Supplementary Material to paper "Interactive Visual Exploration of Rule-Based Model Logic" published in *IEEE Transactions on Visualization and Computer Graphics*, 2026 http://dx.doi.org/10.1109/TVCG.2026.3689736: Expert Evaluation of RuleSense

Natalia Andrienko, Gennady Andrienko, and Bahavathy Kathirgamanathan

The evaluation of RuleSense [1] followed a multi-study approach designed to assess the utility of visual analytics techniques for model auditing and domain-driven sensemaking. To evaluate the potential of RuleSense for logic-centered model assessment, we conducted two complementary expert studies. Rather than aiming at usability testing or performance benchmarking, both studies focused on assessing whether the proposed visual analytics techniques support meaningful reasoning about model logic, trustworthiness, and domain alignment.

1 Model Logic Auditing by ML Experts

This study investigated whether RuleSense supports expert reasoning about the internal logic of a trained model. The study used the vessel movement behavior classification model as a representative large rule-based model derived from a Random Forest.

Procedure. Participants were invited via email and provided with a richly illustrated slide deck (42 slides) demonstrating RuleSense capabilities using the vessel behavior model. The slides presented the main visual representations and analytical workflows, followed by structured questions focusing on logic auditing, anomaly detection, and trust calibration.

Invitation to ML Experts

The following letter was sent to multiple machine learning experts.

Subject: Need your expert help for paper revising

> Dear [. . .],
>
> We hope everything is going well with you.
>
> We are writing to ask for a bit of help. We're currently revising a paper on "RuleSense" — a set of techniques we've been working on to help audit complex ML models and figure out if they can actually be trusted. The key requirement of the paper reviewers has been to conduct an expert evaluation.

Since you're an expert in ML and its practical applications, your perspective would be incredibly valuable to us. We would like you to judge if our techniques actually appear useful for real-world auditing.

What is the goal?
We aren't looking for a usability study or a performance benchmark. Instead, we want to evaluate the potential of RuleSense techniques to support:

- **Auditing model logic**: Does it actually help you see how a model makes decisions?
- **Catching weird behaviour**: Is it good for spotting when a model is being unreliable or totally implausible?
- **Trust calibration**: Does it help you decide when (and how much) to trust a trained model?

What are we asking of you?
We've prepared a slide deck to make this as efficient as possible for you:

- **Slides 1–34**: These provide the background and walk through the tools using an example (analysing a model trained to recognise vessel movement behaviours).
- **Slides 37–43**: These contain specific questions for each tool and an overall assessment.

How to provide feedback:
You don't need to provide a formal answer to every single question on the slides. We've included them primarily to orient you toward the themes we're interested in, specifically *Sensemaking* (understanding logic), *Assessment* (judging risk), and *Actionability* (making decisions about the model).

We will be happy if you find suitable time to share your thoughts with us in oral or written form, whichever is more convenient to you. If you prefer to talk, we are generally available on Mondays for a meeting in person at the institute, or we may arrange an online or in-person meeting on another day if that's better for your schedule.

We kindly ask you to give your feedback by **January 15, 2026**.

Would you be open to taking a look? We've attached the slides to this email. For your convenience, we also attach a plain text file with all questions.

Looking forward to hearing from you!

[Authors]

Participants. Four machine learning experts agreed to participate in this study:

- **E1:** Male, about 60 years old, with a strong mathematical background and extensive experience in ML consultancy and deployment of Random Forest models in industrial settings.
- **E2:** Male, about 35 years old, working on the development and deployment of AI systems in healthcare applications.
- **E3:** Female, about 30 years old, primarily engaged in ML research while also contributing to applied model development projects.

- **E4:** Male, about 35 years old, ML researcher and data scientist involved in applied industrial and business projects.

All participants had substantial experience with machine learning models but little or no prior exposure to visual analytics tools for model inspection.

For E1, the study took the form of an in-person, semi-structured interview lasting approximately three hours, involving three authors. The session included extensive discussion, clarification of visual representations, and reflective assessment of each technique. The other experts provided written feedback. In the case of E3, this feedback was preceded by a detailed written exchange and an online discussion, primarily addressing questions about the interpretation and role of topic modeling in understanding rule interactions.

1.1 Interview with E1

1.1.1 Expert Profile

The interviewed expert (MM) is a senior male practitioner (approximately 60 years old) with a strong mathematical background and extensive experience in machine learning. His professional background includes many years of consultancy work focused on the development and deployment of ML models in real-world industrial and business contexts. While highly experienced in model construction, validation, and auditing, he had no prior experience with visual analytics systems or interactive visualization techniques for model inspection. This background provided a valuable perspective on the potential added value of visual, logic-centered model analysis tools for experienced ML practitioners.

1.1.2 Evaluation Format

The evaluation was conducted as an in-person expert interview involving three paper authors and one ML expert. The session lasted approximately three hours (originally planned for 1.5 hours).

The interview consisted of two main parts:

- **Part 1 (approximately 2.5 hours):** Introduction of the goals of RuleSense, followed by a detailed walkthrough of each visual, interactive, and computational technique using the vessel behavior classification case study. This part involved extensive discussion, clarification questions, and suggestions from the expert.
- **Part 2 (approximately 0.5 hours):** Structured assessment of individual tools and the overall approach, guided by the evaluation questions included in the slide deck.

1.1.3 Summary of expert's feedback

Perceived utility. The expert found the feature-value distribution overview particularly useful for gaining an initial global understanding of model logic. Interactive filtering was considered essential for isolating and inspecting questionable or contradictory rules, although clearer visual differentiation after filtering was suggested. Automatic detection of contradictions and redundancies was viewed as valuable, provided that flagged rules could be visually inspected and any modifications validated on data.

