

DEPARTMENT: DISSERTATION IMPACT

Visual Analytics for Explainable and Trustworthy Artificial Intelligence

Angelos Chatzimparmpas , Utrecht University, 3584 CC, Utrecht, The Netherlands

FROM THE EDITOR

Angelos Chatzimparmpas received the 2024 EuroVis Best Ph.D. Dissertation Award.

Our society increasingly depends on intelligent systems to solve complex problems, ranging from recommender systems suggesting the next movie to watch to AI models assisting in medical diagnoses for hospitalized patients. With the iterative improvement of diagnostic accuracy and efficiency, AI holds significant potential to mitigate medical misdiagnoses by preventing numerous deaths and reducing an economic burden of approximately € 450 billion annually. However, a key obstacle to AI adoption lies in the lack of transparency, that is, many automated systems provide predictions without revealing the underlying processes. This opacity can hinder experts' ability to trust and rely on AI systems. Visual analytics (VA) provides a compelling solution by combining AI models with interactive visualizations. These specialized charts and graphs empower users to incorporate their domain expertise to refine and improve the models, bridging the gap between AI and human understanding. In this work, the author defines, categorizes, and explores how VA solutions can foster trust across the stages of a typical AI pipeline. The author proposes a design space for innovative visualizations and presents an overview of our previously developed VA dashboards, which support critical tasks within the various pipeline stages, including data processing, feature engineering, hyperparameter tuning, understanding, debugging, refining, and comparing models.

The rapid improvement in hardware technology and the increased availability of computational resources have revamped the field of AI in recent years. These advancements did not go

unnoticed by companies and other organizations wanting to use automated methods to enhance performance and provide better services to their clients. However, employing such AI solutions often comes with a tradeoff: these models lack accountability for their predictions, and their decision-making processes are opaque. This lack of transparency is problematic in high-stakes domains where experts need to understand the reasoning behind specific predictions to establish trust in AI systems.¹ For example, in healthcare, diagnostic errors contribute significantly to preventable harm, with misdiagnoses accounting for

0272-1716 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.

Digital Object Identifier 10.1109/MCG.2025.3533806

Date of current version 16 June 2025.

€ 446 billion in wasted costs in the EU (as reported by the Symptoma EU project) and \$750 billion in the U.S. Annually, around 795,000 Americans die or suffer permanent disability due to misdiagnoses of severe illnesses, such as cancer, one of the “big three” most frequently misdiagnosed conditions.² While AI holds the potential to revolutionize medical practice, its successful integration into application domains, such as digital pathology, remains limited.³

Artificial Intelligence Research

Initial efforts of AI researchers primarily targeted methods for collecting and labeling large datasets to serve as inputs for progressively more sophisticated AI models. Within the AI community, most studies were centered around achieving better predictive performance. After significant advancements in hyper-parameter optimization and algorithmic development, there appears to be a renewed focus on data-centric AI.⁴ Intrinsic algorithmic interpretability and posthoc explainability have also become indispensable components of modern AI pipelines and are now more critical than ever as sanity checks and assessment tools of an AI model’s overall knowledge.

While the achievements from the AI domain are remarkable, particularly with fully automated computational methods tailored for specific tasks, much of the research has fallen short of addressing human needs in this area. Moreover, disciplines, such as psychology and cognitive science, have long explored concepts, such as trust, interpretability, and explainability. The information visualization (InfoVis) and visual analytics (VA) communities also intersect here, offering techniques that leverage human perception and cognitive abilities.

VA Research

InfoVis and VA are subfields of the broader area of human-computer interaction (HCI) research. The former focuses on leveraging human cognitive abilities through the use of interactive, computer-supported data visualizations,⁵ while the latter centers on using interactive visual interfaces to facilitate analytical reasoning processes.⁶ In our context, a typical VA approach integrates methods from both the InfoVis/VA and AI domains, aiming to combine computational analyses with interactive visualizations. By doing so, the VA dashboard collaborates with the human analyst to interpret source data, comprehend models and predictions, and acquire knowledge, thereby completing the sensemaking loop. Consequently, a human-centered strategy for explainable AI (XAI) supports

analysts in understanding data when fully automated methods either fall short or are unsuitable.

The primary issue in this research domain is that most methodologies either emphasize understanding AI models while neglecting the initial phases of the AI pipeline, such as data processing and feature engineering, or focus exclusively on generating explanations for complex models, such as neural networks (NNs). However, deep learning may not be particularly effective for tabular data comprising meaningful features,⁷ especially when compared to ensemble learning methods that are often more powerful. Therefore, the main goals (Gs) of this work are to: (G1) define the concept of trust in the context of AI, (G2) examine and classify existing research works, (G3) propose integrated workflows with interactive visualization (VIS) for XAI across all stages, (G4) create model-agnostic VA dashboards aligned with the ensemble learning paradigm, and (G5) implement a model-specific dimensionality reduction solution that enhances the robustness of the results.

OBJECTIVES

To fulfill these goals, the author defines the following four research objectives (Os) that are also visible in Figure 1.

O1 involves conducting a meta-analysis of the current state of interactive visualization use for XAI (G1). This is achieved by manually collecting survey papers and analyzing their perspectives on individual peer-reviewed papers, summarizing key insights, and creating a guide to help researchers navigate this vast domain. This step establishes fundamental knowledge of the field and highlights research opportunities. It confirms the growing trend of using visualization to interpret and explain AI models in recent years, showing its potential to enhance trust in such models.

O2 focuses on a comprehensive literature review of how interactive visualization contributes to building trust in AI models (G2). The author establishes a design space by providing background on trustworthiness, reviewing relevant literature, categorizing visualization techniques extensively, and identifying future research directions. Our work offers a categorization of trust across various stages of interactive AI (cf., Table 1).

