AAAI-HUMAN.AI Symposium Nov 4-6, 2021: Preprint.

Collaborative Human-ML Decision Making Using Experts' Privileged Information Under Uncertainty

Mansoureh Maadi[1], Hadi Akbarzadeh Khorshidi[2], Uwe Aickelin[3]

The University of Melbourne[1,2,3]

mmaadi@student.unimelb.edu.au[1], hadi.khorshidi@unimelb.edu.au[2], uwe.aickelin@unimelb.edu.au[3]

Abstract

Machine Learning (ML) models have been widely applied for clinical decision making. However, in this critical decision making field, human decision making is still prevalent, because clinical experts are more skilled to work with unstructured data specially to deal with uncommon situations. In this paper, we use clinical experts' privileged information as an information source for clinical decision making besides information provided by ML models and introduce a collaborative human-ML decision making model. In the proposed model, two groups of decision makers including ML models and clinical experts collaborate to make a consensus decision. As decision making always comes with uncertainty, we present an interval modelling to capture uncertainty in the proposed collaborative model. For this purpose, clinical experts are asked to give their opinion as intervals, and we generate prediction intervals as the outputs of ML models. Using Interval Agreement Approach (IAA), as an aggregation function in our proposed collaborative model, pave the way to minimize loss of information through aggregating intervals to a fuzzy set. The proposed model not only can improve the accuracy and reliability of decision making, but also can be more interpretable especially when it comes to critical decisions. Experimental results on synthetic data shows the power of the proposed collaborative decision making model in some scenarios.

Introduction

Machine Learning (ML) has experienced a surge in recent years. They have been used to develop models for facilitating decision making processes in different areas. The development of these models is based on the idea that computers can process big data and make predictions whilst it is often hard for human experts. However, humans are more skilled to work with unstructured information and deal with uncommon situations. That is the reason that human decision making is still prevalent in many areas like clinical decision making, defence commanding, criminal punishment prediction, etc. In some of these areas like clinical decision making, ML models have been applied largely. However, the reliability of these models is always under question. It is reported that ML models are not enough in critical clinical decision making (Itani, Lecron, and Fortemps 2019), (Zerilli et al. 2019).

Human-in-the-loop ML models pave the way to implement human expertise in ML models. In these models, human experts can interact and collaborate in different stages of ML process to improve the performance of ML models. Clinical experts can collaborate in three stages of data producing and data processing (Huang et al. 2020), (Wrede, Hellander, and Wren 2019), ML modelling (Cai et al. 2019), and ML evaluation and refinement (Alahmari et al. 2019) to improve the performance of ML methods for clinical decision making (Maadi, Khorshidi, and Aickelin 2021). However, in human-in-the-loop ML approaches, ML models are principal decision makers and clinical experts can guide the models according to their expertise and experience.

ML models are generated using training data and used to predict the test samples. So, these models decide based on training data information. However, there are some information that are available at the training stage but not available for test data. This information called privileged (hidden) information (Vapnik, Vashist and Pavlovitch 2009), (Vapnik and Vashist 2009). In this study, we introduce another type of privileged information that is available at the testing stage but not available (recorded) for training. For example, at the time of diagnosing a patient's disease, clinical experts have an estimation about the diagnosis (with different level of confidence) based on their experience, patient's appearance and reviewing documents and test results. These estimations are not normally recorded so that they cannot be used in training models. However, they can be captured for each patient at the diagnosis (testing) stage. In this paper, we

propose a framework to capture experts' privileged information and integrate with trained ML models in a collaborative decision making.

As ML models and clinical experts use different sources of information to make decision, a consensus decision making approach can improve decision making. In this paper, unlike human-in-the-loop ML models, we introduce a collaborative human-ML decision making model where both clinical experts and ML models are decision makers. Clinical decisions should not be made individually. We believe clinical experts and ML models can help each other through a group decision making process. Collaborative decision making approach can provide a trust between ML models and clinical experts to have trustable and explainable models. In addition, this approach provides an opportunity to have a precise investigation on the ML models' and clinical experts' performance in clinical decision making process individually and jointly. Besides, this approach improves decision making especially when a decision maker is not available due to disruption in connectivity or information is provided intermittently for decision making.

