Evidential Random Forests

Arthur Hoarau, Arnaud Martin, Jean-Christophe Dubois, Yolande Le Gall

Evidential Random Forests

Hoarau Arthur[a,*], Martin Arnaud[a], Dubois Jean-Christophe[a], Le Gall Yolande[a]

[a] *Univ Rennes, CNRS, IRISA, DRUID, Lannion, 22000, France*

Abstract

In machine learning, some models can make uncertain and imprecise predictions, they are called evidential models. These models may also be able to handle imperfect labeling and take into account labels that are richer than the commonly used hard labels, containing uncertainty and imprecision. This paper proposes an Evidential Decision Tree, and an Evidential Random Forest. These two models use a distance and a degree of inclusion to allow the model to group observations whose response elements are included in each other into a single node. Experimental results showed better performance for the presented methods compared to other evidential models and to recent Cautious Random Forests when the data is noisy. The models also offer a better robustness to the overfitting effect when using datasets that are effectively uncertainly and imprecisely labeled by the contributors. The proposed models are also able to predict rich labels, an information that can be used in other approaches, such as active learning.

Keywords: Decision Tree, Random Forest, Classification, Rich labels, Dempster-Shafer Theory

1. Introduction

In supervised learning one of the best known approaches is the decision trees: amongst the different classification methods, they have the advantage of being easily understandable, and their interpretation is within the reach of a larger number of people (Quinlan, 1987). They can be used in both classification and regression for quantitative and qualitative variables.

*Corresponding author

The two most popular decision tree models are C4.5 (Quinlan, 1993) and CART (Breiman et al., 1984). The effectiveness of these models is recognized, they are simple to define, have good interpretability and can be used in exploratory analysis (Siciliano, 1998). However, decision trees are prone to overfitting, this occurs when any learning process over-optimizes the error on the learning set at the expense of generalization (Bramer, 2013).

The labeling process is often carried out by humans (Fredriksson et al., 2020) and may therefore be subject to imperfection. Using hard labels might be convenient for many machine learning and deep learning problems but is never completely representative of reality. On the other hand, imperfection can help us fill in this lack of information using richer labels[1]. It can be represented by many criteria (Smets, 1997) but only uncertainty and imprecision will be discussed in this paper. *Uncertainty* can be considered as a partial knowledge of the real value of the data (*e.g. This might be a cat*). *Imprecision* measures a quantitative defect of knowledge (*e.g. This is a cat or a dog*). Ignorance is thus derived from imprecision. This ignorance and imprecision can be modeled with the theory of belief functions (Dempster, 1967; Shafer, 1976). This framework, commonly used for data fusion, allows to represent several degrees of ignorance, and generalizes other approaches, like probabilities or possibilities.

Evidential decision trees can handle this imperfection in labels. Several authors (Denoeux et al., 2019; Elouedi et al., 2001; Trabelsi et al., 2019; Sutton Charani et al., 2013) have offered to couple the theory of belief functions with decision trees in order to handle these imperfectly labeled data. However, these models tend to be overfitted due to the richness of these labels, resulting in large trees with small leaves. In this paper we suggest an Evidential Decision Tree for rich labels robust to overfitting and based on a conflict measure introduced by Martin (2019). Other evidential decision trees do not address the problem of overfitting (other than by pruning), the use of conflict allows to group observations with similar response elements and to reduce this effect. The goal is both to come up with a model able to work with imperfectly labeled observations, but also to address the problem of overfitting.

Moreover, we extend the proposed model to Evidential Random Forests to

[1]Rich labels are answer elements given by a source that may have several degrees of imprecision (*c.f.* section 2.1).

overcome the high variance of decision trees and gain performance. Random Forests are first introduced by Leo Breiman (1996, 2001), they combine the predictions of a large number of trees using *bagging* and *random feature selection*.

Several models based on belief functions have been proposed to deal with data imperfections. Cautious Random Forests (Zhang et al., 2023; Moral-García et al., 2020) is a recent model that can produce uncertain and imprecise predictions, as well as cautious ones, but cannot take into account all the uncertainties present in the rich labels, unlike Evidential Random Forests.

Recent evidential models, such as Evidential SVMs (Xu et al., 2016; Kadir et al., 2019) or Evidential Deep Neural Networks (Yuan et al., 2020) can handle some information about the uncertainty inherent in the quality of the observation(s) - for example, an image may represent a cat uncertainly because it strongly resembles a dog -, while Evidential Random Forests take into account label uncertainty - the image undoubtedly represents a cat, but the source that labelled it lacks knowledge and thus induces uncertainty in its labelling.

The Evidential K-Nearest Neighbors (Denœux, 1995; Denoeux et al., 2019) is able to take into account these rich labels in the same way as the proposed competitive model, the Evidential Random Forests.

In the experiments conducted for this study, the proposed method is compared with most of these models

The rest of the paper is organized as follows. Section 2 reviews richer labels as well as the theory of belief functions, Decision Trees and Random forests. Section 4 describes the proposed conflict-based Evidential Decision Tree as well as the Evidential Random Forest. They both use a distance and a degree of inclusion to allow the model to group observations whose response elements are included in each other into a single node. By doing so, the built tree is shallower and less over-trained. Experiments on imperfectly labeled and noisy datasets are discussed in section 5. Finally, section 6 and 7 conclude the article.

2. Background

2.1. Richer labeling

Most of the datasets used for classification consider only hard labels (*i.e.* "*Dog*" or "*Cat*"). In this paper, we refer as rich labels the elements of response given by a source that may include several degrees of imprecision

(*i.e.* "*This might be a cat*", "*I don't know*" or "*I am hesitating between dog and cat, with a slight preference for cat*)". In this document, these uncertain and soft labels, will be called *rich* labels as opposed to hard labels and they are modeled using the theory of belief functions.

2.2. Theory of belief functions

The theory of belief functions, also called Dempster-Shafer theory (Dempster, 1967; Shafer, 1976), is used in this paper in order to model uncertainty and imprecision, for both labeling and prediction.

Let $\Omega = \{\omega_1, \ldots, \omega_M\}$ be the frame of discernment for M exclusive and exhaustive hypotheses. The power set 2^Ω is the set of all subsets of Ω. A Basic Belief Assignment is the belief that a source may have about the elements of the power set of Ω, this function assigns a mass to each element of this power set such that the sum of all masses is equal to 1.

$$\begin{aligned} & m : 2^\Omega \rightarrow [0,1], \\ & \sum_{A \in 2^\Omega} m(A) = 1. \end{aligned} \tag{1}$$

Each subset $A \in 2^\Omega$ such that $m(A) > 0$ is called a *focal element* of m. The uncertainty is therefore represented by a mass $m(A) < 1$ on a focal element A and the imprecision is represented by a non-null mass $m(A) > 0$ on a focal element A such that $|A| > 1$.

A mass function m is called *categorical mass function* when it has only one focal element such that $m(A) = 1$. In the case where A is a set of several elements, the knowledge is certain but imprecise. For $|A| = 1$, the knowledge is certain and precise.

A mass function m is called *simple support mass function* when it has two focal elements, one of which is Ω:

$$\begin{aligned} & m(A) = 1 - \boldsymbol{w}, \ A \in 2^\Omega, \\ & m(\Omega) = \boldsymbol{w}, \\ & m(B) = 0, \ B \in 2^\Omega \backslash \{A, \Omega\}. \end{aligned} \tag{2}$$

with $\boldsymbol{w} \in [0,1]$, the mass function m can then be noted $A^{\boldsymbol{w}}$.

On decision level, the pignistic probability $BetP$ helps decision making on singletons:

$$BetP(\omega) = \sum_{A \in 2^\Omega, \, \omega \in A} \frac{m(A)}{|A|}. \tag{3}$$

It is also possible to combine several mass functions into a single body of evidence. The normalized conjunctive combination of the mass functions m_j derived from N sources is given as follows:

$$\begin{cases} m(A) = \dfrac{1}{1-\kappa} \displaystyle\sum_{B_1 \cap \ldots \cap B_N = A} \prod_{j=1}^{N} m_j(B_j), \ \text{ if } A \neq \emptyset, \\ m(\emptyset) = 0, \end{cases} \tag{4}$$

with $\emptyset$ the empty set and:

$$\kappa = \sum_{B_1 \cap \ldots \cap B_N = \emptyset} \prod_{j=1}^{N} m_j(B_j). \tag{5}$$

This combination rule will not be used, because for several datasets, the labels and therefore the mass functions are not independent. Also, a combination of mass functions will be used on the prediction of the different estimators (*e.g.* the aggregated decision trees for the prediction of Random Forest). Equation (4) is unstable when used on a large number of estimators and for the output of the trees used. A simple average of the mass functions will be preferred and is defined as follows:

$$m(A) = \frac{1}{N} \sum_{j=1}^{N} m_j(A), \ \ A \in 2^{\Omega}. \tag{6}$$

Example 1:

Let $\Omega = \{Cat, Dog\}$ be a frame of discernment. An observation labeled "Cat" by a source can be modeled in the framework of belief functions by the mass function m_1 such that: $m_1(\{Cat\}) = 1$ and $m_1(A) = 0, \ \forall A \in 2^{\Omega} \backslash \{Cat\}$.

Example 2:

An observation labeled "Cat or Dog" by a source can be modeled by the mass function m_2 such that: $m_2(\{Cat, Dog\}) = 1$ and $m_2(A) = 0$, $\forall A \in 2^{\Omega} \backslash \{Cat, Dog\}$.

Example 3:

The mean mass function $\bar{m}$ of m_1 and m_2 is: $\bar{m}(\{Cat\}) = 0.5$, $\bar{m}(\{Cat, Dog\}) = 0.5$ and $\bar{m}(A) = 0$ for all other subsets A in 2^{Ω}. Its pignistic probability $BetP$, used for decision making is: $BetP(\{Cat\}) = 0.75$ and $BetP(\{Dog\}) = 0.25$.

