Applied Intelligence (2025) 55:319
https://doi.org/10.1007/s10489-024-06127-0

# A meta-heuristic approach to estimate and explain classifier uncertainty

**Andrew Houston**[1,2] · **Georgina Cosma**[1]

Accepted: 30 November 2024


**Abstract**
Trust is a crucial factor affecting the adoption of machine learning (ML) models. Qualitative studies have revealed that end-users, particularly in the medical domain, need models that can express their uncertainty in decision-making allowing users to know when to ignore the model's recommendations. However, existing approaches for quantifying decision-making uncertainty are not model-agnostic, or they rely on complex mathematical derivations that are not easily understood by laypersons or end-users, making them less useful for explaining the model's decision-making process. This work proposes a set of class-independent meta-heuristics that can characterise the complexity of an instance in terms of factors that are mutually relevant to both human and ML decision-making. The measures are integrated into a meta-learning framework that estimates the risk of misclassification. The proposed framework outperformed predicted probabilities and entropy-based methods of identifying instances at risk of being misclassified. Furthermore, the proposed approach resulted in uncertainty estimates that proves more independent of model accuracy and calibration than existing approaches. The proposed measures and framework demonstrate promise for improving model development for more complex instances and provides a new means of model abstention and explanation.

**Keywords** Uncertainty quantification · Meta-learning · Fuzzy clustering · Complexity theory · Explainable AI

## 1 Introduction

As the use of AI continues to grow in fields where incorrect decisions can have serious consequences. Despite their potential benefits, the adoption of such tools remains a challenge, with trust being cited as a primary barrier [1]. Complex 'black-box' algorithms that cannot explain when they may be correct are often cited as a source of distrust [1]. Therefore there is a requirement for tools which help to equip end-users with the means to facilitate an appropriate trust relationship with ML and AI tools, termed trust calibration. The design of methods to facilitate trust calibration requires consideration of the needs and perspectives of the end user. Qualitative research has shown, that in the medical sector, clinicians have emphasised the importance of models indicating uncertainty or abstaining when their confidence is low [2, 3]. Amann et al. [4] emphasise that the effectiveness of explanations depends on the end user's ability to comprehend their meaning, thus, the interpretability of the methods themselves is a crucial factor in facilitating trust calibration.

In decision-making, regardless of whether the decision maker is human or an AI, uncertainty is often classed as arising from one of two types, epistemic or aleatoric [5]. *Epistemic uncertainty* refers to uncertainty arising from limitations in our knowledge. In human decision-making, this could be the result of an individual having limited experience in a given situation, such as a doctor attempting to treat a patient with a rare and ill-understood disease. For an AI, epistemic uncertainty could originate from the insufficiency of similar training data [5], which can manifest in various forms, such as the absence of underrepresented groups in facial recognition datasets, resulting in a decline in recognition performance for these groups [6], or the occurrence of rare circumstances within a dataset. *Aleatoric uncertainty* refers to uncertainty arising from inherent variability or randomness in the system itself. In human decision-making,

✉ Andrew Houston
a.houston@lboro.ac.uk

[1] Department of Computer Science, Loughborough University, Epinal Way, Loughborough LE11 3TU, UK

[2] Academic Department of Military Rehabilitation, Defence Medical Rehabilitation Centre, Stanford Hall, Loughborough, UK

Published online: 14 January 2025

aleatoric uncertainty might arise from unpredictable external factors, such as environmental conditions affecting agricultural crop yields or the response of a patient to a certain medication. For AI systems, aleatoric uncertainty can stem from inherent stochasticity in the data or the process being modeled [7]. This could include factors such as sensor noise in data collected from physical sensors, variability in human behaviour captured by social media data, or randomness in financial markets. By accounting for these types of uncertainty, end-users and AI developers can better assess risks and implement strategies to mitigate the impact of uncertainty on decision outcomes.

Existing approaches for characterising instance complexity and quantifying decision-making uncertainty have limitations, including post-deployment applicability [8, 9] and failure to consider multiple sources of complexity [10]. The most easily applied variations of uncertainty estimation, such as entropy of the predicted probabilities, suffer from an inherent link to classifier calibration and accuracy that could result in poor-quality explanations. Additionally, existing approaches typically rely on complex mathematical derivations that are not easily understood by laypersons or end-users, making them less useful for explaining the model's decision-making process [11].

This paper presents a novel approach to address the problem of misclassification detection and explanation, and makes the following three key contributions:

- A suite of model-agnostic, interpretable meta-heuristics are proposed, which characterise instances in terms of the sources of complexity which are shared between human and AI decision-making. These meta-heuristics are designed to be interpretable, ensuring that the explanations resonate with end-users.
- A simple yet effective uncertainty estimation system is developed, employing a Bayesian-optimised, weighted fuzzy c-means algorithm. This system provides more reliable uncertainty assessments compared to entropy-based methods and shows performance comparable to the Trust score while being less dependent on model calibration and accuracy.
- A robust investigation into the optimal construction of a knowledge base using a combination of real and synthetic datasets. The investigation highlights a number of actionable insights which help to enhance the quality of the estimates made by the misclassification detection system, leading to better risk estimation and explanation of potential misclassifications.

The remainder of paper is structured as follows: Section 2 provides a descriptive overview of existing approaches to characterising the complexity of instances and methods for misclassification detection on tabular datasets; Section 3 describes the proposed meta-heuristics, outlines the design of the fuzzy clustering system for estimating misclassification risk, describes the design of a knowledge-base used to improve the quality of the estimates made by the misclassification detection system; Section 4 describes the datasets used in the experiments, the process used to produce synthetic data to improve the robustness of the evaluation and the methods used to train the models and apply the proposed methods; Section 5 presents the experimental evaluation of the suite of meta-heuristics, the findings from the investigation into the knowledge base construction and the comparison of the proposed methods against the existing methods discussed in the related works section; Lastly, Section 6 discusses potential applications of the proposed methods and provides a discussion of future directions.

## 2 Related works

This section provides an overview of the sources of complexity in classification problems and reviews existing methods for the quantification of classifier uncertainty and misclassification detection, with a focus on the explainability of each method.

### 2.1 Types of uncertainty

Uncertainty refers to the model's lack of confidence or reliability in its predictions. Uncertainty can be classed as one of two types, epistemic or aleatoric. Epistemic uncertainty can be described as a type of uncertainty originating from the insufficiency of similar training data [5]. Epistemic uncertainty can manifest in various forms, such as the absence of underrepresented groups in facial recognition datasets, resulting in a decline in recognition performance for these groups [6], or the occurrence of rare circumstances within a dataset. Aleatoric uncertainty reflects uncertainty arising from a degree of randomness which cannot be explained away, such as the roll of a dice, flip of a coin, noise in a signal or low resolution of an image [7].

In ML research, several factors have been identified for increasing the epistemic and aleatoric uncertainty of classification problems, such as class imbalance, class overlap and outliers. These are described below.

#### 2.1.1 Class imbalance

Class imbalance is defined as an unequal distribution of instances between classes in a dataset and is a common problem in many domains, spanning medical predictions [12], sentiment analysis [13] and information retrieval [14]. Large class imbalances can result in models that are highly accurate but lack sensitivity [15]. Common approaches for address-

ing class imbalance include re-sampling techniques, such as synthetic minority over-sampling (SMOTE) [16], that either increases the instances in the minority class or reduces the majority class, and cost-sensitive learning, which gives higher weight to errors made on specific classes [17]. However, class imbalance may have a limited impact on the classifier performance, depending on other factors such as the well-defined class boundaries [9, 18].

#### 2.1.2 Class overlap

Class overlap, defined as the overlap in the feature space between instances of multiple classes, is a well-recognised challenge in classification tasks [9, 18]. The complexity of an instance in a region of class overlap is higher compared to instances in regions dominated by a single class [9]. This overlap can introduce noise in a classification, increasing the aleatoric uncertainty of decisions made on such instances. While class overlap itself may not be inherently stochastic, instances lying within overlapping regions are at a higher risk of misclassification, reflecting an uncertainty akin to aleatoric uncertainty [9]. The overlap can cause the model to assign classes based on subtle feature variations that may not be robust, thus heightening the unpredictability of the classification outcome.

The relationship between class overlap and class imbalance is often discussed in the literature, with some studies suggesting that the impact of class overlap is greater than class imbalance, while the influence of class imbalance on the complexity of a classification task increases in problems with high overlap [9, 18, 19]. In the context of clinical decision-making, class overlap may present in two forms: a general overlap of all features, making it difficult to distinguish between the diagnoses or prognoses of patients, or in patients where different aspects of their presentation align closely with different class outcomes. Figure 1 demonstrates both forms of class overlap, where instance D represents a patient with features $x1$ and $x2$ falling within the global overlap of the blue and red classes, and instance B represents a patient where feature $x1$ aligns closely with the blue class and feature $x2$ aligns with the red class.

#### 2.1.3 Outliers

Outliers are instances that lie in a region of low neighbourhood density and can result in higher levels of epistemic uncertainty in predictions due to the absence of similar instances for comparison [5]. The impact of outliers on the complexity of a prediction is dependent on various factors, such as the location of the instance in the feature space and the classifier used. For instance, in the hypothetical dataset shown in Fig. 1, two instances may have similar levels of outlierness but their location in the feature space may result in differing levels of complexity. While instance A has high

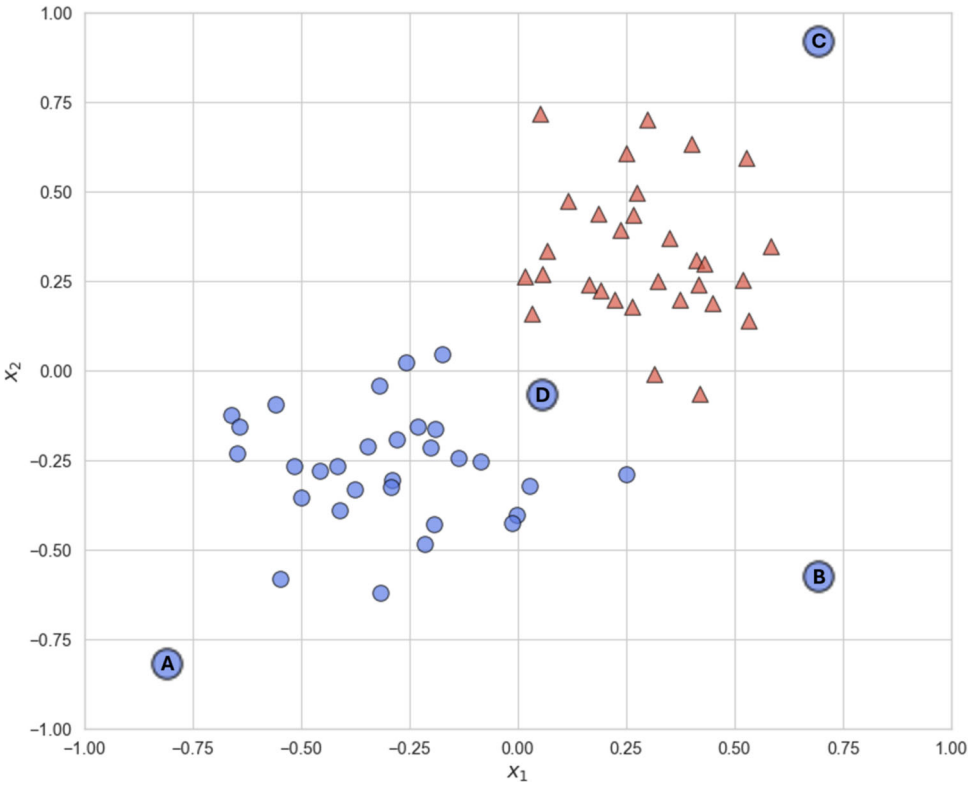



**Fig. 1** A hypothetical 2-feature dataset where points A and C are instances where a prediction could be considered to have a high degree of epistemic uncertainty, point B reflects an instance where a prediction could be considered to have both high epistemic and aleatoric uncertainty, and point D is an instance where a prediction could be considered to have a high degree of aleatoric uncertainty

