Uncertainty Estimation in Medical Machine Learning : Application to Pancreatic Cancer Diagnosis

Amal Chaoui, Slim Hammadi

June 2021

Abstract - Pancreatic Cancer has a substantial potential of malignity and early detection of this type of disease is challenging. Its survival rates have not changed significantly in nearly 40 years. The lack of standardised international guidelines in assessing suspicious pancreatic masses is the main cause behind its high mortality rate. Artificial Intelligence (AI) technologies have emerged to provide better diagnosis and treatment of diseases based on patients clinical, personal, and medical information. However, Uncertainty Quantification (UQ) is usually neglected when dealing with Machine Learning predictions. In this paper, we present two different approaches to uncertainty estimation. The first one uses a recent published work based on Bayesian Ensembles. The second one is based upon the Dempster Shafer theory of evidence. The obtained results are then compared in terms of uncertainty precision as well as the way of combining evidence from different sources and suggest that the second method gives a way of combining evidence based on their reliability degree and provides better uncertainty handling with respect to every patient individually. This way, physicians can better take advantage from ML models predictions for patients diagnosis and treatment decision making.

1 Introduction

Pancreatic Cancer (PC) remains one of the leading causes of cancer-related death in the world, in both men and women. Early detection of PC is difficult due to the lateness of PC-related symptoms to advanced stages of cancer in most patients (Klein et al., 2013 [1]). In France, the incidence rate of PC increased at an average rate of 2.7% per year in men and more steadily in women by 3.8% from 1990 to 2018 [2]. Furthermore, identifying patients at high risk for PC or

with early-stage cancer is challenging due to the lack of reliable screening tools and sensitive biomarkers (Yu et al., 2006 [3]).

Detection of pancreatic lesions is a key point, and identifying lesions at risk of degeneration from benign to malignant cysts is more important, since it provides subsequent management of patients' health status. However, surgical resection is the only treatment for PC. Unfortunately, due to late presentation, only low rates (15% to 20%) of patients are candidates for pancreatectomy [5] Furthermore, the majority of diagnosis imaging fail to accurately capture lesions <3cm.

Therefore, there is an enhanced need for the use of Machine Learning to improve patient care and support physicians in decision making for diagnosis and treatment. However, the estimation of uncertainty regarding the prediction is usually neglected. Noise in observations and incomplete coverage of the problem domain, known as aleatoric uncertainty, is inevitable and Machine Learning models will always introduce some error (epistemic uncertainty). Uncertainty quantification is still a new field of research that will help physicians abstain from considering the predicted results and make more principled decisions. Different approaches to uncertainty have since then emerged for potential use in healthcare settings (Alizadehsani et al., 2021 [4]).

In this paper, we examine two methods for uncertainty estimation in medical Machine Learning, applied to PC lesions predictions based on patients personal and medical data. The first method has recently been introduced by Malinin et al. [6], and can be used in many practical, high-risk applications. This method assesses both aleatoric and epistemic uncertainty based on a probabilistic ensembles-based approach that deduces uncertainty estimates from the predictions gradient boosting classification. However, the Bayesian approach suffers certain limits, especially when handling different sources of information. The second method uses Dempster Shafer's theory of evidence to settle this problem and provides different metrics that measure conflict of evidence. The two methods were tested on medical data of patients and compared to each other. Both aleatoric and epistemic uncertainty estimates are assessed based on predictions provided by Machine Learning models (known as predictive uncertainty). This provides physicians with a way to approach predictions of Machine Learning models in terms of uncertainty quantifying to support their decision based on the past history of decisions they, or other physicians have made. These two approaches can be used widely in other areas other than medical healthcare.

This paper is organised as follows: section 2 provides a medical background on pancreatic cystic lesions, section 3 presents related work pertaining to Machine Learning models as well as uncertainty estimation approaches used in medicine. Section 4 analyses two methods of uncertainty estimation; the bayesian ensemble method and a novel method developed using Dempster Shafer's theory,

section 5 presents the results of uncertainty estimates using the two approaches and finally section 6 presents conclusions and future work.

2 Medical Background

The pancreas is an organ of the digestive and endocrine system that has two main functions: an exocrine function that produces enzymes that are important to digestion. The endocrine function consists of releasing important hormones (insulin and glucagon) into the bloodstream to monitor blood sugar level [7].

Pancreatic cysts are saclike pockets of fluid on or in the pancreas [8]. The cysts are not often cancerous or do not cause symptoms. However, some pancreatic cysts, known as borderline cysts, can become cancerous. In both cases, patients' health status requires evaluation and monitoring. Physicians rely upon tests such as endoscopic procedures, CT scans, MRI, laboratory tests and biopsies.

Imaging is not really accurate in the early stages of PC. The techniques involved require significant advance, both in screening and technology application [9]. Given these difficulties, patients' lives are at stake. Taking the wrong diagnostic and treatment decision is risky. In this regard, considering Machine Learning as a means of support to physicians can be of a huge assistance in order to address these challenges.