The projection of rules by condition similarity was rated as highly informative. The observed clustering and proximity of certain classes aligned well with domain expectations and increased trust in the model logic. Topic modeling was regarded as a meaningful way to reason about higher-order feature interactions, especially given the impracticality of inspecting individual rules. The

feature–topic matrix was seen as an effective alternative to common topic visualizations, though interpreting topics required substantial cognitive effort.

Overall assessment and limitations. Overall, the expert concluded that RuleSense meaningfully supports logic-centered model auditing and trust calibration, particularly as a complement to performance-based evaluation. He emphasized that unexpected patterns should trigger further data-based investigation rather than immediate conclusions. Suggested extensions included better support for assessing model uncertainty, robustness, and the impact of small changes on predictions.

1.1.4 Detailed report of the discussion

Introduction and Scope The session began with a clarification of the overarching goal of RuleSense: supporting global and intermediate-level understanding of model logic, rather than local instance-level explanations. The expert raised several foundational questions:

- Whether global explanations are feasible in practice,
- Whether RuleSense requires a specific rule format,
- Whether the approach is applicable beyond decision trees (e.g., association rules),
- Whether logical operators such as OR and NOT are supported,
- Whether the approach applies to Random Forest models.

We clarified that RuleSense currently operates on rule sets consisting of conjunctions of conditions over numeric features, using comparison operators ($<$, $>$). OR and NOT are not supported. The approach is designed for rules extracted from Random Forests. The expert noted that all of his consultancy projects are based on Random Forest models, confirming the practical relevance of the approach.

Feature Distribution Overview The expert engaged in an extended discussion about the feature–value distribution overview:

- He discussed the distinction between the importance of a feature for a class and the importance of specific value intervals of that feature.
- He initially found the heatmap representations difficult to interpret, especially after filtering, when all cells appeared strongly colored.
- The discussion highlighted the impact of feature binning on interpretability.

The expert suggested improving textual formulations in the filtering interface, for example replacing *"select rules not including the feature"* with *"exclude rules including the feature"*, to reduce ambiguity.

Interactive Filtering Filtering was seen as particularly important for isolating and pruning contradictory or confusing rules. However, the expert again noted difficulties interpreting uniformly colored heatmap cells after filtering and suggested showing, for filtered subsets, the percentage of rules per class relative to the full rule set to provide better context.

Rule Projection by Condition Similarity The expert asked detailed questions about the distance function used for projecting rules. We explained that it is a specialized variant of the Hausdorff distance adapted to sets of numeric intervals representing rule conditions.

He found the projection view highly informative, noting that:

- The overall clustering structure helped form a mental model of the classifier's logic.
- The proximity of certain classes (e.g., anchoring and port enter/exit behaviors) aligned well with domain expectations and increased trust in the model logic.

He suggested that in mixed regions of the projection it would be useful to more explicitly highlight the distinguishing conditions between nearby clusters. The added value of interactive selection within the projection was less immediately clear to him and required further explanation.

Topic Modeling During the discussion of topic modeling, the expert explicitly stated that inspecting individual feature distributions is insufficient for understanding model logic, making higher-order analysis approaches such as topic modeling reasonable.

Key discussion points included:

- Interpretation of the feature–topic matrix and mapping topics back to concrete rules.
- The difficulty of interpreting aggregated line plots without training; to address this, we showed the non-aggregated version and explained the aggregation into 5% quantile bands.
- A question as to why only the dominant topic is not used; we explained that similar topic weights may lead to misleading conclusions.

The expert expressed a desire to identify decisive combinations of feature intervals for specific classes. He noted that, for this model, topic modeling did not reveal strongly class-exclusive combinations, even when certain topics were dominant for a class. He suggested considering frequent itemset mining as a complementary technique.

He positively evaluated the feature–topic matrix visualization, noting that it is more informative than word clouds commonly used in text-based topic modeling. He suggested assigning meaningful names to topics and discussed possibilities for automatic topic labeling. He also recommended simplifying the aggregated line plots.

1.1.5 Expert Assessment of RuleSense Capabilities

Overall Usefulness The expert stated that the approach is clearly useful and increases confidence in the model. However, upon detecting unexpected or suspicious rules, he would not draw conclusions immediately. Instead, he would investigate corresponding borderline cases in the data, emphasizing the importance of validation on labeled data when available.

Assessment by Tool

- **Feature distribution overview:** Very useful as an initial global overview.
- **Interactive filtering:** Very important for pruning and inspecting problematic rules; visual interpretation could be improved.

- **Contradiction and redundancy detection:** Useful, but visual inspection is necessary for understanding and verification. The expert emphasized that model modification should always be followed by evaluation on data.
- **Rule projection:** Highly informative; clustering and class proximity aligned well with domain logic and increased trust.
- **Topic modeling:** Helpful for understanding composite logic and feature interactions; requires cognitive effort but offers meaningful insights.

1.1.6 Limitations and Suggested Extensions

The expert identified several limitations and potential extensions:

- Support for assessing model uncertainty through overlapping or conflicting rules.
- Tools for investigating model robustness, such as identifying minimal changes that alter predictions.
- Better integration of data-driven validation to complement logic-based inspection.

Overall, the expert viewed RuleSense as a promising approach for logic-centered model auditing, particularly as a complement to traditional performance-based evaluation methods.

1.2 Feedback from Expert E2

1.2.1 Summary of E2's feedback

E2 found the visual analytics framework broadly helpful for assessing model logic, identifying questionable patterns, and supporting auditing workflows. The rules overview enables quick checks of feature reasonableness but lacks insight into feature combinations. Interactive filtering is considered highly valuable, offering clearer interpretation than static rule lists, though E2 notes that exploring many feature ranges manually can be demanding.

