Journal of Biomedical Informatics 127 (2022) 103996

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Evaluating pointwise reliability of machine learning prediction

Giovanna Nicora [a,*], Miguel Rios [b], Ameen Abu-Hanna [b], Riccardo Bellazzi [a]

[a] Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy
[b] Department of Medical Informatics, Amsterdam UMC, University of Amsterdam, the Netherlands

ARTICLE INFO

Keywords:
Machine learning trustworthiness
Predictive reliability
Uncertainty

ABSTRACT

Interest in Machine Learning applications to tackle clinical and biological problems is increasing. This is driven by promising results reported in many research papers, the increasing number of AI-based software products, and by the general interest in Artificial Intelligence to solve complex problems. It is therefore of importance to improve the quality of machine learning output and add safeguards to support their adoption. In addition to regulatory and logistical strategies, a crucial aspect is to detect when a Machine Learning model is not able to generalize to new unseen instances, which may originate from a population distant to that of the training population or from an under-represented subpopulation. As a result, the prediction of the machine learning model for these instances may be often wrong, given that the model is applied outside its "reliable" space of work, leading to a decreasing trust of the final users, such as clinicians. For this reason, when a model is deployed in practice, it would be important to advise users when the model's predictions may be unreliable, especially in high-stakes applications, including those in healthcare. Yet, reliability assessment of each machine learning prediction is still poorly addressed.

Here, we review approaches that can support the identification of unreliable predictions, we harmonize the notation and terminology of relevant concepts, and we highlight and extend possible interrelationships and overlap among concepts. We then demonstrate, on simulated and real data for ICU in-hospital death prediction, a possible integrative framework for the identification of reliable and unreliable predictions. To do so, our proposed approach implements two complementary principles, namely the density principle and the local fit principle. The density principle verifies that the instance we want to evaluate is similar to the training set. The local fit principle verifies that the trained model performs well on training subsets that are more similar to the instance under evaluation. Our work can contribute to consolidating work in machine learning especially in medicine.

1. Introduction

Machine Learning (ML) for making predictions and extracting information from data is increasingly applied in many different fields, from medicine [53,63] to finance [32]. Thanks to these algorithms we are able to analyze the increasing mass of biological data produced by next-generation sequencing technologies, to make inference about genomic mutation pathogenicity [69,5,67] or cell type [33], to identify therapy targets, and to design new therapeutic compounds [23]. ML can also be used to extract information in EHR data to predict patients' diagnosis [59], post-hospitalization risks [49] or heart failure [73], and it can also support the analysis of emerging diseases, such as outcome severity during COVID-19 infection [4].

Moreover, there is a fast increase in the number of FDA approved products that are now available on the market; most of them developed for the fields of Radiology, but several other areas are represented, such as cardiology and internal medicine [7]. The transition from research to bedside poses several challenges in order to provide systems with safeguards that can allow a widespread clinical use [50,97]. Those challenges include logistical and regulatory issues, with focus on trustworthiness and fairness [21,97]. Recently, the European Union regulations, in compliance with the General Data Protection Regulation (GDPR), have defined minimal standards for implementing ML systems in public health, requiring, among others, the explainability of the model. Explainable AI (XAI) is becoming a promising field of research within the AI community. Its purpose is to investigate methods for analyzing or complementing AI black box models to make the internal logic and output of algorithms transparent and/or interpretable.

* Corresponding author.
E-mail address: giovanna.nicora01@universitadipavia.it (G. Nicora).

https://doi.org/10.1016/j.jbi.2022.103996
Received 31 October 2021; Received in revised form 7 January 2022; Accepted 11 January 2022
Available online 15 January 2022
1532-0464/© 2022 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Table 1
Definition of different concepts to evaluate a classifier on a test set, along with possible measures and relevant probability distributions.

Concepts	Definition	Measures	Relevant Distributions
Calibration	The predictive probability distribution of a classification model honestly expresses its observed probability distribution [94]	Calibration (reliability) diagram	$P(Y \in \cdot \mid f(X))$ $f(X)$
Robustness	How accurate is a ML algorithm on new independent data even if one or more of the input independent variables (features) or assumptions are drastically changed due to unforeseen circumstances	Accuracy on adversarial samples or on small perturbations of inputs	$p(Y = C\|X;\theta)$ $p(X)$
Sharpness	Variability of predictions as described by distributions of predictions [64]	$\frac{\sum_{i=1}^{N} P(y_i\|x_i)(1 - P(y_i\|x_i))}{N}$	$p(Y = C\|X;\theta)$

Table 2
Definition of different concepts to evaluate a classifier **on a single instance,** along with possible measures and relevant probability distributions.

Concepts	Definition	Measures	Relevant Distributions
Uncertainty (of a single ML prediction)	Variability in a model's prediction due to different types of uncertainty (epistemic/ aleatoric uncertainty)	Variance, entropy, confidence intervals	$p(Y = C\|X;\theta)$
Reliability (of single ML prediction)	Degree of trust that the prediction is correct $p(\hat{y}_i = = y_i)$	Local fit principle, density principle	$p(Y = C\|X;\theta)$ $p(X)p(Y)$

Explainability can be thus perceived as a property through which ML can gain trust from the user [74]. However, when dealing with a model's prediction, we trust the classifier not only if we *understand* its logic but, most importantly, if its output classification is likely to be *correct*. XAI may help in identifying incorrectly classified examples by identifying artifacts or undesirable correlations from training [74], but it does not in itself enable trust [42]. For instance, XAI may be less useful to identify critical situations in the data when they are not known *a priori* by users. In other words, XAI can be used to help understand the classification process made by the model, but it cannot assess reliability of single ML predictions *per se*. Understanding whether we can trust the prediction of ML systems is crucial to identify failures, since ML inherently suffers from dataset shifts and poor generalization ability across different populations [50,41]. As a matter of fact, examples of fooling deep learning (DL) networks with adversarial instances are widely reported in the literature [96,24]. Generalization ability under data perturbation (i.e. the robustness of the algorithm) can be promoted by using stable ML methods and tested before deployment [86] and strategies for model updating need to be considered for deployed models [22]. In addition to that, we need methods and metrics to identify potential failures, due to perturbations or poor generalization ability, *a posteriori,* as part of a monitoring strategy of our model.

Assessing the reliability of ML predictions may increase the trust of potential users when using these tools in many high stakes applications [42]. In a ML pipeline, model performance is evaluated on a set of instances (test or validation set) through performance metrics such as accuracy, sensitivity and specificity. These metrics represent an aggregate prediction capability of the model on that test set, since they are computed (in the case of a binary classification task) from the number of true/false positive examples and true/false negative examples in the test set, or from the complete set of predicted posterior probabilities (to compute AUC and PRC). When the model is deployed, we would expect that it will perform (on average) as well as it did on the test/validation set. However, this may not be the case [76]. Moreover, the overall performance metrics, mentioned above, do not provide a way to measure the performance of single new cases whose class is predicted by the model during deployment [54]. Additionally, we may not know the true class of these new future examples. For these reasons, it is crucial to develop new types of metrics and methodologies that can aid us to understand whether the classification can be wrong *case by case*. For instance, suppose that we have trained a ML model to predict whether a patient has a tumor based on the results of laboratory testing and MRI images. Suppose that the classifier predicts, for a certain patient, that his/her probability of having a tumor is 0.89. The probability threshold for classification is 0.5, therefore the classifier assigns the patient to the positive class. Through reliability assessment, we would like to assign a degree of reliability of such prediction. The pipeline may report that the predicted class for our patient is "positive" with a certain reliability, for instance 0.7, or a binary value (0 or 1, indicating an unreliable or reliable classification). In this article, we aim at discussing potential approaches to evaluate the reliability of ML model predictions that can be integrated in clinical decision support systems. We use the term "reliability" to indicate the degree of trust that we have on the prediction made by the ML model on a single example, as used in Saria et al. [76] and in [54]. The application of reliability can support the identification of correct and incorrect cases, mentioned in the definition of "trust in model correctness" [42]. Naively, the probability of the predicted class could be seen as the classifier's trust on that prediction, but this approach can be misleading since it does not account for the algorithm's biases [54]. Other strategies can be pursued, but we observe a lack of standardization in the usage of different terms such as reliability, uncertainty or robustness in the literature. For this reason, we will first start with a background section where we list the relevant concepts pertaining to model evaluation, both in terms of performance ability and reliability, and we then systematically review relevant papers that, to some extent, cover our principles of interest. Afterwards, we will present our approach to integrate different concepts to perform reliability assessment on simulated and real clinical data, and we will conclude our work with a discussion and some closing remarks.

2. Background

2.1. Terminology and problem definition

Let us consider the supervised classification problem, where a ML model $f : R^m \rightarrow R^c$

has been trained on a training set $D = (x_i, y_i)$, where $x_i \in R^m$ is the vector containing the attribute values for the i-th example and $y_i \in (1, .., c_j, ...C)$ is the class of the training examples. In particular, for a discriminative ML algorithm:

$$f := p(Y = c_j | X \in D)$$

is the predicted posterior probability of the class being c_j given the training data D. In the development phase of our model, we need to evaluate the performance of the ML classifier on a new set of data, called test set $T = (x_j, y_j)$ (alternatively, we can use bootstrap, cross-validation or other resampling approaches to estimate errors). On the test set, the

model is commonly evaluated in terms of its predictive accuracy, i.e. the fraction of correctly classified examples in the test set. In many applications, such as in medicine, accuracy alone should not be the only evaluation metric, especially when the classes are imbalanced and/or the cost of misclassification for a class is higher with respect to the other (s) [58]. For this reason, other metrics, such as precision (i.e. positive predictive value), recall (i.e. sensitivity), specificity, F1 score and Matthew's Correlation Coefficient should be reported and analyzed as well [43]. These metrics are applied to the discretized output of the classifier, which is the posterior predicted probability of the class given the data. The computation of such measurements on a test set represents the most common approach to evaluate the performance of a classifier. Yet, other measurements, reflecting different concepts, can be used during validation and deployment. Here, we provide an overview of these concepts, reported in Table 1 and Table 2. Table 1 highlights metrics that are computed on the entire test or validation set, while in Table 2 we reported metrics applied to a single machine learning prediction.

Calibration measures whether the distribution of the target class conditional on any given prediction of our model is equal to that prediction [94]. It is worth noting that the concept of calibration is called "reliability" in the weather forecast community [64], but in this paper we will use "calibration" since with "reliability" we are referring to a comprehensive concept that aims at assessing whether the predicted class is equal to the true class of a single example. Through the variability of the predicted probability distributions we can evaluate the sharpness [64], for instance by computing the variance of the predicted posterior probabilities [90]. These "overall" metrics measure the ability of the classifier on the complete test set. If the test set is representative of the entire underlying population of interest, then we will expect that the classifier will perform as well as in the test on future data coming from the population. Another property that can be evaluated on a set of examples is the robustness of the classifier, i.e. how accurate is a ML algorithm on new independent data even if one or more of the input independent variables (features) or assumptions are drastically changed due to unforeseen circumstances. Out-of-distribution (OOD) samples and adversarial samples are often generated to fool DL [12]. Robustness against adversarial attacks should be guaranteed in high stakes applications, such as medical ML [25].