3 Related Work

3.1 Machine Learning methods used in medical decision making

There have been many Machine Learning models that have been used in cancer research for the development of predictive models. Apart from being able to effectively identify pancreatic cysts, this enables physicians to classify patients into high and low risk groups and therefore facilitating clinical management of patients.

A variety of techniques has been widely used in cancer detection applications. Many works considered using both clinical and genomic data. Kourou et al. [10] reviewed recent and commonplace ML methods for predictive modeling of cancer progression: Artificial Neural Networks (ANNs), Bayesian Networks (BNs), Support Vector Machines(SVMs) and Decision Trees (DTs).

ANNs are often used for handling pattern recognition and classification problems [10]. Their power lies in the hidden layers performing neural connections through mathematical operations between the layers. However, there are certain drawbacks to using ANNs. The learning process proves to be lengthy. This is critical in medical decision making since it might put patients' lives at risk. They are also quite sensitive to overfitting. Another disadvantage to it is that the method is often referred to as " a black-box technology". This implies that ANNs' decisions are not always understandable and finding out how classification has been performed is challenging. This makes room for introducing a large amount of uncertainty.

DTs are predominantly used for classification problems through a tree-structured scheme. They have long been used for the purpose of decision analysis in different areas. However, decisions trees are considered as weak learners and often needs to be used along other ML techniques or through Ensemble Learning (e.g. Gradient Descent algorithms).

SVMs have recently been adopted for cancer diagnosis [10]. They are used for classification, regression and outlier detection. They identify a hyperplane that separates data points into different classes, in an attempt to data clustering. They can be used for tumor classification. The generalizability of the resulting classifier enables for reliable classification of new data samples.

BNs are probabilistic graphical models that use a directed acyclic graph to represent variables with probabilistic dependencies . They have been used for reasoning and knowledge representation.

Most importantly, Gradient Boosting method, known for achieving several winning solutions in data science competitions, has been vastly used for its ability to capture not only linear but also nonlinear relations in complex data with heterogeneous features (Dorogush et al., 2018 [11]). It has been used by Taninaga et al. [12] to classify patients with gastric cancer [13] one of the most common cancers in the world with high incidences, into high and low risk patients. It can perform both classification and regression tasks. For the example of gastric cancer, Gradient Boosting outperformed Logistic Regression, and the output performance increased when adding more input features.

For cancer prediction, ML models are often used for diagnosis and prognosis: cancer susceptibility, recurrence or survival. The accuracy of cancer predictions has remarkably improved by up to 20% in the last few years [9]. Nevertheless, careful features selection and dataset quality are of huge importance for accurate prediction outcome. As for classification problems, for instance, small training samples can result in misclassifications and biased models.

Although several ML models proved efficient in predicting outcomes, hardly few of them penetrated the clinical practice [9]. Physicians usually abstain from

considering predictions due to the lack of knowledge pertaining to uncertainty estimates. Consequently, it is essential to include a way of introducing uncertainty in ML predictions to support physicians in decision making. In this regard, the next section will expose the most prominent approaches to uncertainty quantification in medical machine learning from a recent published work.

Research Paper	Addressed Problem	Methods	Results	Conclusion and Critics
Machine learning applications in cancer prognosis and prediction Prediction of cancer susceptibility, recurrence, and survival (Kourou K. et al. Computational and Structural Biotechnology Journal, 2015)	Predicting cancer susceptibility (Risk estimation for breast cancer in women)	Model used: ANNs -Discrimination of benign and malignant cysts -Use of a large number of hidden layers - Estimated performance using 10-fold cross-validation - Use of the ES (Evolution Strategies) approach to control overlearning	- AUC-ROC of the model is 0.965 - A unique model in that it incorporates mammographic data	- ANNs and SVMs are widely used for their accurate predictive performance Limited volume of data ⇒ difficulty in generalizing results to other samples - Better statistical analysis of heterogeneous data would provide more accurate results
	Predicting cancer recurrence (Case of oral squamous cell carcinoma OSCC)	Several ML methods were used for patient classification	The BN classifier (the Bayesian network) gives the best performance (82.8% accuracy)	
	Predicting cancer survival (Evaluation of survival in women with breast cancer)	3 ML models were considered: ANN, SVM, SSL (Semi-Supervised Learning)	Accuracy of the 3 models ANN, SVM, SSL: 65%, 51% and 71% respect.	
Prediction of future gastric cancer risk using a machine learning algorithm and comprehensive medical check-up data (Taninaga J. et al. *Nature, 2019*)	Stomach cancer is a cancer with a high incidence and mortality rate → Need to classify patients according to their risk of developing cancer	- Classification of patients into two classes: low and high risk patients - Method used: XGBoost	- Best performance is obtained with a model that incorporates as many features as possible (AUC-ROC value of 0.899, sensitivity of 0.933) - XGBoost produces better results than linear regression methods	- Risk factors alone are not enough to produce a powerful classifier. - It is necessary to integrate all the features relevant to the type of prediction

Figure 1: Comparison of techniques used for medical Machine Learning in the scientific literature ([10], [12])

3.2 Uncertainty Estimation in Healthcare

Uncertainty Quantification (UQ) is the science of quantitative characterization and reduction of uncertainties in both computational and real world applications [15]. It is about determining how likely a certain outcome can be produced knowing input features.