Automatic detection of contradictions and redundancies is seen as useful, with contradictions prompting deeper inspection, while redundancies are less critical. Rule projection becomes informative mainly when clusters are interactively explored, especially in cases of class mixing or outliers. Topic modelling provides limited added value unless topics remain small and interpretable.

Overall, E2 believes these tools reveal intermediate-level model behaviours that would remain hidden when relying solely on accuracy or instance-level explanations. They can influence deployment decisions, though some issues, such as label leakage, data bias, calibration, or fairness, are not well captured. E2 sees potential applicability in other domains, including healthcare.

1.2.2 Original responses from E2

Rules overview

Does this overview allow you to quickly assess whether the model relies on reasonable features and value ranges for this domain?

- Yes, it does help me assess reasonable feature and value ranges.
- An optimal usage of this view requires intuitive feature names and good knowledge about the class numbering.

Would this view help you decide whether the model merits deeper inspection, or whether there are early warning signs?

- Yes, this view would motivate me to dig deeper, especially if I detect early warning signs.

What kinds of anomalies, surprises, or inconsistencies would you expect to detect at this stage?

- Unexpectingly large of small values of a feature/class combination.
- Something like a 'model collapse', indicated by the rules of a class only using a single feature for predicting this class.

What important aspects of the model's logic are not yet visible from this overview alone?

- This view doesn't offer insight into feature combinations. A single class may consist of rules, that only rely on single or very few features (i.e. 'disjuct rules').

Interactive filtering

Can interactive filtering help you isolate parts of the model logic that you would consider questionable, critical, or worth closer inspection?

- I think the filtering definitely gives me the opportunity to detect questionable parts, but I have to work out the specific ranges out myself. Personally, I would be a little hesitant to check hundreds of feature range combinations and would hope for some suggestions.

Can filtering by features and their value ranges make the model's reasoning clearer or easier to interpret?

- Yes, the filtering supports me by checking pre-existing expectations and thereby my interpreation of the model.

Would this capability support decisions such as "this logic is acceptable" versus "this part of the model is problematic"?

- This capability can support the notion of "this part of the model is problematic" very well, by highlighting non-sensical parts.
- The decision for "this logic is acceptable" would grow in me after checking multiple expectations and finding no problematic relationships. I would still be cautios, because I dont know how good the rule surrogate approximates the original model and whether an acceptable logic of the surrogate translates to the original model.

Compared to static summaries or rule lists, does interactive filtering provide additional insight that you consider important for model auditing?

- Yes, the interactive filtering is far superior to mere rule lists, as it aggregates multiple rules and contextualizes them.

Automatic detection of contradictions and redundancies

Can automatically detected contradictions between general and more specific rules help you identify potentially unreliable or illogical model behaviour?

- Identified contradictions definitly would make me listen up. Still, I would be asking myself if this is a quirk of the rule model (maybe an artifact of the rule learning algorithm) or a deeper flaw of my original model.
- I would expect contradictory rules to indicate some data bias or class edge-cases.

Do you consider any contradictions problematic, or some could be tolerable? What domain or contextual factors would influence your judgment?

- I would mainly inspect the number of instances for rules involved in contradictions and redudancies to assess whether this is a major problem or a minor.
- I would inspect features more thoroughly if domain experts of my project indicated that they are highly relevant/discriminatory features.

Is it useful to automatically flag rules that are subsumed by others and predict the same class as candidates for redundancy or simplification?

- Yes, I would prefer a button to automatically subsume all redundant rules.

Would this kind of automated detection influence your decision to inspect further, modify the model logic, or reduce trust in certain parts of the model?

- The detection of redundancies would not influence my decision. Contradictions for multiple instances would definitely make me inspect the model more.

Rule projection

Does projecting the rule set by condition similarity help you form a mental model of the classifier's overall logic?

- The projection alone doesnt help me a lot. But selecting e.g. overlapping clusters and looking at the included rules might help me a lot.

Do visually distinct clusters, overlaps among classes, or isolated rules draw your attention to areas that may require closer inspection from a domain or logic perspective?

- Yes, especially overlapping classes draw my attention. I would like to inspect how similar those rules are.

Is the ability to interactively select clusters and immediately inspect their shared conditions and predicted classes useful for auditing large rule sets?

- Yes this ability is useful. It might be even more helpful so somehow automatically highlight the most relevant differences between instances of multuple classes in a selected cluster.

Which projection patterns would you consider most concerning (and why): class mixing, outliers, or overly dense clusters?

- Class mixing and outliers.

Topic modelling

Does topic modelling provide added value for understanding higher-order interactions in the model's logic?

- The benefit is not immediately clear to me. E.g. what do I get from the topics that I cannot see in the clusters?

Does it help you identify meaningful and recurring combinations of conditions?

- I think it could help me, but only if it is a low number of features in a topic, e.g. maximum of 3 or 4, as I would need to interpret not only the features but also their co-occurrence.

Do any topics reveal condition combinations that you find surprising, implausible, or worth questioning from a domain perspective?

- Hard for me to judge, as I dont understand all features, e.g. what is SpeedQ3?

Would a model dominated by a small number of interpretable and domain-consistent topics increase your confidence? Conversely, would many unclear or questionable topics reduce it?

- Yes, the first would increase my confidence. The second one would not necessarily reduce my confidence, just increase my need to search for better explanations.

Overall assessment

To what extent can analyses of this kind support decisions about whether a model is trustworthy and safe to deploy?