Uncertainty is inevitable in decision making. Decision makers may have different levels of uncertainty based on their knowledge and level of access to information. When it comes to human decision makers, decision making accompany with two kinds of uncertainty named inter-expert uncertainty and intra-expert uncertainty. Inter-expert uncertainty shows the variation among different decision makers and intra-expert uncertainty is related to the change of the mind of each decision maker during the time on the same situation (Havens, Wagner, and Anderson 2017). To capture decision maker's opinion with its uncertainties, linguistic variables and computing with words paradigms have attracted the researchers recently (Khorshidi and Aickelin 2020). In these techniques, opinions as words encode to fuzzy sets (Borovička 2019), cloud models (Khorshidi and Aickelin 2020), intervals (Wu, Mendel, and Coupland 2012), to name a few, and computational analysis on them provides decisions. Expecting exact values from clinical decision makers is unrealistic. They should be given an opportunity to express their opinions with a level of uncertainty. So, in the proposed collaborative decision making approach, we capture the uncertainty using intervals. We ask each clinical expert to give their opinion as an interval and show the level of the uncertainty using the width of the interval.

In ML models, data uncertainty and model uncertainty are two important sources of uncertainty. To capture uncertainty of ML models by intervals, we recently have introduced an interval modelling technique to capture uncertainty in ensemble learning (Maadi, Aickelin, and Khorshidi 2020). In this technique, for each ML model in an ensemble, an interval is generated as a prediction. In the proposed collaborative decision making model, we use this technique to capture uncertainty of each ML model by intervals.

Interval Agreement Approach (IAA) is an aggregation method that generates fuzzy sets from interval-valued data to minimize the loss of information in aggregation process. (Havens, Wagner, and Anderson 2017; Khorshidi and Aickelin 2020). This approach is introduced by (Wagner et al. 2015). In the proposed collaborative model, we use IAA as the aggregation function to improve decision making through capturing more uncertainty and minimizing the loss of information.

Thus, in this paper, we make two important contributions: (1) we present a collaborative human-ML decision making model to use two important sources of information in clinical decision making from two groups of decision makers, ML models and clinical experts, and (2) we measure uncertainty of both decision maker groups through intervals and capture decision uncertainty in the collaborative model using interval modelling and IAA.

The structure of the paper is as follows. In the next section, we describe the technique to generate intervals as the outputs of ML models to capture ML models' uncertainty. Also, IAA is described in this section. Then, the proposed collaborative decision making model is explained. After that, the performance of the proposed model is investigated using a synthetic dataset. Finally, conclusions of the paper are presented.

Preliminary

Generating Uncertainty Intervals for ML Models

To capture uncertainty of clinical experts' opinion, we use interval data. When decision makers are ML models, we can capture the uncertainty of ML models using prediction intervals. Here, we describe how we can generate intervals as the prediction of ML models based on the approach we introduced in (Maadi, Aickelin, and Khorshidi 2020). Let C be an ML classifier such as decision tree or logistic regression. From a training dataset, we can generate different training datasets using bagging method. Bagging is a sampling strategy proposed by Breiman (Breiman 1996). In this method, some samples are elicited randomly from the training dataset with replacement and generate a new training dataset named as bag. By training the ML model on different bags, we have different classifiers. Applying them on the test dataset generates multiple probabilities related to class prediction. Using these probabilities, we can generate an uncertainty interval (*UI*) for the prediction of the ML model. *UI* captures the uncertainty of the ML model (Maadi, Aickelin, and Khorshidi 2020).

Suppose $p_1, p_2, \ldots, p_b$ as probabilities determined by b classifiers (generated using bags), an *UI* for ML model is generated by calculating the first quartile (Q_1) and the third quartile (Q_3) of the probabilities as (1).

