Proceedings of the 35th European Safety and Reliability & the 33rd Society for Risk Analysis Europe Conference
Edited by Eirik Bjorheim Abrahamsen, Terje Aven, Frederic Bouder, Roger Flage, Marja Ylönen
©2025 ESREL SRA-E 2025 Organizers. *Published by* Research Publishing, Singapore.
doi: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P6499-cd

Increasing Confidence in AI Models by Explaining Uncertainty in Predictions

K. Darshana Abeyrathna

Group Research and Development, DNV, Høvik, Norway. E-mail: darshana.abeyrathna.kuruge@dnv.com

Andreas Hafver

Group Research and Development, DNV, Høvik, Norway. E-mail: andreas.hafver@dnv.com

A major challenge in AI is that models are sometimes confidently wrong, which can have severe consequences in critical decision-making. One way to address this issue is through interpretable models or explainability methods that provide reasons for predictions. These reasons can be scrutinized by humans to determine trust in the model; however, explanations can be convincing yet incorrect. Another approach is uncertainty quantification, which provides a measure of confidence in predictions. However, uncertainty alone is of limited value unless we understand its basis.

In this paper, we recognize that explanations of predictions and confidence measures are useful for decision-makers. However, we hypothesize that decision-makers could benefit even more from explanations of uncertainty. This paper introduces an approach based on the Tsetlin Machine that provides predictions, confidence measures, and explanations for both predictions and their uncertainty to assess how confidence explanations add value. Additionally, we propose incorporating uncertainty explanations with "human-in-the-loop" feedback in a continuous cycle to improve the model. This approach enhances both the technical and practical aspects of AI, making it more reliable and trustworthy in high-stakes applications such as healthcare, energy, transport, and finance. Using real-world data, we explore the importance of local interpretability—ensuring decision-makers gain relevant insights for individual predictions and uncertainty—and global interpretability, which provides a comprehensive understanding of the model's decision process. This global understanding, enriched by expert feedback, enables further model refinement.

Keywords: Uncertainty Quantification, Interpretable AI, Explainable AI, XAI, Trust in AI, Tsetlin Machines.

1. Introduction

The rapid advancement of AI has transformed healthcare, finance, transportation, and energy, aiding in diagnosis, market prediction, traffic optimization, and energy management. However, AI models often function as ”black boxes,” lacking transparency. This opacity hinders trust and adoption, especially in high-stakes scenarios where errors can have severe consequence (Kaminski, 2021).

AI models, particularly deep neural networks, have achieved remarkable accuracy in various tasks, yet they can still produce incorrect predictions with a high degree of confidence (Nguyen et al., 2015). This phenomenon underscores a critical challenge in AI: ensuring that model predictions are not only accurate but also trustworthy. Models that make confident but wrong decisions can erode trust and reduce the willingness of users to rely on automated systems (Ovadia et al., 2019).

To address the challenge of trust in AI, the field has seen a growing interest in interpretable and explainable AI (XAI). XAI aims to render AI decisions more transparent by elucidating the logic behind model predictions. By providing comprehensible reasons for decisions, stakeholders can better understand the model's strengths and limitations (Molnar, 2020). This understanding is crucial for high-stakes decisions where the rationale for a prediction is as important as the prediction itself.

While interpretability helps, it's not enough on its own. Explanations can be convincing but still wrong, so it's important to also measure a model's confidence (Lakkaraju et al., 2016). Uncertainty Quantification (UQ) tackles this by showing how confident a model is in its predictions, helping spot cases where it lacks enough training data or might be unreliable (Hüllermeier and Waegeman, 2021).

Although UQ offers valuable insight, a static measure of uncertainty is limited in its utility. To fully exploit the benefits of UQ, decision-makers must understand the basis of the uncertainty. In most existing AI models, explanations are provided for predictions but not for the uncertainties associated with them (Pearce et al., 2020). This gap leaves decision-makers with half the picture: they know the model is uncertain but lack insights into why.