Uncertainty Quantification carries out a major role in imprecision and ignorance reduction, in particular when using ML models to predict sensitive outcome, which is the case of healthcare applications. Models providing uncertainty relative information permits better and robust decision making.

In research work, uncertainty is usually classified into two categories : aleatoric and epistemic uncertainty. One way to distinguish the two is to ask whether or not uncertainty is reducible.

Aleatoric uncertainty refers to the irreducible part of uncertainty. This is usually due to the intrinsic noise in data qualified as inevitable, since it is related to the non-deterministic nature of input/output stochastic dependency.

On the other side, epistemic uncertainty (sometimes referred to as model or knowledge uncertainty) is due to the lack of information or knowledge about the perfect model which can represent accurately the relationship between input and output data. It can pertain, for example, to uncertainty about estimated model parameters. In essence, epistemic uncertainty can be reduced by adding more knowledge to the model.

A recent paper was published by Alizadehsani et al. [4] reviewing different most common methods for uncertainty handling in medical data.

Bayesian Inference is the most widely used method in terms of uncertainty measurement in ML models. It is based on the principle that we can update the probability of a hypothesis, whenever further information is available, through Bayes' theorem. In this respect, Bayesian Neural Networks (BNNs) are used to handle uncertainty in NNs. The major reason for this is that NNs use maximum likelihood MLE to assess point estimates and ignore the uncertainty regarding proper weight values. One principal disadvantage to this method is the computation of integrals for probability measurement which turns out to be intensive.

Fuzzy Logic is likewise introduced to handle uncertainty estimation. While in classic logic, decisions are based upon a binary system (either yes or no), Fuzzy logic finds room for more precision through the use of membership functions. It has been used in risk assessment prediction for coronary heart disease, preventing and reducing the progression of renal failure, but also in the medical internet of things. The downside of this method lies in the subjectivity of defining fuzzy membership functions as well as the difficulty of validating the results [14].

Monte Carlo Simulation helps in explaining the impact of uncertainty in prediction models. It helps to predict all potential results of a system for better decision making [4]. The method addresses the intractability of integrals analytically encountered with Bayesian Inference. Furthermore, it provides results that are easily understandable. Meanwhile, its computational cost depends highly on the number of repeated runs.

Dempster-Shafer's theory combined with fuzzy logic appealed to many researchers in the scientific community. The method settles some limitations of Bayesian methods, such as lack of ignorance description. Bayesian methods deal with single evidence while DS's method deals with sets of elements. What is more, it offers different metrics to measure, imprecision, ignorance and confusion. It is used along with other uncertainty assessment methods (mainly fuzzy sets) to help the clinicians in medical diagnosis, notably in medical imaging.

AI methods automate prediction processes which make them vastly in various domains. Yet, they are not being used to their full potential in clinical practice as they lack considering uncertainty quantification.

Research Paper	Addressed Problem	Methods	Results	Conclusion and Critics
Handling of uncertainty in medical data using machine learning and probability theory techniques (Alizadehsani R., *Ann Oper Res*, 2021)	Estimating uncertainty provides valuable information for decision making 2 types of uncertainty : - Epistemic uncertainty (relative to the model) - Noise on the data	Methods and theories used : - Bayesian inference - Imprecise probability - Fuzzy logic - Approximate classification - Monte Carlo simulation - Dempster-Shafer theory	The most commonly used uncertainty estimation theories in the literature: 1 - Bayesian inference 2 - Fuzzy logic 3 - Monte Carlo simulation 4 - Dempster-Shafer theory	Successful learning can be achieved even with an arbitrary amount of noise Awareness of the importance of uncertainty in health care should start with physicians who must strive to find new approaches to reduce the amount of noise in the data There is a need for some sort of trade-off between reducing epistemic uncertainty and data noise

Figure 2: Approaches of uncertainty handling in medical data used in the scientific literature [4]

ML models are said to be complex if they describe reality more accurately. For a complex model, the uncertainty on the output decreases and output relative uncertainty increases consequently since much data should be analysed specifically to produce better results. Therefore, it is important to find a compromise between data and model uncertainty.

4 Uncertainty Quantification: Proposed Approach

4.1 First Approach : Bayesian Ensembles

In a recent work, Malinin et al. [6] examined a new technique for uncertainty measuring in the Gradient Boosting model using a Bayesian ensemble-based framework. They had shown that the method was successfully able to detect anomalous inputs in the data. Moreover it is capable of quantifying both data and knowledge uncertainty. The work conducted in the paper can be applied to any supervised machine learning model, be it a regression or classification task.