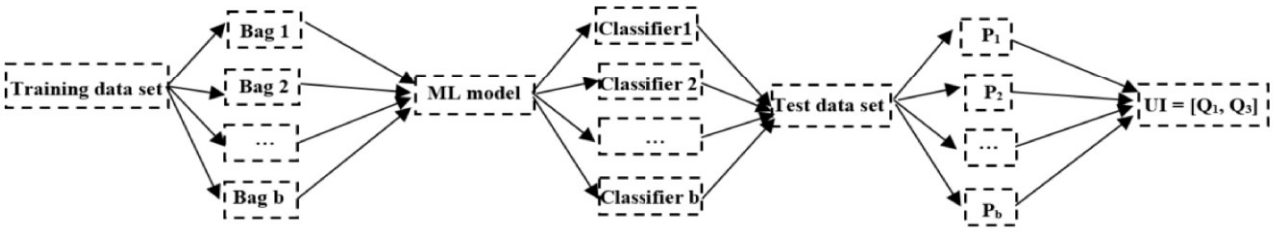

Figure 1: Generating an Uncertainty Interval for an ML Model on a Dataset to Capture Uncertainty

$$UI = [Q_1, Q_3] \quad (1)$$

According to (1), *UI* is defined as the inter quartile range of $p_1, p_2, \ldots, p_b$. In Figure 1, the framework of generating uncertainty intervals for ML models is shown.

Interval Agreement Approach (IAA)

IAA is an aggregation function to aggregate decision makers' opinions while the opinions are presented as intervals. IAA aggregates intervals to a fuzzy set. Let $A = \{A_1, \ldots, A_m\}$ be a set of intervals given by *m* decision makers as their opinions where $A_i = [l_i, r_i]$ *(i = 1, 2, ..., m)*. Aggregating intervals of set A generates a Type 1 Fuzzy Set (T1 FS) in IAA with the membership function of μ_A which is defined as (2) (Wagner et al. 2015).

$$\mu_A = \sum_{i=1}^{m} {}^{y_i} \Big/ \left(\bigcup_{j_1=1}^{m-i+1} \bigcup_{j_2=j_1+1}^{m-i+2} \ldots \bigcup_{j_i=j_{i-1}+1}^{m} (A_{j_1} \cap \ldots \cap A_{j_i}) \right) \quad (2)$$

In (2), $y_i = {}^{i}/_{m}$ is the degree of membership and '/' refers to assignment of degree of membership. The degree of membership in IAA is related to the number of intervals overlapped in a point. So, the value of one for the degree of membership in a point shows all intervals overlap at that point. To simplify (2), μ_A can be written as (3) for a point like x.

$$\mu_A(x) = \frac{\sum_{i=1}^{m} \mu_{\bar{A}_i}(x)}{m} \quad (3)$$

Where $\mu_{\bar{A}_i}(x) = \begin{cases} 1 & L_{\bar{A}_i} \le x \le r_{\bar{A}_i} \\ 0 & else \end{cases}$

Collaborative Human-ML Decision Making Model

We consider clinical decision making problem as a classification problem. Considering three classifiers (ML models) and three clinical experts as decision makers, the process of collaborative decision making is depicted in Figure 2. In this process, decision makers determine the probability that a test sample belongs to the main class and present it as an interval. For example, in a cancer diagnosis problem, decision makers determine the probability that the test sample is malignant.

In the proposed model, both classifiers and clinical experts have access to the electronic health records. The records are used to train classifiers. Using generating uncertainty intervals explained before, we have an interval as the output of each classifier. This interval shows the probability of belonging a test sample to the main class. Also, clinical experts are asked to give their opinions about the test sample as intervals. IAA aggregates all intervals to a T1 FS. This fuzzy set shows how much all decision makers are in agreement. A T1 FS can be shown as a list of tuples which each tuple indicates different region of change over the membership function as (4).

$$TFS = [R_1, R_2, \ldots, R_v],\ R_i = ([R_{il}, R_{ir}], R_{ih}) \quad (4)$$

Where l is the left point, r is the right point and h is the height or the membership function value of the tuple R_i calculated using (3).