O3 targets the development of unified workflows addressing key research questions at different stages of the AI pipeline (G3). For example, during the initial stage, the author enables users to identify and sample training instances—both easy and challenging to classify—from all classes rather than uniformly applying undersampling or oversampling. In the next stages,

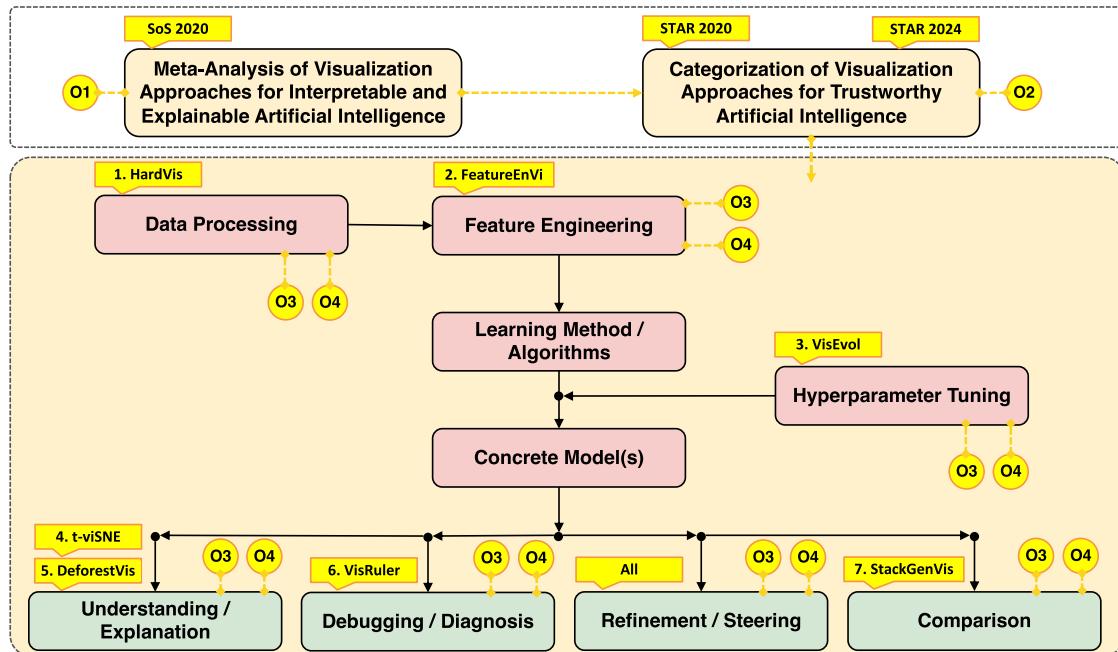


FIGURE 1. Overview of the main results of our research. The first block at the top presents three consensus reviews, encoding the generated knowledge in light yellow. This knowledge is used to identify research gaps and contribute to the field with our own works (1–7), as shown in the second block. The red boxes present a typical AI pipeline with four trust components in green. Lastly, the vivid yellow nodes refer to the research objectives (O1–O4) associated with the related results.

the author introduces methods, such as data space slicing for feature exploration and model generation, through evolutionary hyperparameter optimization processes. Users are also supported in evaluating global accuracy and local quality with projection-based visualizations, facilitating better understanding and debugging of ensemble models. Lastly, the author aids users in comparing models to streamline the creation of model stacks by eliminating underperforming or overly complex models.

O4 focuses on translating analytical tasks into specific design goals for building concrete VA dashboards (G4 and G5). The author contributes by designing and implementing multiple interactive VA techniques. These proof-of-concept dashboards are validated using both quantitative and qualitative methods by AI and domain experts. By practically implementing theoretical ideas, our dashboards aim to cover the defined design space.

To explore the existing work in the design space of VIS for XAI, the author begins by examining a set of 18 survey papers. The new insights obtained during this process help us pinpoint gaps in the literature that may appeal to early-stage and senior visualization researchers as well as AI experts.

Online Training Processes

A core future challenge is related to real-time online training processes (11 out of 18). This challenge arises because current methods mainly focus on analyzing the outcomes of AI techniques, neglecting the potential for an interactive training process that could enhance trust in the results. For example, consider an expert training an NN, a process that may take hours to complete. While the expert could use an output visualization to interpret the final results, a VA dashboard that supports the online training process could provide preliminary insights and thereby enable the expert to guide the training more effectively.

STATE OF THE ART

To lay the foundation for this research, the author draws on findings from a survey of surveys⁸ and two literature reviews (STARs^{9,10}) that he conducted.

Enhancing Trust in Artificial Intelligence

Another key challenge is how to build trust in AI algorithms and models (10 out of 18).¹¹ Achieving this

TABLE 1. AI Visualization techniques categorization in the 2024 STAR¹⁰ (including comparison with the 2020 STAR data).

