Assessor Models for Predicting and Explaining Aleatoric Uncertainty in Classification

Pedro B. S. Lima[1], **Ricardo B. C. Prudêncio**[1],
Ana Carolina Lorena[2], **Maria Gabriela Valeriano**[2]

[1]Centro de Informática
Universidade Federal de Pernambuco (UFPE) – Recife, PE – Brazil

[2]Instituto Tecnológico de Aeronáutica
Departamento de Ciência e Tecnologia Aeroespacial (DCTA)
São José dos Campos, SP – Brazil

{rbcp, pbsl}@cin.ufpe.br, {aclorena, mariagabrielavaleriano}@gmail.com

Abstract. *Machine Learning (ML) models have been successfully adopted in several applications. However, despite their success, high levels of uncertainty can be observed depending on the problem being solved. Specifically, aleatoric uncertainty in classification is related to the inherent randomness associated with predictor and target attributes, usually observed for instances close to an area of class overlap. The literature on ML has developed different methods for quantifying uncertainty. The current paper addresses the complementary task of explaining uncertainty. For this, we proposed using assessors, which are meta-models trained to predict the performance of a given base model on a problem of interest. In our work, we adapted the concept of assessors to predict the aleatoric uncertainty of instances, measured by the average class entropy across a pool of diverse base models. Once built, eXplainable AI (XAI) techniques are adopted to extract explanations from the assessor. Experiments were performed in two case studies, using both a simulated and a real dataset for predicting the severity of COVID-19 patients. A pool of 12 base models was adopted to measure aleatoric uncertainty, and Random Forest was adopted as the assessor. The explanations extracted from the assessor were useful to identify specific features that cause high uncertainty in the classification problems.*

1. Introduction

Machine Learning (ML) techniques have been successfully adopted in different applications, covering a variety of problems like classification, regression, anomaly detection, and language understanding, among others. Despite the usefulness of ML models, their quality is usually not homogeneous in a given problem [Burnell et al. 2023]. Specifically, in supervised classification, some instances in a dataset may be harder than others to classify due to different reasons [Lorena et al. 2023]. In such cases, the confidence of an ML model may be low, reflected, for example, in terms of its predicted class probabilities.

The literature distinguishes two general types of uncertainty in ML: aleatoric and epistemic [Hüllermeier and Waegeman 2021]. Aleatoric uncertainty is intrinsic to a problem, caused mainly by a non-deterministic relation between predictor and target attributes. For example, in a binary classification problem, an instance may lie close to a class boundary and hence the *true* class conditional probability $p(y|\mathbf{x})$ is close to 0.5. When aleatoric

uncertainty is high, no model would be correctly confident about its prediction. Epistemic uncertainty, in turn, arises due to the lack of knowledge related to the available data and model. This is the case, for example, when training sets are scarce or a wrong model is chosen at development time.

There are different methods in literature to quantify uncertainty in ML, like standard uncertainty measures (e.g., maximum predicted probability and class entropy, conformal prediction, credal sets) as well as specific methods to decompose the aleatoric and epistemic components of uncertainty [Bhatt et al. 2021, Hüllermeier and Waegeman 2021]. Beyond quantification, the current paper addresses a complementary task: explaining uncertainty. So this paper is directly related to the area of eXplanaible AI [Molnar 2022], but instead of explaining model predictions, we aim to derive explanations for model uncertainty.

Previous works have investigated XAI techniques to explain uncertainty in the context of ML with reject option [Hendrickx et al. 2024]. In this context, given an instance, the model uncertainty is quantified, and its prediction is rejected if the measured uncertainty is higher than an acceptance threshold. Once rejected, post-hoc explanations of uncertainty are derived by adapting local methods like LIME and counterfactuals [Artelt et al. 2023, Antorán et al. 2020]. The current paper investigates a different direction, which is to build a global assessor [Hernández-Orallo et al. 2022] to predict and explain the aleatoric uncertainty of instances in a problem, estimated in our work by the average class entropy across a pool of diverse base models.

In our proposal, given a pool of models and a test set of instances, initially the aleatoric uncertainty is quantified for each instance. Then a set of meta-examples is built, each one storing: (1) the instance's attributes; and (2) the quantified uncertainty, which is now the target attribute for the assessor. Based on a set of meta-examples, the assessor is learned to model the relationship between the predictor attributes and the aleatoric uncertainty. Once the assessor is a model itself, standard XAI techniques can be used to derive both global and local explanations, which are more flexible than previous works.