levels of outlierness, its location in the feature space relative to other instances of the same class may result in highly accurate predictions. On the other hand, instance C, though having similar levels of outlierness, may not be predicted accurately as its location in the feature space aligns closely with the opposing class. The impact of outliers on decision-making complexity is therefore specific to the classifier and the dataset [20]. In clinical decision-making, the availability of past evidence and experience plays a critical role in supporting predictions. Qualitative work found that outlying, abstract patient presentations can increase uncertainty in decision-making [21]. Furthermore, due to the core underpinning of clinical decision-making being the availability of evidence and past experience [22, 23], in the case of rare and complex patient presentations, much like the presence of outliers within the context of ML, uncertainty in the decisions made increases [24].

## 2.2 Existing approaches to characterising complexity and uncertainty

### 2.2.1 Meta-heuristic approaches

In the context of this paper, a meta-heuristic is defined as an approximation method used to assess the complexity of instances without relying on prior knowledge of class labels or classifications. They provide a way to evaluate the difficulty of instances, in a specific complexity domain, based on their inherent characteristics, in terms of the training set.

Smith and Martinez [8] proposed a series of hardness measures to identify instances that are likely to be misclassified in typical classification problems and defined thresholds for their use as a means of data cleaning during model development. A subsequent paper further explored the application of the proposed measures and their relevance to the complexity of instance-level decision-making [9] finding class overlap to be the most significant factor in characterising complex instances. The paper presented several approaches for integrating the proposed hardness measures into the learning process, including the augmentation of error functions within multi-layer perceptrons to reduce the weight assigned to more complex instances and the removal of complex instances to reduce unwanted noise within the training set. The advantage of the measures proposed by Smith et al. [9] is that they allow for the accurate characterisation of the complexity of individual instances, enabling robust model development and improved reasoning as to why misclassifications occur, as demonstrated in Houston et al. [12]. Additionally, the measures are easy to calculate, and the sources of complexity they reflect are well-defined, making them easily understood by end-users. However, post-deployment, most of the proposed measures are limited in their utility because they rely on class knowledge, which is unavailable in prospective cases.

In 2022, Barandas et al. [25] proposed a method for characterising aleatoric and epistemic uncertainty in traditional classification problems and applied the methods to aid in abstaining from making a prediction. They applied three measures to characterise the domains of uncertainty. Entropy [26] was used to measure aleatoric uncertainty, variation ratios were applied to evaluate epistemic uncertainty resulting from the model, and a density estimation technique was applied to measure epistemic uncertainty resulting from a lack of data. The methods proved successful in utilising uncertainty measures to improve the performance of a series of classifiers using uncertainty-based rejection. The authors augmented the way two models can be compared by looking at both actual performance metrics, such as accuracy, and the uncertainty values associated with the predictions. Additionally, the interpretable nature of the proposed measurements shows promise in acting as a facilitator of trust to the end user.

One of the most successful methods for inferring whether to trust a prediction is the trust score, proposed by Jiang et al. [10]. The trust score is determined by evaluating the proximity of a test sample to the densest regions of both the predicted class and the nearest alternative class, providing insight into the model's confidence in its prediction. If the test sample is closer to the densest region of a class other than the predicted one, it suggests potential uncertainty or ambiguity in the classification. Conversely, if the test sample is closer to the densest region of the predicted class, it indicates higher confidence in the classification. The trust score method was proven to consistently outperform other existing methods in terms of precision versus percentile curve across various datasets and models, however, the magnitude of this improvement diminished as model accuracy improved.

### 2.2.2 Model-dependent approaches

A commonly used method to estimate misclassification risk is to use the prediction probability of the predicted class for the instance. Typically, prediction probability refers to the conditional probability of the predicted class, although it may differ depending on the model. For example, for support vector machines, Platt scaling is commonly used to calibrate the output probabilities to better reflect the true likelihood of a particular class [27]. Translation of the predicted probabilities into a confidence metric can be expressed simply, as:

$$\text{Confidence} = \left| \frac{1}{C} - \max_{c} P(x_c) \right| \tag{1}$$

where $C$ is the number of classes in the dataset and $P(x_c)$ denotes the predicted probability for class $c$. Investigations into the effectiveness of such simplistic measures found that predictions of a ML mode accompanied by confi-

dence scores can help end users adjust their trust levels, i.e. end users tend to trust a model more when its confidence level is elevated [28]. Findings were further substantiated by Rechkemmer and Yin [29] who found that higher predicted probabilities increased users' willingness to follow the model's predictions and improved their self-reported trust in the model. Whilst such a method is simple to estimate it lacks interpretability as it does not explain why a model is less confident. Furthermore, given the somewhat abstract nature of the probability estimates to non-technical users, the use of such methods for calibrating trust may be hindered by the end-user's math and logic skills [11], highlighting the need for more explainable ways to model to express their confidence.

While representing confidence/uncertainty according to (1) focuses on the difference between the maximum predicted probability and the uniform distribution where all classes have equal probability. Other approaches have applied entropy to the predicted probability distribution providing a measure of disorder in the distribution of probability [25, 30–32]. Entropy is typically applied as:

$$H(x) = -\sum_{c=1}^{C} P(x_c) \log P(x_c), \tag{2}$$

where $C$ is the number of classes in the dataset and $P(x_c)$ denotes the predicted probability for class $c$. Higher entropy values indicate higher uncertainty, while lower entropy values indicate more confidence in the prediction. Nicora et al. [30] applied entropy to the predicted posterior probabilities of test instances as a means of categorising them as "reliable" or "unreliable", proving such an approach to be an effective means of abstention. Results demonstrated by improvements in the performance of a classifier on "reliable" instances, with respect to accuracy and area under the receiver operating characteristic curve (AUROC), and found the entropy-based approach to outperform a resampling-based approach that estimates the amount that a prediction would change if the model had been fit on different training data [33]. Similarly, the application of entropy to the predicted probability of test instances was explored as a means of setting the abstention thresholds. Findings highlighted the calibration of a classifier being better among instances where uncertainty was low, raising an important limitation of the use of predicted probabilities for inferring misclassification risk. Poor calibration can lead to inaccurate uncertainty estimates using predicted probabilities, potentially underestimating or overestimating the uncertainty associated with model predictions [34].

The use of both conditional probabilities and entropy are common sampling methods within active learning. Active learning is an approach designed to reduce the computational cost and time required to train a learning algorithm by identifying the most useful instances to train a classifier. A popular method of active learning is uncertainty sampling, introduced by Lewis and Gale [35], which identifies instances that a model is most uncertain about, learning the representation of such challenging cases to create a decision boundary. However, while uncertainty sampling is effective in identifying instances that a model is most uncertain about, like prediction probabilities, it does not identify the reasoning behind the uncertainty.

Sharma and Bilgic [36] proposed an evidence-based framework for explaining uncertainty, targeting two specific sources of uncertainty which they termed 'conflicting evidence uncertainty' and 'insufficient evidence uncertainty'. 'Conflicting evidence uncertainty' refers to the presence of strong evidence for an instance belonging to more than one class, whereas 'insufficient evidence uncertainty' refers to a lack of evidence for an instance belonging to any class. Their evaluations on real-world datasets found that 'conflicting evidence uncertainty' appeared to be a more effective means of active learning, outperforming traditional active learning and 'insufficient evidence uncertainty'. The benefit of using such measures, unlike the hardness measures proposed by Smith et al. [9], is their lack of reliance on class knowledge. However, a downside to the approach of Sharma and Bilgic [36] is the requirement to modify the evidence-based framework for each classifier, given the differences in the ways classifiers compute their predictions. Additionally, active learning approaches typically fail to capture concepts such as noise and outlierness [37].

The primary concern with model-dependent approaches is that their performance is tied to specific classifier properties such as calibration and accuracy, as demonstrated in Section 5 of this paper. Where this becomes an issue is in cases where a classifier might be overly optimistic about their own accuracy, especially if they are complex models prone to overfitting. As a result, the assessment of misclassification risk by methods which are dependent on model outputs may provide unrealistic estimates. The same argument could be made with respect to classifier calibrations, wherein a classifier might predict the right class labels but could be poorly calibrated, meaning the probabilities it outputs do not match the actual likelihoods [38]. With methods that make use of predicted probabilities, the assessment of misclassification risk could be biased by the poorly calibrated probabilities. Beyond model performance, there are limitations associated with methods which work solely for a given classifier. This lack of generalisability reduces the methods' practical utility, especially in cases where multiple models are used or where models need to be frequently updated or changed [39]. Furthermore, the inflexibility of these model-dependent measures is a major drawback in meta-learning contexts, where

there is a requirement to apply meta-information across various tasks and models to understand the learning processes.

### 2.3 Three-way decisions

Three-way decision (TWD) is a framework developed to manage uncertainty in machine learning, particularly in classification tasks. Proposed by Yao et al. [40] as an alternative to binary classification, TWD introduces a third category for uncertain cases, in addition to the typical positive and negative categories. By assigning ambiguous instances to a boundary set, TWD can provide a clearer representation of uncertainty in data.

Within the TWD framework, both Decision-Theoretic Rough Sets (DTRS) [41] and Pawlak Rough Sets [42] offer structured ways to implement this tripartition, dividing data into the three categories. Pawlak Rough Sets originally established the concept by defining a boundary region for instances that cannot be clearly classified due to overlapping attributes [42]. DTRS extends this approach by including probabilistic thresholds based on decision costs and benefits, dynamically adjusting the boundary region to reflect nuanced uncertainty management [41]. This cost-sensitive approach allows models to balance the trade-offs in classifying uncertain cases, which can enhance the framework's effectiveness in scenarios where ambiguity is common and errors carry greater consequences.