Domain	542	
Biology	49	-5% ↓
Business	87	+6% ↑
Computer Vision	180	+3% ↑
Computers	9	-1% ↓
Health	90	+2% ↑
Humanities	81	-6% ↓
Nutrition	16	-1% ↓
Simulation	19	0% -
Social / Socioeconomic	58	0% -
Other	230	-5% ↓
Target Variable	542	
Binary (categorical)	85	-4% ↓
Multi-class (categorical)	297	-9% ↓↓
Multi-label (categorical)	21	-1% ↓
Continuous (regression problems)	60	0% -
Other	147	+8% ↑
AI Methods	542	
Convolutional Neural Network (CNN)	72	0% -
Deep Convolutional Network (DCN)	12	-2% ↓
Deep Feed Forward (DFF)	11	-3% ↓
Deep Neural Network (DNN)	55	0% -
Deep Q-Network (DQN)	11	-3% ↓
Generative Adversarial Network (GAN)	13	-3% ↓
Long Short-Term Memory (LSTM)	26	-2% ↓
Recurrent Neural Network (RNN)	31	-3% ↓
Variational Auto-Encoder (VAE)	22	-3% ↓
Other (DL methods)	92	+6% ↑
Linear (DR)	81	-14% ↓↓
Non-linear (DR)	94	-9% ↓↓
Bagging (ensemble learning)	74	0% -
Boosting (ensemble learning)	30	0% -
Stacking (ensemble learning)	10	-1% ↓
Other (generic)	233	-6% ↓
AI Types	500	
Classification (supervised)	288	-3% ↓
Regression (supervised)	55	0% -
Other (supervised)	32	+2% ↑
Association (unsupervised)	14	0% -
Clustering (unsupervised)	69	-8% ↓
Dimensionality Reduction (unsupervised)	104	-14% ↓↓
Classification (semi-supervised)	40	0% -
Clustering (semi-supervised)	16	0% -
Classification (reinforcement)	4	0% -
Control (reinforcement)	11	0% -
AI Processing Phase	542	
Pre-processing / Input	120	+4% ↑
In-processing / Model	156	+6% ↑
Post-processing / Output	363	-14% ↓↓
Treatment Method	542	
Model-agnostic / Black Box	395	+1% ↑
Model-specific / White Box	165	-5% ↓
Dimensionality	542	
2D	538	+1% ↑
3D	9	-1% ↓
Visual Aspects	542	
Computed	497	-6% ↓
Mapped	273	-5% ↓
Visual Granularity	542	
Aggregated Information	468	-6% ↓
Instance-based / Individual	377	-3% ↓
Visual Representation	542	
Bar Charts	243	+4% ↑
Box Plots	39	+1% ↑
Matrix	153	+3% ↑
Glyphs / Icons / Thumbnails	131	-8% ↓
Grid-based Approaches	59	+1% ↑
Heatmaps	149	+4% ↑
Histograms	174	+4% ↑
Icicle Plots	9	-1% ↓
Line Charts	154	0% -
Node-link Diagrams	128	0% -
Parallel Coordinates Plots (PCPs)	73	-3% ↓
Pixel-based Approaches	20	0% -
Radial Layouts	90	+6% ↑
Scatterplot Matrices (SPLOMs)	26	-4% ↓
Scatterplot / Projections	294	-4% ↓
Similarity Layouts	190	+21% ↑↑↑
Tables / Lists	201	-6% ↓
Treemaps	13	-1% ↓
Other	207	+8% ↑
Interaction Technique	525	
Select	491	+9% ↑↑
Explore / Browse	445	-3% ↓
Reconfigure	220	+4% ↑
Encode	294	-2% ↓
Filter / Query	314	+1% ↑
Abstract / Elaborate	354	-24% ↓↓↓
Connect	395	+9% ↑↑
Guide / Shepherd	169	+7% ↑
Verbalize	31	+1% ↑
Visual Variable	542	
Color	540	+2% ↑
Opacity	294	+12% ↑↑
Position / Orientation	83	-14% ↓↓
Shape	89	-3% ↓
Size	164	-4% ↓
Texture	30	-3% ↓
Evaluation	542	
Standard	132	+5% ↑
Comparative	27	-1% ↓
Before / During Development	140	+8% ↑
Alter Development	175	+8% ↑
Not Evaluated	196	-15% ↓↓
Trust Levels (TL) 1–5	542	
Source Reliability	43	+2% ↑
Transparent Collection Process	21	+1% ↑
Uncertainty Awareness	84	0% -
Equality / Data Bias	57	+3% ↑
Comparison (of Structures)	271	+7% ↑
Guidance / Recommendations	142	+3% ↑
Outlier Detection	174	-2% ↓
Familiarity	24	+2% ↑
Understanding / Explanation	314	+10% ↑↑
Debugging / Diagnosis	129	-3% ↓
Refinement / Steering	164	-5% ↓
Comparison	114	-10% ↓↓
Knowledgeability	47	+4% ↑
Fairness	29	+2% ↑
Experience	31	+2% ↑
In Situ Comparison	135	-2% ↓
Performance	338	+8% ↑
What-if Hypotheses	87	-4% ↓
Model Bias	58	-1% ↓
Model Variance	34	-2% ↓
Agreement of Colleagues	21	-1% ↓
Visualization Evaluation	301	+12% ↑↑
Metrics Validation / Results	413	+11% ↑↑
User Bias	24	0% -
Target Group	542	
Beginners	104	-2% ↓
Practitioners / Domain Experts	424	-3% ↓
Developers	113	+3% ↑
AI Experts	204	+1% ↑

Color Legend: 0 papers 271 papers 542 papers

Symbol Legend:
 +(-1 - 8% ↑↓
 +/-[9 - 16% ↑↑↓↓
 +/-[17 - 24% ↑↑↑↓↓↓

Note: Each row shows the total count of techniques per category (with heatmap-style icons), and the % difference compared to the 2020 STAR.

requires integrating interactive visualizations with traditional quantitative metrics into VA dashboards, thereby creating an “ecosystem” that allows expert users to oversee and manage AI models at various levels. For example, a VA dashboard might display how standard validation metrics (e.g., accuracy, precision, or recall) of classifiers evolve during each step of an AI algorithm’s execution, enabling experts to detect deviations in the model’s behavior. Thus, experts can intervene by inserting their knowledge into the process. In essence, VA dashboards should support timely interventions when needed.

To first define what trust in AI is, the author proposes a more detailed, multilevel model of trust in AI that consists of five trust levels (TLs): the raw data (TL1: source reliability and transparent

collection process), the processed data (TL2: uncertainty awareness, equality/data bias, comparison of structures, guidance/recommendation, and outlier detection), the algorithm/learning method (TL3: familiarity, understanding/ explanation, debugging/ diagnosis, refinement/steering, comparison, knowledgeability, and fairness), the concrete model(s) for a particular task (TL4: experience, in situ comparison, performance, what-if hypotheses, model bias, and model variance), and the evaluation and the subjective users’ expectations (TL5: agreement of colleagues, visualization evaluation, metrics validation/results, and user bias).

