This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2026.3689736

Interactive Visual Exploration of Rule-Based Model Logic

Natalia Andrienko, Gennady Andrienko, and Bahavathy Kathirgamanathan

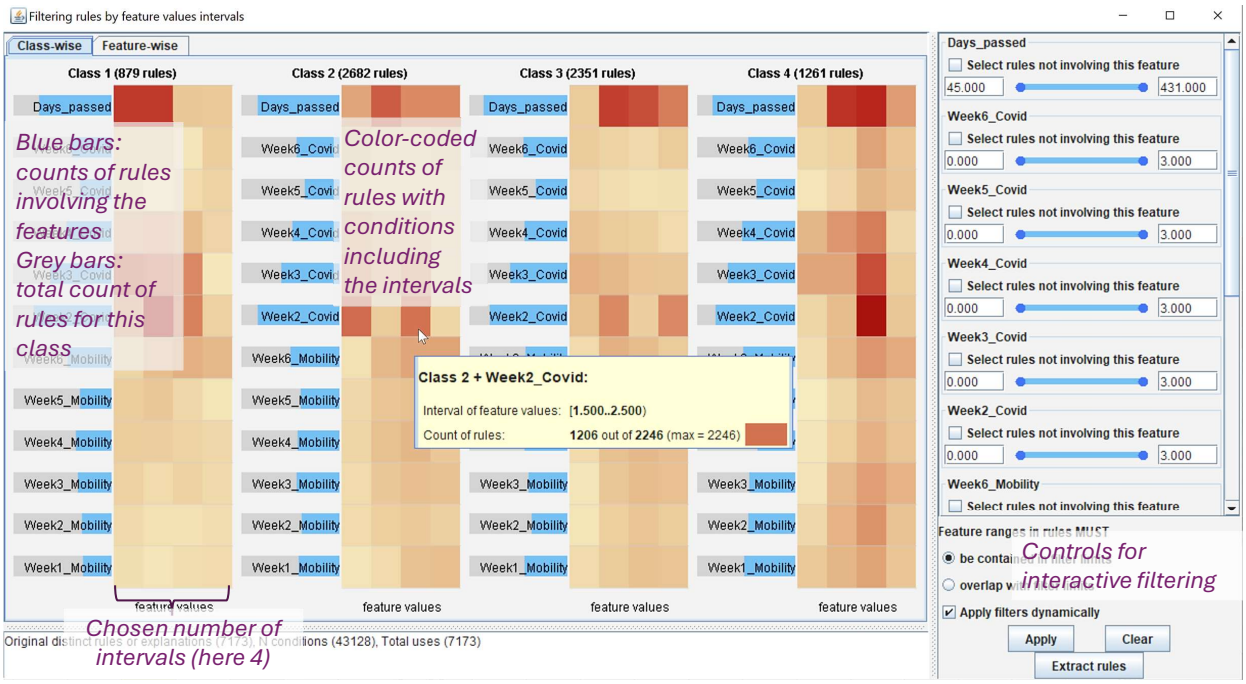

Fig. 1: An overview of the distribution of the features and feature value intervals across subsets of decision rules from a model.

Abstract—Rule-based machine learning models, including those derived from decision trees or forests, are often considered inherently interpretable. However, human understanding is hindered by model size, rule complexity, and interdependencies between features. Moreover, rule sets extracted from ensemble models can contain contradictory, incomplete, or counterintuitive logic, even when the overall model achieves high predictive accuracy. This paper introduces a visual analytics methodology designed to support systematic exploration of rule-based model logic and its alignment with domain knowledge. Our approach integrates overview visualizations, interactive filtering, contradiction analysis, and topic modeling. This enables analysts to detect illogical or implausible rules, assess their potential impact, and refine the model to improve its interpretability and trustworthiness. A key distinction of our method is its ability to support reasoning about model behavior both with and without access to labeled data. We demonstrate the approach through two real-world case studies: evaluating logical consistency in a vessel movement classifier and analyzing feature relationships in a COVID-19 prediction model. These studies show how visual analytics can facilitate logic-focused model critique beyond traditional performance metrics and enable valuable domain-relevant insights.

Index Terms—Model explainability, Model trustworthiness, Decision rules, Decision trees, Random Forest

1 Introduction

In many real-world applications of machine learning, especially in high-stakes domains such as healthcare, policy development, or infrastructure planning, it is not enough for a model to perform well on training or test data. Domain experts must be able to check whether the model's logic aligns with established knowledge and reasoning patterns, so that it can behave sensibly when confronted with new, unseen situations. A model that produces correct predictions for wrong or unclear reasons can undermine trust and lead to serious consequences when deployed.

This paper addresses the fundamental need to **evaluate the internal logic of ML models** for the correspondence to human reasoning and domain understanding. This is particularly important in applications where:

- the model may be applied to new data not represented in the training set,
- domain constraints or regulatory requirements demand explanation and justification, or
- the model is intended not only to make predictions but to reveal insights about the underlying data and processes.

We introduce RuleSense, a visual analytics *approach* and conceptual–methodological *framework* designed to support interactive exploration of model logic represented by decision rules. RuleSense enables users to examine the model beyond input–output behavior and detect inconsistencies, unclear or redundant reasoning patterns, and poten-

- *Natalia and Gennady Andrienko are with Fraunhofer Institute IAIS and with City, University of London. E-mail: {natalia\gennady}.andrienko@iais.fraunhofer.de.*
- *Bahavathy Kathirgamanathan is with Fraunhofer Institute IAIS.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

tially unsound or surprising decision logic. The approach is designed to work primarily on rule sets derived from models without requiring access to training data, although additional functionality is available when labeled ground truth data are present.

Our methodology integrates:

- Multi-level visual representations of large rule sets, supporting analysis from overview to detail [6],
- Interactive filtering and drill-down by features, outcomes, and value intervals,
- Computational techniques for detecting contradictions, identifying redundant rules, and uncovering feature co-occurrence patterns through topic modeling.

Unlike most existing XAI tools, which focus on explaining individual predictions (e.g., via SHAP values or counterfactuals) or improving performance through feedback, our methodology enables systematic examination of the model as a whole, checking its alignment with domain logic rather than statistical performance. Addressing the needs of domain experts and model auditors, RuleSense supports model soundness verification, domain-relevant insight extraction, and reliability assessment. We demonstrate the methodology through two real-world case studies: vessel movement classification and COVID-19 pandemic prediction. These studies show how the approach identifies logically flawed rules, validates model behavior, and uncovers latent patterns in the learned logic.

To validate the utility and trustworthiness of our approach, we conducted two complementary empirical studies with expert users (Section 7). Four ML experts evaluated RuleSense's capacity to support logic-focused model auditing, demonstrating that the methodology enables reasoning about model behavior at global and intermediate levels and assessment of model trustworthiness beyond accuracy metrics. An epidemiology domain expert applied the methodology to the COVID-19 prediction model, confirming that RuleSense can serve as a "lens" for understanding real-world phenomena, revealing strong alignment between the model's decision patterns and known epidemiological dynamics while also identifying structural model limitations.

Our contributions are:

- A visual analytics methodology for evaluating rule-based model logic and its alignment with human reasoning and domain expectations,
- Interactive and computational techniques for analyzing and refining rule sets, applicable with or without labeled data,
- Evidence from case studies and expert evaluation that logic-focused examination supports both identification of model inconsistencies and domain knowledge discovery, enabling a shift from performance-centered to logic-centered model validation.

The remainder of this paper is organized as follows. Section 2 positions our work within related research on interpretable ML, visual analytics, and topic modeling. Section 3 defines the analytical tasks our methodology addresses. Section 4 describes how RuleSense supports these tasks through overview visualization, interactive filtering, computational analysis, and topic modeling. Sections 5 and 6 demonstrate task execution through case studies on vessel movement and pandemic prediction. Section 7 validates the methodology through expert evaluation focusing on logic auditing and domain-driven discovery. Section 8 reflects on contributions and limitations.

2 Related Work

Recent visual analytics research emphasizes bridging machine learning models and human reasoning. Frameworks integrating VA with explainable AI address conceptual mismatches between ML outputs and human mental models, leveraging visualization for model development and interpretability [3].

2.1 Interpretability in Rule-Based Models

Interpretability in ML refers to explaining model decisions in human-understandable terms. While rule-based models are often considered inherently interpretable due to their explicit logical structure, complexity quickly exceeds human cognitive limits with increasing rule counts or complex feature interactions [3].

Efforts to enhance interpretability typically focus on optimizing rule structures and explanation mechanisms. Mullins citemullins2023shape analyzes how rule structures influence transparency. The Interpretable Decision Sets framework [20] improves comprehensibility through independent, non-overlapping rules. Adilova et al. [1] enable flexible exploration by representing large rule sets through simpler models at adjustable abstraction levels.

For Random Forests [7], interpretability techniques include feature importance rankings and rule extraction [16]. LionForests [30] simplifies rule sets using unsupervised methods to reduce paths and features. TreeExplainer [24] uses Shapley values to explain individual predictions and capture feature interactions, deriving global understanding from local insights. Rudin & Shaposhnik [41] introduce summary explanations that locally interpret predictions from globally interpretable models.

Despite these advances, most approaches lack interactive methods enabling deep exploration of rule sets **from a logical consistency perspective** rather than predictive performance alone.

2.2 Visual Analytics for Model Interpretation

Visual analytics techniques support interpretation and exploration of increasingly complex ML models, offering interactive tools to understand model logic, evaluate predictions, and refine models [28]. work specifically focuses on approaches supporting interpretation of decision tree or rule-based models, excluding neural networks and other black-box models [19].

Foundational systems. RuleMatrix [29] pioneered visual exploration of rule sets, arranging rules in structured matrix formats to facilitate feature interaction inspection. SuRE [45] organizes surrogate rule sets hierarchically, visualizing them as sparse matrices where rows represent tree layers and columns denote features. ExplainExplore [11] enables active manipulation of explanations by selecting surrogate models and adjusting parameters. Specialized techniques have been developed for industrial applications, such as RfX for fault detection in electrical engines [13].

Tree ensemble visualization. Models comprising multiple decision trees pose specific interpretability challenge. To visualize ensemble components and their contributions to overall model results, Colorful Trees represent decision trees in a botanically inspired visual form [34], while Ribeiro et al. [40] employ dimensionality reduction to visualize the results of classifier ensembles.

Systems specifically designed for random forests often transform them to rule sets. iForest [48] combines feature-level and instance-level views for global and local explanation, visualizing importance, split distributions, partial dependencies, and decision path clusters. Explainable Matrix [32] uses matrix layouts to show rule-feature relationships. RFMap [27]] extends this with rule-instance and instance-rule matrices supporting dimensionality reduction and embedding-based similarity exploration, emphasizing class separation and misclassification patterns.

Scalability and Refinement. RuleExplorer [22] addresses scalability through anomaly-biased model reduction and hierarchical organization, using logistic regression proxy models to compute anomaly scores. Users explore flagged rules and verify their prediction impact. Other tools simulate "what-if" scenarios by extracting simple rules and suggesting feature modifications [15].

Design Principles. Research in rule set visualization explores how design choices, such as feature alignment, predicate encoding, and rule ordering, affect human understanding [46, 47]. Visual frameworks for analyzing decision tree collections enable quality assessment through performance metrics and efficient parameter tuning [35].

Distinction of Our Approach. While these systems provide valuable approaches for improving interpretability and supporting refinement, most prioritize enhancing predictive performance and model tuning. Our approach distinguishes itself by focusing on understanding inherent model logic and checking consistency with domain knowledge

Table 1: Key distinctions between RuleSense approach and representative systems.

Aspect	Existing Systems (e.g., iForest, RFMap, SuRE)	RuleSense Approach
Primary Goal	Performance inspection, local explanation	Evaluate logic, align with domain reasoning
Data Dependence	Strong (instance-based metrics, coverage, etc.)	Works independently of data (data-optional)
Target Audience	Model developers	Domain experts, model reviewers
Rule Reduction	Optimized for fidelity/performance	Optional, logic-driven cleaning and simplification
Contradiction Detection	Rarely addressed	Explicitly supported
Use Cases	Debugging, instance-level analysis	Post-hoc audit, model understanding, trust support

and human reasoning, enabling not only logical issue identification but also domain insight discovery.

2.3 Applications of Topic Modeling

Topic modeling identifies latent thematic structures through patterns of co-occurrence [44]. Its application extends beyond text analysis to DNA sequences, software repositories, user interactions, and time series [5,8,9,23] by conceptualizing data elements as terms and records as documents. The primary goal is revealing latent structure through shared usage patterns, supporting exploration of variable relationships rather than identifying individual frequent combinations. Unlike traditional clustering techniques that typically enforce a 'hard' assignment (where a data item must belong to a single group), topic modeling allows for probabilistic membership. This is more suitable for complex, structured data [18] where where items may relate to multiple thematic contexts simultaneously.