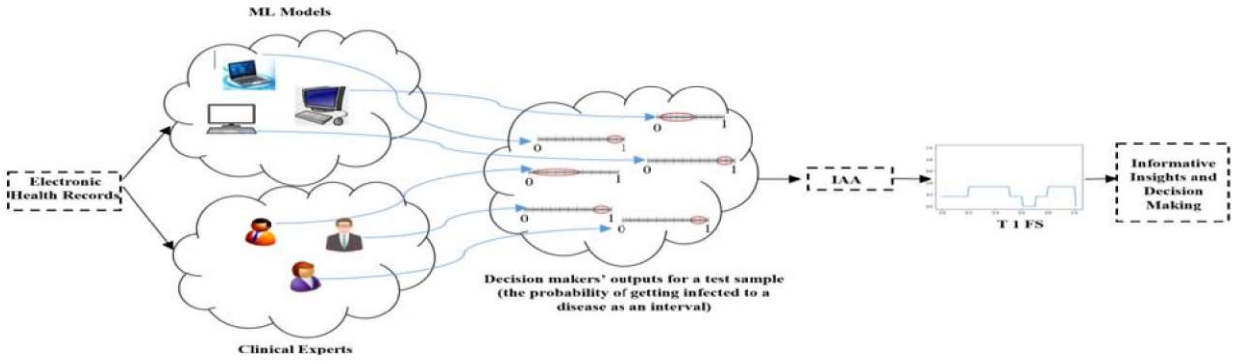

Figure 2: Collaborative Human-ML Decision Making Framework

To make the collaborative decision about a test sample, we calculate the centroid of this fuzzy set. The centroid of the fuzzy set is calculated using (5).

$$centroid\ (TFS) = \frac{\sum_{i=0}^{v}(R_{ih}\times R_{il})+(R_{ih}\times R_{ir})}{\sum_{i=o}^{v} 2(R_{ih})} \quad (5)$$

If the value of the centroid is more than 0.5, it shows the test sample belongs to the main class. For example, in the cancer diagnosis problem, if the centroid is more than 0.5, it shows the test sample is malignant.

In the proposed approach, accompany with making a collaborative decision, we can evaluate the performance of both groups of decision makers separately. Specially in high risk conditions, this provides us important information to make decision in a timely manner.

In the proposed collaborative decision making approach, we can calculate the width of the intervals presented by clinical experts and ML models as a measure for uncertainty. The width of intervals provides us information about how much certain clinical experts and ML models are in their decision making separately and collectively. So, we can recognize decision makers that are highly deviated from the consensus to extract informative insights for updating the decision-making process.

In Algorithm 1, the steps of the proposed collaborative decision making approach for *m* ML models (classifiers) and *n* clinical experts is described.

Algorithm 1: Collaborative expert-ML decision making process

Input:

1. Electronic health records
2. Test sample u
3. Collection $C_1, C_2, \ldots, C_m$ of classifiers
4. Clinical experts' opinions about the test sample u (*n* clinical experts)
5. The number of randomly picked samples in bagging (V)
6. The number of bags generated using bagging (H)

Output: The decision about getting infected to the disease for test sample u

1. For i from 1 to m do
2. For j from 1 to H do
3. Select a bag of training samples (V samples with replacement) using bagging
4. Train classifier C_i on the selected bag
5. Compute the probability for the given test sample u using the trained classifier and assign it to P_j
6. end
7. Compute the first quartile of $\{P_1, P_2, \ldots, P_H\}$ and assigns it to $Q1_i$
8. Compute the third quartile of $\{P_1, P_2, \ldots, P_H\}$ and assigns it to $Q3_i$
9. Determine the uncertainty interval $[Q1_i, Q3_i]$ for classifier C_i regarding test sample u
10. end
11. Ask the probability of belonging the test sample u to the main class as an interval from clinical expert j and named it $[DL_j, DR_j]$ $(j = 1, 2, \ldots, n)$
12. Aggregate $[Q1_1, Q3_1]$, $[Q1_2, Q3_2]$, …, $[Q1_m, Q3_m]$, $[DL_1, DR_1]$, $[DL_2, DR_2]$, …, $[DL_n, DR_n]$ using IAA and generate a T1 FS
13. Calculate the centroid of the T1 FS
14. If the centroid >= 0.5