The author then selects 542 peer-reviewed papers that introduce a large variety of visualization techniques for increasing trust in AI models and their results.

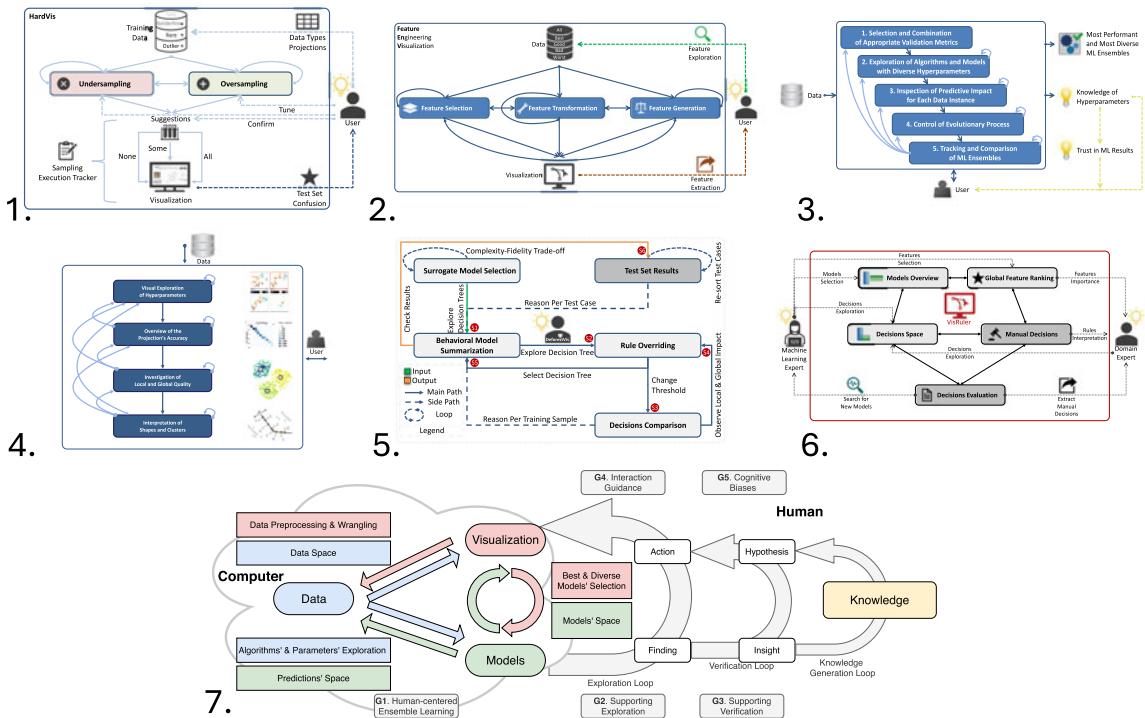


FIGURE 2. Author designed coherent workflows for our seven technical contributions, covering each stage of the AI pipeline.

Based on them, the author proposes a fine-grained categorization comprising eight high-level design space aspects partitioned into 18 category groups that, on their part, contain 119 categories in total (see Table 1).

Our categorization served as a foundation for the design space of VA approaches for AI, which the author introduces in the subsequent section.

VA WORKFLOWS AND DASHBOARDS

Data Processing

Despite significant advancements in AI, handling imbalanced data remains a challenge in many real-world scenarios. Among various approaches, sampling algorithms are considered effective solutions. However, many studies emphasize the major role of instance hardness, that is, the challenge of managing noisy or difficult-to-classify instances that lead to misclassification and degraded performance.

The proposed workflow begins with splitting the training data into four types based on the user's visual inspection of nine alternative projections [see Figure 2(1)]. Data undergo iterative undersampling or oversampling, with suggestions provided continuously

and user confirmation sought after exploratory visual analysis.

HardVis¹² is a VA dashboard designed to address instance hardness by employing highly configurable undersampling and oversampling methods [cf., Figure 3(1)]. The dashboard provides multiple coordinated views, enabling users to determine the optimal distribution of data types iteratively, selectively remove safe-to-discard samples, and interactively oversample others as needed. It also supports the analysis of algorithmic recommendations by using diverse visual cues to validate the safety of proposed removals or additions. This VA approach makes the entire process transparent by being effective in tackling instance hardness and class imbalance problems.

A limitation identified by AI experts is the scalability due to inadequate support for feature engineering.

Feature Engineering

Feature engineering can significantly benefit AI by enhancing predictive performance, minimizing computational costs, reducing unnecessary noise, and improving the transparency of decisions made during training. However, although various VA dashboards exist to monitor and control the different stages of the AI life cycle (e.g., those concerning data and