The proposal is adopted in two case studies. First, we applied it in the Two Moons dataset, a non-linear synthetic binary classification problem, which was convenient for inspecting results. Additionally, we applied the proposal to a real dataset to predict the severity of COVID-19 in patients based on laboratory tests. In the case studies, 12 diverse ML models were used for classification. The aleatoric uncertainty of each instance is measured by the average class entropy across the models. Then, the Random Forest (RF) regression algorithm was adopted to learn the assessor model. In the experiments, Permutation Feature Importance (PFI) and Partial Dependence Plots (PDP) were adopted to derive explanations from the assessors. In the experiments, we evaluated the assessor and discussed the insights gained from the derived explanations.

The remaining of this paper is organized as follows. Section 2 presents the background and related work for contextualizing the current paper. This is followed by Section 3, which presents the proposal and the performed case studies. Finally, Section 4 concludes the paper with final considerations and future work.

2. Background and Related Work

In this section, we introduce the background on the problem of uncertainty quantification in ML (Section 2.1), followed by the topic of assessors (Section 2.2), which is the base of the proposed solution. Finally, in Section 2.3, we discuss previous works on using XAI techniques to explain uncertainty in ML, which is closely related to our proposal.

2.1. Uncertainty in ML Classification

In the ML context, uncertainty refers to the lack of confidence in the output of a predictive model [Bhatt et al. 2021]. Let $\mathcal{Y}$ be the set of possible classes and let $p(y|\mathbf{x}_i, h)$ be the conditional probability predicted by model h for the class $y \in \mathcal{Y}$. The uncertainty related to the predicted class probabilities can be computed in different ways [Bhatt et al. 2021]. The most common measure is the maximum class probability returned by the model (see Eq. 1), or alternatively the entropy of the class probabilities (see Eq. 2). Class entropy tends to be more informative, especially in multi-class problems.

$$U_h\left(\mathbf{x}_i\right) = 1 - \max_{y \in \mathcal{Y}} p(y|\mathbf{x}_i, h) \tag{1}$$

$$U_h\left(\mathbf{x}_i\right) = -\sum_{y \in \mathcal{Y}} p(y|\mathbf{x}_i, h) \log(p(y|\mathbf{x}_i, h)) \tag{2}$$

A high uncertainty can result from different factors, distinguished as either aleatoric or epistemic [Hüllermeier and Waegeman 2021]. Aleatoric uncertainty is associated with the intrinsic difficulty of an instance, which is usually the byproduct of class overlap, i.e., a non-deterministic relation between the predictors and the class attribute. Epistemic uncertainty, in turn, usually refers to the lack of knowledge about the adequate model for a problem or the lack of representative data during the model training. The total uncertainty is the sum of the aleatoric and epistemic uncertainty [Bhatt et al. 2021].

A pool of models $\mathcal{H}$ can be used to estimate the total uncertainty of an instance, trying to overcome the choice of a single model to make predictions. So instance uncertainty can be measured by the *predictive entropy* as follows:

$$U\left(\mathbf{x}_i\right) = -\sum_{y \in \mathcal{Y}} \overline{p}(y|\mathbf{x}_i) \log(\overline{p}(y|\mathbf{x}_i)) \tag{3}$$

in which $\overline{p}(y|\mathbf{x}_i)$ is the average class probability for class y and instance $\mathbf{x}_i$:

$$\overline{p}(y|\mathbf{x}_i) = \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}} p(y|\mathbf{x}_i, h) \tag{4}$$

By following [Depeweg et al. 2018], the predictive entropy can be decomposed into its aleatoric and epistemic components. Aleatoric uncertainty can be measured by the average entropy across models:

$$U_a\left(\mathbf{x}_i\right) = \frac{1}{H} \sum_{h} U_h\left(\mathbf{x}_i\right) \tag{5}$$

Table 1. Example of decomposing the total uncertainty (predictive entropy) into aleatoric and epistemic components, for two instances and five models.

Model	$p(+\|\mathbf{x}_1)$	$U_h(\mathbf{x}_1)$	$p(+\|\mathbf{x}_2)$	$U_h(\mathbf{x}_2)$
h_1	0.90	0.46	0.60	0.97
h_2	0.80	0.72	0.55	0.99
h_3	0.50	1.00	0.50	1.00
h_4	0.20	0.72	0.45	0.99
h_5	0.10	0.46	0.40	0.97
$\overline{p}(+\|\mathbf{x})$	0.50		0.50	
$U(\mathbf{x})$	1.00		1.00	
$U_a(\mathbf{x})$	0.67		0.98	
$U_e(\mathbf{x})$	0.33		0.02	