These metrics are especially useful during the validation phase, to understand how the model performs in general. In deployment, we may want to understand how the classifier is performing case by case. In this context, **uncertainty** is an important concept related to AI trust. In statistics, uncertainty is a statement about a single prediction, that can be for instance a confidence interval. Ideally, especially in the medical field, a predictive model should have the ability to abstain from prediction when the uncertainty is high [52]. Uncertainty can be related to the model's "self-confidence" on the prediction: for instance, given a binary discriminative classifier outputting a probability, we can label the model as "certain" when the predicted probability is near 0 or 1. In this sense, uncertainty is a measure of sharpness on a single example. From now on, we will refer to this type of uncertainty as "unsharpness", to distinguish it from the predictive uncertainty, which is due to the epistemic and aleatoric uncertainty [65]. Aleatoric uncertainty refers to uncertainty in the data, such as noise, while epistemic uncertainty refers to the uncertainty in the model, such as the uncertainty in the model's parameters due to lack of knowledge. Epistemic uncertainty should increase when the prediction is made on an instance which is "far" from the training set (i.e. out-of-distribution (OOD) samples). The total uncertainty is the sum of aleatoric and epistemic uncertainty. Aleatoric uncertainty represents the *irreducible* part of the uncertainty, while epistemic uncertainty is the *reducible* part of (the total) uncertainty. In ML, the sources of uncertainty (aleatoric or epistemic) are not usually distinguished [40]. Different methodologies allow for total uncertainty quantification, either in terms of building a confidence interval around a prediction, or by computing a single score. Gaussian processes can also estimate uncertainty by computing the variance of predictions around a single point, and they have been used to select small molecules based on predicted binding affinity with kinases [39]. A widely used method to measure the uncertainty is by using multiple classifiers. Intuitively, the more the classifiers disagree, the higher the uncertainty. Ensembles of classifiers can be used to estimate both aleatoric and epistemic uncertainty in a formal way, by computing the entropy or the relative likelihood of the posterior probabilities predicted by each weak classifier [82]. In Deep Networks, uncertainty is estimated using drop-out methods or through ensembles of networks. In [56], the authors build a DL network that includes drop-out Bayesian uncertainty estimation, that can be used to diagnose diabetic retinopathy, showing that decisions informed by uncertainty estimation can improve diagnostic performance. Aleatoric and epistemic uncertainty of Bayesian deep networks have been estimated for computer vision tasks in [51]. The reader can refer to [2] for a review of uncertainty quantification of DLs and to [40] for formal definitions of uncertainty and methodological approaches to estimate total, epistemic and aleatoric uncertainty. Another assumption that has been made is that the predicted probability is the measure of (un)certainty, provided that the model is calibrated [55]. However, calibration itself can suffer from dataset shift, thus making this type of uncertainty estimation less trustworthy in case of OOD examples [71].

In 2009, [10] defined reliability in ML as "*any qualitative property or ability of the system which is related to a critical performance indicator (positive or negative) of that system, such as accuracy, inaccuracy, availability, downtime rate, responsiveness, etc*". In this paper, we are interested in the assessment of individual (pointwise) reliability, thus we will use the term reliability to indicate a property of a single ML prediction made on a given sample. In this context, [76] put together principles from reliability engineering to assure that a ML system performs as intended in deployment, i.e. **without failure** and **without specified performance limits**. These principles encompass (1) prevention of failures, i.e. adoption of techniques, during development, to avoid possible incorrect cases, (2) identification of failures during deployment and (3) maintenance of the model, i.e. fix or prevent the failures when they occur. To identify failures, the concept of point-wise reliability should be applied, by evaluating the model's prediction of a single instance and thus possibly rejecting the classification when such prediction is deemed as "unreliable". To compute pointwise reliability, two principles should be considered according to [76]:

1. **density principle**: it asks whether the test sample is **close** to the training set (similar to outlier detection or out-of-distribution samples detection)
2. **local fit principle**: it asks whether the model was **accurate** on training samples closest to the test case.

By evaluating these two principles we should be able to identify the "context" in which the model is expected to be performant, thus promoting the "trust in model correctness". In other words, the model should perform well on samples that (1) are similar to the training set and (2) are similar to training samples on which the model performed well. The density principle only refers to the distribution on the attribute space of the training set and the new instance to be classified, while the local fit principle also refers to the averaged performance of the classifier. The first principle acknowledges that the model may be unreliable when the distribution of the new data is changing for any reason. The second principle acknowledges that the model was not "perfectly accurate" in the first place (during model development and evaluation) on specific subsets of the training set.

A formal definition of pointwise reliability was proposed by [54] as the probability that the predicted class is equal to the true class value of an instance. To calculate this probability, the authors proposed a transductive approach that involves the re-training of the algorithm: each time a new instance needs to be classified, the posterior predicted probability of the original model is stored, the new instance is included

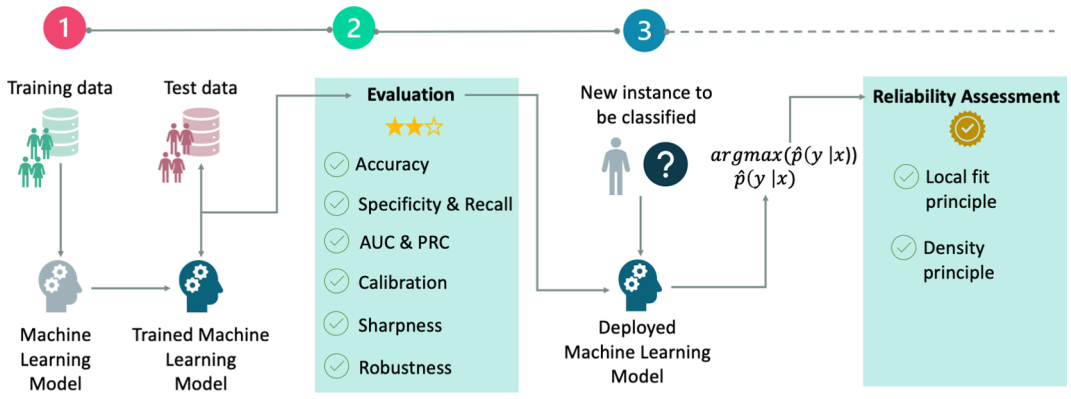


Fig. 1. Supervised Machine Learning model life cycle. First, a model is trained on an available set of training samples (1). Then, the model is evaluated on a test set in terms of different metrics (2). Eventually, the model is deployed, and it can be used for prediction on new cases (e.g. patients). In such single cases, the true class is usually unknown, but it is necessary to understand whether the ML prediction should be trusted. Reliability assessment can support the identification of unreliable predictions (3).

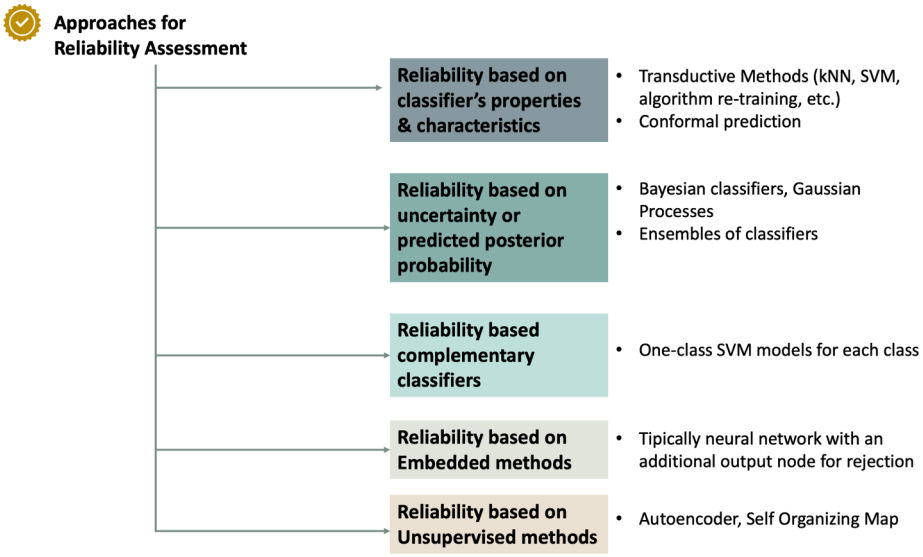


Fig. 2. Taxonomy of approaches implementing reliability assessment reported in the literature.

in the training set and the model is re-trained. The posterior probability is predicted by this second model, and it is compared with the stored predicted probability. Intuitively, if the posterior probability changes with some degree after the inclusion of the instance in the training set, such instance is adding knowledge that was previously unavailable. Therefore, the first predicted probability should be deemed as unreliable. However, this approach implies the re-training of the algorithm, which is sometimes unfeasible due to computational time and/or unavailability of the training set. Uncertainty may be used to evaluate the reliability of a classifier, following the principle that we will be more willing to trust a predictive model that is "certain" about a prediction. For instance, through uncertainty estimation, in [56] the authors were able to detect difficult cases for further inspection, resulting in substantial improvements in detection performance in the remaining data. Moreover, as the epistemic uncertainty can be caused by OOD data, it may seem that evaluating epistemic uncertainty naturally includes the evaluation of the density principle. Yet, as studies have shown, uncertainty evaluation through the predicted posterior probability is not trustworthy under dataset shift [71], which is exactly the case in which we would like to have a reliable approach to detect possible failures. Also the robustness of the model is related to its reliability, since it promotes the development of accurate models on adversarial samples. Yet, robustness is limited to the evaluation of adversarial samples, while a pointwise reliability estimation deals with any kind of future input samples.

Fig. 1 shows a workflow for ML model development and deployment, highlighting validation and continuous monitoring through reliability assessment.

2.2. Related work

Several works incorporate the notions of reliability, uncertainty, and robustness to provide classifiers able to detect potential failures. Some classifiers naturally provide a way to compute a form of reliability measure, while other work exploits or adapts well-known classifiers to compute a form of reliability measure.

In particular, a branch of pattern recognition research, named ***Learning with Reject Option*** (LRO), focuses on the development of classifiers with reject ability. In LRO, the aim is to find a subset of the test set for which the model has higher performances, thus identifying non-reliable examples for which the classifier may withhold the classification [6]. The concept of LRO is equivalent to the concept of "selective prediction", in which a model can choose to abstain itself from the classification when the uncertainty is high [31]. For instance, Bayes classifiers were exploited to detect reliable regions in gene expression data [34], while posterior probability and contextual information are used to

classify images from teratoma tissues and reject non-reliable portions [15]. By identifying samples for which the classification may be wrong, classification reliability and classification with reject option can be seen as synonyms [3,27,61]. Learning with rejection often implies the definition of a reject threshold. In pattern recognition, an optimal reject threshold based on the rejection cost was formulated in 1970 [14], and a reject threshold was computed based on the known costs of misclassification and rejection in text categorization [27]. An optimal reject rule for a binary classification problem was also formulated by [89], in which error costs peculiar for the particular problem are used to compute optimal reliability thresholds based on ROC curves. [75] demonstrated that the approaches from [14] and from [89] are actually equivalent. Another important aspect is the performance evaluation of LRO classifiers. [66] introduced the *accuracy-rejection curve* (ARC), which represents how the accuracy varies as function of the rejection rate. To account for misclassification and rejection costs in applications such as medical diagnosis, [1] recently extended the ARC curves.

A binary classifier with reject option (or reliability estimation) can be developed in different ways: 1) by choosing a classifier whose characteristics can be exploited to evaluate reliability 2) by thresholding the predicted posterior probability or an uncertainty estimation 3) by designing independent and complementary classifiers and 4) by designing classifiers that intrinsically learn to classify with rejection (also known as *embedded* methods) [61,85]. Also 5) unsupervised methods, such as generative models, can be used to capture statistical properties of the dataset to be used for reliability estimation. In the following subsections we will discuss these different approaches (as reported in Fig. 2). Note that we are interested in LRO methods that identify unreliable regions on the feature space. Another application of LRO methods is the detection of samples in *outlier* classes, in situations where new classes can appear [47,88]. Moreover, we focus on reliability assessment for classification problems, while also regression problems may benefit from reliability estimation [9].

2.2.1. Reliability based on classifier's properties and characteristics

Transductive algorithms, such as k-Nearest Neighbor and Support Vector Machines, rather than output a general classification rule or function, classify instances with respect to the available training set. Therefore, reliability can be determined based on how much the new case is "close" to the training set, an assumption that is included in the "density principle" as well as in other previous work [30]. For instance, using k-Nearest Neighbour, one can rely on the "purity" of the k nearest neighbours for an example: if less than k' neighbors are in the predicted class, then the classification is rejected (i.e. unreliable) [37]. In [78] the *confidence* of a SVM classifier is computed from the values of the Lagrange multipliers, since they reflect how "odd" is a particular example. The term *confidence* is sometimes used to refer to the same concept of reliability [54,26,48,44]. Distance from the separating hyperplane of the SVM was used by [93] to detect arrhythmia from complex electrocardiogram (ECG) signals and a SVM with l1 penalty and reject option was developed by [13] to perform feature selection and classify cancer samples based on gene expression profiles. Among neural network approaches, LVQ (Learning Vector Quantisation) networks consist of two layers for input and output and an array of codebook vectors. The output is defined as the distance between the input vector and the codebook vectors for every class. The classification for an input sample x would be the class whose representative codebook vector has the minimum distance to the sample x. The distances between the sample and the codebook vectors of each class are used by [3] to estimate the probability densities for codebook vectors, which in turn are used to select the threshold for rejection. LVQ networks for LRO were first proposed by [17,18] and were also used for reliable footstep identifications [87]. It has also been argued that LVQ combines model interpretability with the reject option [11]. Along with transductive algorithms, conformal prediction can also be used to compute reliability. Conformal prediction is a general framework that uses past experience to predict the confidence intervals for a new prediction in an on-line setting, in which the prediction of a new case is made based on the previous classified example. Given a predicted label $\hat{y}$ and the true label y, conformal prediction can compute a 95% confidence interval that contains the y label with at least 95% probability. Usually, such regions also include the predicted label $\hat{y}$ In case of regression, $\hat{y}$ and y are continuous values, while for classification they are discrete values [81]. Conformal prediction is often applied for reliable drug discovery (see [20] for a review) and can be extended to DL as well [36,62].

