knowledge, no studies in the literature compare uncertainty methods using a multi-label setting in ECG analysis. However, in single-label ECG analysis scenarios, Vranken et al. [78] obtained similar conclusions. In different application modalities such as images, text, and categorical data, Ovadia et al. [101] conducted a comprehensive comparison of uncertainty methods under dataset shift and also reported better results for ensemble-based methods.

Regarding the quality of uncertainty sources, aleatoric uncertainty estimations achieved better results in internal validation, while epistemic uncertainty estimations yielded superior results in external validation in terms of rank-based measures. Concerning OOD detection, ensemble-based methods using epistemic or total uncertainty outperformed other methods, achieving an approximately 0.80 AUROC. Surprisingly, the methods designed for OOD detection, which have shown good results in other studies in the literature [6,72,73], obtained poor results in our ECG classification problem. Although OOD and OSR are similar concepts and OOD is often used in the literature to represent a broad view of anomaly, outlier, or novelty detection, in our setting, OOD datasets are more related to the OSR problem since the samples are composed of classes from the same datasets. For this reason, OSR problems are typically associated with more challenging scenarios where the statistics of unseen classes can be similar to the statistics of known classes in the dataset.

6.3. Rejection and active learning

While uncertainty evaluation measures are important to compare different uncertainty estimates, they do not take into consideration the real impact of using said measures when implementing new technologies into clinical practice. The notion of uncertainty and the ability to abstain from predicting a sample should be considered key features of any ML model to be used in clinical practice. Although, in the ECG classification field, none or few works address this important concept. In our analysis, we showed that by using such techniques, the ML-based models were able to abstain from predicting samples with high uncertainty, reducing the wrongly classified samples and consequently increasing the overall classification performance. Applying a 15% rejection threshold in the training set leads to more than double the rejection rate in external datasets, along with a 10% increase in classification performance.

This high rejection rate indicates potential dataset shift effects and the need to retrain models. When deploying a ML model, it is crucial to consider the dynamic environment it operates in, where concept drifts and unknown medical conditions may arise during testing. To cope with the cost of data labeling, selecting informative samples for labeling is essential. We found that uncertainty estimation is a viable method for selecting such samples within the active learning concept. By retraining the DeepEnsemble model using the rejected samples with higher uncertainty, the model learned the new data, achieving performance similar to internal validation with approximately 2000 new samples added.

6.4. Limitations and future work

Although this study provides valuable insights and advancements in uncertainty quantification for multi-label ECG classification, it has some limitations that should be acknowledged. The research relies on a single DNN architecture, which may limit the generalization of results to other models or architectures. Moreover, to properly evaluate performance on external validation sets, the training and test sets should ideally contain the same classes between datasets to ensure fairness in comparing different combinations of training and test sets. However, this restriction led us to reduce the number of classes to only the common ones, resulting in conclusions limited to the selected 9 classes. Considering the potential impact of foundation models in the field of uncertainty quantification [102] and medical AI [103], future research could explore the use of foundation models, tailored to the automatic diagnosis of ECG pathologies supported by UQ. This exploration may enhance the generalization capabilities of current DL models that address this task.

Moving forward, there are more research opportunities to explore. We intend to investigate the combination of both aleatoric and epistemic uncertainty for rejection, a topic that remains underexplored in

the literature. Additionally, selecting an appropriate uncertainty threshold is non-trivial, and existing studies often use arbitrary thresholds without solid reasoning. Hence, proper threshold selection is another important direction for future research.

7. Conclusions

Our study emphasizes the crucial role of uncertainty quantification in clinical decision-making, with a specific focus on multi-label classification, a largely overlooked topic in the literature. We use ECG classification as a case study. As a key contribution, we present the evaluation of state-of-the-art uncertainty estimation methods for multi-label classification, which has broad practical applications. Our results demonstrate that uncertainty estimation methods can aid in the machine learning process. However, current methods still have limitations in accurately quantifying uncertainty, particularly in the case of dataset shift. On external validation, a significant decrease in performance was noticed, accompanied by a decline in the quality of uncertainty estimates. Nevertheless, incorporating uncertainty estimates with a classification with rejection option improves the ability to detect such changes. After deploying a ML model, the data may change rapidly due to various reasons, such as a shift in the population, use of different medical equipment, or limited or unrepresentative training data. These changes often occur when new technologies are introduced in clinical practice, and retraining the ML models may become necessary. In such situations, where labeling a large amount of data may be impractical, we demonstrated that using uncertainty estimates as a criterion for sample selection can significantly reduce the number of samples that need to be labeled, and therefore, the frequency of model retraining compared to random sampling.