In an applied setting TWD frameworks offer a means of managing uncertainty in classification tasks, especially in fields such as healthcare, where the consequences of overconfident predictions can be severe. Abdar et al. [43] implemented a TWD-based approach within Bayesian deep learning models to classify skin cancer images more reliably. Their approach used a multi-phase classification process with uncertainty quantification techniques like Monte Carlo dropout and deep ensembles, enhancing the model's ability to identify uncertain cases. Similarly, Chen et al. [44] applied a TWD-supported decision system in the diagnosis of focal liver lesions, tri-partitioning cases into certain benign, certain malignant, and uncertain categories. The tri-partitioning allowed the system to handle difficult cases carefully offering a means of balancing diagnostic accuracy and risk. In concurrence, Campagner et al. [45] highlighted the use of TWD for handling ambiguous data, enabling models to abstain from uncertain classifications when necessary. By supporting these more cautious classifications, TWD frameworks effectively reduce errors in complex applications and facilitate safer, more reliable decision-making.

A core limitation of TWD frameworks is their bucketing of predictions into "certain" and "uncertain" categories, which can result in an oversimplification of uncertainty [46]. For instance, instances that lie close to decision boundaries might represent mild uncertainty, while others may show high degrees of ambiguity due to conflicting evidence or sparse data. More sophisticated methods, like those that quantify degrees of uncertainty, can offer a richer, more nuanced picture, supporting more tailored responses. Thus, while TWD provides a structured way to handle uncertainty, it sacrifices nuance, limiting its effectiveness in highly complex or risk-sensitive applications where understanding distinctions in uncertainty levels is useful.

## 3 Methodology

### 3.1 Characterising class diversity

Several methods have been proposed to quantify class imbalance in binary and multi-class problems measure like the imbalance-ratio commonly used [47]. However, such measures can be difficult to examine in highly multi-class problems and are not single-value representations. To overcome such limitations, we modify the multi-class imbalanced degree, proposed by Zhu et al. [48]. We first calculate the likelihood-ratio imbalance degree (LRID):

$$\mathrm{LRID}(y, C) = \sum_{\substack{i \in c \\ i > 0}} i \cdot \log\left(\frac{m}{C \cdot i}\right), \tag{3}$$

where $y$ is a vector containing the class labels of a set of instances, $c$ is a set containing the number of instances in the sample for each unique class, $C$ represents the total number of classes in the dataset and, $m$ is the total number of instances in the sample. To obtain the diversity of the sample, the LRID is normalised to the LRID of a sample, equal in size and number of classes, with minimal diversity:

$$\mathrm{Diversity} = 1 - \frac{\mathrm{LRID}(y, C)}{\mathrm{LRID}(y_{\mathrm{null}}, C)}, \tag{4}$$

where $y_{\mathrm{null}}$ is a set of class labels, equal in length to $y$, where the number of instances in one class is equal to $m - (C - 1)$, with a single instance existing in the remaining classes. Therefore, a normalised diversity score closer to zero reflects less diversity being present within our sample, and a diversity closer to one reflects more a equal balance of classes.

### 3.2 Characterising instance complexity

This section presents the meta-heuristics used to characterise the complexity of a given instance. Seven meta-heuristics are proposed each reflecting a unique component of instance complexity, presented in Table 1.

**Table 1** Descriptions of each meta-heuristic and lay-person explanations which illustrate how each heuristic could be used to explain a different source of complexity within a clinical context

| Meta-heuristic | Description | Lay-person Explanation |
|---|---|---|
| $k$-Diverse Neighbours | Reflects the localised overlap of classes among similar instances based on all features. | *Patients with a similar presentation to the current patient don't all have the same clinical outcome.* |
| Disjunct size | Reflects the complexity of the decision boundary. | *The decision process is complex and requires consideration of multiple factors for a diagnosis to be made.* |
| Disjunct class diversity | Reflects the localised overlap of classes among similar instances based on a subset of features. | *Patients with some of the similar symptoms have different underlying conditions.* |
| Hyperplane Distance | Indicates the closeness of an instance to the decision boundary. | *The patient's presentation means their prognosis could go either way.* |
| Outlierness | Indicates the rarity of the instance with respect to the training data. | *The patient's presentation is rare and not one for which much data exists.* |
| Class-Level Outlierness | Indicates whether the instance is more rare for one class than another. | *The patient's presentation does not provide an indication towards any single diagnosis.* |
| Evidence Conflict | Reflects whether the features of the instance give an indication to more than one class label. | *The patient's BMI and blood pressure are high which would indicate a poor surgical outcome, but they are also young, which often results in good outcomes.* |

#### 3.2.1 *k*-Diverse Neighbours

$k$-Diverse Neighbours ($k$-DN) reflects the local overlap of an instance within the task space, relative to its nearest neighbours from the training set and is calculated:

$$k\text{-DN}(x) = \text{Diversity}([c_1, c_2, \ldots, c_k], C), \tag{5}$$

where $[c_1, c_2, \ldots, c_k]$ are the class labels of nearest $k$ instances, derived using a nearest neighbours algorithm fitted to the training set and $C$ represents the total number of classes in the dataset.

#### 3.2.2 Disjunct size

Disjunct size is a class-independent measure proposed by Smith et al. [9] and calculates the size of the disjunct an instance is classified into by an unpruned decision tree, formed from a training set. The disjunct size of the returned instance is then normalised by dividing it by the size of the largest disjunct within the dataset:

$$\text{DS}(x) = \frac{|distjunct(x)|}{\max_{d \in D} |disjunct(d)|}, \tag{6}$$

where $|distjunct(x)|$ returns size of the disjunct for instance $x$, $D$ set of all disjuncts in the tree formed from the training set $X$, and the denominator $\max_{d \in D} |disjunct(d)|$ returns size of the largest disjunct in the tree formed from the training set $X$.

#### 3.2.3 Disjunct class diversity

Disjunct class diversity (DCD) reflects the diversity of the class of instances within the disjunct which an instance is classified into, ie those that are similar based on a subset of their features. Contrary to the methods of DS, the decision tree applied when calculating DCD is pruned:

$$\text{DCD}(x) = \text{Diversity}(\{c \mid c \in \text{disjunct}(x)\}, C), \tag{7}$$

where $\{c \mid c \in \text{disjunct}(x)\}$ is the class labels of the instances contained in the same disjunct as instance $x$, derived using the decision tree fitted to the training set and $C$ represents the total number of classes in the dataset.

#### 3.2.4 Hyperplane distance

Hyperplane distance (HD) reflects the distance of an instance from a fitted linear decision boundary within the feature space. When more than two classes are present, a one-vs-one approach is taken, calculating the distance from the hyperplane fitted to each pairwise class combination. This distance is calculated as follows:

$$\text{HD}(x) = \min_h \left( \frac{\text{decision\_function}(x)}{\|\text{coefficients}_h\|} \right), \tag{8}$$

where decision function value of instance, $x$, is normalised by dividing it by the L2 norm of the coefficients of each SVM classifier. Here, $h$ refers to a single distance from the decision boundary. The decision function values are determined using:

$$\text{decision\_function}(x) = w^T x + b, \tag{9}$$

where $w$ is the weight vector associated with the hyperplane $h$, $T$ represents the transpose operation and $b$ is the intercept. After calculating the distance from the decision boundary, the distance is normalised using a min-max scaler fitted to the training data.

#### 3.2.5 Outlierness

Outlierness (OL), reflects the degree to which an instance is similar to the instances contained within the training set. The outlierness of an instance is calculated using the density-based approach proposed by Tang and He [49], where the metric captures the ratio of the average neighbourhood density, to the density of instance $x$:

$$OL(x) = \frac{\sum_{x' \in S(x)} p(x')}{|S(x)|p(x)}, \tag{10}$$

where $S(x)$ represents the set of instances in the neighborhood of $x$, consisting of its $k$-nearest neighbors, reverse nearest neighbors, and shared nearest neighbors. Here, $x'$ denotes an arbitrary instance within this neighborhood $S(x)$, while $x$ is the specific instance for which outlierness is being calculated. The function $p(x)$ returns the density at the location of instance $x$, where density is calculated using:

$$p(x) = \frac{1}{|S(x)|+1} \sum_{X' \in S(x) \cup \{x\}} \frac{1}{h^N} k\left(\frac{x - X'}{h}\right), \tag{11}$$

where $N$ is the number of features in the dataset, $X'$ is the neighbourhood of $x$ and $|S(x)|$ is the size of the neighborhood of $x$. Here, $k\left(\frac{x-X}{h}\right)$ represents a Gaussian kernel function applied to the normalised distance between $x$ and each neighbor in the neighborhood, scaled by the kernel width $h$. The Gaussian kernel $k$ is defined as follows:

$$k\left(\frac{x - X'}{h}\right) = \frac{1}{(2\pi)^{\frac{N}{2}}} \exp\left(-\frac{\|x - X'\|^2}{2}\right), \tag{12}$$

where $\|x - X\|^2$ is the squared L2 norm between $x$ and each instance in the $x$'s neighbourhood $X'$. Following calculation of the outlierness for a test instance, the distance is normalised using a min-max scaler fitted to the training data.

#### 3.2.6 Class-level outlierness

Class-Level Outlierness (CL-OL), reflects the disparity in the level of outlierness calculated for an instance, across each class within the dataset. Given an instance, an outlierness score is calculated for each class in the dataset. Afterward, the entropy of the outlierness scores is calculated using (2).

#### 3.2.7 Evidence conflict

Evidence conflict (EC) reflects degree to which an instance's features fall within the distribution of multiple classes. To calculate EC, we first determine the volume of evidence for each class $c$. The volume of evidence (EV) for a class is computed as follows:

$$EV = \prod_{i \in N} \text{evidence}(x_i, c, i), \tag{13}$$

where $N$ is the total number of features. The function $\text{evidence}(x_i, c, i)$, quantifies the evidence supporting class $c$ across all features and is calculated as:

$$\text{evidence}(x_i, c, i) = \begin{cases} (1 - |(0.5 - \text{ECDF}(x_i; X_{ic}))|) \times w & \text{if feature } i \text{ is continuous} \\ \frac{|X_{ic} == x_i|}{|X_c|} \times w & \text{otherwise} \end{cases}, \tag{14}$$

where the function $\text{ECDF}(x_i; X_c)$ is the empirical cumulative distribution function for the feature value $x_i$, based on the class-level training data, $X_{ic}$, and $w$ is the feature importance for feature $i$, calculated using the mutual information score, applied to the training set:

$$\text{importance}(X, y) = \sum_{i=1}^{N} I(X_i; y), \tag{15}$$

where $I(X_i; y)$ is the measure of dependence between feature $i$ and the class labels, $y$, and $N$ is the total number of features in the dataset. Lastly, the entropy of the evidence volumes across all classes is calculated to give the evidence conflict score, using (2).

