Journal of Biomedical Informatics 101 (2020) 103350

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Targeting the uncertainty of predictions at patient-level using an ensemble of classifiers coupled with calibration methods, Venn-ABERS, and Conformal Predictors: A case study in AD

Telma Pereira[a,b,*], Sandra Cardoso[c], Manuela Guerreiro[c], Alexandre Mendonça de[c], Sara C. Madeira[a,**], for the Alzheimer's Disease Neuroimaging Initiative[1]

[a] LASIGE, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal
[b] Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal
[c] Laboratório de Neurociências, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portugal

ARTICLE INFO

Keywords:
Prognostic prediction
Mild cognitive impairment
Alzheimer's disease
Uncertainty at patient-level
Venn-ABERS
Conformal prediction

ABSTRACT

Despite being able to make accurate predictions, most existing prognostic models lack a proper indication about the uncertainty of each prediction, that is, the risk of prediction error for individual patients. This hampers their translation to primary care settings through decision support systems. To address this problem, we studied different methods for transforming classifiers into probabilistic/confidence-based predictors (here called uncertainty methods), where predictions are complemented with probability estimates/confidence regions reflecting their uncertainty (uncertainty estimates). We tested several uncertainty methods: two well-known calibration methods (Platt Scaling and Isotonic Regression), Conformal Predictors, and Venn-ABERS predictors. We evaluated whether these methods produce valid predictions, where uncertainty estimates reflect the ground truth probabilities. Furthermore, we assessed the proportion of valid predictions made at high-certainty thresholds (predictions with uncertainty measures above a given threshold) since this impacts their usefulness in clinical decisions. Finally, we proposed an ensemble-based approach where predictions from multiple pairs of (classifier, uncertainty method) are combined to predict whether a given MCI patient will convert to AD. This ensemble should putatively provide predictions for a larger number of patients while releasing users from deciding which pair of (classifier, uncertainty method) is more appropriate for data under study. The analysis was performed with a Portuguese cohort (CCC) of around 400 patients and validated in the publicly available ADNI cohort. Despite our focus on MCI to AD prognosis, the proposed approach can be applied to other diseases and prognostic problems.

1. Introduction

Machine learning is at the core of major advances in healthcare and medical domains. In the particular case of Alzheimer's disease, the leading cause of dementia [1], researchers have sought for robust supervised learning models to predict whether a patient with Mild Cognitive Impairment (MCI) is likely to convert to AD in the future [2,3]. These prognostic models might then be used to guide clinical decisions in real-world situations concerning patients' treatment, participation in cognitive rehabilitation programs, and selection for clinical trials. Nevertheless, despite the promising results attained by advanced machine learning methods [2,3], some issues have hampered their practical application in clinical settings [4]. In fact, most models output the most likely prediction for new patients without providing an indication of the uncertainty of each prediction, i.e., the risk of prediction error for an individual patient. This precludes their usability in risk-sensitive

Abbreviations: AD, Alzheimer's Disease; MCI, Mild Cognitive Impairment; PS, Platt Scaling; IR, Isotonic Regression; CPs, Conformal Predictors; VAs, Venn-ABERS; CCC, Complaints Cognitive Cohort; ADNI, Alzheimer's Disease Neuroimaging Initiative

* Corresponding author at: LASIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal.
** Corresponding author at: LASIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal.
E-mail addresses: telma.pereira@tecnico.ulisboa.pt (T. Pereira), madeira@ciencias.ulisboa.pt (S.C. Madeira).

[1] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

https://doi.org/10.1016/j.jbi.2019.103350
Received 26 April 2019; Received in revised form 25 November 2019; Accepted 1 December 2019
Available online 06 December 2019
1532-0464/

decisions, typical in the medical domain.

Despite the above scenario, most classifiers provide measures that could be used to estimate the uncertainty of predictions at instance-level, in this case, at patient-level [5]. Some compute the probability distribution over classes directly, such as Naïve Bayes (NB), Neural Networks (NN), and Logistic Regression (LR). Others, such as Decision Trees (DT), estimate the probability distribution *ad hoc* of the classification task using measures related with inherent characteristics of the classifier itself [6,7]. In this case, the predicted class is that with the highest probability, and the probability assigned to the predicted (the most likely) class reflects its degree of uncertainty. However, several studies [7–10] have shown that some of these probability estimates are not well-calibrated, i.e., they do not reflect the empirical probabilities as the number of predictions goes to infinity [11]. NB, for instance, is known to produce poor probabilities (often pushed towards 0 and 1) due to the independence assumption, despite its generally accurate performance [8,12,13]. Moreover, the leaf frequency-based probability estimated by DT has a similar behavior [7,14]. This is prompted by the effect of small leaves, potentially providing optimistic estimates and thus introducing bias in the predictions [14,5]. Other classifiers, such as LR and NN, produce well-calibrated probabilities [8,12,13]. In this context, the reliability of these probability estimates strongly depends on the classifier. Other supervised learning algorithms rely on distance-based scores [15]. An example are Support Vector Machines (SVM), where the distance to the hyperplane may be interpreted as a measure of predictions' reliability [16]. In this work, we refer to these measures of uncertainty, directly computed by the classifier, as Direct Probability Estimates (DPE).

Alternatively, calibration methods [6,7,9,14,17,18] perform an *ad-hoc* mapping of the original probability estimates into probabilities that better represents the true likelihood of predictions, thus correcting the biased probabilities computed by some classifiers. Platt Scaling (PS) [17] and Isotonic Regression (IR) [18,7] are the most commonly used methods for probability calibration. Platt scaling was originally proposed to transform SVM outputs into probabilities by fitting them through a Logistic Regression [17]. It can be used with other classifiers as well [13]. Isotonic Regression is a more flexible calibration method since it only requires the mapping function to be isotonic (monotonically increasing), making no assumption on its shape. Different classifiers produce probability estimates characterized by different distortions, that will, in turn, influence the benefit of each calibration method. A comprehensive comparison between PS and IR using different classifiers is presented in [13]. In theory, PS is more appropriate for classifiers whose prediction curve is sigmoid-shaped. That is the case of maximum margin methods, such as SVM and boosted trees, which tend to push predicted probabilities away from 0 and 1 [13]. IR can also be used to calibrate probabilities with sigmoid-shape distortions. However, it tends to overfit when using small calibration sets. IR outperforms PS when calibrating probabilities estimated by NB, which are closer to 0 and 1, or DT, which have a non-characteristic distortion curve [13]. Some classifiers, such as LR and NN, produce well-calibrated probabilities and thus do not benefit from probability calibration [13,19].

Although useful in practice, calibration methods do not provide any validity guarantee or risk of error bound [20,15]. On the other hand, Venn-ABERS predictors (VAs) [21–23] and Conformal Predictors (CPs) [24] are theoretical-backed frameworks guaranteed to always produce valid predictions in the form of perfect calibration, based on the randomness assumption. Validity is the property of a predictor to output probability distributions performing well against statistical tests as the number of empirical probabilities goes to infinity. These methods differ in the sense that VAs are probability-based predictors while CPs are confidence-based predictors. In this context, VAs output the probability of assigning a given class to a particular instance (patient in this case) while CPs output a prediction region containing the predicted class within a calibrated confidence interval.

We refer to the aforementioned methods as uncertainty methods from now on for ease of readability. Several studies have theoretically and empirically studied these uncertainty methods regarding the validity of the produced predictions when combined with different classifiers and datasets. Namely, researchers have compared VAs with calibration methods (PS and IR) [21,22], the probability estimates produced by classifiers with calibration methods [13] and with CPs [25], and VAs with CPs [26]. Notwithstanding, there is a lack of studies systematically evaluating different aspects of such methods using a common dataset and with greater variability in the number of training examples. For instance, most experiments are pursuit on large datasets [13,22,27], while in some application domains, such as in dementia-related research, data is scarce. Moreover, in order to give actionable insight in risk sensitive decisions, predictions must be assigned with low uncertainty estimates, while middle-values (around 0.5) are not very informative. Therefore, choosing the most appropriate method for the problem under study is not straightforward.

Despite the clinical relevance of evaluating the uncertainty of prognostic predictions at patient-level, to our knowledge, defining which method is more suitable for this purpose is still poorly studied. In an exploratory study [25], we showed that credibility scores computed by CPs are more informative than the posterior probabilities estimated by NB when targeting uncertainty of predictions on the MCI to AD conversion problem, using neuropsychological data. In general, CPs produced few misclassifications for high-certainty levels. Nevertheless, their efficiency was challenging, since many predictions were associated to uncertainty estimates below the predefined certainty threshold, and thus, they were considered as uncertain predictions (or unpredictable cases). A different approach used to tackle prediction's uncertainty relies on survival methods. In [28], the authors provide a framework, based on prognostic models built with the Cox Hazard Proportional Regression, to interpret individual patient data. For each MCI patient, it outputs the probability of progression to AD at 1 and 3 years follow-up, using demographic, cerebrospinal fluid, and imaging-based biomarkers. Yet, once again, predictions with probability values near 0.5 are not informative, since they lead to unpredictable cases.

In this work, we aim at exploring the most suitable approach to complement predictions at patient-level with a valid measure of prediction uncertainty for prognostic models in MCI. In this context, the uncertainty methods should be compared regarding their interpretability, calibration guarantees, and efficiency (number of predictions with uncertainty estimates above the predefined certainty threshold) since it directly impacts the usefulness of this approach in the clinical settings. We thus provide an outright comparison between different methods to measure the uncertainty of individual predictions for different classifiers. We studied Platt Scaling and Isotonic Regression since these have been proved to work well empirically for a range of classifiers [13]. Moreover, we investigated VAs and CPs due to their theoretically-backed foundations and successful application in health-related domains [29–31]. Furthermore, we tested the Cox Hazard Proportional Regression model for the sake of comparability with the study reported in [28].

As aforementioned, we are particularly interested in evaluating predictions made at high-certainty thresholds, since these give actionable insight. Cases that cannot be predicted given the certainty threshold are referred, in this study, as unpredictable cases. In this line, one drawback of the aforementioned prognostic models [28,25] regards the limited number of predictions produced at high-certainty thresholds, leaving many patients without a prognostic. Notwithstanding, we hypothesize that unpredictable cases may differ across uncertainty methods and classifiers used. In this context, predictions could be complemented between them improving the number of predictions outputted. With this in mind, we propose an ensemble-based approach, where predictions from multiple classifiers and multiple methods able to address the uncertainty of predictions are combined to predict whether a given MCI patient will convert to AD. This ensemble should

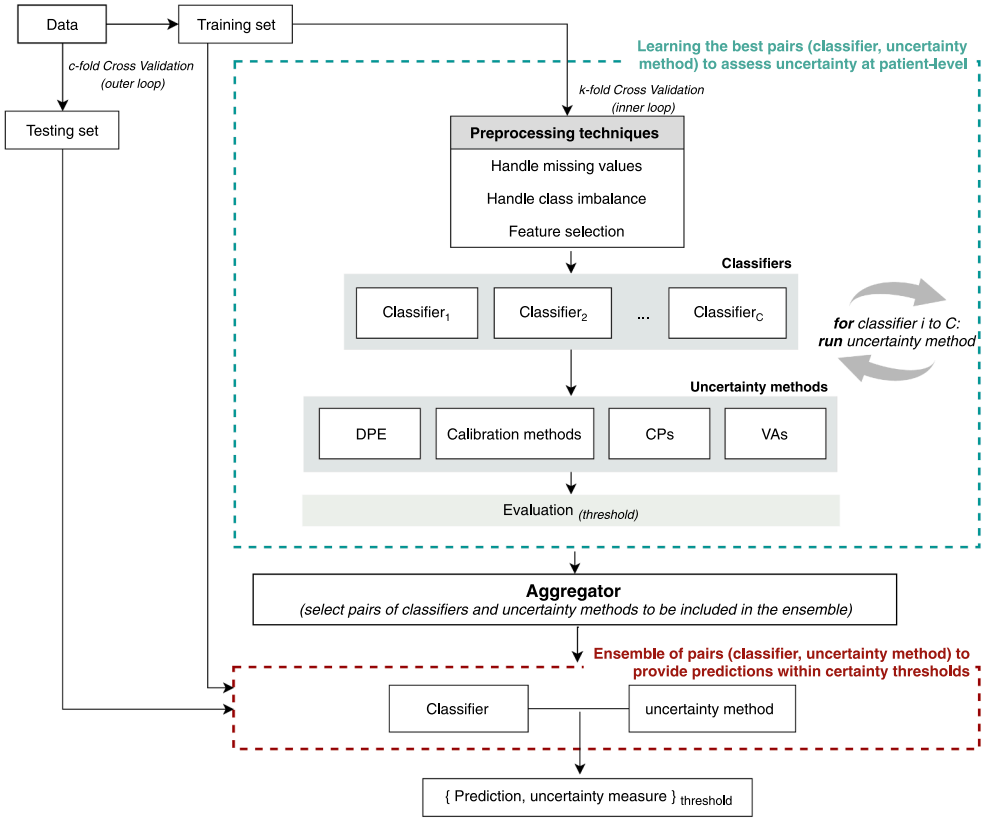

Fig. 1. Workflow of the ensemble-based approach using different classifiers and methods to target uncertainty of predictions at patient-level.

putatively provide predictions for a larger number of patients while releasing users from deciding which pair of (classifier, uncertainty method) is more appropriate for the data under study.