- When I can check the models behavior on a more detailed level, I can be more confident if it relies on actual patterns or spurious correlations. This influences my trust into the model.

Could this kind of analysis change your decision about deploying a model that appears highly accurate?

- Yes, if I detect suprising combinations or implausities, I would hesitate to deploy and consult domain experts first.

Do these tools expose problems that would likely remain hidden if you relied only on accuracy, validation on test datasets, or instance-level explanations?

- Yes, definitely. Global model metrics are very high level, instances very low level. This tool gives me the possibility to inspect 'intermediate' model behaviors.

Do the tools of RuleSense allow you to reason about the model at meaningful levels of abstraction (global and intermediate), given that inspecting individual rules is infeasible?

- Yes, see above.

Are there situations where the insights obtained might be misleading or over-interpreted?

- If clusters or topics are small, the influence of the clustering/dimensionality reduction or topic modeling algorithm might be larger than the actual model. This could lead to misleading interpretations or false alarms.

Which types of model issues do you feel are not well captured by these techniques?

- Label leakage, Data selection bias, calibration errors (this tool only looks at the singular class predictions, not on probabilities, right?), out-of-distribution errors, fairness problems.

Do you see these techniques as applicable to other domains or models you work with?

- Yes, in healthcare e.g. rule predictors for deteriorating hospital conditions.

1.3 Feedback from Expert E3

1.3.1 Summary of E3's feedback

E3 sees RuleSense as a helpful tool for high-level inspection of rule-based models, particularly for identifying which features and value ranges contribute to each class. The rules overview supports spotting suspicious feature ranges, though deeper inspection is always required and interactions between rules remain hidden. Interactive filtering is useful mainly for focusing on domain-relevant feature intervals and identifying problematic rules. It enhances detail compared to static rule lists but does not, in E3's view, clarify the model's reasoning.

Contradiction and redundancy detection is informative but model-dependent: in random forests, contradictions are expected, though they may still reveal corner cases or unstable reasoning. Automatic removal of redundant rules helps streamline analysis.

Rule projection helps assess consistency of class descriptions, with class overlaps being the most concerning pattern. Topic modeling provides a coarse similarity structure and can reveal recurring condition combinations, though its meaningfulness is uncertain.

Overall, E3 believes RuleSense can reveal problematic inputs, class confusion, and expected domain logic, supporting trust decisions and potentially influencing deployment if confusion or vulnerabilities are exposed. The approach offers a suitable abstraction level but may introduce biases through projection or topic modeling and does not capture blind spots or out-of-distribution behavior. Applicability to deep learning remains unclear.

1.3.2 Original responses from E3

Rules overview

Does this overview allow you to quickly assess whether the model relies on reasonable features and value ranges for this domain?

- This overview allowed me to see which features participate in the rules for each particular class and and which ranges of values are there across all the rules. It would be interesting to know which overlaps in the rules are there - which should be seen in topic modeling later. This view can allow for high level assesment if some particular features have very wrong range for a particular class.

Would this view help you decide whether the model merits deeper inspection, or whether there are early warning signs?

- I would see it as a way to see early warning signs, but it would always require deeper inspection.

What kinds of anomalies, surprises, or inconsistencies would you expect to detect at this stage?

- See the first answer in the group.

What important aspects of the model's logic are not yet visible from this overview alone?

- See the first answer in the group.

Interactive filtering

Can interactive filtering help you isolate parts of the model logic that you would consider questionable, critical, or worth closer inspection?

- I see it rather as a way to concentrate on features and feature intervals that I know from domain knowledge should be relevant for each class; or other way around that should not be there in any case.

Can filtering by features and their value ranges make the model's reasoning clearer or easier to interpret?

- I do not think so, because model's reasoning is based more on the intercation between different rules. It can help to identify better problematic rules though.

Would this capability support decisions such as "this logic is acceptable" versus "this part of the model is problematic"?

- I think it will support the second statement.

Compared to static summaries or rule lists, does interactive filtering provide additional insight that you consider important for model auditing?

- From my point of view, it improves level of details that we can grasp.

Automatic detection of contradictions and redundancies

Can automatically detected contradictions between general and more specific rules help you identify potentially unreliable or illogical model behaviour?

- Not in the case of random forests, because in this particular case it is implied that there will be overlapping rules and possibly contradicting. It is still interesting to see contradictions, because it can mean that there are corner case samples in the dataset that lead to such contradicting rules. It can signal about unstable reasoining in rules set models or in decision tree.

Do you consider any contradictions problematic, or some could be tolerable? What domain or contextual factors would influence your judgment?

- I see it as dependant on the model type and amount of corner cases in the data.

Is it useful to automatically flag rules that are subsumed by others and predict the same class as candidates for redundancy or simplification?

- Since the task of the RuleSense is analysis, such automatic cleaning helps to concentrate on the interesting details of the model interpretation.

Would this kind of automated detection influence your decision to inspect further, modify the model logic, or reduce trust in certain parts of the model?

- None of the listed (on the condition of good performance of the model on the selected data). Amount of contradictory rules would signal me that there can be problems with the training data.

Rule projection

Does projecting the rule set by condition similarity help you form a mental model of the classifier's overall logic?

- I would rather say that for me this projection will illustrate consistency of the class descriptions formed by a model.

Do visually distinct clusters, overlaps among classes, or isolated rules draw your attention to areas that may require closer inspection from a domain or logic perspective?

- I guess most interesting are overlaps, because they should be responsible for confusion between classes. Isolated rules can be interesting if they can directly isolate some samples of a particular class, so maybe these samples are very special.