More recently, a new method computes the reliability of a trained model provided that the gradient of its loss function is accessible [80].

2.2.2. Reliability based on uncertainty or predicted posterior probability

The most straightforward way to develop LRO classifiers is adding a rejection rule to any type of classifier outputting a posterior probability or an uncertainty estimation. The reject threshold is set according to such value, based on some trade-off between misclassification and rejection [27,34,8,95]. This approach is also called *plug-in* [14,35]. When using the predicted posterior probability, we assume that such probability is calibrated. In this case, when the predicted posterior probability for the predicted class is lower than a threshold, then the classification is deemed as unreliable. In [72] authors define rejection intervals, based on misclassification rates, for the predicted posterior probability of a classifier that predicts the effect of amino acid substitutions on protein stability. In order to use a measure of uncertainty, we can use Gaussian Process classifiers, or we can compute epistemic, aleatoric and total uncertainty as the relative likelihood, or the entropy, of the posterior probabilities predicted by each weak classifier in an ensemble [40,82]. Alternatively, [48] exploited different "fusion methods" for posterior probabilities of ensembles, for instance by summing or taking the maximum value of the probabilities, to develop an anti-diabetic drug failure classifier with a reject option.

Recently, a variety of work has been published dealing with uncertainty and rejecting options for DL networks. In [19], deep networks ensembles are used to predict drug efficacy. It has been observed that DL tends to provide high predicted probabilities also for misclassified examples. This is due to the fact that the softmax function, which is usually used by DL to output a predicted probability, is fast-growing exponentially [38]. Regarding the usage of the DL posterior probability as an indicator of reliability, published results are contradictory [38] observed that the predicted probability of misclassified examples can be usually lower with respect to the correctly identified instances, and therefore DL probabilities may be evaluated to reject a classification. Lakshminarayanan et al. [55] successfully used an ensemble of networks to estimate the uncertainty from the predicted posterior probability. However [71] performed a large-scale evaluation of different methods for quantifying predictive uncertainty under dataset shift and shows that, along with accuracy, also the calibration of the classifier deteriorates under dataset shift. Consequently, using the classifier's prediction to detect unreliable classification may not be trustworthy. On tabular medical data, a recent study found that, while ensemble methods for prediction and uncertainty quantification are the best performing in detecting correlation between uncertainty and performance, they are less suited for the identification of out-of-distribution examples. In this latter case, novelty detection methods, for instance based on Variational Autoencoder (VAE) performs better [60]. The limitation of uncertainty estimation to detect OOD examples is also reported by [92], where the authors investigated different methodologies for uncertainty estimation to detect OOD examples in medical data.

2.2.3. Complementary classifier

Another possible solution to design a pipeline that learns with rejection in a binary classification problem is the development of two independent classifiers. The first classifier is trained to output C1 only when the posterior probability for C1 classification is high, the second is

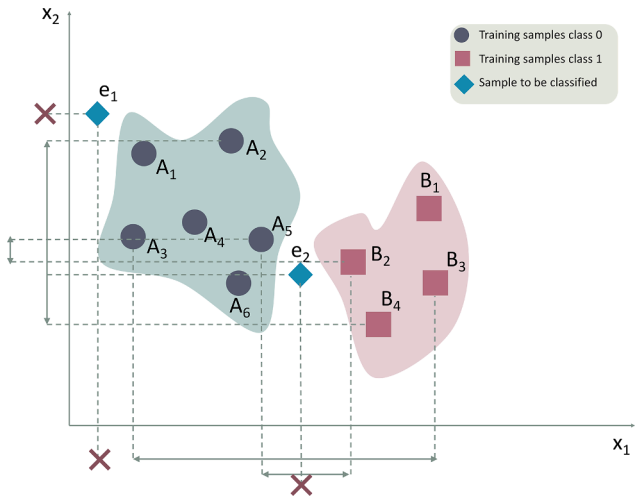


Fig. 3. Examples of inner and outer border samples on two attributes. A1-A6 belongs to class 0, B1-B4 to class 1. For attribute x_1, A5 and B2 are inner borders, A3 and B3 outer borders. For attribute x_2, A5 and B2 are inner borders, while A2 and B4 are outer borders A4 and B1 are outer borders. The test example e_2 would be inner border for attribute x_1, while the e_1 example would be outer border for both for x_1 and x_2 attributes. Therefore, the reliability computed for e_1 would be 1-½=0.5, for e_2 the reliability would be 1–2/2 = 0.

trained to output C2 only when the posterior probability for C2 classification is high. The classification of a single new instance is seen as reliable if and only if the two classifiers agree. This approach for classification with the reject option was proposed by [85] and a similar methodology was developed by [61] for software defect prediction. [35] fitted two one-class support vector machines (one for each class in a binary problem) that detect the two unreliable regions (when examples of different classes are overlapping in the feature space and when samples are actually outlier) for rejections. One-class complementary classifiers were also used in [44].

2.2.4. Embedded methods

Embedded methods intrinsically learn to classify with rejection [28,30,31,57,84]. Back in 1992, [57] proposed a neural network able to compute reliability of classification by adding additional output nodes that provide uncertainty in the classification. [16] included the reliability estimation within a DL approach by defining a new target criterion for the network training. Rules for reliability assessment of multilayer perceptron were also introduced [17,18], where the authors assumed that the function that describes the performance of the networks is a linear function combining the recognition rate, the misclassification rate and the rejection rate. The threshold for rejection is computed according to the values of output nodes of the network.

2.2.5. Unsupervised methods

To capture similarities in data, also unsupervised methods, such as Self Organizing Map (SOM), can be used to summarize the statistical properties of the training data, and they are therefore suitable for detecting patterns of reliable and unreliable instances [29]. Generative models, like GANs, have been used to learn a reliable region in the Learning with Reject Option (LRO) framework by [30]. GANs (Generative Adversarial Networks) are particular types of unsupervised networks, which consist of two models: the generative model generates samples from a latent distribution learned from the data, and the adversarial model predicts whether the generated samples belong to the original dataset. For this reason, the discriminative model of GAN can be used to detect OOD and therefore potentially unreliable samples. An other type of neural networks, named autoencoder (AE), can learn efficient codings of unlabeled data by attempting to regenerate the input from the encoding. AE reconstruction error can be used as a measure of uncertainty [60].

All the aforementioned work relies on the usage of particular types of classifiers, forcing the user to choose among ensembles, transductive algorithms etc, thus limiting them in the design of the model. Fewer work has been developed to deal with any types of classifiers. Given that the assumption for rejecting or assigning a low confidence/reliability to an example is often that this example is "distant" from the training set, the parallelism becomes clear between learning with rejection and data subset selection, such as instance selection methods [29]. Instance selection methods are usually exploited to remove from the training set the non-useful instances, thus speeding runtime training while not affecting classification performance [70]. Such non-useful instances may be those similar to already well represented examples in the training set, and they are not adding new relevant information to the dataset. Therefore, unlike LRO, which is performed *after* model training, instance selection is applied *before* training to select the most relevant examples in the training set. In a recent work, we suggested the exploitation of instance selection methods for learning with rejection [68]. We do not use instance selection methods to actually select training examples, but rather to assess whether a new unseen example would be selected as "relevant" in comparison with the training set: if so, the new element may not come from the training set population distribution, and therefore we cannot trust the model prediction (following the density principle). Therefore, instance selection is performed *after* the training phase, by comparing the training relevant examples with the new example(s) to be classified. This approach can be applied to any type of classifier. Another "classifier-independent approach" has been proposed by [45]: first the authors detect high density region for each class by filtering out training samples that may be outliers, and then a "trust score" is calculated as the ratio between the distance from the testing sample to the-high-density-set of the nearest class different from the predicted class, and the distance from the test sample to the high-density-set of the class predicted.

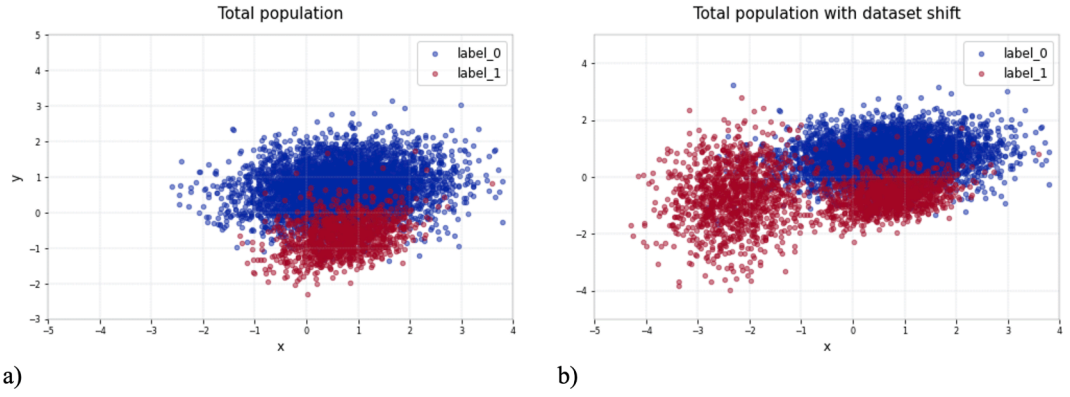


Fig. 4. a) Simulated binary dataset b) Simulated binary dataset with dataset shift.

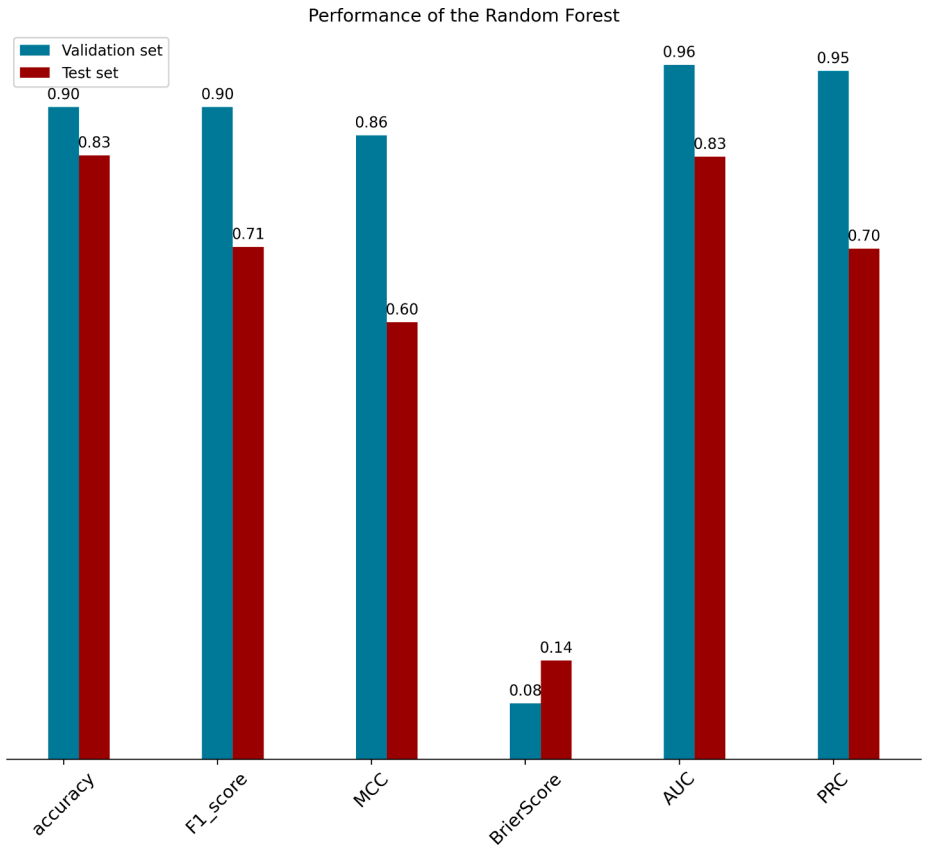


Fig. 5. Performances of the RF (100 trees) on the validation (IID) and test set (OOD and IID), in terms of accuracy, F1 score, Matthews Correlation Coefficient, Brier Score, Area Under The Receiving Operating Curve (AUC) and Area Under the Precision Recall Curve (PRC).