### 3.3 Proposed uncertainty estimation system

The proposed framework for estimating uncertainty is shown in Fig. 2, with each component explained throughout this section. An open-source repository containing the proposed methods, along with a simple example demonstrating their application on a clinical dataset, can be found at: https://github.com/andrewhouston113/Estimating-and-Explaining-Classifier-Uncertainty

#### 3.3.1 Knowledge-base formation

Meta-learning is the field of study focused on algorithms and techniques that enable models to learn and adapt across various tasks or datasets [50]. In the context of uncertainty estimation, meta-learning involves processes to gauge when

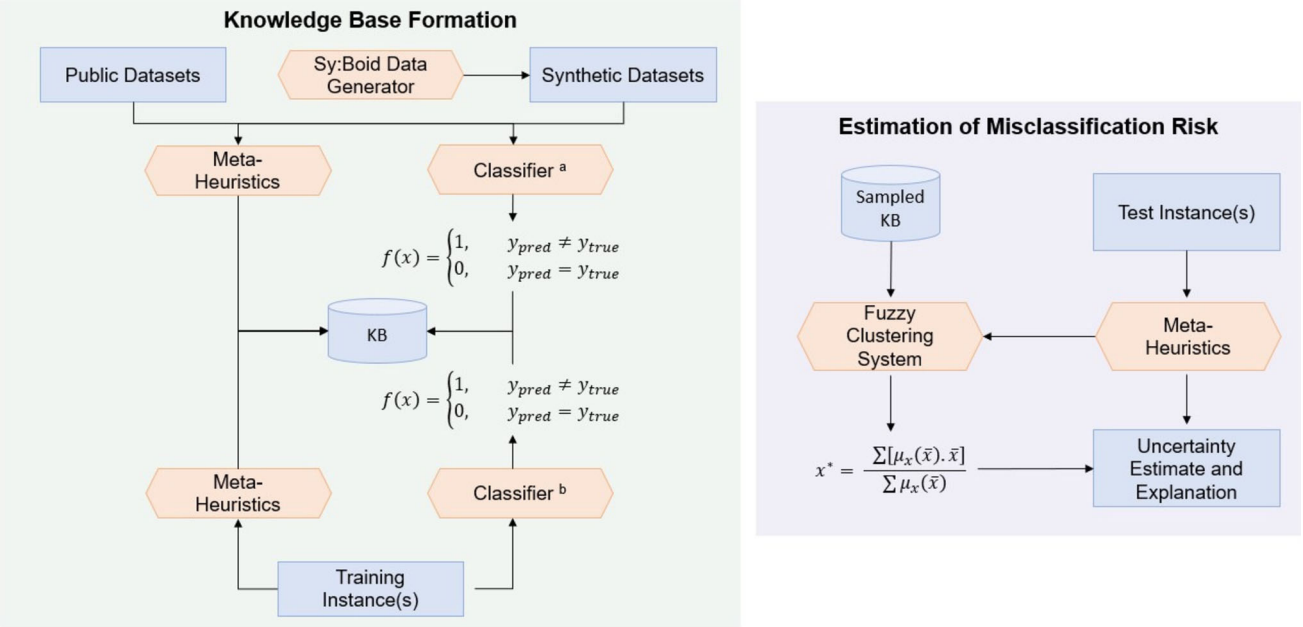



**Fig. 2** A framework for estimating the uncertainty of a prediction made on instances for a given classifier, employing a knowledge-base and fuzzy clustering system. [a] The classifier is the same as the classifier for which uncertainty is being assessed on test instances. [b] The classifier for which misclassification risk is being assessed on test instances

a model is more prone to errors based on its historical performance on analogous instances. To integrate meta-learning into the uncertainty estimation system, we first need to create a knowledge base. A knowledge base refers to a repository or collection of information gathered from previous tasks or problems. To construct a knowledge base, first, using 5-fold stratified cross-validation the meta-heuristics and classification model are fitted to the training fold of a series of public datasets and synthetic datasets, generated using the process described in Section 4.2, and the training instances of the current classification tasks. Thereafter, meta-heuristics are calculated and predictions made for each instance in the training fold. Misclassifications are identified using the equation:

$$f(x) = \begin{cases} 1 & y_{pred} \neq y_{true} \\ 0 & y_{pred} = y_{true} \end{cases}, \tag{16}$$

where $y_{pred}$ is the predicted class of an instance, and $y_{true}$ is the actual class of an instance. Misclassification events are stored in the knowledge base along with their corresponding meta-heuristics. The resulting knowledge base contains an $M \times 7$ matrix where $M$ is the number of instances, with a column containing the values of each of the seven meta-heuristics described in Subsection 3.2, and a single vector, of length $M$ containing a binary label reflecting whether the instance was correctly or incorrectly classified.

A static or dynamic approach to constructing the knowledge base can be taken. Should a static approach be taken, the knowledge base remains the same for all instances on which uncertainty estimations are to be made. Should a dynamic approach be taken, the knowledge base is altered to generate a more optimal knowledge base for the instances on which uncertainty estimations are to be made (the process for select said instances is described in Algorithm 2).

### 3.3.2 Estimation of misclassification risk

To prospectively estimate the misclassification risk of an instance, a weighted fuzzy c-means clustering algorithm is used. The rationale for selecting fuzzy clustering lies in its ability to capture the inherent uncertainty in data, acknowledging that classification complexity varies by instance rather than fitting into a binary easy or hard category. Each cluster represents a different presentation of complexity in terms of the meta-heuristics presented in Subsection 3.2, with each cluster assigned a value reflecting the number of instances in each cluster that were misclassified. Fuzziness quantifies ambiguity, membership degrees indicating the degree of alignment of an instance with all clusters. Such membership values provide clear insights into the strength and distribution of associations with different complexity profile, facilitating informed decision-making, while circumventing the need for complex algorithms which could further obscure the transparency of machine learning model's predictions.

To train the clustering algorithm, the algorithm is fit to the generated knowledge base. The clustering algorithm works by iteratively updating cluster centres and membership val-

ues until convergence. To achieve this the following steps are performed:

1. *Initialisation:* The membership matrix is randomly initialised, ensuring that the sum of membership values for each data point across all clusters is 1.
2. *Cluster Centre Calculation:* The cluster centres are computed based on the current membership values, by taking a weighted average of the data points, where the weights are the membership values raised to the power $m$:

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m x_i}{\sum_{i=1}^{N} u_{ij}^m} \tag{17}$$

   where $c_j$ is the centre of cluster $j$, $u_{ij}$ is the membership value of instance $x_i$ in cluster $j$, and $m$ is the fuzziness parameter.
3. *Membership Update:* Membership values are updated based on the distances of each data point from the cluster centres. The membership value $u_{ij}$ is updated as:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{d_{ij}}{d_{ik}}\right)^{2/(m-1)}} \tag{18}$$

   where $d_{ij}$ is the distance between data point $x_i$ and cluster centre $c_j$, and $C$ is the number of clusters.
4. *Repeat:* Steps 2 and 3 are repeated until the maximum number of iterations is reached.

Following fitting, of the clustering system, the misclassification rate is calculated for each cluster using all instances contained in the respective cluster. Misclassification rates were calculated using:

$$Misclassification\ Rate = 1 - \frac{TP + TN}{TP + FP + TN + FN}, \tag{19}$$

where $TP$ refers to true positive prediction, $FN$ refers to a false negative prediction, $FP$ refers to false positive prediction and $TN$ refers to a true negative prediction.

When a new instance, $x$ is input into the system their fuzzy membership is calculated, returning the membership of that instance to each cluster. To obtain a single value uncertainty estimate, $x^*$, reflecting the risk of misclassification of the classifier applied to $x$, the output is defuzzified using the weighted average method:

$$x^* = \frac{\sum[\mu_x(\bar{x}) \times \bar{x}]}{\sum \mu_x(\bar{x})}, \tag{20}$$

where $\mu_x$ is the multiplication of each weighting function of instance $x$ and $\bar{x}$ is the misclassification rate associated with each membership value.

To tune the clustering system, nested-cross-validation approach is applied, embedding Bayesian optimisation. Bay esian optimisation is a probabilistic model-based approach for optimising objective functions that can be expensive to evaluate [51]. In the proposed framework, Bayesian optimisation is applied within the inner-loop to determine the optimal weightings of the seven heuristics and the number of clusters, in order to maximise the area under the precision-recall curve (AUPRC), where the input is the misclassification risk scores and the target is the binary misclassification event. The outer-loop is used to evaluate the quality of the misclassification risk scores for identifying misclassification.

## 4 Experimental methodology

### 4.1 Datasets

Twenty-seven publicly available datasets were used in experiments to evaluate the relationships between the proposed methods and misclassification events. A description of the datasets and their source is provided in the Supplementary file. The datasets vary in their number of instances and dimensionality (i.e. the number of features). The selected datasets include continuous, ordinal and categorical features, to ensure the proposed meta-heuristics and developed system for estimating uncertainty are evaluated across a range of datasets.

### 4.2 Synthetic data generation

However, given that datasets from public repositories have drawn criticism for not being representative of the classification problems that may exist in the real world [52] and that their homogeneity, from a complexity standpoint, can result in a biasing of algorithm development, the framework embeds a synthetic data generator, termed Sy:Boid [53]. The process for generation is described fully in Houston et al. [53], but briefly, the synthetic data generator takes inputs relating to the number of instances, number of features, number of classes, and the desired complexity, in this paper, the F1 and N1 measures of Ho and Basu [54] are used. A series of random points are generated and labelled, and then, using a modified Boid algorithm points are moved about the feature space and the complexity of the dataset is measured. The algorithm is embedded within a genetic algorithm which optimises the rule's weightings to generate a dataset that closest meets the desired complexity. To generate synthetic datasets to complement the real data within the uncertainty system, Sy:Boid was tasked with generating datasets for ran-

dom combinations of F1 and N1, where both F1 and N1 are values between 0 and 1, N1 is not larger than F1 as this rarely occurs in real data [53]. The dataset specifications were randomly determined with the number of instances ranging between 200 and 2000, the number of features between 5 and 50, and the number of classes between 2 and 10. 100 synthetic datasets were generated.

### 4.3 Model development and application of the proposed methods

The method for model development and the application of the proposed methods is outlined in Algorithm 1. First, a dictionary of models and hyperparameters is defined. This study explored the relationship between misclassification and uncertainty estimates across 10 commonly used models, namely: Logistic regression, Support Vector Machines, $k$-Nearest Neighbours, Random Forest, AdaBoost, Naive Bayes, Linear and Quadratic Discriminant Analysis, a multilayer perceptron, and XGBoost. The hyper-parameters tuned are presented in the Supplementary file. To store the outputs of the model for evaluation an empty DataFrame is initialised (Algorithm 1, line 3). Each model in the model dictionary undergoes a 5-fold stratified cross-validation. Within each fold, the model's hyperparameters are tuned using RandomizedSearchCV on the training data (Algorithm 1, line 8). Then, an uncertainty estimator is initialised with the tuned model and fitted to the training data and the sample knowledge base (Algorithm 1, line 10). To ensure the fitting of the uncertainty estimator does not take a long time, a method for constructing a knowledge base which is efficient to train on is discussed in Section 5.2. Meta-heuristic scores are calculated for each instance in the test set using the methods described in Section 3.2. Thereafter, predictions and uncertainty estimates are made on the test set, and misclassifications are identified (Algorithm 1, lines 12 - 13). Finally, all information is stored in preparation for analysis (Algorithm 1, line 14).