This paper hypothesizes that by providing explanations for both predictions and the uncertainties associated with them, decision-makers are better equipped to make informed decisions. We propose an innovative approach that leverages the Tsetlin Machine, a transparent, logic-based learning model Granmo (2018), to explain not only model predictions but also the uncertainties inherent to them. By categorizing predictions into four groups—confidently correct, unconfidently correct, confidently wrong, and unconfidently wrong—we explore how a comprehensive explanation of model outputs and their uncertainties can significantly enhance the quality of decision-making.

Furthermore, we expand the approach by integrating “human-in-the-loop” feedback, creating a continuous improvement cycle for the model. This feedback loop not only heightens AI reliability but also promotes an iterative dialogue between the model and domain experts, ultimately refining the decision-making process (Bansal et al., 2021). We demonstrate the effectiveness of our methodology through empirical evaluations using real-world data, emphasizing both local and global interpretability, and underscore its potential impact across various critical sectors.

2. Tsetlin Machines

The basic Tsetlin Machine (TM), discussed in this section, is the foundation for other variants of TMs, including the Probabilistic TM. TMs consist of clauses that capture patterns in data in the form of conjunctions of input binary variables or their negation (together called literals). The number of clauses, m, is set by the user and affects the learning process.

Each clause, c_j, is defined as:

$$c_j = 1 \wedge \left(\bigwedge_{k \in I_j^I} x_k \right) \wedge \left(\bigwedge_{k \in \bar{I}_j^I} \neg x_k \right),$$

where x_k and $\neg x_k$ refer to the literals chosen for clause j. The set I_j^I stores the indices of non-negated literals, while $\bar{I}_j^I$ stores the indices of negated literals.

Inclusion or exclusion of literals in clauses are decided by Tsetlin Automata (TAs), with $2N$ memory states, correspond to each literal in each clause. The clause outputs 1 if all included literals are true; otherwise, it outputs 0. The clauses are divided into two groups. Clauses with odd indices are given positive polarity, c^+ are responsible for learning patterns for class 1, while even-indexed clauses are assigned negative polarity c^- and learn the patterns for class 0. The final output of the TM is based on the majority of the clause outputs.

Learning in a TM involves guiding TAs in clauses to correctly classify inputs using two types of feedback: Type I and Type II (Granmo, 2018). Type I feedback reinforces the patterns learned when clauses output 1 when they should output 1, while erasing incorrectly recognized patterns. Type II feedback combats false positive clause outputs by systematically turning the clause output from 1 to 0.

3. The Probabilistic Tsetlin Machine

The Probabilistic Tsetlin Machine (PTM) (Darshana Abeyrathna et al., 2024) extends the standard TM by modeling the states of the TAs as probability distributions rather than fixed states. States of TA which represents k^{th} literal in j^{th} clause in the PTM are represented by state probability vector $SPV_{j,k}$, $SPV_{j,k} \in [0,1]^{2N}$, representing the likelihood of the automaton being in any of its $2N$ possible states. This allows the PTM to more flexibly update its knowledge.

The clause output in the PTM is calculated in the same way as in the TM, but with the states of the TAs sampled from their SPVs. Type I and Type II feedback are used in the same manner as in the TM, but now, instead of updating states directly,

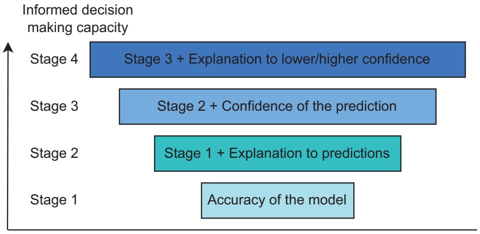
Informed decision making capacity
Stage 4
Stage 3 + Explanation to lower/higher confidence
Stage 3
Stage 2 + Confidence of the prediction
Stage 2
Stage 1 + Explanation to predictions
Stage 1
Accuracy of the model

Fig. 1. Increasing the information for the decision maker in order to make informed decisions.

feedback is incorporated into Transition Probability Matrices (TPMs), which update the SPVs. This process eliminates the need for explicit state transitions and simplifies the learning process.