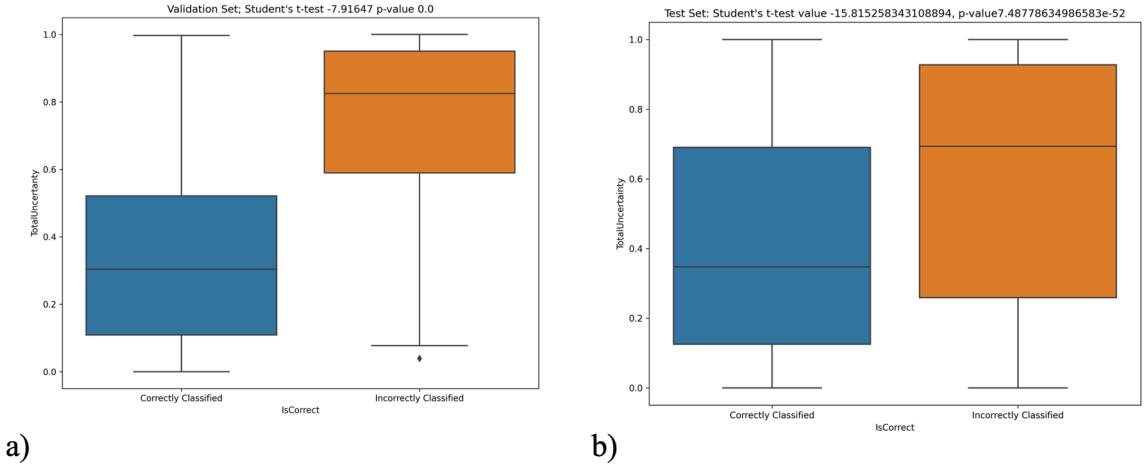


Fig. 6. a) Boxplots of the RF total uncertainty on correctly classified examples and incorrectly classified examples in the validation set b) Boxplots of the RF total uncertainty on correctly classified examples and incorrectly classified examples in the test set.

3. Reliability estimation following the density and local fit principles

Previous work has shown that the usage of "classifier-related" metrics (such as posterior predicted probability) to assess pointwise reliability may be biased and misleading given the intrinsic data-driven nature of the classifiers. Moreover, "a classifier may simply not be the best judge of its own trustworthiness" [45]. Motivated by this, we are interested in investigating the use of approaches for reliability estimation that are independent of the classifier chosen, and that does not rely on the classifier itself to compute its own reliability. To do so, we believe that the road to follow for reliability estimation should be based on the local and density principles defined by [76]. In the following, we present a possible strategy which is based on the implementation of these two principles, and we demonstrate the utility of this approach on a simulated and real-case medical dataset.

Given any type of classifier and the training set used to train the classifier, we first apply the instance selection-based approach proposed in [68]. Briefly, this approach first detects the "border" instances of the feature spaces, that is, for each attribute, we identify those training examples that are either inner or outer border for their membership class. When a new example will be classified, the "**density principle**"

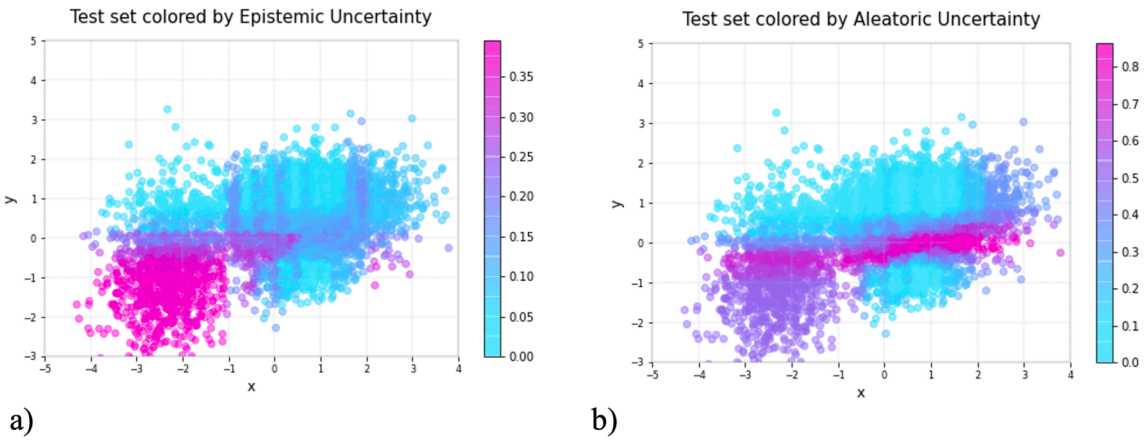


Fig. 7. a) Test samples colored by the epistemic uncertainty value. b) Test samples colored by the aleatoric uncertainty value.

will be evaluated by comparing the example to the training border, and its "density-based reliability" will be calculated by the simple formula:

$$drel(x) = 1 - \frac{m_x}{m}$$

where m_x is the number of attributes for which the value of the new example falls outside the training border and m is the total number of attributes. The example will be considered as reliable according to the density principle if the computed *drel* is greater than a given threshold. An example on a simple dataset with 2 attributes is shown in Fig. 3. More details can be found in [68].

After evaluating the density principle, we will evaluate the **local fit principle**, i.e. we will check whether the algorithm performs well on training samples close to the example. To do so, we will select the k nearest examples in the training set (or in a validation set coming from the same exact distribution of the training set), and we will calculate average performance metrics, such as accuracy or F1 score, on these neighbors. If the performance metrics are above a predefined threshold, we will evaluate the classifier as "trustful" on that example following the local fit principle. The threshold should be selected by the user, according to the desired and realistic performance that can be achieved within the specific problem. Eventually, the example will be considered as reliable if it is reliable according to the density principle and to the local fit principle.

We then compare this approach to an uncertainty-based estimation, by using a Random Forest (RF). RFs are widely used also in clinical application since they have shown high performances in different prediction tasks [79,79]. In addition, it has been recently shown how it is possible to compute prediction uncertainty as the entropy of the predicted posterior probabilities as reported in [82]. For continuous measure of reliability (e.g. uncertainty and density based-reliability), the optimal threshold to label an example as "reliable" or "unreliable" can be selected as follows: for different thresholds, we will compute the entropy of the reliable and unreliable sets with respect to the label "correctly classified" and "incorrectly classified" by the classifier. The threshold with lowest entropy will be selected. By doing so, we select the threshold that is more able to distinguish correctly classified and incorrectly classified examples. In particular, we will train a RF with 100 trees and at least 10 samples per leaf. We apply the two approaches for pointwise reliability estimation to detect incorrectly classified examples on a simulated dataset and on MIMIC-III, a widely used medical dataset which collects de-identified health data associated with thousands of intensive care unit (ICU) admissions [46]. Code is available as Google Colaboratory notebooks (https://github.com/GiovannaNicora/reliability).

3.1. Simulated dataset

We simulate a binary classification problem under dataset shift. First, we generate normally distributed samples (std = 1) with 2 features and classes overlaps. Each class is represented by a cluster in the 2d space, as reported in Fig. 4a. The dataset is made of 6000 samples, 80% of which in class 0 and the remaining in class 1. This dataset represents the true underlying population. We then want to simulate 1) that the available class distribution in the training set is different from the true population and 2) that dataset shift occurs. Therefore, we selected about 800 samples for training and about 300 for validation. In both training and validation, samples are equally distributed in the two classes. The remaining samples are kept for final testing. We then simulate dataset shift of the red class by shifting the "red samples" along the x axis and by adding random noise to the y axis (Fig. 4b). The shifted samples will be part of the final test set. In this context, the validation set represents an IID population, while the test set will gather both IID and OOD instances. The validation and the test set will only be used to evaluate the performance of the classifier, as well as reliability estimation, and they will not be used for hyperparameter tuning or model selection. The validation set will be used to select thresholds for reliability based on uncertainty and based on the density principle, while the threshold for local fit principle was empirically set to local accuracy equal or higher than 0.75. For additional information about the experiments, code is available (https://github.com/GiovannaNicora/reliability). After developing our ML classifier, errors can occur for different reasons: on one hand, the classifier may be wrong when there is overlap between the feature values of the two different classes (*aleatoric uncertainty*) and when the dataset shift occurs (*epistemic uncertainty*). To detect reliable and unreliable examples, we will apply and compare the approaches detailed in previous sections: reliability will be evaluated (1) in terms of local fit and density principles, (2) in terms of the predictive uncertainty.

3.2. MIMIC-III dataset

The MIMIC-III dataset is a freely available resource that contains clinical data and vital signs of thousands of patients in ICU. In particular, we used the preprocessed dataset made available by the PhysioNet 2012 challenge [83]. We are interested in developing a model to predict in-hospital death from clinical data. Moreover, we will examine whether bias in the training set will affect test performance and if reliability assessment can support the user to identify such unreliable predictions.

After removing features with at least 90% of missing values and patients with at least a missing value, we obtain a dataset of 4480 patients that survived ("In-Hospital death" = 0) and 768 that died in the hospital ("In-Hospital death"=1). More details about the features set can be found in [83] and in https://physionet.org/content/challenge-2012/

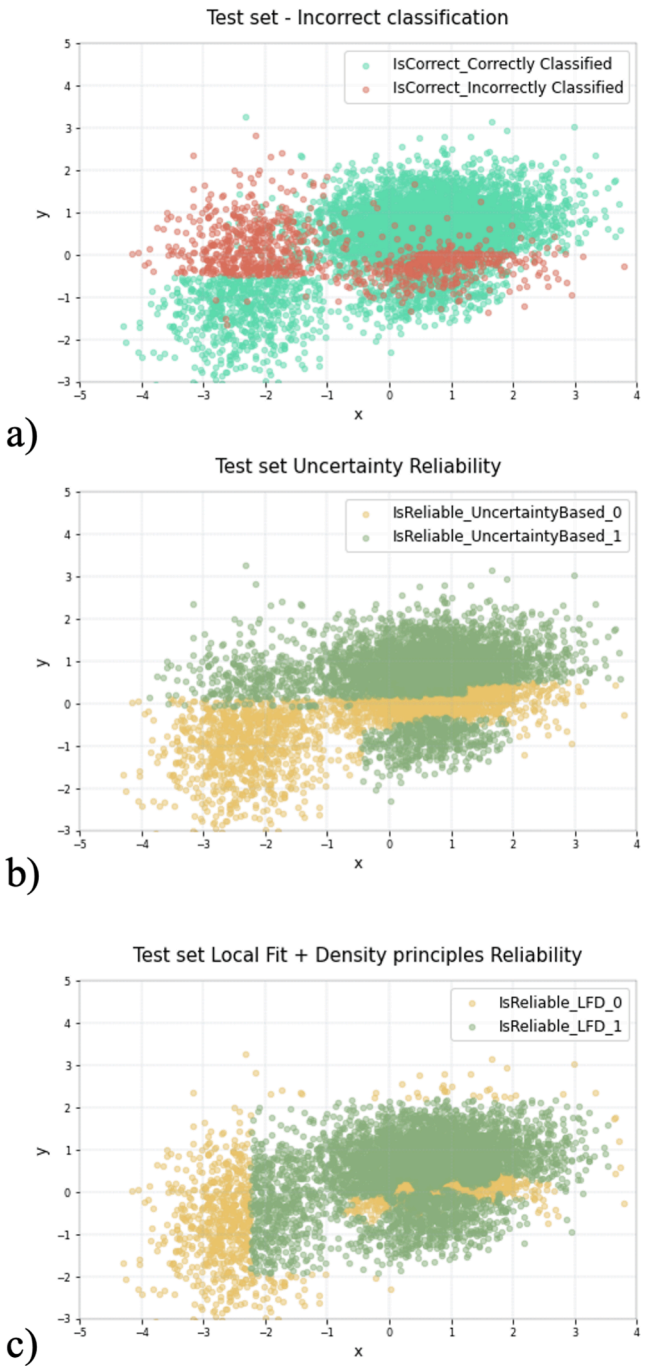


Fig. 8. a) Test samples colored in red if incorrectly classified, colored in green if correctly classified. b) Reliable and unreliable samples according to the total uncertainty. c) Reliable and unreliable samples according to the density and local fit principles. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

1.0.0/. In this case, we would like to simulate the situation in which some sub-populations of patients are not represented, or they are under-represented in the training set. To do so, we simulated the extreme case in which only male patients are available in the training set. We therefore select 70% of male patients for training and validation (that will be exploited for reliability and uncertainty thresholds selection), while the remaining, plus the female cohort, will be used for testing. The i.d,d set (1716 male patients in class 0 (alive) and 277 male patients in class 1 (deceased)) will be divided into 70% training and 30% validation. The test set will be made of more than 2273 female patients and 893 male patients. Therefore, in the test set, around 72% of patients are female. In the training set, 14% of patients are in class 1, while in the test the percentage of patients in class 1 is higher (23%). We select as a classifier a Random Forest, which allows us to compute uncertainty, as explained in the previous section. Best thresholds for uncertainty and density-based reliability will be computed on the results of the predictions made on the validation (i.i.d.) set. Also in this case, as in the simulated dataset, the validation set is similar to the training set. We carried out an additional experiment on the MIMIC dataset, reported in the Supplementary Material, where the data shift is simulated using age groups. In this case, we exploited as classifier a Lasso logistic regression, and we used the predicted posterior probability as uncertainty estimation (see Supplementary Material).