In turn, epistemic uncertainty can be measured by the complement of the predictive entropy as follows:

$$U_e(\mathbf{x}_i) = U(\mathbf{x}_i) - U_a(\mathbf{x}_i) \tag{6}$$

Table 1 illustrates the computation of uncertainty for two instances and five models in a binary classification problem, in which total uncertainty is measured by the predictive entropy (Eq. 3). For both instances, the pool of models is highly uncertain: $\overline{p}(+|\mathbf{x}_1) = \overline{p}(+|\mathbf{x}_2) = 0.5$, which results on the maximum predictive entropy ($U = 1$). There is a high variation in the uncertainty across the models (U_h) for instance $\mathbf{x}_1$. Some models are more confident than others. In turn, for instance $\mathbf{x}_2$, all models are uncertain. These differences across models are reflected in the aleatoric and epistemic uncertainty values. While instance $\mathbf{x}_2$ has a higher aleatoric uncertainty, instance $\mathbf{x}_1$ has a higher epistemic uncertainty, which could be reduced by choosing the right models. We could say that instance $\mathbf{x}_2$ is more challenging, as it has a high aleatoric uncertainty, which is irreducible in principle. For this, the current paper focuses on aleatoric uncertainty.

2.2. Assessor Models

The concept of assessor models was introduced in [Hernández-Orallo et al. 2022], aiming to generalize the results collected from evaluating a base AI system on different tasks and then predict the performance for new tasks. The idea is straightforward. Each training example for the assessor is related to a specific task, labeled with the system performance measured in a testing battery. So, a supervised ML model is learned using training examples from a set of tasks. Once built, the assessor can predict the system performance for new tasks based on their characteristics. Assessor models have already been applied for different practical applications [Zhou et al. 2022, Da Costa et al. 2023].

Closely related to our work, in [Prudencio et al. 2024], assessors were adopted to predict the hardness of instances in classification tasks. In the assessor modeling, the AI system is a classifier pool and the AI task is an instance to predict. So the paper relies on a test dataset $\mathcal{D} = \{\langle \mathbf{x}_i, y_i \rangle\}_{i=1}^{N}$ of N instances. For each instance, the pool is applied and the hardness $IH(\langle \mathbf{x}_i, y_i \rangle)$ is measured for that instance, defined as the average error obtained by the pool. Then a training example for the assessor is built by storing: (1) the

original instance's features, $\mathbf{x}_i$; and (2) the value of $IH\left(\langle \mathbf{x}_i, y_i \rangle\right)$ as the assessor's target attribute. Finally, based on such examples, the assessor is learned to predict the IH for an instance according to its features. The current paper differs from [Prudencio et al. 2024] because the assessor predicts the uncertainty as the target attribute, instead of predicting average error. This is convenient, for example, to build an assessor when a pool of models is available to make predictions, but a good labeled test set is not.

2.3. Related Work

In the literature, we identified some previous works that used XAI methods for explaining instance uncertainty, mainly developed in the context of ML with reject option [Hendrickx et al. 2024].In this context, given a model and an instance, initially the model uncertainty is quantified and the prediction is rejected if the uncertainty is greater than a predefined threshold. Once the instance is rejected, an XAI method is used to explain the causes of rejection. In [Antorán et al. 2020], this is done by producing counterfactual explanations of uncertainty, identifying features that distinguish the uncertain instance from a certain one. In [Watson et al. 2024], the Shappley method is adapted. In [Artelt et al. 2023] in turn, the LIME method is adapted to explain uncertainty measured by conformal prediction. As in the standard LIME, a local surrogate is learned from data points surrounding the instance to be explained. However, the local points are labeled with their uncertainty, and hence the local surrogate is an interpretable model to predict the uncertainty measure.

Our paper is related to the above papers, but we can point out some differences. First, previous works have adapted single XAI techniques like LIME to explain uncertainty. In turn, the assessor is a model itself and hence existing XAI techniques can be used in a flexible way to provide different types of explanations. Notably, global or semi-local explanations can be derived at development time to inspect the potential uncertain areas in a problem. Also, local explanation methods like LIME can be eventually adopted on the top of the assessor, to derive explanations for particular instances if necessary, similarly to [Artelt et al. 2023]. Second, the assessor is anticipative, i.e., it can be used to predict uncertainty before the base models are actually dispatched, as pointed out in [Hernández-Orallo et al. 2022]. Finally, regarding uncertainty quantification, in the experiments, we focused on quantifying, predicting, and then explaining aleatoric uncertainty, which can be critical in revealing intrinsic uncertain situations in a classification task. The proposed methods, however, can be adapted to other types of uncertainty.