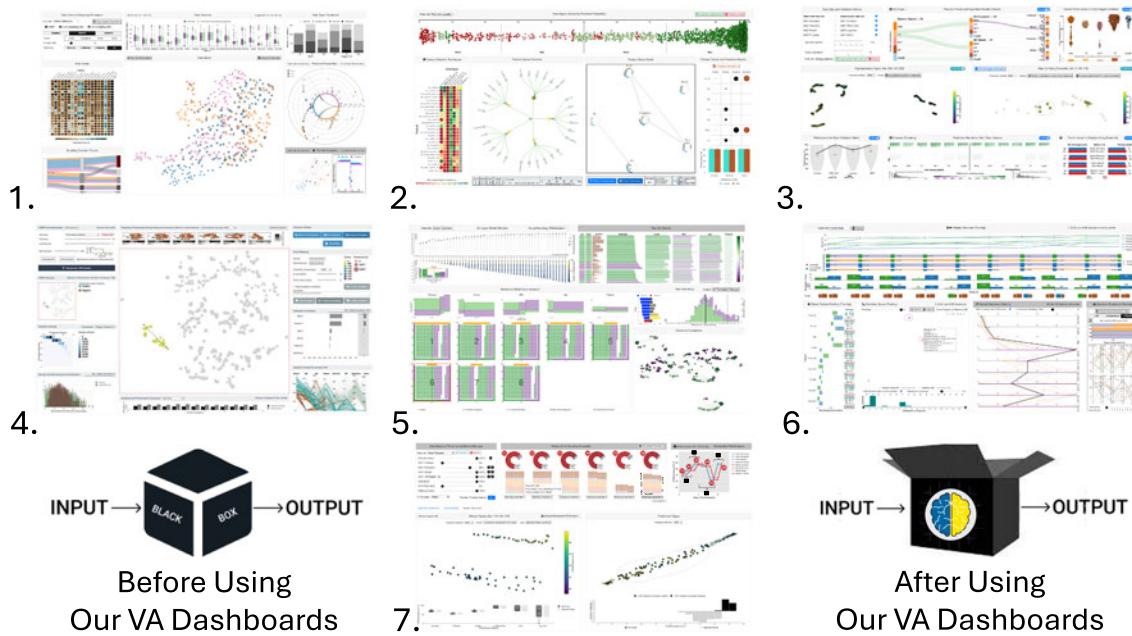


FIGURE 3. Author developed seven VA dashboards to address challenges at every stage of the AI pipeline effectively.

algorithms), support for feature engineering is yet insufficient.

The proposed workflow starts by dividing the dataset based on the prediction probabilities of individual samples [see Figure 2(2)]. The data then undergo three distinct feature engineering processes (selection, transformation, and generation) that are performed iteratively, with the user in control of the process.

FeatureEnVi¹³ is a VA dashboard designed to support users in engineering features through stepwise feature selection and semiautomatic feature extraction methods [cf., Figure 3(2)]. The dashboard employs multiple coordinated views to assist users in selecting, transforming, and creating new features as part of a highly iterative workflow. By examining the impact of features using various statistical metrics and automated feature selection techniques, users can enhance predictive performance, minimize computational resource requirements, and shorten training time. This VA approach increases adaptability and transparency by improving the feature engineering process.

A limitation is the lack of experimentation with alternative hyperparameter tuning techniques.

Hyperparameter Tuning

During the training phase of AI models, multiple hyperparameters typically need to be configured. This task is computationally demanding and involves a

comprehensive search to identify the optimal hyperparameter set for a specific problem. The difficulty increases due to the fact that each hyperparameter in an AI algorithm is often interdependent with others, and modifying one may lead to unpredictable effects on the rest. Evolutionary hyperparameter optimization offers a promising approach to tackle these challenges. This strategy retains high-performing models while enhancing others through crossover and mutation techniques inspired by genetic algorithms.

The proposed workflow enables users to build effective and diverse AI ensembles while gaining insights into the hyperparameter selection process facilitated by evolutionary optimization [see Figure 2(3)]. Users can control the process by specifying the number of models involved in crossover and mutation operations for various algorithms.

VisEvol¹⁴ is a VA dashboard designed to facilitate hyperparameter search using evolutionary optimization [cf., Figure 3(3)]. By using multiple coordinated views, the dashboard enables users to create new hyperparameter sets and preserve robust ones within a majority-voting ensemble. Examining the effects of adding or removing algorithms and models in a majority-voting ensemble from various perspectives, along with monitoring the crossover and mutation processes, provides users with clarity on selecting hyperparameters for a single model or complex ensembles that require a balance of performance and diversity.

A concern for this dashboard is its reliance on projection-based views [e.g., produced by t-distributed stochastic neighbor embedding (t-SNE)].

Model Understanding

t-SNE is a widely used method for visualizing multidimensional data, with numerous successful applications across various fields. However, interpreting t-SNE projections can be challenging and, at times, misleading, which can undermine the reliability of the results. Gaining insight into the workings of t-SNE and the underlying causes of specific patterns in its output can be particularly difficult, especially for those unfamiliar with dimensionality reduction techniques.

The proposed workflow supports the selection of hyperparameters through visual exploration and quality metrics, offers a rapid assessment of projection accuracy, enables deeper investigation into quality by distinguishing the trustworthiness of various projection regions, and facilitates the interpretation of visible patterns in the projection relative to the original dataset's dimensions in an iterative process [see Figure 2(4)].

t-viSNE¹⁵ is a VA dashboard designed for exploring t-SNE projections visually [cf., Figure 3(4)]. By partially unveiling the inner workings of the t-SNE algorithm, the author empowers users to evaluate the quality of the projections and gain insights into the algorithm's decision-making process when forming clusters. The author also reveals typically overlooked information from within the algorithm, such as point densities, and highlights regions that are suboptimally optimized by t-SNE.

With the foundation for assessing visual embeddings established, the author now addresses understanding, debugging, and comparing supervised AI algorithms.

A simple, model-agnostic approach for XAI models involves training surrogate models, such as decision trees and rule sets, which approximate the original models while remaining simpler and more explainable. However, rule sets can become excessively lengthy with numerous if-else statements, and decision trees may quickly grow deep when attempting to replicate complex AI models accurately. Consequently, both methods can struggle to achieve their primary objective of delivering model interpretability to users.

The proposed workflow allows users to select a surrogate model that aligns with their preferred trade-off between fidelity and simplicity, examine decisions derived from the surrogate model (acting as a simplified version of the complex AI model), and adjust

individual rules based on visual feedback and personal experience [see Figure 2(5)]. To complete the cycle, users can iteratively reason through specific test cases while reviewing previously generated explanations for the training data.

DeforestVis¹⁶ is a VA dashboard designed to employ decision stumps (single-level decision trees) derived from the AdaBoost algorithm, which are inherently interpretable, to create surrogate models that approximate the behavior of a target model [see Figure 3(5)]. The dashboard's linked views enable users to iteratively balance complexity and fidelity while reducing the precision (measured in decimal places) used in decision thresholds without compromising accuracy. Moreover, DeforestVis allows users to explore decision stumps in various ways, such as evaluating their purity and impact, and it summarizes the surrogate model's behavior by aggregating the predictive outcomes and influence of all decision stumps associated with each data feature. Users can modify rules and compare them at both local and global levels.