Visual analytics has expanded topic modeling utility through interactive exploration frameworks [10, 14, 31] and broader co-occurrence structure analysis emphasizing meaningful subspace discovery [18].

An important distinction exists in this context between topic modeling and pattern mining techniques, such as frequent itemset or association rule mining. Pattern mining identifies explicitly recurring combinations exceeding frequency thresholds. Topic modeling uncovers latent relationships through partial and indirect co-occurrence patterns. When item A frequently co-occurs with B and C, and B co-occurs with C, but the full combination (A,B,C) is rare, pattern mining may miss the structure while topic modeling reveals their common latent theme.

Our work is, to our knowledge, the first to **apply topic modeling to rule sets** from ML models. We encode rule conditions as terms and rules as documents, enabling automated discovery of recurring themes and higher-order relationships between conditions within large rule sets. This complements feature-centric analysis with a holistic view of model logic, especially where explicit frequent combinations are sparse. However, when identifying explicit, outcome-discriminative condition combinations is the objective, pattern mining remains more appropriate. For example, when applied to rules from a model predicting potency of chemical compounds, pattern mining helped us to discover specific combinations of chemical substructures that frequently occur in rules predicting high compound potency but are rare or absent in rules predicting low potency.

2.4 Comparative Positioning

Our approach builds on prior research in interpretable ML, rule-based modeling, and visual analytics, but **emphasizes examination of model logic and its alignment with domain reasoning** rather than predictive performance, instance explanations, or model reduction. Table 1 summarizes key distinctions.

Performance- and Instance-Oriented Exploration. Systems such as iForest [48], Explainable Matrix [32], RFMa [27], and RuleExplorer [22] provide interactive visualizations to analyze random forests and derived rules. They support revealing feature-prediction relationships, investigating decision paths, identifying misclassifications, and evaluating feature importance using statistical measures. These tools rely heavily on labeled data, evaluating rule certainty and importance through instance coverage, confidence measures, or proxy models.

RuleSense takes an essentially different approach by supporting model understanding **regardless of data availability**, suitable for restricted-access domains or third-party audits. Rather than relying on instance coverage and statistical agreement, our methodology enables reasoning about rule semantics, detecting presence/absence of relevant features, identifying counterintuitive conditions, and uncovering rule contradictions.

Rule Simplification and Surrogate Models. Systems like SuRE [45] build compressed hierarchical representations to approximate black-box models. RuleExplorer [22] introduces anomaly-aware reduction. Surveys like Haddouchi & Berrado [16] highlight methods balancing fidelity and complexity for performance explanation. These systems improve usability through simplification but aren't primarily designed for domain-inconsistent logic detection or manual logic validation. Simplification techniques often discard statistically weak rules that may hold significant value for domain experts.

RuleSense allows interactive inspection and refinement aiming not at improving performance but at **eliminating rules violating domain expectations or introducing illogical behavior**. Modifications are applied only upon explicit expert decision, not automatically. When validation data are available, experts can compare model accuracy before and after changes and undo modifications if needed.

Domain-Centered Knowledge Discovery. Among reviewed systems, RfX [13] most closely aligns with our emphasis on domain expert reasoning, supporting experts in discovering meaningful feature combinations and translating them into domain insights. Similarly, RuleSense is intended not just for explaining models but for extracting insights from their structure: identifying unexpected rules, discovering feature relationships, and validating domain constraint compliance. However, RfX focuses on editing tree structures and requires data access, while **RuleSense works at the rule set level** and **functions even in data-sparse or data-inaccessible contexts**, supporting flexible use cases like post-hoc audits or exploratory reviews.

Design and Visualization. RuleSense incorporates matrix- and glyph-based representations similar to Explainable Matrix [32], RuleExplorer [22], and Visualizing Rule Sets [47]. This design aligns with findings that feature alignment and visual predicate encoding improve understanding. However, we extend these with

- **High-level visual summaries** of the relationships between feature values and model outcomes for large subsets of rules,
- **Interactive filtering tools** for detecting rules missing relevant features or exhibiting unexpected combinations,
- **Contradiction analysis** to spot potential logical inconsistencies,
- **Topic modeling** applied to rules to uncover latent structures in the decision logic.

Summary of Distinctions. This positioning highlights our unique focus on evaluating whether model decision logic is reasonable, trustworthy, and aligned with human expectations, contributing to VA and XAI by shifting focus from input-output behavior exploration to internal mechanism understanding through structural analysis.

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2026.3689736

3 Problem Statement

This section formalizes the problem that our research addresses and the tasks that our methodology is designed to support. While RuleSense is designed to **work with IF-THEN rule sets derived from any ML model**, we demonstrate the methodology primarily using Random Forest-derived rule sets. Such models inherently include redundancies, contradictions, and over-specificity, which makes them particularly suitable for developing and testing logic assessment techniques. Section 3.1 defines the data structures and terminology. Section 3.2 describes the analytical tasks.

3.1 Background

A **decision rule** in machine learning is a logical statement of the form "IF *conditions* THEN *outcome*". Conditions are typically formulated as comparisons between feature values and thresholds (e.g., IF age > 50 AND blood pressure $\leq$ 140 THEN risk = low). Decision rules can be used for classification, where the outcome is a class label, or regression, where the outcome is a numerical value.

A **decision tree** is a hierarchical structure that uses decision rules to recursively partition data into subsets based on feature conditions. Each internal node represents a condition on a feature, each branch corresponds to an outcome of that condition, and each leaf node holds a final prediction. The path from the root to a leaf forms a rule representing the combination of conditions encountered along that path. Therefore, a decision tree can be transformed into an equivalent set of rules by extracting a separate rule for each leaf node, capturing the conditions from the root to the leaf.

A **random forest** is an ensemble model consisting of multiple decision trees, typically trained on different subsets of the data through bootstrap sampling. Each tree independently generates predictions, and the forest aggregates these predictions by majority vote in classification tasks or by averaging in regression tasks. Transforming a random forest into a set of rules involves collecting rules from all individual trees, resulting in a potentially large and diverse rule set. The complexity and size of such rule sets make them challenging to explore and interpret.

Additional challenges arise from the way the random forest algorithm works. Each decision tree is trained on a randomly sampled subset of data and a random subset of features at each split. While this randomness improves robustness and accuracy, it also leads to redundant and contradictory rules when the trees are transformed into a unified rule set. Contradictions occur because different trees may learn conflicting patterns from their respective samples, while redundancies arise when multiple trees produce identical or very similar rules.

Scope and applicability. Our methodology is applicable to rule sets from any source: decision trees, outputs of rule induction algorithms, or expert-crafted systems. However, we have primarily focused on **random forest-derived rule sets** because they **inherently exhibit the challenges our approach addresses: redundancy across trees, contradictory predictions, and over-specific conditions**. This makes random forest models well suited for both validating and demonstrating our logic assessment techniques. The principles and methods generalize to any rule-based model where logical consistency and domain alignment are important.

3.2 Objectives and Tasks

Based on discussions with domain experts and our own experience on applying machine learning in various domains, we have identified analytical tasks involved in examining rule-based models. These tasks focus on understanding model logic, identifying inconsistencies, and assessing domain alignment.

- **T1: Overview Visualization** – Provide a comprehensive, scalable view of the entire rule set, displaying feature frequencies, condition distributions, and relationships between rules and predicted outcomes.
- **T2: Interactive Filtering and Drill-Down** – Enable filtering of rules by features or outcome values and drilling down into specific subsets to examine how feature conditions influence predicted outcomes and how well model logic aligns with domain knowledge.
- **T3: Analysis of Feature Relationships and Interactions** – Explore joint feature effects and co-occurrence patterns as expressed in the rule set.
- **T4: Identification of Logically Inconsistent Rules** – Detect logical flaws in the model's decision-making process: rules that lack essential features, contradict domain logic, or exhibit reasoning patterns deviating from human expectations.

Tasks T1–T4 are designed to be performed without labeled data. When such data are available (possibly different from those used to develop the model), the following additional tasks can be supported:

- **T5: Rule Testing on Labeled Data** – Apply the rule set or selected subsets to labeled data and examine correct and incorrect predictions to identify potential flaws in model behavior or data inconsistencies.
- **T6: Impact Assessment and Refinement** – Evaluate how removal or modification of problematic rules affects overall model performance to determine whether logical improvements can be made without degrading predictive accuracy. All modifications can be applied only upon explicit expert decision and are reversible.

4 Methodology

Building on our earlier work [1], which introduced a distance function for rule similarity and an algorithm for rule aggregation and generalization, and following approaches like RuleMatrix [29], Explainable Matrix [32], and others [47], RuleSense visualizes individual rules in table (matrix) format with sorting by features, outcomes, rule characteristics, or similarity. For examining small rule groups, we provide compact glyph-based representations [1]. This paper focuses on new components constituting our conceptual and methodological framework:

- **overview visualizations** of feature and value-interval usage across large rule sets (T1);
- **interactive filtering** and subset extraction (T2);
- structural **analysis of contradictions, redundancy, and similarity** (T3, T4);
- **topic modeling** of rule conditions to reveal higher-order feature co-occurrences (T3).

A software prototype instantiates this methodology; implementation details are not our primary contribution.

4.1 Overview Visualization

To provide a comprehensive view of how features and value intervals are used across a rule set (T1), the methodology employs heatmap-based overviews of feature distributions (Figs. 1 and 2). To create such a display, the analyst selects a number of equal-length intervals for the feature value ranges, e.g., 4, as in Fig. 1, or 10, as in Fig. 2. Coarser discretizations give a general overview; finer ones reveal more detailed patterns.

Two complementary layouts are used, presenting the same information in different arrangements to provide complementary perspectives on the rule set. Both variants consist of multiple heatmap matrices that visually summarize the distribution of feature values across the rule set. The columns of the matrices correspond to intervals of feature values. In the **class-wise view**, each matrix corresponds to a class, and rows correspond to features (Figs. 1 and 2, top). For regression models, the analyst defines classes interactively by dividing the range of predicted values into intervals. In the **feature-wise view**, each matrix corresponds to a feature, and rows correspond to classes (Fig. 2, bottom). Regardless of the view, each row represents a particular class-feature combination, with rows grouped either by class or by feature.

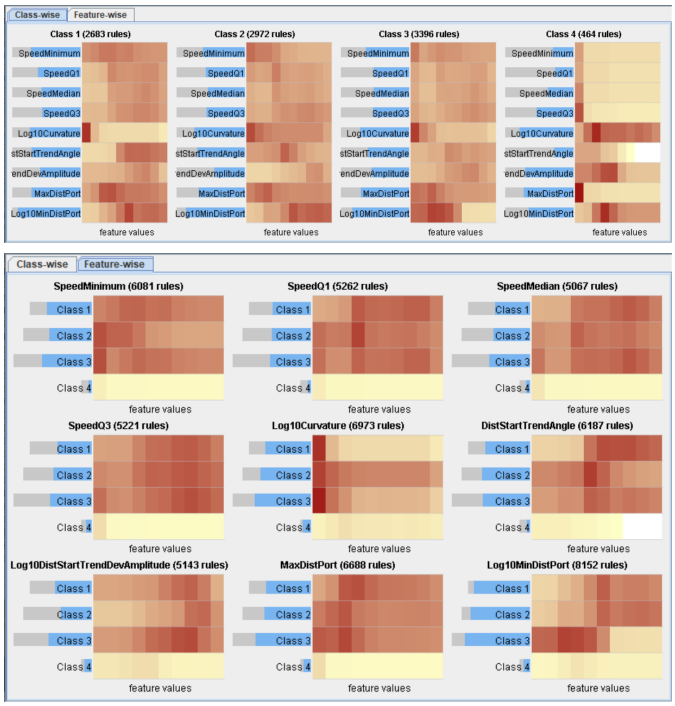

Fig. 2: Class-centered (top) and feature-centered (bottom) overviews of the distribution of feature value intervals across the rule set. Gray bars indicate the total number of rules for a class; blue bars show the fraction of those rules containing the corresponding feature.

A matrix row for a combination (C_i, F_j) appears as a heatmap of colored rectangles corresponding to the intervals of the feature F_j. The color intensity represents how often a particular value interval appears in rule conditions for class C_i. On the left of a heatmap, a gray horizontal bar representing all rules predicting that class is overlaid by a blue bar showing how many rules for that class include the specific feature. The ratio of the blue to gray bar visually indicates how much feature F_j is involved in the rules predicting class C_i.

This overview serves two diagnostic purposes:

- **Feature use diagnostics**: Check whether domain-relevant features are used to recognize a class and identify underused or missing expected features.
- **Interval diagnostics**: Check whether the value ranges used in rules for a class agree with domain expectations, potentially flagging suspicious intervals for closer inspection.