3. Proposal

The main goal of this paper is to investigate the use of assessor models to predict and explain uncertainty in classification problems. Specifically, the current paper is interested in predicting aleatoric uncertainty, estimated in our work by the average entropy across a pool of classification models. Figure 1 presents the architecture of the proposed solution. By following the notation of [Prudencio et al. 2024], the proposed solution relies on a pool of models $\mathcal{H}$ and a test set $\mathcal{D} = \{\langle \mathbf{x}_i, y_i \rangle\}_{i=1}^{N}$ of N instances. For each instance, the uncertainty is quantified using the pool of models. As already mentioned, this is done by computing the average entropy $U_a\left(\mathbf{x}_i\right)$ (see Eq. 5). Then a meta-example stores: (1) the instance's features, $\mathbf{x}_i$; and (2) the value of $U_a\left(\mathbf{x}_i\right)$ as the target attribute. The assessor is then a regression model to predict the U_a for the instances according to their features.

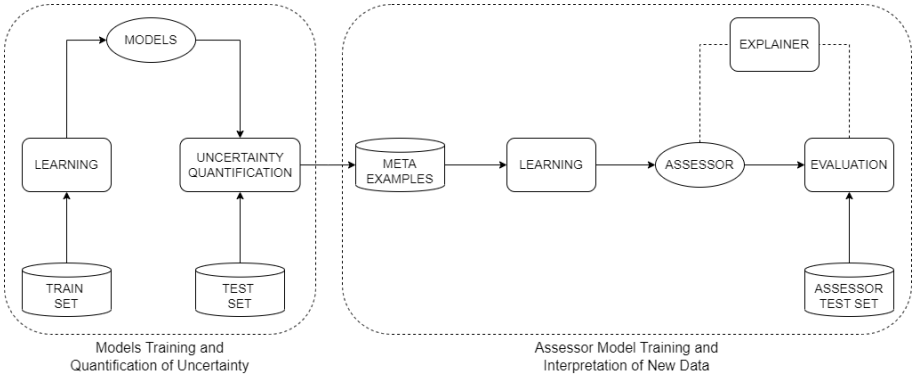


Figure 1. Proposed architecture

As the assessor is a model, various XAI methods can be used to extract explanations. As discussed in Section 2.3, the assessor is a global surrogate of uncertainty, and hence valuable for extracting both global and local explanations of uncertainty. The assessor is anticipative, i.e., it can predict the uncertainty of instances without the need to use the base models in the pool. We highlight that although our focus is to model aleatoric uncertainty, the proposed solution can be adopted with other definitions of uncertainty and to model the uncertainty of individual black-box models.

3.1. Experiments: Two Moons Problem Dataset

In a first battery of experiments, we adopted the proposed solution to the Two-Moons problem, a simulated non-linear binary classification task. A dataset with 5000 examples and a noise level of 0.22, was generated (as illustrated in Figure 2a) using the Scikit-learn library[1]. The dataset was equally divided into training and test. The first set was used to train a pool of 12 base models, while the holdout test set $\mathcal{D}$ was used for testing the trained ML models, and then for producing the meta-examples to train the assessor. This process was repeated 30 times, each one with different randomized seeds.

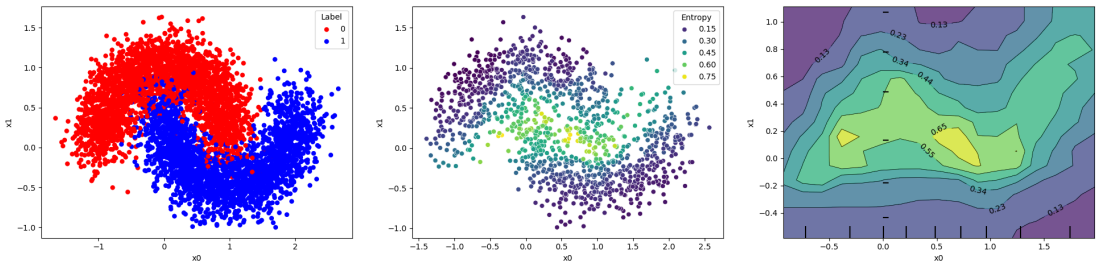
3.1.1. Uncertainty Quantification

For quantifying the aleatoric uncertainty of each test instance, the predicted probability was collected from a pool composed of 12 different models: two Decision Trees with different maximum depths (2 and 10) and minimum samples per split (50 and 10), two Support Vector Machines with different kernels (linear and RBF), Random Forests, k-Nearest Neighbors, Gaussian Process, Gaussian Naive-Bayes, Logistic Regression, Stochastic Gradient Descent, Multi-layer Perceptron and Kolmogorov-Arnold Network (KAN) [Liu et al. 2024]. Except for the KAN, which required its own library, Pykan[2], all the other models were adopted using their respective implementations in the Scikit-learn library[1]. Once the predicted probabilities are collected, uncertainty is measured by adopting the expected entropy of each instance, as defined in Eq. 5.