Despite the fact that uncertainty estimation is a fundamental feature for every ML model to be applied to clinical practice and there is a wealth of research on multi-label ECG analysis, very few studies address uncertainty estimations in their methodology. Our main motivation with this work is to spark future research on how to consider uncertainty quantification as a tool to improve the ML model development and their application to clinical decision-making, ultimately promoting the safe deployment of ML in various applications.

CRediT authorship contribution statement

Marília Barandas: Conceptualization, Formal analysis, Methodology, Investigation, Writing – original draft, Writing – review & editing. **Lorenzo Famiglini:** Formal analysis, Investigation, Writing – review & editing. **Andrea Campagner:** Formal analysis, Writing – review & editing. **Duarte Folgado:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Raquel Simão:** Formal analysis, Methodology, Investigation, Data curation, Writing – review & editing. **Federico Cabitza:** Conceptualization, Supervision, Writing – review & editing. **Hugo Gamboa:** Conceptualization, Supervision, Project administration, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The used datasets are already public available.

Acknowledgments

This work was supported by European funds through the Recovery and Resilience Plan, project ”Center for Responsible AI”, project number C645008882-00000055.

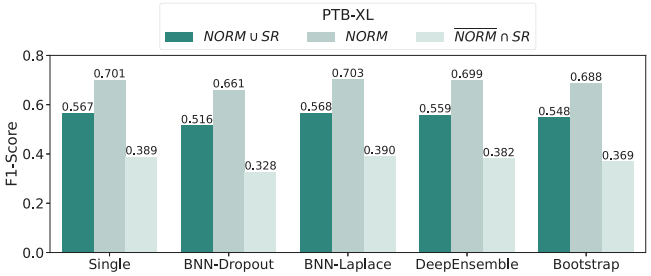

Fig. A.14. F1-Score for three subsets of the PTB-XL dataset. $NORM \cup SR$ represents the full dataset, $NORM$ is the subset with Normal Class and $\overline{NORM} \cap SR$ is the subset with only SR annotations.

Appendix A. Datasets annotations

The similarity distance analysis between labels from internal and external datasets led us to a more detailed examination of annotations between datasets, with a special focus on the NSR class from PTB-XL. We found that CPSC and G12EC datasets do not contain multi-label annotations with NSR class, unlike the PTB-XL dataset that contains 2704 multi-label annotations associated with NSR class. To avoid the differences related to different annotation protocols, the annotations provided by PhysioNet/CinC Challenge 2020 were used. However, originally PTB-XL had both NORM (normal ECG) and SR (Sinus Rhythm) label annotations that were merged and relabeled to NSR (Normal Sinus Rhythm). Contrarily, the CPSC dataset had originally only the Normal label that was relabeled to NSR. For the G12EC dataset, since it was first used on PhysioNet/CinC Challenge 2020, no additional information was found.

Following this finding, we proceed with an evaluation of a subset of the PTB-XL dataset that contains only Normal labels to understand whether the mentioned differences in annotation affected the classification performance. Fig. A.14 compares the F1-Score using the entire dataset ($NORM \cup SR$), a subset with Normal class ($NORM$) containing 13,932 recordings and the subset without Normal class ($\overline{NORM} \cap SR$) that contains 8544 recordings. The sum of recordings exceeds the number of PTB-XL records because of multi-label annotations per record. As expected, the subset with only Normal classes resulted in a significant improvement in performance across all methods. With this subset, both external validation sets obtained comparable performance.