4. Results

4.1. Simulated data

Fig. 5 shows the performance of a RF on the simulated validation set (which is identically distributed, IID) and on the test (which also contains out-of-distributed, i.e. OOD samples).

As expected, the performances on the IID validation set are high, while all the metrics, including calibration and AUC, decrease when the properties of the dataset are different from the training set. In this case, we have both imbalanced data and OOD data.

Aleatoric, epistemic and total uncertainties are calculated on each validation and test samples. As we can see in the boxplot of Fig. 6a, the total uncertainty of the correctly identified validation samples are distributed around lower values in comparison to the incorrectly classified validation samples, and the uncertainty means are statistically different according to a *t*-test result. In Fig. 6b we can observe that, for the test set, the distributions of correctly and incorrectly classified examples are more similar, even though their means are significantly different.

By focusing on the test samples, we can see how the aleatoric and epistemic uncertainty varies in the feature space (Fig. 7).

These results confirm that the epistemic uncertainty is higher for the OOD samples, while the aleatoric uncertainty is capturing the noise when the two classes are overlapping.

Fig. 8 compares the correctly and incorrectly classified samples with the reliable and unreliable examples as labeled by the uncertainty method and by the combination of density and local fit principles. Ideally, the incorrectly classified examples should be labeled as unreliable.

As we can see, the RF fails to correctly classify IID samples that fall in the feature space where there is an overlap of the two classes. The reliability based on uncertainty estimation, as well as the reliability based on the local fit and density principle, are able to label as "unreliable" these samples. Regarding the OOD samples, which all belong to the class 1, only those points whose y value is roughly higher than −0.5 are incorrectly classified. On these shifted data, the uncertainty estimation labels as "reliable" the majority of misclassified points while samples in the area with the majority of correctly classified points are labeled as "unreliable". Therefore, the uncertainty estimation seems to be less trustworthy on shifted data. Instead, the approach based on the local fit and density principles labels as "reliable" those OOD samples which are nearer to the IID clusters along the x axis.

We then investigated whether the performances of the classifier on the reliable and unreliable sets differ (Fig. 9). Ideally, the reliable set should consist of samples for which the classification is more trustworthy. Therefore, performance metrics on the reliable sets should be higher. All the performance metrics, except for the area under the precision-recall curve and the F1 score, decrease in the unreliable set in comparison to the reliable one. The performance decrease of the unreliable set is higher when using the reliability estimation with local fit and density principles. For instance, the Matthews Correlation

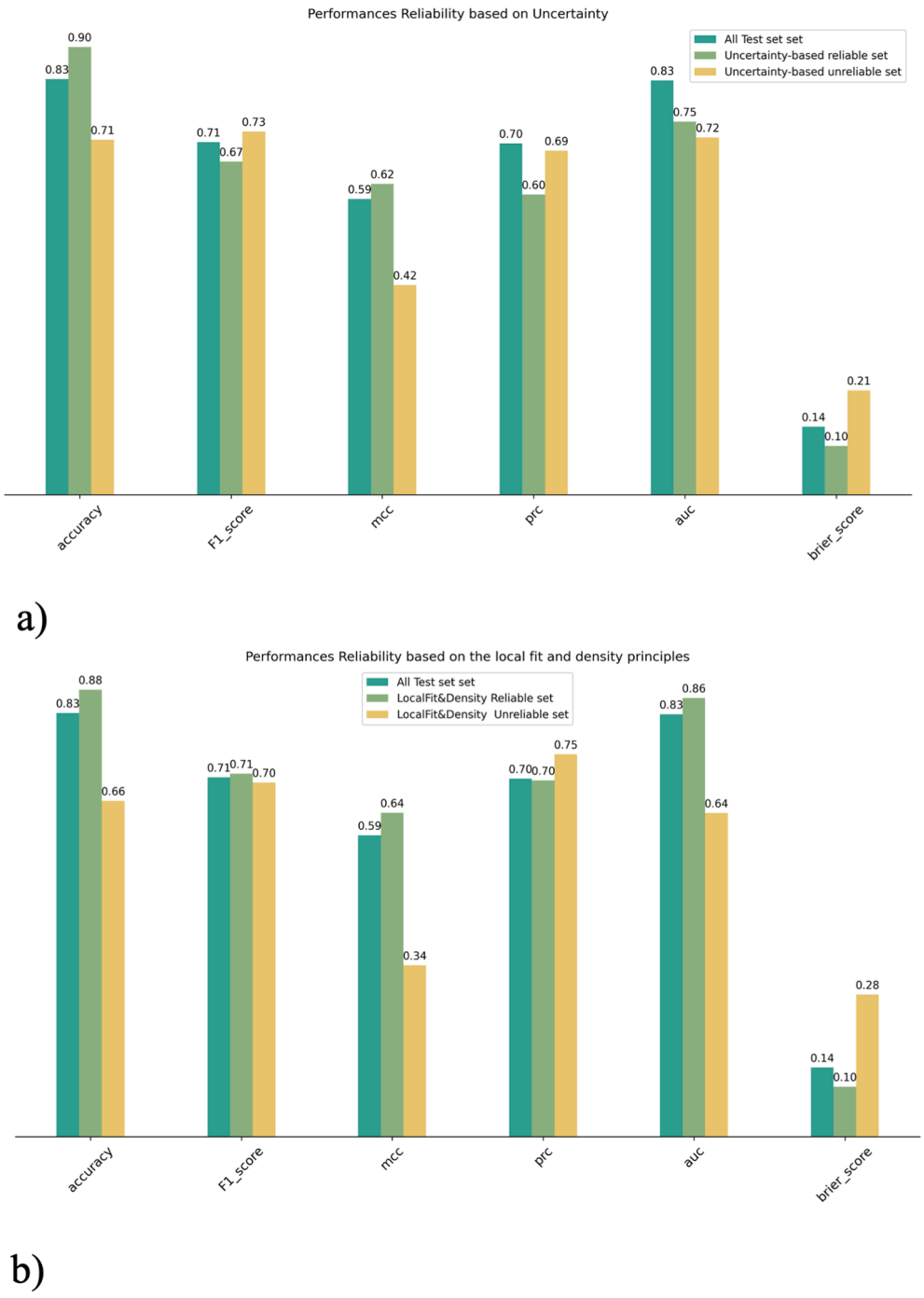


Fig. 9. a) Performance of the classifier on the entire test set, on the reliable set and on the unreliable set identified through the uncertainty estimation. b) Performance of the classifier on the entire test set, on the reliable set and on the unreliable set identified through the application of the local fit and density principles.

coefficient for the unreliable set according to the uncertainty method is 42%, while it drops to 34% for the unreliable set detected with the local fit and density principle. The AUC of the reliable and unreliable sets according to the uncertainty are comparable (75% vs 72%), while their difference is higher with the second method (86% vs 64%). Therefore, in this simulated case a reliability estimation based on local fit and density principles seems to perform better than the uncertainty-based estimation in the identification of reliable and unreliable instances.

4.2. MIMIC dataset

Fig. 10 reports the results of a RF trained on male patients.

The performance on the test set, both on the IID and the OOD sets, are lower than those reported on the training set through a 5-fold cross validation. Yet, on the IID set, the results are slightly higher in comparison to the OOD set (female patients). We then calculated the uncertainty and the reliability to evaluate whether we would be able to identify a subset of patients with a higher chance of being correctly classified. By analyzing the performance metrics computed on the reliable and unreliable sets according to the uncertainty (Fig. 11a and Table 3), for this particular dataset, the uncertainty-based method did not label any true positive example as reliable. On the contrary, the reliability based on the local fit and density principles labels all the false negative and false positive samples in the unreliable sets, while the

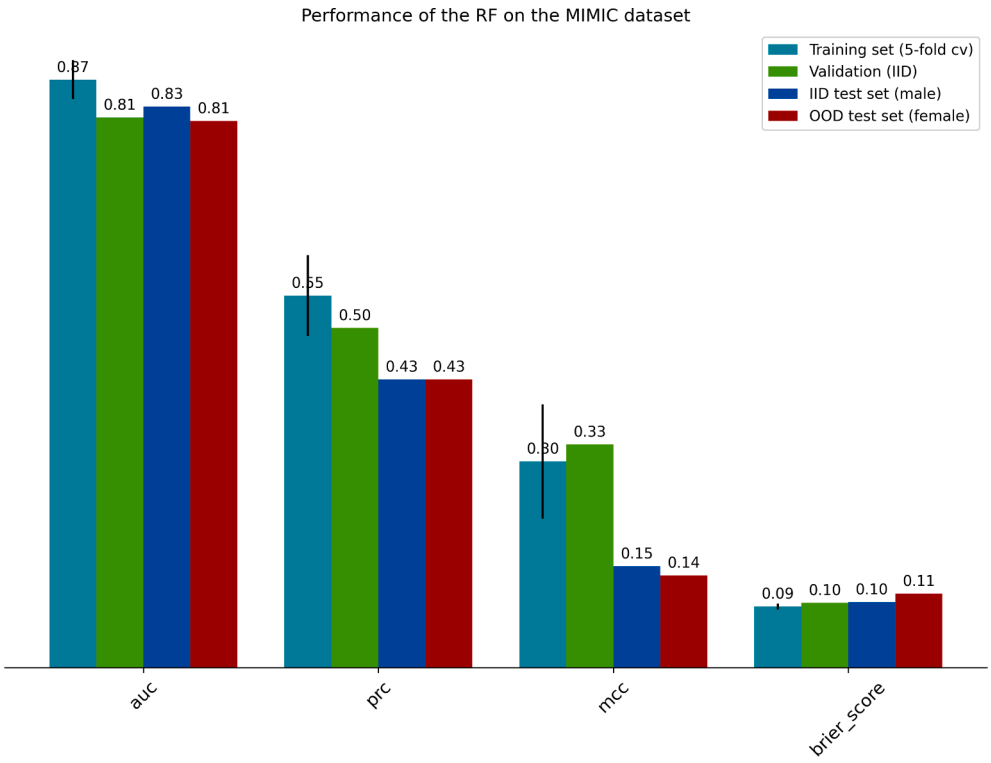


Fig. 10. Performance of the classifier on the training set (5-fold cross validation mean and standard deviation), on the validation set and on the IID test set and on OOD test set.

reliable set is entirely made of correctly classified examples (Fig. 11b and Table 4). Some correctly classified samples are labeled as unreliable. Reliability estimation in a region informs us about the distance from the training set, independently of the labels. This implies that reliability estimation provides a warning sign about the machine learning prediction by computing how far a particular sample is "distant" from the training set and/or whether it falls in a region where a high number of errors occur. Still, the model can correctly classify unreliable samples, even if this should occur at a lower rate.