The following section explores the possible advantages of incorporating human involvement in debugging ensembles.

Model Debugging

Bagging and boosting are widely used ensemble techniques in AI that generate multiple individual decision trees. Because of their ensemble nature, these methods often achieve better predictive performance compared to single decision trees or other AI models. However, each decision tree creates numerous decision paths, increasing the model's complexity and limiting its application in fields requiring reliable and explainable outcomes. Therefore, as the number of decisions increases, the interpretability of algorithms, such as random forest and AdaBoost, diminishes.

The proposed workflow enables AI experts to identify high-performing and diverse models, prioritize important features, and refine models by adjusting hyperparameters [see Figure 2(6)]. Meanwhile, domain experts can investigate reliable decisions, benchmark them against global standards, identify local decisions for a specific test instance, and extract these insights.

VisRuler¹⁷ is a VA dashboard that enables users to examine a range of rules derived from bagged and boosted decision trees in order to reach a consensus on the final decision for each case [cf., Figure 3(6)]. The dashboard's multiple coordinated views support the selection of various high-performance models, analysis of feature contributions, management of

multiple decisions, exploration of global decisions, and facilitation of case-based reasoning.

Having set the basis for combining bagging and boosting methods in various voting schemes, the next subsection explores the potential benefits of involving humans in the construction of stacking ensembles.

Model Comparison

Stacking is an ensemble method that integrates various base models, organized in at least one layer, and uses a metamodel to aggregate their predictions. While it can significantly improve the predictive accuracy of AI, creating a model stack from the ground up often involves a time-consuming trial-and-error process. This difficulty arises due to the vast range of possible solutions, including varying sets of data instances and features for training, a wide selection of algorithms, and different parameter settings (i.e., models) that yield different performances based on various metrics.

The proposed workflow is built on a knowledge generation model for ensemble learning using VA [see Figure 2(7)]. On the left-hand side, it shows how a VA dashboard can facilitate data and model exploration through visualization. On the right-hand side, several design objectives support the user in exploring, verifying, and generating knowledge for ensemble learning.

StackGenVis¹⁸ is a VA dashboard designed for aligning data, algorithms, and models in stacking ensemble [cf., Figure 3(7)]. Through multiple coordinated views, users can create an effective stacking ensemble from scratch. By exploring the algorithms, data, and models from various perspectives and monitoring the training process, users can confidently determine the next steps in developing complex stacks of models, requiring a combination of both the highest performing and the most diverse individual AI models.

LESSONS LEARNED

Ensemble Learning for Tabular Data

The VA community has devoted significant attention to XAI in recent years, yet much of the focus remains on NNs for image data. Out of 542 visualization papers surveyed on enhancing trust in AI, 345 primarily addressed NN models, while only 114 concentrated on ensembles. Although image analysis is undeniably an important research domain—especially in fields, such as medicine, where NNs excel with high-dimensional data—other application areas should not be overlooked. Tabular data still

represent the majority of datasets employed in AI workflows, where ensemble learning models often surpass NNs in performance.⁷

Ideas Beyond Developing VA Dashboards

The visualization approaches discussed in this work can also be employed as standalone tools to support various stakeholders (refer to Table 1) in addressing complex challenges related to different phases of the AI pipeline (cf., Figure 1). However, the individual VA workflows proposed are highly adaptable and could be integrated to create custom solutions that address increasingly difficult supervised and unsupervised learning tasks. This integration enables the creation of a versatile design space with personalized approaches for managing AI models designed for tabular data, explored from multiple perspectives, ranging from local explanations of single instances to global summaries of model behavior across all or specific subsets of data.

Common Sources of Bias

In this work, the assumption of ground truth in supervised classification tasks was maintained across all use cases and usage scenarios. However, evaluating the quality and fairness of the data, as well as identifying potential biases or errors from the annotators responsible for labeling, could prove critical. Also, visualizations have the potential to obscure additional biases. For example, in VisEvol, the author developed a bar chart that represents the uncertainty introduced by the force-based layout algorithm used for the beeswarm plot, as shown in Figure 3(3). Additionally, incorporating domain knowledge into visual explanation dashboards can introduce bias, as data scientists might fine-tune parameters to align explanations with their own assumptions.¹⁹ This issue is not exclusive to our approach but is inherent in the application of domain-specific expertise. Hence, it is essential to exercise caution and explicitly signal instances where explanations diverge from the actual model by using provenance and uncertainty visualizations.

Online Survey Browser

Our publicly accessible survey browser has received considerable positive feedback from the VA community for facilitating the discovery of new papers. Our web-based survey browser, as shown in Figure 4, has been visited by nearly 5000 unique users from 122 countries since its launch in Spring 2020 (as reported by Google Analytics). The author considers



FIGURE 4. Our interactive survey browser,^{9,10} for exploring the 542 identified VA techniques (available at <https://trustmlvis.lnu.se/>).

TrustMLVis Browser a valuable resource across multiple disciplines and remains committed to maintaining and further developing it in the future. For example, the initial dataset of 200 entries has recently grown to 542, as given in Table 1.

Teaching Applications

Several VA dashboards developed by the author have demonstrated their potential in teaching fundamental AI concepts to computer science students who are comfortable with visualizations. For two consecutive years, students in the advanced InfoVis course at Linnaeus University (Sweden) were tasked with exploring Google’s embedding projector and providing feedback on its key strengths, weaknesses, and potential improvements. Next, t-viSNE was introduced, enabling students to compare the two dashboards in addressing the shortcomings of t-SNE highlighted by Wattenberg et al.²⁰ During the t-viSNE paper’s presentation at the IEEE VIS 2020 conference, attendees noted the clear t-viSNE’s applicability in educational settings. Also, an AI expert evaluating VisEvol suggested that features, such as the Sankey diagram, could assist students in grasping the concept of evolutionary

optimization. Another AI expert acknowledged the educational value of HardVis.