The PTM is also probabilistic during inference. When making predictions on new data, the states of the TAs are sampled from their SPVs, and the clause outputs are determined based on these sampled states. This introduces variability in the output, similar to Bayesian Neural Networks, where multiple predictions may be generated for the same input sample.

The PTM thus enhances the TM by providing a probabilistic framework that allows for better handling of uncertainty during both training and inference, making it more suitable for tasks where uncertainty quantification is important.

4. Methodology

The primary objective of this work is to enhance the information available to decision-makers beyond traditional approaches. Model performance, commonly evaluated on a validation set in terms of validation accuracy, is a widely used metric to assess the reliability of an AI model. With advancements in explainable and interpretable AI, decision-makers are now provided with insights into model behavior, either at a global level (global interpretability/explainability) or for specific predictions (local interpretability/explainability). Providing reasons for predictions in a comprehensible format has been shown to improve trust in decision-making processes.

Uncertainty quantification (UQ) is another critical aspect, offering a measure of the model's confidence in its individual predictions. When combined with validation accuracy and interpretability, uncertainty measures further enhance trust in the decision-making process.

This work proposes an additional layer of information by offering explanations for the observed levels of uncertainty (low or high) in predictions. By providing decision-makers not only with the prediction, the rationale behind the prediction, and the model's confidence, but also the reasons for varying levels of confidence, this approach aims to significantly improve decision-making capacity, flexibility, and trust in AI systems. We try to illustrate this in Fig. 1.

The accuracy of AI models, the explainable and interpretable AI, and the UQ are well-studied areas of research. This paper explains how the reasons for different levels of uncertainties can be obtained with TMs.

In this study, a TM and a PTM are trained in parallel. Specifically, the PTM updates its state probability vectors using the same feedback provided during the TM training. This approach enables the extraction of rules from the trained TM for direct classification while obtaining uncertainty measures from the trained PTM. Using PTM, we predict each test sample a predefined number of times, d. In a binary classification scenario, if we divide the number of times a specific test sample outputs 1 by d, we get the probability that that sample is classified into class 1, $p(\hat{y} = 1)$. This probability can also be used to measure the prediction entropy as follows:

$$H = -\sum_{i=0}^{1} p(\hat{y} = i) \cdot \log_2(p(\hat{y} = i)) \quad (1)$$

where $p(\hat{y} = 0) = 1 - p(\hat{y} = 1)$.

During inference, the probability of assigning each sample to class 1, $p(\hat{y} = 1)$, is computed. Based on a user-defined threshold q, these probabilities are used to categorize samples into high-uncertainty and low-uncertainty classes. This process results in a new dataset that retains the original features but incorporates these uncertainty-based labels. The newly labeled dataset is subsequently used to train another TM, which learns the patterns associated with high and low uncertainties. The complete process is illustrated in Fig. 2.

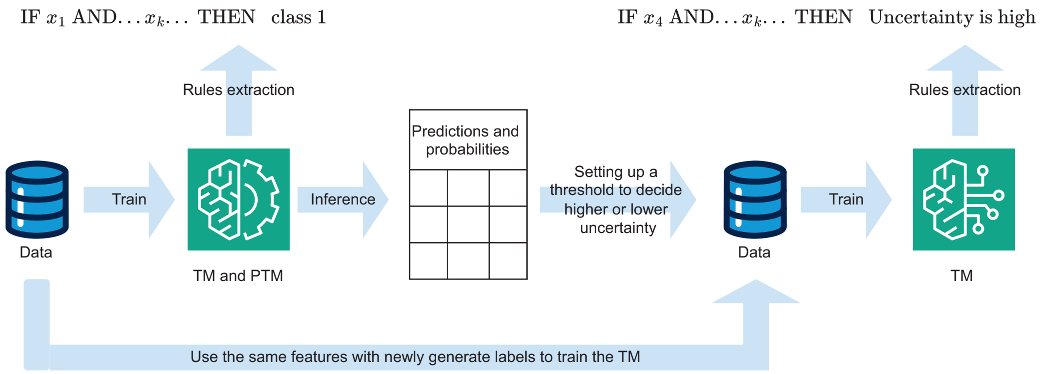
IF x_1 AND... x_k... THEN class 1
IF x_4 AND... x_k... THEN Uncertainty is high
Rules extraction
Rules extraction
Predictions and probabilities
Train
Inference
Setting up a threshold to decide higher or lower uncertainty
Train
Data
TM and PTM
Data
TM
Use the same features with newly generate labels to train the TM

Fig. 2. The proposed approach to explain uncertainties.