**Algorithm 1** Training and application of meta-heuristics and uncertainty estimation.

```
1  Inputs: Dataset X, Class labels y, Model dictionary model_dict
2  Outputs: Results data frame results
3  Initialise an empty DataFrame, results, to store evaluation metrics
4  for v in model_dict do
5      Divide X and y into 5 stratified folds
6      for fold k_i in the 5 folds do
7          Set fold k_i as the test set
8          Tune model hyperparameters using RandomizedSearchCV on the remaining 4 folds
9          Initialise an uncertainty estimator, with the tuned model as the classifier.
10         Fit the uncertainty estimator to training data and sampled knowledge base
11         Calculate the meta-heuristic scores for each instance in the test set
12         Make predictions and uncertainty estimates on the test set
13         Identify misclassifications on the test set
14         Append misclassifications, meta-heuristics and uncertainty estimates to results
15     end
16 end
17 return results
```

## 5 Experiments

### 5.1 Evaluation of the proposed meta-heuristics

#### Experimental methods

The intention of each meta-heuristic is to characterise the complexity of an instance, with the sources of complexity being independent of each other. Therefore, to assess the independence of each meta-heuristic, a correlation analysis was performed using Spearman's Rank Correlation. To evaluate the direction and magnitude of the association between each meta-heuristic and misclassification occurrences, separate univariate binary logistic regression analyses were conducted. In these analyses, misclassification events served as the dependent variable, while each meta-heuristic was treated as the independent variable. Before conducting the logistic regression, standardisation through Z-score transformation was employed for each meta-heuristic to aid the interpretation of odds ratios (OR). Furthermore, the discriminative ability of each meta-heuristic in detecting misclassification events was assessed using the AUROC curve. Recognising that misclassification events are often rare compared to correct classifications, the AUPRC was also computed. This metric offers a more nuanced evaluation of the meta-heuristics effectiveness in identifying misclassification events, particularly in scenarios where the overall model accuracy is high.

#### 5.1.1 Experimental results

Results of the correlation analysis are presented in Fig. 3. R esults showed minimal co-linearity between meta-heuristics, meaning the source of uncertainty is that each meta-heuristic measure is independent of one another, therefore reflecting unique components of complexity.

Figure 4 A shows the results of the logistic regression analysis. When aggregated across all models, each meta-heuristic demonstrates a significant association with misclassification events. An increase in the KDN, DCD, OL, CL-OL and

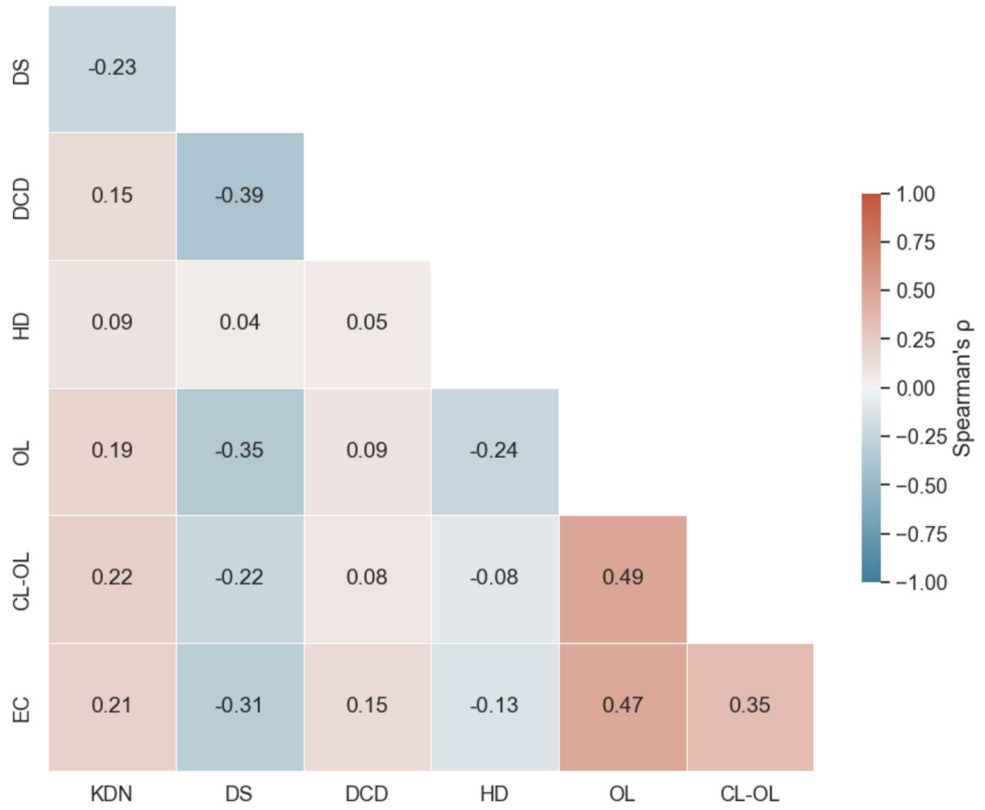



**Fig. 3** Spearman's rank correlation matrix showing the relationship between each meta-heuristic. A negative value, highlighted in blue, represented a negative correlation and a positive value, highlighted in red, represents a positive correlation

EC scores of an instance results in a significant increase in the likelihood of a misclassification event occurring ($p < 0.001$). Contextually, this means that instances with more class diversity among similar instances, which are outliers when compared to the training set, are equally outlying to each class and have feature values which align with multiple classes and have a greater risk of being misclassified. In contrast, an increase in the DS and HD scores of an instance results in a significant decrease in the likelihood of a misclassification event occurring ($p < 0.001$). Contextually, this means that instances with less complex decision boundaries and those which fall further away from the decision hyperplane are less at risk of being misclassified.

Figure 4B and C show the discriminatory capability of each meta-heuristic. KDN, OL and EC proved to be the most robust discriminators, with AUROC values of 0.64, 0.66 and 0.63 respectively, and AUPRC values of 0.45, 0.50 and 0.43, respectively. HD appears to offer limited discriminatory ability, with AUROC and AUPRC values of only 0.49 and 0.34, respectively. This poor performance is likely due to most classification problems, particularly those with a large number of features, being unlikely to be separable with a linear hyperplane without additional data transformation.

## 5.2 Determining the optimal knowledge base

### 5.2.1 Experimental methods

To construct an optimal knowledge base, several factors must be considered: 1) the number of instances to include, 2) the size of the knowledge base relative to the size of the training set, 3) the composition of the dataset, whether formed of exclusively real or synthetic instances, or a combination of both, 4) the density of the knowledge base about a given instance, and 5) the ratio of correct and incorrectly classified instances in the knowledge base, and the difference between these ratios. Whilst it is hypothesised that larger more densely populated knowledge bases will perform better given the greater number of examples available, in scenarios involving a smaller dataset, an excessively large knowledge base might dilute the training set and weaken the effectiveness of the uncertainty estimator in addressing the current classification problem.

To better understand the influence of the knowledge base specification on estimation performance, the following approach was adopted: For each test set, knowledge bases comprised $m$ instances, randomly selected from the remain-

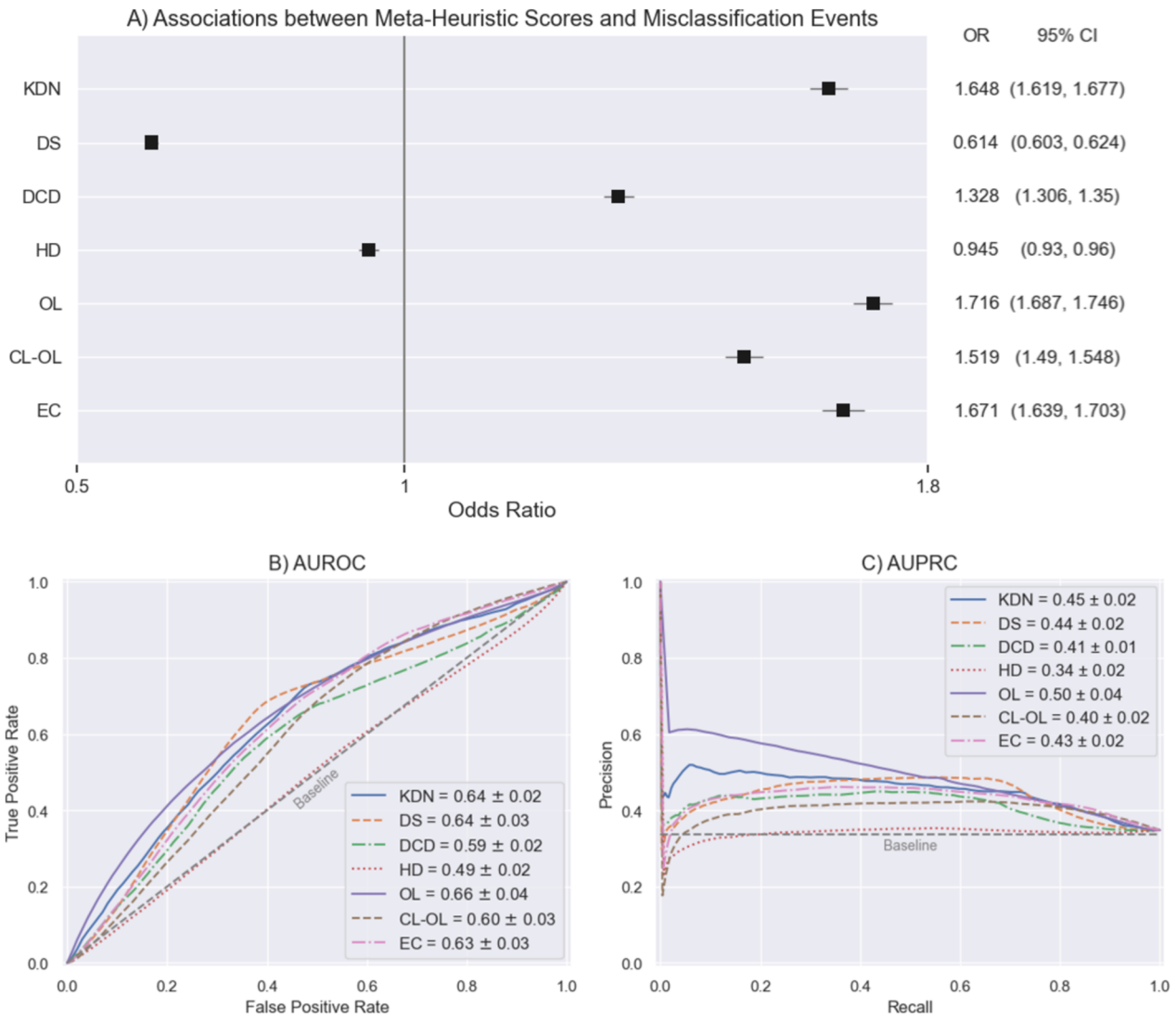



**Fig. 4** Aggregated results evaluating the relationship with misclassification events and discriminatory capability of each meta-heuristic. A) Forest plot displaying aggregated odds ratios and confidence intervals; B) Aggregated Receiver Operating Characteristic (ROC) curves; C) Aggregated Precision-Recall curves

ing datasets, were appended to the respective training set, where $m$ is an integer ranging from zero to 10,000. Thereafter, an uncertainty estimator was fitted to the training set and knowledge base, and an uncertainty estimate was determined for each test instance. The uncertainty estimate and whether the instance was misclassified was stored, along with the following meta-data:

- size of the knowledge base
- % of the knowledge base comprised of synthetically generated instances
- density of each the knowledge base about each instance in the test set, calculated using (11)
- ratio of correctly and incorrectly classified instances in the knowledge base
- ratio of correctly and incorrectly classified instances in the training set
- difference between the number of correctly and incorrectly classified instances in the knowledge base compared to the training set

#### 5.2.2 Experimental results

To understand how each factor might influence the quality of the estimations made by the uncertainty estimation system, logistic regression models were applied, with the classification outcome as the dependent variable and the inputs being the uncertainty estimate, the co-variate and the interaction term between the co-variate and the uncertainty estimate. Results indicated that larger knowledge bases

resulted in marginally poorer quality uncertainty estimates (OR = 0.983 [0.971, 0.995]), with a denser knowledge base about each instance in the test set improving the quality of the uncertainty estimates (OR = 1.319 [1.302, 1.335]). A more balanced knowledge base, in terms of the number of correctly and incorrectly classified instances, resulted in marginally poorer quality uncertainty estimates (OR = 0.983 [0.971, 0.995]), with the comparability of the imbalance of the knowledge base compared to the training set being of more importance with larger differences resulting in much poorer quality uncertainty estimates (OR = 0.842 [0.832, 0.853]). Knowledge bases comprised of more synthetic instances resulted in small improvements in the quality of the uncertainty estimates (OR = 1.050 [1.037, 1.064]). The ratio of the knowledge base size to the training set size did not affect the relationship between uncertainty estimates and misclassification events (OR = 1.008 [0.995, 1.022]).

Considering that multiple factors were found to influence the performance of the uncertainty system, a multivariate model was employed to identify which factors should be most focused on when constructing the knowledge base (Table 2). Results found analysis revealed significant interaction effects between uncertainty and all factors on misclassification. Notably, density exhibited the greatest effect on the quality of the uncertainty estimates (OR = 1.300 [1.283, 1.318]). Conversely, the difference in the imbalance of the knowledge base and training set displayed the strongest negative influence on the quality of the uncertainty estimates (OR = 0.860 [0.849, 0.871]), emphasising the need to ensure cohesion between the percentage of instances which are misclassified in the knowledge base compared to the training set.

#### 5.2.3 Implications for knowledge base design

In light of the aforementioned findings related to the specification of the knowledge base, Algorithm 2 is proposed. Findings suggested that the density of the knowledge base about each test set instance was most influential on the quality of the uncertainty estimates. Therefore, to ensure the meta-feature space in the knowledge base was sufficiently populated, a nearest neighbours algorithm was applied to retrieve the most similar instances within the knowledge base (Algorithm 2, Lines 6 - 10). Given that larger knowledge bases than the training set detrimentally affect the quality of uncertainty estimates, the number of neighbours sampled for each test instance is limited to 10. Moreover, to maintain a balance between correctly and incorrectly classified instances in the knowledge base comparable to that of the training set, the selection of knowledge base instances is weighted accordingly (Algorithm 2, Lines 3 and 4).

### 5.3 Evaluation of the proposed approach and comparison with existing methods

#### 5.3.1 Experimental methods

The relationship between the uncertainty scores of each method and misclassification events were determined using a logistic regression analysis to derive the odds ratio. The discriminative ability of the proposed uncertainty system in detecting misclassification events was assessed using the AUROC and AUPRC. To understand whether the proposed

**Table 2** Results of the multivariate logistic regression analysis, to understand the influence of knowledge base specification on quality of the uncertainty estimates

| Variable | Odds Ratio (95% CI) | std err | p-value |
|---|---|---|---|
| Constant | 0.34 (0.335, 0.344) | 0.002 | > 0.001 ** |
| Uncertainty | 1.985 (1.956, 2.012) | 0.014 | > 0.001 ** |
| Knowledge Base Size | 1.056 (1.038, 1.075) | 0.010 | > 0.001 ** |
| Knowledge Base Imbalance | 1.246 (1.23, 1.262) | 0.009 | > 0.001 ** |
| Training Set Size/Knowledge Base Size | 1.121 (1.1, 1.141) | 0.010 | > 0.001 ** |
| Density | 0.807 (0.797, 0.818) | 0.006 | > 0.001 ** |
| Synthetic Instance Ratio | 1 (0.986, 1.014) | 0.007 | 0.976 |
| $\Delta$ Knowledge Base Imbalance - Training Set Imbalance | 1.435 (1.416, 1.454) | 0.010 | > 0.001 ** |
| Uncertainty x Knowledge Base Size | 0.932 (0.915, 0.948) | 0.008 | > 0.001 ** |
| Uncertainty x Knowledge Base Imbalance | 0.977 (0.965, 0.99) | 0.007 | 0.001 * |
| Uncertainty x Training Set Size/Knowledge Base Size | 0.961 (0.945, 0.979) | 0.009 | > 0.001 ** |
| Uncertainty x Density | 1.3 (1.283, 1.318) | 0.009 | > 0.001 ** |
| Uncertainty x Synthetic Instance Ratio | 1.036 (1.021, 1.05) | 0.007 | > 0.001 ** |
| Uncertainty x $\Delta$ Knowledge Base Imbalance - Training Set Imbalance | 0.86 (0.849, 0.871) | 0.006 | > 0.001 ** |

**Algorithm 2** Training and application of meta-heuristics and uncertainty estimation.

**Inputs:** Training set $X\_train$, Test set $X\_train$, Knowledge Base $KB$, knowledge base instances per test instance $k$
**Outputs:** Sample Knowledge Base DataFrame $sampled\_KB$
Set $Balance$ as the misclassification rate of the model on $X\_train$ using cross validation
Set $n\_incorrect$ as the $k \times Balance$ rounded the nearest whole number
Initialise DataFrame $sampled\_KB$
Fit a nearest neighbours algorithm to the incorrectly classified instances in $KB$
Return the nearest $n\_incorrect$ instances for each instance in $X\_test$ and remove duplicates
Append knowledge base data of the returned instances to $sampled\_KB$
Fit a nearest neighbours algorithm to the correctly classified instances in $KB$
Return the nearest $k - n\_incorrect$ instances for each instance in $X\_test$ and remove duplicates
Append knowledge base data of the returned instances to $sampled\_KB$
**return** $sampled\_KB$

approach improves upon existing work, the following methods for estimating uncertainty were applied in the same manner to the same datasets:

- absolute difference of the maximum predicted probability from a uniform distribution (1)
- entropy of the predicted probabilities (2)
- trust score [10]
- aleatoric and epistemic uncertainty measures provided in the uncertainty-aware machine learning (UAML) library [5] (combined using a logistic regression equation derived from the training set)

A prominent limitation of model-dependent methods such as the use of predicted probabilities, is their susceptibility to over/underinflated estimates of confidence as a result of poor calibration. Therefore, to understand how each method is sensitive to factors such as model performance and calibration, logistic regression models were applied, with the classification outcome as the dependent variable and the inputs being the uncertainty estimate, the co-variate and the interaction term between the co-variate and the uncertainty estimate derived from each method. The accuracy score and Brier score of each respective test set were used as covariates.

### 5.3.2 Experimental results

All methods derived uncertainty scores which were significantly associated with misclassification events ($p < 0.001$). The uncertainty estimation method which was most strongly associated with misclassification events was the trust score [10] (OR = 2.543 [2.489, 2.599], $p < 0.001$), followed by the UAML estimates (2.043 [2.006, 2.081], $p < 0.001$) and the proposed methods (2.002 [1.968, 2.038], $p < 0.001$). The approaches calculating the absolute difference of the predicted probability from a uniform distribution and the entropy of the predicted probabilities, while still significant, demonstrated weaker relationships with misclassification events (OR = 1.869 [1.835, 1.903] and OR = 1.542 [1.499, 1.587], respectively). Figure 5 presents the discriminatory capability of the proposed system against existing methods. Results found that the trust score [10] proved optimal, both in terms of AUROC and AUPRC. However, the proposed approach outperformed all other methods in terms of AUROC and AUPRC.

Results indicated significant interactions between accuracy and all uncertainty estimation methods. The interaction between the trust score and accuracy was the greatest (OR = 2.069 [2.014, 2.126]), indicating that the relationship between the trust score and misclassification events is more dependent on the level of accuracy than other examined methods. The proposed methods recorded the smallest interaction (OR = 1.841 [1.804, 1.879]), demonstrating its relationship with misclassification events to be the most independent of model accuracy than other examined methods. Interactions between accuracy and predicted probabilities, entropy and the UAML-derived recorded odds ratios of 1.896 (1.858, 1.934), 1.982 (1.939, 2.026) and 1.973 (1.932, 2.015), respectively.

Significant interactions were observed between model calibration, assessed using the Brier score, and all uncertainty estimation methods. The interaction between the entropy-based uncertainty estimates and model calibration was the greatest (OR = 1.214 [1.201, 1.228]), indicating that the relationship between the entropy-based uncertainty estimates and misclassification events is more dependent on the calibration of the model than other examined methods. The proposed methods recorded the smallest interaction (OR = 1.019 [1.009, 1.029]), demonstrating its relationship with misclassification events to be the most independent of model calibration than other examined methods. Interactions between model calibration and predicted probabilities, the trust score and the UAML-derived recorded odds ratios of 1.042 (1.031, 1.052), 1.039 (1.024, 1.053) and 1.105 (1.094, 1.116), respectively.