2.3. Evidential Learning

Evidential learning is defined here as the set of models able to both use these rich labels and represent their prediction with belief functions. Such information can be used to represent knowledge or to increase the performance of the model (Hoarau et al., 2022). Evidential decision trees have been introduced (Denoeux & Bjanger, 2000; Elouedi et al., 2001; Trabelsi et al., 2019), but evidential learning is not limited to decision trees, other models like K-Nearest Neighbors have been extended to the evidential world (Denœux, 1995; Denoeux et al., 2019) and even evidential neural networks are modeled (Yuan et al., 2020). Some of these models are presented and used for comparison during the experiments. Here, we are interested in proposing an Evidential Random Forest based on a new Evidential Decision Tree.

2.4. Decision Trees

In classification, Decision Trees are used to predict the class of an incoming observation based on a structure of nodes and splits. The tree is previously built on a training set and the different splits define the path to follow for prediction.

A *node* contains observations, it can have one parent and two or more child nodes (depending on the chosen architecture). A *split* divides a node into child nodes using *edges* corresponding to the possible values of an attribute. The choice of the best split attribute is made by maximizing a gain function. At first, every observation used for training is part of a single node called the *root node*.

The most popular decision trees are the versions C4.5 (Quinlan, 1993) and CART (Breiman et al., 1984). They are referred to as *top down induction decision trees* and defined by an *attribute selection criterion* used to find the best attribute for a split, a *partitioning strategy* to divide the node using the selected attribute and *stopping criteria* to stop splitting at a node and make it a *leaf*.

Partitioning strategy

The selected attribute to split a node is the one that maximizes a gain function. Let the node S be a set of observations (if all the training observations are in S, it is the root node). Let $\Omega = \{\omega_1, \ldots, \omega_M\}$ be the set of all possible classes for each element of S. Let $\mathcal{A}$ be an attribute from its finite domain $\mathcal{D}_{\mathcal{A}}$. The information gain $Gain(S, \mathcal{A})$ of splitting on $\mathcal{A}$ is defined by

Quinlan:

$$Gain(S, \mathcal{A}) = Info(S) - Info_{\mathcal{A}}(S), \tag{7}$$

where $Info(S)$ is the information of the node S, and $Info_{\mathcal{A}}(S)$ is the weighted sum of the child nodes information considering a split on attribute $\mathcal{A}$:

$$Info_{\mathcal{A}}(S) = \sum_{v \in \mathcal{A}} \frac{|S_v|}{|S|} Info(S_v), \tag{8}$$

with S_v the subset of S for which the attribute $\mathcal{A}$ has the value v (*i.e.* the child node resulting from the split on attribute A for the value v). A split is performed on each node, starting from the root and recursively for each child node until one of the stopping criteria is reached.

Attribute selection criterion

Maximizing the *Gain* means choosing the best attribute $\mathcal{A}$ to split S. To do this, Quinlan proposes the Shannon (1948) entropy as a selection criterion:

$$Info(S) = -\sum_{\omega \in \Omega} p_\omega(S) \log_2 p_\omega(S), \tag{9}$$

with $p_\omega(S)$ the proportion of observations in S belonging to the class ω. The Gini criterion is also commonly used as a selection criterion:

$$Info(S) = 1 - \sum_{\omega \in \Omega} p_\omega(S)^2. \tag{10}$$

Stopping criteria

The construction of the tree is stopped when one of these criteria is reached:

- Only one observation is part of the current node.
- Observations belong to the same class.
- The remaining attributes have a *Gain* less or equal to zero.

Pruning is a compression technique reducing the size of a tree. Pre-pruning (pruning during the construction of the tree) will be lightly discussed in this document but when not specified and with respect to Random Forest, consider that the tree is fully grown.

In the case of continuous variables, the method used is the following; for an attribute $\mathcal{A}$, the observations are sorted in ascending order on this attribute and for each pair of consecutive values c_1 and c_2, a threshold v will be used such that $v = c_1 + (c_2 - c_1)/2$. The threshold maximizing the information gain will be used to split a node into two child nodes, one with values strictly lower than v and the other with values greater than v.

Prediction

Once the tree is built, the unlabeled observations will cross the tree from the root based on their attributes. When a leaf is reached, observations will be given a probability of belonging to each class based on the proportion representing the class in the node. The class maximizing this probability is the predicted class (*e.g.* if a new observation reaches a leaf $\mathcal{I}$, composed of 9 observations of class ω_1 and 1 observation of class ω_2, the predicted probability associated with ω_1 is 0.9 and that associated with ω_2 is 0.1).

2.5. Random Forests

Introduced by Breiman (1996) the Bagging is the first step towards Random Forest, later defined by Breiman (2001) by adding a random selection of features. The principle is to overcome the weakness of decision trees, the high variance, by combining the predictions of a large number of trees, the *forest*. This section explains the operation of bagging and random feature selection, both of which are used in the most commonly used version of Random Forest (Breiman, 2001).

2.5.1. Bagging

This definition is largely based on Breiman's publication (Breiman, 1996) where bagging on decision trees is first introduced. Let $\Omega = \{\omega_1, \ldots, \omega_M\}$ be a collection of M different classes and let $\mathcal{L} = \{x_k|1, \ldots, K\}$ be a learning set of K samples where each element is associated to a label $y_k \in \Omega$. Given an estimator $\varphi(x, \mathcal{L})$, the objective is to create a $\{\mathcal{L}_n\}$ sequence of new learning sets to improve the performance of the predictor on the unique learning set. A new sequence of predictors $\{\varphi(x, \mathcal{L}_n)\}$ is then introduced. Bagging stands for **B**ootstrap **Agg**regat**ing** where the bootstrap allows to create the $\{\mathcal{L}_n\}$ learning sets, the result is then given by aggregating the $\{\varphi(x, \mathcal{L}_n)\}$ predictors.

Bootstrap: The $\{\mathcal{L}_n\}$ sets are composed of K elements (*i.e.* in Bagging, the size of the $\{\mathcal{L}_n\}$ learning set is the same as $\mathcal{L}$). Each element is drawn

at random *with replacement*, in $\mathcal{L}$. Which means that an observation x_k can appear several times or not at all in any $\mathcal{L}_n$ set.

Aggregating: When $\varphi(x, \mathcal{L})$ predicts a class (*c.f.* (Breiman, 1996) for continuous predictions) a method for aggregating the $\{\varphi(x, \mathcal{L}_n)\}$ predictions is to use a majority vote or a weighted vote. The results of the N trees are then aggregated, to form the prediction of the unique Random Forest model.

2.5.2. Random feature selection

Inspired by Amit & Geman (1997), Breiman introduces the random feature selection for Random Forest (Breiman, 2001). The first objective is to be competitive with the Adaboost model. The author says that "using a random selection of features to split each node yields error rates that compare favorably to Adaboost, but are more robust with respect to noise". It works by using only a small number of variables, randomly selected at each split, when training a Decision Tree. In this paper, for Random Forest, Cautious Random Forest and the proposed Evidential Random Forest, $\sqrt{p}$ variables will be randomly selected at each split, with p the number of variables and the dimension of the vector x (Hastie et al., 2009). During a split, a Decision Tree used in a Random Forest only has access to a reduced number of variables in order to find the best possible split.

3. Motivation from other Evidential Decision Trees

For observations that are no longer perfectly labeled (*i.e.* labeled by a mass function in this case), the *Gain* (7) cannot be calculated with Entropy (9) or Gini criterion (10). The proportion p_ω of observations belonging to the class ω no longer exists as an observation is no longer characterized by a class but by a mass function. To make the gain calculation possible again, a new information criterion $Info$ must be used, such models are presented here.

3.1. The uncertainty approach

Some authors introduced a new decision tree (Denoeux & Bjanger, 2000) based on the theory of belief functions and Klir uncertainty (Klir & Wierman, 1998). Observations in child nodes are grouped while reducing as much as possible the uncertainty present in each node. Two information are combined to calculate the information gain, first the non-specificity:

$$N(m) = \sum_{A \subseteq \Omega} m(A) \log_2(|A|), \tag{11}$$

and an extension of the Shannon entropy, the degree of discord:

$$D(m) = -\sum_{A \subseteq \Omega} m(A) \log_2(BetP(A)). \tag{12}$$

Using both non-specificity and discord, the authors propose the following information criterion:

$$Info(S) = (1 - \lambda) N(\bar{m}^S) + \lambda D(\bar{m}^S), \tag{13}$$

with $\lambda \in [0, 1]$ a positive coefficient and $\bar{m}^S$ the mass function representing the subset S. In this paper we choose the average of the mass functions of the observations belonging to node S. The value of $\bar{m}^S$ is therefore the combination of each rich label (*i.e.* every mass function) in the node according to the average. Dempster's combination is not used because for some datasets, the labels and thus the masses are not independent. The λ parameter acts as a cursor on the degree of non-specificity or discord that one wants as a criterion.

This method will be called the Uncertainty-EDT, with $\lambda = 0.5$. For all the experiments presented, more non-specificity in the criterion (lowering λ) only lowered the performance of the model while more discord (increasing λ) increased performance with a threshold reached at about 0.5.

3.2. The Euclidean distance approach

Another approach, based on the Euclidean distance between mass functions is presented by Elouedi et al. (2001) (the authors originally suggested a one-to-one correspondence vector instead of the mass). It is proposed to minimize the intra-class distance in the child nodes (*i.e.* the mass functions close in the sense of the Euclidean distance will be grouped to form the splits). The Euclidean distance $d(m_i, m_j)$ between two mass functions m_i and m_j is defined as follows:

$$d(m_i, m_j) = \sqrt{\sum_{A \subseteq \Omega} (m_i(A) - m_j(A))^2}. \tag{14}$$

The information $Info(S)$ in the subset S is then given by the mean of the distances between the mass functions of all observations of node S and the model using a Euclidean distance as a criterion will be called Euclidean-EDT. As a reminder, the mass functions are the labels of the observations, given in the dataset.