Appendix B. Statistical analysis

Table B.8 presents a statistical analysis comparing the distribution of uncertainty values for correctly and incorrectly classified samples using DeepEnsemble model. The analysis employs the Mann–Whitney U test to assess the differences in the distributions. Before applying the Mann–Whitney U test we performed the analysis to validate the assumptions for the two samples' t-test. Firstly, the Kolmogorov–Smirnov test was performed for normality assumptions (in both scenarios the normality is met). Secondly, we computed the Levene test to evaluate if there was an equal variance between the analyzed groups. In this case, the equal variance assumption is not met. For this reason, we performed the non-parametric Mann–Whitney U test. Ultimately, Benjamini–Hochberg correction was applied to the p-values within each dataset. In a multi-label setting, we consider two scenarios: a label dependence scenario, where the entire label combination is either correct or incorrect, and a label independence scenario, where each class is treated as a binary classification problem. Both approaches are included in the table.

Table B.8
Statistical comparison of average uncertainty values for correctly classified and wrongly classified samples using the non-parametric Mann–Whitney U test. P-values, P-values adjusted with Benjamini–Hochberg procedure, and the absolute value of Cohen's d effect sizes (Δ) are shown for each comparison.

Dataset	Metric	Label Independence			Label Dependence		
		p_{value}	$p_{valueadj}$	Δ	p_{value}	$p_{valueadj}$	Δ
CPSC	p_{max}	<.001	<.001	1.700	<.001	< .001	1.485
	u_t	<.001	<.001	1.893	<.001	<.001	1.344
	u_a	<.001	<.001	1.835	<.001	<.001	1.283
	u_e	<.001	<.001	1.060	<.001	<.001	1.022
	σ^2	<.001	<.001	1.490	<.001	<.001	1.246
	vr	<.001	<.001	1.147	<.001	<.001	1.292
G12EC	p_{max}	<.001	<.001	1.566	<.001	<.001	1.100
	u_t	<.001	<.001	1.689	<.001	<.001	0.977
	u_a	<.001	<.001	1.584	<.001	<.001	0.849
	u_e	<.001	<.001	1.050	<.001	<.001	0.906
	σ^2	<.001	<.001	1.442	<.001	<.001	1.062
	vr	<.001	<.001	1.184	<.001	<.001	1.117
PTB-XL	p_{max}	<.001	<.001	1.409	<.001	<.001	0.877
	u_t	<.001	<.001	1.503	<.001	<.001	0.762
	u_a	<.001	<.001	1.439	<.001	<.001	0.672
	u_e	<.001	<.001	0.965	<.001	<.001	0.736
	σ^2	<.001	<.001	1.310	<.001	<.001	0.875
	vr	<.001	<.001	1.050	<.001	<.001	0.936

References

[1] E. Hüllermeier, W. Waegeman, Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods, Mach. Learn. 110 (3) (2021) 457–506.

[2] E. Begoli, T. Bhattacharya, D. Kusnezov, The need for uncertainty quantification in machine-assisted medical decision making, Nat. Mach. Intell. 1 (1) (2019) 20–23.

[3] B. Kompa, J. Snoek, A.L. Beam, Second opinion needed: Communicating uncertainty in medical machine learning, NPJ Digit. Med. 4 (1) (2021) 1–6.

[4] F. Rewicki, J. Gawlikowski, Estimating uncertainty of deep learning multi-label classifications using Laplace approximation, in: IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium, 2022, pp. 1560–1563.

[5] J.-Y. Jiang, W.-C. Chang, J. Zhong, C.-J. Hsieh, H.-F. Yu, Uncertainty in extreme multi-label classification, 2022, arXiv preprint arXiv:2210.10160.

[6] H. Wang, W. Liu, A. Bocchieri, Y. Li, Can multi-label classification networks know what they don't know? in: Advances in Neural Information Processing Systems, Vol. 34, 2021, pp. 29074–29087.

[7] A.H. Kashou, W.-Y. Ko, Z.I. Attia, M.S. Cohen, P.A. Friedman, P.A. Noseworthy, A comprehensive artificial intelligence–enabled electrocardiogram interpretation program, Cardiovasc. Digit. Health J. 1 (2) (2020) 62–70.

[8] A.M. Alqudah, S. Qazan, L. Al-Ebbini, H. Alquran, I.A. Qasmieh, ECG heartbeat arrhythmias classification: A comparison study between different types of spectrum representation and convolutional neural networks architectures, J. Ambient Intell. Humaniz. Comput. 13 (10) (2022) 4877–4907.