In the case of a medical domain, physicians are usually confronted with the problem of proposing the right diagnosis for their patients. In Machine Learning, this type of problem is qualified as a Classification Task. It refers to a predictive modeling problem where a class label is predicted for a given input data.

4.1.1 Preliminaries

Every ML model has a certain amount of uncertainty due the lack of its knowledge about inputs far from the region from which the training dataset is from. This is known as knowledge uncertainty (or epistemic uncertainty).

One way of capturing knowledge uncertainty is Bayesian Ensembles. Their main power stems from their ability to decompose total uncertainty into data uncertainty (aleatoric uncertainty) and knowledge uncertainty (epistemic uncertainty).

This starts with defining multiple models with different parameters θ. Here we only consider classification problems. For this type of tasks, the ML models yield the predictive probabilities of class labels along with the target outcome. The models are generated through the sampling of the posterior probability $p(\theta|D)$, where $D = \{x_i, y_i\}$ is the training dataset, since it should produce explanations that are consistent with observations in the training dataset.

Each model yields an amount of data uncertainty which can be expressed by the entropy of the predictive outcome.

$$Uncert_{data} = E_{p(\theta|D)}[H[P(y|x;\theta)]] \tag{1}$$

Whereas knowledge uncertainty is expressed as the spreadness of models in the ensemble considered. However, the main disadvantage to using bayesian inference is the intractability of integrals. Consequently, we consider taking the mean of expectations with respect to each model in the ensemble :

$$P(y|x,D) = E_{p(\theta|D)}[P(y|x;\theta)] = \frac{1}{N}\sum_{n=1}^{n=N} P(y|x;\theta^{(n)}) \tag{2}$$

where $\theta^{(n)} \sim p(\theta|D)$ and N is the number of ML models used. Then, the entropy of the predictive posterior is calculated. It provides an estimate of the overall uncertainty (knowledge uncertainty + epistemic uncertainty).

$$Uncert_{total} = H[P(y|x,D)] \tag{3}$$

Yet, it is important to differentiate data uncertainty from knowledge uncertainty. The latter is obtained through subtracting the estimated entropy of each model from the total uncertainty.

$$Uncert_{knowledge} = H[P(y|x,D)] - E_{p(\theta|D)}[H[P(y|x;\theta)]] \tag{4}$$

4.1.2 Application: Pancreatic Cancer

In this section, we evaluate the performance of the Bayesian ensembles approach on a PC dataset, in order to estimate uncertainty over the predicted cyst type of each patient provided by the model. We define the frame of discernment as

Model N°	Data uncertainty
1	0.567
2	0.674
3	0.53
4	0.645
5	0.73
6	0.573
7	0.556
8	0.665
9	0.672
10	0.65

Table 1: Data uncertainty computed for 10 Machine Learning models

$\Omega = \{$*cadkm, cam, cas, tipmp-deg, tpps, tipmp, kh, tne, pseudocyst, adk*$\}$, which contains patients PC cysts present in the dataset.

The Dataset consists of 73 patients with different features, such as personal information, symptoms, clinical history, medical test results (including MRI, CT scans, endoscopic procedures, and biopsies). The dataset is then splitted to a training dataset and a test dataset (consisting of 15 patients).

Since the dataset's volume is quite restricted, some features were eliminated even despite their importance for an accurate diagnosis, lest they can bias the model.

For this classification task, we chose to use the gradient boosting model Catboost, on account of its high prediction performance as explained in section 3.1. In order to create ensemble models, we generate CatBoost models with random seed values in a bid to explore more areas of knowledge. The ensembles consists of 10 models for which we obtain the following results :

The values of data uncertainty are computed using Shannon Entropy of the predicted distribution:

$$H(X) = -\sum_{A \in \Omega} p(A) \log p(A) \tag{5}$$

For a binary system, entropy values are located in the interval [0,1]. As for a classification problem using multiple classes, the entropy values could exceed 1. In order to normalize its values, we divide by the maximum entropy, which is obtained with a uniform probability distribution :

Patient N°	Entropy using the mean model
1	0.463
2	0.427
3	0.71
4	0.645
5	0.853
6	0.656
7	0.646
8	0.49
9	0.942
10	0.522
11	0.825
12	0.865
13	0.76
14	0.88
15	0.487

Table 2: Calculated Shannon entropy with the mean model for every patient in the test dataset respectively

$$H_{max}(X) = -\sum_{i=1}^{n} \frac{1}{n} \log \frac{1}{n} = \log n = \log(10) = 2.3 \tag{6}$$

where n is the number of PC cysts types in the dataset(in this case $n = 10$).

Hence, the computed values of data uncertainty are clearly in the interval of [0, 1]. (see table 1). These values show the extent to which a model can handle noise and bias present in data.

The overall uncertainty is computed using the equation 3. Since each model provides its own predictive distribution on cysts classes with respect to each patient in the test dataset, we first estimate the mean distribution of the predicted probabilities. Next, entropy values are assessed over each patient in the obtained mean distribution.

Finally, the overall uncertainty is obtained through the mean value of the calculated entropies.