The information related to Stage 1, Stage 2, and Stage 3 in Fig. 1 can be obtained after the first round of training and inference. From the second TM trained on the newly generated dataset, we learn the reasons behind the higher or lower uncertainties associated with individual samples needed in Stage 4.

5. Experiments, Results, and Discussion

In this section, we present the experiments conducted to evaluate the proposed approach, analyze the results to assess its effectiveness, and discuss the implications of the findings.

5.1. *Experiments*

In the context of finance, accurately predicting bankruptcy is crucial for mitigating economic losses. For this reason, interpretable machine learning algorithms are often preferred over black-box methods to enhance transparency and trust in predictions (Kim and Han, 2003).

The bankruptcy dataset, which contains historical records of 250 companies, is used to evaluate our approach. Each record includes six categorical features relevant to bankruptcy prediction: **Industrial Risk**, **Management Risk**, **Financial Flexibility**, **Credibility**, **Competitiveness**, and **Operation Risk**. Each feature is classified into one of three states: Negative (N), Average (A), or Positive (P). The target variable consists of two classes: **Bankruptcy** and **Non-bankruptcy**. For this study, we focus on the three most influential features—**Management Risk**, **Financial Flexibility**, and **Competitiveness**—selected through a secondary analysis. To align with our method, the ternary features are binarized using the thresholding approach outlined in (Darshana Abeyrathna et al., 2019), resulting in a dataset with 9 binary features used for prediction.

A TM with merely four clauses is constructed to classify companies into the **Bankruptcy** and **Non-bankruptcy** classes. Each TA within each clause is configured with 100 memory states per action ($N = 100$). The Probabilistic TM, which learns the state probabilities in parallel, is of the same size as the regular TM. The precision parameter, s, and the target parameter, T, are set to 2.

At the end of training, the PTM predicts each sample 100 times ($d = 100$), providing the opportunity to measure both the probability of assigning each sample to class 1 and the corresponding prediction entropy. In this experiment, we set the threshold q to 0.8, which defines the level of uncertainty in classification. Specifically:

- Samples classified into class 1 with a probability $\geq$ 0.8 are considered **low-uncertainty** classifications for class 1.
- For samples classified into class 0, we use the threshold $1 - q = 0.2$. That is, if the probability of a sample belonging to class 1 is $\leq$ 0.2, we safely classify the sample as class 0, marking it as a **low-uncertainty** classification for class 0.
- All samples with a probability of being in class 1 between 0.2 and 0.8 are categorized as **high-uncertainty** samples.

The newly generated dataset retains the original features but now includes updated labels. These labels correspond to the two new classes: **low-uncertainty** and **high-uncertainty**. We train a second TM to identify the factors associated with these new classes. For this task, a TM similar to the one used previously is deployed.

5.2. *Results*

We organize the results according to the different stages illustrated in Fig. 1.

Stage 1: In the first stage, we evaluate the classification accuracy of the TM. Despite using merely four clauses, the TM achieves an accuracy of 98.4%.

Stage 2: The second stage focuses on providing explanations for the predictions. To accomplish this with the TM, we analyze the patterns learned by the clauses for their respective classes. At the end of training, the clauses have converged to the following patterns:

- Clause 1: Always outputs 0.
- Clause 2: Competitiveness is Average.
- Clause 3: Management Risk is **NOT** Positive **AND** Financial Flexibility is **NOT** Positive **AND** Competitiveness is **NOT** Positive.
- Clause 4: Competitiveness is Positive.