Figure 2b presents an example test dataset, collored according to the aleatoric uncertainty. As expected, the uncertainty is more prominent in the areas of overlap between

[1]https://scikit-learn.org
[2]https://github.com/KindXiaoming/pykan

the two classes, reaching $U_a(\mathbf{x}_i) \approx 0.5$ on the tips of both moons. Uncertainty can be even higher, e.g., $U_a(\mathbf{x}_i) \approx 0.75$, for noisier instances. For this example, the mean and standard deviation (SD) values for the expected entropy are, respectively, 0.263 and 0.166, while across all tests, the average for both metrics were 0.264 and 0.173, respectively; indicating a general intermediate difficulty of the problem.



(a) Original Two-Moons Problem Data **(b) Expected Entropy of Test Set** **(c) PDP of the Trained Assessor**

Figure 2. The Two-Moons Problem dataset example

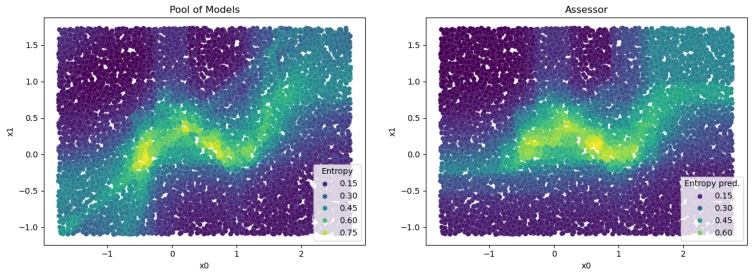
3.1.2. Assessor Model

This paper adopted a Random Forest Regressor model as the assessor, trained and evaluated with the meta-examples from the holdout set of instances. It was also evaluated using 10-fold cross-validation. The performance of the assessor in this experiment was measured by the Root Mean Squared Error (RMSE), which was 0.003. To verify how the assessor extrapolates its predictions, we evaluated it in regions of data not seen during its training (i.e., out-of-distribution data). For this, we produced a new test set sampled from a uniform distribution, as illustrated in Figure 3. With the newly generated set of 10000 instances, the aleatoric uncertainty was then calculated just like in Section 3.1.1, so it could be used for the measurement of the assessor's performance predictions, which resulted in an average RMSE of 0.086 (with 0.019 of standard deviation). Although this value is not high in absolute terms (considering the range of values for class entropy), it is greater than the RMSE observed for the in-distribution data (i.e., average RMSE = 0.003).

As seen in Figures 3a and 3b, the actual and predicted uncertainties on the top-right and the bottom-left regions of the distribution have the most discrepant values. For analysing the top-right region, we considered instances with attributes $x_0 \geq 1.4$, and $x_1 \geq 0.6$, resulting on the average RMSE of 0.141. In turn, for the bottom-left, we considered the instances with attributes $x_0 \leq -0.4$ and $x_1 \leq -0.1$, resulting in the average RMSE of 0.177. In both cases, the instances correspond to out-of-distribution data. Especially in the bottom-left region, the assessor tended to underestimate the uncertainty of instances (see Figure 4b). So, despite the potential of assessors, its use must be carefully judged for out-of-distribution instances, which is actually the case for ML models in general.

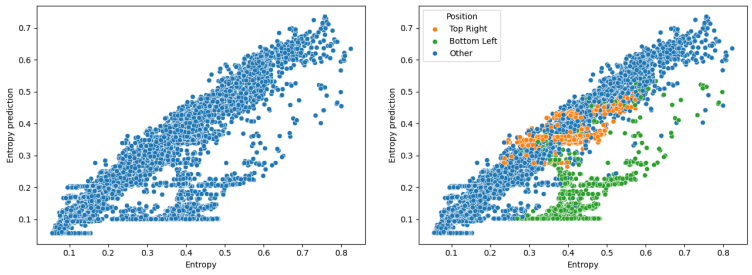
3.1.3. Explanations

Figure 2c presents the PDP extracted from the assessor model. The PDP actually shows that the assessor precisely identified that the instances located in the regions of class over-



(a) Pool Entropy Values **(b) Assessor Predictions**

Figure 3. Example evaluation of assessor's predictions of newly generated data



(a) All instances **(b) Highlighted positions**

Figure 4. True Uncertainty Values versus Predictions for the Two Moons Dataset

lap have a higher aleatoric uncertainty than the other instances. Notice that uncertainty is even higher in the extreme positions of the two moons. Uncertainty progressively decays as instances move away from the class frontier.

