Are We Closing the Loop Yet? Gaps in the Generalizability of VIS4ML Research

Hariharan Subramonyam ⃝iD, and Jessica Hullman ⃝iD

Abstract— Visualization for machine learning (VIS4ML) research aims to help experts apply their prior knowledge to develop, understand, and improve the performance of machine learning models. In conceiving VIS4ML systems, researchers characterize the nature of human knowledge to support human-in-the-loop tasks, design interactive visualizations to make ML components interpretable and elicit knowledge, and evaluate the effectiveness of human-model interchange. We survey recent VIS4ML papers to assess the generalizability of research contributions and claims in enabling human-in-the-loop ML. Our results show potential gaps between the current scope of VIS4ML research and aspirations for its use in practice. We find that while papers motivate that VIS4ML systems are applicable beyond the specific conditions studied, conclusions are often overfitted to non-representative scenarios, are based on interactions with a small set of ML experts and well-understood datasets, fail to acknowledge crucial dependencies, and hinge on decisions that lack justification. We discuss approaches to close the gap between aspirations and research claims and suggest documentation practices to report generality constraints that better acknowledge the exploratory nature of VIS4ML research.

Index Terms—VIS4ML, Visualization, Machine learning, Human-in-the-loop, Human Knowledge, Generalizability, Survey.

---

1 INTRODUCTION

Visualization for machine learning (VIS4ML) research aims to support human involvement in the machine learning (ML) process by making ML models interpretable to humans [62]. The underlying assumption is that by providing experts such as ML engineers and domain specialists with appropriate *visual representations* of the modeling pipeline, they will be able to combine their relevant *prior knowledge* with the machine representation toward positive ends - i.e., human-in-the-loop (HITL) machine learning. For instance, DataDebugger [88] supports human correction of mislabeled training data, INFUSE [37] enables domain expert involvement in feature engineering, and ConceptExplainer [27] allows analysts to extract concept-based explanations for explainable AI tasks. In an ideal scenario, the knowledge generated through bespoke VIS4ML contributions can influence real-world ML workflows, in which practitioners can use these tools to interpret and develop performant models.

In this paper, we consider evidence of the *generalizability* of VIS4ML research. Generalizability concerns the alignment between the general claims made about the effectiveness and applicability of VIS4ML contributions and the quantitative or qualitative evidence presented to validate those claims. In VIS4ML research, this depends on how *design hypotheses* – propositions about effective visualization and interaction choices for meeting HITL task needs – are operationalized in light of the researchers' generalization goals. This includes understanding how papers go from aspirations about how VIS4ML systems can enable HITL tasks, to specific intended effects of involving human knowledge in a pipeline, to particular design instantiations meant to realize these effects, to the evaluation of those design artifacts, to the conclusions that are ultimately drawn about effective VIS4ML strategies. When design hypotheses involve unstated assumptions that are overlooked in interpreting the results–for example, about the degree of knowledge people have of model components–we should not expect claims to generalize to settings where those assumptions are not in place. Further, we would expect claims entailed by the design hypotheses to be directly validated by evidence of use, including eval-

uating the validity and robustness of human-generated insights and how these ultimately affect the target learning pipeline or downstream outcomes. However, causal inferences that users generate to explain model performance may not be verifiable without further data collection, or researchers may lack visibility into the larger lifecycle of a model that they intend to affect. When research claims are based on unstated dependencies and overlooked gaps in evidence, the claimed effects of applying VIS4ML contributions are unlikely to realize in practice.

In this work, we critically examine a set of 52 VIS4ML papers to characterize the space of design hypotheses and evaluation practices and identify gaps that could hinder the adoption of VIS4ML research in practice. Our analysis surfaces patterns in how papers envision the role of human knowledge in VIS4ML, the knowledge assumptions made of system users, the algorithms and interpretability approaches they rely on, and the approaches they take to evaluate their hypotheses. We find that a majority of VIS4ML papers aim to combine visualizations and interactivity to bring about concrete improvements to an ML pipeline. Yet, more often than not, these improvements are not directly evaluated. We also identify common dependencies, for example, on the same small group of experts during development and evaluation, on well-known datasets, and on post-hoc interpretability methods that often lack faithfulness guarantees — that may threaten the ability of independent authors or practitioners to experience the same gains when they apply the contributed approaches in related contexts.