[9] Z. Ahmad, A. Tabassum, L. Guan, N.M. Khan, ECG heartbeat classification using multimodal fusion, IEEE Access 9 (2021) 100615–100626.

[10] I. Neves, D. Folgado, S. Santos, M. Barandas, A. Campagner, L. Ronzio, F. Cabitza, H. Gamboa, Interpretable heartbeat classification using local model-agnostic explanations on ECGs, Comput. Biol. Med. 133 (2021) 104393.

[11] Q. Yao, R. Wang, X. Fan, J. Liu, Y. Li, Multi-class arrhythmia detection from 12-lead varied-length ECG using attention-based time-incremental convolutional neural network, Inf. Fusion 53 (2020) 174–182.

[12] A.H. Ribeiro, M.H. Ribeiro, G.M. Paixão, D.M. Oliveira, P.R. Gomes, J.A. Canazart, M.P. Ferreira, C.R. Andersson, P.W. Macfarlane, W. Meira Jr., et al., Automatic diagnosis of the 12-lead ECG using a deep neural network, Nat. Commun. 11 (1) (2020) 1–9.

[13] T.-M. Chen, C.-H. Huang, E.S. Shih, Y.-F. Hu, M.-J. Hwang, Detection and classification of cardiac arrhythmias by a challenge-best deep learning neural network model, Iscience 23 (3) (2020) 100886.

[14] V. Gupta, M. Mittal, V. Mittal, N.K. Saxena, A critical review of feature extraction techniques for ECG signal analysis, J. Instit. Eng. (India): Series B 102 (5) (2021) 1049–1060.

[15] A.Y. Hannun, P. Rajpurkar, M. Haghpanahi, G.H. Tison, C. Bourn, M.P. Turakhia, A.Y. Ng, Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network, Nat. Med. 25 (1) (2019) 65–69.

[16] G.B. Moody, R.G. Mark, The impact of the MIT-BIH arrhythmia database, IEEE Eng. Med. Biol. Mag. 20 (3) (2001) 45–50.

[17] A. Ullah, S.M. Anwar, M. Bilal, R.M. Mehmood, Classification of arrhythmia by using deep learning with 2-D ECG spectral image representation, Remote Sens. 12 (10) (2020) 1685.

[18] R. He, Y. Liu, K. Wang, N. Zhao, Y. Yuan, Q. Li, H. Zhang, Automatic cardiac arrhythmia classification using combination of deep residual network and bidirectional LSTM, IEEE Access 7 (2019) 102119–102135.

[19] F. Liu, C. Liu, L. Zhao, X. Zhang, X. Wu, X. Xu, Y. Liu, C. Ma, S. Wei, Z. He, et al., An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection, J. Med. Imag. Health Inform. 8 (7) (2018) 1368–1373.

[20] D. Zhang, S. Yang, X. Yuan, P. Zhang, Interpretable deep learning for automatic diagnosis of 12-lead electrocardiogram, Iscience 24 (4) (2021).

[21] N. Strodthoff, P. Wagner, T. Schaeffter, W. Samek, Deep learning for ECG analysis: Benchmarks and insights from PTB-XL, IEEE J. Biomed. Health Inf. 25 (5) (2020) 1519–1528.

[22] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F.I. Lunze, W. Samek, T. Schaeffter, PTB-XL, a large publicly available electrocardiography dataset, Sci. Data 7 (1) (2020) 1–15.

[23] L.T. Duong, T.T. Doan, C.Q. Chu, P.T. Nguyen, Fusion of edge detection and graph neural networks to classifying electrocardiogram signals, Expert Syst. Appl. 225 (2023) 120107.

[24] S. Gustafsson, D. Gedon, E. Lampa, A.H. Ribeiro, M.J. Holzmann, T.B. Schön, J. Sundström, Development and validation of deep learning ECG-based prediction of myocardial infarction in emergency department patients, Sci. Rep. 12 (1) (2022) 1–14.

[25] A. Ballas, C. Diou, A domain generalization approach for out-of-distribution 12-lead ECG classification with convolutional neural networks, in: 2022 IEEE Eighth International Conference on Big Data Computing Service and Applications, BigDataService, 2022, pp. 9–13.