According to the reliability computed with the local fit and density principles, 45% of the test set is considered reliable (70% females) and 55% unreliable (73% females). The percentage of reliable predictions on female patients is 44%, while 47% of predictions on male patients was considered reliable. If we consider the reliable and unreliable sets identified with uncertainty, we see that 67% of the test set is considered reliable (71% females) and 33% unreliable (73% females). The percentage of predictions supposed to be reliable on female patients is 66%, while 70% of predictions on male patients was considered reliable. From these results, we can observe that the strategies behave as expected, and that, while both are able to highlight that predictions on females are less reliable than one on males, local fit and density principles are able to label a higher percentage of predictions on o.o.d. samples (odds ratio 1.144) as unreliable than uncertainty (odds ratio 1.135).

5. Discussion

ML applications to medical-related problems need not only to show high performance during validation, but they also need to demonstrate their trustworthiness throughout the entire deployment phase. Intrinsic trust can be achieved through explainable methods that pinpoint the reasoning process performed by black box predictive models. Reliability monitoring is as crucial as explainability. Our claim is that the implementation of reliability monitoring mechanisms should be a requirement of any ML applications to achieve trust. Reliability is also a requirement of the recent guidelines from the European Commision for the development of trustworthy AI systems (https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai). In such guidelines, it is stated that "*a reliable AI system is one that works properly with a range of inputs and in a range of situations*". We need therefore to define (1) when a system is considered to work properly and (2) what is the acceptable range of inputs and situations. In this paper we concentrate our attention to point-wise prediction reliability, which is the degree of trust that the predicted class is equal to the true class of a single instance. In other words, it is the degree of trust that the classification is correct. Reliability can be judged by analyzing whether the instance to be classified falls into a "trustworthy" region of the feature space, where the classifier should be confident and sure about the prediction. This region contains samples coming from the same distribution of the training data (i.i.d. samples) and for which the classifier showed "good fit", for instance high accuracy. Other approaches to estimate reliability rely on uncertainty estimation, or using particular types of classifiers, such as conformal methods.

Yet, incorporating reliability estimation within a ML pipeline also adds computational complexity. For instance, when the reliability is computed by comparing the training set to each new instance to be classified, the user needs to have access to the training data. Access to training data can be unfeasible or even impossible, due to the dimensionality of the training set or to privacy issues.

A well-known paradigm to deal with these issues is represented by instance selection methods, which select only the most informative training examples to be stored. We suggest using such methods for reliability assessment. In this case, instance selection methods choose the most informative training samples that will be compared with new instances for reliability assessment. The results may depend on the specific method used to evaluate if the example of the test set is OOD.

In our paper, we have presented an approach that is even more parsimonious, since it performs OOD reliability assessment by defining boundaries attribute-by-attribute, both in the original or in a transformed feature space. In this case, the additional information to be retained can be compactly represented by O(m) number of variables.

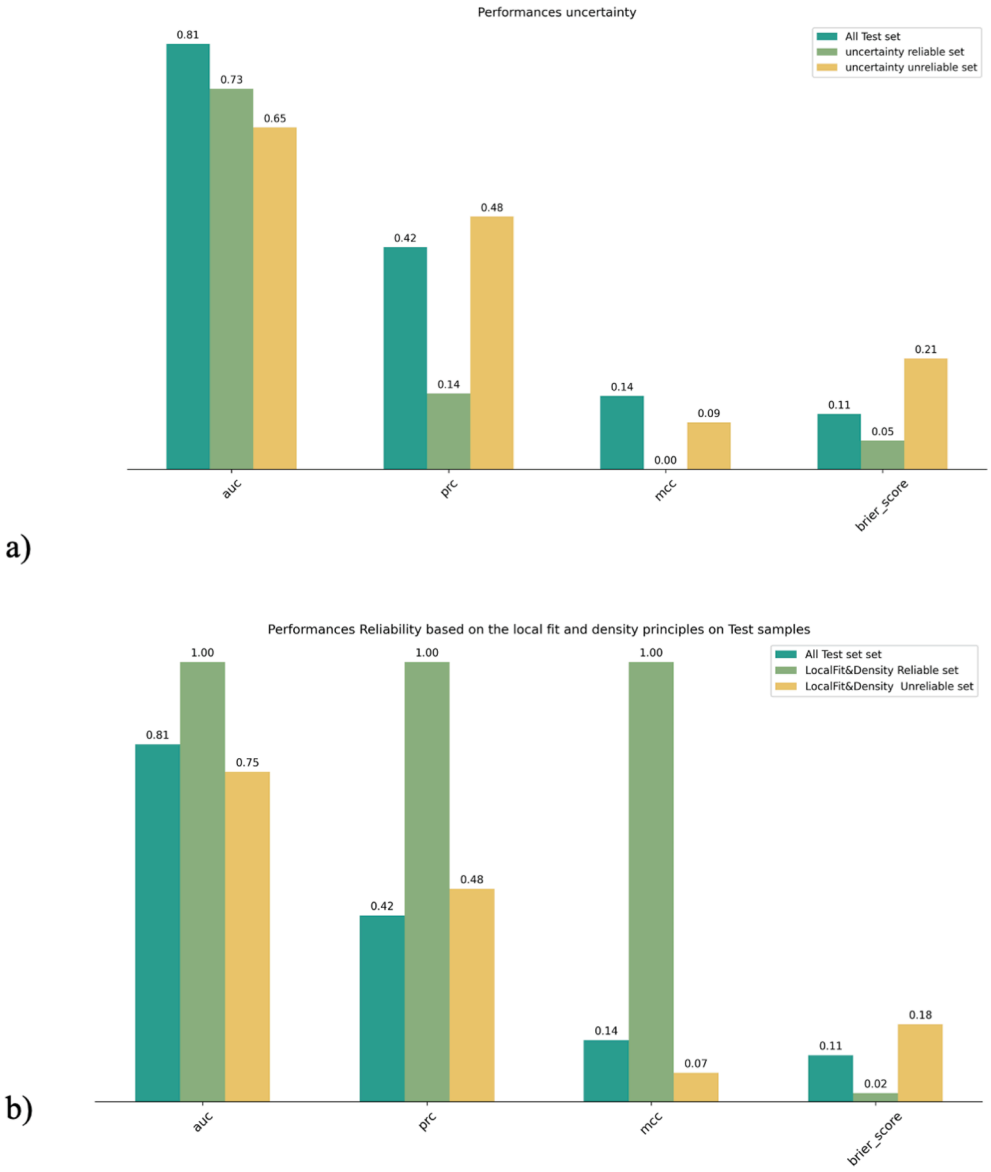


Fig. 11. a) Performance of the classifier on the entire test set, the reliable and unreliable set according to the value of uncertainty. b) Performance of the classifier on the entire test set, the reliable and unreliable set according to the local fit and density principles.

Table 3
Confusion matrix of the test set, the reliable set and the unreliable set detected with the local fit and density-based method.

	TN	TP	FP	FN
All test set	2658	31	30	447
Reliable set	1416	8	0	0
Unreliable set	1242	23	30	437

Table 4
Confusion matrix of the test set, the reliable set and the unreliable set detected with the uncertainty method.

	TN	TP	FP	FN
All test set	2658	31	30	447
Reliable set	2010	0	0	129
Unreliable set	648	31	30	318

Despite its simplicity, our approach has some limitations. First of all, by detecting border examples attribute-by-attribute, we do not take into account possible correlations between features. Secondly, the results of reliability assessment may depend on parameters that are tuned by the user, such as the threshold for reliable/unreliable labelling, the number of neighbors and the distance function chosen to select those neighbors within the application of the local fit principle. When applying the density principle, the values of the features in the training set are used to select the borders. Before testing, users can select a threshold for reliability, by either using another dataset (which we called "validation set" in our experiments), or, in absence of a third dataset, by choosing a predefined threshold. We showed how to perform reliability assessment on a simple simulated dataset and on the well-known medical dataset MIMIC. For illustrative purposes, in the first simulated binary dataset of two dimensions, we added a shift that affects only one of the two classes i.e. the positive class (red samples). In medical applications, this may happen for instance, when interventions are taken to treat a disease. Consider for example an application where the aim is to detect cancer cells in a patient based on single-cell gene expression profiles. Suppose that the classifier is trained on data coming from patients before

treatment. If a new treatment, targeting cancer cells only, is available, such treatment can modify the gene expression profiles of the targeted cancer cells before they die. Therefore, we may observe a shift in gene expression features that affect only the positive class. Yet, in the majority of applications the shift will affect the entire population, regardless of the class. The investigation of different types and combinations of covariate shifts is an important topic. Another possible approach can be to simulate a dataset shift that corresponds to a change of the very nature of the distribution itself. An extensive benchmarking of this method applied to different Machine Learning classifiers, along with the previously published approaches mentioned in the Background sections, will be performed as future work. Additionally, the qualitative evaluation of the different methodologies that can compute a form of a reliability measure is complex, and standard procedures to obtain these measures are missing to the best of our knowledge. In our results, we computed different metrics on the reliable and unreliable sets to qualitatively evaluate whether reliable predictions were associated with better performance metrics. This method implies that the ground truth of test data is available. In the scenario in which reliability assessment is deployed within a machine learning pipeline, and it is applied to new prediction (s), the ground truth may not be available. Therefore, it may be impossible to monitor the reliability performance as we reported in the paper, and new metrics should be developed. Yet, it is when the ground truth is not available (during deployment), that an approach for reliability estimation can guide and support the user, by "raising an alarm" when predictions are labeled as unreliable. Reliability may also support monitoring, by highlighting possible shifts and bias in the data. Coupled with XAI, reliability can promote trustworthiness in a model's prediction. While local XAI can show the relationship between feature values and the prediction on a single instance in an interpretable way, reliability assures that the model is being used as intended, i.e. on data coming from a region of the features space where the classifier should be confident.

An interesting, improved approach is to resort to generative models. Such models can be flexibly used to perform data distribution assessment. By capturing the statistical properties of the training data, they can be either used to compute goodness of fit measures or to generate synthetic data [91] that can be used to substitute real data during reliability assessment.

6. Conclusion

We formally defined different metrics for reliability estimation and comprehensively reported different methodologies that have been used to evaluate reliability, not only in healthcare applications.

We then showed a simple approach to include a reliability estimation to any type of classifier. Such reliability is computed following the local fit and density principles defined by Saria et al. [76]. We compared this method to reliability according to the level of the uncertainty, and we investigated which of the two methodologies is best suited for the identification of incorrectly classified examples, both on a simulated dataset and a clinical dataset. In these two experiments, the reliability based on the local fit and density principles seems to perform better in the identification of samples for which the classifier is correct at a higher rate. We also compared the local fit and density principles with reliability based on the predicted posterior probability. Our results confirm that the predicted posterior probability as a measure of reliability of the classification can be less useful under dataset shift (see Supplementary Material). We conclude that the usage of "classifier-related" metrics (such as posterior predicted probability, uncertainty or conformal prediction) to assess pointwise reliability may be biased and misleading given the intrinsic data-driven nature of the classifier. Reliability estimation should evaluate the classifier's output independently of the classifier itself. However, additional experiments should be carried out on different types of both simulated and medical datasets for further evaluation.

CRediT authorship contribution statement

Giovanna Nicora: Methodology, Formal analysis, Conceptualization, Writing – original draft. **Miguel Rios:** Conceptualization, Writing – review & editing. **Ameen Abu-Hanna:** Conceptualization, Writing – review & editing, Supervision. **Riccardo Bellazzi:** Methodology, Conceptualization, Writing – review & editing, Supervision.

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.

Acknowledgements

We thank the anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions. This work was partially supported by the Department of Electrical, Computer and Biomedical Engineering of University of Pavia and by the European Commission as part of the PERISCOPE project (Grant Agreement 101016233), coordinated by the University of Pavia.