3.3. The Jousselme distance approach

Trabelsi et al. (2019) introduce a decision tree using the Jousselme distance (Jousselme et al., 2001). Close to the Euclidean approach, the distance used differs by considering the mass functions as bodies of evidence and not simple vectors. The Jousselme distance $d_J(m_i, m_j)$ is defined by:

$$d_J(m_i, m_j) = \sqrt{\frac{1}{2}(m_i - m_j)^T \mathcal{D}(m_i - m_j)}, \tag{15}$$

with $\mathcal{D}$ the $2^M \times 2^M$ matrix of M classes computed as follows:

$$\mathcal{D}(A, B) = \frac{|A \cap B|}{|A \cup B|}, \quad A, B \in \Omega. \tag{16}$$

The information $Info(S)$ can be obtained by computing the mean of Jousselme distances between mass functions in node S. This model differs from the Euclidean one by assigning a smaller distance between two response items that contain some of the same information, it will be referenced as the Jousselme-EDT.

All these methods have the advantage of working with imperfectly labeled data, but are prone to overfitting, as will be shown in the section 5. This leads to our proposal for a more robust criterion.

4. Evidential Decision Trees and Evidential Random Forest

Few models are able to take uncertainty and imprecision into account. This imperfection in richer labels can be used to approximate what a source thinks about an observation, and can even increase the performance of a model (Hoarau et al., 2022). In this section we propose a new separation criterion and a new Evidential Random Forest composed of Evidential Decision Trees using this criterion.

4.1. Conflict-based Evidential Decision Trees

Amongst the models presented in section 3, two properties can be discussed. The first is that the degree of discord at equation (12) can be greater than zero for a single response from a single user, and thus cannot best represent the contradiction between several responses. The second is that for the use of distances, at equations (14) and (15), the impurity in a node is non-null when answer elements are included in each other. To address these two problems, the notion of conflict in the theory of belief functions can be used.

4.1.1. The Conflict approach

In this paper, we propose to use a conflict measure as a split criterion. This conflict measure based on inclusion degree and distance is introduced by Martin (2019), and two definitions of inclusion are given. The first is a *strict inclusion* saying that a mass function m_i is included in m_j when all focal elements of m_i are included one by one in each focal element of m_j. A *strict degree of inclusion* $\delta_s^{i\subseteq j}(m_i, m_j)$ of m_i in m_j is then given by:

$$\delta_s^{i\subseteq j}(m_i, m_j) = \frac{1}{|\mathcal{F}_i||\mathcal{F}_j|} \sum_{A\in\mathcal{F}_i} \sum_{B\in\mathcal{F}_j} Inc(A, B), \tag{17}$$

with $Inc(A, B) = 1$ if $A \subseteq B$ and 0 otherwise, $\mathcal{F}_i$ and $\mathcal{F}_j$ respectively are the set of focal elements of m_i and m_j.

The second definition is a *light inclusion* saying that a mass function m_i is included in m_j if all the focal elements of m_i are included in at least one element of m_j. The *light degree of inclusion* $\delta_l^{i\subseteq j}(m_i, m_j)$ is given as follows:

$$\delta_l^{i\subseteq j}(m_i, m_j) = \frac{1}{|\mathcal{F}_i|} \sum_{A\in\mathcal{F}_i} \max_{B\in\mathcal{F}_j}(Inc(A, B)). \tag{18}$$

We introduce and define in this paper $\delta^{i\subseteq j}(m_i, m_j)$, a slightly less strict degree of inclusion than equation (17), without taking into account the inclusion on ignorance.

Definition 1. *A fair inclusion between two mass functions is defined by: a mass function m_i is fair included in m_j if all the focal elements of m_i on $2^\Omega\backslash\Omega$ are included one by one in each focal element of m_j on $2^\Omega\backslash\Omega$.*

Definition 2. *The* fair degree of inclusion $\delta^{i\subseteq j}(m_i, m_j)$ *is given as follows:*

$$\delta^{i\subseteq j}(m_i, m_j) = \frac{1}{|\mathcal{L}_i||\mathcal{L}_j|} \sum_{A\in\mathcal{L}_i} \sum_{B\in\mathcal{L}_j} Inc(A, B), \tag{19}$$

with $\mathcal{L}_i$ and $\mathcal{L}_j$ respectively the set of focal elements on $2^\Omega\backslash\Omega$ of m_i and m_j.

This equation is used instead of the strict inclusion equation because the ignorance is only included in itself.

Example. Let y_1, y_2 and y_3 be rich labels with $\Omega = \{dog,\ cat,\ bird\}$ according to Section 2.1 such that:

y_1**:** $m_1(\{dog\}) = 1$, *"This is a dog"*.

y_2**:** $m_2(\{cat\}) = 1$, *"This is a cat"*.

y_3**:** $m_3(\{cat,\ bird\}) = 1$, *"This is a cat or a bird"*.

Fair degrees of inclusion for some evidence body couples are $\delta^{1\subseteq 1}(y_1, y_1) = 1$, $\delta^{1\subseteq 2}(y_1, y_2) = 0$, $\delta^{1\subseteq 3}(y_1, y_3) = 0$, $\delta^{2\subseteq 3}(y_2, y_3) = 1$ and $\delta^{3\subseteq 2}(y_3, y_2) = 0$.

4.1.2. Computing the information criterion

Based on the fair degree of inclusion, a degree of inclusion $\delta(m_i, m_j)$ of m_i and m_j is:

$$\delta(m_i, m_j) = \max(\delta^{i\subseteq j}(m_i, m_j), \delta^{j\subseteq i}(m_j, m_i)), \tag{20}$$

and the conflict measure $\mathcal{C}(m_i, m_j)$, used as a criterion in our proposed Evidential Decision Tree is defined by:

$$\mathcal{C}(m_i, m_j) = (1 - \delta(m_i, m_j))d_J(m_i, m_j), \tag{21}$$

with $d_J(m_j, m_i)$ the Jousselme distance between m_i and m_j. The calculation of $\mathcal{C}$ gives the conflict, the information $Info(S)$ is then, the average two by two conflict of node S:

$$Info(S) = \frac{\sum_{x_i \in S} \sum_{x_j \in S} \mathcal{C}(m_i, m_j)}{|S|^2 - |S|}. \tag{22}$$

The gain is calculated identically to decision trees with equation (7). This approach differs from the Jousselme-EDT by allowing two observations to belong to the same node, without loss of gain, if one response is *included* in the other. The proposed model based on conflict will be referenced in this paper as the Conflict-EDT.

4.1.3. Prediction

Once the tree is built, a new observation will cross the tree from the root according to the value of its attributes. When a leaf is reached, the observation will be assigned a mass function equal to the mean mass function in the node (*e.g.* if a new observation reaches a leaf $\mathcal{I}$, composed of 10 elements of masses $\{m_1, \ldots, m_{10}\}$; then the predicted mass of the observation will be $\bar{m}$ the average of the 10 mass functions). The class maximizing the pignistic probability (3) of this mass function is the predicted class.

Example. Let $\Omega = \{\omega_1, \omega_2\}$ be the set of possible classes in a classification problem. Let S be a leaf reached by observation x and composed of two mass functions m_1 and m_2 such that:

m_1**:** $m_1(\{\omega_1\}) = 0.3$, $m_1(\Omega) = 0.7$,

m_2**:** $m_2(\{\omega_1\}) = 0.2$, $m_2(\Omega) = 0.8$.

The average mass function $\bar{m}$ in S is the predicted rich label of x by the Evidential Decision Tree: $\bar{m}(\{\omega_1\}) = 0.25$, $\bar{m}(\Omega) = 0.75$. On decision level the class maximizing the pignistic probability is ω_1 and is then chosen (*i.e.* the hard label used to calculate the accuracy of the model).

4.1.4. Robustness to overfitting

When handling perfectly labeled data and classical decision trees, many observations have the same class and the tree stops growing when all observations of a node are of the same class. There is no better split possible reducing the impurity of the node and thus increasing the *Gain*.

Now, when dealing with imperfectly labeled data with the theory of belief functions, almost no observation has an identical label. The decision tree therefore has a harder time to stop growing, resulting in an over-trained tree that will have difficulty generalizing. With Conflict-EDT, a node of several masses can have a null conflict if the mass functions are included in each other. This allows, as in the case of perfectly labeled data, to have a null *Gain* and to stop the growth of the tree.

Example. Let $\Omega = \{\omega_1, \omega_2\}$ be a frame of discernment. Let three observations x_1, x_2 and x_3 respectively labeled m_1, m_2 and m_3 be part of a root node such that:

m_1**:** $m_1(\omega_1) = 0.9$, $m_1(\Omega) = 0.1$,

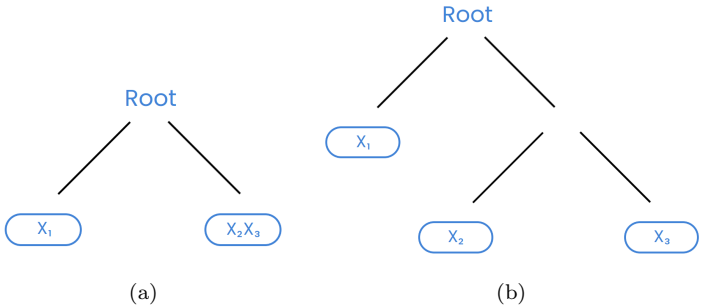


Figure 1: A shallower decision tree (a) and a deeper decision tree (b) of three observations x_1, x_2 and x_3.

m_2**:** $m_2(\omega_2) = 0.8$, $m_2(\Omega) = 0.2$,

m_3**:** $m_3(\omega_2) = 0.9$, $m_3(\Omega) = 0.1$.

The mass function m_1 can be seen as a strong belief that observation x_1 belongs to ω_1. The two mass functions m_2 and m_3 can be seen as strong beliefs that x_2 and x_3 belong to ω_2 with a lower share of ignorance for x_3.

Using the Uncertainty-EDT model and if there are enough attributes to separate the nodes, the discord calculated in equation (12) by combining masses m_2 and m_3 will be non-null, resulting in the deepest possible tree, visible in Figure 1b. The same property is present for Euclidean-EDT and Jousselme-EDT models, both Euclidean and Jousselme distances between m_2 and m_3 are non-null, resulting in the same tree. However, with Conflict-EDT, m_2 is included in m_3 and the conflict, equation (21), between m_2 and m_3 is zero. A shallower tree with larger leaves is then created (see Figure 1a). To summarize, the other evidential models do not group x_2 and x_3 in the same node because their labels are different, while the proposed model groups them because the two responses are included in each other.