Generally, the patients with the lowest entropy values are the ones for whom the model is quite certain about their outcome (cyst type). On the other hand, those with highest entropy values are the ones for whom provided features are fuzzy, thus inducing a substantial amount of uncertainty.

4.1.3 Results and Critics

After the computing the Bayesian ensemble, the estimated uncertainty values are as follow :

$$\text{Total uncertainty} = 0.887$$
$$\text{Knowledge uncertainty} = 0.06$$
$$\text{Data uncertainty} = 0.82$$

The obtained results regarding uncertainty types are consistent with the PC dataset used for training the model. In fact, the dataset is relatively limited in volume, whereas the features are numerous. On the other hand, knowledge uncertainty is very low, and this can be explained by the fact that the ML model (Gradient Boosting) is a deterministic model that makes decisions based on a logic reasoning scheme.

Using Bayesian Ensembles enables us to separate knowledge uncertainty and data uncertainty. This is important for realizing the source of uncertainty and therefore monitoring it.

However, this approach has few drawbacks. Indeed, in some areas like healthcare, it is important to be capable of estimating the amount of uncertainty with regard to every patient in the test dataset. This can enable the physician to be more cautious with the patient identified by the highest level of uncertainty in terms of the proposed diagnosis and treatment.

Using a probabilistic model does not always allow for a convenient way of combining evidence. In the case of the method used here, we are only considering the mean model of the Bayesian ensemble. This can lead to biased estimates of uncertainty since not all models are equally reliable.

Additionally, distinction between knowledge and data uncertainty is not sufficient. It is more suitable to capture other levels of uncertainty, especially in knowledge uncertainty.

4.2 Second Approach : Dempster Shafer's theory of evidence

In this section, we propose an approach to settle the aforementioned disadvantages while leveraging the strengths of the Bayesian Ensembles approach.

4.2.1 Preliminaries

Dempster-Shafer theory is a generalized scheme for expressing uncertainty. It uses the same standard of probability calculus as in the Bayesian theory. Yet, instead of defining a probability value on each element of the frame of discernment Ω (the set of all possible results), Dempster Shafer Theory considers assigning a mass function to each subset of Ω, noted 2^{Ω}. In fact, unlike Bayesian Inference which assesses a probability as a single value, DS assignes to each set an interval within which the degree of belief for the set must lie [16].

The Dempster Shafer theory is often referred to as a generalization of the Bayesian inference method. In fact, when it deals with individual elements of the frame discernment, Dempster Shafer's theory coincides with probability measure.

Assume a set of possible outcome results Ω (also called frame of discernment). We define a Basic Probability Assignment (BPA, also known as mass function) : $m : 2^{\Omega} \rightarrow [0, 1]$ as a function satisfying the following statements :

$$m(\emptyset) = 0, \tag{7}$$

and

$$\sum_{A \in 2^{\Omega}} m(A) = 1 \tag{8}$$

$m(A)$ is interpreted as the belief allocated to hypothesis x in A and to no other more specific hypothesis.

The Belief Function: The belief function of m assessed in A is defined as follows :

$$Bel_m(A) = \sum_{B \in 2^{\Omega}, B \subseteq A} m(B) \tag{9}$$

It measures the strength of the evidence in favor of proposition A.

The Plausibility Function: The plausibility function Pl of the BPA m is defined as follows :

$$Pl_m(A) = \sum_{B \in 2^{\Omega}, B \cap A = \emptyset} m(B) \tag{10}$$

It measures to what extent the evidence in favor of A leaves room for belief in A.

The Dempster Shafer rule of combination: Dempster Shafer's rule of combination is defined as follows :

$$m(A) = \frac{1}{1 - K} \sum_{B \cap C = A} m_1(B) m_2(B) \tag{11}$$

where $K = \sum_{B \cap C = \emptyset} m_1(B) m_2(C)$ is the conflicting degree between defined BPAs, A, B and C are elements in 2^Ω.

If $K = 1$ the two pieces of evidence are totally conflicting and the dempster shafer's rule of combination cannot be applied. If $K = 0$, the 2 pieces of evidence are non-conflicting.

Jousselme's distance: Jousselme's distance is one of the rational and popular definitions of distance used to measure the distance between two BPAs (mass functions). It is defined as following :

$$d_{ij} = \sqrt{\frac{1}{2}(\vec{m_i} - \vec{m_j})^T D(\vec{m_i} - \vec{m_j})} \tag{12}$$

with m_i, m_j being two BPAs under the frame of discernment. D is the matrix defined as $D(A, B) = \frac{|A \cap B|}{|A \cup B|}$, $A, B \in \Omega$ and $|A|$ represents cardinality. Jousselme's distance is an efficient tool to quantify the dissimilarity of two BPAs.

Shannon Entropy: Shannon entropy of a BPA m is defined as follows :

$$H(X) = -\sum_{A \in 2^\Omega} m(A) \log m(A) \tag{13}$$

As in the previous section, we consider normalising Shannon entropy by dividing by the maximum entropy (obtained for a uniform distribution).