    Return 'Test sample u belongs to the main class'

    Else

    Return 'Test sample u belongs to the subordinate class'

Experimental Analysis Using Synthetic Data

To show how the proposed collaborative approach works, we use Liver Disorders dataset from UCI as electronic health records (Dua 2019). This dataset has 345 samples and 5 features that are all blood tests which are related to liver disorders arise from alcohol consumption. In this dataset, target value is alcohol consumption and features' values are integer numbers. To have a bi-class classification dataset, we follow the strategy described in (Turney 1994). We assign class 0 to the number of drinks less than 3, and class 1 to the number of drinks equal or more than 3. This dataset is a balanced dataset.

In this experiment, three classifiers of Decision Tree (DT), Logistic Regression (LR) and Gaussian Naïve Bayes (GNB) are considered as ML models. We split the dataset to training dataset and test dataset with the ratios of 75% and 25% respectively. Regarding the proposed approach, we calculate uncertainty intervals for classifiers. For clinical experts' opinion, we generate synthetic data as their decision (the probability of assigning one sample to class 1, here as the main class) to examine different scenarios in collaborative decision making approach.

In the first scenario, we assume that both groups of decision makers (ML models and clinical experts) make similar decisions and have close uncertainties. So, we construct three synthetic intervals as clinical experts' opinions using three uncertainty intervals generated by classifiers. To generate one opinion interval from one uncertainty interval, we randomly select two numbers from the range of numbers between endpoints of the uncertainty interval added by an ε and these numbers are the endpoints of the opinion interval. Results of this scenario in three different experiments are shown in Table 1.

Experiments	Performance measure	Collaborative model	ML models	Clinical experts
1	Accuracy	0.790	**0.795**	0.784
	F-score	0.870	**0.873**	0.866
	G-mean	0.563	**0.564**	0.555
	Width of the intervals	0.156	0.151	0.161
2	Accuracy	**0.775**	0.770	0.770
	F-score	**0.859**	0.856	0.855
	G-mean	**0.560**	0.556	0.556
	Width of the intervals	0.150	0.144	0.155
3	Accuracy	**0.773**	0.771	**0.773**
	F-score	**0.858**	0.856	0.857
	G-mean	0.539	0.537	**0.553**
	Width of the intervals	0.154	0.149	0.158

Table 1: Collaborative and Individual Results for Two Groups of Decision Makers (ML Models and Clinical Experts) in Scenario 1

In Table 1, three experiments show possible situations in scenario 1 where each group of decision makers (collaborative model and individual models) has relatively better performance than others. However, the difference among the performance of these three groups of decision makers is not

significant. According to this table, the decision-making results in terms of accuracy, F-score and G-mean are similar to each other for three decision maker groups in all experiments. Also, in this scenario the widths of intervals are close together for all groups. We know that always the width of the intervals for collaborative model is between the width of the intervals for clinical experts and ML models. Totally, we can say that the performance of the collaborative model in the case that the performance of two individual groups is like each other is better than at least one of them.

In the second scenario, we examine the effect of the bad performance of one of decision makers' groups on collaborative decision making. We generate opinions' intervals so that the predictions for clinical experts are far from the real class of the test samples to some extent. In experiment 1 of this scenario, we generate intervals with wider width and in experiment 2, we generate intervals with narrower width. The results related to this scenario are shown in Table 2.