4.2. Aggregating evidential trees : The Evidential Random Forest

Evidential Decision Trees suffers from the same problems as classical decision trees such as the high variance (Breiman, 1996). We also propose here an Evidential Random Forest, based on the Evidential Decision Tree, taking advantage of both high performance on imperfectly labeled data and increased performance due to variance reduction.

Algorithm 1 Evidential Decision Tree

Require: A training set S and features X of attribute domain $D_{\mathcal{A}}$.

```
function EvidentialDecisionTree(S)
    T ← ∅
    if stopping criteria met then
        leaf ← Create node with S
        T ← Attach leaf
    else
        for 𝒜 in D_𝒜 do                                  ▷ Browse attributes
            Info_𝒜(S) ← Compute conflict info criterion
        end for
        Best_𝒜 ← Best attribute according to Info_𝒜(S)
        for v in Best_𝒜 do                               ▷ Browse values
            S_v ← Split S according to value v
            node ← EvidentialDecisionTree(S_v)
            T ← Attach node
        end for
    end if
    return T
end function
function PredictEDT(T, X)
    S_X ← Find the leaf of T with respect to the attributes of X
    m ← Combine with average all rich labels in S_X
    return m
end function
```

4.2.1. Evidential Random Forest

In this section both bagging and random feature selection are used on Evidential Decision Trees to model Evidential Random Forests.

The *Bootstrap* step in the proposed model involves richer labels: Let $\Omega = \{\omega_1, \ldots, \omega_M\}$ be a collection of M different classes and let $\mathcal{L} = \{x_k|1, \ldots, K\}$ be a learning set of K samples where each element is associated to a mass (*i.e.* the label) $m_k \in 2^\Omega$. Let the estimator $\varphi(x, \mathcal{L})$ be the Evidential Random Forest, and $\{\varphi(x, \mathcal{L}_n)\}$ the sequence of Evidential Decision Tree predictors. The $\{\mathcal{L}_n\}$ sets are the new bootstrapped learning sets composed of K elements drawn at random *with replacement*, in $\mathcal{L}$. The number N (with $n \in N$) of *bags* is discussed in section 5 and is set to 50 for experiments.

When *Aggregating* is performed, each estimator $\varphi(x, \mathcal{L}_n)$ predicts a mass m_n (*i.e.* the soft class predicted by Evidential Decision Trees) for an unlabeled observation. We propose to combine all the masses with the average:

$$m_{ERF}(A) = \frac{1}{N} \sum_{n=1}^{N} m_n(A), \quad A \in 2^\Omega, \tag{23}$$

with m_{ERF} the prediction of the Evidential Random Forest.

4.2.2. Decision making

The motivation is the representation of evidence as an output of the model, but a decision on the set of classes Ω can be made. The class ω_{ERF} maximizing the pignistic probability (3) of m_{ERF} is the predicted class:

$$\omega_{ERF} = \underset{\omega \in \Omega}{\operatorname{argmax}}(BetP(\omega,\ m_{ERF})), \tag{24}$$

with $BetP(\omega,\ m)$ the pignistic probability of ω according to m.

Example. Let $\Omega = \{\omega_1, \omega_2\}$ be the set of possible classes in a classification problem. Let $\varphi(x, \mathcal{L})$ be an Evidential Random Forest of $N = 3$, Evidential Decision Tree estimators $\varphi(x, \mathcal{L}_1)$, $\varphi(x, \mathcal{L}_2)$ and $\varphi(x, \mathcal{L}_3)$ respectively predicting m_1, m_2 and m_3 for a new unlabeled incoming sample such that:

m_1**:** $m_1(\omega_1) = 0.8$, $m_1(\Omega) = 0.2$,

m_2**:** $m_2(\omega_1) = 0.9$, $m_2(\Omega) = 0.1$,

m_3**:** $m_2(\omega_2) = 0.2$, $m_2(\Omega) = 0.8$.

The estimators $\varphi(x, \mathcal{L}_1)$ and $\varphi(x, \mathcal{L}_2)$ are strongly supporting class ω_1 while $\varphi(x, \mathcal{L}_3)$ supports very slightly ω_2. The prediction m_{ERF} for Evidential Random Forest is then:

m_{ERF}**:** $m_{ERF}(\omega_1) = 0.57$, $m_{ERF}(\omega_2) = 0.07$, $m_{ERF}(\Omega) = 0.36$.

On decision level, we compute the pignistic probabilities:

$BetP(\omega_1) = 0.75$,

$BetP(\omega_2) = 0.25$.

And ω_1 is chosen as the hard predicted label because it maximizes the pignistic probability on m_{ERF}.

Algorithm 2 Evidential Random Forest

Require: A training set $\mathcal{L}$, features X and N estimators in forest $\varphi(\mathcal{L})$.

```
function EvidentialRandomForest(L, X)
    φ(L) ← ∅
    for n in 1, ..., N do                          ▷ Create the forest
        L_n ← A bootstrapped sample of L
        φ(L_n) ← EvidentialDecisionTree(L_n, X_n)
        φ(L) ← φ(L) ∪ φ(L_n)
    end for
    return φ(L)
end function
function PredictERF(φ(L), X)
    for φ(L_n) in φ(L) do                          ▷ Browse the trees
        m_n ← PredictEDT(φ(L_n), X_n)
    end for
    m ← Combine m_{n∈N}                            ▷ With average
    return m
end function
```

5. Experiments

In this section, experiments are conducted with the proposed evidential decision trees and evidential random forests.

5.1. Design of the experiments

As the objective is to show the robustness of the models to imperfections, the experiments have been carried out on several datasets containing between 2 and 10 classes and with a number of observations ranging from 40 to 700. First, we used 10 well-known datasets available on the UCI Machine Learning Repository (Dua & Graff, 2017), to which noise has been applied.

Imprecision noise: An observation is chosen at random and the corresponding label loses one degree of precision, with another class chosen at random in Ω (*e.g.* If a source labeled an observation *Virginica*, the noisy label becomes either *Virginica or Setosa* or *Virginica or Versicolor*). A 50% noisy dataset would mean that half of the labels have lost a degree of precision. For non-evidential models, the hard label is the class maximizing the pignistic probability (see equation (3)).

We also performed tests on uncertain and imprecise labelled datasets. In a previous study, Thierry et al. (2021) used these data to validate Smets' hypothesis according to which the more imprecise humans are, the more certain they are (Smets, 1997). Such datasets, offering users the possibility of expressing their imprecision and confidence in their answers, are not common. We used Credal Bird-10, Credal Bird-2, Credal Dog-7, Credal Dog-4 and Credal Dog-2, the only ones, to our knowledge, that have actually been imperfectly labelled in crowdsourcing campaigns. They have a uniform class distribution and are published by Hoarau et al. (2023)[2]. Details from all datasets are presented in Table 1. Each experiment is performed 100 times to obtain an estimation of the actual mean accuracy of the model for each dataset. An iteration corresponds to a random draw of 20% of the dataset as a test set, the rest is used for training. Due to some dataset imbalance, the F1-score is also added as a comparison metric. The highest value of the F1-score is 1 and indicates a perfect precision and recall. The AUC (Area Under ROC Curve) is also used, it provides an overall measure of performance for all possible classification thresholds.

The experiments are separated in two parts, the first one focuses on Evidential Decision Tree and the second on Evidential Random Forest (composed of evidential trees).

[2]Link to the datasets *https://data.mendeley.com/datasets/4hz3wx6wm5*.

Table 1: Datasets description, with total number of observations, number of classes and number of explanatory variables (Features).

Dataset	Observations	Classes	Features
Breast cancer	569	2	30
Ionosphere	351	2	34
Post-operative	86	2	8
Sonar	208	2	60
Liver	345	2	6
Balance scale	625	3	4
Iris	150	3	4
Wine	178	3	13
Glass	214	6	9
Ecoli	336	8	7
Credal Dog-2	200	2	42
Credal Dog-4	400	4	47
Credal Dog-7	700	7	43
Credal Bird-2	40	2	17
Credal Bird-10	200	10	30

5.2. Experiments on Evidential Decision Trees

This section compares the proposed Evidential Decision Trees with Decision Trees using scikit-learn default parameters (Pedregosa et al., 2011) and other Evidential Decision Trees. The Uncertainty-EDT (Denoeux & Bjanger, 2000), the Euclidean-EDT (Elouedi et al., 2001) and the Jousselme-EDT (Jousselme et al., 2001) are used for comparison. The proposed model is noted Conflict-EDT. For experiments where this is not specified, the trees are grown to the maximum depth and are not pruned (because Random Forest uses fully grown trees).

5.2.1. Performance over several datasets

This part first focuses on the relevance of the model. Table 2 represents the performance of Decision Trees as well as Euclidean-EDT, Uncertainty-EDT, Jousselme-EDT and the proposed Conflict-EDT over several rich label datasets. The ten first datasets are noisy by imprecision at 50% (*i.e.* half of the labels have lost a degree of precision). Credal datasets did not need to be noisy as they inherently implement rich labels. Tables 3 and 4 propose two other comparison criteria, the F1 score and the Area Under the ROC Curve for the two-class datasets. The results show similar performance to the accuracy criterion, except for the post-operative dataset, where Conflict-

Table 2: Mean accuracy on 50% noisy dataset (± a 95% confidence interval for mean estimation). A Decision Tree (DT) and four Evidential Decision Trees (EDT) are used for comparison, the proposed model is the Conflict-EDT (Welch's t-test significance at p-value < 0.05 indicated by *****).