It expresses the level of randomness or chaos in information or evidence. As explained in the previous section, entropy values can exceed 1. For the sake of easy interpretation, we consider normalising its values by dividing by the maximum of Shannon entropy (obtained for a uniform probability distribution).

4.2.2 The proposed method for uncertainty estimation in PC diagnosis

In this section, we draw special attention to total uncertainty estimation (hence knowledge uncertainty), due to its reducibility potential. Thus, data uncertainty will not receive much of our focus.

The approach developed by Malinin et al. (section 4.1) does not allow for a proper combination of evidence. This might lead to biasing the results, specifically as not all pieces of evidence are reliable in the same way. We suggest handling this problem through appropriate methods of evidence combination.

As with the Bayesian ensembles approach, we choose a Gradient Boosting model for training the dataset. The model predicts patients cysts as well as the

associated predictive probabilities.

In the case of our PC dataset, there are 10 different patient's PC cysts . The frama of discernment can then be defined as $\Omega = \{$*cadkm, cam, cas, tipmp-deg, tpps, tipmp, kh, tne, pseudocyst, adk*$\}$. To take advantage of Dempster Shafer's theory of evidence, we define the mass function of every model and every patient in the test dataset. For computation complexity reasons implied by the DS combination rule, we limit our analysis to 2 Gradient Boosting models whose parameters are close to the optimal values to ensure the consistency of results.

In this regard, for each model and patient in the test dataset, we only consider the cyst for which the maximum predictive probability was allocated:

$$m(cystA) = max(\{p(x_i)\})_{i=1}^{i=N} \tag{14}$$

where N is the total number of possible PC cyst types. This way, we draw attention to the main PC cyst type predicted, while the remainder of mass is assigned to the whole frame of discernment as a way of expressing fuzziness regarding the probability of having other cysts types.

This way, each model defines a mass function for every patient in the dataset based on the results predicted. Afterwards, we use the Dempster Shafer rule of combination to combine the pieces of evidence from each model and for every patient.

For example, for a patient *patientX* in the dataset, we define the combined BPA m_{comb}(or mass function) as follows :

$$m_{comb,patientX}(A) = \frac{1}{1 - K_{patientX}} \sum_{B \cap C = A} m_{1,patientX}(B) m_{2,patientX}(B), \tag{15}$$

where :

$$K_{patientX} = \sum_{B \cap C = \emptyset} m_{1,patientX}(B) m_{2,patientX}(C) \tag{16}$$

A, B, and C represent PC cysts from the frame of discernment Ω.

Finally, we measure the total uncertainty by summing and averaging the entropies of the BPA issued from the DS combination for every patient in the dataset.

$$Uncert_{total} = \sum_{patients} H_{combined}(patient) \tag{17}$$

where : $H_{combined}(patient) = -\sum_{A\in\Omega} m_{comb,patientX}(A) \times \log m_{comb,patientX}(A)$

4.2.3 Improvement of the rule of combination

The previous approach provides a better way to combine information while incorporating taking into consideration potential conflict between pieces of evidence. However, DS rule of combination has its computational complexity as its main drawback. The more models we add to the system, the more uncertainty we will be able to detect and the more time it will take to combine knowledge. This can be critical while making decisions about patients diagnosis based on ML models outcome.

An improvement to this has been suggested by Wang et al. (2016, [17]) through a weighted averaging combination rule. This method is based upon assigning a credibility degree to each of the ML models used as a way to get around the DS combination rule.

Given a set of BPAs, the mass function (BPA) issued from the Weighted Averaging combination of the models can be clarified as:

$$m_{WAE} = \sum_{i=1}^{i=n} w_i \times m_i \tag{18}$$

Where w_i is the corresponding weight degree of the BPA m_i. The weights are calculated considering the following steps

1. Calculating similarity between every 2 BPAs using Jousselme's distance

$$Sim(m_i, m_j) = 1 - d(m_i, m_j) \tag{19}$$

2. Measuring the support degree of each BPA (expression)

$$Sup(m_i) = \sum_{j=1, j\neq i}^{k} Sim(m_i, m_j) \tag{20}$$

3. Assessing the credibility degree of BPAs (expression)

$$Crd_i = \frac{Sup(m_i)}{\sum_{j=1}^{k} Sup(m_j)} \tag{21}$$

After computing the m_WAE, the total uncertainty is then assessed using the Shannon entropy (Eq. 13).

5 Results and Discussion

The key feature of using Bayesian Ensembles for uncertainty estimation is its capability of not only measuring the total uncertainty, but also splitting it into data and knowledge uncertainties.