Broadly, our analysis finds the current scope of human-in-the-loop VIS4ML is somewhat limited and draws attention to potential threats to the practical adoption of VIS4ML contributions and the generalizability of research claims. To bridge the gap between bespoke VIS4ML contributions and its use in ML production workflows, we make short and longer-term recommendations for action, including transparent documentation of unstated assumptions and constraints, tightening loose derivation chains in the logical progression from aspirations for human knowledge integration to designs and their evaluation, and exploring partnerships with the broader human-centered AI research community.

2 BACKGROUND

2.1 Taxonomizing VIS4ML

Existing surveys of visual analytics for ML research [19, 24, 62, 63, 70, 71, 91] taxonomizes goals, activities, and human inputs to the modeling pipeline (separate from the indirect use of ML for improving visual analytics pipelines, e.g., [6]). This includes data quality and feature engineering before model building, understanding and diagnosing issues with parameters or training dynamics during model fitting

---

- *Hariharan Subramonyam is with Stanford University. E-mail: harihars@stanford.edu.*
- *Jessica Hullman is with Northwestern University. E-mail: jhullman@northwestern.edu.*

and selection, and reasoning about results after model building. Taxonomies also capture differences in intended audiences for VIS4ML tools, from ML experts to non-experts or domain experts [24, 63, 91], and commonly used visualization and interaction techniques [24, 63].

Most relevant to our work is Sperrle et al.'s [70] survey of human-centered evaluations of human-centered machine learning, which characterizes heterogeneity in evaluation styles and assesses data types, analysis tasks, and interactivity in VIS4ML. They differentiate the knowledge requirements of VIS4ML users, including ML versus domain expertise. However, because their scope is broader than ours (nearly half of the 71 papers they survey study explainable AI techniques in lab settings, similar to several other recent surveys [52, 71]), their results are less targeted to visualization-specific research. Additionally, the aim of our work is unique in that, we are interested in the alignment between researchers' aspirations and claims and their methods, including how knowledge is claimed to be produced through moving from research aspirations to specific design hypotheses to the validation of those hypotheses. Hence we focus on the epistemic status of VIS4ML and the generalizability of results than prior surveys.

2.2 Knowledge generation through visual data analysis

We investigate the forms of knowledge and insight authors describe to motivate and evaluate VIS4ML systems. Our work relates to knowledge generation (KG) models used in visual analytics [18, 20, 64], which aim to explain the process by which analysts generate knowledge in working with interactive visualizations of data or models. For instance, Sacha et al.'s KG model [64] adapts sensemaking concepts to describe how an analyst engages in iterative exploration and verification loops with an interactive visualization system to generate knowledge. The process is conceptualized through loops in which the analyst takes *actions*, referring to tasks that generate tangible, unique responses from the visual analytics system, to explore visualized evidence for *findings*, or visual patterns, perhaps driven by an analytical goal. The identification of *findings* leads either to further interaction with the system or to new *insights* when the analyst applies their prior knowledge to interpret the results within the domain-specific setting.

Specific to VIS4ML, Sacha et al. [62] contribute an ontology that breaks complex sequences of human interactions with a VIS4ML system into K-Driven processes, which take in human input to control the process, K-Oriented processes, which output information for humans to process, and K-Centered processes, which are designed for human interaction and cannot be easily further broken down. While they contribute a language for representing interactions, we empirically investigate how such processes are studied in the VIS4ML literature and the sorts of human knowledge and capabilities they assume. Others have studied the variety of techniques by which prior knowledge can be integrated into machine learning systems [26, 36, 77], for example, how the integration of domain knowledge in machine learning pipelines more broadly is often informal and under-described in applied ML research [36], which our results corroborate for VIS4ML.