[26] H. Zhu, C. Cheng, H. Yin, X. Li, P. Zuo, J. Ding, F. Lin, J. Wang, B. Zhou, Y. Li, et al., Automatic multilabel electrocardiogram diagnosis of heart rhythm or conduction abnormalities with deep learning: A cohort study, Lancet Digit. Health 2 (7) (2020) e348–e357.

[27] M. Kent, L. Vasconcelos, S. Ansari, H. Ghanbari, I. Nenadic, Fourier space approach for convolutional neural network (CNN) electrocardiogram (ECG) classification: A proof-of-concept study, J. Electrocardiol. 80 (2023) 24–33.

[28] A.A. Rawi, M.K. Elbashir, A.M. Ahmed, Deep learning models for multilabel ECG abnormalities classification: A comparative study using TPE optimization, J. Intell. Syst. 32 (1) (2023) 20230002.

[29] Y. Gal, R. Islam, Z. Ghahramani, Deep bayesian active learning with image data, in: International Conference on Machine Learning, 2017, pp. 1183–1192.

[30] A. Sadafi, N. Koehler, A. Makhro, A. Bogdanova, N. Navab, C. Marr, T. Peng, Multiclass deep active learning for detecting red blood cell subtypes in brightfield microscopy, in: International Conference on Medical Image Computing and Computer Assisted Intervention, 2019, pp. 685–693.

[31] V.-L. Nguyen, M.H. Shaker, E. Hüllermeier, How to measure uncertainty in uncertainty sampling for active learning, Mach. Learn. 111 (1) (2022) 89–122.

[32] R. Senge, S. Bösner, K. Dembczyński, J. Haasenritter, O. Hirsch, N. Donner-Banzhoff, E. Hüllermeier, Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty, Inform. Sci. 255 (2014) 16–29.

[33] P. Tabarisaadi, A. Khosravi, S. Nahavandi, Uncertainty-aware skin cancer detection: The element of doubt, Comput. Biol. Med. 144 (2022) 105357.

[34] M. Abdar, M. Samami, S.D. Mahmoodabad, T. Doan, B. Mazoure, R. Hashemifesharaki, L. Liu, A. Khosravi, U.R. Acharya, V. Makarenkov, et al., Uncertainty quantification in skin cancer classification using three-way decision-based Bayesian deep learning, Comput. Biol. Med. 135 (2021) 104418.

[35] H. Asgharnezhad, A. Shamsi, R. Alizadehsani, A. Khosravi, S. Nahavandi, Z.A. Sani, D. Srinivasan, S.M.S. Islam, Objective evaluation of deep uncertainty predictions for Covid-19 detection, Sci. Rep. 12 (1) (2022) 1–11.

[36] M. Abdar, S. Salari, S. Qahremani, H.-K. Lam, F. Karray, S. Hussain, A. Khosravi, U.R. Acharya, V. Makarenkov, S. Nahavandi, UncertaintyFuseNet: Robust uncertainty-aware hierarchical feature fusion model with ensemble Monte Carlo dropout for COVID-19 detection, Inf. Fusion 90 (2023) 364–381.

[37] K. Wickstrøm, M. Kampffmeyer, R. Jenssen, Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps, Med. Image Anal. 60 (2020) 101619.

[38] G. Carneiro, L.Z.C.T. Pu, R. Singh, A. Burt, Deep learning uncertainty and confidence calibration for the five-class polyp classification from colonoscopy, Med. Image Anal. 62 (2020) 101653.

[39] Z. Huang, H. Lam, H. Zhang, Quantifying epistemic uncertainty in deep learning, 2021, arXiv preprint arXiv:2110.12122.

[40] A. Nguyen, J. Yosinski, J. Clune, Deep neural networks are easily fooled: High confidence predictions for unrecognizable images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 427–436.

[41] D. Heaven, et al., Why deep-learning AIs are so easy to fool, Nature 574 (7777) (2019) 163–166.

[42] J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. Torr, P. Dokania, Calibrating deep neural networks using focal loss, in: Advances in Neural Information Processing Systems, Vol. 33, 2020, pp. 15288–15299.