Between the model-agnostic approaches to uncertainty estimation (the proposed methods and the trust score), their dependency on training data was assessed, in the same manner as performed with accuracy and model calibration. Results indicated significant interactions between the training set size and the uncertainty estimates. However, the magnitude of the interaction effect for the proposed methods was far smaller than that of the trust score (OR = 1.010 [1.000, 1.019] vs. 1.182 [1.165, 1.200]). This finding indi-

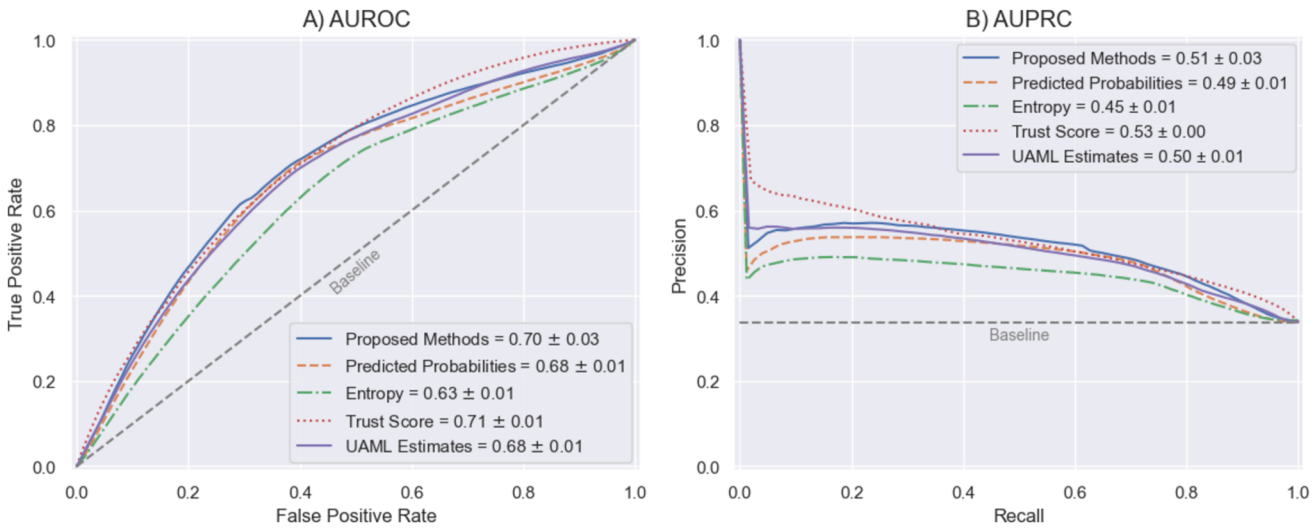



**Fig. 5** Aggregated results evaluating the discriminatory capability of each method for estimating uncertainty. A) Aggregated Receiver Operating Characteristic curves; B) Aggregated Precision-Recall curves

cates the quality of the uncertainty estimates generated using the proposed methods is largely independent of the size of the training set, and is likely due to the proposed methods benefiting from a meta-learning approach to forming the knowledge base.

## 6 Discussion

The proposed method proved to be an effective means of estimating the risk of a misclassification occurring within classification problems, which can be a crucial factor in instilling trust in AI applications, particularly in high-stakes domains. Compared with other approaches, the proposed methods outperform model-dependent approaches such as predicted probabilities and entropy-based approaches, whilst only being marginally outperformed by the trust score. A notable benefit of the proposed approach is that it proved to be most independent of model accuracy and calibration, compared with other existing methods.

There are several clear avenues for the application of such methods to improve both the robustness of AI development and the facilitation of trust calibration. In terms of improving model robustness, the estimated uncertainty can be used as a method for abstention, preventing decisions from being shown to the end-user when uncertainty surpasses a given threshold. Although such an approach can reduce the number of misclassifications, the number of instances for which the model is applicable also reduces. Therefore, the application of such a method may not be appropriate for all use cases. A second application of the proposed methods is for the improved explanation of predictive uncertainty. Shapley additive explanations (SHAP) can be applied to the calculated meta-heuristics and the estimated uncertainty to uncertainty to demonstrate how each concept of complexity influences the level of certainty in decision-making. A hypothetical example is presented in Fig. 6 as an example of how force plots may be employed to visualise the causes of increasing and decreasing uncertainty.

Given the performance of the proposed methods and their relevance towards the end user, there are some exciting opportunities for further research and exploration in enhancing AI reliability and trustworthiness. An emerging area in the field of precision medicine is the personalisation of model development to the patient instead of the overall task. Studies have investigated meta-learning techniques for dynamic classifier selection [55] and ensemble generation [56, 57] in this domain. Within the area of complexity, such methods could be applied to select models based on their ability to handle more challenging instances to maximise the likelihood of a correct classification [58]. In domains where large datasets are scarce, meta-learning approaches have the potential to overcome data limitations and learn viable solutions to parameter tuning and model architecture from similar problems before applying the learned principles to the current problem to arrive at a more optimal solution.

Existing end-user evaluation of AI explainability measures have demonstrated the effectiveness of presenting uncertainty values to improve trust calibration [28, 29]. However, the universal applicability of such explanations is lacking due to variations in end users' math and logic skills [11]. Given the interpretability of the proposed methods, for understanding how the various sources of complexity influence prediction uncertainty, there is scope for future work to focus on the application of GPT models for explaining uncertainty in terms end-users can understand. Existing work, such as the Talk2Model system, implemented a question-and-answer

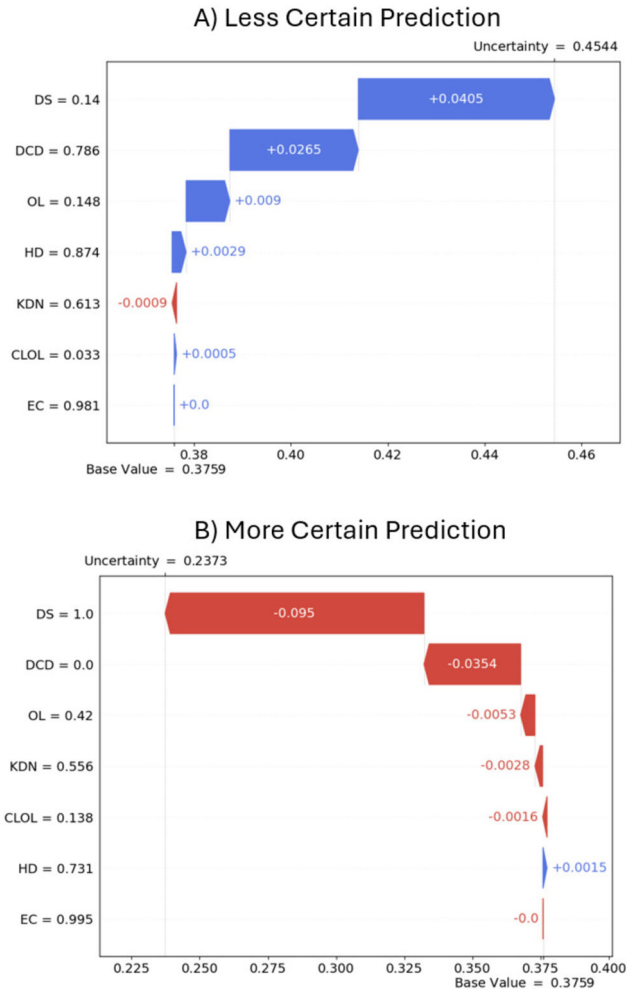



**Fig. 6** Force plots demonstrating the influence of each meta-heuristic on the uncertainty of decisions made by the classifier

system allowing end users to question why the model made the prediction that it did [59]. A real-world evaluation of the system applied to a disease prediction problem, found that 78% of healthcare workers felt confident in using the system and 73% of those expressed a willingness to use the Talk2Model system above other systems. By providing end-users with a more transparent means of engaging with and comprehending model uncertainty, it is hypothesised that trust calibration will be enhanced.

Despite the promise demonstrated by the proposed methods, there a some clear limitations which introduce further avenues for improvement. Investigations revealed meta-heuristics such as the hyperplane distance, offered negligible predictive value of misclassification risk when applied across a large number of datasets and classifiers. The reasoning behind this is likely due to the overly simplistic nature of the meta-heuristic's design. The rationale was to ensure that each meta-heuristic could be easily explained to an end-user, whilst accurately reflecting the the heuristics derivation. That said, it should be noted that linearly separable datasets are not common and therefore the applicability of the metric may not suit every classification problem, particularly in those within larger numbers of features. Therefore, explorations into the non-linear transformations of the input data could be explored as a means of improving the predictive capabilities of the meta-heuristic. Additionally, the datasets used in this paper are all structured tabular datasets. It is widely recognised that healthcare dataset often come in many modalities and therefore future work could look to modify and refactor the proposed methods to be applicable to other modalities such as imaging and natural language.

In conclusion, the proposed method of uncertainty estimation was effective in identifying instances that are more likely to be misclassified. Furthermore, the proposed meta-heuristics offer additional opportunities to improve the explainability of AI decision-making. Future research directions include the application of the proposed meta-heuristics to enhance model development through meta-learning and studying the impact of natural language explanations of uncertainty on the trust calibration of end-users.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s10489-024-06127-0.

**Acknowledgements** We wish to acknowledge the financial support of DMRC and Loughborough University who jointly funded the project.

**Author Contributions** A.H. and G.C. contributed to the design of the study. A.H. and G.C. conceived the experiments, A.H. conducted the experiment(s). A.H. analysed the results. A.H. and G.C. interpreted the findings. G.C. supervised the project. A.H. drafted the initial version of the manuscript. A.H. and G.C. reviewed the manuscript.

**Data Availability** All data used in this study are available publicly with sources provided in the Supplementary file.

## Declarations

**Ethical and Informed consent for Data Used** This study used only publicly-available data. As such, ethical approval and informed consent for data usage were not required.

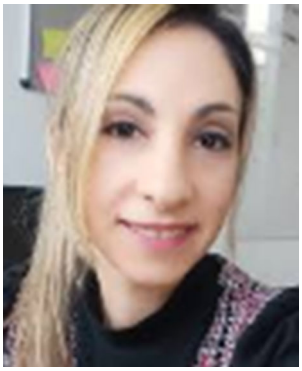
**Competing Interests Statement** The authors declare they have no competing interests to disclose.