Dataset	DT	EDT			
		Euclidean	Uncertainty	Jousselme	Conflict
Breast cancer	70.1 ± 0.9	72.4 ± 0.8	71.4 ± 0.7	72.8 ± 0.9	**91.1*** ± 0.5
Ionosphere	67.6 ± 1.2	71.0 ± 0.9	68.4 ± 1.1	71.9 ± 1.0	**87.4*** ± 0.8
Post-operative	55.7 ± 2.4	**60.4** ± 2.4	56.7 ± 2.2	57.8 ± 2.4	59.9 ± 2.3
Sonar	61.1 ± 1.6	59.5 ± 1.5	60.2 ± 1.6	59.2 ± 1.4	**66.8*** ± 1.4
Liver	53.8 ± 1.2	55.4 ± 1.2	54.8 ± 1.1	56.0 ± 1.2	**58.0*** ± 1.1
Balance scale	59.4 ± 1.0	69.8 ± 0.8	57.5 ± 0.9	69.9 ± 0.7	**75.1*** ± 0.6
Iris	63.6 ± 2.0	74.0 ± 1.8	68.5 ± 1.7	73.9 ± 1.7	**90.4*** ± 1.2
Wine	70.4 ± 1.9	70.1 ± 1.4	68.3 ± 1.6	68.3 ± 1.7	**88.1*** ± 1.3
Glass	50.7 ± 1.5	51.4 ± 1.5	51.8 ± 1.5	53.1 ± 1.6	**60.5*** ± 1.5
Ecoli	56.8 ± 1.3	58.8 ± 1.1	57.1 ± 1.2	59.0 ± 1.2	**69.9*** ± 1.0
Credal Dog-2	82.9 ± 1.3	81.2 ± 1.4	82.4 ± 1.2	81.8 ± 1.3	**83.0** ± 1.2
Credal Dog-4	57.7 ± 1.3	58.0 ± 1.2	57.7 ± 1.0	58.2 ± 1.2	**59.2** ± 1.1
Credal Dog-7	50.1 ± 0.9	50.1 ± 0.9	48.8 ± 0.9	50.0 ± 0.8	**53.1*** ± 1.0
Credal Bird-2	50.8 ± 3.6	**62.8** ± 3.7	57.3 ± 3.7	59.8 ± 3.3	52.4 ± 3.4
Credal Bird-10	42.0 ± 1.6	43.1 ± 1.5	45.0 ± 1.7	42.7 ± 1.6	**45.4** ± 1.6

EDT scores poorly in terms of F1-score, due to the high class imbalance for this dataset.

The proposed model exibits better performance on both noisy and imperfectly labeled datasets. Conflict-EDT obtains better results, with sometimes important differences in performance, like on Breast cancer, Iris or Wine. The reason for this gap is explained in the following experiments.

5.2.2. Robustness to noise

This section focuses particularly on the robustness to imprecision and to the evolution of model performance with respect to noise increase. Figure 2 shows the results of the experiment on imprecision noise.

Iris (2a), Wine (2b), Glass Identification (2c), Balance Scale (2d), Ionosphere (2e) and Ecoli (2f) datasets are noised from 0% to 100% and mean accuracies are presented.

On noisy Iris, Balance Scale and Ionosphere datasets, distance based models Euclidean-EDT, Jousselme-EDT and Conflict-EDT have better performance than Decision Tree and Uncertainty-EDT. However, amongst these models using distances, the proposed Conflict-EDT is the most robust to

Table 3: Mean F1-score (± a 95% confidence interval) on 2-class datasets. A Decision Tree (DT) and four Evidential Decision Trees (EDT) are used for comparison, the proposed model is the Conflict-EDT (Welch's t-test significance indicated by *****).

Dataset	DT	EDT			
		Euclidean	Uncertainty	Jousselme	Conflict
Breast cancer	75.2 ± 0.9	77.3 ± 0.7	75.0 ± 0.8	77.6 ± 0.9	**92.9*** ± 0.4
Ionosphere	75.5 ± 1.1	76.7 ± 0.9	74.1 ± 1.0	77.5 ± 1.0	**90.3*** ± 0.6
Post-operative	27.8 ± 3.2	**30.5** ± 3.7	29.1 ± 3.8	29.0 ± 3.4	21.7 ± 3.6
Sonar	57.9 ± 2.0	57.5 ± 1.7	57.3 ± 1.8	56.8 ± 1.8	**63.5*** ± 1.8
Liver	58.7 ± 1.2	59.7 ± 1.2	59.0 ± 1.1	60.7 ± 1.3	**61.2** ± 1.3
Credal Dog-2	81.5 ± 1.4	80.4 ± 1.5	81.6 ± 1.3	80.8 ± 1.5	**82.2** ± 1.5
Credal Bird-2	53.1 ± 4.6	**61.3** ± 4.2	58.5 ± 4.5	57.8 ± 4.1	51.8 ± 4.4

Table 4: Mean AUC (Area Under the ROC Curve) on 2-class datasets. A Decision Tree (DT) and four Evidential Decision Trees (EDT) are used for comparison, the proposed model is the Conflict-EDT (Welch's t-test significance indicated by *****).

Dataset	DT	EDT			
		Euclidean	Uncertainty	Jousselme	Conflict
Breast cancer	0.69	0.83	0.81	0.83	**0.94***
Ionosphere	0.67	0.79	0.77	0.79	**0.90***
Post-operative	0.48	**0.53**	0.50	0.53	0.48
Sonar	0.59	0.64	0.65	0.65	**0.70***
Liver	0.55	0.58	0.57	0.59	**0.60**
Credal Dog-2	0.84	0.87	0.88	0.88	**0.92***
Credal Bird-2	0.52	**0.72**	0.68	0.70	0.58

noise with about 90% of good predictions on the half-noisy Iris dataset against about 75% of good predictions for the second best model. On the other three datasets Wine, Glass Identification and Ecoli, Euclidean-EDT and Jousselme-EDT, do not show better performances than models that do not use distances (DT and Uncertainty-EDT) but Conflict-EDT is still as efficient compared to all other presented models.

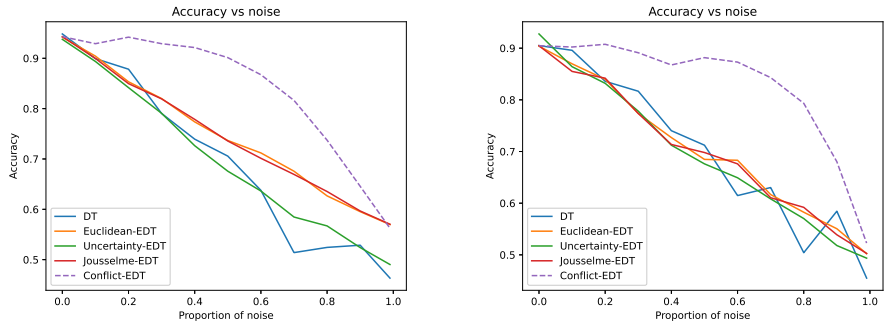
For imprecision noise, the proposed Conflict-EDT obtained better results on all presented datasets and at all noise levels. In fact, the model is not really more robust to imprecision than other Evidential Decision Trees, but to overfitting, due to equation (21). Another experiment showed that by pre-pruning all the trees, the performance of the other models increased and the performance gap with the proposed model disappeared. An example is shown in Figure 3, this is the same experiment as the previous one (2d) for the Balance Scale dataset but this time the models have been pruned.

The pruning used is a pre-pruning that prevents the tree from growing too deep during its learning phase. The number of observations is limited and the tree cannot create a leaf with less than 5 elements. Here, the Uncertainty-EDT, Euclidean-EDT, Jousselme-EDT and Conflict-EDT models are all robust to imprecision noise. The Conflict-EDT model no longer performs much better than the other models because its advantage is to limit overfitting. The pruning of the trees prevents all models from being overfitted, and therefore the difference in performance is no longer noticeable. Only the classical decision tree has very low performance, because data are labeled with imprecision, an information from which this non-evidential model cannot benefit.

In the absence of pruning, Conflict-EDT shows high robustness to overfitting when the data is imprecisely labeled.

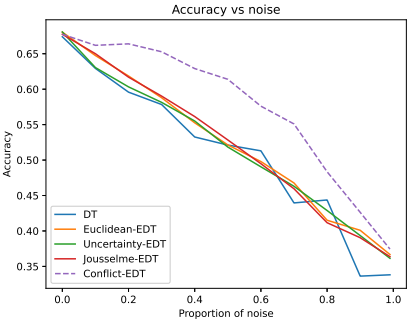
5.2.3. Tree growth

In the previous experiment, the robustness of the model to overfitting was deduced from its performance. In this experiment, two additional criteria are presented to demonstrate the benefits of the proposed method. With uncertainly and imprecisely labeled data, Evidential Decision Trees tend to overfit the training data and therefore to deliver worse performance on the test set. This results in large over-trained trees with small leaves. The *Depth* of the tree is the maximum number of divisions between the root of the tree and a leaf. The greater the depth, the more the tree is trained on the data. *Leaf size* is the average number of observations present in the leaves of the

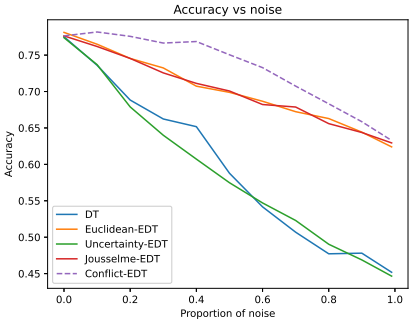


(a) Iris

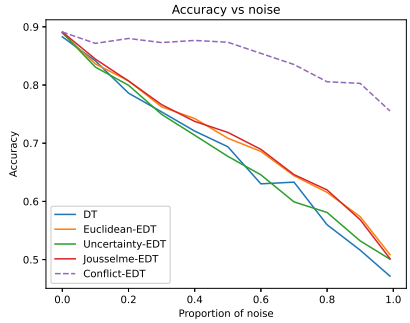
(b) Wine



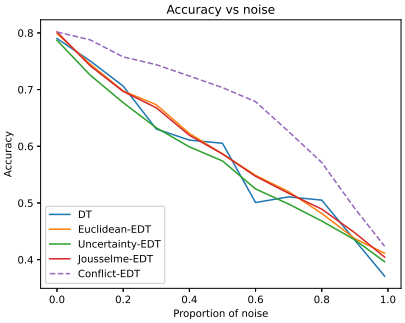
(c) Glass Identification



(d) Balance Scale



(e) Ionosphere



(f) Ecoli

Figure 2: Mean accuracy by amount of noise on several datasets for Decision Tree (DT) and Evidential Decision Trees (EDT).