Is the ability to interactively select clusters and immediately inspect their shared conditions and predicted classes useful for auditing large rule sets?

- It maight help to get selected view on the similar rules, but I suspect that the interactive filtering can give similar insights as well.

Which projection patterns would you consider most concerning (and why): class mixing, outliers, or overly dense clusters?

- Class mixing - explained in the answer to the question 2 in this group.

Topic modelling

Does topic modelling provide added value for understanding higher-order interactions in the model's logic?

- It helps to see more coarse similarity structure between rules for separate classes.

Does it help you identify meaningful and recurring combinations of conditions?

- I cannot be sure about meaningful - I would actually say that it can be an insightful "test", to see if there is any topic found that is described by very meaningful set of conditions - but it should find recurring combinations.

Do any topics reveal condition combinations that you find surprising, implausible, or worth questioning from a domain perspective?

- I cannot be sure that topics can reveal something implausible, but we can be reassured in the model's logic, if we can find enough topics that have meaningful combinations of conditions.

Would a model dominated by a small number of interpretable and domain-consistent topics increase your confidence? Conversely, would many unclear or questionable topics reduce it?

- (i) Yes, (ii) no - it is hard to expect that the model creates conditions exactly in a way that a human (domain expert) would.

Overall assessment

To what extent can analyses of this kind support decisions about whether a model is trustworthy and safe to deploy?

- Such analysis can help to see (i) problematic data inputs (ii) reasons for confusion between classes (iii) presence of expected domain logic in the decisions. I think two last points can increase the trust in the model.

Could this kind of analysis change your decision about deploying a model that appears highly accurate?

- If I see a lot of confusion in the model I will first try to produce counterfactuals - data points that belong to my input domain, but will lead to wrong predictions. And if I can reliably do this, then I will not deploy such model.

Do these tools expose problems that would likely remain hidden if you relied only on accuracy, validation on test datasets, or instance-level explanations?

- Yes.

Do the tools of RuleSense allow you to reason about the model at meaningful levels of abstraction (global and intermediate), given that inspecting individual rules is infeasible?

- Yes, such approach gives a sufficient level of abstraction.

Are there situations where the insights obtained might be misleading or over-interpreted?

- I would suspect that tSNE and topic modeling can introduce own biases, so one should be very careful with making conclusions from the location of the rules in visualization or topics.

Which types of model issues do you feel are not well captured by these techniques?

- Blind spots or unknown unknowns, so with this analysis we cannot see what the model will do on the outlier samples.

Do you see these techniques as applicable to other domains or models you work with?

- Not sure about deep learning.

1.4 Feedback from Expert E4

1.4.1 Summary of E4's feedback

E4 found that RuleSense provides a useful overview of rule distributions and feature usage, though interaction effects between features remain hidden. The overview can highlight potential anomalies or spurious correlations and help decide whether deeper inspection is warranted.

Interactive filtering was viewed as essential for narrowing the rule set, clarifying feature interactions, and making auditing feasible. While it supports identifying questionable logic, final judgments still depend on domain expertise.

Automatic detection of contradictions and redundancies was considered helpful for managing large rule sets, although contradictions are expected in ensembles and may reflect uncertainty

rather than illogical behaviour. Redundancy flags aid exploration but removing rules could distort ensemble predictions.

Rule projection offers broad insight but was seen as too complex to reveal clear structure; class mixing was noted as the most concerning pattern. Cluster-based inspection is useful but would benefit from tighter control over variation.

Topic modelling adds value by exposing recurring condition patterns, though many topics are dominated by single features and often span multiple classes, limiting interpretability. Confidence would depend on topic coherence and explanatory power.

Overall, E4 considered RuleSense effective for global and intermediate-level reasoning about large rule sets and capable of revealing issues missed by accuracy or instance-level explanations. Key limitations include lack of information on rule support, rule accuracy, and ensemble effects. The techniques generalize well to tree-based models and domains with structured, interpretable features.

1.4.2 Original responses from E4

Rules Overview

Does this overview allow you to quickly assess whether the model relies on reasonable features and value ranges for this domain?

- It provides a concise overview on which intervals and feature regions rules are located and onto what they're focusing. As with many explainability methods, this visual representation also relies on whether or not features are sufficiently interpretable by humans, and whether differences among features can be understood in concise ways. For example for class one in the presented figure both median and Q3 speed features appear to be used in a similar fashion.

Would this view help you decide whether the model merits deeper inspection, or whether there are early warning signs?

- To continue from above, it's precisely these type of observations which can prompt deeper inspections of the model.

What kinds of anomalies, surprises, or inconsistencies would you expect to detect at this stage?

- While likely not relevant for the dataset presented, the visualization would be a good way to find spurious correlations in terms of simple clever hunts effects early on. For example, if the dataset would be contaminated by simple cues and correlations indicative of specific classes without any underlying causality.

What important aspects of the model's logic are not yet visible from this overview alone?

- Rules typically rely on interaction, meaning the values of feature A which are taken into account depend on the values used for feature B. This type of interplay cannot be represented in this reduced view, although it might is important to understand the interaction of features in more detail.

- A second visual problem is the range of rules. While the view condenses the rules on coverage of specific subintervals, one might be mistaken in whether the subintervals are covered by separate rules or by larger rules spanning the entire region.

Interactive filtering

Can interactive filtering help you isolate parts of the model logic that you would consider questionable, critical, or worth closer inspection?

- Filtering rules affects also the display of all other rules in question, and in this way allows one to study a bit more the interaction mentioned as limitation above. It also highlights the support of given rules. For example, in the presented figure, although the log curvature is set close to maximum, most rules still cover the entire support of the feature. This might be indicative that these rules are rather designed to exclude very specific cases as they are not sensitive on most values of said feature.