References

[1] M.R. Abbas, M.S.A. Nadeem, A. Shaheen, A.A. Alshdadi, R. Alharbey, S.-O. Shim, W. Aziz, Accuracy Rejection Normalized-Cost Curves (ARNCCs): A Novel 3-Dimensional Framework for Robust Classification, IEEE Access 7 (2019) 160125–160143, https://doi.org/10.1109/ACCESS.2019.2950244.
[2] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, et al., A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges, 2021. ArXiv:2011.06225 [Cs], January. http://arxiv.org/abs/2011.06225.
[3] A. Ahmadi, S. Omatu, T. Fujinaka, T. Kosaka, Improvement of Reliability in Banknote Classification Using Reject Option and Local PCA, Inf. Sci. 168 (1) (2004) 277–293, https://doi.org/10.1016/j.ins.2004.02.018.
[4] A. Alimadadi, S. Aryal, I. Manandhar, P.B. Munroe, B. Joe, X. Cheng, Artificial Intelligence and Machine Learning to Fight COVID-19, Physiol. Genomics 52 (4) (2020) 200–202, https://doi.org/10.1152/physiolgenomics.00029.2020.
[5] N. Alirezaie, K.D. Kernohan, T. Hartley, J. Majewski, T.D. Hocking, ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants, Am. J. Human Genet. 103 (4) (2018) 474–483, https://doi.org/10.1016/j.ajhg.2018.08.005.
[6] P.L. Bartlett, M.H. Wegkamp, Classification with a Reject Option Using a Hinge Loss, J. Machine Learn. Res. 9 (59) (2008) 1823–1840.
[7] S. Benjamens, P. Dhunnoo, B. Meskó, The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database, NPJ Digit Med. 11 (3) (2020) 118, https://doi.org/10.1038/s41746-020-00324-0.
[8] A. Benso, S. Di Carlo, G. Politano, A. Savino, H. Hafeezurrehman, Building Gene Expression Profile Classifiers with a Simple and Efficient Rejection Option in R, BMC Bioinf. 12 (Suppl 13) (2011) S3, https://doi.org/10.1186/1471-2105-12-S13-S3.
[9] Z. Bosnić, I. Kononenko, Estimation of Individual Prediction Reliability Using the Local Sensitivity Analysis, Appl. Intell. 29 (3) (2008) 187–203, https://doi.org/10.1007/s10489-007-0084-9.
[10] Z. Bosnić, I. Kononenko, An Overview of Advances in Reliability Estimation of Individual Predictions in Machine Learning, Intell. Data Anal. 13 (2) (2009) 385–401.
[11] J. Brinkrolf, B. Hammer, Interpretable machine learning with reject option, Automatisierungstechnik 66 (4) (2018) 283–290, https://doi.org/10.1515/auto-2017-0123.
[12] I. Buzhinsky, A. Nerinovsky, S. Tripakis, Metrics and Methods for Robustness Evaluation of Neural Networks with Generative Models'. ArXiv:2003.01993 [Cs, Stat], 2020 March, http://arxiv.org/abs/2003.01993.
[13] H. Choi, D. Yeo, S. Kwon, Y. Kim, Gene Selection and Prediction for Cancer Classification Using Support Vector Machines with a Reject Option, Comput. Stat. Data Anal. 55 (5) (2011) 1897–1908, https://doi.org/10.1016/j.csda.2010.12.001.
[14] C. Chow, On Optimum Recognition Error and Reject Tradeoff, IEEE Trans. Inf. Theory 16 (1) (1970) 41–46, https://doi.org/10.1109/TIT.1970.1054406.
[15] F. Condessa, J. Bioucas-Dias, C.A. Castro, J.A. Ozolek, J. Kovačević, Classification with Reject Option Using Contextual Information, in: 2013 IEEE 10th International Symposium on Biomedical Imaging, 2013, pp. 1340–1343, https://doi.org/10.1109/ISBI.2013.6556780.
[16] C. Corbière, N. Thome, A. Bar-Hen, M. Cord, P. Pérez, Addressing Failure Prediction by Learning Model Confidence, ArXiv:1910.04851 [Cs, Stat], 2019 October. http://arxiv.org/abs/1910.04851.
[17] L.P. Cordella, C. De Stefano, C. Sansone, M. Vento, An Adaptive Reject Option for LVQ Classifiers, in: C. Braccini, L. DeFloriani, G. Vernazza (Eds.), Image Analysis

and Processing. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 1995, pp. 68–73, https://doi.org/10.1007/3-540-60298-4_238.
[18] L.P. Cordella, C. De Stefano, F. Tortorella, M. Vento, A Method for Improving Classification Reliability of Multilayer Perceptrons, IEEE Trans. Neural Networks 6 (5) (1995) 1140–1147, https://doi.org/10.1109/72.410358.
[19] I. Cortés-Ciriano, A. Bender, Deep Confidence: A Computationally Efficient Framework for Calculating Reliable Prediction Errors for Deep Neural Networks, J. Chem. Informat. Model. 59 (3) (2019) 1269–1281, https://doi.org/10.1021/acs.jcim.8b00542.
[20] I. Cortés-Ciriano, A. Bender, Concepts and Applications of Conformal Prediction in Computational Drug Discovery, ArXiv:1908.03569 [Cs, q-Bio], 2019b August, http://arxiv.org/abs/1908.03569.
[21] C.M. Cutillo, K.R. Sharma, L. Foschini, S. Kundu, M. Mackintosh, K.D. Mandl, Machine Intelligence in Healthcare—Perspectives on Trustworthiness, Explainability, Usability, and Transparency, Npj Digital Medicine 3 (1) (2020) 1–5, https://doi.org/10.1038/s41746-020-0254-2.
[22] S.E. Davis, Stabilizing Calibration of Clinical Prediction Models in Non-Stationary Environments: Methods Supporting Data-Driven Model Updating, 2019 October. https://ir.vanderbilt.edu/handle/1803/14327.
[23] Z. Dlamini, F.Z. Francies, R. Hull, R. Marima, Artificial Intelligence (AI) and Big Data in Cancer and Precision Oncology, Comput. Struct. Biotechnol. J. (2020), https://doi.org/10.1016/j.csbj.2020.08.019.
[24] G.F. Elsayed, I. Goodfellow, J. Sohl-Dickstein, Adversarial Reprogramming of Neural Networks'. ArXiv:1806.11146 [Cs, Stat], 2018 November. http://arxiv.org/abs/1806.11146.
[25] S.G. Finlayson, J.D. Bowers, J. Ito, J.L. Zittrain, A.L. Beam, I.S. Kohane, Adversarial Attacks on Medical Machine Learning, Science 363 (6433) (2019) 1287–1289, https://doi.org/10.1126/science.aaw4399.
[26] L. Fischer, L. Ehrlinger, V. Geist, R. Ramler, F. Sobieczky, W. Zellinger, B. Moser, Applying AI in Practice: Key Challenges and Lessons Learned, in: A. Holzinger, P. Kieseberg, A. Min Tjoa, E. Weippl (Eds.), Machine Learning and Knowledge Extraction. Lecture Notes in Computer Science, Springer International Publishing, Cham, 2020, pp. 451–471, https://doi.org/10.1007/978-3-030-57321-8_25.
[27] G. Fumera, I. Pillai, F. Roli, Classification with Reject Option in Text Categorisation Systems, in: 12th International Conference on Image Analysis and Processing, 2003 Proceedings, 2003, pp. 582–587, https://doi.org/10.1109/ICIAP.2003.1234113.
[28] G. Fumera, F. Roli, Support Vector Machines with Embedded Reject Option, in: S.-W. Lee, A. Verri (Eds.), Pattern Recognition with Support Vector Machines. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2002, pp. 68–82, https://doi.org/10.1007/3-540-45665-1_6.
[29] G. Sousa, A.R. Ricardo, R. Neto, J.S. Cardoso, G.A. Barreto, Robust Classification with Reject Option Using the Self-Organizing Map, Neural Comput. Appl. 26 (7) (2015) 1603–1619, https://doi.org/10.1007/s00521-015-1822-2.
[30] J. Gao, J. Yao, Y. Shao, Towards Reliable Learning for High Stakes Applications, Proc. AAAI Conf. Artif. Intell. 33 (01) (2019) 3614–3621, https://doi.org/10.1609/aaai.v33i01.33013614.
[31] Y. Geifman, R. El-Yaniv, SelectiveNet: A Deep Neural Network with an Integrated Reject Option, ArXiv:1901.09192 [Cs, Stat], June 2019, http://arxiv.org/abs/1901.09192.
[32] H. Ghoddusi, G.G. Creamer, N. Rafizadeh, Machine Learning in Energy Economics and Finance: A Review, Energy Econ. 81 (June) (2019) 709–727, https://doi.org/10.1016/j.eneco.2019.05.006.
[33] F.K. Hamey, B. Göttgens, Machine Learning Predicts Putative Hematopoietic Stem Cells within Large Single-Cell Transcriptomics Data Sets, Exp. Hematol 78 (2019) 11–20, https://doi.org/10.1016/j.exphem.2019.08.009.
[34] B. Hanczar, E.R. Dougherty, Classification with Reject Option in Gene Expression Data, Bioinformatics 24 (17) (2008) 1889–1895, https://doi.org/10.1093/bioinformatics/btn349.
[35] B. Hanczar, M. Sebag, Combination of One-Class Support Vector Machines for Classification with Reject Option, in: T. Calders, F. Esposito, E. Hüllermeier, R. Meo (Eds.), Machine Learning and Knowledge Discovery in Databases. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2014, pp. 547–562, https://doi.org/10.1007/978-3-662-44848-9_35.
[36] Y. Hechtlinger, B. Póczos, L. Wasserman, Cautious Deep Learning, ArXiv:1805.09460 [Cs, Stat], February 2019, http://arxiv.org/abs/1805.09460.
[37] M.E. Hellman, The Nearest Neighbor Classification Rule with a Reject Option, IEEE Trans. Syst. Sci. Cybernet. 6 (3) (1970) 179–185, https://doi.org/10.1109/TSSC.1970.300339.
[38] D. Hendrycks, K. Gimpel, A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. ArXiv:1610.02136 [Cs], October 2018, http://arxiv.org/abs/1610.02136.
[39] B. Hie, B.D. Bryson, B. Berger, Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design, Cell Syst. 11 (5) (2020) 461–477.e9, https://doi.org/10.1016/j.cels.2020.09.007.
[40] E. Hüllermeier, W. Waegeman, Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods, Machine Learn. 110 (3) (2021) 457–506, https://doi.org/10.1007/s10994-021-05946-3.
[41] E.J. Hwang, S. Park, K.-N. Jin, J.I. Kim, S.Y. Choi, J.H. Lee, J.M. Goo, et al., Development and Validation of a Deep Learning-Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs, JAMA Network Open 2 (3) (2019), e191095, https://doi.org/10.1001/jamanetworkopen.2019.1095.
[42] A. Jacovi, A. Marasović, T. Miller, Y. Goldberg, Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. ArXiv:2010.07487 [Cs], January 2021, http://arxiv.org/abs/2010.07487.
[43] L.A. Jeni, J.F. Cohn, F. De La Torre, Facing Imbalanced Data Recommendations for the Use of Performance Metrics, in: International Conference on Affective Computing and Intelligent Interaction and Workshops : [Proceedings]. ACII (Conference) 2013, 2013, pp. 245–251, https://doi.org/10.1109/ACII.2013.47.
[44] F.C. Jiang, Study on a Confidence Machine Learning Method Based on Ensemble Learning, Cluster Comput. 20 (4) (2017) 3357–3368, https://doi.org/10.1007/s10586-017-1085-z.
[45] H. Jiang, B. Kim, M.Y. Guan, M. Gupta, To Trust Or Not To Trust A Classifier. ArXiv:1805.11783 [Cs, Stat], October 2018, http://arxiv.org/abs/1805.11783.
[46] A.E.W. Johnson, T.J. Pollard, L.u. Shen, L.-W. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L.A. Celi, R.G. Mark, MIMIC-III, a Freely Accessible Critical Care Database, Sci. Data 3 (1) (2016) 160035, https://doi.org/10.1038/sdata.2016.35.
[47] J. Kang, C.D. Yoo, Learning of a Multi-Class Classifier with Rejection Option Using Sparse Representation, in: The 18th IEEE International Symposium on Consumer Electronics (ISCE 2014), 2014, pp. 1–2, https://doi.org/10.1109/ISCE.2014.6884541.
[48] S. Kang, S. Cho, S.-J. Rhee, Y. Kyung-Sang, Reliable Prediction of Anti-Diabetic Drug Failure Using a Reject Option, Pattern Anal. Appl. 20 (3) (2017) 883–891, https://doi.org/10.1007/s10044-016-0585-4.
[49] E. Kawaler, A. Cobian, P. Peissig, D. Cross, S. Yale, M. Craven, Learning to Predict Post-Hospitalization VTE Risk from EHR Data, AMIA Annual Symp. Proc. 2012 (November) (2012) 436–445.
[50] C.J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, D. King, Key Challenges for Delivering Clinical Impact with Artificial Intelligence, BMC Med. 17 (1) (2019) 195, https://doi.org/10.1186/s12916-019-1426-2.
[51] A. Kendall, Y. Gal, What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?' ArXiv:1703.04977 [Cs], October 2017, http://arxiv.org/abs/1703.04977.
[52] B. Kompa, J. Snoek, A.L. Beam, Second Opinion Needed: Communicating Uncertainty in Medical Machine Learning, Npj Digital Med. 4 (1) (2021) 1–6, https://doi.org/10.1038/s41746-020-00367-3.
[53] I. Kononenko, Machine Learning for Medical Diagnosis: History, State of the Art and Perspective, Artif. Intell. Med. 23 (1) (2001) 89–109, https://doi.org/10.1016/S0933-3657(01)00077-X.
[54] M. Kukar, I. Kononenko, Reliable Classifications with Machine Learning, in: Proceedings of the 13th European Conference on Machine Learning, ECML '02, Springer-Verlag, Berlin, Heidelberg, 2002, pp. 219–231.
[55] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. ArXiv:1612.01474 [Cs, Stat], 2017 November, http://arxiv.org/abs/1612.01474.
[56] C. Leibig, V. Allken, M.S. Ayhan, P. Berens, S. Wahl, Leveraging Uncertainty Information from Deep Neural Networks for Disease Detection, Sci. Rep. 7 (1) (2017) 17816, https://doi.org/10.1038/s41598-017-17876-z.
[57] J.A. Leonard, M.A. Kramer, L.H. Ungar, A Neural Network Architecture That Computes Its Own Reliability, Comput. Chem. Eng., Int. J. Comput. Appl. Chem. Eng. 16 (9) (1992) 819–835, https://doi.org/10.1016/0098-1354(92)80035-8.
[58] C.X. Ling, V.S. Sheng, C. Sammut, G.I. Webb, Cost-Sensitive LearningCost-Sensitive Learning, in: Encyclopedia of Machine Learning, Springer US., Boston, MA, 2010, pp. 231–235, https://doi.org/10.1007/978-0-387-30164-8_181.
[59] S. Malakouti, M. Hauskrecht, Predicting Patient's Diagnoses and Diagnostic Categories from Clinical-Events in EHR Data, in: D. Riaño, S. Wilk, A. ten Teije (Eds.), Artificial Intelligence in Medicine. Lecture Notes in Computer Science, Springer International Publishing, Cham, 2019, pp. 125–130, https://doi.org/10.1007/978-3-030-21642-9_17.
[60] L. Meijerink, G. Cinà, M. Tonutti, Uncertainty Estimation for Classification and Risk Prediction on Medical Tabular Data, ArXiv:2004.05824 [Cs, Stat], May 2020. http://arxiv.org/abs/2004.05824.
[61] D.P.P. Mesquita, L.S. Rocha, J.P.P. Gomes, A.R. Rocha Neto, Classification with Reject Option for Software Defect Prediction, Appl. Soft Comput. 49 (December) (2016) 1085–1093, https://doi.org/10.1016/j.asoc.2016.06.023.
[62] S. Messoudi, S. Rousseau, S. Destercke, Deep Conformal Prediction for Robust Models, Informat. Process. Manage. Uncertainty Knowledge-Based Syst. 1237 (May) (2020) 528–540, https://doi.org/10.1007/978-3-030-50146-4_39.
[63] S.J. Mooney, V. Pejaver, Big Data in Public Health: Terminology, Machine Learning, and Privacy, Annu. Rev. Public Health 39 (1) (2018) 95–112, https://doi.org/10.1146/annurev-publhealth-040617-014208.
[64] A.H. Murphy, What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting, Weather Forecasting 8 (2) (1993) 281–293, https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.
[65] K. Murphy, Probabilistic Machine Learning: An Introduction, Accessed 8 April 2021, n.d., https://probml.github.io/pml-book/book1.html.
[66] M.S. Nadeem, J.-D. Ahmed, B. Hanczar, Accuracy-Rejection Curves (ARCs) for Comparing Classification Methods with a Reject Option, in: Machine Learning Systems Biology, PMLR, 2009, pp. 65–81, in: http://proceedings.mlr.press/v8/nadeem10a.html.
[67] P.M. do Nascimento, I.G. Medeiros, R.M. Falcão, B. Stransky, J.E.S. de Souza, A Decision Tree to Improve Identification of Pathogenic Mutations in Clinical Practice, BMC Medical Informat. Decision Making 20 (1) (2020) 52, https://doi.org/10.1186/s12911-020-1060-0.
[68] G. Nicora, R. Bellazzi, A Reliable Machine Learning Approach Applied to Single-Cell Classification in Acute Myeloid Leukemia, AMIA Annual Symp. Proc. 2020 (January) (2021) 925–932.
[69] G. Nicora, S. Marini, I. Limongelli, E. Rizzo, S. Montoli, F.F. Tricomi, R. Bellazzi, A Semi-Supervised Learning Approach for Pan-Cancer Somatic Genomic Variant Classification, in: D. Riaño, S. Wilk, A. ten Teije (Eds.), Artificial Intelligence in Medicine. Lecture Notes in Computer Science, Springer International Publishing, Cham, 2019, pp. 42–46, https://doi.org/10.1007/978-3-030-21642-9_7.