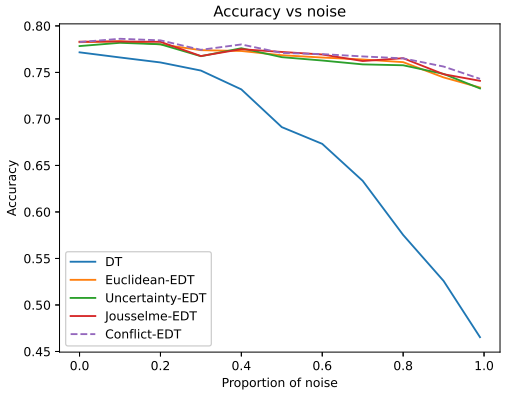


Figure 3: Mean accuracy by noise on pre-pruned models for Balance Scale dataset.

tree. The wider the leaves, the less the tree is overfit.

In Table 5, the average depths are presented for each Evidential Decision Tree presented in this paper. The ten first datasets are arbitrarily noisy at 30%, as studied in the previous experiment, any noise value can be used for Conflict-EDT to perform better. Imprecision noise is used, and the values presented are averaged over 50 experiments, rounded to the unit. Credal datasets have not been noised, because they have already been labeled in an uncertain and imprecise way by contributors. On the Iris and Wine datasets, Euclidean-EDT, Uncertainty-EDT and Jousselme-EDT have an average depth between 12 and 18 while Conflict-EDT has an average depth of 8. This difference is present on all datasets, and particularly notable on datasets with a large number of classes.

Table 6 represents the average size of tree leaves following the same specifications. On Wine and Iris, leaf size is larger for Conflict-EDT with an average of 7 to 9 versus 2 to 4 for the other Evidential Decision Trees. The trend is the same for other datasets, with an even larger gap in leaf size for 2-class datasets.

Overall, Conflict-EDT has grown less deeply and has larger leaves than the other Evidential Decision Trees. This robustness to overfitting on uncertain labels allows to obtain better generalization results.

5.3. Experiments on Evidential Random Forest

This section compares the proposed Evidential Random Forest to other models and shows its contribution compared to the Evidential Decision Tree.

Table 5: Trees mean depth rounded to the unit, on 30% noisy datasets (except for Credal datasets) for each Evidential Decision Tree, the proposed version is the Conflict.

Dataset	Euclidean	Uncertainty	Jousselme	Conflict
Breast cancer	19	33	19	**7**
Ionosphere	17	24	18	**10**
Post-operative	11	13	11	**9**
Sonar	12	17	12	**7**
Liver	18	18	17	**13**
Balance scale	13	14	13	**12**
Iris	13	14	13	**8**
Wine	12	18	13	**8**
Glass	21	16	20	**14**
Ecoli	26	19	26	**17**
Credal Dog-2	15	36	15	**9**
Credal Dog-4	25	41	23	**20**
Credal Dog-7	48	42	36	**24**
Credal Bird-2	12	18	11	**6**
Credal Bird-10	26	24	20	**14**

Table 6: Trees mean number of observations in leaves, rounded to the unit on 30% noisy datasets (except for Credal datasets) for each Evidential Decision Tree, the proposed version is the Conflict.

Dataset	Euclidean	Uncertainty	Jousselme	Conflict
Breast cancer	5	3	5	**36**
Ionosphere	5	2	5	**18**
Post-operative	3	2	3	**5**
Sonar	5	3	5	**12**
Liver	3	2	3	**6**
Balance scale	3	1	3	**5**
Iris	3	2	3	**7**
Wine	4	2	4	**9**
Glass	2	2	2	**3**
Ecoli	2	2	2	**3**
Credal Dog-2	2	1	2	**8**
Credal Dog-4	1	1	1	**2**
Credal Dog-7	1	1	1	**2**
Credal Bird-2	1	1	1	**5**
Credal Bird-10	1	1	1	**2**

Table 7: Performance gain of Evidential Random Forests compared to Evidential Decision Trees on noisy and imperfectly labeled datasets. The difference in accuracy between the two models is presented as well as the percentage gained.

Dataset	Accuracy gain	% Improvement
Breast cancer	3.5	3.8
Ionosphere	5.2	6.0
Post-operative	11.1	18.6
Sonar	10.0	15.0
Liver	8.5	14.7
Balance scale	9.5	12.7
Iris	5.3	5.8
Wine	9.5	10.8
Glass	14.5	23.9
Ecoli	15.6	22.3
Credal Dog-2	10.8	13.1
Credal Dog-4	18.0	30.4
Credal Dog-7	26.1	49.2
Credal Bird-2	0.2	0.5
Credal Bird-10	15.8	34.9

Also, Random Forests and Evidential Random Forests are trained with 50 estimators (*i.e.* 50 decision trees trained on bagged samples). Breiman (1996) shows that the performance of the model peaks around 25 estimators, and does not increase anymore at 50. We obtain similar results, with a stop of the performance increase beyond 30 decision trees. A random feature selection is made at each node and, when not specified, the parameters of the models used are those by default present in scikit-learn (Pedregosa et al., 2011).

5.3.1. Gain in performance over Evidential Decision Tree

Random forests are low variance models, defined to reduce the error of Decision Trees. This experiment proposes to restate the accuracy gain of the proposed Evidential Random Forest model compared to an Evidential Decision Tree, the model used in the forest creation process. This gain in performance is represented in Table 7, the flat value of accuracy as well as the percentage of performance gained are given. The means are estimated over 100 iterations and the noise used is the noise over imprecision (except for Credal datasets), with a noise rate of 50% of the dataset.

As expected, the model increases the performance of decision trees by aggregating the predictions of multiple estimators. On datasets with few classes (Iris, Wine, Balance, Breast cancer, Ionosphere) Evidential Random

Forest increases less, but still significantly, the performance of Evidential Decision Tree. When the number of classes increases (Glass, Ecoli, Credal Dog-7, Credal Bird-10) the performance gain becomes very important, with an augmentation of more than 20%. When the datasets are imperfectly labeled, without using noise, and on a large number of classes (Credal Bird-10 and Credal Dog-7) the performance gain is the most impressive (with respectively a 35% and 49% increase).

5.3.2. Comparison with Random Forest

The proposed model reduces the high variance of decision trees as theoretically expected. However, it remains to be shown whether the performance of Evidential Random Forest outperforms Random Forest. The average accuracy of the two models, as well as F1-scores and AUCs, are compared in this experiment to demonstrate the benefits of the proposal. Mean accuracies, as well as confidence intervals are presented in Table 8. Mean F1-scores and AUCs are respectively presented in Tables 9 and 10 on 2-class datasets.

For each dataset, whether noisy or really labeled in an uncertain and imprecise way, the performance of Random Forest is largely improved by the proposed evidential version. The only exception is on the post-operative dataset, where the F1-score is drastically lower, the cause being the imbalance of classes for very few observations. The increase is up to 10 accuracy points for Iris and Balance Scale datasets. When the imperfection linked to human labeling can be represented in the data, a gain in accuracy is also noticed.

5.3.3. Other cautious and evidence-based models

In this section, a recent Cautious Random Forest (Zhang et al., 2023) as well as two other evidential models, the Evidential K-Nearest Neighbors (Denœux, 1995) and a state-of-the-art Eviential SVM (Kadir et al., 2019) are used for comparison.

Cautious Random Forest uses the imprecise Dirichlet model at the estimator prediction level and the theory of belief functions (Dempster, 1967; Shafer, 1976) for aggregation. This means that this model has some similarities with our proposed Evidential Random Forest. It is allowed to make *cautious* predictions by combining probability intervals. However, the model is only defined for datasets with two classes. Each decision tree in the forest produces an interval-valued probability estimate for the positive class, thanks to the imprecise Dirichlet model. The Cautious Random Forest then

Table 8: Mean accuracy for Random Forest (RF) and the proposed Evidential Random Forest (ERF) on 50% noisy datasets (± a 95% confidence interval for mean estimation and Welch's t-test significance at p-value < 0.05 indicated by *****).

Dataset	RF	ERF
Breast cancer	90.5 ± 0.5	**94.5*** ± 0.4
Ionosphere	84.4 ± 1.0	**92.6*** ± 0.6
Post-operative	60.5 ± 2.3	**71.0*** ± 2.0
Sonar	72.0 ± 1.3	**76.8*** ± 1.2
Liver	58.2 ± 1.2	**66.5*** ± 0.9
Balance scale	75.1 ± 0.6	**84.5*** ± 0.5
Iris	84.4 ± 1.3	**95.3*** ± 0.7
Wine	91.5 ± 1.0	**97.5*** ± 0.5
Glass	68.2 ± 1.5	**75.1*** ± 1.3
Ecoli	77.6 ± 0.9	**85.5*** ± 0.8
Credal Dog-2	91.4 ± 1.0	**93.8*** ± 0.9
Credal Dog-4	72.3 ± 1.0	**77.1*** ± 0.9
Credal Dog-7	77.4 ± 0.7	**79.1*** ± 0.8
Credal Bird-2	45.0 ± 3.2	**52.6*** ± 3.2
Credal Bird-10	52.8 ± 1.4	**61.2*** ± 1.5

Table 9: Mean F1-score for Random Forest (RF) and the proposed Evidential Random Forest (ERF) on 2-class datasets (± a 95% confidence interval for mean estimation and Welch's t-test significance at p-value < 0.05 indicated by *****).

Dataset	RF	ERF
Breast cancer	93.3 ± 0.4	**95.6*** ± 0.3
Ionosphere	88.0 ± 0.9	**94.2*** ± 0.4
Post-operative	**27.2*** ± 3.4	6.6 ± 2.6
Sonar	64.9 ± 1.6	**74.4*** ± 1.6
Liver	62.5 ± 1.3	**70.4*** ± 1.0
Credal Dog-2	91.3 ± 1.3	**92.8** ± 1.0
Credal Bird-2	43.0 ± 4.5	**51.5*** ± 4.0

Table 10: Mean AUC (Area Under ROC Curve) for Random Forest (RF) and the proposed Evidential Random Forest (ERF) on 2-class datasets (± a 95% confidence interval for mean estimation and Welch's t-test significance at p-value < 0.05 indicated by *****).