Feature occurrence counts indicate feature prominence within rule subsets but differ from methods like `LIME` [39] and `SHAP` [25], which measure feature contributions to specific predictions. In RuleSense, rule counts serve as logic-oriented diagnostic cues rather than direct predictive importance measures. In tree-based models, frequent occurrence typically reflects use in higher-level splits propagating into many rules; in ensembles like random forests, it indicates repeated feature selection across independently trained trees, suggesting robust reliance in learned decision logic. However, high counts may signal different phenomena, such as meaningful reliance, redundancy, or potential inconsistencies, which require interpretation through further interactive exploration.

We emphasize rule counts rather than instance coverage because our primary goal is auditing and understanding model logic independently of specific datasets. Coverage-based measures depend on underlying data distributions and may hide logically implausible or brittle rules affecting few instances yet critical for trust and safety. The overview visualization serves as an entry point for understanding how and where the model reasons, highlighting rule set parts requiring closer inspection.

4.2 Interactive Filtering and Subset Extraction

Interactive filtering refines the overview by restricting attention to rule subsets that match user-defined criteria, supporting tasks T2–T4. The methodology provides basic operations:

- **Filtering by outcome**: Select rules predicting one or more specified outcomes (classes or value ranges).
- **Filtering by feature inclusion or exclusion**: Retain rules that include (or exclude) one or more features. This allows examination of the role of particular features and identification of rules that omit features considered essential by domain experts.
- **Filtering by feature value intervals**: Restrict rules to those whose conditions intersect selected intervals for chosen features. This supports targeted analysis of how certain value ranges influence predictions.
- **Subset extraction and removal**: Interactively defined subsets can be extracted for detailed inspection (e.g., in rule-table or glyph views [1]) or removed from the rule set for hypothetical refinement or performance testing.

All filters are reflected immediately in the overview visualizations, which always remain linked to an unfiltered reference view. This linkage supports iterative exploration: analysts can progressively narrow down subsets while preserving a mental model of how these subsets relate to the rule set as a whole.

4.3 Computational Techniques for Rule Analysis

To reveal structural properties of the rule set and support T3–T4, the methodology incorporates three computational analysis functions (building on [1]):

- **Contradiction detection**: Identify pairs of rules where conditions in one rule are strictly more general than in another, yet the two rules predict different outcomes. Such contradictions indicate unstable or illogical regions in the decision logic and are prime candidates for closer scrutiny.
- **Subsumption detection**: Detect rules that are strictly more specific than others with the same outcome. Such *subsumed* rules are potentially redundant: they do not contribute new logic beyond what more general rules already express. Removing or merging subsumed rules can simplify the rule set while preserving its logical structure and, typically, its predictive behavior.
- **Grouping and generalization of similar rules**: Using a distance measure on rule conditions [1], rules with highly overlapping conditions can be grouped and generalized into more compact representatives. This provides intermediate-level abstractions: analysts can reason about aggregated "rule fragments" rather than thousands of individual rules.

These analyses do not prescribe automatic pruning. Instead, they provide structural signals (contradictions, redundancy, clusters of similar rules) that guide expert judgment about which parts of the logic deserve attention or potential modification.

4.4 Topic Modeling on Rule Conditions

To move beyond pairwise feature co-occurrence and capture higher-order condition patterns (T3), we apply topic modeling to the rule set. Each rule is treated as a short "document" and each encoded condition as a "term." Directly using raw continuous feature values ((e.g., SpeedMinimum_5.145_infinity)) would create an unmanageably large vocabulary of unique terms, so we proceed as follows:

1. **Discretization**: For each feature, its value range in the rule set is divided into a small number of quantile-based intervals (e.g., terciles or quartiles). Intervals are defined so that conditions are distributed roughly evenly across them.
2. **Condition encoding**: A rule is considered to include an interval if the numeric range in its condition overlaps that interval. For each feature, we represent the selected intervals as a binary code of

fixed length (one bit per interval), with 1 denoting overlap. Each condition is encoded as textttfeature__code. For example, with four intervals, a condition `SpeedMinimum` $\in [5.145, \infty)$ becomes `SpeedMinimum__0011`.

3. **Rule representation**: Each rule is converted into a space-separated string of encoded conditions, e.g.: `SpeedMinimum__0011 Log10Curvature__1000 DistStartTrendAngle__0111 Log10MinDistPort__0011`.

We then apply Non-negative Matrix Factorization (NMF), as it has been found effective for analyzing short-text documents [12,26], which aligns with the typical length of our extracted rules. NMF output consists of

- a **topic–term matrix**, where each topic is a weighted combination of encoded conditions, and
- a **document–topic matrix**, where each rule has a vector of topic weights.

Conditions that frequently co-occur across rules receive high weights within the same topic. This reveals latent "themes" in the rule logic, i.e., combinations of features and value ranges that recur across many rules.

Selecting the number of topics. NMF requires specifying the number of topics k in advance. In our approach, this parameter is determined empirically by running NMF iteratively across a range of candidate values and evaluating the quality of resulting topic models. Excessively high values produce degenerate topics characterized by single dominant terms with negligible weights for others, failing to capture meaningful co-occurrence patterns. Conversely, too few topics may fail to capture the diversity of condition patterns in the rule set.

The document–topic matrix reveals how strongly each rule associates with discovered topics. For each rule, the highest topic weight indicates its dominant topic. When the dominant weight is low and all topic weights are similarly small, the rule lacks strong association with any topic, suggesting it contains a rare or unusual condition combination not captured by the learned topics. A well-suited number of topics minimizes such low-association rules while avoiding topic degeneracy.

To select an appropriate number of topics, we recommend analyzing how dominant topic weights are distributed across rules for each candidate value of k. This can be visualized using quantile plots (Fig. 7) or histograms showing the distribution of maximum topic weights. Higher lower-quantile boundaries indicate that even weakly associated rules maintain reasonable topic affiliations, suggesting better model quality. The case studies (Sections 5–6) demonstrate this selection process in practice.

4.5 Task Support Mapping

The components described above jointly support the **data-independent tasks** defined in Section 3.2:

- **T1: Overview Visualization:** Class-wise and feature-wise heatmaps provide a scalable summary of feature usage and value-interval distributions, highlighting unexpected or missing feature–class relationships.
- **T2: Interactive Filtering and Drill-Down:** Outcome-based, feature-based, and interval-based filters allow analysts to isolate subsets of rules that reflect specific hypotheses (e.g., "rules for class 3 that ignore port distance") and to compare these subsets against the full rule set.
- **T3: Analysis of Feature Relationships and Interactions:** Co-occurrence patterns can be explored visually through the evolving heatmaps under different filters and computationally through topic modeling and rule similarity analysis.
- **T4: Identification of Logically Inconsistent Rules:** Contradiction and subsumption analyses identify structurally problematic rules; filtering operations support targeted inspection of rules that omit expected features or use implausible value ranges.

These techniques have been instantiated in a prototype implementation of the RuleSense framework, which provides concrete views and interactions for the case studies presented in the Sections 5 and 6. The focus of this paper, however, is on the methodology, i.e., the combination of visual, interactive, and computational techniques for logic-centered analysis of rule-based models, rather than on software architecture or engineering details.

The case studies described below were conducted by the authors, who had acquired sufficient familiarity with the respective application domains through earlier collaborations with domain experts.

5 Case Study 1: Recognition of Vessel Movement Patterns

5.1 Domain, Task, and Model

We examine a random forest model developed for recognizing movement patterns of fishing vessels. The model inputs are numeric attributes describing trajectory segments. The training data for model derivation were prepared as described in [4]. The model recognizes four distinct movement classes: **Forward movement** (class 1), **Trawling** (class 2), **Port enter/exit** (class 3), and **Anchoring** (class 4). The trajectory patterns corresponding to these classes are illustrated in Figs. 3 and 4.

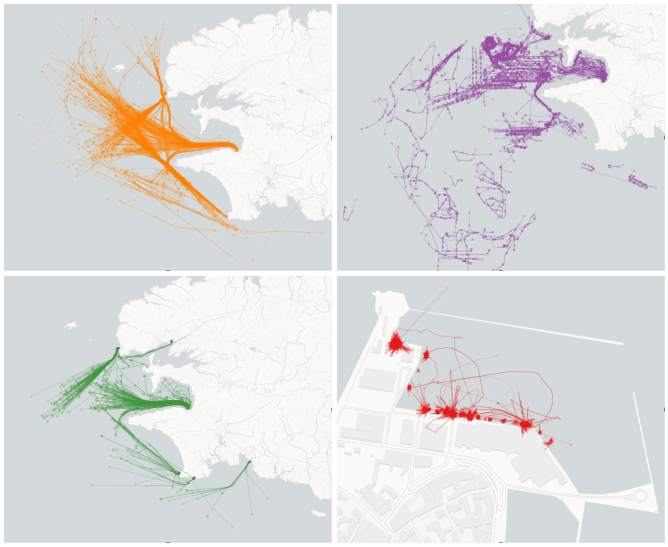

Fig. 3: Shapes of trajectory segments for four activity types: forward movement (top left), trawling (top right), port entering or existing (bottom left), and anchoring (bottom right).

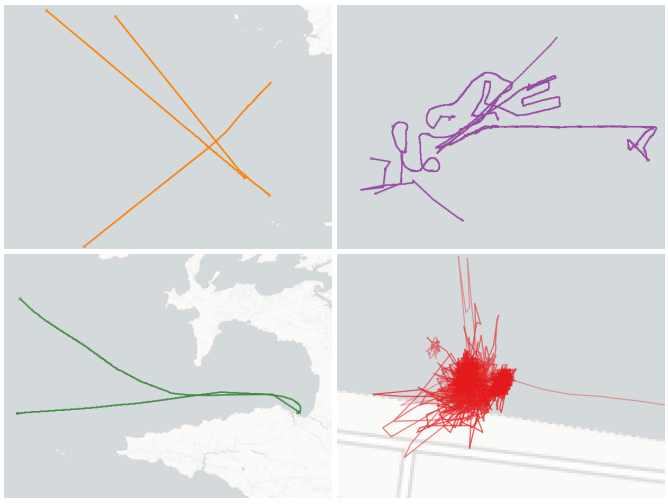

Fig. 4: Selected example trajectories for each movement pattern.

The model utilizes **nine features** describing vessel speed characteristics, trajectory shape, and proximity to ports. Logarithmic transformation was applied to features with highly skewed value distributions in the training data.

- **Speed statistics**: `SpeedMinimum`, `SpeedQ1`, `SpeedMedian`, `SpeedQ3`. These features summarize the speed distribution, from minimum observed speed to quartile-based spread.
- `Log10Curvature`: The logarithm of the curvature of the time series of the vessel's distance from the starting point. The curvature is computed as the ratio between the sum of absolute consecutive changes in the time series and the amplitude of values. A value close to 1 indicates nearly straight movement, while higher values indicate the presence of turns.
- `DistStartTrendAngle`: The angle of the linear trend fitted to the time series of the vessel's distance from its starting point. A higher angle indicates a stronger trend of moving away, while a lower angle suggests slower movement or a return toward the starting position.
- `Log10DistStartTrendDevAmplitude`: The logarithm of the amplitude of deviations from the trend line of the distance to the starting point. It quantifies path tortuosity; higher values indicate more zigzagging movement.
- **Port proximity**: `MaxDistPort`, `Log10MinDistPort`. These features represent the maximum and (log-transformed) minimum distance from the nearest port, which are useful for recognizing anchoring and port maneuvering.

Together, `Log10Curvature` (global shape) and `Log10DistStartTrendDevAmplitude` (local tortuosity) help distinguish steady outward movement, looping, and in-place maneuvering.

The random forest consists of 100 decision trees and achieves 99.33% accuracy on a test dataset with 4,798 labeled instances. The model was transformed into 9,939 unique rules, with duplicates removed.

5.2 Computational Analysis and Cleaning

Given the complexity of the ensemble-derived rule set, we first applied our computational techniques to clean the rules and improve interpretability while maintaining accuracy:

1. **Contradictory rules removal**: 113 rules with conflicting conditions leading to different predictions were detected, examined, and eliminated with no impact on accuracy.
2. **Subsumed rules removal**: 311 rules were identified as fully covered by more general rules and removed. The accuracy slightly improved to 99.35%.

The cleaned rule set consists of 9,515 rules.

5.3 Exploration of the Feature Value Distributions

For task **T1**, we create an overview display (Section 4.1) dividing feature value ranges into 10 equal-length intervals (Fig. 2). In the class-wise view (upper image), the heatmap matrices correspond to the classes with the rows representing the features. In the feature-wise view (lower image), the matrices correspond to the features and their rows to the classes.

The exploration begins with perceiving all matrices together and noting the most prominent differences. In the class-wise view, class 4 strikingly differs from the other classes. It is predominantly recognized by very low speed values and low `MaxDistPort`, aligning with domain knowledge about anchoring. `Log10Curvature` also shows notably different distribution, tending toward higher intervals. This feature and `Log10MinDistPort` are strongly involved in rules for all classes.