We perform this study using a Portuguese dataset, the Cognitive Cohort Study (CCC) [32], and validated the proposed ensemble-based approach targeting the uncertainty of predictions at patient-level using the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset [33]. The goal is to predict conversion from MCI to AD using neuropsychological data. Despite our focus on MCI to AD prognosis, we note that the proposed approach can be applied to other diseases or prognostic problems.

The paper is organized as follows. Section 2 describes the methods used to target the uncertainty of predictions at patient-level as well as the proposed ensemble-based approach. Section 3 presents data and discusses the results obtained. Finally, Section 4 concludes the paper.

2. Methods

Fig. 1 illustrates the proposed ensemble-based approach using different classifiers and methods to target the uncertainty of predictions at patient-level (named here as uncertainty methods). This approach comprises two phases *(1)* Learning the best pairs (classifier, uncertainty method) to assess the uncertainty at patient-level and *(2)* Ensemble of pairs (classifier, uncertainty method) to provide predictions within certainty thresholds. The learning process follows a nested cross-validation (CV) procedure repeated with fold randomization to access model generalization. Each step of the proposed ensemble approach is described in detail in the following subsections. We describe the methods used to target uncertainty of predictions: DPE, calibration methods (PS and IR), VAs and CPs. In the case of CPs several conformity measures are used. We describe this measure and justify the proposal of a new conformity measures for SVM. Finally, we describe the step where methods and classifiers are combined into an ensemble.

2.1. Targeting the uncertainty of predictions at patient-level

This section contains a brief description of the methods used to assess the uncertainty of predictions at patient-level in order to make this paper self-contained. All methods are based on previously published theory [17,18,24,21]. Let us assume that we are given a training set (x_1, y_1), ...,(x_{n-1}, y_{n-1}), where $x_i \in X$ is a vector of d attributes and $y_i \in Y$ is the class label (we assume a binary classification problem). Given a new test instance (x_n) we aim to predict its class $(\hat{y}_n)$ and an estimation on the likelihood of its prediction being true: $\hat{y}_n = y_n$, where y_n is the true label for . In this context, we study the performance of *1)* probability estimates given by standard classifiers, referred in this study as Direct Probability Estimates, *2)* calibration methods (Platt Scaling and Isotonic Regression), *3)* Conformal Predictors, and *4)* Venn-ABERS predictors. We discuss these methods regarding *i)* the interpretability of the measure used to assess the risk of error for individual predictions, *ii)* whether this measure is well-calibrated, and *iii)* the number of predictions made per certainty threshold.

2.1.1. Direct probability estimates (DPE)

Some supervised learning algorithms are *scoring classifiers* [22,5] in the sense they estimate a real-valued score, $s(x_i)$, per instance, which is then compared to a threshold, c, to obtain a categorical prediction (0 if $s(x_i) < c$ and 1, otherwise). This is the case of NN and NB, for instance. This score can be used as a measure of the likelihood of the predicted class. So, $s(.)$ is hereby called the scoring function. Formally, a scoring classifier is a mapping from the instance space I to a k-vector of real numbers ($\mathbf{s}$: $I \rightarrow \mathbb{R}^k$), where k is the number of classes. In binary classification, it usually suffices to consider the score of only one class. As an example, NB computes the posterior probability distribution over two possible classes. The predicted class is as likely as the *maximum posterior probability* obtained. The probability of assigning the other class is given by the complement to 1 of that *maximum posterior*

probability. In this work, we use $s(x_i)$ to denote the score of the positive class ($y = 1$).

Other supervised learning algorithms, despite not being naturally scoring classifiers, provide measures that can be used to compute probability estimates ($p_i(x_n)$, $i = 0, 1$) for each of the two classes [6,7]. This is the case of DT, kNN, and SVM, for instance. While the first two classifiers rely on the class distribution on the leaves or the nearest neigbohrs, respectively, to compute probability estimates over classes, the latter relies on distance-based scores, using the distance to the hyperplane as a measure of predictions' likelihood [16]. We refer to these measures as DPE.

Several studies [13,8–10] have shown that these probability estimates are not well-calibrated, i.e., the predicted class probabilities do not reflect the empirical probabilities as the number of predictions goes to infinity [11].

The interpretability of the score, in this case, depends on the classifier. Probability estimates are easily interpreted in probabilistic classifiers, such as NB. For non-probabilistic classifiers, such as SVM, interpretability depends on the knowledge of the users on this algorithm. Probabilistic predictors produce well-calibrated predictions under strong assumptions whereas simple (non-probabilistic) predictors have no concept of guaranteed calibration.

2.1.2. Calibration methods

Calibration methods consist on *ad hoc* mappings of the original probability estimates or prediction scores ($s(x_i)$) into more accurate posterior probabilities [6,25,9,14,17,18]. Platt Scaling [17] and Isotonic Regression [18] are the most prevalent methods for probability calibration. In a nutshell, these calibration methods pass the prediction scores (estimated by a scoring classifier) through a function that searches for the argument space that minimizes loss in order to improve the probability estimates. The calibrated probability estimates are restricted to a range between 0 and 1, being seen as the probability of assigning the class 1 to instance x_i. When given a test instance, x_n, the respective prediction score $s(x_n)$ is computed with the calibration model, thus returning the calibrated probability estimate.

Platt Scaling fits a Logistic Regression to the prediction scores $s(x_i)$:

$$p(y_n = 1|s(x_n)) = \frac{1}{1 + exp(\alpha s(x_n) + \beta)}, \quad (1)$$

where parameters α and β are fitted using maximum likelihood estimation from the training (or calibration) set: $(s(x_i), y_i)$. For more details, we refer to [17].

Isotonic Regression [18] is a more flexible method for probability calibration. Contrarily to PS, where the distortion correction follows a sigmoid shape, IR only requires the mapping function to be isotonic, i.e., monotonically increasing. Therefore, given the training set $(s(x_i), y_i)$, the IR problem is to find a mapping function that best fits the training set according to the mean-squared error criterion. The Pair-Adjacent Violators (PAV) algorithm [34] is a possible solution to find a stepwise constant isotonic function.

In order to avoid bias, the models should be learned with a calibration set independent from the training set used to obtain the DPE scores. We note that the quality of the calibrated probability estimates is as good as the quality of the DPE scores given as input. PS is known to perform better on classifiers whose prediction scores are sigmoid-shaped, such as SVM or boosted trees [13]. However, when the calibration set is large enough, IR performs similarly to PS in classifiers with sigmoid-shaped curves and outperforms it for a broader range of classifiers, such as NB, kNN, and DT [13]. Both calibration methods give probabilistic predictions, which are straightforwardly interpretable. However, as far as we know, they are not proved to be well-calibrated in a general sense.

2.1.3. Conformal predictors (CPs)

Conformal Prediction is a machine learning framework built on top of standard classifiers that, for a given test instance, produces a prediction set guaranteed to include the true class, at a pre-specified confidence level. CPs are valid under the randomness (i.i.d) assumption, which states that instances are independently drawn from the same distribution [24]. The new test instance (x_n) is thus expected to have the class label (y_n) that makes it similar to the training instances of the same class. The degree to which that similarity holds amongst the known instances determines how confident the classifier is in that prediction.

We introduce the idea behind the conformal prediction framework. For a more formal description we refer to [24]. Let us assume that we are given a training set $(x_1, y_1), \ldots, (x_{n-1}, y_{n-1})$, where $x_i \in X$ is a vector of attributes and $y_i \in Y$ is the class label (binary classification problem). Given a new test instance (x_n) we aim to predict its class. Intuitively, we assign each class $y_n \in Y$ to x_n, at a time, and then evaluate how (dis)similar the instance (x_n, y_n) is in comparison with the training data. A (non-)conformity measure, that assesses (dis-)similarity between instances by means of a numerical (non-)conformity score (α_n) is defined using the underlying classifier. To evaluate how different x_n is from the training set we compare its non-conformity score with those of the remaining training instances $x_{j:j=0,\ldots,n-1}$, using the *p*-value function:

$$p(\alpha_n) = \frac{|\{j = 1, \ldots, n: \alpha_j \geqslant \alpha_n\}|}{n}, \quad (2)$$

where α_n is the non-conformity score of x_n, assuming it is assigned to the class label y_n. If the *p*-value is small, then the test instance (x_n, y_n) is non-conforming, since few instances (x_i, y_i) had a higher non-conformity score when compared with α_n. If the *p*-value is large, x_n is very conforming, since most instances (x_i, y_i) had a higher non-conformity score when compared with α_n. Once *p*-values are computed, CP can be used in one of the following ways:

1. *Using prediction regions (CP-PR)*. For a given significance level (ϵ), CPs output a prediction region, T^{ϵ}: set of all classes with $p(\alpha_n) > \epsilon$, contrarily to the single predictions given by standard classifiers. These prediction regions have a guaranteed error rate (guaranteed validity [24]). This means that the frequency of errors (fraction of true values outside T^{ϵ}) does not exceed ϵ, at a confidence level $1 - \epsilon$. Prediction regions may, therefore, comprise more than one class (uncertain prediction), no class (empty prediction) or a single class (certain prediction). Multiple predictions are not mistakes but a reflection of the classifier not being confident enough to predict a certain class. The smaller the prediction region, the more efficient the conformal predictor [24]. Efficiency depends on the underlying classifier and on the (non-)conformity measures used.
2. *Using forced predictions (CP-FP)*. To have single predictions rather than prediction regions, CPs predict the class with the highest p-value (forced prediction), together with its credibility (the largest p-value) and confidence (complement to 1 of the second highest p-value). This has the cost of losing the guaranteed error bounded confident interval. Confidence reveals how likely the predicted classification is compared with the other classes. Credibility reveals how suitable the CP is for classifying the given instance. Low credibility means that either the training set is non-random or the test instance is not representative of the training set. The probability that the credibility is less than some threshold ϵ is less than ϵ (randomness assumption) [35,24]. The higher the values of both confidence and credibility the more reliable the prediction is. Confidence may also be interpreted as the largest $1 - \epsilon$ at which T^{ϵ} has a single prediction (certain prediction) and credibility as the largest for which T^{ϵ} is empty [24,36].

CPs have the advantage of enjoying guaranteed validity (i.e., perfect calibration) [24]. Nevertheless, p-values and confidence concepts are not as easily interpretable as probabilities.

Conformity measure for SVM

We recall that the CP framework requires the definition of a non-conformity (or conformity) measure to quantitatively estimate the strangeness (or similarity) of an instance compared with other instances. This measure arises from the underlying classifier, and it is critical to the quality and efficiency of the predictions returned by CPs. Different non-conformity measures have been proposed for widely used classifiers [37–41]. For instance, it is common to use the posterior probability estimated by NB to calculate the respective non-conformity measure. When using SVM, the non-conformity score may be computed using the weights assigned to the support vectors [41], the distance to the separating hyperplane [38], or a compositive measure of the distance to the margin boundary of the class under consideration [42]. However, most of these non-conformity measures were restricted to the Mondrian Conformal Prediction framework [24], a variant of CPs that deals with imbalanced data, since the score of a given instance x_n needs to be computed with respect to its own class. In this context, we propose a new conformity measure to be used with SVM, which, to our knowledge, was not used before. Follows a description of the proposed conformity measure for SVM based on the distance to the margin boundaries of each possible class.