Dataset	RF	ERF
Breast cancer	0.96	**0.99***
Ionosphere	0.93	**0.97***
Post-operative	**0.51**	0.49
Sonar	0.76	**0.87***
Liver	0.65	**0.71***
Credal Dog-2	0.97	**0.98***
Credal Bird-2	0.48	**0.58***

predict the class of unlabeled observations either precisely (*i.e.* a prediction on ω_1 or ω_2) or imprecisely (*i.e.* a prediction on $\omega_1 \cup \omega_2$) on $\Omega = \{\omega_1, \omega_2\}$. We computed accuracy for Cautious Random Forests classically for precise predictions (*i.e.* the proportion of correct predictions), and took the most plausible answer according to their defined *plausibility* score when the model gives a *cautious* prediction. To adapt to the model's ability to give a *cautious* answer the authors propose to use the u_{65} criterion (Zaffalon et al., 2012) which rewards an imprecise prediction on $\omega_1 \cup \omega_2$ by 0.65 (instead of 1 for a precise prediction).

Evidential K-Nearest Neighbors is a version of the K-Nearest Neighbors that is a standard in literature to both handle rich labels and produce soft predictions with the theory of belief functions. The best K number of neighbors is estimated according to a 5-fold cross-validation and the parameters used are those defined for γ-EKNN by Hoarau et al. (2022).

Evidential SVM[3] (Kadir et al., 2019) is a recent evidential classification model that uses both classical SVMs, K-means clustering and belief function theory to handle model uncertainty and imprecision in the form of a mass function. All parameters used for this model are those proposed by the authors in their publication. This version is also restricted to binary classification problems.

Table 11 presents the mean accuracies for Cautious Random Forest, Evidential SVM, Evidential K-Nearest Neighbors and the proposed Evidential Random Forest. The u_{65} score is also present for Cautious Random Forest to reward cautious predictions. Note that the two first models can only handle 2-class datasets, hence the absence of results in the table for datasets with more than 2 classes. As for the other experiments, the first 10 datasets are noisy and the Credal datasets have inherently rich labels. F1-scores and AUCs are also respectively present in Tables 12 and 13.

The ability for Cautious Random Forest to provide *cautious* predictions does not compensate for the drop in performance induced by label richness. The Evidential SVM does not perform as well as the other two evidential models, except for the post-operative and credal bird-2 datasets, in terms of F1-score. This is due to its inability to take into account all the uncertainty present in the labels. The Evidential K-Nearest Neighbors and the proposed Evidential Random Forest can both benefit from the rich labels

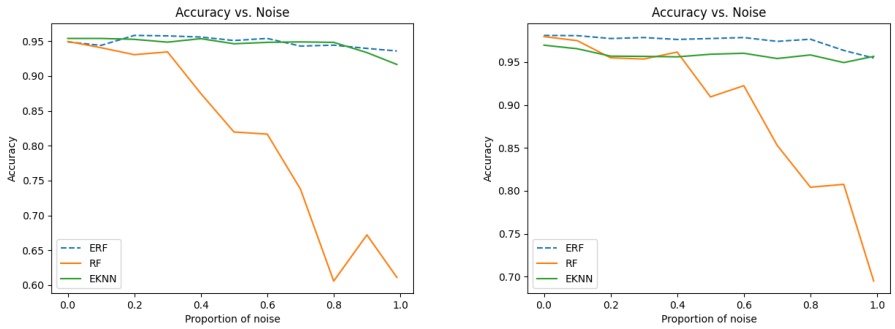
[3]SVM stands for Support Vector Machine.

Table 11: Mean accuracy for Cautious Random Forest (CRF), Evidential Support Vector Machine (ESVM), Evidential K-Nearest Neighbors (EK-NN) and the proposed Evidential Random Forest (ERF) on 50% noisy datasets ($\pm$ a 95% confidence interval for the estimate of the mean. The u_{65} score is also present for CRF to reward cautious predictions and Welch's t-test significance at p-value < 0.05 is indicated by *****).

Dataset	CRF		ESVM	EK-NN	ERF
	Acc	u_{65}	*Acc*	*Acc*	*Acc*
Breast cancer	91.2 ± 0.6	91.7 ± 0.5	91.9 ± 0.5	**95.4*** ± 0.4	94.5 ± 0.4
Ionosphere	85.6 ± 1.0	84.3 ± 0.8	90.2 ± 0.8	83.2 ± 0.9	**92.4*** ± 0.5
Post-operative	59.8 ± 2.3	59.2 ± 1.9	58.6 ± 2.3	65.7 ± 2.0	**71.2*** ± 1.9
Sonar	68.3 ± 1.5	74.4 ± 1.2	65.3 ± 1.4	**77.3** ± 1.3	76.5 ± 1.3
Liver	60.3 ± 1.1	61.2 ± 1.0	55.4 ± 1.0	58.2 ± 1.0	**66.1*** ± 0.9
Balance scale				**88.2*** ± 0.5	84.5 ± 0.5
Iris				94.6 ± 0.8	**95.3** ± 0.7
Wine				95.9 ± 0.7	**97.5*** ± 0.5
Glass				64.2 ± 1.5	**75.1*** ± 1.3
Ecoli				84.8 ± 0.8	**85.5** ± 0.8
Credal Dog-2	90.9 ± 1.1	92.8 ± 0.8	63.4 ± 1.4	73.8 ± 1.5	**93.3** ± 0.9
Credal Dog-4				69.3 ± 1.0	**77.1*** ± 0.9
Credal Dog-7				75.8 ± 0.7	**79.1*** ± 0.8
Credal Bird-2	48.3 ± 3.6	47.3 ± 3.1	53.8 ± 3.4	**58.4*** ± 3.4	53.1 ± 3.2
Credal Bird-10				60.6 ± 1.5	**61.2** ± 1.5

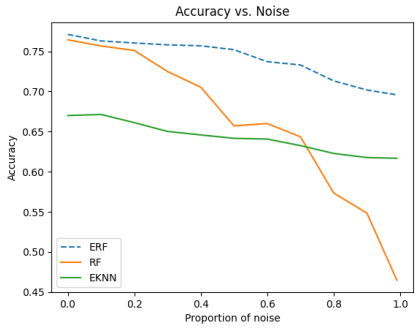
Table 12: Mean F1-score for Cautious Random Forest (CRF), Evidential Support Vector Machine (ESVM), Evidential K-Nearest Neighbors (EK-NN) and the proposed Evidential Random Forest (ERF) on 2-class datasets ($\pm$ a 95% confidence interval for the estimate of the mean. Welch's t-test significance at p-value < 0.05 is indicated by *****).

Dataset	CRF	ESVM	EK-NN	ERF
Breast cancer	93.0 ± 0.5	93.6 ± 0.4	**96.4*** ± 0.3	95.6 ± 0.3
Ionosphere	88.2 ± 0.9	92.0 ± 0.7	88.1 ± 0.7	**94.2*** ± 0.4
Post-operative	**27.5** ± 3.5	27.3 ± 3.5	15.1 ± 3.5	6.6 ± 2.6
Sonar	65.2 ± 1.8	50.0 ± 2.1	72.9 ± 1.6	**74.4** ± 1.6
Liver	61.9 ± 1.4	57.1 ± 1.3	62.3 ± 1.1	**70.4*** ± 1.0
Credal Dog-2	90.0 ± 1.2	44.3 ± 2.5	66.0 ± 2.2	**92.8** ± 1.0
Credal Bird-2	42.5 ± 4.9	**54.2** ± 4.6	41.4 ± 5.1	51.5 ± 4.0

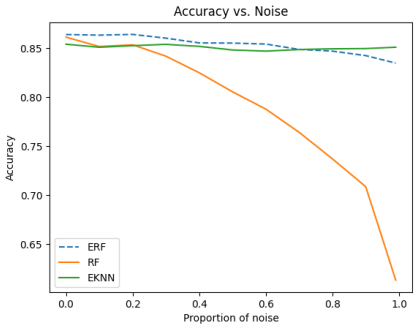


(a) Iris

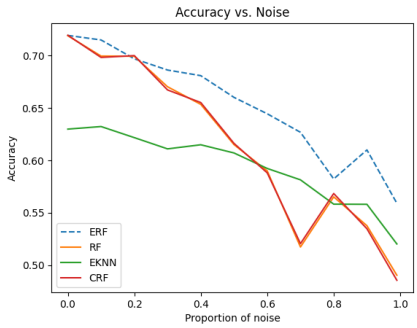
(b) Wine



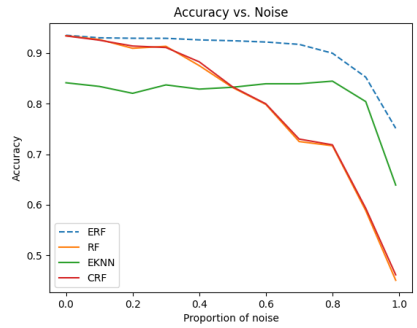
(c) Glass Identification



(d) Ecoli



(e) Liver



(f) Ionosphere

Figure 4: Mean accuracy by amount of noise on several datasets for Random Forest (RF), Cautious Random Forest (CRF), Evidential K-Nearest Neighbors (EKNN) and the proposed Evidential Random Forest (ERF).

Table 13: Mean AUC (Area Under ROC Curve) for Cautious Random Forest (CRF), Evidential Support Vector Machine (ESVM), Evidential K-Nearest Neighbors (EK-NN) and Evidential Random Forest (ERF) on 2-class datasets ($\pm$ a 95% confidence interval for the estimate of the mean. Welch's t-test significance at p-value < 0.05 is indicated by *****).

Dataset	CRF	ESVM	EK-NN	ERF
Breast cancer	0.90	0.94	0.98	**0.99***
Ionosphere	0.86	0.89	0.95	**0.97***
Post-operative	0.51	**0.53**	**0.53**	0.49
Sonar	0.68	0.70	0.86	**0.87**
Liver	0.61	0.58	0.60	**0.71***
Credal Dog-2	0.91	0.75	0.88	**0.98***
Credal Bird-2	0.50	0.55	**0.74***	0.58

and are therefore more competitive. Overall on the 10 noisy datasets, Evidential Random Forest performs slightly better with better performance on 6 datasets. On the credal datasets, the proposed model also performs better.