[43] J. Gawlikowski, C.R.N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, et al., A survey of uncertainty in deep neural networks, 2021, arXiv preprint arXiv:2107.03342.

[44] C. Blundell, J. Cornebise, K. Kavukcuoglu, D. Wierstra, Weight uncertainty in neural network, in: International Conference on Machine Learning, 2015, pp. 1613–1622.

[45] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: International Conference on Machine Learning, 2016, pp. 1050–1059.

[46] Z. Eaton-Rosen, F. Bragman, S. Bisdas, S. Ourselin, M.J. Cardoso, Towards safe deep learning: Accurately quantifying biomarker uncertainty in neural network predictions, in: Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention, 2018, pp. 691–699.

[47] M. Rußwurm, M. Ali, X.X. Zhu, Y. Gal, M. Körner, Model and data uncertainty for satellite time series forecasting with deep recurrent models, in: IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium, 2020, pp. 7025–7028.

[48] A. Graves, Practical variational inference for neural networks, in: Advances in Neural Information Processing Systems, Vol. 24, 2011, pp. 2348–2356.

[49] A. Mobiny, P. Yuan, S.K. Moulik, N. Garg, C.C. Wu, H. Van Nguyen, Dropconnect is effective in modeling uncertainty of Bayesian deep networks, Sci. Rep. 11 (1) (2021) 1–14.

[50] P. McClure, N. Kriegeskorte, Robustly representing uncertainty through sampling in deep neural networks, 2016, arXiv preprint arXiv:1611.01639.

[51] M.A. Kupinski, J.W. Hoppin, E. Clarkson, H.H. Barrett, Ideal-observer computation in medical imaging with use of Markov-chain Monte Carlo techniques, J. Opt. Soc. Amer. A 20 (3) (2003) 430–438.

[52] N. Ding, Y. Fang, R. Babbush, C. Chen, R.D. Skeel, H. Neven, Bayesian sampling using stochastic gradient thermostats, in: Advances in Neural Information Processing Systems, Vol. 27, 2014, pp. 3203–3211.

[53] J. Denker, Y. LeCun, Transforming neural-net output levels to probability distributions, in: Advances in Neural Information Processing Systems, Vol. 3, 1990, pp. 853–859.

[54] A. Kristiadi, M. Hein, P. Hennig, Learnable uncertainty under Laplace approximations, in: Uncertainty in Artificial Intelligence, PMLR, 2021, pp. 344–353.

[55] Z. Deng, F. Zhou, J. Zhu, Accelerated linearized Laplace approximation for Bayesian deep learning, Adv. Neural Inf. Process. Syst. 35 (2022) 2695–2708.

[56] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, in: Advances in Neural Information Processing Systems, 2017, pp. 6402–6413.

[57] E.J. Herron, S.R. Young, T.E. Potok, Ensembles of networks produced from neural architecture search, in: International Conference on High Performance Computing, 2020, pp. 223–234.

[58] I. Osband, J. Aslanides, A. Cassirer, Randomized prior functions for deep reinforcement learning, Adv. Neural Inf. Process. Syst. 31 (2018).

[59] B. He, B. Lakshminarayanan, Y.W. Teh, Bayesian deep ensembles via the neural tangent kernel, in: Advances in neural information processing systems, Vol. 33, 2020, pp. 1010–1022.

[60] V. Dwaracherla, Z. Wen, I. Osband, X. Lu, S.M. Asghari, B. Van Roy, Ensembles for uncertainty estimation: Benefits of prior functions and bootstrapping, 2022, arXiv preprint arXiv:2206.03633.

[61] A. Malinin, M. Gales, Predictive uncertainty estimation via prior networks, in: Advances in Neural Information Processing Systems, Vol. 31, 2018, pp. 7047–7058.

[62] L. Oala, C. Heiß, J. Macdonald, M. März, W. Samek, G. Kutyniok, Interval neural networks: Uncertainty scores, 2020, arXiv preprint arXiv:2003.11566.

[63] M. Możejko, M. Susik, R. Karczewski, Inhibited softmax for uncertainty estimation in neural networks, 2018, arXiv preprint arXiv:1810.01861.