Let us assume a kernel function $k(x)$ and the decision boundary of a kernel-based binary SVM $y = w{\cdot}k(x) + b$. Given a test instance, x_n, we assign the positive class ($y_p = +1$) if the decision boundary is greater than zero and the negative class ($y_p = -1$), otherwise. The signed distance of a given instance, x_n, to the separating (maximum-margin) hyperplane is given by

$$d_n^h = \frac{y}{\|w\|}, \tag{3}$$

where y is the output of the SVM and $\|w\|$ is the weighted sum of support vectors (using the dual formulation of SVM). The distance to the margin boundary of the class under consideration (predicted class) is given by

$$d_i^{mb} = |d_n^h - \frac{1}{\|w\|}. \tag{4}$$

We note that $\frac{1}{\|w\|}$ is the distance between the margin boundary and the hyperplane. In this work, we define the conformity measure for SVM as the distance between the instance and the margin boundary of the class under consideration. Therefore, the conformity score ($\alpha_n^{y_p}$) of instance x_n belonging to class y_p is given by

$$\alpha_n^{y_p} = \begin{cases} \|d_n^h\| - \frac{1}{\|w\|} & , if\,(y < 0 AND\, y_p = -1)\, OR\, (y > 0 AND\, y_p = +1). \\ -\left(\|d_n^h\| + \frac{1}{\|w\|}\right) & , if\,(y < 0 AND\, y_p = +1)\, OR\, (y > 0 AND\, y_p = -1). \end{cases} \tag{5}$$

Fig. 2 provides an illustrative example of how the proposed conformity score is computed. As aforementioned, given an example (in this case "A") CPs assign, at a time, each possible class. Then, CPs predict the class for which the training instances conform better with. If we consider "A" to belong to the positive class ($y_p = +1$), α_n^{+1} is the distance to the margin boundary corresponding to the positive class $\left(\alpha_n^{+1} = \|d_A^h\| - \frac{1}{\|w\|}\right)$. On the other hand, if we consider "A" to belong to the negative class ($y_p = -1$), α_n^{-1} is the distance to the margin boundary corresponding to the negative class. In case SVM place "A" in the negative side of the hyperplane, we penalize the conformity score with a negative weight $\left(\alpha_n^{-1} = -\left(\|d_A^h\| + \frac{1}{\|w\|}\right)\right)$. The same holds when a negative instance is hypothesized as being of the positive class.

2.1.4. Venn-ABERS predictors (VAs)

Venn-ABERS [21], similarly to Conformal Prediction, is a machine learning framework built on top of scoring classifiers, that produces (multi)probabilities (instead of CPs p-values) to estimate the uncertainty of predictions at patient-level. VAs are an adaptation of IR and a special case of Venn Predictors [24,43], from which they inherit guaranteed validity in the form of well-calibrated probability

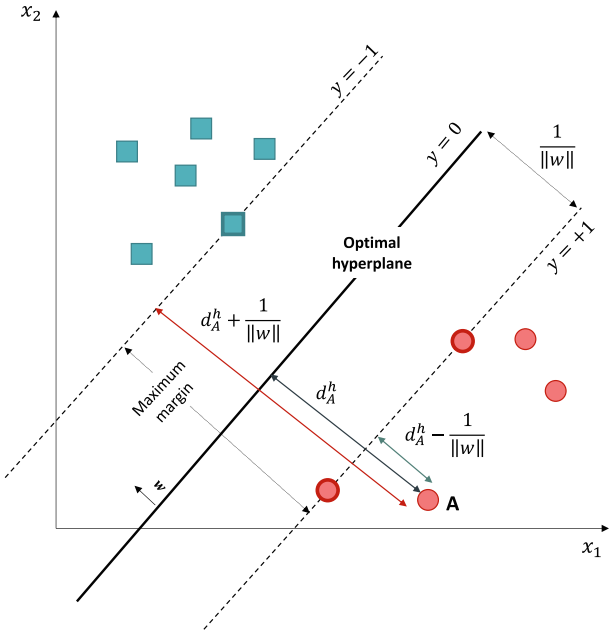

Fig. 2. Illustration of the proposed non-conformity measure for SVM.

predictions [22]. This means that *"in the long term the relative frequency of examples with the desired property, i.e., $P\{y = 1|p_{pred}(x) = p\}$ among those with predicted probability p of having that property is indeed p"* [27]. IR is shown to overfit, especially when using small calibrations sets, while VAs atenuate this tendency [13].

Analogously to CPs, given a test instance x_n, we assign each class $y_n \in \{0, 1\}$ to x_n, at a time, and compute the respective prediction scores ($s(x_n)$) using the underlying classifier. Let $s_0(x)$ and $s_1(x)$ represent the scoring function learned with the training set including the instances $(x_n, 0)$ and $(x_n, 1)$, respectively. We fit an isotonic regression to the series $(s_0(x_i), y_i)$, and $(s_1(x_i), y_i)$, $i = 0, ..., n$ generating the function $f_0(x)$ and $f_1(x)$, respectively. The multi-probabilistic prediction outputted by VAs is $(p_0, p_1) := (f_0(s_0(x_n)), f_1(s_1(x_n)))$, which corresponds to the lower and higher probability of x_n being from class 1. The difference between p_0 and p_1 reflects how reliable the prediction is (the smallest the best) [22].

Single-valued probabilities rather than multi-probabilities are preferable in real-world applications [27,22]. In this context, we can merge these probabilities by minimizing a loss function at the cost of losing the validity guarantee, although empirical evidence shows that high accuracy is maintained [22]. A commonly used metric is the log-loss in which multi-probabilities are combined by

$$p = \frac{p_1}{1 - p_0 + p_1}. \tag{6}$$

VAs are advantageous over CPs in the sense that probabilities give a more intuitive estimation of the uncertainty of predictions. Moreover, they are easier to combine to find optimal decisions than confidence measures (p-values) [26].

2.2. Ensemble-based approach to target the uncertainty of predictions at patient-level

There are several methods to address the uncertainty of individual predictions, each with its own strengths and flaws, which might depend on data and on the classifier used to learn the model. In this context, deciding which pair of (classifier, uncertainty method) is more suitable to the problem at hand is not trivial, requiring a deep understanding of both data and algorithms. Bayesian classifiers, for instance, produce accurate confidence values for predictions when priors are known [44]. Yet, for real-world data, prior knowledge on generating data

distribution is missing, leading to incorrect confidence values. Moreover, the benefit of probability calibration methods depends on shape of the predicted probability estimates given by each classifier and on data size [13]. For instance, PS is typically more effective for classifiers whose predicted probabilities produces sigmoid-shaped distortions (such as those produced by SVM). In addition, PS works well on small datasets while IR requires larger calibration sets to prevent overfitting. In the case of VAs and CPs, the only requirement to have perfect calibration is that data are independently and identically distributed. Nevertheless, the number of predictions made at high-certainty thresholds by CPs depends on the conformity measure, which in turn, depends on the underlying classifier [45].

Moreover, one challenge when imposing certainty thresholds is the low efficiency obtained (i.e., small number of predictions with uncertainty estimates above the predefined threshold), and thus high number of unpredictable cases. We hypothesize these unpredictable cases may differ across the uncertainty methods and classifiers used. Therefore, a higher number of reliable predictions are putatively achieved if we complemented the results from different methods and classifiers. In this scenario, we propose an ensemble-based approach that combines multiple methods following different strategies to target the uncertainty of predictions and different classifiers.

In the first step of the ensemble-based approach depicted in Fig. 1 ("Learning the best pairs (classifier, uncertainty method) to assess uncertainty at patient-level" module), different classifiers are combined with different strategies to assess the uncertainty of predictions: DPE, calibration methods (PS and IR), CPs, and VAs. Then, in the second step ("Ensemble of pairs (classifier, uncertainty method) to provide predictions within certainty thresholds" module), the aggregator selects which of these pairs should be included in the ensemble, according to a predefined rule (described onward in Sections 2.3 and 3.3). Then, for each patient (in the outer CV loop), each selected pair of (classifier, uncertainty method) provides a prediction and the respective uncertainty value, per threshold. The final prediction is that with the highest uncertainty estimate given the certainty threshold. In this context, we aim at quantitatively and qualitatively improve predictions at patient-level.

2.3. Classification settings

We compared different strategies on the task of targeting uncertainty of individual predictions (uncertainty methods): probability estimates obtained directly by classifiers, calibration methods, conformal predictors, and Venn-ABERS predictors. In order to assess the robustness of our comparative analysis, we used classifiers relying on different approaches to tackle the classification problem and to compute the likelihood of each prediction: Gaussian Naïve Bayes, Decision Tree with J48 algorithm, Gaussian (SVM RBF) and Polynomial-kernel (SVM Poly) Support Vector Machines using SMO implementation, Logistic Regression and Neural Networks.

Table 1 depicts the metrics used, per classifier, to compute scores from which uncertainty of predictions are calculated, either by using them directly (DPE), by calibrating them (PS and IR), or by feeding them into CPs and VAs frameworks. The non-conformity measure $-\log(p)$ was previously used in conformal prediction studies using NB [12,38]. We extended it to DT, LR and NN since these classifiers also produce probability distributions over classes. For SVM we used the signed output of SVM to be calibrated (PS and IR) or fed into VAs. Moreover, we used the normalized signed output of SVM as DPE to confine scores to a common range ($[0, 1]$), for the sake of comparability. The conformity measure used in CPs is given by the proposed metric defined in (5). Since data under study is not high dimensional, we used the Transductive Conformal Prediction framework. When using calibration methods (PS and IR), we used an inner 3-fold CV of the training set to calibrate the models. We studied also whether the cumulative distribution function (denoting the probability of progression to AD occurring up to a given time), estimated by Cox Proportional Hazard Regression models, represents a good indicator of the uncertainty of predictions.

2.3.1. Evaluation

To evaluate the overall performance of the ensemble approach we used a nested cross-validation procedure. First, an external 5-fold CV was implemented in which data were randomly divided into 5 subsets while maintaining class proportion (stratified CV). At each step, one subset was left for testing (validation set) and the remaining subsets were used for training. Then, we used the training set in a 10×5-fold stratified cross-validation scheme for each experiment: <classifier, uncertainty method, threshold>. We tested three thresholds ($\tau \in 0.80, 0.90, 0.95$) considered useful in clinical practice. The performance of each pair of classifier and method to assess uncertainty (per threshold) was evaluated in the 10×5-fold of the inner CV loop. To evaluate classification performance, we assessed the Area Under the ROC Curve (AUC), sensitivity (proportion of actual converting patients, cMCI, which are correctly classified), and specificity (proportion of non-converting patients, sMCI, which are correctly classified). Then, the aggregator selects which pairs (classifier, uncertainty method) to include in the ensemble. In this study, we used two general rules and two specific rules that are defined in the aftermath of the results described in Section 3.3. The former rules consist in *1)* using all pairs (classifier, uncertainty method) to assess the uncertainty of predictions, and *2)* using the uncertainty method with highest AUC per classifier. Finally, the ensemble-based approach is evaluated using the validation subset of the outer CV loop. We pursued the experiments using CCC data and used ADNI to validate the final ensemble-based approach.

Dealing with a large number of (possibly irrelevant) features may have a significant impact on both classification performance and model simplicity and interpretability. In this context, we selected the most relevant set of features by using the feature selection ensemble algorithm proposed in [46]. This feature selection approach starts by ranking features according to their relevance as assessed by a consensus of different feature selection algorithms using a heterogeneous ensemble. The best subset of features is composed by the top-ranked features maximizing both predictability and stability performances (we refer to [46] for details).

The statistical significance of classification results was evaluated

Table 1
Parameters and metrics used by each uncertainty method.

Classifier	Parameters	DPE	CP	VA-PS-IR	Comment
NB	Gaussian	p	$-\log p(y_i = c\|x_i)$	p	p is the posterior probability estimated by NB
DT	*confidence* = 0.25	p	$-\log p(y_i = c\|x_i)$	p	p is the frequency-based probability estimated by DT
SVM Poly	$c = 1$ *degree* = 2	signed y (normalized)	5	signed y	y is the output estimated by SVM
SVM RBF	$c = 1$ *degree* = 2	signed y (normalized)	5	signed y	y is the output estimated by SVM
LR	*ridge* = 10^{-8}	p	$-\log p(y_i = c\|x_i)$	p	p is the probability estimated by LR
NN	$l = 0.3$ $m = 0.2$ *No. Layers* = (No. features + No. classes)/2	p	$-\log p(y_i = c\|x_i)$	p	p is the probability estimated by NN

Table 2
Baseline demographic characterization data.

	CCC			ADNI		
	sMCI	*cMCI*	*p-value*	*sMCI*	*cMCI*	*p-value*
No. of instances (%)	227 (56%)	175 (44%)	–	143 (54%)	122 (46%)	-
Age, years (M ± SD)	65.5 ± 9.1	71.9 ± 8.3	< 10–12*	72.1 ± 7.3	74.8 ± 7.6	< 0.004*
Formal Education, years (M ± SD)	10.4 ± 4.7	8.8 ± 4.8	< 0.001*	16.1 ± 2.8	16.0 ± 2.6	0.895
Gender (male/female)	85/142	60/115	0.995	82/61	70/52	0.995

using the averaged AUC across the 10 × 5-fold CV (using CCC dataset). Friedman Tests [47] were used to check whether results obtained across different uncertainty methods were significantly different. Pairwise comparisons (using the Wilcoxon Signed Rank Test) were then performed (with Bonferroni correction for multiple testing) to assess which methods performed statistically better. We used IBM SPSS Statistics 24 to execute the statistical tests.

The described methodology, including the comparative analysis using different classifiers and methods to target uncertainty of predictions to the ensemble-based approach using the best-performing uncertainty methods was implemented in Java using WEKA's functionalities. The survival analysis was performed in Python using *Scikit-survival* package.