As we discussed in Section 2, Clause 1 and Clause 3 learn patterns for class 1 (**Bankruptcy**), while Clause 2 and Clause 4 learn patterns for class 0 (**Non-bankruptcy**). After a careful analysis of these patterns, we can derive a single global rule that the TM has effectively learned for its classifications, as follows:

$$\text{Class} = \begin{cases} \text{Bankruptcy} & \textbf{IF} \text{ Competitiveness} \\ & \text{is Negative} \\ \text{Non-bankruptcy} & \textbf{OTHERWISE}. \end{cases} \tag{2}$$

Stage 3: Now, it is time to measure the confidence of the predictions. First, we analyze the predictions of the training samples. We measure how many of the predictions are confidently correct, unconfidently correct, confidently wrong, and unconfidently wrong, as we know the correct labels and also using the threshold $q = 0.8$. The results are summarized in Table 1.

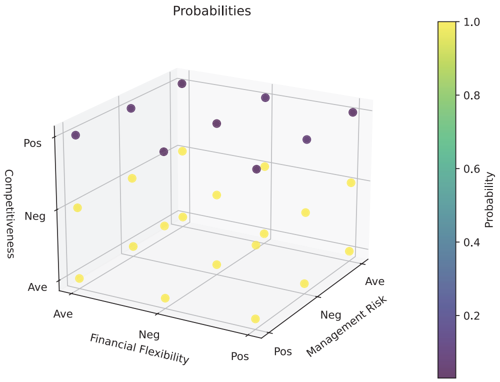

Fig. 3. Prediction probabilities of different validation points.

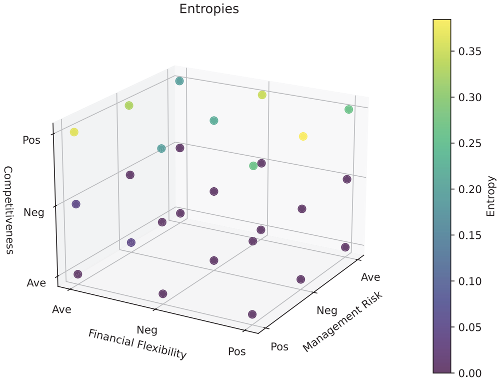

Fig. 4. Prediction Entropies of different validation points.

Additionally, we created a new set of tests without labels, covering all possible test cases that can be generated using the three selected features, each with three categories of values. The probabilities and entropies associated with classifying all 27 samples into the **Bankruptcy** class were measured. Fig. 3 and Fig. 4 display these probabilities and entropies, respectively. These measurements can be used to determine which class label should be predicted for each test case and with what level of confidence.

Stage 4: This stage is dedicated to uncovering the factors contributing to higher and lower uncertainties. The final TM trained at this stage was fed with 198 samples from the low-uncertainty

Table 1. Predictions and their confidence captured with the PTM.

Group	Percentage	Actual class	
		Bankruptcy	Non-Bankruptcy
Confidently correct	78.4%	107	89
Unconfidently correct	0%	0	0
Confidently wrong	0.8%	0	2
Unconfidently wrong	20.8%	0	52

category (class 0) and 52 samples from the high-uncertainty category (class 1). After training, the TM achieved a training accuracy of 98.4%.

To gain insights into the patterns associated with lower and higher uncertainties, we analyze the clauses learned by the TM:

- Clause 1: Financial Flexibility is Positive **AND** Competitiveness is Average.
- Clause 2: Competitiveness is **NOT** Average.
- Clause 3: Financial Flexibility is Average **AND** Competitiveness is Average.
- Clause 4: Competitiveness is **NOT** Average.

Through a detailed analysis of the patterns learned by the aforementioned clauses, we derive the following global rule for classifying samples into low-uncertainty and high-uncertainty categories:

$$\text{Uncertainty} = \begin{cases} \text{Low} & \textbf{IF } \text{Competitiveness is} \\ & \textbf{NOT } \text{Average} \\ \text{High} & \textbf{OTHERWISE}. \end{cases} \tag{3}$$

These findings are elaborated upon in the subsequent section.

5.3. *Discussion*

To ensure reliable classification explanations, it is crucial to calibrate the machine learning model during Stage 1. In our study, the TM achieves a robust accuracy of 98.4% at this stage. Provided with this measure of accuracy, the decision maker can trust the explanation obtained at Stage 2.