Can filtering by features and their value ranges make the model's reasoning clearer or easier to interpret?

- Especially the capability to exclude specific features from the rule set makes it easier to obtain shorter and therefore more interpretable rules as interactions, and in the same way the depth of the considered tree(s), are limited. This can be seen on slide 22, where excluding the top features led to much sharper distributions on, e.g., `MaxDistPort` for class 1, despite not restricting this specific feature. However, there might be a trade-off between the depth of the forest ensemble members and the capability of excluding specific features.

Would this capability support decisions such as "this logic is acceptable" versus "this part of the model is problematic"?

- It allows the user or observer to narrow down on a much smaller range of rules, allowing for a much clearer investigation. Stating whether a logic is acceptable or problematic is, on the other hand, challenging. Given domain knowledge, one could likely prescribe such an attribute to one specific rule.

Compared to static summaries or rule lists, does interactive filtering provide additional insight that you consider important for model auditing?

- To continue from above, only the ability to filter down on subset of rules makes, in my opinion, such investigations feasible. There is a trade-off between too many individual rules (on the order of ten thousand) to investigate each in detail, and a too abstract representation when showcasing all of them. Therefore I believe that interactive exploration is an essential part to RuleSense.

Automatic detection of contradictions and redundancies

Can automatically detected contradictions between general and more specific rules help you identify potentially unreliable or illogical model behaviour?

- Given that the underlying model is a random forest consisting of one hundred trees, I assume the focus is more on unreliability than illogical behavior. The model is an ensemble, implying that for any given input X, it is quite unlikely that all members of the ensemble will come to the same conclusion. It necessarily follows that for the specific conditions presented by the input, several trees and therefore rules need to contradict one another.

Do you consider any contradictions problematic, or some could be tolerable? What domain or contextual factors would influence your judgment?

- The contradictions in the rules are likely a measure of uncertainty or unclear data domains such that different outcomes are plausible. This of course does not mean that rules might be illogical or defy domain knowledge (due to, e.g., errors in the model or data).

Is it useful to automatically flag rules that are subsumed by others and predict the same class as candidates for redundancy or simplification?

- On the one hand, this approach is quite helpful for exploration of the rule set. If one can reason about the general rule, investing the specific ones in detail might be unnecessary. In terms of model prediction and outcome, on the other hand, I believe that removing such rules bears the risks of changing the underlying "probabilistic" distribution of the overall ensemble model.

Would this kind of automated detection influence your decision to inspect further, modify the model logic, or reduce trust in certain parts of the model?

- Yes. One of the core challenges is to make sense out of a huge rule set, and therefore any reduction of a set due to redundancy can reduce the necessary effort. On the other hand, contradicting decisions, especially if the rules are not only overlapping but highly similar, can indicate decision boundaries within the data distribution over the model, and therefore might also help to uncover corner cases and peculiarities.

Rule projection

Does projecting the rule set by condition similarity help you form a mental model of the classifier's overall logic?

- While the projections to a 2D space provides an interesting overview on all the rules, the entire space appears too complex for good structure to emerge. While there are some indications for clusters, most of them are mixed with respect to predicted classes, and the resulting rule sets are still very broad. This might serve an additional purpose of better following which parts of the rule space were already explored, as it is easier for a human to keep track of a 2D space.

Do visually distinct clusters, overlaps among classes, or isolated rules draw your attention to areas that may require closer inspection from a domain or logic perspective?

- From perspective of exploration, of course isolated points could be interesting. There is, though, the risk that they result from some idiosyncrasy in, e.g., splitting data among ensemble members. Based on the provided figures, almost all classes, with the exception of the red one (class 4), intersection with one another in terms of similar rules.

Is the ability to interactively select clusters and immediately inspect their shared conditions and predicted classes useful for auditing large rule sets?

- While useful in principle, it would be beneficial if the degree of variation within the selected subset would be more easily controllable.

Which projection patterns would you consider most concerning (and why): class mixing, outliers, or overly dense clusters?

- More concerning to me is the overlap in classes, as it implies that the distance metric used among rule sets is mostly meaningless with respect to descriptions of classes. Either necessary distances to distinguish classes are not resolved, or most of the rules are what is called "contradictory" above.

Topic modelling

Does topic modelling provide added value for understanding higher-order interactions in the model's logic?

- Yes. In terms of sparseness, it is comparable to the filtered rule sets used before, which makes it easier to see correlations among feature ranges. In contrast to the filtered rule sets, topics can be attributed, as fragments, to a larger set of rules allowing a much broader overview.

Does it help you identify meaningful and recurring combinations of conditions?

- It should be noted that many of the topics have one clear focus element, such as `Log10MinDistPort` (Topic 1) or `SpeedQ3` (Topic 4), while other feature ranges are of lesser "intensity".

Do any topics reveal condition combinations that you find surprising, implausible, or worth questioning from a domain perspective?

- Yes. Consider, e.g., Topic 8 where the dominant feature are higher ranges for `SpeedQ1`, while at the same time maximal ranges for `SpeedMinimum` seem to be excluded. This seems to suggest that "acceleration" or at least variability within the overall speed range is of some importance.

Would a model dominated by a small number of interpretable and domain-consistent topics increase your confidence? Conversely, would many unclear or questionable topics reduce it?

- As far as I understand the concept of topic modeling the number of models is a tunable hyper-parameter. To build confidence solely on the basis of topics it would therefore also be necessary to understand their explanatory power for given rules.
- A second challenge is interpreting the topics in context of the predictions. As observed on the slides most topics correspond to most classes. Therefore the inner "logic" of a topic cannot be interpreted solely in the context of a class.