3. Results and discussion

In this section, we first describe data used in this study: ADNI and CCC datasets. Then, we compare, qualitatively and quantitatively, the performance of different uncertainty methods combined with different classifiers using CCC dataset. Thereafter, we present the results obtained with the ensemble-based approach to target uncertainty of individual predictions (Fig. 1) using CCC and ADNI datasets. The set of rules used in the aggregator are drawn based on CCC results and validated with ADNI data.

3.1. Data

Participants were selected from two large prospective studies: ADNI project (http://adni.loni.usc.edu/) [33] and Cognitive Complaints Cohort (CCC) [32]. Participants with the clinical diagnosis of MCI at the baseline (first) assessment and who had at least one follow-up assessment were chosen. Demographic and neuropsychological data from different cognitive domains were selected in both datasets. Informed consent to participate in the study was obtained from all participants.

We followed the strategy to create learning instances using time windows described in [2]. Based on this, for a given time-window, we label patients that convert to dementia within a predefined interval (progressed from MCI to AD in one of the yearly assessments up until the limit of the window) as cMCI (converter MCI). Patients that didn't convert to AD during that period and presented a diagnosis of MCI at the limit of the window or afterward are included in the learning set labeled as sMCI (stable MCI). We chose a follow-up of 4-years since it corresponds to the maximum time width without skewed class proportions, for both CCC and ADNI datasets.

3.1.1. Cognitive complaints cohort

The Cognitive Complaints Cohort [32] is a prospective study conducted at the Faculty of Medicine of Lisbon to investigate the progression to dementia in subjects with cognitive complaints based on extensive neuropsychological evaluation. The inclusion criteria for admission to CCC were the presence of cognitive complaints and completing assessment with a neuropsychological battery, designed to evaluate multiple cognitive domains and validated for the Portuguese population (Bateria de Lisboa para Avaliação das Demências – BLAD [48]). The exclusion criteria for admission to CCC were a diagnosis of dementia (according to DSM-IV [49]) or other disorders that may cause cognitive impairment. For the purpose of this study, participants were diagnosed with Mild Cognitive Impairment when fulfilling the criteria of the MCI Working Group of the European Consortium on Alzheimer's disease [50]. Participants could later be diagnosed with AD according to the DSM-IV [49] criteria at follow-up.

The dataset included 41 features covering demographic and neuropsychological data. The feature selection approach described in [46] selected 20 features (we refer to [46] for more information on the selected features). From a total of 402 MCI patients, 227 (56%) patients remained stable and 175 (44%) converted to dementia within the follow-up period. Table 2 presents demographic characterization data. Converting patients are older and have fewer years of formal education than those who remained MCI while no statistical differences were found concerning gender.

3.1.2. ADNI

ADNI was launched in 2003 as a public-private partnership led by Principal Investigator Michael W. Weiner [33]. It aims at finding relevant biomarkers in all stages of AD to guide clinical trials for novel drugs or treatments. ADNI includes several biomarkers of Alzheimer's disease beyond neuropsychological tests, such as cerebrospinal fluid, structural Magnetic Resonance Imaging (MRI), functional-MRI, Positron Emission Tomography, and other biological data. Data is collected from every ADNI participant at the baseline assessment, as well as annual follow-up consultations. Participants were diagnosed with Mild Cognitive Impairment in the presence of a self-report (or via an informant) memory complaints without severe interference on daily live activities, objective memory deficit and absence of significant impairment on non-memory cognitive domains and of dementia. The NINCDS/ADRDA criteria were used to classify patients with probable AD.

In this work, we used 79 demographic and neuropsychological features from ADNI-2 patients (accessed in June 2017). After running the feature selection approach described in [46], 32 features were chosen (we refer to [46] for more information on the selected features). From a total of 265 MCI patients, 143 (54%) patients remained stable and 122 (46%) converted to dementia within the follow-up period. No differences were found in gender and years of formal education between converting and non-converting patients. On the other side, converting patients are older than those who remained MCI during the follow-up period.

3.2. Qualitative comparison between uncertainty methods using different classifiers

Fig. 3 illustrates the distribution of the estimated uncertainty values for correctly and incorrectly predictions as histograms, using six classifiers. Each histogram represents the level of uncertainty (i.e. the likelihood of the prediction being incorrect) to whom each patient is classified in the positive class ($y = 1$), as assessed by probabilities (PS, IR, and VAs), credibility (CPs) or direct probability estimates. Optimal results are those where correctly classified patients are either close to 0 or close to 1 and incorrectly classified cases are close to 0.5. The prediction space was discretized into fifty bins. For each bin, the number of

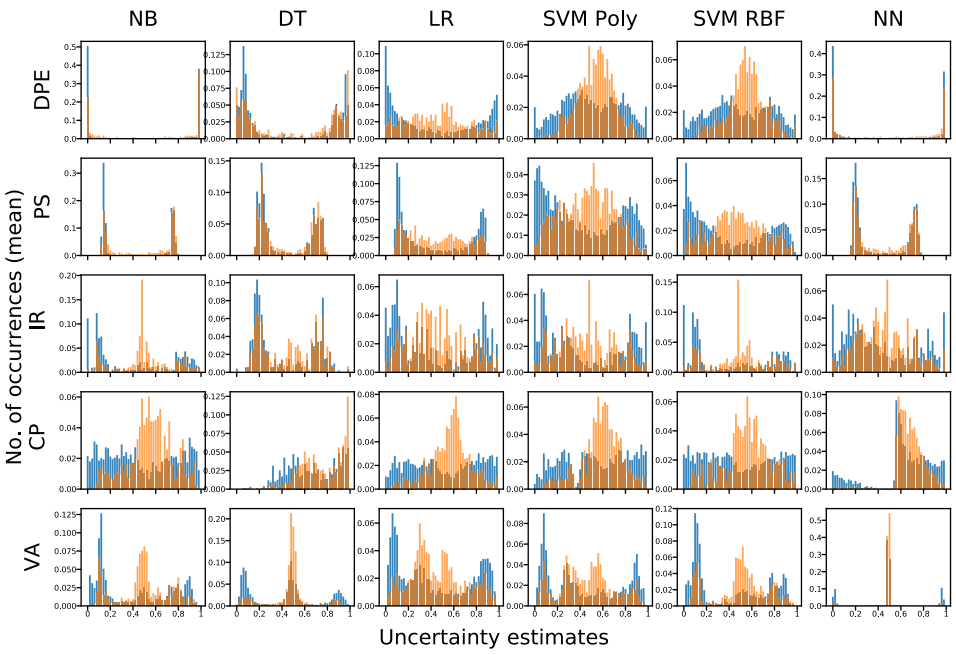

Fig. 3. Histogram plots showing the distribution of the uncertainty estimates for correctly (blue/dark) and incorrectly (orange/light) predictions for each classifier and uncertainty method. Values of Y-axis vary amongst the subplots to maximize the view in each subplot. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

occurrences was averaged over experiments (repetitions and CV-folds).

As expected, the posterior probabilities (name here as DPE) estimated by NB are shifted toward 0 and 1, for both correct and incorrect predictions, which is useless from a clinical decision point of view. Similar behavior was observed for NN and DT, although in the latter the two peaks have a wider distribution (Fig. 3). PS (and IR in the case of DT) pushed these peaks to the central region and is thus not adequate to calibrate predictions. IR represents a better choice to calibrate NB, as most of the incorrectly predicted values lie in the central region, while correctly predictions fall near 0 or 1. In the case of NB, VAs produced probability distributions similar to IR, while CPs (despite placing credibility values of most wrong predictions between 0.4 and 0.8) spread the correct prediction over the entire range of values. When coupled with NN, VAs moved the probability mass to the center, including nearly all the estimated probabilities associated with misclassified cases. Moreover, they isolate two peaks of correct predictions close to 0 and 1, despite compromising a small number of cases.

Conversely to NB and DT, the scores computed by SVM (here denoted as DPE) tend to lie in the central region with few predictions reaching 0 or 1. We expected Platt Scaling to effectively calibrate these probabilities, as maximum-margin algorithms produce sigmoid-shaped distortions. Indeed, PS moved the probability mass away from the middle prediction values, in particular those that correspond to correct predictions. However, post-calibration using IR, which corrects any monotonic distortion, achieved similar results. Moreover, VAs also produced a distribution of correct predictions with two peaks near 0 and 1, and the mass of incorrectly probabilities is mainly positioned in the center. Once again, CPs successfully identified wrongly classified cases as uncertain (peak of credibility values range from 0.4 to 0.8) but failed to push correct predictions toward the maximum (or minimum when the negative class is assigned) credibility values. Finally, LR produced well-calibrated probabilities and thus post-calibration methods (PS and IR) did not seem to help. CPs and VAs produced similar transformations of prediction scores for all classifiers. These methods shifted incorrectly classified cases to the center and thereby label those as uncertain, while forcing some correct predictions (assigned previously with high probability values by DPE) to middle-uncertainty estimates (close to 0.5). Our findings are in line with the results shown in [13,22,21].

Fig. 4 is analogous to Fig. 3 when using the probability of progression to AD within 4-years computed by the Cox Hazard Regression model. Two peaks of probability (of correct and incorrect predictions) are evidenced close to the tails. However, some misclassified cases are also spread over all probability values.

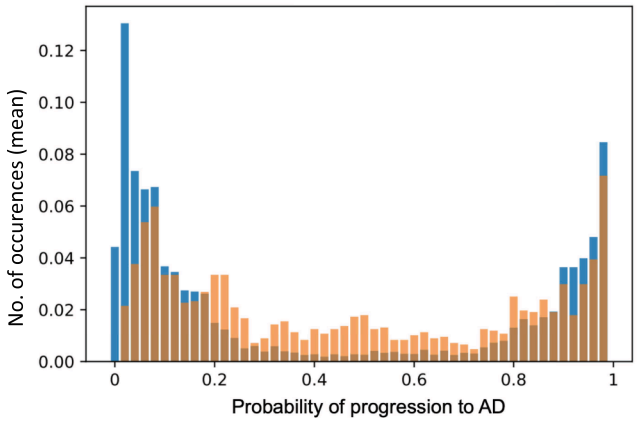

Fig. 4. Histogram with the probability of progression to AD within 4-years for correct (blue/dark) and incorrect (orange/light) predictions, computed by the Cox Hazard Regression model, using CCC. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.3. Quantitative comparison between uncertainty methods using different classifiers

We now discuss the performance of each prognostic model learned with different classifiers and uncertainty methods per threshold ($\tau \in 0.80, 0.90, 0.95$), as assessed by AUC, sensitivity, specificity, and number of predictions made above the predefined certainty thresholds (Tables 3–5, respectively). Statistically significant differences in the classification performance (AUC) amongst different uncertainty methods were found, per classifier, as assessed by the Friedman Test [47] ($p < 0.0005$). Pairwise comparisons (using the Wilcoxon Signed Rank Test [47]) were then performed (with Bonferroni correction for multiple testing, $p<0.002$) to compare such approaches across thresholds, per classifier.

When using NB as the underlying classifier, CPs and VAs

Table 3
AUC computed for each classifier and uncertainty method, per threshold, using CCC. The percentage of predictions above the respective threshold is reported within brackets.