Figure 4 shows the robustness of the models to imprecision noise. Random Forest always seems to start with a high accuracy score when there is no noise and drops in performance very quickly as the noise increases. Cautious Random Forest, when evaluated on accuracy (and not u_{65}) gives results very close to Random Forest. Evidential models (Evidential K-Nearest Neighbors and Evidential Random Forest) evolve in the same way, they start from different accuracies on perfect data depending on their ease in generalizing the dataset, but both have performances that decrease very slowly with the addition of noise. Overall, on the presented datasets, the proposed Evidential Random Forests performs consistently better.

6. Discussion

If the proposed Evidential Random Forest shows good performance according to accuracy, F1-score, AUC and robustness to imprecision, the underlying motivation is not only performance but also representation of evidence. We proposed models[4] that are both able to take into account rich labels and to produce an evidential prediction modeled by the theory of belief functions. Such a soft prediction is shown in Table 14. This information can for example be used in uncertainty sampling (Hoarau et al., 2022), to represent both the

[4]The proposed models are Evdiential Decision Trees and Evidential Random Forests.

Table 14: Two observations respectively from Credal Dog-2 and Credal Bird-2 with possible model prediction. The classes respectively are $\{\omega_1$: Brittany, ω_2: Beagle$\}$ and $\{\omega_1$: Western Jackdaw, ω_2: Rook$\}$, for both observations the true class is ω_1.

Observation	Prediction				
	RF	CRF	ESVM	EK-NN	ERF
	ω_2	$\{\omega_1, \omega_2\}$	$\omega_1 = 0.4$ $\omega_2 = 0.3$ $\{\omega_1, \omega_2\} = 0.3$	$\omega_1 = 0.2$ $\omega_2 = 0.7$ $\{\omega_1, \omega_2\} = 0.1$	$\omega_1 = 0.6$ $\omega_2 = 0.2$ $\{\omega_1, \omega_2\} = 0.2$
	ω_2	$\{\omega_1, \omega_2\}$	$\omega_1 = 0.5$ $\omega_2 = 0.2$ $\{\omega_1, \omega_2\} = 0.3$	$\omega_1 = 0.3$ $\omega_2 = 0.5$ $\{\omega_1, \omega_2\} = 0.2$	$\omega_1 = 0.7$ $\omega_2 = 0.1$ $\{\omega_1, \omega_2\} = 0.2$

uncertainty of the model predictions and the uncertainty in the labels. Our current work focuses on the reduction of labeling costs (*i.e.* active learning) thanks to these evidential models in uncertainty sampling. The ability of the model to represent its uncertainty in a complete way makes it possible to target better observations to label. In addition, we are working to separate the uncertainty of the model to address the exploitation-exploration problem in active learning with rich labels.

For Evidential Decision Trees, some criteria have been studied such as the Euclidean distance, the Jousselme distance or the more robust conflict criterion. However some authors (Zhang et al., 2022) have shown an inconsistency in the use of distances with the theory of belief functions. Perhaps the notion of distance is therefore not the most appropriate criterion for the problem at hand. The use of another criterion thus remains an open track to propose new Evidential Random Forests.

The ability of the model to take into account uncertain and imprecise data is at the cost of complexity. Even if the proposed model can run on a machine without a GPU, we have not managed to reach a level of optimization close to that of scikit-learn for example. Note that the python code of the two proposed models is available to the scientific community.

7. Conclusion

In this paper, we propose an overfitting robust Evidential Decision Tree based on a conflict measure. By using a distance between masses and a degree of inclusion, this criterion allows to group in the same node, observations with elements of response included in each other. By doing so, the tree built is shallower and less over-trained. We also propose an Evidential Random Forest which allows to overcome the high variance of decision trees. The model has been compared with other recent models and in particular with the reference in the evidential models literature. The results show high performance both on noisy data and on datasets that are actually labeled in an uncertain and imprecise manner.

The underlying motivation, other than robustness and performance, is the representation of evidence. The goal is to couple evidential learning with active learning, in order to work with imperfect data but also with as few labeled observations as possible.

Python code for Evidential Decision Trees is available on Github at: https://github.com/ArthurHoa/conflict-edt.

Python code for Evidential Random Forests is available on Github at: https://github.com/ArthurHoa/evidential-random-forest.

Funding: This work was supported by the Brittany region and the Côtes-d'Armor department.

Special thanks to Vincent Lemaire from Orange Labs for proofreading this article.

References

Amit, Y., & Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural Computation*, *9*, 1545–1588.

Bramer, M. (2013). Avoiding overfitting of decision trees. In *Principles of Data Mining* (pp. 121–136). Springer London.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Breiman, L., Friedman, J., Stone, C., & Olshen, R. (1984). *Classification and Regression Trees*. Taylor & Francis.

Dempster, A. P. (1967). Upper and Lower Probabilities Induced by a Multivalued Mapping. *The Annals of Mathematical Statistics*, *38*, 325 – 339.

Denoeux, T., & Bjanger, M. (2000). Induction of decision trees from partially classified data using belief functions. *Systems, Man, and Cybernetics, 2000 IEEE International Conference*, *4*, 2923–2928.

Denoeux, T., Kanjanatarakul, O., & Sriboonchitta, S. (2019). A New Evidential K-Nearest Neighbor Rule based on Contextual Discounting with Partially Supervised learning. *International Journal of Approximate Reasoning*, *113*, 287–302.

Denœux, T. (1995). A k-nearest neighbor classification rule based on dempster-shafer theory. *Systems, Man and Cybernetics, IEEE Transactions on*, *219*.

Dua, D., & Graff, C. (2017). UCI machine learning repository.

Elouedi, Z., Mellouli, K., & Smets, P. (2001). Belief decision trees: theoretical foundations. *International Journal of Approximate Reasoning*, *28*, 91–124.

Fredriksson, T., Mattos, D. I., Bosch, J., & Olsson, H. H. (2020). Data labeling: An empirical investigation into industrial challenges and mitigation strategies. In M. Morisio, M. Torchiano, & A. Jedlitschka (Eds.), *Product-Focused Software Process Improvement* (pp. 202–216). Cham: Springer International Publishing.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer.

Hoarau, A., Martin, A., Dubois, J.-C., & Le Gall, Y. (2022). Imperfect labels with belief functions for active learning. In *Belief Functions: Theory and Applications*. Springer International Publishing.

Hoarau, A., Thierry, C., Martin, A., Dubois, J.-C., & Le Gall, Y. (2023). Datasets with rich labels for machine learning. In *FUZZ-IEEE*.

Jousselme, A.-L., Grenier, D., & Bossé, E. (2001). A new distance between two bodies of evidence. *Information Fusion*, *2*, 91–101.

Kadir, M. E., Saha Akash, P., Ali, A. A., Shoyaib, M., & Begum, Z. (2019). Evidential svm for binary classification. In *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)* (pp. 1–6).

Klir, G. J., & Wierman, M. J. (1998). Uncertainty-based information: Elements of generalized information theory. In *Springer-Verlag*.

Martin, A. (2019). Conflict management in information fusion with belief functions. In E. Bossé, & G. L. Rogova (Eds.), *Information quality in information fusion and decision making* Information Fusion and Data Science (pp. 79–97).

Moral-García, S., Mantas, C. J., Castellano, J. G., Benítez, M. D., & Abellán, J. (2020). Bagging of credal decision trees for imprecise classification. *Expert Syst. Appl.*, *141*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Quinlan, J. (1987). Simplifying decision trees. *International Journal of Man-Machine Studies*, *27*, 221–234.

Quinlan, J. R. (1993). C4.5: Programs for machine learning. *Morgan Kaufmann, San Mateo*, .

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton: Princeton University Press.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423.

Siciliano, R. (1998). Exploratory versus decision trees. In R. Payne, & P. Green (Eds.), *COMPSTAT* (pp. 113–124). Heidelberg: Physica-Verlag HD.

Smets, P. (1997). Imperfect information: Imprecision and uncertainty. In A. Motro, & P. Smets (Eds.), *Uncertainty Management in Information Systems: From Needs to Solutions* (pp. 225–254). Boston, MA: Springer US.

Sutton Charani, N., Destercke, S., & Denoeux, T. (2013). Learning decision trees from uncertain data with an evidential EM approach. In *12th International Conference on Machine Learning and Applications (ICMLA 2013)* (pp. 1–6). Miami, United States.

Thierry, C., Martin, A., Dubois, J.-C., & Le Gall, Y. (2021). Validation of Smets' hypothesis in the crowdsourcing environment. In *6th International Conference on Belief Functions*. Shanghai, China.

Trabelsi, A., Elouedi, Z., & Lefevre, E. (2019). Decision tree classifiers for evidential attribute values and class labels. *Fuzzy Sets and Systems*, *366*, 46–62.

Xu, P., Davoine, F., Zha, H., & Denoeux, T. (2016). Evidential calibration of binary svm classifiers. *International Journal of Approximate Reasoning*, *72*, 55–70.

Yuan, B., Yue, X., Lv, Y., & Denoeux, T. (2020). Evidential Deep Neural Networks for Uncertain Data Classification. In *Knowledge Science, Engineering and Management (Proceedings of KSEM 2020)* Lecture Notes in Computer Science. Springer Verlag.

Zaffalon, M., Corani, G., & Mauá, D. (2012). Evaluating credal classifiers by utility-discounted predictive accuracy. *International Journal of Approximate Reasoning*, *53*, 1282–1301. Imprecise Probability: Theories and Applications (ISIPTA'11).

Zhang, H., Quost, B., & Masson, M.-H. (2023). Cautious weighted random forests. *Expert Systems with Applications*, *213*.

Zhang, Y., Destercke, S., Zhang, Z., Bouadi, T., & Martin, A. (2022). On computing evidential centroid through conjunctive combination: An impossibility theorem. *IEEE Transactions on Artificial Intelligence*, *PP*, 1–10.