Overall assessment

To what extent can analyses of this kind support decisions about whether a model is trustworthy and safe to deploy?

- The tool provides in-depth visualizations of the underling rule-sets of Random Forest models. These can be used to analyze sub-groups of the rules via interactive filtering. A clear limitation is the humans capability to keep track of large sets (10k rules). This is somewhat offset by implementing contradictions discovery, 2D embedding (t-sne), and topic modeling.

Could this kind of analysis change your decision about deploying a model that appears highly accurate?

- Potentially yes. For example, due to the discovery of spurious correlations. But also contradictory use of features, both among rules and w.r.t. domain knowledge, could prompt further investigation. Here, it would be advantageous if rules could be traced back to matching data instances.

Do these tools expose problems that would likely remain hidden if you relied only on accuracy, validation on test datasets, or instance-level explanations?

- RuleSense provides a visually-assisted interactive way to explore sets of rules. This allows for a significantly deeper model understanding compared to aggregated metrics. While instance-level explanations might be comparable in explanatory power, they suffer the same drawbacks individual rules have: they are isolated examples and investigating all in detail is typically not feasible.

Do the tools of RuleSense allow you to reason about the model at meaningful levels of abstraction (global and intermediate), given that inspecting individual rules is infeasible?

- The overall level of abstraction is appropriate, especially as dynamic zoom-ins are possible for both the main rule-view (by filtering constraints) or the 2D embeddings (by selecting regions).

Are there situations where the insights obtained might be misleading or over-interpreted?

- Some feature ranges cover almost most of the available feature range, suggesting that their main purpose is not to include specific inputs, but rather exclude some of the values. Without going back to the underlying data, this is however a speculative statement.
- Another example are contradictions. While some of the rules seem to come to opposing classifications for similar data points, this is more a consequence of the used model class than necessarily a failure on the model itself.

Which types of model issues do you feel are not well captured by these techniques?

- As RuleSense investigates the set of rules in isolation, it is unclear how well supported the respective rules are by data. In a similar fashion, it is also unclear how accurate a given rule is. The second aspect not covered is the ensemble nature of the random forest.

Do you see these techniques as applicable to other domains or models you work with?

- It naturally applies to Random Forests, individual Decision Trees, or Subgroup based approaches. It might also be extendable to Gradient Boosted Trees, but boosted architectures no longer treat all rules equally, which likely requires to re-think some of the approaches.
- Concerning the data domain, the approach should work for most fields where inputs can be reduced to small-dimensional structured data (ideally with interpretable features), e.g., in industrial application for sensor analytics.

2 Domain-Oriented Interpretation with an Epidemiology Expert

This study examined whether RuleSense can support domain experts in using a trained model as a lens for understanding real-world phenomena encoded in the data. This study focused on the COVID-19 incidence prediction model based on disease history and population mobility in Spain.

Participant. The participant was a male epidemiology expert, approximately 40 years old, with deep domain knowledge of COVID-19 dynamics and public health responses in Spain.

Focus. The study explored:

- how well the model logic aligns with the expert's understanding of pandemic evolution,
- whether temporal shifts in model behavior correspond to known phases of the pandemic, and
- whether interactive rule exploration supports knowledge discovery about the interplay between disease progression and mobility patterns.

2.1 Compact summary of the study

We conducted a 90-minute online session with a domain expert in epidemiology and mobility analysis to assess the utility of RuleSense in exploring a rule-based Random Forest model predicting weekly COVID-19 incidence levels in Spain (2020–2021). The discussion focused on interpreting model behavior, identifying strengths and limitations, and evaluating the relevance of the visualizations for domain-expert reasoning.

Key observations from the session include:

- The model captures broad pandemic phases and short-term momentum in disease incidence, but fails to represent temporal trends consistently for all classes.
- Mobility effects are underrepresented, as only internal mobility within provinces was included; inter-regional mobility, which strongly influences pandemic dynamics, was not captured.
- Topic modeling revealed stable high and low incidence patterns but did not provide clear evidence of local trends, reinforcing that temporal dynamics are poorly represented in the model.
- RuleSense visualizations were effective in exposing model logic, highlighting structural limitations, and supporting expert-led interpretation.

The expert recommended incorporating explicit trend features and inter-regional mobility in future models. Overall, the evaluation confirmed the value of RuleSense for model-centered exploration and diagnosis, while also identifying clear directions for improving the underlying predictive model.

2.2 Detailed interview report

This report summarizes the discussion with the domain expert regarding the RuleSense analysis of the Spanish COVID-19 dataset. The session focused on interpreting model behaviour, identifying its strengths and shortcomings, and assessing the usefulness of RuleSense visualizations for exploring the underlying decision logic.

- **Duration:** 90 minutes
- **Format:** Online session (Microsoft Teams)
- **Participants:** RuleSense development team and one domain expert in epidemiology and mobility analysis (male, about 40 years old)
- **Materials Used:** Slide deck sith 16 slides

Procedure A 90-minute expert evaluation session was conducted to assess how effectively RuleSense supports the exploration and interpretation of a rule-based model predicting weekly COVID-19 incidence levels in Spain (2020–2021). The session combined a structured presentation (based on a prepared slide deck) with interactive discussion and clarification of data and modelling choices. The primary aim was to understand how well the model reflects real-world epidemiological dynamics and how the visual analytics capabilities of RuleSense facilitate domain-expert reasoning.

2.2.1 Introduction to RuleSense and the Model

The session began with an overview of RuleSense and the structure of the COVID-19 prediction model (Slides 1–6). RuleSense was introduced as a visual analytics tool for logic-focused exploration of rule-based machine learning models, supporting the assessment of how well the model's decision logic aligns with domain knowledge, even when access to the training data is limited.