	DPE	CP - FP	CP – PR	VA	Platt	IR
NB						
All	0.869 ± 0.0	0.869 ± 0.00	–	0.858 ± 0.0	0.825 ± 0.01	0.853 ± 0.0
0.80	0.876 ± 0.01 (96%)	0.933 ± 0.00 (40%)	0.936 ± 0.0 (39%)	0.885 ± 0.02 (57%)	0.614 ± 0.07 (49%)	0.885 ± 0.01 (65%)
0.90	0.884 ± 0.0 (90%)	0.964 ± 0.01 (21%)	0.969 ± 0.0 (19%)	0.957 ± 0.02 (17%)	0.0 ± 0.0 (0%)	0.901 ± 0.03(30%)
0.95	0.891 ± 0.0 (86%)	0.993 ± 0.01 (9%)	0.964 ± 0.0 (8%)	0.986 ± 0.51 (5%)	0.0 ± 0.0 (0%)	0.935 ± 0.05(12%)
DT						
	DPE	**CP - FP**	**CP – PR**	**VA**	**Platt**	**IR**
All	0.745 ± 0.02	0.697 ± 0.02	–	0.662 ± 0.01	0.732 ± 0.02	0.751 ± 0.02
0.80	0.748 ± 0.02 (87%)	0.699 ± 0.03 (77%)	0.692 ± 0.46 (1%)	0.878 ± 0.02 (43%)	0.502 ± 0.27 (12%)	0.444 ± 0.16 (31%)
0.90	0.726 ± 0.03 (61%)	0.651 ± 0.02 (53%)	0.0 ± 0.0 (0%)	0.869 ± 0.04 (27%)	0.0 ± 0.0 (0%)	0.287 ± 0.31 (2%)
0.95	0.664 ± 0.05 (30%)	0.591 ± 0.32 (31%)	0.0 ± 0.0 (0%)	0.775 ± 0.29 (6%)	0.0 ± 0.0 (0%)	0.154 ± 0.25 (1%)
LR						
	DPE	**CP - FP**	**CP – PR**	**VA**	**Platt**	**IR**
All	0.866 ± 0.01	0.866 ± 0.01	–	0.861 ± 0.01	0.849 ± 0.01	0.846 ± 0.01
0.80	0.916 ± 0.01 (61%)	0.939 ± 0.01 (42%)	0.951 ± 0.01 (31%)	0.919 ± 0.01 (50%)	0.895 ± 0.04 (58%)	0.901 ± 0.02 (56%)
0.90	0.931 ± 0.02 (41%)	0.959 ± 0.01 (21%)	0.969 ± 0.01 (13%)	0.925 ± 0.04 (26%)	0.570 ± 0.40 (4%)	0.900 ± 0.05 (25%)
0.95	0.943 ± 0.02 (27%)	0.968 ± 0.01 (10%)	0.976 ± 0.0 (5%)	0.921 ± 0.01 (6%)	0.0 ± 0.0 (0%)	0.817 ± 0.29 (11%)
SVM Poly						
	DPE	**CP - FP**	**CP – PR**	**VA**	**Platt**	**IR**
All	0.856 ± 0.01	0.859 ± 0.00	–	0.792 ± 0.01	0.852 ± 0.01	0.846 ± 0.01
0.80	0.953 ± 0.02 (23%)	0.931 ± 0.01 (42%)	0.946 ± 0.01 (26%)	0.907 ± 0.01 (43%)	0.914 ± 0.02 (46%)	0.903 ± 0.02 (54%)
0.90	0.990 ± 0.02 (9%)	0.956 ± 0.02 (21%)	0.980 ± 0.02 (9%)	0.910 ± 0.02 (28%)	0.925 ± 0.03 (23%)	0.918 ± 0.04 (29%)
0.95	0.997 ± 0.01 (5%)	0.975 ± 0.02 (11%)	1.0 ± 0.0 (3%)	0.980 ± 0.04 (5%)	0.928 ± 0.07 (11%)	0.929 ± 0.04 (12%)
SVM RBF						
	DPE	**CP - FP**	**CP – PR**	**VA**	**Platt**	**IR**
All	0.866 ± 0.01	0.868 ± 0.00	–	0.853 ± 0.00	0.868 ± 0.00	0.855 ± 0.00
0.80	0.939 ± 0.01 (27%)	0.925 ± 0.00 (41%)	0.928 ± 0.00 (39%)	0.868 ± 0.01 (59%)	0.910 ± 0.01 (51%)	0.882 ± 0.01 (65%)
0.90	0.957 ± 0.02 (10%)	0.952 ± 0.01 (20%)	0.954 ± 0.01 (20%)	0.907 ± 0.04 (17%)	0.915 ± 0.03 (27%)	0.902 ± 0.03 (29%)
0.95	0.945 ± 0.03 (5%)	0.959 ± 0.01 (11%)	0.961 ± 0.01 (9%)	0.288 ± 0.46 (4%)	0.956 ± 0.03 (13%)	0.882 ± 0.06 (12%)
NN						
	DPE	**CP - FP**	**CP – PR**	**VA**	**Platt**	**IR**
All	0.778 ± 0.02	0.772 ± 0.01	–	0.691 ± 0.02	0.765 ± 0.02	0.789 ± 0.02
0.80	0.794 ± 0.02 (89%)	0.866 ± 0.03 (47%)	0.945 ± 0.03 (18%)	0.932 ± 0.02 (22%)	0.589 ± 0.05 (15%)	0.867 ± 0.03 (39%)
0.90	0.802 ± 0.02 (83%)	0.902 ± 0.03 (21%)	0.949 ± 0.03 (12%)	0.932 ± 0.02 (21%)	0.0 ± 0.0 (0%)	0.891 ± 0.04 (16%)
0.95	0.811 ± 0.02 (77%)	0.917 ± 0.04 (11%)	0.945 ± 0.05 (7%)	0.932 ± 0.0 (21%)	0.0 ± 0.0 (0%)	0.872 ± 0.07 (10%)

outperformed the remaining methods in the task of predicting conversion from MCI to AD ($p < 0.0005$) for all certainty thresholds. No differences were found between CP-FP or CP-PR frameworks ($p < 0.861$), or between CPs and VAs ($p < 0.026$). However, CPs and VAs achieved a smaller number of predictions (Table 3). Not surprisingly, using a sigmoid transformation is not appropriate to calibrate NB models as PS attained the poorest results. On the other hand, IR obtained results equivalent to DPE ($p < 0.02$). All uncertainty methods, excepting for PS, produced good and well-balanced results in terms of sensitivity and specificity.

Of all classifiers, DT yielded the weakest classification results. Both calibration methods (PS and IR) and CP-PR failed to produce predictions at high-certainty levels ($\tau \in 0.90, 0.95$). This could be foreseen by observing Fig. 3 (2nd column, 2nd and 3rd rows), where the mass probability is placed in two peaks pulled to the center. VAs outperformed the remaining methods ($p < 0.0005$). However, they had the lowest AUC across all tested classifiers. In this context, in general, DT does not seem to be suitable for this prognostic problem.

LR is known to provide well-calibrated probabilities [13]. Indeed, calibration methods such as IR and PS did not offer any further benefit ($p < 0.0005$). Nevertheless, CPs outperformed DPE predictions in terms of classification performance (AUC, sensitivity, and specificity), at the cost of providing predictions for a smaller number of MCI patients (lower number of predictions for all thresholds). No significant differences were found between VAs and DPE ($p = 0.586$). However, once again, DPE showed the advantage of producing a higher number of predictions.

NN are also claimed to typically produce well-calibrated predictions [13]. However, in this work, the probabilities estimated by NN had similar behavior to those produced by NB (with values pushed towards 0 and 1). This miscalibration of NN was also reported in [51], with authors assigning this issue to the increase of model capacity (increase in width and depth of the neural network) and lack of regularization. Therefore, NN benefited from calibration with CP-PR and VAs,

Table 4
Sensitivity computed for each classifier and uncertainty method, per threshold, using CCC. The percentage of predictions above the respective threshold is reported within brackets.

NB

	DPE	CP - FP	CP – PR	VA	Platt	IR
All	0.789 ± 0.01	0.823 ± 0.01	–	0.760 ± 0.02	0.783 ± 0.01	0.751 ± 0.02
0.80	0.809 ± 0.01 (96%)	0.928 ± 0.01 (40%)	0.868 ± 0.01 (39%)	0.801 ± 0.02 (57%)	0.04 ± 0.144 (49%)	0.842 ± 0.02 (65%)
0.90	0.827 ± 0.01 (90%)	0.995 ± 0.01 (21%)	0.918 ± 0.01 (19%)	0.925 ± 0.06 (17%)	0.0 ± 0.0 (0%)	0.792 ± 0.06 (30%)
0.95	0.843 ± 0.00 (86%)	1.0 ± 0.0 (9%)	0.936 ± 0.01 (8%)	1.0 ± 0.0 (5%)	0.0 ± 0.0 (0%)	0.911 ± 0.12 (12%)

DT

	DPE	CP - FP	CP – PR	VA	Platt	IR
All	0.688 ± 0.03	0.710 ± 0.02	–	0.615 ± 0.02	0.668 ± 0.02	0.668 ± 0.02
0.80	0.713 ± 0.02 (87%)	0.658 ± 0.04 (77%)	0.825 ± 0.33 (1%)	0.768 ± 0.04 (43%)	0.050 ± 0.12 (12%)	0.449 ± 0.17 (31%)
0.90	0.703 ± 0.04 (61%)	0.612 ± 0.07 (53%)	0.0 ± 0.0 (0%)	0.773 ± 0.07 (27%)	0.0 ± 0.0 (0%)	0.287 ± 0.31 (2%)
0.95	0.761 ± 0.05 (30%)	0.556 ± 0.09 (31%)	0.0 ± 0.0 (0%)	0.623 ± 0.31 (6%)	0.0 ± 0.0 (0%)	0.167 ± 0.25 (1%)

LR

	DPE	CP - FP	CP – PR	VA	Platt	IR
All	0.741 ± 0.01	0.821 ± 0.01	–	0.729 ± 0.01	0.725 ± 0.01	0.699 ± 0.01
0.80	0.844 ± 0.02 (61%)	0.935 ± 0.02 (42%)	0.919 ± 0.02 (31%)	0.888 ± 0.03 (50%)	0.814 ± 0.01 (58%)	0.834 ± 0.04 (56%)
0.90	0.881 ± 0.03 (41%)	0.962 ± 0.02 (21%)	0.947 ± 0.03 (13%)	0.885 ± 0.05 (26%)	0.279 ± 0.40 (4%)	0.840 ± 0.12 (25%)
0.95	0.898 ± 0.03 (27%)	1.0 ± 0.0 (10%)	0.956 ± 0.05 (5%)	0.829 ± 0.14 (6%)	0.0 ± 0.0 (0%)	0.793 ± 0.29 (11%)

SVM Poly

	DPE	CP - FP	CP – PR	VA	Platt	IR
All	0.741 ± 0.01	0.835 ± 0.01	–	0.744 ± 0.01	0.743 ± 0.01	0.709 ± 0.02
0.80	0.969 ± 0.02 (23%)	0.937 ± 0.01 (42%)	0.925 ± 0.02 (26%)	0.860 ± 0.03 (43%)	0.835 ± 0.03 (46%)	0.846 ± 0.03 (54%)
0.90	0.987 ± 0.03 (9%)	0.961 ± 0.01 (21%)	0.965 ± 0.04 (9%)	0.837 ± 0.05 (28%)	0.826 ± 0.09 (23%)	0.837 ± 0.10 (29%)
0.95	0.992 ± 0.02 (5%)	0.991 ± 0.02 (11%)	1.0 ± 0.0 (3%)	0.956 ± 0.07 (5%)	0.849 ± 0.15 (11%)	0.920 ± 0.06 (12%)

SVM RBF

	DPE	CP - FP	CP – PR	VA	Platt	IR
All	0.737 ± 0.00	0.834 ± 0.01	–	0.793 ± 0.02	0.733 ± 0.03	0.753 ± 0.03
0.80	0.954 ± 0.01 (27%)	0.932 ± 0.00 (41%)	0.868 ± 0.01 (39%)	0.782 ± 0.02 (59%)	0.835 ± 0.03 (51%)	0.806 ± 0.06 (65%)
0.90	0.995 ± 0.01 (10%)	0.955 ± 0.01 (20%)	0.923 ± 0.01 (20%)	0.797 ± 0.12 (17%)	0.816 ± 0.08 (27%)	0.818 ± 0.07 (29%)
0.95	1.0 ± 0.0 (5%)	1.0 ± 0.0 (11%)	0.917 ± 0.03 (9%)	0.30 ± 0.48 (4%)	0.764 ± 0.02 (13%)	0.788 ± 0.29 (12%)

NN

	DPE	CP - FP	CP – PR	VA	Platt	IR
All	0.677 ± 0.03	0.799 ± 0.02	–	0.615 ± 0.10	0.685 ± 0.02	0.618 ± 0.05
0.80	0.693 ± 0.02 (89%)	0.846 ± 0.03 (47%)	0.938 ± 0.04 (18%)	0.924 ± 0.04 (21%)	0.0 ± 0.0 (15%)	0.798 ± 0.07 (39%)
0.90	0.708 ± 0.03 (93%)	0.865 ± 0.05 (22%)	0.946 ± 0.05 (12%)	0.924 ± 0.04 (21%)	0.0 ± 0.0 (0%)	0.846 ± 0.11 (16%)
0.95	0.718 ± 0.03 (77%)	0.859 ± 0.06 (11%)	0.939 ± 0.05 (7%)	0.922 ± 0.04 (21%)	0.0 ± 0.0 (0%)	0.834 ± 0.15 (10%)

outperforming the remaining uncertainty methods ($p < 0.001$). IR and CP-FP methods performed equally ($p = 0.041$).

CP-PR and DPE achieved the highest classification performance over the three thresholds when using SVM with the polynomial kernel as the underlying classifier ($p < 0.0005$), despite the small number of predictions. CP-FP, VAs, and PS, whose performance was statistically comparable ($p < 0.909$), followed. When using SVM with the Gaussian (RBF) kernel, PS performed as well as CPs and DPE, obtaining a higher number of predictions. In this case, and as expected, PS is more suitable to calibrate SVM predictions than IR ($p < 0.0005$). No differences were found between IR and VAs ($p < 0.057$). The calibration methods and VAs yielded, in general, slightly higher values of specificity than sensitivity, while CP-FP and DPE showed the inverse relationship between sensitivity and specificity.