Experiments	Performance measure	Collaborative model	ML models	Clinical experts
1	Accuracy	0.766	**0.774**	0.333
	F-score	0.849	**0.859**	0.238
	G-mean	**0.602**	0.534	0.353
	Width of the intervals	0.327	0.154	0.500
2	Accuracy	0.773	**0.777**	0.336
	F-score	0.855	**0.861**	0.251
	G-mean	**0.639**	0.524	0.182
	Width of the intervals	0.121	0.153	0.09

Table 2: Collaborative and Individual Results for Two Groups of Decision Makers in Scenario 2

The results in Table 2 shows the power of interval modelling as well as IAA as interval aggregation function in decision making. In the cases that one group of decision makers makes bad decisions, results show it does not affect the collaborative decision significantly. So, the proposed model is robust and is suitable for critical collaborative clinical decision making. Considering Table 2, we conclude that when it comes to the width of the intervals as a measure for uncertainty, we can investigate the decision made by the group with high value of width of intervals. Decision making by this group of decision makers can affect decision quality. So, we can ignore them in decision making process.

In the third scenario, we investigate the effect of interval modelling in the proposed model on decision making performance. We believe that the proposed interval modelling improves decision making through capturing uncertainty. To test it, we compare the performance of the proposed collaborative model with its equivalent point prediction model. In this scenario, we use majority voting as the most common used aggregation function in point prediction approaches. To create point prediction model, for each ML model like DT, we create different classifiers using bagging, then we use majority voting to determine the class label of each test sample using classifiers. To create synthetic point prediction data as experts' opinions, we use the strategy mentioned in previous scenarios to generate opinions intervals. Then, we consider the middle points of the intervals as clinical experts' point predictions to determine the class label of test samples. Finally, using majority voting on point predictions generated by ML models and clinical experts, we determine consensus point prediction for test samples. The results on comparing the proposed collaborative model to its equivalent point prediction model are shown in Table 3. In the first experiment, the opinion intervals are generated according to scenario 1 and in the second experiment opinion intervals are generated according to scenario 2.

Experiments	Performance measure	Proposed model	Point prediction model
1	Accuracy	**0.780**	0.774
	F-score	**0.862**	0.858
	G-mean	**0.555**	**0.555**
2	Accuracy	**0.766**	0.715
	F-score	**0.849**	0.813
	G-mean	**0.602**	**0.602**

Table 3: Proposed Collaborative Model Compared to Its Equivalent Point Prediction Model in Scenario 3

The results in Table 3 show better performance of the proposed model than its equivalent point prediction model in terms of accuracy and F-score. It means that interval modelling in the proposed collaborative model improves decision making through capturing uncertainty.

Regarding all described scenarios and experiments, we conclude that the proposed collaborative model can be an effective model to capture uncertainty, use clinical experts' privileged information and implement the power of ML models in clinical decision making. However, real datasets and scenarios can present more accurate analysis on the performance of the proposed collaborative human-ML decision making model.

Conclusion

In this paper, we use expert's privileged information in addition to ML models in clinical decision making and present a collaborative human-ML decision making model. In the proposed model, both clinical experts and ML models are considered as decision makers. To handle the uncertainty of decision making in collaborative model, we use intervals. Clinical experts' opinions are asked as intervals, and we develop an approach to have intervals as the outputs of ML models. IAA as a powerful interval-based aggregation function is used to aggregate decision makers' interval predictions to a T1 FS. The generated FS is used to determine the final consensus decision. In the proposed collaborative model, the width of the intervals is a measure for decision makers' uncertainty. So, it provides information about the uncertainty of each group of decision makers to improve decision making. To show how the proposed model works, we

consider different scenarios and experiments using synthetic data and test the performance of the proposed model. Results show the power of intervals and interval modelling in the proposed collaborative model to capture uncertainty and making more effective and robust decision in clinical decision making. Also, as a significant result, we observe that weak performance of a group of decision makers does not affect the collaborative decision in the proposed model significantly. For the future work, we are collecting real data to examine our proposed model in different scenarios. Also, we will develop the proposed method for multi-class classification decision making problems.

References

Alahmari, S.; Goldgof, D.; Hall, L.; Dave, P.; Phoulady, A.H.; and Mouton, P. 2019. Iterative Deep Learning Based Unbiased Stereology with Human-in-the-Loop. In *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA*, 665–670. Orlando, FL, USA.