IMPACTS AND FUTURE WORK

Under the General Data Protection Regulation and the AI Act, data controllers must provide clear explanations of the logic behind automated decisions. In addition, individuals have the right to avoid being subjected to decisions made solely through automated processes. Enhancing trust in human decision-making influenced by AI algorithms is a crucial step in addressing concerns about such outcomes, particularly when aiming to minimize biases. For example, in healthcare, medical professionals and data analysts who use AI models to diagnose diseases based on sensitive medical data bear both ethical and legal responsibilities for the resulting decisions. In modern practice, failing to explain how a specific medical prognosis was reached is increasingly viewed as unacceptable. As the volume of digital data continues to expand and AI technology plays a growing role in societal functions, computational and VA methods can offer powerful tools for addressing core challenges. This work demonstrates how VA dashboards can foster greater explainability

and trust in high-stakes decision making like healthcare, where people's lives are at stake.

The methods described in this work are applicable to both academic research and business intelligence. For example, in the financial sector, declining a loan application should involve greater transparency, clearly outlining the reasons for rejection. Building trust between humans and algorithms depends on understanding how the algorithm operates and being able to explain the reasoning behind its predictions.

Finally, the research opportunities (R1–R8) identified in our 2024 STAR¹⁰ offer valuable insights for visualization researchers, practitioners across diverse disciplines, and AI experts. The author summarizes them below.

R1: Improving Popular XAI Methods

Many widely recognized XAI methods using visualization to communicate results were developed by experts from disciplines outside of visualization research. The growing appeal of these model-agnostic techniques, such as LIME and SHAP, stems from their perspective of a model as a mathematical function. These functions can be broken down into simpler components, whose behavior can then be evaluated using interpretation methods. Visualization researchers can enhance these methods by designing VA dashboards that improve their interactivity and visual representations or by directly refining the algorithmic components of the techniques with their specialized knowledge.

R2: New NN Approaches and Self-Supervision

The AI community consistently develops new NN architectures or adapts existing ones to different contexts, whereas visualization researchers have been slower in adopting and exploring these advancements. For example, transformer NNs have shown promising results in computer vision tasks compared to convolutional NNs, yet the VA community has rarely engaged in visually supporting their exploration. Also, self-supervision represents a rapidly growing field that warrants greater attention from the VA community, alongside prompt engineering and model checkers to address "hallucinations" in large language models.

R3: Multiverse and Confirmatory VA

Recently, most VA solutions have predominantly emphasized exploratory visual analysis. However, there is a growing trend toward employing visual

representations for confirmatory analysis, hypothesis testing, and causal reasoning (e.g., counterfactual reasoning, causal learning, and causal inference), which could benefit from enhanced visualization. Another relevant question is how VA dashboards can assist users in conducting a comprehensive multiverse analysis to identify optimal model strategies, enabling them to understand complex models better and make informed decisions about a specific course of action.

R4: Input/Output Uncertainty Quantification

Most VA dashboards are designed to assist users in understanding models, but two significant challenges remain: quantifying input and output uncertainties, and evaluating robustness and sensitivity to changes, as users aim to deploy dynamic, well-calibrated models in practical real-world scenarios. While uncertainty quantification and sensitivity analysis are widely used in other disciplines, visualization research still lacks comprehensive approaches for examining model inputs and outputs, such as the conformal prediction method, which offers statistical guarantees.

R5: Rigorous Evaluation and Benchmarking

Collaboration with domain experts from various disciplines often encounters significant communication challenges. These stem from differences in terminology, expectations, and expertise, requiring considerable effort to align goals and solutions. To reduce reliance on expensive user evaluations, it is crucial to explore AI-assisted, simulated evaluation methods. Enhancing the rigor of state-of-the-art benchmarking datasets is equally important, as they may contain notable errors. Additionally, the development of rational agent benchmarks that assess the necessity and advantages of visualization can support the evaluation of XAI methods. Such benchmarks can potentially reduce the dependence on extensive human involvement, thereby lowering associated costs.

R6: Model Deployment and Visual Channels

Scaling VA dashboards to handle a large number of instances, features, and algorithms is particularly challenging, especially in multiclass classification contexts. While progressive VA combined with incremental learning offers a promising approach, developing simpler yet more robust systems is critical for real-world applications, where issues, such as covariate and label shifts, frequently occur. The simultaneous

use of visual channels, such as color and opacity, for encoding tasks adds to the complexity. Many users struggle to navigate these sophisticated dashboards, particularly during their initial interactions. Visualization literacy can help address this issue by educating users. In addition to existing tutorials, storytelling can effectively be used to demonstrate a dashboard's functionalities, while sonification and verbalization can enhance user comprehension and simplify the use of VA dashboards.

R7: Advancing the Impact and Reproducibility

Many application-specific VA dashboards fail to transition into real-world deployment, leaving them vulnerable to concept and distribution drifts. This underscores the necessity of investigating the applicability of models in out-of-domain/distribution scenarios, stressing the importance of adapting specialized models into more flexible AI solutions. Broadening the accessibility and usability of VA dashboards can be achieved through various approaches, such as incorporating visualizations within computational notebooks and providing hybrid solutions that combine programmatic APIs with interactive UIs to gather user feedback over extended periods. Adopting open-source practices plays a major role in encouraging community participation and addressing key challenges related to the replication and reproducibility of AI models used in VA dashboards.