[64] J. Van Amersfoort, L. Smith, Y.W. Teh, Y. Gal, Uncertainty estimation using a single deep deterministic neural network, in: International Conference on Machine Learning, 2020, pp. 9690–9700.

[65] M. Sensoy, L. Kaplan, M. Kandemir, Evidential deep learning to quantify classification uncertainty, in: Advances in Neural Information Processing Systems, Vol. 31, 2018, pp. 3179–3189.

[66] M. Raghu, K. Blumer, R. Sayres, Z. Obermeyer, B. Kleinberg, S. Mullainathan, J. Kleinberg, Direct uncertainty prediction for medical second opinions, in: International Conference on Machine Learning, 2019, pp. 5281–5290.

[67] T. Ramalho, M. Miranda, Density estimation in representation space to predict model uncertainty, in: International Workshop on Engineering Dependable and Secure Machine Learning Systems, 2020, pp. 84–96.

[68] J. Lee, G. AlRegib, Gradients as a measure of uncertainty in neural networks, in: 2020 IEEE International Conference on Image Processing, ICIP, 2020, pp. 2416–2420.

[69] F.T. Liu, K.M. Ting, Z.-H. Zhou, Isolation-based anomaly detection, ACM Trans. Knowl. Discov. Data (TKDD) 6 (1) (2012) 1–39.

[70] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT Press, 2016.

[71] M.M. Breunig, H.-P. Kriegel, R.T. Ng, J. Sander, LOF: Identifying density-based local outliers, in: Proceedings of the ACM SIGMOD, International Conference on Management of Data, 2000, pp. 93–104.

[72] D. Hendrycks, S. Basart, M. Mazeika, M. Mostajabi, J. Steinhardt, D. Song, Scaling out-of-distribution detection for real-world settings, 2019, arXiv preprint arXiv:1911.11132.

[73] K. Lee, K. Lee, H. Lee, J. Shin, A simple unified framework for detecting out-of-distribution samples and adversarial attacks, in: Advances in Neural Information Processing Systems, Vol. 31, 2018, pp. 7167–7177.

[74] W. Liu, X. Wang, J. Owens, Y. Li, Energy-based out-of-distribution detection, in: Advances in Neural Information Processing Systems, Vol. 33, 2020, pp. 21464–21475.

[75] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U.R. Acharya, et al., A review of uncertainty quantification in deep learning: Techniques, applications and challenges, Inf. Fusion 76 (2021) 243–297.

[76] S. Hong, W. Zhang, C. Sun, Y. Zhou, H. Li, Practical lessons on 12-lead ECG classification: Meta-analysis of methods from PhysioNet/Computing in cardiology challenge 2020, Front. Physiol. (2022) 2505.

[77] J. Belen, S. Mousavi, A. Shamsoshoara, F. Afghah, An uncertainty estimation framework for risk assessment in deep learning-based AFib classification, in: 2020 54th Asilomar Conference on Signals, Systems, and Computers, IEEE, 2020, pp. 960–964.

[78] J.F. Vranken, R.R. van de Leur, D.K. Gupta, L.E. Juarez Orozco, R.J. Hassink, P. van der Harst, P.A. Doevendans, S. Gulshad, R. van Es, Uncertainty estimation for deep learning-based automated analysis of 12-lead electrocardiograms, Eur. Heart J.-Digit. Health 2 (3) (2021) 401–415.

[79] A.O. Aseeri, Uncertainty-aware deep learning-based cardiac arrhythmias classification model of electrocardiogram signals, Computers 10 (6) (2021) 82.

[80] Y. Elul, A.A. Rosenberg, A. Schuster, A.M. Bronstein, Y. Yaniv, Meeting the unmet needs of clinicians from AI systems showcased for cardiology with deep-learning–based ECG analysis, Proc. Natl. Acad. Sci. 118 (24) (2021) e2020620118.

[81] W. Zhang, X. Di, G. Wei, S. Geng, Z. Fu, S. Hong, A deep Bayesian neural network for cardiac arrhythmia classification with rejection from ECG recordings, 2022, arXiv preprint arXiv:2203.00512.