However, the `Log10MinDistPort` distribution for class 4 initially appears counterintuitive: medium to high values prevail, contrary to the left-skewed `MaxDistPort` distribution. Considering that values are logarithms of distances, the left half represents negative logarithms

Fig. 5: Interactive filtering of rules. Top: Class-wise distributions after hiding rules including any of the speed attributes. Bottom: Result of limiting `DistStartTrendAngle` to values below 0.

(distances < 1 km). The hot spot in intervals 3–5 corresponds to very small port distances, expected for anchored vessels. Consistently, class 3 rules cover the left part of this feature's range, while classes 1 and 2 tend toward higher values.

Class 1 (forward movement) is expectedly distinguished by higher speed values and `DistStartTrendAngle`, indicating fast progression from the starting point. `Log10Curvature` shows prominent prevalence of the lowest interval, reflecting relatively straight paths.

Class 2 (trawling) exhibits no strong hot spots, indicating high behavioral variety. This aligns with the knowledge that trawling vessels operate in diverse modes depending on fishing technique, environmental conditions, and vessel type. The overall distribution may hide subpatterns requiring further analysis.

In the feature-wise view, the light coloring of bottom matrix rows for class 4 immediately reveals the low rule count compared to other classes, suggesting class 4 is easier to identify with fewer rules.

Other observations include the strong role of `Log10Curvature` in differentiating class 2 from classes 1 and 3, `Log10MinDistPort` in distinguishing class 3, and `DistStartTrendAngle` in differentiating class 1. These align with domain expectations.

However, almost all value intervals appear in rules predicting multiple classes, except high `DistStartTrendAngle` values (absent for class 4). This indicates complex class boundaries relying on feature combinations rather than individual features, suggesting trajectory shape and movement intensity interplay is crucial. Deeper exploration is needed to understand these interactions.

5.4 Interactive Filtering and Drill-Down

Interactive filtering (**T2**) enables detailed examination of model logic and supports identifying potential inconsistencies (**T3**, **T4**). Filtering operations include selecting rules by outcome, feature inclusion/exclusion, and value intervals. Our prototype implementation provides filtering controls linked to the feature value distribution overview (Fig. 1, right panel). The interface includes checkboxes for selecting rules based on feature inclusion or exclusion, and two-sided sliders for setting value ranges.

One use of filtering is to check whether rules for a class depend on features that are considered domain-relevant (**T4**). For example, for class 3 (port entering/exiting), we expect relevant rules to include `MaxDistPort` or `Log10MinDistPort`. Filtering isolates rules omitting both features, revealing a group contradicting domain logic. Similarly, we explore whether other classes can be distinguished without speed features. Filtering out all speed attributes leaves few rules, many

relying on DistStartTrendAngle, though not always within plausible ranges.

Figure 5 shows filtering results. In the top image, the rules involving speed attributes are hidden to examine remaining non-speed-based rules. In the bottom, DistStartTrendAngle is further restricted to low values, highlighting class 1 rules not complying with expectations.

These exploratory operations do not modify the rule set but identify potentially illogical patterns for further investigation.

5.5 Rule Set Editing and Evaluation

To assess the impact of rule modifications, we apply the rules to a labeled dataset of 4,798 instances (**T5**) and examine how rule removal affects performance determining whether simplification can be achieved without sacrificing accuracy (**T6**). The original cleaned rule set consisted of 9,515 rules; after the refinement process described below, 8,903 rules remain.

Identified issues and actions:

- Class 3 rules omitting port features: 144 rules omit both MaxDistPort and Log10MinDistPort, contradicting domain logic. Removal does not reduce accuracy.
- Class 1 rules lacking speed features: 23 rules predict forward movement without speed features. Two allow too-broad DistStartTrendAngle ranges (from from $-\infty$ to high values), inconsistent with the class definition. Removal has no accuracy impact.
- Class 2 rules with atypically high speed: 7 rules allow SpeedMinimum $\geq$ 9.95 km/h, which is atypically high for trawling. These apply to very few instances (2–6) with mostly incorrect predictions. Removal yields no accuracy loss.
- Class 4 rules lacking upper speed bound: 53 rules predicting anchoring lack upper speed bounds, which is an implausible omission, were removed.
- Class 3 rules with unconstrained port distance: 353 rules have unconstrained minimum port distance. Removal causes minor accuracy drop from 99.35% to 99.33% due to one misclassification. The affected instance with Log10MinDistPort = 0.57 ($\approx$ 3.715 km) labeled as class 3 may be mislabeled; the revised prediction (class 1) may be more accurate.
- Class 3 rules with large port distances: 53 remaining rules have Log10MinDistPort lower bound exceeding -0.035 ($\approx$ 0.92 km). Removal does not affect performance.

The remaining model consists of 8,903 rules.

These examples illustrate how rule editing, informed by interactive filtering and supported by evaluation on labeled data, can refine model logic without significantly affecting performance. While effective, this process is labor-intensive. Further research is needed to explore approaches for reducing manual effort, such as enabling domain experts to define high-level semantic constraints that could be automatically verified against the rule set, potentially helping flag inconsistencies and guide refinement more systematically.

5.6 Exploration of Feature Interrelationships

Feature interrelationships (**T3**) can be explored interactively using filtering. Restricting the value range of one feature and observing how the distributions of other features respond identifies co-occurrence patterns and conditional dependencies. Figure 6 shows an example. In the top, SpeedMedian limited to low intervals reveals co-occurrence with lower intervals of other speed features and lower DistStartTrendAngle. In the bottom, SpeedMedian restricted to high intervals shows broader variety in other speed features but tendency to co-occur with lower Log10Curvature and higher DistStartTrendAngle.

While informative, purely interactive exploration is time-consuming and does not scale well. To address this, we apply topic modeling (Section 4.4).

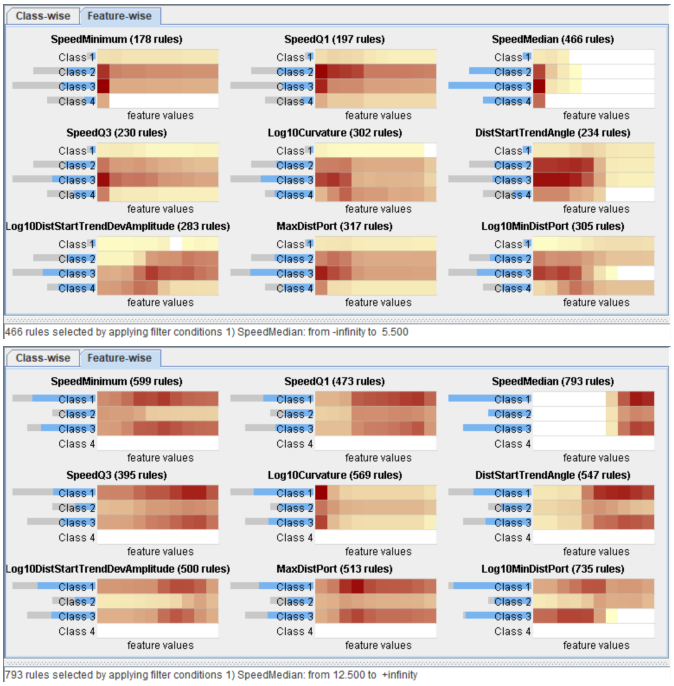

Fig. 6: Exploration of the relationships between SpeedMedian and the other features by limiting the value range of SpeedMedian to low (top) and high (bottom) value intervals.

Topic Modeling Application

We apply Non-negative Matrix Factorization (NMF) topic modeling to the 9,903 rules remaining after editing. Rule conditions are encoded as described in Section 4.4, discretizing feature value ranges into quartiles with each condition represented by a four-digit binary code.

To determine the appropriate number of topics k (see Section 4.4 for methodology), we run NMF iteratively for k ranging from 5 to 25. Values above 25 produce degenerate topics. Using a quantile plot (Fig. 7), we analyze the distributions of the dominant topic weights across the models. The horizontal axis of the plot corresponds to the number of topics (5–25), and the vertical axis represents the dominant topic weights, which in our experiments range from 0.0017 to 0.2802.

The distribution is visualized by plotting quantile bands: each band covers an inter-quantile interval containing approximately equal portion of the rules ordered by increasing dominant topic weight. In our example, we divided the rule set into 20 quantiles, i.e., each portion includes approximately 5% of the rules. The quantile bands are painted in alternating dark and light gray. The top border of each band represents a particular quantile; e.g., the top of the first (dark gray) band traces the 5th percentile of dominant topic weights across all topic models, the second band reaches the 10th percentile, and so on. Higher vertical positions of these lower quantile boundaries indicate that even the least topically associated rules maintain stronger topic affiliations, which is preferable for our analysis.

Considering only the 5th percentile (bottom band), the highest values occur at $k = 16$ and $k = 22$. However, when we examine the next bands (10th and 15th percentiles), the best results are obtained for $k = 13$, marked by a vertical line in Fig. 7. Since 13 topics are also easier to examine and interpret than larger numbers, we adopt this configuration for further analysis.

Topic–Term Analysis

The topic–term matrix for the 13-topic model is visualized in Fig. 8. In this display, rows correspond to features, columns to topics, and cells contain visual representations of the binary codes indicating the presence or absence of quartile-based intervals in rule conditions. Within each code, darker and lighter shades represent interval presence and

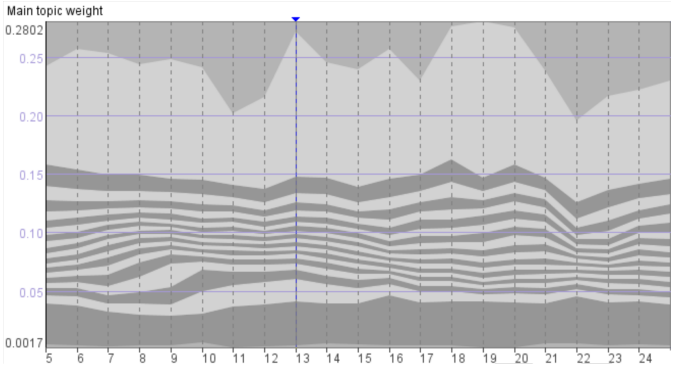

Fig. 7: Distributions of the dominant topic weights for the number of topics ranging from 5 to 25. The bands painted in alternating dark and light gray shades represent the 0.05 quantiles (5th percentiles) of the weight distributions. The vertical line marks the distribution for 13 topics.

absence, respectively. To maintain clarity, only terms with weights exceeding a user-defined threshold (adjustable via a slider below the table) are shown. The opacity of the symbols is proportional to the weight of the respective term. Relationships between features can be studied by interpreting the contents of each column, which reflect feature conditions that tend to co-occur.

Following observations can be made:

- **Topic 00**: Higher `SpeedMinimum` and `SpeedQ1` co-occur with lowest `Log10Curvature` and higher `DistStartTrendAngle`. While aligned with domain knowledge, association with low to medium port distances is less obvious. Other topics (**01, 02, 08, 12**) suggest that higher speeds/low curvature can also occur at medium to large port distances.
- **Topic 04**: Low `SpeedMinimum` may co-occur with medium to high `SpeedQ1` and low `Log10Curvature`, consistent with expectations, often observed far from ports.
- **Topics 05 and 10**: Lowest speed intervals associate with lower `DistStartTrendAngle` and higher `Log10Curvature`, agreeing with the observation that curvy movements occur at lower speeds.
- **Topics 08 and 12**: Less intuitive association between zigzagging (`Log10DistStartTrendDevAmplitude`) and low trajectory curvature.

Topics and Predicted Classes

To further clarify these findings, we also examine how the discovered feature associations (i.e., topics) relate to the predicted classes. Figure 9 presents a line chart where each rule is represented by a line connecting its weights across 13 topics, colored by the predicted class. This reveals multiple rule clusters per class associated with different dominant topics. No topic is exclusive to a single class, except class 4 (anchoring), predominantly associated with topic 10, suggesting this topic captures distinctive anchoring behavior.

Figure 10 shows quantile plots constructed separately for rule subsets per outcome class. Lower quantile bands are often invisible due to near-zero weights, meaning many rules have negligible weights for particular topics. Thus, the number of visible bands indicates the proportion of rules where a topic plays substantial roles.

We see the following class-specific patterns:

- **Class 1 (forward movement)**: 30–40% of rules have substantial weights for topics **1–4**, **8**, and **9**, characterized by high speed intervals and low `Log10Curvature`.
- **Class 2 (trawling)**: ∼20–25% of rules associate with topics **1** and **4** (suggesting straight-line trawling movements), and ∼25% with topics **5** and **7** (lower speeds). Topics **8**, **9**, and **11** involving higher `Log10DistStartTrendDevAmplitude` (zigzag movements) show moderate contributions (15–25%).
- **Class 3 (port enter/exit)**: Most dominant topics are **0**, **3**, and **6** (lower port distances), as expected. However, notable numbers associate with topics **2**, **4**, **8**, and **9** (higher port distances), supporting our earlier observation that the model's class 3 definition doesn't fully align with domain logic.
- **Class 4 (anchoring)**: Clear dominance of topic **10** (low speeds, high `Log10Curvature`, low `DistStartTrendAngle` and `MaxDistPort`), entirely consistent with domain expectations.