In the literature, VAs proved to produce improved calibrated probabilities when compared to calibration methods (PS and IR), as assessed by Brier Loss and Log Loss scores, using a variety of classifiers and datasets [19,21,22]. Moreover, DPE outperformed IR sometimes, mainly when using Bagging DT, LR, and NN [21]. Comparisons between VAs and CPs are scarce. In a report [26], authors compare the performance between conformal and probabilistic results on predicting bioactivity with an estimate of its uncertainty. In particular, they used (confident) p-values and (VAs) probabilities to rank instances for the purpose of screening. CPs and VAs produced similar results. Still, the ranking obtained from CPs had a finer granularity than that outputted by VAs, in the sense that more instances shared the same probability than the same p-value.

Table 6 reports the classification performance obtained using the Cox Proportional Hazard Regression model. It achieved comparable results to VAs ($p < 0.052$) and PS ($p < 0.011$) built on top of DT and LR, respectively. Moreover, it produced superior performances to PS and VAs, when using NB and NN, and SVM RBF, respectively ($p > 0.002$). Finally, it underperformed the remaining pairs of (classifier, uncertainty method). In [52], this survival method also underperformed

Table 5
Specificity computed for each classifier and uncertainty method, per threshold, using CCC. The percentage of predictions above the respective threshold is reported within brackets.

	NB					
	DPE	**CP - FP**	**CP – PR**	**VA**	**Platt**	**IR**
All	0.788 ± 0.0	0.786 ± 0.00	–	0.784 ± 0.01	0.786 ± 0.01	0.777 ± 0.01
0.80	0.807 ± 0.01 (96%)	0.897 ± 0.00 (40%)	0.901 ± 0.01 (39%)	0.886 ± 0.02 (57%)	0.792 ± 0.02 (49%)	0.882 ± 0.01 (65%)
0.90	0.824 ± 0.00 (90%)	0.957 ± 0.01 (21%)	0.957 ± 0.01 (19%)	0.935 ± 0.03 (17%)	0.0 ± 0.0 (0%)	0.897 ± 0.01(30%)
0.95	0.836 ± 0.00 (86%)	0.981 ± 0.02 (9%)	0.968 ± 0.00 (8%)	0.968 ± 0.02 (5%)	0.0 ± 0.0 (0%)	0.940 ± 0.02(12%)
	DT					
	DPE	**CP - FP**	**CP – PR**	**VA**	**Platt**	**IR**
All	0.729 ± 0.01	0.673 ± 0.02	–	0.658 ± 0.02	0.727 ± 0.01	0.719 ± 0.02
0.80	0.755 ± 0.02 (87%)	0.702 ± 0.03 (77%)	0.926 ± 0.33 (1%)	0.861 ± 0.02 (43%)	0.670 ± 0.16 (12%)	0.759 ± 0.04 (31%)
0.90	0.769 ± 0.02 (61%)	0.645 ± 0.04 (53%)	0.0 ± 0.0 (0%)	0.898 ± 0.02 (27%)	0.0 ± 0.0 (0%)	0.0 ± 0.0 (2%)
0.95	0.737 ± 0.04 (30%)	0.566 ± 0.07 (31%)	0.0 ± 0.0 (0%)	0.0 ± 0.0 (6%)	0.0 ± 0.0 (0%)	0.0 ± 0.0 (1%)
	LR					
	DPE	**CP - FP**	**CP – PR**	**VA**	**Platt**	**IR**
All	0.788 ± 0.01	0.766 ± 0.01	–	0.795 ± 0.01	0.779 ± 0.02	0.772 ± 0.02
0.80	0.892 ± 0.01 (61%)	0.916 ± 0.02 (42%)	0.935 ± 0.01 (31%)	0.904 ± 0.02 (50%)	0.877 ± 0.01 (58%)	0.875 ± 0.02 (56%)
0.90	0.922 ± 0.02 (41%)	0.949 ± 0.01 (21%)	0.961 ± 0.01 (13%)	0.932 ± 0.02 (26%)	0.0 ± 0. 0 (4%)	0.912 ± 0.03 (25%)
0.95	0.935 ± 0.02 (27%)	0.961 ± 0.01 (10%)	0.976 ± 0.03 (5%)	0.923 ± 0.03 (6%)	0.0 ± 0.0 (0%)	0.939 ± 0.04 (11%)
	SVM Poly					
	DPE	**CP - FP**	**CP – PR**	**VA**	**Platt**	**IR**
All	0.789 ± 0.01	0.762 ± 0.01	–	0.791 ± 0.01	0.783 ± 0.01	0.774 ± 0.01
0.80	0.943 ± 0.02 (23%)	0.913 ± 0.01 (42%)	0.928 ± 0.02 (26%)	0.894 ± 0.01 (43%)	0.897 ± 0.02 (46%)	0.875 ± 0.01 (54%)
0.90	0.982 ± 0.03 (9%)	0.941 ± 0.02 (21%)	0.945 ± 0.02 (9%)	0.907 ± 0.02 (28%)	0.928 ± 0.02 (23%)	0.914 ± 0.03 (29%)
0.95	0.992 ± 0.02 (5%)	0.966 ± 0.02 (11%)	1.0 ± 0.0 (3%)	0.974 ± 0.04 (5%)	0.954 ± 0.04 (11%)	0.934 ± 0.04 (12%)
	SVM RBF					
	DPE	**CP - FP**	**CP – PR**	**VA**	**Platt**	**IR**
All	0.788 ± 0.01	0.792 ± 0.00	–	0.781 ± 0.00	0.787 ± 0.01	0.785 ± 0.01
0.80	0.920 ± 0.02 (27%)	0.901 ± 0.00 (41%)	0.902 ± 0.00 (39%)	0.885 ± 0.00 (59%)	0.899 ± 0.01 (51%)	0.877 ± 0.01 (65%)
0.90	0.959 ± 0.02 (10%)	0.938 ± 0.01 (20%)	0.945 ± 0.01 (20%)	0.897 ± 0.03 (17%)	0.929 ± 0.01 (27%)	0.909 ± 0.02 (29%)
0.95	0.950 ± 0.02 (5%)	0.954 ± 0.01 (11%)	0.961 ± 0.01 (9%)	0.978 ± 0.04 (4%)	0.948 ± 0.02 (13%)	0.927 ± 0.02 (12%)
	NN					
	DPE	**CP - FP**	**CP – PR**	**VA**	**Platt**	**IR**
All	0.711 ± 0.01	0.679 ± 0.01	–	0.604 ± 0.04	0.732 ± 0.01	0.723 ± 0.01
0.80	0.738 ± 0.01 (89%)	0.815 ± 0.03 (47%)	0.938 ± 0.03 (18%)	0.934 ± 0.03 (22%)	0.704 ± 0.06 (15%)	0.857 ± 0.02 (39%)
0.90	0.753 ± 0.01 (83%)	0.867 ± 0.04 (22%)	0.948 ± 0.03 (12%)	0.934 ± 0.03 (21%)	0.0 ± 0.0 (0%)	0.881 ± 0.04 (16%)
0.95	0.763 ± 0.01 (77%)	0.885 ± 0.04 (11%)	0.949 ± 0.04 (7%)	0.933 ± 0.03 (21%)	0.0 ± 0.0 (0%)	0.881 ± 0.01 (10%)

Table 6
Results of the Cox Proportional Hazard Regression model and respective probability of progression to AD used to measure the uncertainty of predictions, per threshold, using CCC. The percentage of predictions above the respective threshold is reported within brackets.

τ	**AUC**	**Sensitivity**	**Specificity**
All	0.783 ± 0.1	0.743 ± 0.01	0.850 ± 0.01
0.80	0.826 ± 0.01 (54%)	0.858 ± 0.01	0.919 ± 0.01
0.90	0.845 ± 0.01 (34%)	0.898 ± 0.01	0.931 ± 0.01
0.95	0.863 ± 0.02 (18%)	0.952 ± 0.01	0.917 ± 0.01

other commonly used classifiers when predicting the need for non-invasive ventilation in patients with Amyotrophic Lateral Sclerosis. In this context, we decided to pursue the analysis without this method.

In a nutshell, if we would have to select a single method to output calibrated probability estimates for a similar classification problem and data, independently of the classifier, it would be CP-PR. It outperformed the remaining methods for most classifiers and enjoys guaranteed validity (perfect calibration). VAs are also a valid option given that often achieved results comparable to CPs while outputting a probability, which is easily interpretable. We recall that while CPs answer the question: "What is the probability of a given instance, being as, or more, contrary than the test instance to the randomness assumption, given the training set?", VAs address the following question: "What is the probability of assigning a certain class to the test instance, given the training set?". If users prefer a higher number of predictions rather than validity, then the well-calibrated probability estimates outputted by LR are a better option.

3.4. Ensemble-based approach to target the uncertainty of predictions at patient-level

In this Section, we report the results of the proposed ensemble-based

Table 7
Results obtained with the ensemble-based approach to target predictions uncertainty using different rules in the aggregator step, per threshold, using CCC data. The percentage of predictions above the respective threshold are reported within brackets.

AUC				
τ	**R1: All**	**R2: Max**	**R3: All but some**	**R4: CPs**
0.80	0.724 ± 0.06 (99%)	0.882 ± 0.04 (64%)	0.852 ± 0.08 (86%)	0.917 ± 0.03 (52%)
0.90	0.757 ± 0.06 (98%)	0.902 ± 0.04 (40%)	0.866 ± 0.07 (53%)	0.947 ± 0.05 (28%)
0.95	0.793 ± 0.08 (95%)	0.912 ± 0.10 (13%)	0.867 ± 0.08 (34%)	0.929 ± 0.09 (14%)
Sensitivity				
τ	**R1: All**	**R2: Max**	**R3: All but some**	**R4: CPs**
0.80	0.713 ± 0.14 (99%)	0.836 ± 0.11 (64%)	0.816 ± 0.19 (86%)	0.941 ± 0.07 (52%)
0.90	0.714 ± 0.14 (98%)	0.829 ± 0.13 (40%)	0.854 ± 0.16 (53%)	0.936 ± 0.06 (28%)
0.95	0.731 ± 0.12 (95%)	0.971 ± 0.06 (13%)	0.839 ± 0.16 (34%)	1.0 ± 0.0 (14%)
Specificity				
τ	**R1: All**	**R2: Max**	**R3: All but some**	**R4: CPs**
0.80	0.756 ± 0.05 (99%)	0.854 ± 0.04 (64%)	0.815 ± 0.08 (86%)	0.903 ± 0.03 (52%)
0.90	0.757 ± 0.06 (98%)	0.901 ± 0.05 (40%)	0.902 ± 0.07 (53%)	0.940 ± 0.02 (28%)
0.95	0.768 ± 0.05 (95%)	0.948 ± 0.05 (13%)	0.905 ± 0.06 (34%)	0.958 ± 0.06 (14%)

approach, as assessed by the averaged AUC over the outer CV loop, using CCC and ADNI data. Four rules are used in the aggregator step. Two of these are general rules, independent of the results discussed in Section 3.3: *(1)* using all pairs of classifiers and methods to assess the uncertainty of predictions (labeled *"R1: All"*), and *(2)* using the uncertainty method with highest AUC per classifier (labeled *"R2: Max"*). The two additional rules were defined in the aftermath of results discussed in Section 3.3. Namely, rule *(3)* uses *1)* but excluding pairs (classifier, uncertainty method) whose results are known to harm the performance of the ensemble, namely, using DT, DPE with NB and NN, and PS with NB and NN (labeled *"R3: All but some"*). Rule 4 uses CPs (PR and FP frameworks) built on top of NB, LR and SVM (Poly and RBF), and is labeled as *"R4: CPs"*. We excluded DT as it underperforms all other classifiers. Although these rules are drawn using CCC data (Section 3.3), we extrapolate them to ADNI data. We note, however, that other rules may be tested to improve the results. We recall that the final prediction is that with the highest uncertainty value among the uncertainty methods and classifiers tested.

3.4.1. Using CCC data

Table 7 reports the results obtained with the ensemble-based approach to target uncertainty of predictions for each rule implemented in the aggregator step, per threshold, using CCC data. The best pairs according to the rule "R2: Max" are the following: <(NB, CP-PR), (LR, CP-PR), (DT, VAs), (SVM Poly, CP-PR), (SVM RBF, DPE), (NN, CP-PR)>.

Not surprisingly, feeding the ensemble with all pairs of classifiers and uncertainty methods led to the poorest results. This is possibly due to the uncalibrated probabilities estimated by NB and NN, which, as illustrated in Fig. 3, assumed values close to 0 and 1, even for misclassified cases. By removing this source of noise, using rule "R3: All but some", the classification performance improved, although with a drop in the number of predictions made. Compared to Table 3, these results are competitive with those obtained with LR using DPE, being superior to the remaining, regarding the balance between AUC and the number of predictions. Some predictions are made by CPs or VAs, enjoying guaranteed validity, representing an advantage over the DPE produced by LR.