2.3 Challenges in Evaluating Human-in-the-loop ML

Prior work has noted challenges in choosing success metrics for VIS4ML tools to identify whether a human-machine collaboration is successful [5]. When metrics or "signals" of performance that a human-in-the-loop system surfaces are locally relevant but poorly connected to the downstream application for which the model is intended, then human attempts to optimize performance for these metrics, such as by cleaning input data, may not effect or even hurt downstream outcomes [55]. Similarly, when the specific contributions of a human versus an automated component are not well defined, it is difficult to identify the proper evaluation for the research claims [5, 70].

Our aim to uncover hidden dependencies in VIS4ML aligns with calls for more reproducible, replicable, and robust ML research [23], especially ML system evaluations [40]. Similar to the replication crisis in experimental research, overlooked dependencies, sources of variance, and conventions that encourage bold claims can lead researchers to misattribute performance differences to parts of an ML pipeline [29].

3 METHODOLOGY

To characterize the nature of VIS4ML research contributions with a particular emphasis on human-in-the-loop machine learning, we conducted a qualitative survey of recent research papers proposing and evaluating VIS4ML tools. Figure 1 shows an overview of our six month-long paper collection process, analysis, and discussion of findings.

3.1 Paper Selection

We took a two-fold approach to identify papers of interest. First, we seeded our list of papers with Yuan et al.'s set of 259 papers published between 2010-2020 at InfoVis, VAST, Vis (later SciVis), EuroVis, PacificVis, IEEE TVCG, CGF, and CG&A [91] used in their survey of *VA techniques* for ML. Second, we applied a keyword search to retrieve papers between 2020-2022 from the same venues. Our keyword selection was expansive and included visualization-specific terms such as *visualization, visual analytics, human-in-the-loop, exploratory data analysis, techniques, tools*, and ML keywords, including *artificial intelligence, machine learning, neural network, deep learning, intelligent system, and intelligent agents*. This resulted in an additional 155 papers, totaling 414 papers.

The first author reviewed all papers initially by reading the abstract and main contributions. Both authors then discussed and defined the inclusion and exclusion criteria. Given the exponential growth of research in this space and our objective for rigorous analysis, we conservatively focused on papers emphasizing the role of human expertise and tasks in VIS4ML design. To be included in our sample, we required that (1) the paper contribute at least one *interactive visualization* for the purpose of facilitating human analysis of ML, (2) the paper clearly articulated VIS4ML tasks (i.e., how the visualization tool *can effect change* to ML components through a HITL approach), (3) the paper specified one or more *design goals* for the proposed visualization system rationalizing the type of human tasks or activities in the ML pipeline, and (4) the paper included some form of evaluation.

Consequently, this eliminated a large number of papers that (1) did not state any clear design hypotheses, (2) focused on downstream usage and understanding of trust and fairness, or (3) presented tools designed to stimulate reflection on visualization techniques that could be applied in a machine learning pipeline, but which could also be used for other purposes (e.g., [69]). After filtering, we had 76 papers in our set. Of these, we sampled 52 papers. Hence, our analysis is not intended to be comprehensive or produce a taxonomy. Instead, we sought to conduct a relatively deep analysis of all of the papers in the sample (see Section 3.3) toward assessing the generalizability gaps of VIS4ML research contributions.

3.2 Codebook Development

Initially, both authors independently coded the same set of 5 papers with the high-level goal of identifying 1) how researchers motivated the incorporation of human knowledge in the ML pipeline and what forms this knowledge took, 2) what knowledge outputs or "insights" the system intended to help users reach, 3) what sorts of visualizations and automated processes they relied on, and 4) how the design hypotheses implied by the specific motivating claims were evaluated. In this initial read of the papers, we incorporated top-down influences in the form of codes adapted from prior taxonomies (section 2.1) and bottom-up influences where we identified codes from our observations.