Topic modeling proves valuable by uncovering feature interdependencies and frequently co-occurring conditions. Extracted topics reveal key interaction patterns and class associations, complementing synoptic overviews and interactive filtering by enabling analysts to move beyond individual feature examination toward holistic understanding of model logic and structure. This aligns with RuleSense's emphasis on intermediate and overall analysis levels [6], essential for models with thousands of rules.

5.7 Summary and Reflections

This case study illustrates how the RuleSense methodology enables analysts to evaluate alignment of a model with domain logic and expectations. Through overview visualizations, interactive filtering, and rule removal, we identified and eliminated logically flawed rules, such as those missing essential features or violating class definitions, without harming performance. Labeled data, when available, served as validation but was not necessary for detecting inconsistencies. More importantly, ensuring that the model's logic is sound increases the likelihood that it will behave reasonably on previously unseen inputs, whereas high accuracy on test data does not guarantee this. In many applications, correct and transparent logic is a critical requirement for high-stakes decision-making.

6 Case Study 2: Exploring Pandemic-Mobility Relationships

In this case study, we use the prototype implementation of the RuleSense approach to investigate the relationships between COVID-19 incidence levels and population mobility across Spanish provinces from April 2020 to May 2021. Our goal is not to examine the model itself, but to use it as a lens for understanding real-world dynamics reflected in the data.

6.1 Model and Data

We analyze a random forest model trained on publicly available data describing daily COVID-19 incidence and population mobility for 52 Spanish provinces [36–38]. The daily time series were transformed into sequences of discrete events lasting 1 to 21 weeks, where each event represents a stable incidence level (1–4). Each event was characterized by features reflecting disease and mobility levels in its temporal context, covering the six preceding weeks [2]. For disease incidence, only weeks -6 to -2 are used, excluding week -1 to account for the typical delay in implementing mobility policies in response to changing case numbers. We also include `Days_passed`, the number of days since the pandemic onset, to capture temporal evolution. Because the first wave had distinct characteristics (high incidence and strict lockdown), data before April 2020 were excluded from training.

The model predicting the next disease level achieved 91.45% accuracy on a set of 468 labeled instances. From this model, 7,173 rules were extracted and reduced to 4,097 after removing contradictions and subsumed rules.

6.2 Overview Visualization and Exploration of Rule Groups

Figure 11 (top left) shows a projection of rules by condition similarity [1], with dots colored by predicted disease level (1 in blue to 4 in red). Rules for levels 1 and 4 form relatively distinct clusters, while levels 2 and 3 are more intermixed. On the right, the feature-distribution view indicates that class 1 rules largely refer to the first half of the study period (roughly until late September 2020), whereas most class 4 rules pertain to July 2020–January 2021. Rules for class 3 peak in the second quarter of the period and decline thereafter. Level 2 often transitions to

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2026.3689736

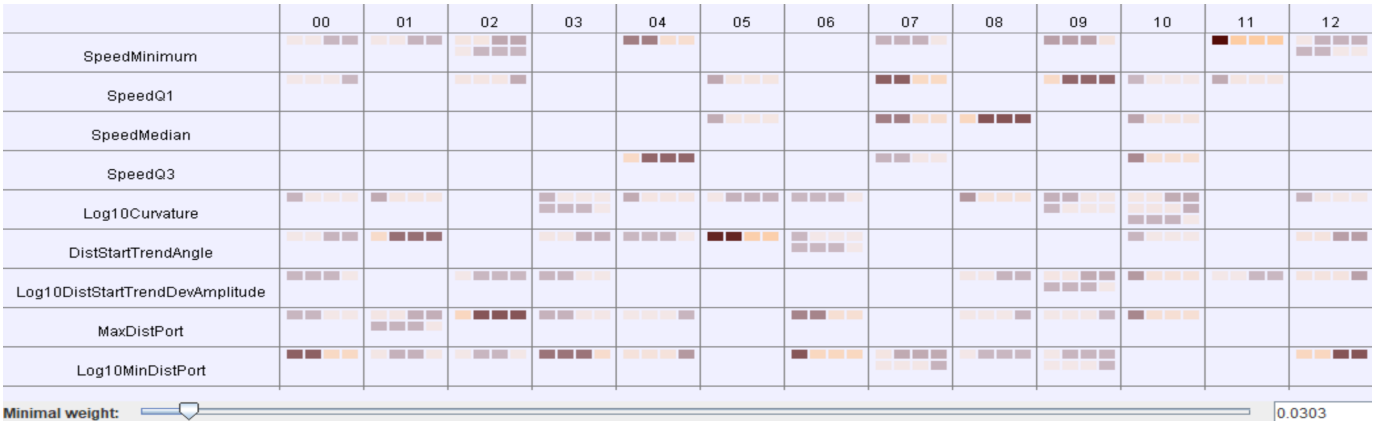

Fig. 8: Visualization of the topic–term matrix for a model with 13 topics. Columns represent topics, rows correspond to features, and cells contain sequences of light and dark rectangles encoding the binary representations of feature conditions.

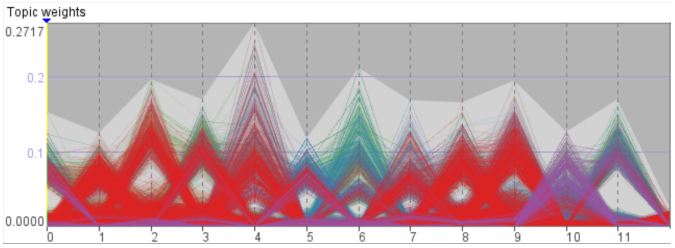

Fig. 9: Combinations of topic weights for individual rules, represented as lines in a line plot. The horizontal axis corresponds to topics (0–12), and topic weights are mapped onto vertical positions. Each line represents one rule and is colored according to its predicted class.

level 1 in the first half of the period and more often to level 3 from the second quarter onward. Overall, features representing recent disease levels, especially in week -2, appear more influential than mobility features.

To explore the meaning of distinct rule clusters, we selected subsets from the projection.

In the middle row of Fig. 11, we selected a group of rules predicting class 4. These rules mostly correspond to August 2020 onward and describe situations where a high pandemic level (e.g., level 3 in week -2) combined with high mobility led to further increase (class 4). In contrast, rules predicting class 2 in this period often featured lower mobility, consistent with expectations.

In the bottom row, we selected a cluster of class 1 rules, primarily from spring–summer 2020. These rules describe scenarios of decreasing pandemic levels (from 2 to 1 or 3 to 2), typically preceded by gradually decreasing mobility. These relationships are consistent with pandemic control dynamics in early 2020.

6.3 Assessing Feature Roles

Using interactive filtering and subset exploration in our RuleSense prototype, we examined the roles of specific features in the model logic.

Role of `Days_passed` The distribution overview in Fig. 11 (top right) shows that `Days_passed`, the temporal progression feature, is frequently used across all classes. To examine the model without this feature, we removed all rules involving `Days_passed`, obtaining a subset of 784 rules. As shown in Fig. 12, the distributions of other features remain similar to the full rule set, but mobility features, particularly for class 2, exhibit more distinct usage patterns, suggesting a non-negligible role in determining trends.

We then segmented the rule set into four periods by `Days_passed` (Fig. 13).

- **Period 1 (days 45–134, April–June 2020)**: Dominated by class 1 rules, corresponding to early post-lockdown recovery and decreasing incidence.
- **Period 2 (days 135–211, July–mid-September)**: Class 2 and 3 predictions increase as mobility rises.
- **Period 3 (days 212–349, mid-September 2020–January 2021)**: Peak of the second wave; class 4 dominates and class 1 disappears entirely.
- **Period 4 (days 350–431, February–May 2021)**: Class 4 declines; class 2 and 3 resurge despite continued high mobility.

These patterns illustrate how the model reflects the pandemic evolution and shifting behavioral norms over time.

Role of mobility features We also examined how mobility contributes to model predictions. Of the 4,097 cleaned rules, only 22 include all six weekly mobility features, while 101 include none. Most rules rely on mobility from only one or two weeks, most often week -6 or week -1. Filtering rule subsets by mobility inclusion shows that class distributions change only slightly, indicating that mobility, while frequently present, plays a secondary role compared to recent incidence.

To probe this further, we constructed a 2D projection of the rule set using only mobility-related features for similarity. In Fig. 14, we show three selections: rules with high mobility in week -1 (top), rules with low mobility in weeks -5 to -1 but higher values in week -6 (middle), and rules with elevated mobility in weeks -5 and -2 (bottom). Although these clusters differ in mobility patterns, their class compositions remain mixed. Mobility conditions alone thus do not strongly determine predicted outcomes; rather, they act as contextual features whose predictive role depends on the pandemic stage.

6.4 Exploration of Feature Interrelationships

We apply topic modeling to uncover feature co-occurrence patterns (see Section 4.4 for methodology). Running NMF iteratively for topic counts from 5 to 25 and assessing dominant topic weight distributions reveals that, while 15 topics maximize the 5th and 10th percentiles, 10 topics yield substantially higher values for the 15th–95th percentiles, suggesting they capture the most significant relationships. We therefore select the 10-topic model.

Figure 15 displays the topic-term matrix, while the histogram in Fig. 16 shows how frequently each topic is dominant across rules. The bars of the histogram are divided into segments corresponding to the counts of the rules predicting different disease levels. We immediately observe that Topic 0 is dominant in very few rules, indicating that it captures rare patterns that don't play significant role in predictions.

The topics presented in Fig. 15 mostly reveal relationships among either the COVID features or the mobility features while relationships between these groups of features are less pronounced. Topics **6** and

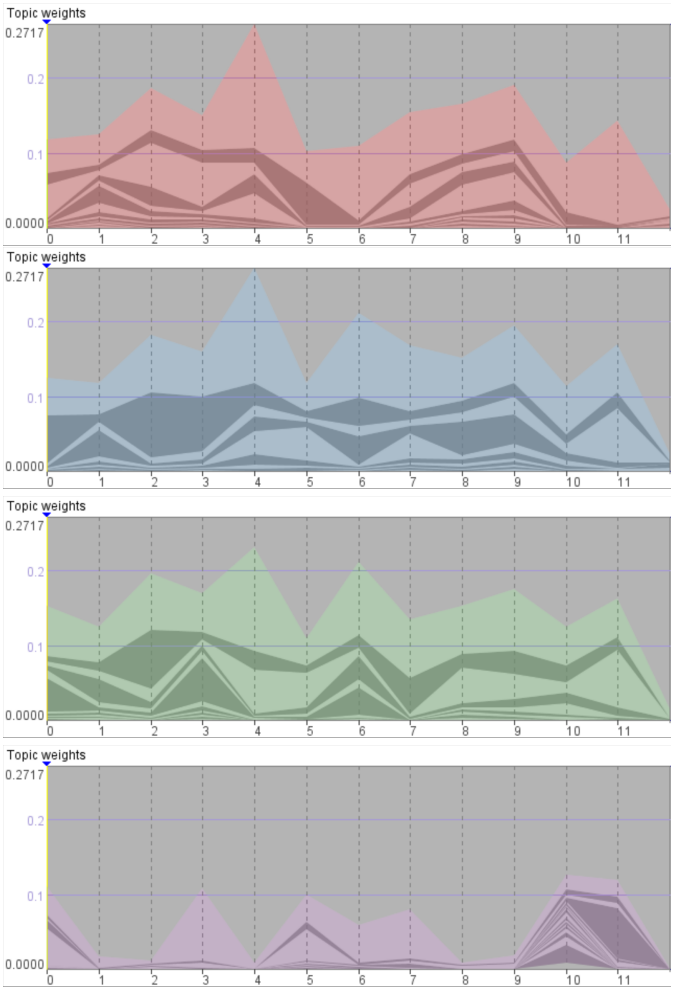

Fig. 10: Quantile plots showing the distributions of topic weights for rule subsets predicting different classes: forward movement (red), trawling (blue), port enter/exit (green), and anchoring (purple). Each plot displays 20 quantiles of topic weights per class, with alternating dark and light color bands representing increasing quantile levels.