[82] V. Jahmunah, E. Ng, R.-S. Tan, S.L. Oh, U.R. Acharya, Uncertainty quantification in DenseNet model using myocardial infarction ECG signals, Comput. Methods Programs Biomed. 229 (2023) 107308.

[83] J. Park, K. Lee, N. Park, S.C. You, J. Ko, Self-attention LSTM-FCN model for arrhythmia classification and uncertainty assessment, Artif. Intell. Med. 142 (2023) 102570.

[84] D. Hendrycks, K. Gimpel, A baseline for detecting misclassified and out-of-distribution examples in neural networks, 2016, arXiv preprint arXiv:1610.02136.

[85] S. Liang, Y. Li, R. Srikant, Enhancing the reliability of out-of-distribution image detection in neural networks, 2017, arXiv preprint arXiv:1706.02690.

[86] J. Mena, O. Pujol, J. Vitrià, Uncertainty-based rejection wrappers for black-box classifiers, IEEE Access 8 (2020) 101721–101746.

[87] M. Barandas, D. Folgado, R. Santos, R. Simão, H. Gamboa, Uncertainty-based rejection in machine learning: Implications for model development and interpretability, Electronics 11 (3) (2022) 396.

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

[89] A. Malinin, L. Prokhorenkova, A. Ustimenko, Uncertainty in gradient boosting via ensembles, 2020, arXiv preprint arXiv:2006.10562.

[90] M.H. Shaker, E. Hüllermeier, Aleatoric and epistemic uncertainty with random forests, in: International Symposium on Intelligent Data Analysis, Springer, 2020, pp. 444–456.

[91] E.A.P. Alday, A. Gu, A.J. Shah, C. Robichaux, A.-K.I. Wong, C. Liu, F. Liu, A.B. Rad, A. Elola, S. Seyedi, et al., Classification of 12-lead ecgs: The physionet/computing in cardiology challenge 2020, Physiol. Meas. 41 (12) (2020) 124003.
[92] G. Scalia, C.A. Grambow, B. Pernici, Y.-P. Li, W.H. Green, Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction, J. Chem. Inform. Model. 60 (6) (2020) 2697–2717.
[93] P. Tabarisaadi, A. Khosravi, S. Nahavandi, M. Shafie-Khah, J.P. Catalão, An optimized uncertainty-aware training framework for neural networks, IEEE Trans. Neural Netw. Learn. Syst. (2022) 1–8.
[94] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, in: International Conference on Machine Learning, 2017, pp. 214–223.
[95] F. Cabitza, A. Campagner, F. Soares, L.G. de Guadiana-Romualdo, F. Challa, A. Sulejmani, M. Seghezzi, A. Carobene, The importance of being external. methodological insights for the external validation of machine learning models in medicine, Comput. Methods Programs Biomed. 208 (2021) 106288.
[96] P.E. McKnight, J. Najab, Mann-Whitney U test, Corsini encyclop. psychol. (2010) 1.
[97] D. Thissen, L. Steinberg, D. Kuang, Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons, J. Educ. Behav. Statist. 27 (1) (2002) 77–83.
[98] S.S. Sawilowsky, New effect size rules of thumb, J. Modern Appl. Statist. Methods 8 (2) (2009) 26.
[99] R. Roady, T.L. Hayes, R. Kemker, A. Gonzales, C. Kanan, Are out-of-distribution detection methods effective on large-scale datasets?, 2019, arXiv preprint arXiv:1910.14034.
[100] E.W. Steyerberg, F.E. Harrell, Prediction models need appropriate internal, internal–external, and external validation, J. Clin. Epidemiol. 69 (2016) 245–247.
[101] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, J. Snoek, Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift, in: Advances in Neural Information Processing Systems, Vol. 32, 2019, pp. 13991–14002.
[102] M. Sun, W. Yan, P. Abbeel, I. Mordatch, Quantifying uncertainty in foundation models via ensembles, in: NeurIPS 2022 Workshop on Robustness in Sequence Modeling, 2022.
[103] M. Moor, O. Banerjee, Z.S.H. Abad, H.M. Krumholz, J. Leskovec, E.J. Topol, P. Rajpurkar, Foundation models for generalist medical artificial intelligence, Nature 616 (7956) (2023) 259–265.