At Stage 3, the results summarized in Table 1 reveal that the model demonstrates high confidence when its predictions are correct. Uncertainty arises primarily in instances of incorrect predictions. However, there are only two cases where the TM confidently predicts Non-Bankruptcy when the correct label should have been Bankruptcy.

Interestingly, the predictions that can be made from the probabilities in Fig. 3 do not align with the rule found in Eq. (2). This is more evident from the plot, which shows that when 'Competitiveness is Average **OR** Negative, **THEN** it is a Bankruptcy'. This discrepancy arises due to the way TAs in TM and PTM learn their include and exclude actions differently. At the end of training, the clauses in PTM output 1 as follows:

- Clause 1 outputs 1 around 99% of the time.
- Clause 2 outputs 1 when Competitiveness is Average and still outputs 1 around 98% of the time when Competitiveness is **NOT** Average.
- Clause 3 outputs 1 when Competitiveness is Negative.
- Clause 4 outputs 1 when Competitiveness is Positive.

From this, it is evident that when Competitiveness is Average, the Bankruptcy class is outputted most of the time, while the Non-bankruptcy class is outputted only occasionally, thereby behaving differently from the rule in Eq. (2).

This is already a strong indication that explanations of predictions and uncertainties alone do not provide a satisfactory level of clarity regarding the predictions, rather they could create confusions. Therefore, we proceed to analyze the reasons behind the uncertainty. We already observed that both TM and PTM agree on what to output when the Competitiveness is Negative (Bankruptcy) and when the Competitiveness is Positive (Non-Bankruptcy). Hence the difference resides on Competitiveness is Average. This has

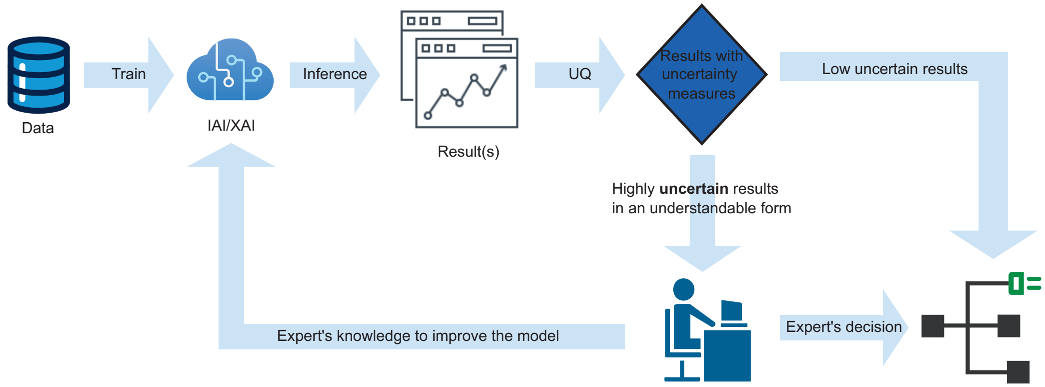
Data
Train
IAI/XAI
Inference
Result(s)
UQ
Results with uncertainty measures
Low uncertain results
Highly **uncertain** results in an understandable form
Expert's knowledge to improve the model
Expert's decision

Fig. 5. Human-in-the-loop framework to improve the decision making and model performance.

been learned by the second TM as summarized in Eq. (3).

To verify the above reasoning, we further examine the dataset. Here, we observe that when Competitiveness is Negative, all samples belong to the Bankruptcy class, and when Competitiveness is Positive, all samples belong to the Non-Bankruptcy class. However, when Competitiveness is Average, 52 samples belong to the Non-Bankruptcy class, while only 4 samples belong to the Bankruptcy class. Given this distribution, it is reasonable for the first TM to classify all samples into the Non-Bankruptcy class when Competitiveness is Average, as this occurs 92.8% of the time in the training data. More importantly, this confusion is recognized by the PTM in its rule in Eq. (3), even though it is very easy to overlook. This can be further connected to the numbers in Table 1, where the 52 Unconfidently wrong classifications of the Non-Bankruptcy class correspond to the samples where Competitiveness is Average.