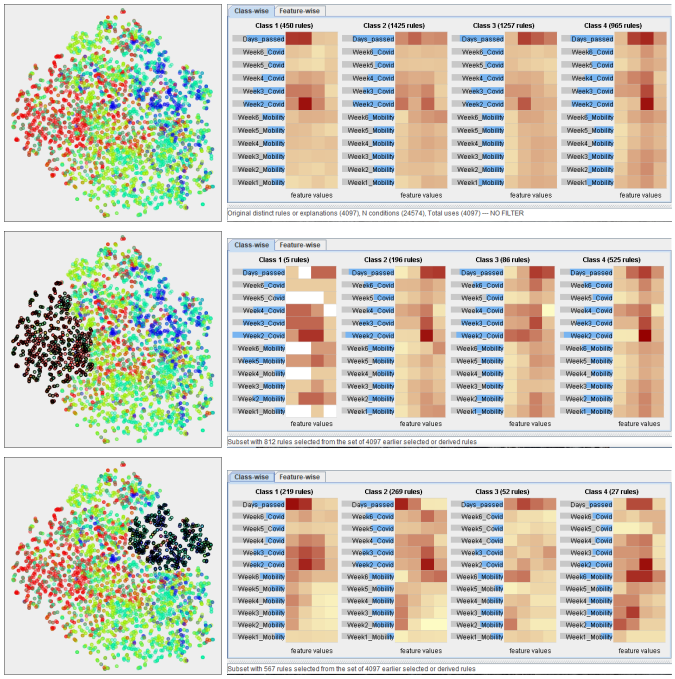

Fig. 11: Exploration of the COVID-19 and mobility model. Left: projection of rule set by similarity of rule conditions. Right: feature value distributions for the entire rule set (top) and for two selected clusters (middle and bottom).

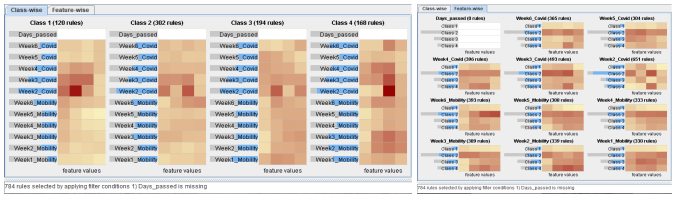

Fig. 12: Feature distributions over rules that do not involve `Days_passed`.

7, for example, encode stable COVID-19 incidence at levels 4 and 1, respectively, over five weeks. These correspond to the final and initial quarters of the study period. Topic **6** also links sustained high incidence with varied mobility in weeks -5 and -4. According to the histogram, this pattern often leads to predictions of level 3 and less often level 2. Topic **7**, on the other hand, is most frequently associated with transitions to level 2, though it also appears in rules predicting levels 3 and 1.

Other topics reflect various temporal trends in COVID-19 incidence. Topics **1** and **2** suggest that high incidence in week -6 may temporarily decrease before rising again, a pattern equally present in rules predicting levels 2 and 4. Topics **3** and **8** represent an increase from weeks -6 to -4 followed by a slight decline by week -2. Topic **9** captures a relationship between a high incidence (level 3) in week -2 and lower levels in week -6, without information about the intervening weeks. Generally, the model tells us that similar temporal patterns of the disease incidence may end up in both increase or decrease of the disease level.

Only topics **0** and **5** involve relationships between mobility features, but topic **0** is too rare to be of major interest. Topic **5** identifies a drop in mobility to the lowest level (1), primarily at the beginning of the study period. Its rules most often predict level 2, but levels 1, 3, and 4 also occur, indicating that mobility reduction alone is not strongly predictive of a specific outcome.

Finally, we consider how different topics relate to predictions of extreme outcomes. Rules predicting level 1 often involve either decreasing trends in incidence (topics 3 and 4), consistently low incidence (topic 7), or low mobility (topic 5). Predictions of level 4 are most frequently associated with increasing incidence (topics 1, 2, and 9), although topic 8 with a slight recent decline can also lead to such predictions.

This topic-based exploration confirms that the model primarily captures temporal dynamics in COVID-19 incidence, with relatively weak linking to mobility trends. Feature co-occurrence patterns disclose the complexity of the decision logic: different combinations of conditions can lead to similar outcomes, and similar temporal trends can lead to different predictions depending on subtle variations. Topic modeling thus serves as a useful complement to direct rule filtering and visual inspection by capturing more complex patterns of feature interaction.

6.5 Summary and Reflections

This case study demonstrates how the RuleSense approach can help analysts in using rule-based models to explore and understand real-world phenomena. Through overview visualizations, filtering, and rule projection, we identified shifts over time in the model's reliance on mobility versus incidence features, broadly aligning with epidemiological understanding: mobility influenced predictions more in early 2020, while disease trends dominated later.

Topic modeling complemented these insights by revealing latent patterns in rule structure, such as persistent high or low incidence, short-term trend reversals, and varying roles of mobility. It showed that similar temporal trajectories can lead to different outcomes, highlight-

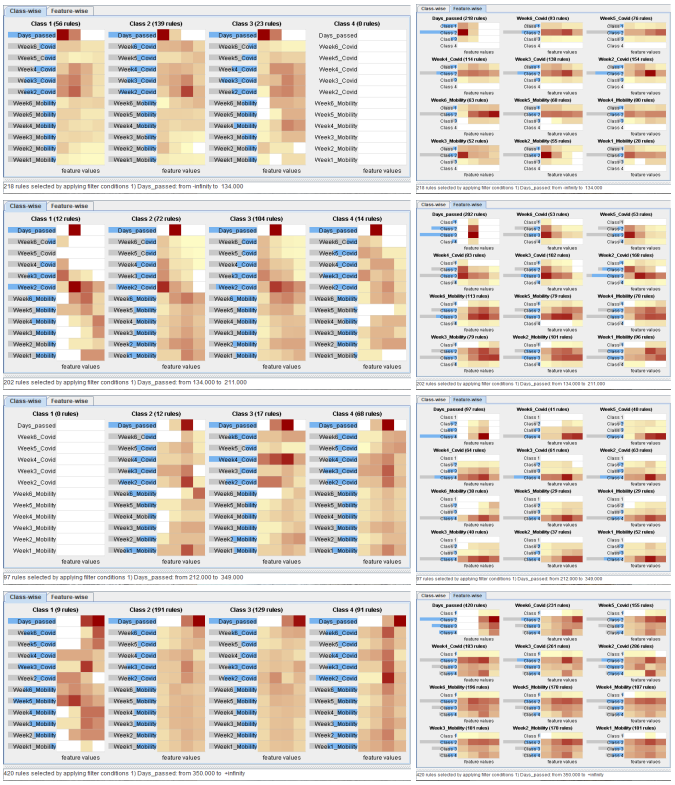

Fig. 13: Feature distributions in rules grouped by different ranges of `Days_passed`.

ing the complexity of the model's internal logic.

This kind of post-hoc exploration of decision rules goes beyond traditional instance-based explainability. With the RuleSense methodology, analysts can reason about model behavior at a conceptual level and relate patterns in the rule logic to underlying epidemiological dynamics.

7 Expert Evaluation

We evaluated the potential of RuleSense for logic-centered model exploration using a "Small-N" qualitative approach involving five high-level experts. This strategy aligns with the MILCs framework [43], which emphasizes that the value of visual analytics lies in supporting complex sensemaking and insight discovery rather than low-level usability [33]. In our context, the "data" being analyzed and explored is the trained model itself.

7.1 Study Design and Participants

Following the Lam et al. taxonomy [21], we conducted two studies focusing on "Visual Data Analysis and Reasoning" and "Discovery and Insight."

7.1.1 Study 1: Model Logic Auditing (ML Experts)

The first study investigated whether our approach supports expert reasoning about model internal logic, using the vessel movement classification model as a representative large rule-based model derived from a random forest. We assessed the methodology's capacity for **sensemaking** (understanding model internal logic), **trust calibration** (evaluating model plausibility and risk), and **actionability** (supporting model intervention decisions). This focus on high-level cognitive tasks follows guidelines for evaluating visual data analysis and reasoning emphasizing exploration and knowledge discovery rather than task completion efficiency [42].

Participants and Procedure. Four ML experts participated:

- **E1**, a mathematician and industrial consultant;
- **E2**, a healthcare AI developer;
- **E3**, a ML researcher participating in applied model development projects; and
- **E4**, a ML researcher and data scientist specializing in applied industrial and business analytics.

None had prior exposure to interactive visual techniques for model exploration. Participants reviewed an illustrated slide deck and responded to structured questions. E1 engaged in a three-hour interview; E2, E3, and E4 provided detailed written feedback, with E3 also participating in follow-up discussions. A report about the interview with E1 and the original responses of E2–E4 are provided in the supplementary materials.

Focus. The study aimed to determine if RuleSense enables experts to: (1) reason about logic at global and intermediate levels; (2) identify contradictions or implausible patterns; and (3) judge model trustworthiness beyond accuracy metrics.

7.1.2 Study 2: Domain-Oriented Interpretation (Epidemiology Expert)

The second study examined the use of a trained model (COVID-19 incidence in Spain) as a "lens" for understanding real-world phenomena. This case study approach is considered a "gold standard" for assessing how VA tools align with complex domain knowledge [17].

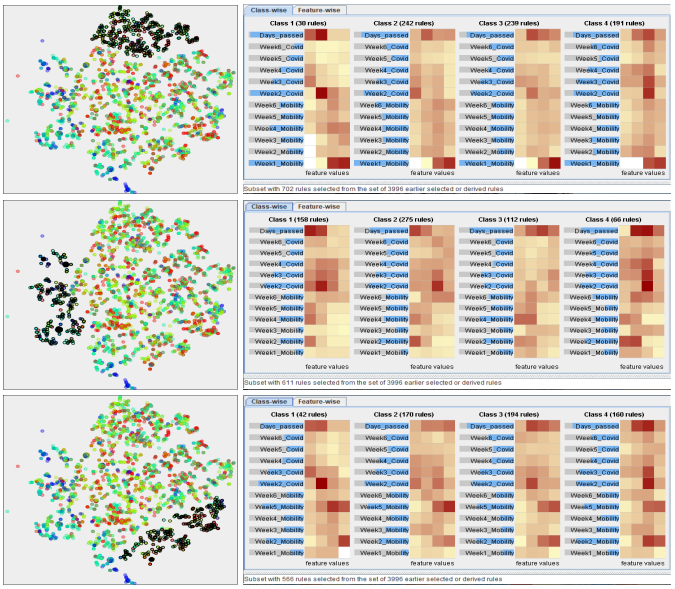

Fig. 14: Mobility-only projection and distributions for selected rule subsets.

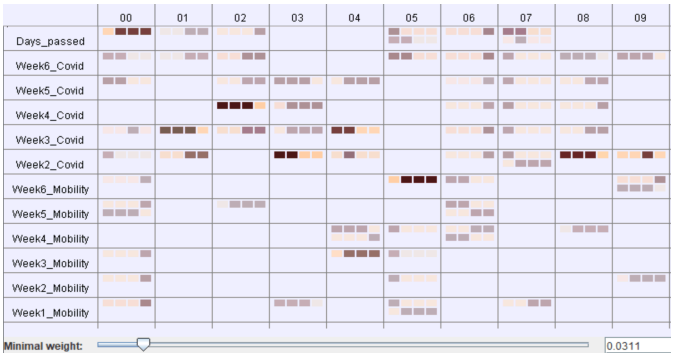

Fig. 15: Ten topics extracted from the COVID-19 prediction rule set.

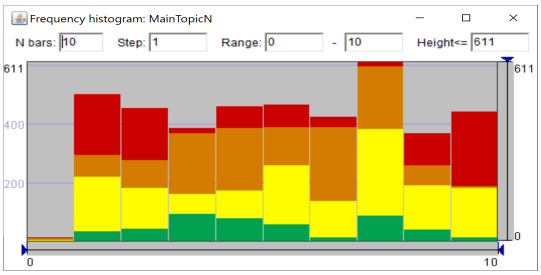

Fig. 16: Histogram of dominant topic counts across rules. Bar segments are color-coded by predicted disease level: 1 (green), 2 (yellow), 3 (orange), 4 (red).

Participant and Procedure. A epidemiology expert (**E5**) with deep knowledge of the pandemic dynamics in Spain participated in a 1.5-hour online session. Using a tailored slide deck, we explored the model's alignment with known public health responses and pandemic phases. A detailed report about the interview with E5 is provided in the supplementary materials.

Focus. The evaluation centered on "domain alignment": whether the model's decision patterns reflected known epidemiological dynamics and if interactive exploration enabled meaningful discovery regarding the interplay of disease and mobility.

7.2 Evaluation Results: Model Logic Auditing (ML Experts)

The feedback from the ML experts (E1–E4) confirmed that RuleSense provides a necessary "intermediate" level of abstraction for model auditing, filling the gap between high-level global metrics and low-level instance explanations.

7.2.1 Sensemaking and Logic Understanding

All experts agreed that the system supports a deep understanding of model logic. E1 initially found the visual representations non-intuitive but eventually concluded they were "clearly useful" for increasing model confidence.

Feature distribution overview. E2, E3, and E4 noted that the feature distribution overview allows for a quick assessment of whether a model relies on reasonable feature ranges in class predictions. E3 specifically noted that this view serves as an "early warning system" for detecting if a model uses wrong intervals for a particular class. E4 highlighted that it is particularly effective for detecting *spurious correlations* and "Clever Hans" effects, where a model might be contaminated by simple cues without underlying causality.