Total uncertainty = 0.887
Knowledge uncertainty = 0.06
Data uncertainty = 0.82

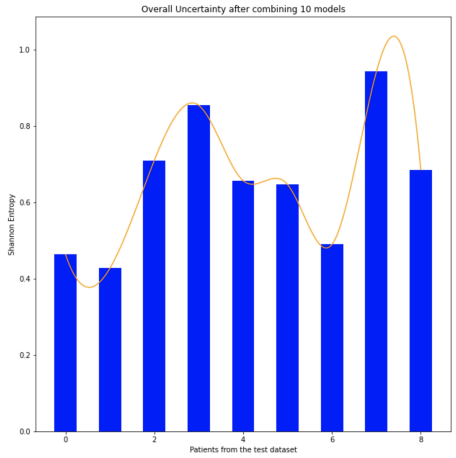

Figure 3: Uncertainty values for some patients in the test dataset using Bayesian Ensembles

The obtained results are consistent since we used 10 Gradient Boosting models with different parameters to capture knowledge from different sources, hence the low value of knowledge uncertainty. However, noise data is massive due to restricted amount of data used for training models. Yet, the main drawbacks of this method lies in the fact it only calculates the overall data and knowledge uncertainty, and provides no way of determining these quantities with respect to every patient.

Using the Dempster Shafer rule of combination, we were able to obtain the following uncertainty values, of individual BPAs and the combined mass function (see figure 3).

Combining mass functions from different models permits for confronting evidence and therefore decide the degree of uncertainty regarding the result predicted by the ML model.

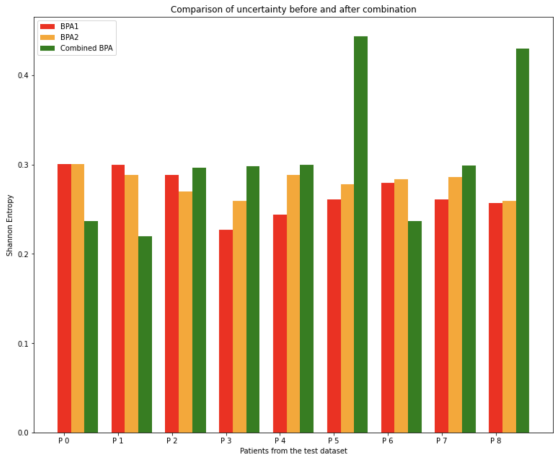

Figure 4: Comparison of uncertainty values between individual and DS combined mass functions for some patients in the test dataset

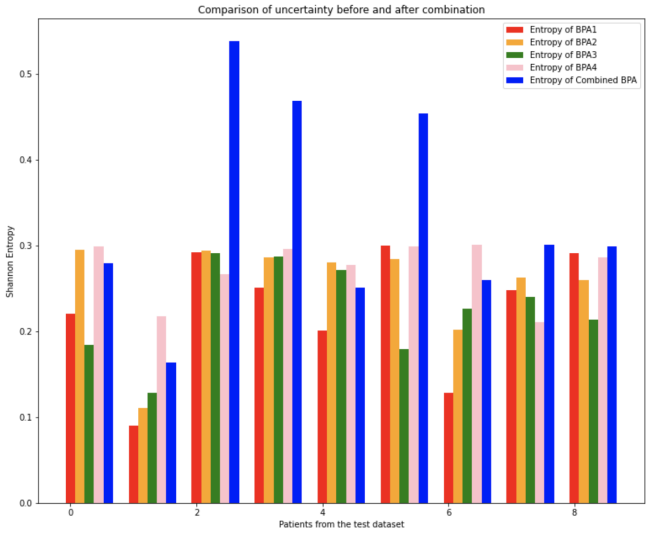

Figure 5: Uncertainty estimation using weighted averaging combination rule for some patients in the test dataset

As the figure 3 shows, combining evidence from different sources allows for better uncertainty estimation. When the two individual BPAs predict the same outcome for a patient, the combined BPA estimates a low level of uncertainty since the pieces of evidence are consistent. By contrast, when medical and personal data of a patient do not approve the same PC cyst diagnosis, the combined mass function assigns a high uncertainty level to the predictions.

In this case, this allows physicians to abstain from considering the predicted outcome of AI models.

However, we only used a limited number of models to validate our results using Dempster Shafer's rule of combination due to its computation complexity. This drawback might generate biased uncertainty estimates. Therefore, it is of an absolute necessity to consider more models for the quantification of uncertainty.

Using the Weighted Averaging combination rule for combining evidence, this approach allocates less uncertainty value to patients for whom models are in accordance (ex. patient "0" and patient "1"), as figure 5 shows. On the other side, patients' PC cyst prediction for whom ML models issued much conflict are assigned with a high value of uncertainty.

Simply averaging the predictions obtained from different models as it is the case when using Bayesian Ensembles is not a suitable method of combining evidence. Different models may have different reliability degrees, and this plays a major role on the way the final results should be predicted. Consequently, when comparing uncertainty estimates from the Bayesian approach (figure 3) with the ones issued from the Dempster Shafer's rule of combination (figure 5), we notice that the latter predicts uncertainty more precisely than the former, in that it highlights more clearly the divergence of uncertainty estimates between patients.