[70] J.A. Olvera-López, J. Ariel Carrasco-Ochoa, J. Francisco Martínez-Trinidad, J. Kittler, A Review of Instance Selection Methods, Artif. Intell. Rev. 34 (2) (2010) 133–143, https://doi.org/10.1007/s10462-010-9165-y.
[71] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J.V. Dillon, B. Lakshminarayanan, J. Snoek, Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. arXiv preprint arXiv:1906.02530, 2019.
[72] A. Ozen, M. Gönen, E. Alpaydan, T. Haliloğlu, Machine Learning Integration for Predicting the Effect of Single Amino Acid Substitutions on Protein Stability, BMC Struct. Biol. 9 (October) (2009) 66, https://doi.org/10.1186/1472-6807-9-66.
[73] M. Panahiazar, V. Taslimitehrani, N. Pereira, J. Pathak, Using EHRs and Machine Learning for Heart Failure Survival Analysis, Stud. Health Technol. Informat. 216 (2015) 40–44.
[74] M.T. Ribeiro, S. Singh, C. Guestrin, Why Should I Trust You?": Explaining the Predictions of Any Classifier'. ArXiv:1602.04938 [Cs, Stat], August 2016, http://arxiv.org/abs/1602.04938.
[75] C.M. Santos-Pereira, A.M. Pires, On Optimal Reject Rules and ROC Curves, Pattern Recogn. Lett. 26 (7) (2005) 943–952, https://doi.org/10.1016/j.patrec.2004.09.042.
[76] S. Saria, A. Subbaswamy, Tutorial: Safe and Reliable Machine Learning. ArXiv: 1904.07204 [Cs], 2019 April. http://arxiv.org/abs/1904.07204.
[77] A. Sarica, A. Cerasa, A. Quattrone, Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review, Front. Aging Neurosci. 9 (2017) 329, https://doi.org/10.3389/fnagi.2017.00329.
[78] C. Saunders, A. Gammerman, V. Vovk, Transduction with Confidence and Credibility. Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2, 722–26. IJCAI'99, Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, 1999.
[79] M. Schinkel, K. Paranjape, R.S. Nannan Panday, N. Skyttberg, P.W.B. Nanayakkara, Clinical applications of artificial intelligence in sepsis: A narrative review, Comput. Biol. Med. 115 (2019) 103488, https://doi.org/10.1016/j.compbiomed.2019.103488.
[80] P. Schulam, S. Saria, Can You Trust This Prediction? Auditing Pointwise Reliability After Learning'. ArXiv:1901.00403 [Cs, Stat], 2019. February, http://arxiv.org/abs/1901.00403.
[81] G. Shafer, V. Vovk, A Tutorial on Conformal Prediction, J. Machine Learn. Res. 9 (12) (2008) 371–421.
[82] M.H. Shaker, E. Hüllermeier, Aleatoric and Epistemic Uncertainty with Random Forests, in: M.R. Berthold, A. Feelders, G. Krempl (Eds.), Advances in Intelligent Data Analysis XVIII. Lecture Notes in Computer Science, Springer International Publishing, Cham, 2020, pp. 444–456, https://doi.org/10.1007/978-3-030-44584-3_35.
[83] I. Silva, G. Moody, D.J. Scott, L.A. Celi, R.G. Mark, Predicting In-Hospital Mortality of ICU Patients: The PhysioNet/Computing in Cardiology Challenge 2012, Comput. Cardiol. 39 (2012) 245–248.
[84] R. Sousa, B. Mora, J.S. Cardoso, An Ordinal Data Method for the Classification with Reject Option, in: 2009 International Conference on Machine Learning and Applications, 2009, pp. 746–750, https://doi.org/10.1109/ICMLA.2009.11.
[85] R. Sousa, A.R. Neto, G. Barreto, Jaime S. Cardoso, M. Coimbra, Reject Option Paradigm for the Reduction of Support Vectors, in: ESANN, 2014.
[86] A. Subbaswamy, S. Saria, From Development to Deployment: Dataset Shift, Causality, and Shift-Stable Models in Health AI, Biostatistics 21 (2) (2020) 345–352, https://doi.org/10.1093/biostatistics/kxz041.
[87] J. Suutala, S. Pirttikangas, J. Riekki, J. Röning, Reject-Optional LVQ-Based Two-Level Classifier to Improve Reliability in Footstep Identification, in: A. Ferscha, F. Mattern (Eds.), Pervasive Computing. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2004, pp. 182–187, https://doi.org/10.1007/978-3-540-24646-6_12.
[88] D.M.J. Tax, R.P.W. Duin, Growing a Multi-Class Classifier with a Reject Option, Pattern Recogn. Lett. 29 (10) (2008) 1565–1570, https://doi.org/10.1016/j.patrec.2008.03.010.
[89] F. Tortorella, An Optimal Reject Rule for Binary Classifiers, in: F.J. Ferri, J. M. Iñesta, A. Amin, P. Pudil (Eds.), Advances in Pattern Recognition. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2000, pp. 611–620, https://doi.org/10.1007/3-540-44522-6_63.
[90] K. Tran, W. Neiswanger, J. Yoon, Q. Zhang, E. Xing, Z.W. Ulissi, Methods for Comparing Uncertainty Quantifications for Material Property Predictions, ArXiv: 1912.10066 [Cond-Mat, Physics:Physics], 2020 February, http://arxiv.org/abs/1912.10066.
[91] A. Tucker, Z. Wang, Y. Rotalinti, et al., Generating high-fidelity synthetic patient data for assessing machine learning healthcare software, npj Digit. Med. 3 (2020) 147, https://doi.org/10.1038/s41746-020-00353-9.
[92] D. Ulmer, L. Meijerink, G. Cinà, Trust Issues: Uncertainty Estimation Does Not Enable Reliable OOD Detection On Medical Tabular Data. ArXiv:2011.03274 [Cs, Stat], 2020 November. http://arxiv.org/abs/2011.03274.
[93] A. Uyar, F. Gurgen, Arrhythmia Classification Using Serial Fusion of Support Vector Machines and Logistic Regression, in: 2007 4th IEEE Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, 2007, pp. 560–565, https://doi.org/10.1109/IDAACS.2007.4488483.
[94] J. Vaicenavicius, D. Widmann, C. Andersson, F. Lindsten, J. Roll, T.B. Schön, Evaluating Model Calibration in Classification'. ArXiv:1902.06977 [Cs, Stat], 2019 February. http://arxiv.org/abs/1902.06977.
[95] M.H. Waseem, M.S.A. Nadeem, A. Abbas, A. Shaheen, W. Aziz, A. Anjum, U. Manzoor, M.A. Balubaid, S.-O. Shim, On the Feature Selection Methods and Reject Option Classifiers for Robust Cancer Prediction, IEEE Access 7 (2019) 141072–141082, https://doi.org/10.1109/ACCESS.2019.2944295.
[96] T.-W. Weng, H. Zhang, P.-Y. Chen, J. Yi, D. Su, Y. Gao, C.-J. Hsieh, L. Daniel, Evaluating the Robustness of Neural Networks: An Extreme Value Theory Approach, ArXiv:1801.10578 [Cs, Stat], January 2018. http://arxiv.org/abs/1801.10578.
[97] J. Wiens, S. Saria, M. Sendak, M. Ghassemi, V.X. Liu, F. Doshi-Velez, K. Jung, et al., Do No Harm: A Roadmap for Responsible Machine Learning for Health Care, Nat. Med. 25 (9) (2019) 1337–1340, https://doi.org/10.1038/s41591-019-0548-6.