Interactive filtering. While E1 and E2 emphasized the superiority of interactive filtering over static lists, E3 viewed filtering primarily as a way to verify domain-specific expectations (e.g., ensuring certain features are present or absent where expected). E4 emphasized that interactive filtering is the only way to make the investigation of large rule sets feasible. He noted that filtering allows for studying feature interactions, for instance, observing how excluding top features leads to sharper distributions in others, thereby uncovering hidden dependencies.

Topic modeling. While the *purpose* of using topic modeling - to uncover higher-order feature interactions - was grasped easily, the interpretation of the results proved challenging. Particularly, E1 and E3 experienced significant initial difficulty; we provided extensive explanations during interview sessions and online conversations.

Once explained, E1 praised the feature-topic matrix as an effective alternative to traditional word clouds, though experts noted cognitive load remains high when mapping abstract topics back to specific domain behaviors. Aggregated line plots (quantile-based bands) were difficult to interpret without step-by-step walkthrough. E3 characterized topic modeling as an "insightful test" to see if the model's recurring condition combinations align with human-understandable logic. E4 argued that topic modeling provides significant value by identifying recurring sub-strategies (fragments) across a larger set of rules, offering a broader overview than filtered subsets alone.

7.2.2 Trust Calibration and Anomaly Detection

The experts used the tools to identify "early warning signs" that might be missed by accuracy-based validation:

Rule projection. Opinions on the 2D projection were mixed. E1 and E3 found that the projection illustrates the "consistency of class descriptions." E1 noted that the proximity of domain-related classes (e.g., "anchoring" and "port exit") increased his trust. E3 highlighted that class mixing (overlaps) in the projection is particularly useful for identifying the specific reasons for confusion between classes.

E4 was more critical, noting that the projection appeared too complex for a clear structure to emerge. He observed significant "class mixing" in the projection, which he interpreted as a sign that the distance metric between rules might not always resolve class differences, or that the model itself relies on contradictory logic for similar data regions.

Contradictions and redundancies. Opinions on contradictions varied among the experts. E2 viewed them as a "make me listen up" signal of data bias. E3, however, noted that for Random Forests, contradictions are often implied and expected, but their presence in RuleSense helps signal "unstable reasoning" or "corner cases" in the training data. E4 provided a nuanced view on contradictions, suggesting they should be seen as a measure of *uncertainty* or fuzzy data domains inherent to ensemble models like Random Forests. He noted that while contradictions are expected, RuleSense helps identify if they occur in regions that defy domain knowledge.

E3 noted that "automatic cleaning" (removing redundancies) is helpful because it allows the expert to "concentrate on the interesting details" rather than being overwhelmed by noise. However, E4 argued that, while removing subsumed rules is helpful for cognitive load, it carries a risk of changing the "probabilistic distribution" of the overall ensemble. This technical aspect was not raised by other experts.

Trust beyond accuracy. E3 stated that if the analysis reveals significant "confusion" in the model's logic, they would use the insights to generate counterfactual data points to test the model's failure limits before deployment. E2 similarly remarked that surprising implausibilities would make them hesitate to deploy even a highly accurate model.

7.2.3 Actionability and Limitations

A recurring theme was the relationship between logic auditing and data validation. E1 and E3 emphasized that logic analysis points toward "problematic data inputs" rather than immediate model modifications. E4 argued that interactive exploration is essential because there is a fundamental trade-off between too many individual rules to investigate and a too-abstract representation that hides critical flaws.

The experts also identified potential risks and areas for improvement:

- **Interpretability and training:** E1 and E2 noted that some views (quantile plots) require training. E3 pointed out that it is unrealistic to expect a model to create conditions exactly as a human would, which can make some topics identified by the topic modeling appear unclear.
- **Algorithmic artifacts:** All experts warned about over-interpretation. E3 and E4 specifically warned that dimensionality reduction and topic modeling can introduce their own biases, meaning conclusions drawn from the spatial location of rules or topic groupings must be handled with care. E4 noted that without tracing rules back to ground-truth data, conclusions about why a rule excludes certain ranges remain speculative.
- **Blind spots:** E3 noted that these techniques do not well-capture "unknown unknowns", that is, how the model will behave on outlier samples that fall entirely outside the rules' current logic.
- **The ensemble nature:** E4 pointed out that RuleSense treats rules in isolation and does not fully capture the *weighted ensemble nature* of the final prediction, which might impact the understanding of the model's ultimate decision-making process.

Overall, the ML experts characterized RuleSense as a powerful methodology for intermediate-level reasoning, providing a structured way to verify whether a model's decisions are grounded in "expected domain logic" or "spurious correlations."

7.3 Evaluation Results: Domain-Oriented Interpretation (E5)

The second study evaluated RuleSense as a tool for domain-driven discovery, using a COVID-19 incidence prediction model for Spain as a "lens" to study real-world pandemic dynamics. The participant (E5), an expert in epidemiology, used the system to verify if the model's data-driven logic aligned with known public health phenomena.

7.3.1 Domain Alignment and Pandemic Phases

The expert confirmed that the global feature distributions and temporal rule subsets (segmented by `Days_passed`) reflected the broad stages of the pandemic in Spain. Specifically, RuleSense successfully captured the transition from the early post-lockdown decline (Class 1) to the second wave peak (Class 4). E5 noted that the model correctly identified disease history as a stronger predictor than internal provincial mobility, the latter being most relevant only during the initial strict lockdown.

7.3.2 Model Diagnosis and Structural Limitations

A key outcome of the study was using RuleSense for "model diagnosis." By exploring rule subsets where absolute time was removed, the expert identified a significant structural flaw: the model failed to capture local temporal trends (e.g., the slope of incidence change).

- **Feature interaction issues:** E5 observed that while the model had access to weekly incidence levels, it treated them as independent variables. RuleSense visualizations (heatmaps and topics) exposed that the model logic relied on static "stable" levels rather than coherent transitions.
- **Topic modeling utility:** Similar to the ML experts, E5 found understanding of topic modeling to be a high-effort task. While the results reinforced the lack of temporal trends (showing topics of stable high/low incidence), the expert found it difficult to extract new epidemiological insights from the fragmented topics.
- **Missing context:** The analysis led to a high-level dialogue about feature engineering. E5 pointed out that the "weak" influence of mobility in the model was likely due to the omission of inter-regional flows, a domain insight that surfaced specifically through the inspection of the feature distribution heatmaps.

7.3.3 Accessibility and the Role of Mediation

A significant observation was the steep learning curve associated with the more advanced visual analytics components. E5 noted that both the conceptual underpinnings of topic modeling and the interpretation of the resulting visualizations were challenging. It became evident that a domain expert without a background in machine learning or visual analytics would likely struggle to use these tools and interpret their results independently.

In the current study, the presence and active support of the authors were necessary to facilitate the expert's reasoning process. This highlights a critical area for future research: making advanced model-auditing techniques accessible to domain specialists through improved UI design, automated guidance, or more intuitive representations that do not rely on a mediator to bridge the gap between model-centric data and domain-centric knowledge.

7.3.4 Utility for Expert-Model Dialogue

Despite the identified model weaknesses and accessibility hurdles, the expert's assessment of the RuleSense was highly positive. E5 concluded that the visualizations were "valuable for diagnosing model behavior and understanding what the model has actually learned."

Rather than just confirming accuracy, the session demonstrated that **RuleSense enables a "diagnostic reasoning" workflow**: the expert used the visual evidence to suggest specific model improvements, such as explicitly encoding temporal slopes and incorporating spatial mobility flows. This confirms the methodology's utility in supporting a "meaningful dialogue" between domain experts and complex models, even when those models are imperfect.

8 Discussion and Conclusions

Most prior work on interpretable machine learning and model examination focuses on two dominant goals: improving predictive performance or investigating model behavior based on available labeled data. These efforts have led to many useful tools for debugging models or explaining individual decisions. Less attention has been paid to examining the intrinsic knowledge structures and reasoning patterns learned by a model, especially their alignment with domain logic and their potential for supporting knowledge discovery.

Rather than relying on data-driven metrics like test accuracy or instance coverage, we take a different focus: examine the internal structure and meaning of the rules themselves making decision logic transparent even when ground truth data is unavailable.

Role of Domain Knowledge RuleSense doesn't require formally encoded domain knowledge. Instead, domain expertise enters primarily during interpretation and judgment. The framework helps analysts identify structural properties, such as feature reliance patterns, contradictions, redundancies, recurring conditions, that can be examined even with limited domain background. However, deeper domain knowledge becomes essential when evaluating whether these patterns make sense. Do certain feature combinations align with known mechanisms? Are temporal relationships plausible?

RuleSense is therefore designed to guide expert attention to potentially critical parts of the model logic, not to automate domain validation. Alignment with domain reasoning ultimately requires human judgment, which the framework aims to support rather than replace.

Rule Removal and Logical Inconsistencies Our approach does not prescribe automatic removal or post-hoc pruning of rules. In ensemble models such as Random Forests, overlapping or even contradictory rules are an expected consequence of combining multiple weak learners. In our framework, such rules are treated as diagnostic signals that may indicate unstable regions of the input space, data ambiguity, or potential biases, rather than as errors to be eliminated. Our empirical studies (particularly ML experts E3 and E4 in Study 1) confirmed this perspective: contradictions are often inherent outcomes of the learning process rather than modeling errors.

Decisions to remove or ignore rules should be **made by human experts** and only when representative validation or test data are available, so that the impact on predictive behavior can be assessed. Without such data-based verification, rule removal is risky and should generally be avoided. In such cases, the main value of RuleSense lies in improving model understanding and trust calibration, not in modifying the model itself.

Key Insights from Empirical Studies Our empirical studies yielded several complementary insights regarding the role of logic-centered model exploration in expert reasoning and model assessment.

RuleSense enables diagnostic reasoning, not just accuracy checking. Across both studies, experts emphasized that RuleSense shifted the focus from verifying predictive accuracy to diagnosing what a model has actually learned. In Study 2, this diagnostic capability exposed a fundamental limitation: the model couldn't capture temporal trends. This insight allowed the domain expert to propose concrete improvements, such as adding temporal slope features and incorporating inter-regional mobility data. The tool moved the expert from assessment to actionable refinement. In this sense, RuleSense facilitated a transition from passive assessment to actionable model refinement.

Logical contradictions signal instability rather than errors. ML experts considered RuleSense as a means to identify regions of unstable reasoning, ambiguous decision boundaries, or potential data biases. By making such contradictions visible at intermediate levels of abstraction, the system supported informed judgment and prompted deeper investigation of borderline cases.

Higher-order logic analysis entails substantial cognitive effort and benefits from mediation. Both studies showed that techniques intended to expose higher-order structures, such as rule projection and topic modeling, require considerable interpretive effort. Experts recognized their potential for uncovering feature interactions and recurring decision fragments, but several participants noted that meaningful interpretation often depended on guidance from someone familiar with both the analytical techniques and the domain. This was particularly evident in the epidemiology study, where the domain expert relied on mediation to connect abstract topic structures to epidemiological mechanisms. These findings suggest that such techniques are currently best suited for collaborative or expert-mediated settings and motivate future work on automated guidance, explanation, and more intuitive representations.

Intermediate abstractions are essential for auditing large rule-based models. A recurring theme across both studies was the necessity of operating at global and intermediate levels of abstraction. Experts consistently noted that individual rules are too numerous and granular to support meaningful reasoning, while overly aggregated metrics obscure critical flaws. RuleSense was valued for providing this middle ground, enabling experts to reason about feature usage patterns, rule subsets, and recurring condition combinations in a way balancing scalability with interpretability.

Overall, the studies indicate that RuleSense is most effective when used as a diagnostic and exploratory instrument: it does not replace data-driven validation or domain expertise, but complements them by structuring expert attention, revealing latent logical patterns, and supporting informed judgment about model trustworthiness and limitations.

Contributions This work offers four main contributions:

- **Scalable rule set exploration**: Interactive visualizations that provide high-level overviews and detailed drill-downs for thousands of rules.
- **Investigation of feature interdependencies**: Topic modeling and similarity measures that reveal hidden relationships influencing model decisions.
- **Logical consistency checking**: Tools for detecting inconsistent rules and identifying "corner case" instabilities.
- **Domain-driven model diagnosis**: A human-in-the-loop workflow where experts investigate model logic for discovery and structural refinement.

Limitations Our approach has clear limitations. The investigation process remains labor-intensive and difficult to scale without further automation. Our contradiction detection relies on heuristic assumptions that may not generalize to all domains. Topic modeling, while powerful, demands substantial cognitive effort, as noted by all experts. Moreover, our empirical evaluation involved a limited number of participants and application domains, so broader generalization requires further study.

Software Scalability and Performance Considerations. RuleSense is implemented as a proof-of-concept prototype for exploratory research. For rule sets of several thousand rules, most interactions (filtering, overview updates, linked highlighting) produce immediate responses. Computationally intensive operations (contradiction detection and rule embedding) currently require several minutes and execute on demand in background mode while users continue interacting with available displays. These could be potentially optimized through indexing or incremental computation. As a conceptual framework, RuleSense components can be re-implemented or scaled independently to meet specific application requirements.