The ensemble built exclusively with CPs, using different classifiers, slightly increased the number of predictions and decreased AUC, when compared to results obtained with CPs individually (Table 3). Finally, using the uncertainty method with the highest AUC per classifier ("R2: Max") seems to be a valid alternative to the ensemble using rule "R3: All but some", since it improved AUC, despite having a lower number of predictions. A trade-off between the number of predictions and classification performance of such predictions should be taken into consideration when choosing the best settings to be applied into clinical practice. As an illustrative example, Table 8 shows the output of the ensemble for some patients, picked randomly.

3.4.2. Using ADNI data

The ensemble-based approach, with rules derived for CCC data, was validated using ADNI data, a similar and publicly available dataset (Table 9). The pairs (classifier, uncertainty method) selected accordingly to rule "R2: Max", using ADNI, were the following: <(NB, CP-PR), (LR, CP-PR), (DT, VAs), (SVM Poly, DPE), (SVM RBF, DPE), (NN, CP-PR)>. The overall results are in concordance with the conclusions drawn for CCC data. Again, using all pairs of classifiers and methods to build the ensemble conveyed the worst results. Using rule "R3: All but some" represents a good compromise between a higher number of predictions and an acceptable classification performance. Using rule "R2: Max" yielded similar results to using only CPs in the ensemble, despite the latter having guaranteed validity.

The validation using ADNI data shows that the proposed ensemble-based approach to target uncertainty of predictions can be used with similar datasets and prognostic problems.

4. Conclusions

The main contributions of this work are threefold:

- An outright comparison between different methods to target uncertainty of predictions at patient-level, using different classifiers. The analysis was performed with a clinical dataset (CCC), including demographic and neuropsychological tests and around 400 MCI

Table 8
Example of output from the ensemble-based approach to target uncertainty of individual instances, using $\tau = 0.90$ and CCC.

Patient ID	Predicted class	Uncertainty estimate	Classifier & Uncertainty method
802	sMCI	0.949	LR & DPE
722	cMCI	0.918	NB & CP-PR
17	sMCI	0.915	LR & VAs

Table 9
Results obtained with the ensemble-based approach to target predictions uncertainty using different rules in the aggregator step, per threshold, using ADNI data. The percentage of predictions above the respective threshold are reported within brackets.

AUC

τ	R1: All	R2: Max	R3: All but some	R4: CPs
0.80	0.742 ± 0.03 (100%)	0.942 ± 0.08 (59%)	0.862 ± 0.03 (97%)	0.949 ± 0.03 (57%)
0.90	0.757 ± 0.05 (99%)	0.932 ± 0.08 (43%)	0.890 ± 0.04 (67%)	0.979 ± 0.03 (32%)
0.95	0.797 ± 0.03 (97%)	0.968 ± 0.07 (23%)	0.931 ± 0.04 (49%)	0.976 ± 0.05 (21%)

Sensitivity

τ	R1: All	R2: Max	R3: All but some	R4: CPs
0.80	0.743 ± 0.09 (100%)	0.861 ± 0.16 (59%)	0.762 ± 0.15 (97%)	0.868 ± 0.04 (57%)
0.90	0.760 ± 0.09 (99%)	0.876 ± 0.16 (43%)	0.827 ± 0.09 (67%)	0.985 ± 0.03 (32%)
0.95	0.756 ± 0.09 (97%)	0.900 ± 0.22 (23%)	0.860 ± 0.09 (49%)	1.0 ± 0.0 (21%)

Specificity

τ	R1: All	R2: Max	R3: All but some	R4: CPs
0.80	0.779 ± 0.03 (100%)	0.904 ± 0.07 (59%)	0.834 ± 0.03 (97%)	0.908 ± 0.02 (57%)
0.90	0.789 ± 0.03 (99%)	0.926 ± 0.08 (43%)	0.898 ± 0.03 (67%)	0.969 ± 0.04 (32%)
0.95	0.789 ± 0.02 (97%)	0.959 ± 0.09 (23%)	0.920 ± 0.04 (49%)	0.960 ± 0.06 (21%)

patients.

- An ensemble-based approach combining different classifiers and methods to target uncertainty of predictions with the aim of optimizing the quality and quantity of predictions made. Two datasets were used to validate this approach, CCC and ADNI.
- A new conformity measure for SVM.

Most classifiers produce predictions for new instances accurately, without providing a reliable measure of how uncertain predictions are. In the medical domain, this hampers their integration in decision support systems, which could be useful in the clinical practice. Some strategies have been proposed to calibrate predictions (such as Platt Scaling or Isotonic Regression) or complement predictions with theoretical-backed measures to assess their risk of error (Venn-ABERS or Conformal Predictors). The usefulness of these methods depends on the classifier and data under study, and their choice is not trivial. Despite some of these methods have been compared in the literature [19,21,22,26], to our knowledge, no comparison was made using all methods in a common dataset. Moreover, usually large datasets were studied, while clinical data usually have a reduced number of instances.

According to our experimental results, Platt Scaling is only adequate to calibrate SVM's output, which produces a characteristic sigmoid distortion in their predictions. In contrast with [13], Isotonic Regression performed better than PS for the remaining classifiers, even when using small calibration sets. Despite being superior to PS, calibrating DT scores using IR produced poor results. In fact, DT underperformed all the remaining classifiers, across all methods, showing not to be appropriate to learn the prognostic problem under study. In contrast, LR proved to be well-calibrated, thus outperforming both calibration methods (DPE + LR). CPs and VAs produce similar transformations of the prediction scores among the classifiers. CPs typically pushed the credibility mass values of incorrect predictions to a range of values between 0.4 and 0.8. However, the correctly classified cases are in general spread all over the histogram. In this context, CPs lessen false negative and false positives values and, despite in small number, produce predictions with a low error rate for high credibility thresholds. CPs successfully complement predictions with credibility scores (or within confident prediction regions) for all classifiers, except for DT. Likewise, VAs pushed uncertainty estimates associated with misclassified cases to the center and formed two peaks near the histogram ends (0 and 1) of correctly classified cases. Still, it was underperformed by CPs built on top of LR and SVM RBF, since it failed on further extending the tails closer to 0 and 1. Therefore, for high-certainty thresholds, namely for $\tau = 0.95$, it makes no predictions. With this in mind, and given their guaranteed validity, CPs and VAs are preferable methods to complement predictions with a measure of uncertainty. VAs, outputting a probability, may have an advantage over CPs, which has a less interpretable outcome (p-values). Moreover, considering the simplicity of the approach, and the higher number of predictions made, using the probabilities estimated by LR is also a competitive approach.

In this work, we verified that the drawback of leaving a large number of patients without prognostic (those who cannot be predicted given the certainty threshold) was transversal to most classifiers and uncertainty methods evaluated. To tackle this problem, we proposed an ensemble-based approach to combine predictions from multiple pairs of (classifier, uncertainty method), made at high certainty thresholds, thus supplying a reliable prognosis to more patients. The goal was to maximize the number of predictions outputted by the model, taking advantage of all certain predictions made by the different methods, by aggregating individual predictions using the ensemble. The number of predictions made by the ensemble was, in fact, higher than the number of predictions possible when individual classifiers and uncertainty methods were used. Nevertheless, the considerable number of patients without a final prognostic still warrants considerations. We note however that the number of predictions the model is able to make will always be dependent on the complexity of the problem, the quality of the data used to learn the models and the learning capacity of the base learners (both classifiers and uncertainty methods) used in the ensemble. For example, in our problem, we used two datasets. In CCC dataset a maximum of 14% of cases were predicted at the highest certainty threshold and using rule "R4: CPs" (as opposed to 11% of predictions when using CPs individually) while in ADNI 21% of cases were predicted. This clearly showcases that maintaining the base learning, the ensemble always improves the number of predictions, but the increase in performance depends on the dataset. With this in mind, more powerful learners (and other parameters) can always be used together with more/other data. Another possibility could be to study whether the inclusion of other non-conformity measures can increase CPs efficiency. Moreover, as future work, we plan to implement the inductive version of CPs and VAs to make this approach usable for higher dimensional datasets.

A trade-off between the number of predictions and classification performance should be taken into consideration when choosing the best settings to be applied to clinical practice. For prognostic prediction in

AD, clinicians would rather reduce the number of prediction errors at patient-level than increasing the number of predictions made. Based on that, clinicians could withhold the decision about future dementia status for unreliable predictions and would promote the selection of patients at high risk of conversion for clinical trials for reliable predictions. In this context, classification performance gains priority over the number of predictions. In this scenario, rules "R2: Max" and "R4: CPs" should be preferred. The former selects the uncertainty method that optimizes AUC per classifier. It is a general rule (independent of the previous evaluation of results) and produces high classification performances for a number of predictions superior to those outputted when using rule "R4: CPs". The ensemble built exclusively with CPs ("R4: CPs") has the advantage of having guaranteed validity and achieving the highest classification performance, at the cost of providing the lowest number of predictions. In the case of setting the certainty threshold as 0.95, "R4: CPs" must be used, since it yields higher classification performances, for a number of predictions similar to those produced by rule "R2: Max". Rule "R1: All", where all pairs of classifiers and methods are used to build the ensemble should be avoid, since it consistently convey the worst results. If users focus on a higher number of predictions rather than classification performance, rule "R3: All but some" should be preferred. By removing the pairs of (classifier, uncertainty method) that hampers the predictive ability of the model, it represents a good compromise between a higher number of predictions and an acceptable classification performance.

5. Declarations

5.1. Ethics approval and consent to participate

The study regarding the recruitment of CCC participants was conducted in accordance with the Declaration of Helsinki and was approved by the local (Faculty of Medicine, University of Lisbon) ethics committee. Data access was granted in the context of project NEUROCLINOMICS2 (PTDC/EEI-SII/1937/2014) where the authors' institutions participate. The ADNI study was conducted according to the Good Clinical Practice guidelines, the Declaration of Helsinki, and US 21 CFR: Part 50 (Protection of Human Subjects) and Part 56 (Institutional Review Boards). The ADNI study was conducted in compliance with HIPAA regulations. Written informed consent to participate in the study was obtained from all (CCC and ADNI) participants and/or authorized representatives. Data access was de-identified on both studies.

5.2. Availability of data and material

CCC data is not shared with outside institutions while ADNI data is publicly available (http://adni.loni.usc.edu). Java implementation of the ensemble-based approach to target uncertainty of predictions at patient-level is publicly available.

5.3. Funding

This work was supported by FCT through funding of Neuroclinomics2 (PTDC/EEI-SII/1937/2014), Predict (PTDC/CCI-CIF/29877/2017), research grant (SFRH/BD/95846/2013) to TP, LASIGE Research Unit, ref. UID/CEC/00408/2019, and iCare4U (PTDC/EME-SIS/31747/2017 and LISBOA-01-0145-FEDER-03474).

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Acknowledgments

The authors thank the facilities provided by Memoclínica. Moreover, we thank professor Alexander Gammerman and Paolo Toccaceli for their valuable discussion around Conformal Predictions and Venn-ABERS subjects. This work was supported by FCT through funding of Neuroclinomics2 project, ref. PTDC/EEI-SII/1937/2014, research grant (SFRH/ BD/95846/2013) to TP and LASIGE Research Unit, ref. UID/CEC/00408/2019. Data collection and sharing for this project was funded by the Alzheimers Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimers Association; Alzheimers Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimers Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for NeuroImaging at the University of Southern California.

References

[1] P. Scheltens, K. Blennow, M.M.B. Breteler, B. de Strooper, G.B. Frisoni, S. Salloway, W.M. Van der Flier, Alzheimer's disease, The Lancet 388 (10043) (2016) 505–517, https://doi.org/10.1016/S0140-6736(15)01124-1 http://linkinghub.elsevier.com/retrieve/pii/S0140673615011241.

[2] T. Pereira, L. Lemos, S. Cardoso, D. Silva, A. Rodrigues, I. Santana, A. de Mendonça, M. Guerreiro, S.C. Madeira, Predicting progression of mild cognitive impairment to dementia using neuropsychological data: a supervised learning approach using time windows, BMC Med. Inform. Decis. Mak. 17 (1) (2017) 110, https://doi.org/10.1186/s12911-017-0497-2.

[3] S.I. Dimitriadis, D. Liparas, M.N. Tsolaki, Random forest feature selection, fusion and ensemble strategy: Combining multiple morphological MRI measures to discriminate among healhy elderly, MCI, cMCI and alzheimer's disease patients: From the alzheimer's disease neuroimaging initiative (ADNI) data, J. Neurosci. Methods (2017) 1–10, https://doi.org/10.1016/j.jneumeth.2017.12.010.

[4] M. Grassi, G. Perna, D. Caldirola, K. Schruers, R. Duara, D.A. Loewenstein, A clinically-translatable machine learning algorithm for the prediction of Alzheimer's disease conversion in individuals with mild and premild cognitive impairment, J. Alzheimer's Disease 61 (4) (2018) 1555–1573, https://doi.org/10.3233/JAD-170547.

[5] P. Flach, Machine Learning: The Art and Science of Algorithms that Makes Sense of Data, 1st Edition, Cambridge University Press, 2012.