Borovička, A. 2019. New Approach for Estimation of Criteria Weights Based on a Linguistic Evaluation. *Expert Systems with Applications* 125 (July): 100–111.

Breiman, L. 1996. Bagging Predictors. *Machine Learning* 24 (2): 123–40.

Cai, C.; Reif , E.; Hegde, N.; Hipp, J.; Kim, B.; Smilkov, D.; Wattenberg, M.; Viegas, F.; Corrado, G.; Stumpe, M.; and Terry, M. 2019. Human-Centered Tools for Coping with Imperfect Algorithms during Medical Decision-Making. In *Conference on Human Factors in Computing Systems, CHI*, 1–14. Glasgow, Scotland, UK.

Dua, D. and Graff, C. 2019. UCI Machine Learning Repository: Citation Policy. Irvine, CA: University of California, School of Information and Computer Science.

Havens, T.; Wagner, C.; and Anderson, D. 2017. Efficient Modeling and Representation of Agreement in Interval-Valued Data. In *IEEE International Conference on Fuzzy Systems,*1-6. https://doi.org/10.1109/FUZZ-IEEE.2017.8015466.

Huang, Q.; Chen Y.; Liu, L.; Tao, D.; and Li, X. 2020. On Combining Biclustering Mining and AdaBoost for Breast Tumor Classification. *IEEE Transactions on Knowledge and Data Engineering* 32 (4): 728–38.

Itani, S.; Lecron, F.; and Fortemps, P. 2019. Specifics of Medical Data Mining for Diagnosis Aid: A Survey. *Expert Systems with Applications*118(March):300–314.

Khorshidi, H.A.; and Aickelin, U. 2020. Multicriteria Group Decision-Making under Uncertainty Using Interval Data and Cloud Models. *Journal of the Operational Research Society* 1-15.

Maadi, M.; Aickelin, U.; and Khorshidi, H.A. 2020. An Interval-Based Aggregation Approach Based on Bagging and Interval Agreement Approach in Ensemble Learning. In *IEEE Symposium Series on Computational Intelligence, SSCI 2020*, 692–99. Canberra, Australia.

Maadi, M.; Khorshidi, H.A.; and Aickelin, U. 2021. A Review on Human–AI Interaction in Machine Learning and Insights for Medical Applications. *International Journal of Environmental Research and Public Health* 18 (4): 1–27.

Turney, P.; 1994. Cost-Sensitive Classification: Empirical Evaluation of a Hybrid Genetic Decision Tree Induction Algorithm. *Journal of Artificial Intelligence Research* 2: 369-409.

Vapnik, V.; and Vashist, A.; 2009, A New Learning Paradigm, Learning using privileged information, Neural Networks, 22 (5-6): 544-557.

Vapnik, V.; Vashist, A.; and Pavlovitch, N. 2009. Learning using Hidden Information (Learning with Teacher). In *the international Joint conference on Neural Networks,* 3188-3195

Wagner, C.; Miller, S.; Garibaldi, J.; Anderson, D.; and Havens, C. 2015. From Interval-Valued Data to General Type-2 Fuzzy Sets. *IEEE Transactions on Fuzzy Systems* 23 (2): 248–69.

Wrede, F.; Hellander, A.; and Wren, J. 2019. Smart Computational Exploration of Stochastic Gene Regulatory Network Models Using Human-in-the-Loop Semi-Supervised Learning. *Bioinformatics* 35 (24): 5199–5206.

Wu, D.; Mendel, J.; and Coupland, S. 2012. Enhanced Interval Approach for Encoding Words into Interval Type-2 Fuzzy Sets and Its Convergence Analysis. *IEEE Transactions on Fuzzy Systems* 20 (3): 499–513.

Zerilli, J.; Knott, A.; Maclaurin, J.; and Gavaghan, C. 2019. Algorithmic Decision-Making and the Control Problem. *Minds and Machines* 29 (4): 555–78.