R8: Underexplored Areas

By analyzing the categorization of the 542 visualization techniques, it becomes evident that the underrepresented categories have the potential to inspire new research directions for non-TL areas, indirectly impacting trust in AI. For example, VA dashboards for boosting and stacking ensemble learning methods are notably scarce. Similarly, multilabel and regression problems are less explored compared to classification. In the realm of unsupervised learning, association and pattern mining remain underrepresented in VA. Reinforcement learning also receives minimal attention. Lastly, more emphasis is needed on addressing the needs of beginners and users with varying levels of experience who aim to analyze their data broadly.

ACKNOWLEDGMENTS

The author would like to thank his main Ph.D. advisor, Prof. Dr. Andreas Kerren (Linköping University, Sweden), for the tremendous support and guidance for

this dissertation work. Also, many thanks to his Ph.D. co-advisors, Dr. Rafael M. Martins (Linnaeus University, Sweden) and Dr. Ilir Jusufi (Blekinge Institute of Technology, Sweden), for the feedback received throughout the entire Ph.D. study. The author is thankful to his postdoc mentor, Prof. Dr. Jessica Hullman (Northwestern University, USA), for her efforts in teaching him her perspective on conducting research. Finally, many thanks to Prof. Dr. Alexandru C. Telea for his mentorship in further advancing his academic career.

REFERENCES

1. J. Zhou and F. Chen, "2D transparency space—Bring domain users and machine learning experts together," *Hum. Mach. Learn.*, pp. 3–19, 2018.
2. D. E. Newman-Toker et al., "Burden of serious harms from diagnostic error in the USA," *BMJ Qual. Saf.*, vol. 33, no. 2, pp. 109–120, 2024.
3. A. Sajithkumar et al., "AI in pathology: Current applications, limitations, and future directions," *Ir. J. Med. Sci.*, vol. 193, pp. 1117–1121, 2024.
4. O. H. Hamid, "From model-centric to data-centric AI: A paradigm shift or rather a complementary approach?," *Proc. 8th Int. Conf. Inf. Techn. Trends*, 2022, pp. 196–199.
5. S. K. Card, J. D. Mackinlay, and B. Shneiderman, *Readings in Information Visualization: Using Vision to Think*, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999.
6. J. J. Thomas and K. A. Cook, *Illuminating the Path: The Research and Development Agenda for Visual Analytics*, Los Alamitos, CA, USA: IEEE, 2005.
7. R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Inf. Fusion*, vol. 81, pp. 84–90, 2022.
8. A. Chatzimpampas, R. M. Martins, I. Jusufi, and A. Kerren, "A survey of surveys on the use of visualization for interpreting machine learning models," *Inf. Vis.*, vol. 19, no. 3, pp. 207–233, 2020.
9. A. Chatzimpampas et al., "The state of the art in enhancing trust in machine learning models with the use of visualizations," *Comput. Graphics Forum*, vol. 39, no. 3, pp. 713–756, 2020.
10. A. Chatzimpampas, K. Kucher, and A. Kerren, "Visualization for trust in machine learning revisited: The state of the field in 2023," *IEEE Comput. Graphics Appl.*, vol. 44, no. 3, pp. 99–113, May/Jun. 2024.
11. A. Endert et al., "The state of the art in integrating machine learning into visual analytics," *Comput. Graphics Forum*, vol. 36, no. 8, pp. 458–486, 2017.

12. A. Chatzimpampas, F. V. Paulovich, and A. Kerren, "HardVis: Visual analytics to handle instance hardness using undersampling and oversampling techniques," *Comput. Graphics Forum*, vol. 42, no. 1, pp. 135–154, 2023.
13. A. Chatzimpampas, R. M. Martins, K. Kucher, and A. Kerren, "FeatureEnVi: Visual analytics for feature engineering using stepwise selection and semi-automatic extraction approaches," *IEEE Trans. Vis. Comput. Graphics*, vol. 28, no. 4, pp. 1773–1791, Apr. 2022.
14. A. Chatzimpampas, R. M. Martins, K. Kucher, and A. Kerren, "VisEvol: Visual analytics to support hyperparameter search through evolutionary optimization," *Comput. Graphics Forum*, vol. 40, no. 3, pp. 201–214, 2021.
15. A. Chatzimpampas, R. M. Martins, and A. Kerren, "t-viSNE: Interactive assessment and interpretation of t-SNE projections," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 8, pp. 2696–2714, Aug. 2020.
16. A. Chatzimpampas, R. M. Martins, A. C. Telea, and A. Kerren, "DeforestVis: Behaviour analysis of machine learning models with surrogate decision stumps," *Comput. Graphics Forum*, vol. 43, no. 6, 2024, Art. no. e15004.
17. A. Chatzimpampas, R. M. Martins, and A. Kerren, "VisRuler: Visual analytics for extracting decision rules from bagged and boosted decision trees," *Inf. Vis.*, vol. 22, no. 2, pp. 115–139, 2023.
18. A. Chatzimpampas, R. M. Martins, K. Kucher, and A. Kerren, "StackGenVis: Alignment of data, algorithms, and models for stacking ensemble learning using performance metrics," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 2, pp. 1547–1557, Feb. 2021.
19. E. Dimara, S. Franconeri, C. Plaisant, A. Bezerianos, and P. Dragicevic, "A task-based taxonomy of cognitive biases for information visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 2, pp. 1413–1432, Feb. 2020.
20. M. Wattenberg, F. Viégas, and I. Johnson, "How to use t-SNE effectively," *Distill*, vol. 1, no. 10, 2016, Art. no. e2.

ANGELOS CHATZIMPARMPAS is an assistant professor with the Department of Information and Computing Sciences, Utrecht University, 3584 CC, Utrecht, The Netherlands. His research interests include visual analytics, human-computer interaction, machine learning interpretability, human-centered AI, and XAI systems. Chatzimpampas received his Ph.D. degree in computer and information science from Linnaeus University, Växjö, Sweden. He is a member of IEEE. Contact him at a.chatzimpampas@uu.nl.

Contact department editor Sumanta N. Pattanaik at Sumanta.Pattanaik@ucf.edu.