Conclusion RuleSense demonstrates that logic-centered model exploration is both feasible and valuable. Our studies show that this approach can yield meaningful insights into a model's internal reasoning. To achieve broader applicability, further advances are needed in three areas: abstraction (richer intermediate representations), automation (more guidance and pattern detection), and human-centered design (interfaces and explanations that reduce cognitive load). Progress along these dimensions will help make logic-based model examination more accessible to domain experts without requiring constant technical mediation.

Acknowledgments

This work was supported by Federal Ministry of Education and Research of Germany and the state of North-Rhine Westphalia as part of the *Lamarr Institute for Machine Learning and Artificial Intelligence* (Lamarr22B), and by EU in project *CrexData* (grant no. 101092749).

References

[1] L. Adilova, M. Kamp, G. Andrienko, and N. Andrienko. Re-interpreting rules interpretability. *International Journal of Data Science and Analytics*, Jul 2023. doi: 10.1007/s41060-023-00398-5 2, 4, 5, 9

[2] N. Andrienko and G. Andrienko. Exploring relationships between events in context. In M. El-Assady and H.-J. Schulz, eds., *EuroVis Workshop on Visual Analytics (EuroVA)*. The Eurographics Association, 2024. doi: 10.2312/eurova.20241114 9

[3] N. Andrienko, G. Andrienko, L. Adilova, and S. Wrobel. Visual analytics for human-centered machine learning. *IEEE Computer Graphics and Applications*, 42(1):123–133, 2022. doi: 10.1109/MCG.2021.3130314 2

[4] N. Andrienko, G. Andrienko, A. Artikis, P. Mantenoglou, and S. Rinzivillo. Human-in-the-loop: Visual analytics for building models recognizing behavioral patterns in time series. *IEEE Computer Graphics and Applications*, 44(3):14–29, 2024. doi: 10.1109/MCG.2024.3379851 6

[5] N. Andrienko, G. Andrienko, and G. Shirato. Episodes and topics in multivariate temporal data. *Computer Graphics Forum*, 42(6):e14926, 2023. doi: 10.1111/cgf.14926 3

[6] J. Bertin. *Semiology of Graphics*. University of Wisconsin Press, 1983. 2, 9

[7] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001. doi: 10.1023/A:1010933404324 2

[8] S. Chen, N. Andrienko, G. Andrienko, L. Adilova, J. Barlet, J. Kindermann, P. H. Nguyen, O. Thonnard, and C. Turkay. Lda ensembles for interactive exploration and categorization of behaviors. *IEEE Transactions on Visualization and Computer Graphics*, 26(9):2775–2792, 2020. doi: 10.1109/TVCG.2019.2904069 3

[9] T.-H. Chen, S. W. Thomas, and A. E. Hassan. A survey on the use of topic models when mining software repositories. *Empirical Software Engineering*, 21(5):1843–1919, 2016. 3

[10] J. Choo, C. Lee, C. K. Reddy, and H. Park. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE transactions on visualization and computer graphics*, 19(12):1992–2001, 2013. doi: 10.1109/TVCG.2013.212 3

[11] D. Collaris and J. J. van Wijk. ExplainExplore: Visual exploration of machine learning explanations. In *2020 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 26–35, 2020. doi: 10.1109/PacificVis48177.2020.7090 2

[12] R. Egger and J. Yu. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify twitter posts. *Frontiers in Sociology*, 7, 2022. doi: 10.3389/fsoc.2022.886498 6

[13] J. Eirich, M. Münch, D. Jäckle, M. Sedlmair, J. Bonart, and T. Schreck. RfX: A design study for the interactive exploration of a random forest to enhance testing procedures for electrical engines. *Computer Graphics Forum*, 41(6):302–315, 2022. doi: 10.1111/cgf.14452 2, 3

[14] M. El-Assady, F. Sperrle, O. Deussen, D. Keim, and C. Collins. Visual analytics for topic model optimization based on user-steerable speculative execution. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):374–384, 2019. doi: 10.1109/TVCG.2018.2864769 3

[15] R. Gurung, T. Lindgren, and H. Boström. An interactive visual tool to enhance understanding of random forest predictions. In *European Conference on Data Analysis (ECDA)*, 2019. doi: 10.5445/KSP/1000098011/082

[16] M. Haddouchi and A. Berrado. A survey and taxonomy of methods interpreting random forest models, 2024. doi: 10.48550/arXiv.2407.12759 2, 3

[17] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Möller. A systematic review on the practice of evaluating visualization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2818–2827, 2013. doi: 10.1109/TVCG.2013.126 12

[18] W. Jentner, G. Lindholz, H. Hauptmann, M. El-Assady, K.-L. Ma, and D. Keim. Visual analytics of co-occurrences to discover subspaces in

structured data. *ACM Trans. Interact. Intell. Syst.*, 13(2), article no. 10, 49 pages, June 2023. doi: 10.1145/3579031 3

[19] B. La Rosa, G. Blasilli, R. Bourqui, D. Auber, G. Santucci, R. Capobianco, E. Bertini, R. Giot, and M. Angelini. State of the art of visual analytics for explainable deep learning. *Computer Graphics Forum*, 42(1):319–355, 2023. doi: 10.1111/cgf.14733 2

[20] H. Lakkaraju, S. H. Bach, and J. Leskovec. 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*, KDD '16, 10 pages, p. 1675–1684. Association for Computing Machinery, New York, NY, USA, 2016. doi: 10.1145/2939672.2939874 2

[21] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics*, 18(9):1520–1536, 2012. doi: 10.1109/TVCG.2011.279 12

[22] Z. Li, W. Yang, J. Yuan, J. Wu, C. Chen, Y. Ming, F. Yang, H. Zhang, and S. Liu. RuleExplorer: A scalable matrix visualization for understanding tree ensemble classifiers, 2024. 2, 3

[23] L. Liu, L. Tang, W. Dong, S. Yao, and W. Zhou. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 5:1–22, 2016. 3

[24] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. Explainable ai for trees: From local explanations to global understanding, 2019. doi: 10.48550/arXiv.1905.04610 2

[25] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. NIPS'17. Curran Associates Inc., Red Hook, NY, USA, 2017. 5

[26] M. Luo, F. Nie, X. Chang, Y. Yang, A. Hauptmann, and Q. Zheng. Probabilistic non-negative matrix factorization and its robust extensions for topic modeling. In *Thirty-first AAAI conference on artificial intelligence*, 2017. doi: 10.1609/aaai.v31i1.10832 6

[27] D. Mazumdar, M. P. Neto, and F. V. Paulovich. Random forest similarity maps: A scalable visual representation for global and local interpretation. *Electronics*, 10(22):2862, 2021. doi: 10.3390/electronics10222862 2, 3

[28] C. Maçãs, J. R. Campos, N. Lourenço, and P. Machado. Visualisation of random forest classification. *Information Visualization*, 23(4):312–327, 2024. doi: 10.1177/14738716241260745 2

[29] Y. Ming, H. Qu, and E. Bertini. RuleMatrix: Visualizing and understanding classifiers with rules. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):342–352, 2019. doi: 10.1109/TVCG.2018.2864812 2, 4

[30] I. Mollas, N. Bassiliades, and G. Tsoumakas. Conclusive local interpretation rules for random forests. *Data Mining and Knowledge Discovery*, 36(4):1521–1574, 2022. doi: 10.1007/s10618-022-00839-y 2

[31] L. Moussavi, G. Andrienko, N. Andrienko, and A. Slingsby. Visually-supported topic modeling for understanding behavioral patterns from spatio–temporal events. *Computers and Graphics*, 129:104245, 2025. doi: 10.1016/j.cag.2025.104245 3

[32] M. P. Neto and F. V. Paulovich. Explainable matrix - visualization for global and local interpretability of random forest classification ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1427–1437, 2021. doi: 10.1109/TVCG.2020.3030354 2, 3, 4

[33] C. North. Toward measuring visualization insight. *IEEE Computer Graphics and Applications*, 26(3):6–9, 2006. doi: 10.1109/MCG.2006.70 12

[34] R. H. Nsch, P. Wiesner, S. Wendler, and O. Hellwich. Colorful trees: Visualizing random forests for analysis and interpretation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 294–302, 2019. doi: 10.1109/WACV.2019.00037 2

[35] L. Padua, H. Schulze, K. Matković, and C. Delrieux. Interactive exploration of parameter space in data mining: Comprehending the predictive quality of large decision tree collections. *Computers & Graphics*, 41:99–113, 2014. doi: 10.1016/j.cag.2014.02.004 2

[36] M. Ponce-de Leon, J. del Valle, J. M. Fernández, M. Bernardo, D. Cirillo, J. Sanchez-Valle, M. Smith, S. Capella-Gutierrez, T. Gullón, and A. Valencia. COVID19 Flow-Maps daily cases reports, 2021. doi: 10.5281/zenodo.5217386 9

[37] M. Ponce-de Leon, J. del Valle, J. M. Fernández, M. Bernardo, D. Cirillo, J. Sanchez-Valle, M. Smith, S. Capella-Gutierrez, T. Gullón, and A. Valencia. COVID19 Flow-Maps daily-mobility for Spain, 2021. doi: 10.5281/zenodo.5539411 9

[38] M. Ponce-de Leon, J. del Valle, J. M. Fernández, M. Bernardo, D. Cirillo, J. Sanchez-Valle, M. Smith, S. Capella-Gutierrez, T. Gullón, and A. Valencia. COVID19 Flow-Maps population data, 2021. doi: 10.5281/zenodo.5226351 9

[39] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. ACM, 2016. doi: 10.1145/2939672.2939778 5

[40] P. C. Ribeiro, G. G. Schardong, S. D. Barbosa, C. S. de Souza, and H. Lopes. Visual exploration of an ensemble of classifiers. *Computers & Graphics*, 85:23–41, 2019. doi: 10.1016/j.cag.2019.08.012 2

[41] C. Rudin and Y. Shaposhnik. Globally-consistent rule-based summary-explanations for machine learning models: application to credit-risk evaluation. *Journal of Machine Learning Research*, 24(16):1–44, 2023. 2

[42] M. Sedlmair, M. Meyer, and T. Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2431–2440, 2012. doi: 10.1109/TVCG.2012.213 12

[43] B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization*, BELIV '06, 7 pages, p. 1–7. Association for Computing Machinery, New York, NY, USA, 2006. doi: 10.1145/1168149.1168158 12

[44] I. Vayansky and S. A. Kumar. A review of topic modeling methods. *Information Systems*, 94:101582, 2020. doi: 10.1016/j.is.2020.101582 3

[45] J. Yuan, B. Barr, K. Overton, and E. Bertini. Visual exploration of machine learning model behavior with hierarchical surrogate rule sets. *IEEE Transactions on Visualization and Computer Graphics*, 30(2):1470–1488, 2024. doi: 10.1109/TVCG.2022.3219232 2, 3

[46] J. Yuan, O. Nov, and E. Bertini. An exploration and validation of visual factors in understanding classification rule sets. In *2021 IEEE Visualization Conference (VIS)*, pp. 6–10, 2021. doi: 10.1109/VIS49827.2021.9623303 2

[47] J. Yuan, O. Nov, and E. Bertini. Visualizing rule sets: Exploration and validation of a design space, 2021. doi: 10.48550/arXiv.2103.01022 2, 3, 4

[48] X. Zhao, Y. Wu, D. L. Lee, and W. Cui. iForest: Interpreting random forests via visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):407–416, 2019. doi: 10.1109/TVCG.2018.2864475 2, 3

BIOGRAPHIES

[]Natalia Andrienko is a lead scientist at Fraunhofer Institute for Intelligent Analysis and Information Systems, part-time professor at City St George's University London, and Area Chair at Lamarr Institute for Machine Learning and Artificial Intelligence. Results of her research have been published in two monographs, "*Exploratory Analysis of Spatial and Temporal Data: a Systematic Approach*" (2006) and "*Visual Analytics of Movement*" (2013). Natalia Andrienko is an associate editor of *Computer Graphics Forum*.

[]Gennady Andrienko is a lead scientist responsible for visual analytics research at Fraunhofer Institute for Intelligent Analysis and Information Systems, part-time professor at City St George's University London, and a Co-PI at Lamarr Institute for Machine Learning and Artificial Intelligence. Gennady Andrienko was a paper chair of *IEEE VAST* conference (2015–2016) and associate editor of *IEEE Transactions on Visualization and Computer Graphics* (2012–2016), *Information Visualization* and *International Journal of Cartography*.

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2026.3689736

[ ]Bahavathy Kathirgamanathan is a postdoctoral researcher at the University of Cologne, Germany working on topics such as Explainable AI, Visual Analytics, and Human-Centered AI. She was previously a researcher at Fraunhofer Institute IAIS in Sankt Augustin, Germany. Contact her at bkathirg@uni-koeln.de