## References


1. Asan O, Bayrak AE, Choudhury A et al (2020) Artificial intelligence and human trust in healthcare: focus on clinicians. J Med Int Res 22(6):e15154
2. Kompa B, Snoek J, Beam AL (2021) Second opinion needed: communicating uncertainty in medical machine learning. NPJ Digital Medicine 4(1):4
3. Tonekaboni S, Joshi S, McCradden MD, Goldenberg A (2019) What clinicians want: contextualizing explainable machine learning for clinical end use. In: Machine learning for healthcare conference, PMLR, pp 359–380
4. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q Consortium (2020) Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inf Dec Making 20:1–9
5. Hüllermeier E, Waegeman W (2021) Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Mach Learn 110:457–506
6. Joy Adowaa Buolamwini. *Gender shades: intersectional phenotypic and demographic evaluation of face datasets and gender classifiers*. PhD thesis, Massachusetts Institute of Technology, 2017
7. Spiegelhalter DJ (2008) Understanding uncertainty. Ann Fam Med 6(3):196–197
8. Michael R Smith and Tony Martinez. Improving classification accuracy by identifying and removing instances that should be misclassified. In *The 2011 International Joint Conference on Neural Networks*, pages 2690–2697. IEEE, 2011
9. Smith MR, Martinez T, Giraud-Carrier C (2014) An instance level analysis of data complexity. Mach Learn 95:225–256
10. Jiang H, Kim B, Guan M, Gupta M (2018) To trust or not to trust a classifier. Adv Neural Inf Process Syst 31:1
11. Suresh H, Lao N, Liccardi I (2020) Misplaced trust: Measuring the interference of machine learning in human decision-making. In: Proceedings of the 12th ACM conference on web science, pp 315–324
12. Houston A, Cosma G, Turner P, Bennett A (2021) Predicting surgical outcomes for chronic exertional compartment syndrome using a machine learning framework with embedded trust by interrogation strategies. Sci Rep 11
13. Kushankur Ghosh, Arghasree Banerjee, Sankhadeep Chatterjee, and Soumya Sen. Imbalanced twitter sentiment analysis using minority oversampling. In *2019 IEEE 10th international conference on awareness science and technology (iCAST)*, pages 1–5. IEEE, 2019
14. Chang EY, Li B, Wu G, Goh K (2003) Statistical learning for effective visual information retrieval. In *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*, IEEE, vol 3, pp III–609
15. López V, Fernández A, García S, Palade V, Herrera F (2013) An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Inf Sci 250:113–141
16. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
17. Elkan C (2001) The foundations of cost-sensitive learning. In: International joint conference on artificial intelligence, vol 17, pp 973–978. Lawrence Erlbaum Associates Ltd
18. Vuttipittayamongkol P, Elyan E, Petrovski A (2021) On the class overlap problem in imbalanced data classification. Knowl Based Syst 212:106631
19. Stefanowski J (2013) Overlapping, rare examples and class decomposition in learning classifiers from imbalanced data. Emerging paradigms in machine learning pp 277–306
20. Acuña E, Rodríguez C (2005) An empirical study of the effect of outliers on the misclassification error rate. Submitted to transactions on knowledge and data engineering
21. Islam R, Weir C, Del Fiol G (2014) Heuristics in managing complex clinical decision tasks in experts' decision making. In: 2014 IEEE international conference on healthcare informatics, pp 186–193. IEEE
22. Cioffi J (2001) A study of the use of past experiences in clinical decision making in emergency situations. Int J Nursing Stud 38(5):591–599
23. Qian Yang, Aaron Steinfeld, and John Zimmerman. Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–11, 2019
24. Taruscio D, Mantovani A (2021) Multifactorial rare diseases: Can uncertainty analysis bring added value to the search for risk factors and etiopathogenesis? Medicina 57(2):119
25. Barandas M, Folgado D, Santos R, Simão R, Gamboa H (2022) Uncertainty-based rejection in machine learning: Implications for model development and interpretability. Electronics 11(3):396
26. Shannon CE (1948) A mathematical theory of communication. Bell Syst Techn J 27(3):379–423
27. Platt J (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv Large Margin Classifiers 10(3):61–74
28. Zhang Y, Liao QV, Bellamy RKE (2020) Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In: Proceedings of the 2020 conference on fairness, accountability, and transparency, pp 295–305
29. Rechkemmer A, Yin M (2022) When confidence meets accuracy: exploring the effects of multiple performance indicators on trust in machine learning models. In: Proceedings of the 2022 chi conference on human factors in computing systems, pp 1–14
30. Nicora G, Rios M, Abu-Hanna A, Bellazzi R (2022) Evaluating pointwise reliability of machine learning prediction. J Biomed Inf 127:103996
31. Shaker MH, Hüllermeier E (2020) Aleatoric and epistemic uncertainty with random forests. In: Advances in intelligent data analysis XVIII: 18th international symposium on intelligent data analysis, IDA 2020, Konstanz, Germany, April 27–29, 2020, Proceedings 18, pp 444–456. Springer
32. Meijerink L, Cinà G, Tonutti M (2020) Uncertainty estimation for classification and risk prediction on medical tabular data. arXiv:2004.05824
33. Schulam P, Saria S (2019) Can you trust this prediction? auditing pointwise reliability after learning. In: The 22nd international conference on artificial intelligence and statistics, PMLR, pp 1022–1031
34. Guo C, Pleiss G, Sun Y, Weinberger KQ (2017a) On calibration of modern neural networks. In *International conference on machine learning*, PMLR, pp 1321–1330
35. Lewis DD, Gale WA (1994) A sequential algorithm for training text classifiers. In: Croft BW, van Rijsbergen CJ (Eds) SIGIR '94, London, Springer London, pp 3–12
36. Sharma M, Bilgic M (2017) Evidence-based uncertainty sampling for active learning. Data Min Knowl Disc 31:164–202
37. Roy N, McCallum A (2001) Toward optimal active learning through sampling estimation of error reduction. In: International conference on machine learning, Morgan Kaufmann, pp 441–448
38. Guo C, Pleiss G, Sun Y, Weinberger KQ (2017b) On calibration of modern neural networks. In: International conference on machine learning, PMLR, pp 1321–1330

39. Ribeiro MT, Singh S, Guestrin C (2016) Model-agnostic interpretability of machine learning. arXiv:1606.05386
40. Yao Y (2010) Three-way decisions with probabilistic rough sets. Inf Sci 180(3):341–353
41. Yao Y (2007) Decision-theoretic rough set models. In: Rough sets and knowledge technology: second international conference, RSKT 2007, Toronto, Canada, May 14-16, 2007. Proceedings 2, Springer, pp 1–12
42. Pawlak Z (1982) Rough sets. Int J Comput Inf Sci 11:341–356
43. Abdar M, Samami M, Mahmoodabad SD, Doan T, Mazoure B, Hashemifesharaki R, Liu L, Khosravi A, Acharya UR, Makarenkov V et al (2021) Uncertainty quantification in skin cancer classification using three-way decision-based bayesian deep learning. Comput Biol Med 135:104418
44. Chen Y, Yue X, Fujita H, Fu S (2017) Three-way decision support for diagnosis on focal liver lesions. Knowl Based Syst 127:85–99
45. Campagner Andrea, Cabitza Federico, Ciucci Davide (2020) The three-way-in and three-way-out framework to treat and exploit ambiguity in data. International Journal of Approximate Reasoning 119:292–312
46. Orcutt GH, Watts HW, Edwards JB (1968) Data aggregation and information loss. Am Econ Rev 773–787
47. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F (2011) A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans Syst Man Cybern Part C (Appl Rev) 42(4):463–484
48. Zhu R, Wang Z, Ma Z, Wang G, Xue J-H (2018) LRID: a new metric of multi-class imbalance degree based on likelihood-ratio test. Pattern Recognit Lett 116:36–42
49. Tang B, He H (2017) A local density-based approach for outlier detection. Neurocomputing 241:171–180
50. Vilalta R, Drissi Y (2002) A perspective view and survey of meta-learning. Artif Intell Rev 18:77–95
51. Frazier PI (2018) Bayesian optimization. In: Recent advances in optimization and modeling of contemporary problems, Informs, pp 255–278
52. Macia N, Bernadó-Mansilla E (2014) Towards UCI+: a mindful repository design. Inf Sci 261:237–262
53. Houston A, Cosma G (2022) A genetically-optimised artificial life algorithm for complexity-based synthetic dataset generation. Inf Sci
54. Tin Kam Ho and Mitra Basu (2002) Complexity measures of supervised classification problems. IEEE Trans Pattern Anal Mach Intell 24(3):289–300
55. Cruz RMO, Sabourin R, Cavalcanti GDC (2018) Dynamic classifier selection Recent advances and perspectives. Inf Fusion 41:195–216
56. Cruz RMO, Sabourin R, Cavalcanti GD, Ren TI (2015) Meta-des: A dynamic ensemble selection framework using meta-learning. Pattern Recognit 48(5):1925–1935
57. Guo C, Liu M, Lu M (2021) A dynamic ensemble learning algorithm based on k-means for ICU mortality prediction. Appl Soft Comput 103:107166
58. Cruz RMO, Zakane HH, Sabourin R, Cavalcanti GDC (2017) Dynamic ensemble selection vs k-nn: why and when dynamic selection obtains higher classification performance? In: 2017 Seventh international conference on image processing theory, tools and applications (IPTA), IEEE, pp 1–6
59. Slack D, Krishna S, Lakkaraju H, Singh S (2023) Explaining machine learning models with interactive natural language conversations using talktomodel. Nat Mach Intell 5(8):873–883
60. Czerniak J, Zarzycki H (2003) Application of rough sets in the presumptive diagnosis of urinary system diseases. In: Artificial intelligence and security in computing systems, Springer, pp 41–51
61. Patrício M, Pereira J, Crisóstomo J, Matafome P, Gomes M, Seiça R, Caramelo F (2018) Using resistin, glucose, age and BMI to predict the presence of breast cancer. BMC Cancer 18(1):1–8
62. Dua D, Graff C (2019) UCI machine learning repository. http://archive.ics.uci.edu/ml
63. Antal B, Hajdu A (2014) An ensemble-based system for automatic screening of diabetic retinopathy. Knowl Based Syst 60:20–27
64. Islam MM, Ferdousi R, Rahman S, Bushra HY (2020) Likelihood prediction of diabetes at early stage using data mining techniques. In: Computer vision and machine intelligence in medical image analysis, Springer, pp 113–125
65. Gil D, Girela JL, De Juan J, Gomez-Torres MJ, Johnsson M (2012) Predicting seminal quality with artificial intelligence methods. Expert Syst Appl 39(16):12564–12573
66. Chicco Davide, Jurman Giuseppe (2020) Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med Inform Decis Mak 20(1):1–16
67. Tsanas A, Little MA, Fox C, Ramig LO (2013) Objective automatic assessment of rehabilitative speech treatment in parkinson's disease. IEEE Trans Neural Syst Rehabil Eng 22(1):181–190
68. Sergey E Golovenkin, Jonathan Bac, Alexander Chervov, Evgeny M Mirkes, Yuliya V Orlova, Emmanuel Barillot, Alexander N Gorban, and Andrei Zinovyev. Trajectories, bifurcations, and pseudo-time in large clinical datasets: Applications to myocardial infarction and diabetes data. *GigaScience*, 9(11):giaa128, 2020
69. Van De Vijver MJ, He YD, Van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ et al (2002) A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347(25):1999–2009
70. Sakar CO, Serbes G, Gunduz A, Tunc HC, Nizam H, Sakar BE, Tutuncu M, Aydin T, Isenkul ME, Apaydin H (2019) A comparative analysis of speech signal processing algorithms for parkinson's disease classification and the use of the tunable q-factor wavelet transform. Appl Soft Comput 74:255–263
71. Debernardi S, O'Brien H, Algahmdi AS, Malats N, Stewart GD, Plješa-Ercegovac M, Costello E, Greenhalf W, Saad A, Roberts R et al (2020) A combination of urinary biomarker panel and pancrisk score for earlier detection of pancreatic cancer: A case-control study. PLoS Med 17(12):e1003489
72. Little M, McSharry P, Hunter E, Spielman J, Ramig L (2008) Suitability of dysphonia measurements for telemonitoring of parkinson's disease. Nat Preced 1–1
73. Zięba M, Tomczak JM, Lubicz M, Świtek J (2014) Boosted svm for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. Appl Soft Comput 14:99–108

**Andrew Houston** is a Research Data Scientist at Barts Life Sciences, Barts Health NHS Trust, UK. He holds a PhD in Computer Science from the Loughborough University. His research focuses on improving the robustness and transparency of AI models in healthcare settings.



**Georgina Cosma** is a professor of AI and Data Science at Loughborough University, UK. She holds a PhD in Computer Science from the University of Warwick. Her research focuses on AI in healthcare, and in neural information processing, modelling, and retrieval.