[6] F. Provost, P. Domingos, Tree induction for probability based ranking, Mach. Learn. 52 (3) (2003) 199–215.

[7] B. Zadrozny, C. Elkan, Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers, Icml (May) (2001) 1–8 http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.29.3039&rep=rep1&type=pdf.

[8] P.N. Bennett, Assessing the Calibration of Naive Bayes' Posterior Estimates, Tech. rep., Computer Science Department, School of Computer Science, Carnegie Mellon University, 2000.

[9] N. Chawla, D. Cieslak, Evaluating probability estimates from decision trees, American Association for Artificial Intelligence Workshop, 2006, pp. 18–23. <http://www.aaai.org/Papers/Workshops/2006/WS-06-06/WS06-06-005.pdf>.

[10] U. Johansson, H. Bostrom, T. Lofstrom, Conformal Prediction Using Decision Trees, IEEE 13th International Conference on Data Mining, 2013, pp. 330–339, , https://doi.org/10.1109/ICDM.2013.85.

[11] M.H. DeGroot, S.E. Fienberg, The comparison and evaluation of forecasters, J. Roy. Stat. Soc.: Series D (The Stat.) 32 (1983) 12–22.

[12] T. Pereira, F. Ferreira, A.D. Mendonça, M. Guerreiro, S.C. Madeira, Towards a reliable prediction of conversion from Mild Cognitive Impairment to Alzheimer's Disease: stepwise learning using time windows, in: S. Fodeh, D.S. Raicu (Eds.), Proceedings of The First Workshop Medical Informatics and Healthcare held with the 23rd SIGKDD Conference on Knowledge Discovery and Data Mining, PMLR,

2017, pp. 19–26.
[13] A. Niculescu-Mizil, R. Caruana, Predicting good probabilities with supervised learning, Proceedings of the 22nd international conference on Machine learning (ICML), 2005, pp. 625–632, , https://doi.org/10.1145/1102351.1102430.
[14] E.P. Costa, S. Verwer, H. Blockeel, Estimating prediction certainty in decision trees, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 8207 LNCS, 2013, pp. 138–149. https://doi.org/10.1007/978-3-642-41398-8_13.
[15] R. Hérault, Y. Grandvalet, Sparse probabilistic classifiers, in: Proceedings of the 24th International Conference on Machine Learning, ICML '07, ACM, New York, NY, USA, 2007, pp. 337–344. doi:10.1145/1273496.1273539. URL http://doi.acm.org/10.1145/1273496.1273539.
[16] M. Fauvel, J. Chanussot, J. Benediktsson, A Combined Support Vector Machines Classification Based on Decision Fusion, IEEE International Geoscience and Remote Sensing Symposium, 2006, pp. 2494–2497, , https://doi.org/10.1109/IGARSS.2006.645 http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1704969.
[17] J.C. Platt, Probabilistic outputs for SVM and comparisons to regularized likelihood methods.
[18] B. Zadrozny, C. Elkan, Transforming classifier scores into accurate multiclass probability estimates, Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM New York, New York, USA, 2002, pp. 694–699.
[19] V. Manokhin, Multi-class probabilistic classification using inductive and cross Venn – Abers predictors, in: Proceedings of the Sixth Workshop on Conformal and Probabilistic Prediction and Applications, no. 60, 2017, pp. 228–240.
[20] C. Zhou, Conformal and Venn Predictors for Multi-probabilistic Predictions and Their Applications, Ph.D. thesis University of London, Royal Holloway, 2015.
[21] V. Vovk, I. Petej, Venn-Abers Predictors, in: Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence, 2014, pp. 829–838.
[22] V. Vovk, I. Petej, V. Fedorova, Large-scale probabilistic predictors with and without guarantees of validity, in: C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, R. Garnett (Eds.), Advances in Neural Information Processing Systems 28, Curran Associates Inc, 2015, pp. 892–900. arXiv:1511.00213.
[23] V. Vovk, Venn predictors and isotonic regression, CoRR abs/1211.0025.
[24] V. Vovk, A. Gammerman, G. Shafer, Algorithmic Learning in a Random World, Springer, New York, 2005.
[25] T. Pereira, S. Cardoso, D. Silva, A.D. Mendonça, M. Guerreiro, S.C. Madeira, Trustworthy predictions of conversion from mild cognitive impairment to dementia : a conformal prediction approach, in: Inter. Conference on Practical Applications of Computational Biology & Bioinformatics, Porto, 2017.
[26] P. Toccaceli, I. Nouretdinov, Z. Luo, V. Vovk, L. Carlsson, A. Gammerman, ExCAPE WP1. Probabilistic prediction., Tech. rep.
[27] S. Arvidsson, O. Spjuth, L. Carlsson, P. Toccaceli, Prediction of Metabolic Transformations using Cross Venn-ABERS Predictors, in: Proceedings of the Sixth Workshop on Conformal and Probabilistic Prediction and Applications, Vol. 60, 2017, pp. 118–131. <http://proceedings.mlr.press/v60/arvidsson17a.html>.
[28] I.S. van Maurik, M.D. Zwan, B.M. Tijms, F.H. Bouwman, C.E. Teunissen, P. Scheltens, M.P. Wattjes, F. Barkhof, J. Berkhof, W.M. van der Flier, Interpreting Biomarker Results in Individual Patients With Mild Cognitive Impairment in the Alzheimer's Biomarkers in Daily Practice (ABIDE) Project, JAMA Neurol. 74 (12) (2017) 1481–1491, https://doi.org/10.1001/jamaneurol.2017.2712 http://archneur.jamanetwork.com/article.aspx?doi=10.1001/jamaneurol.2017.2712.
[29] D. Devetyarov, I. Nouretdinov, B. Burford, S. Camuzeaux, A. Gentry-Maharaj, A. Tiss, C. Smith, Z. Luo, A. Chervonenkis, R. Hallett, V. Vovk, M. Waterfield, R. Cramer, J.F. Timms, J. Sinclair, U. Menon, I. Jacobs, A. Gammerman, Conformal predictors in early diagnostics of ovarian and breast cancers, Prog. Artif. Intell. 1 (3) (2012) 245–257, https://doi.org/10.1007/s13748-012-0021-y http://link.springer.com/10.1007/s13748-012-0021-y.
[30] H. Papadopoulos, A. Gammerman, V. Vovk, Reliable diagnosis of acute abdominal pain with conformal prediction, Eng. Intell. Syst. 17 (2–3) (2009) 127–137.
[31] A. Lambrou, H. Papadopoulos, E. Kyriacou, C.S. Pattichis, M.S. Pattichis, A. Gammerman, A. Nicolaides, Assessment of stroke risk based on morphological ultrasound image analysis with conformal prediction, Artif. Intell. Appl. Innovat. (2010) 146–153, https://doi.org/10.1007/978-3-642-16239-8.
[32] J. Maroco, D. Silva, A. Rodrigues, M. Guerreiro, I. Santana, A. De Mendonça, Data mining methods in the prediction of Dementia: a real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests, BMC Res Notes 4 (2011) 229.
[33] S.G. Mueller, M.W. Weiner, L.J. Thal, R.C. Petersen, C.R. Jack, W. Jagust, J.Q. Trojanowski, A.W. Toga, L. Beckett, Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI), Alzheimer's Dementia 1 (1) (2005) 55–66, https://doi.org/10.1016/j.jalz.2005.06.003.
[34] M. Ayer, H.D. Brunk, G.M. Ewing, W.T. Reid, E. Silverman, An Empirical Distribution Function for Sampling with Incomplete Information, Annals Math. Stat. 5 (1955) 641–647, https://doi.org/10.1214/aoms/1177728423 http://www.jstor.org/stable/2236377%5Cnhttp://www.jstor.org/page/.
[35] G. Shafer, V. Vovk, A tutorial on conformal prediction, J. Mach. Learn. Res. (2008) 371–421 arXiv:0706.3188 http://arxiv.org/abs/0706.3188.
[36] C. Saunders, A. Gammerman, V. Vovk, Transduction with confidence and credibility, IJCAI Int. Joint Conf. Artif. Intell. 2 (4) (1999) 722–726.
[37] D. Devetyarov, I. Nouretdinov, Prediction with confidence based on a random forest classifier, IFIP Advances in Information and Communication Technology 339 AICT (2010) 37–44. doi:10.1007/978-3-642-16239-8_8.
[38] P. Toccaceli, I. Nouretdinov, A. Gammerman, Conformal Predictors for Compound Activity Prediction, in: Conformal and Probabilistic Prediction with Applications, 2016, pp. 51–66. arXiv:1603.04506. http://arxiv.org/abs/1603.04506.
[39] F. Yang, H.-Z. Wang, H. Mi, C.-D. Lin, W.-W. Cai, Using random forest for reliable classification and cost-sensitive learning for medical diagnosis, BMC Bioinform. 10 (Suppl 1) (2009) S22, https://doi.org/10.1186/1471-2105-10-S1-S22.
[40] A. Forreryd, U. Norinder, T. Lindberg, M. Lindstedt, Predicting skin sensitizers with confidence — using conformal prediction to determine applicability domain of GARD, Toxicol. In Vitro 48 (August 2017) (2018) 179–187, https://doi.org/10.1016/j.tiv.2018.01.021.
[41] I. Nouretdinov, S. Costafreda-Gonzalez, A. Gammerman, A. Chervonenkis, V. Vovk, V. Vapnik, C. Fu, Machine learning classification with confidence: application of transductive conformal predictors to mri-based diagnostic and prognostic markers in depression, NeuroImage 56 (2) (2011) 809–813, https://doi.org/10.1016/j.neuroimage.2010.05.023.
[42] V. Balasubramanian, R. Gouripeddi, S. Panchanathan, J. Vermillion, A. Bhaskaran, R. Siegel, Support vector machine based conformal predictors for risk of complications following a coronary drug eluting stent procedure, in: 2009 36th Annual Computers in Cardiology Conference (CinC), 2009, pp. 5–8.
[43] V. Vovk, G. Shafer, I. Nouretdinov, Self-calibrating Probability Forecasting, in: S. Thrun, L.K. Saul, B. Scholkopf (Eds.), Advances in Neural Information Processing Systems 16, MIT Press, Cambridge, MA, 2004, pp. 1133–1140. http://papers.nips.cc/paper/2462-self-calibrating-probability-forecasting.pdf%5Cnfiles/3674/Vovk et al.-2004-Self-calibratingProbabilityForecasting.pdf%5Cnfiles/3675/2462-self-calibrating-probability-forecasting.html.
[44] T. Melluish, C. Saunders, I. Nouretdinov, V. Vovk, Comparing the bayes and typicalness frameworks, Proceedings of the 12th European Conference on Machine Learning, EMCL '01, Springer-Verlag, London, UK, UK, 2001, pp. 360–371 http://dl.acm.org/citation.cfm?id=645328.650016.
[45] V. Vovk, I. Nouretdinov, V. Fedorova, I. Petej, A. Gammerman, Criteria of efficiency for set-valued classification, Annals Math. Artif. Intell. 81 (1–2) (2017) 21–46, https://doi.org/10.1007/s10472-017-9540-3.
[46] T. Pereira, F.L. Ferreira, S. Cardoso, D. Silva, A. de Mendonça, M. Guerreiro, S.C. Madeira, f. t. A.D.N. Initiative, Neuropsychological predictors of conversion from mild cognitive impairment to Alzheimer's disease: a feature selection ensemble combining stability and predictability, BMC Med. Inform. Decis. Mak. 18 (1) (2018) 137, https://doi.org/10.1186/s12911-018-0710-y https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-018-0710-y.
[47] J. Demsar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.
[48] M. Guerreiro, Contributo da Neuropsicologia para o Estudo das Demências, Doctoral dissertation, Faculty of Medicine of Lisbon, 1998.
[49] American Psychiatric Association, DSM-IV-TR, fourth ed., APA, Washington DC, 2000.
[50] F. Portet, P. Ousset, P. Visser, G. Frisoni, F. Nobili, P. Scheltens, B. Vellas, J. Touchon, M.W.G. o. t. E.C. o. A.D. (EADC)., Mild cognitive impairment (MCI) in medical practice: a critical review of the concept and new diagnostic procedure. Report of the MCI Working Group of the European Consortium on Alzheimer's Disease., J Neurol Neurosurg Psychiatry 77 (6) (2006) 714–8.
[51] C. Guo, G. Pleiss, Y. Sun, K.Q. Weinberger, On calibration of modern neural networks, ArXiv abs/1706.04599.
[52] A.V. Carreiro, P.M.T. Amaral, S. Pinto, P. Tomás, M. de Carvalho, S.C. Madeira, Prognostic models based on patient snapshots and time windows: Predicting disease progression to assisted ventilation in Amyotrophic Lateral Sclerosis, J. Biomed. Inform. 58 (2015) 133–144, https://doi.org/10.1016/j.jbi.2015.09.021.