3.2. Experiments: COVID Dataset

In this section, we present the experiments performed with a dataset of COVID-19 severity prognosis[3] [Valeriano et al. 2022]. This dataset was produced from the blood tests performed on patients from two large hospitals in São Paulo during the global pandemic. It comprises 17 input features, such as age, sex and 15 blood test results, collected from laboratory tests. As target attribute, it contains a label indicating the severity of each patient case. The severity was defined by either an extended hospitalization of two weeks at least, or the patient's death within 30 days. The data from the first hospital was divided into two sets: a training and a holdout set. The training data was adopted to learn the 12 base models, while the holdout set was used to evaluate them and then to train the assessor. The data from the second hospital was only adopted for testing the assessor.

3.2.1. Uncertainty Quantification

As in Section 3.1.1, the measure of uncertainty used in these experiments was the average class entropy, based on the predicted probabilities from the pool of ML models. Also, the pool was the same as in the previous case study. The aleatoric uncertainty increased drastically from the Two-Moons Problem dataset to the COVID dataset, reaching the

[3]https://repositoriodatasharingfapesp.uspdigital.usp.br/

mean value of 0.7, with a standard deviation of 0.09. About 75% of the instances in the training set have a $U_a(\mathbf{x}_i) \geq 0.65$.

3.2.2. Assessor Model

In this section, the experiment adopted a similar methodology to evaluate the assessor, as in Section 3.1.2. Initially, a Random Forest was trained and evaluated as the assessor model. The aleatoric uncertainty was quantified for each instance, and the assessor was evaluated using 10-fold cross-validation, with an average score of 0.07.

Additionally, the assessor model was tested with the data from the second hospital, obtaining an average RMSE of 0.077 (with a standard deviation of 0.021). The scatterplot of this assessor's entropy predictions against the actual expected entropy values, shown in Figure 5, presents a similar linear pattern observed from the same plot for the Two Moons Problem dataset (see Figure 4), even though both datasets are considerably different. This is relevant as the assessor was trained using data from a specific hospital and produced good predictions for patients from a second hospital.

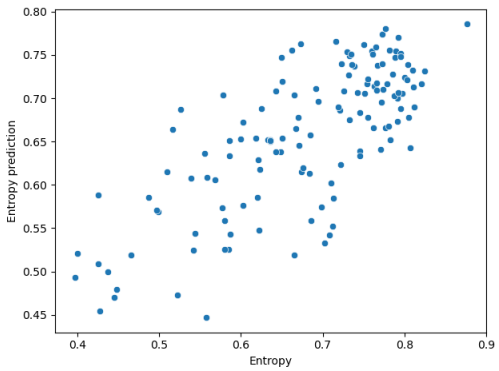


Figure 5. True Uncertainty values vs. Predictions from 2nd Hospital's dataset

3.2.3. Explanations

Figure 6 presents the PFI plot for the trained assessor. The plot shows that the most important features were the level of urea in blood plasma and the age of the patient. To perform a deeper analysis of these attributes, we present their PDPs and class histograms (see Figure 7). Regarding the urea attribute, the uncertainty is low for patients with high levels for this attribute (see Figure 7a). Most patients develop severe cases for values greater than 100 (see Figure 7c). Actually, urea tests often appear in COVID studies, as the disease affects multiple systems, including the kidneys [Peramo-Álvarez et al. 2021]. Uncertainty is high in turn for patients with low levels of urea in isolation ($U_a(\mathbf{x}_i) \approx 0.69$). A combination of other factors then explains severity.

Regarding age, in turn, aleatoric uncertainty is higher for older patients, reaching $U_a(\mathbf{x}_i) \approx 0.70$ on patients older than 50 years (see Figure 7b). The class overlap on higher ages (see Figure 7d) also indicates the increase in uncertainty. Although older people tend to develop more severe cases than young people [Barek et al. 2020], this does not

always happen, which then causes uncertainty in severity prediction. Interestingly, most hospitalized patients were old people. So the prediction uncertainty is not explained, for example, by the lack of training data, but due to causes poorly understood about COVID.