In contrast with the Bayesian Inference approach, rules of combination allow for uncertainty measuring relative to each patient individually. Using the weighted averaging approach allows to combine outcome from several models and therefore without inducing substantial computation complexity. Besides, it combines evidence from different sources based models credibility values. Thus, it detects uncertainty more accurately since it provides room for more models confrontation. This helps to ensure that models with biased results will not affect much the combined mass function.

6 Conclusion and Future Work

Uncertainty handling is still a new domain of research, notably with the advent of AI technologies. In this article, we have presented some approaches to uncertainty handling in medical machine learning. The proposed method is based upon mass functions, key element in uncertainty quantification using Dempster Shafer theory of evidence. We suggested an improvement to using DS combination rule by considering weighted averaging combination rule, while assigning each piece of evidence to its corresponding credibility degree.

The approaches developed in this article are mainly applicable to classification tasks, but can also be easily adapted to regression tasks as well, by considering variance instead of entropy, as Malinin et al. suggested in their paper work.

However, several other questions are yet to be resolved in future works. Needless to say that the limited volume of data at hand is usually the main restraint to developing a reliable ML model. Moreover, data mining plays an important role in organising data and thus, producing reliable results. In this regard, the use of ontologies can be of a valuable support since they serve as means of organising data into information and knowledge. Hence, developing an ontology-based medical AI holds great potential for decreasing aleatoric uncertainty and supporting physicians in the process of decision making regarding both diagnosis and treatment.

References

[1] Klein AP., Lindström S., Mendelsohn JB., Steplowski E., Arslan AA., Bueno-de-Mesquita HB., et al. (2013). *An Absolute Risk Model to Identify Individuals at Elevated Risk for Pancreatic Cancer in the General Population*. PLoS ONE 8(9): e72311.

[2] Bachet JB. *Cancer du pancréas*. Livre Blanc de l'Hépato-gastroentérologie, Paris. www.livre-blanc-cnp-hge.fr/cancer-du-pancreas/

[3] Yu A., Woo SM., Joo J., Yang H-R., Lee WJ., Park S-J., et al. (2016). *Development and Validation of a Prediction Model to Estimate Individual Risk of Pancreatic Cancer*. PLoS ONE 11(1): e0146473.

[4] Alizadehsan R., Roshanzamir M., Hussain S. et al. (1991–2020). *Handling of uncertainty in medical data using machine learning and probability theory techniques: a review of 30 years*. Ann Oper Res (2021)

[5] Zhang L, Sanagapalli S, Stoita A. (2018). Challenges in diagnosis of pancreatic cancer. *Challenges in diagnosis of pancreatic cancer*. World J Gastroenterol. 2018;24(19):2047-2060.

[6] Malinin A., Prokhorenkova L., Ustimenko A., (Apr 2021). *Uncertainty in Gradient Boosting via Ensembles*. arXiv:2006.10562v4 [cs.LG].

[7] *The Pancreas and Its Functions*. The Pancreas Center, COLUMBIA SURGERY. www.columbiasurgery.org/pancreas/pancreas-and-its-functions.

[8] Leblanc S. (2015). *Lesions kystiques du pancreas* Hôpital Cochin, service de gastroenterologie.

[9] Bender E. (2020). *Will a test to detect early pancreatic cancer ever be possible?* Nature 579, S12-S13(2020).

[10] Kourou K., P. Exarchos T., P. Exarchos K., V. Karamouzis M., I. Fotiadis D. (2015). *Machine learning applications in cancer prognosis and prediction.* Computational and Structural Biotechnology Journal, Volume 13, 2015, Pages 8-17, ISSN 2001-0370.

[11] Dorogush A., Ershov V., Gulin A. (2018). *CatBoost: gradient boosting with categorical features support.* ArXiv, abs/1810.11363.

[12] Taninaga J., Nishiyama Y., Fujibayashi K. et al. (2019). *Prediction of future gastric cancer risk using a machine learning algorithm and comprehensive medical check-up data: A case-control study.* Nature, Sci Rep 9, 12384 (2019).

[13] Rawla P., Barsouk A. (2019). *Epidemiology of gastric cancer: global trends, risk factors and prevention* Prz Gastroenterol. (2019). vol. 14(1):26-38.

[14] Eierdanz F., Alcamo J., Acosta-Michlik L. et al. (2008). *Using fuzzy set theory to address the uncertainty of susceptibility to drought.* Reg Environ Change 8, 197–205.

[15] *Uncertainty quantification.* Wikipedia, Wikimedia Foundation (June 2021). en.wikipedia.org/wiki/Uncertainty_quantification.

[16] Das A. (2003). *Knowledge Representation* Editor(s): Bidgoli H., Encyclopedia of Information Systems, Elsevier, 2003, Pages 33-41, ISBN 9780122272400.

[17] Wang J., Hu Y., Xiao F., Deng X., Deng Y. (2016). *A novel method to use fuzzy soft sets in decision making based on ambiguity measure and Dempster–Shafer theory of evidence: An application in medical diagnosis.* Artificial Intelligence in Medicine, Volume 69, 2016, Pages 1-11, ISSN 0933-3657.