For this dataset, uncertain classifications have mid-range probabilities, while confident ones are distinctly high or low. Thus, varying q does not impact results or decision rules, though exploring different thresholds on other datasets could offer insights.

5.4. *A Framework to Further Enhance the Performance*

In this section, we explain how the information we have gathered so far can be used to improve decision-making and enhance the model's performance. Specifically, we propose incorporating a human-in-the-loop (referred to as the decision maker or domain expert) for decision-making, as illustrated in Fig. 5. However, the workload of the decision-making in this framework is significantly reduced by involving them only in uncertain predictions, while certain predictions are directly forwarded to the decision gate. The decision maker is provided with reasoning for the classification, the confidence of individual predictions (measured as probabilities or entropies), and explanations for higher or lower confidence levels.

For example, in the context of the above application, the decision maker only needs to review 20.8% of the samples. When presented with a sample of high uncertainty, the decision maker is given the decision rule Eq. (2), the probabilities for classification into each class, and the reasoning for the high uncertainty. In this case, we expect the classification probability to be between 0.2 and 0.8, with the reason for higher uncertainty being "Competitiveness is Average". Using their expertise and domain knowledge, the decision maker can then decide whether the sample should be classified as Bankruptcy or Non-Bankruptcy when Competitiveness is Average. This approach is expected to yield more accurate decisions when Competitiveness is Average compared to relying solely on the class proportions in the training data.

Additionally, we propose leveraging the expert's knowledge to enhance the model's classification accuracy. To test this, we assume the expert's classification for samples with "Competitiveness is Average" is Non-Bankruptcy. Based on this assumption, we augment the training set with 500 additional samples by manually correcting the labels whenever a sample with "Competitiveness is Average" was originally labeled as Bankruptcy. With this updated training set, the TM achieves a prediction accuracy of nearly 99.5%. This approach also significantly reduces the number of uncertain classifications.

6. Conclusion

This paper introduces a novel approach for providing explanations of uncertainty in AI predictions using Tsetlin Machines. By combining prediction explanations with uncertainty quantification and its explanations, we offer a tool for enhancing decision-making in high-stakes scenarios. In the broader context of safety and reliability science, our idea of explaining uncertainty could be a way to improve the reliability of AI enable systems, because it provides a basis for deciding when to trust such systems or not. Our experiments on the Bankruptcy dataset validate the potential of the proposed methodology for applications requiring high reliability and interpretability. Future work is required to refine the integration of uncertainty explanations with human-in-the-loop systems.

References

Bansal, G., T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. Weld (2021). Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1–16.

Darshana Abeyrathna, K., S. El Mekkaoui, A. Hafver, and C. Agrell (2024). The Probabilistic Tsetlin Machine: A Novel Approach to Uncertainty Quantification. In *2024 The 8th International Conference on Advances in Artificial Intelligence*. ACM.

Darshana Abeyrathna, K., O.-C. Granmo, X. Zhang, and M. Goodwin (2019). A scheme for continuous input to the Tsetlin Machine with applications to forecasting disease outbreaks. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pp. 564–578. Springer.

Granmo, O.-C. (2018). The Tsetlin Machine - A Game Theoretic Bandit Driven Approach to Optimal Pattern Recognition with Propositional Logic. *arXiv:1804.01508*.

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

Kaminski, M. E. (2021). The right to explanation, explained. In *Research Handbook on Information Law and Governance*, pp. 278–299. Edward Elgar Publishing.

Kim, M.-J. and I. Han (2003). The discovery of experts' decision rules from qualitative bankruptcy data using genetic algorithms. *Expert Systems with Applications 25*(4), 637–646.

Lakkaraju, H., S. H. Bach, and J. Leskovec (2016). Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1675–1684.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Nguyen, A., J. Yosinski, and J. Clune (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Ovadia, Y., E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems 32*.

Pearce, T., F. Leibfried, and A. Brintrup (2020). Uncertainty in neural networks: Approximately bayesian ensembling. In *International conference on artificial intelligence and statistics*, pp. 234–244. PMLR.