The above analysis shows distinct situations regarding uncertainty. First, a low level of urea would indicate in principle the non-severity class, but the uncertainty is high on this conclusion. Differently, old age would suggest severity, but again the uncertainty is high. Figure 8 presents the 2-dimensional PDP, illustrating the interaction between these two attributes on the uncertainty. First, uncertainty is low for high urea levels, regardless of the patient's age. Second, an area of high aleatoric uncertainty is observed for patients older than 50 years old but with urea levels below 60. As said, while old age would be a risk factor for COVID-19 severity, a low level of urea would suggest non-severity. Such patients then present two opposite signals for predicting the COVID-19 severity, which was then reflected in a higher level of aleatoric uncertainty.

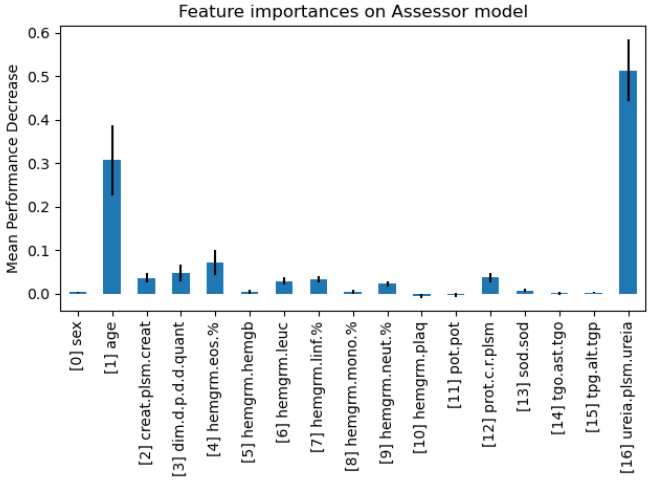


Figure 6. Feature Importance Plot of the Assessor Model

4. Conclusion

In this paper we proposed a novel solution to predict and then explain the aleatoric uncertainty in ML. This solution is composed of 3 major components: (1) measurement of aleatoric uncertainty from instances of the test set by a trained pool of ML models; (2) training of a predictive model (assessor), with a meta-examples set, composed of the test set input features and the aleatoric uncertainty for those instances; and (3) the application of *xAI* techniques to explain the assessor's predictions. Each of these components was implemented using specific methods in both performed experiments. For the first, the uncertainty was quantified by the average entropy of each instance's class probabilities, from the pool of models. The second component was made with Random Forest Regressors as the assessors. And the third mainly used PFI and PDPs, as tools for explanations. The proposal was applied to both artificial and real datasets.

For future works, other implementations could used for each method, which could be more suitable for different datasets. For example, deep artificial neural networks can be used as base models for more complex and bigger datasets. Different methods could be adopted to quantify uncertainty, following the literature on this topic. Specific aspects

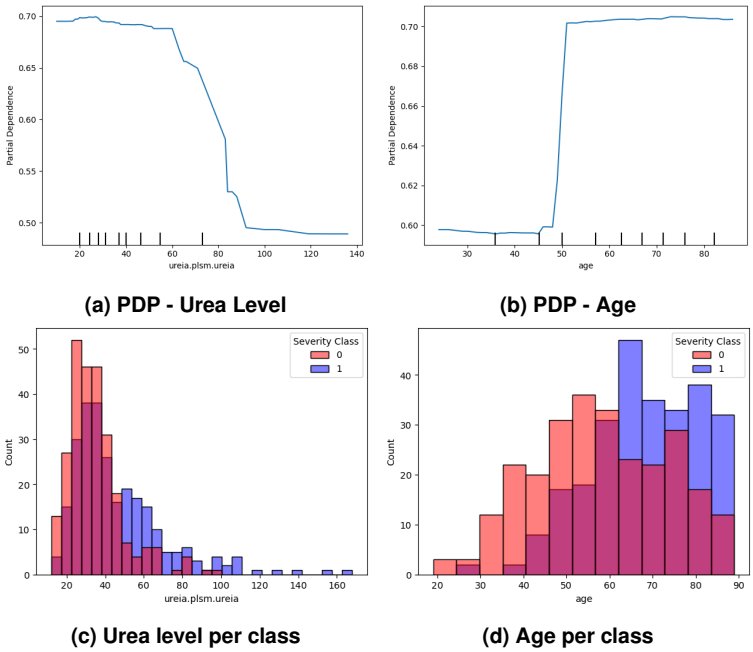


Figure 7. Individual Analysis of Attributes from the 1st Hospital's COVID Dataset

of the construction of the assessor can be investigated. For example, how to deal with out-of-distribution data is a challenging problem. Finally, other XAI tools could be applied to learned assessors, including local methods for explanation.