During this part of the session, the expert asked several clarifying questions concerning:

- the steps involved in data preprocessing,
- the definitions of disease incidence and mobility levels,
- the construction of the temporal context (weeks –6 to –1),
- and the meaning and interpretation of individual features.

To address these questions, additional explanatory materials were presented. After these explanations, the expert confirmed that he understood the data structure and the meaning of the visual representations used in the subsequent parts of the session.

2.2.2 Interpretation of Feature–Value Distributions

Slide 7 introduced the visualization of feature–value distributions across the predicted classes. The expert confirmed his understanding of the heatmaps and how they summarize conditions within the rule set.

We then discussed the patterns observable in these distributions (Slides 8–11). We highlighted our interpretations and asked the expert to evaluate their plausibility based on his knowledge of COVID-19 dynamics in Spain. The discussion focused on several key aspects:

- the dominant influence of recent disease incidence compared to mobility,
- the representation of different pandemic phases through the *Days_passed* feature,
- class-specific patterns corresponding to early, mid, and late stages of the pandemic.

The expert confirmed that the patterns related to *Days_passed*, particularly those shown in Slides 9–11, were broadly consistent with his understanding of how the pandemic evolved.

A further discussion emerged regarding the surprisingly weak influence of mobility in the model. The expert asked whether the mobility features represented internal mobility within provinces or inter-provincial mobility. We clarified that only internal mobility was included. The expert noted that internal mobility had substantial effect only during the strict initial lockdown, whereas later phases of the pandemic were driven primarily by cross-regional movement—an aspect not represented in the current model. He expressed strong interest in developing and exploring a model that incorporates geographic context and inter-province mobility flows.

2.2.3 Rules Without Absolute Time: Evaluating Local Temporal Patterns

Slide 12 presented the feature-value distributions in the subset of rules that did not include the absolute time feature *Days_passed* (used as a proxy for calendar dates). These rules were expected to clarify the influence of local temporal context, such as levels of disease incidence and mobility in the weeks preceding the prediction target week.

The observations were as follows:

- For classes 1 and 4, the rules roughly reflected local trends in disease incidence: decreasing for class 1 and increasing for class 4.
- For classes 2 and 3, no meaningful temporal trends could be identified.

The expert regarded the absence of identifiable trends, especially for Classes 2 and 3, as a significant limitation. He emphasized that representing temporal evolution is essential for interpreting epidemic dynamics and for understanding transitions between incidence levels.

We explained that part of the difficulty arises from the aggregated visualization of the rule subsets in the heatmaps, but the core issue lies in the model itself: it did not effectively capture temporal trends given the input feature design. The weekly incidence values were encoded as separate features. Because the training algorithm treated these features as independent variables, it failed to infer the temporal relationships among them and to learn coherent trend patterns.

Together with the expert, we concluded that the input data should be redesigned to encode temporal trends explicitly (e.g., through difference or slope features). The current representation as a collection of independent weekly indicators hindered the model from learning coherent temporal dynamics.

2.2.4 Topic Modelling Results and Their Interpretation

Slides 13 and 14 presented the results of topic modelling (TM) applied to the rule conditions. We explained that the purpose of TM was to extract recurring combinations of feature conditions that could reveal deeper interrelationships among features, such as implicit temporal patterns or interactions not easily visible in aggregated heatmaps. Thus, if local temporal trends had been effectively captured by the model, they could have appeared as structured patterns within topic distributions.

However, the TM results did not show clear evidence of such trends:

- Topics 6 and 7 reflected stable high and low disease incidence levels, respectively.
- Topics 1–4, 8, and 9 contained only vague and fragmented hints of temporal changes, without forming interpretable patterns.

The expert noted that both the concept of TM and the interpretation of its results were challenging. He found it difficult to determine whether the topics conveyed additional epidemiologically relevant information. Nevertheless, he agreed that the TM results reinforced the earlier conclusion that the model did not adequately capture local temporal trends.

2.2.5 Summary and Expert Reflections

We concluded the session by presenting a synthesis of the pandemic-related findings (Slide 15). The expert confirmed that these findings aligned with his domain knowledge regarding the temporal evolution of COVID-19 in Spain. He also agreed with the statements about the role and utility of RuleSense presented in Slide 16.

During the final discussion, the expert reiterated his key recommendations:

- Future models should explicitly incorporate **temporal trend information**, rather than relying on independent weekly features.
- **Inter-regional mobility** should be included, as it plays a major role in the spatial progression of the pandemic.
- The visualizations in RuleSense are valuable for diagnosing model behaviour, understanding what the model has learned, and identifying limitations.

2.2.6 Conclusions

The expert evaluation session demonstrated the following:

- The current model reflects certain broad pandemic phases but fails to capture key temporal trends.
- Mobility effects are underrepresented due to the restriction to internal mobility data.
- Both the heatmap and topic modelling analyses expose structural limitations of the model's feature representation.
- RuleSense provides effective tools for uncovering these limitations, supporting rich expert analyst dialogue, and enabling model-based reasoning beyond standard performance metrics.

The expert evaluation confirmed that, while RuleSense effectively supports the exploration of model logic, the underlying model suffered from structural limitations related to the representation of temporal information. Future iterations should redesign the feature representation to encode temporal dynamics directly and incorporate mobility patterns.

References

[1] N. Andrienko, G. Andrienko, and B. Kathirgamanathan. Interactive visual exploration of rule-based model logic. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–17, 2026. doi: 10.1109/TVCG.2026.3689736