References

Antorán, J., Bhatt, U., Adel, T., Weller, A., and Hernández-Lobato, J. M. (2020). Getting a clue: A method for explaining uncertainty estimates. *arXiv preprint arXiv:2006.06848*.

Artelt, A., Visser, R., and Hammer, B. (2023). "i do not know! but why?"—local model-agnostic example-based explanations of reject. *Neurocomputing*, 558:126722.

Barek, M. A., Aziz, M. A., and Islam, M. S. (2020). Impact of age, sex, comorbidities and clinical symptoms on the severity of covid-19 cases: A meta-analysis with 55 studies and 10014 cases. *Heliyon*, 6(12).

Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., Melançon, G., Krishnan, R., Stanley, J., Tickoo, O., et al. (2021). Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 401–413.

Burnell, R., Schellaert, W., Burden, J., Ullman, T. D., Martinez-Plumed, F., Tenenbaum, J. B., Rutar, D., Cheke, L. G., Sohl-Dickstein, J., Mitchell, M., et al. (2023). Rethink reporting of evaluation results in ai. *Science*, 380(6641):136–138.

Da Costa, D. C., Prudêncio, R., and Mota, A. (2023). Assessor models with a reject option for soccer result prediction. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1200–1205.

Depeweg, S., Hernandez-Lobato, J.-M., Doshi-Velez, F., and Udluft, S. (2018). Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International conference on machine learning*, pages 1184–1193. PMLR.

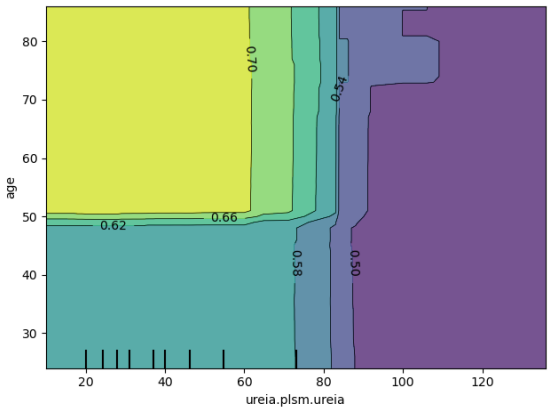


Figure 8. 2-Dimensional PDP for Analysis of Attributes for the COVID Dataset. Blue indicates low uncertainty, while yellow indicates high uncertainty.

Hendrickx, K., Perini, L., Van der Plas, D., Meert, W., and Davis, J. (2024). Machine learning with a reject option: A survey. *Machine Learning*, 113(5):3073–3110.

Hernández-Orallo, J., Schellaert, W., and Martínez-Plumed, F. (2022). Training on the test set: Mapping the system-problem space in ai. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 12256–12261.

Hüllermeier, E. and Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506.

Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T. Y., and Tegmark, M. (2024). Kan: Kolmogorov-arnold networks.

Lorena, A. C., Paiva, P. Y. A., and Prudêncio, R. B. C. (2023). Trusting my predictions: on the value of instance-level analysis. *ACM Computing Surveys*.

Molnar, C. (2022). *Interpretable Machine Learning*. 2 edition.

Peramo-Álvarez, F. P., López-Zúñiga, M. Á., and López-Ruz, M. Á. (2021). Medical sequels of covid-19. *Medicina Clínica (English Edition)*, 157(8):388–394.

Prudencio, R. B., Lorena, A. C., Silva-Filho, T., and Drapal, P. (2024). Assessor models for explaining instance hardness in classification problems. In *2024 IEEE International Joint Conference on Neural Networks*.

Valeriano, M. G., Kiffer, C. R., Higino, G., Zanão, P., Barbosa, D. A., Moreira, P. A., Santos, P. C. J., Grinbaum, R., and Lorena, A. C. (2022). Let the data speak: analysing data from multiple health centers of the são paulo metropolitan area for covid-19 clinical deterioration prediction. In *22nd IEEE CCGrid*, pages 948–951. IEEE.

Watson, D., O'Hara, J., Tax, N., Mudd, R., and Guy, I. (2024). Explaining predictive uncertainty with information theoretic shapley values. *Advances in Neural Information Processing Systems*, 36.

Zhou, L., Martínez-Plumed, F., Hernández-Orallo, J., Ferri, C., and Schellaert, W. (2022). Reject before you run: Small assessors anticipate big language models. In *EBeM IJCAI*.