The authors then discussed the specific aspects of the papers that are important to characterizing threats to the generalizability of VIS4ML research contributions. Through this discussion, we developed our codebook capturing aspirations about *why* the VIS4ML systems were being built, what specific types and examples of insights the papers provide to argue that human knowledge can be extended into the ML pipeline through visualization interfaces, and what datasets and modeling techniques were used in the examples. We also discussed how different sections of the paper map to different codes in our codebook. The codes span across the overall *paper-level* codes such as the paper's motivation and the nature of human expertise required to use
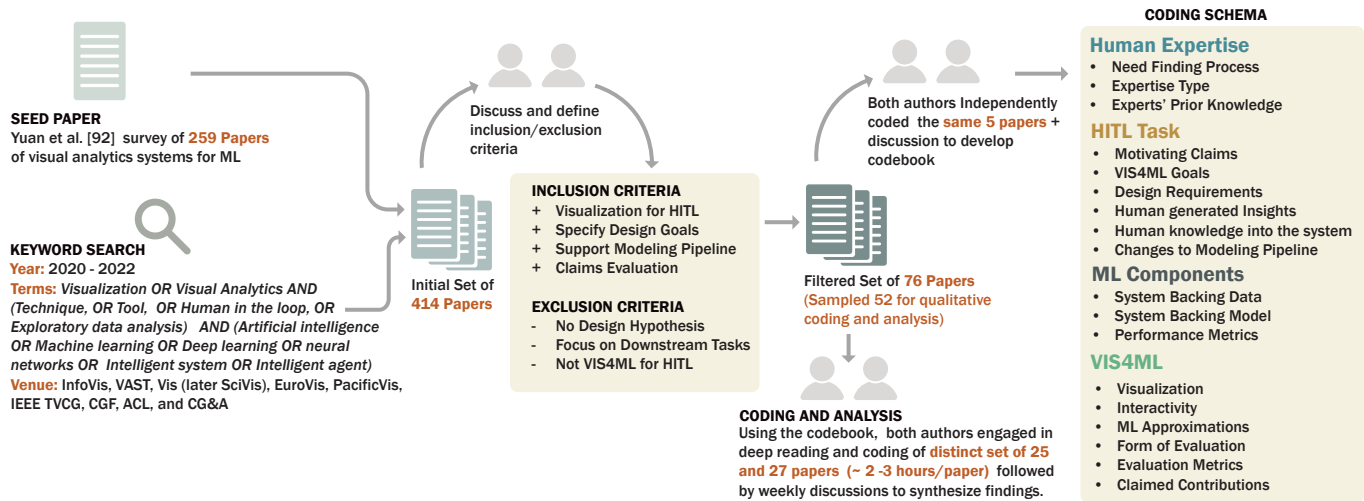
Fig. 1: Method for Paper Selection, Codebook Development, and Coding and Analysis.

the VIS4ML system, and *insight-level* in which the expert inputs their prior knowledge to glean specific insights about the ML components. Our final codebook with descriptions is available in the supplementary materials. Here we describe the main categories using examples from a recent paper on CNNs [43].

At the paper level, we coded elements like the broader *Motivating Claims for VIS4ML* (e.g., "[using the tool] experts can diagnose the potential issues of a model and refine a CNN, which enables more rapid iteration and faster convergence in model construction"), how authors *Identified Support Needs for VIS4ML* (e.g., meeting regularly with six deep learning experts over twelve months and including three as authors), the *Target Generalization Context* for the system (including the properties of the models or datasets it is intended to generalize to, e.g., "CNNs that can be formulated as a DAG with less than 100 classes"), and the overall *Evaluation Approach* (e.g., "visual data analysis and reasoning case study", adapted from an existing taxonomy for visualization evaluation [30]), as well as the specific *Evaluation Metrics* for tracking whether a system was useful (e.g., "expert insights").

We used the insight level as a more specific unit of analysis to differentiate different types of insights described as supported by the system. We coded *Forms of Human Generated Insights* based on descriptions of knowledge gained about the modeling pipeline (e.g., "[expert] identified that neurons in the lower layers learned to detect low-level features such as corners..." [43]). Given the prevalence of claims that integrating human knowledge via VIS4ML leads to improved ML use, we coded any *Actions Taken* that authors described resulting from system use. These could be concrete operations applied to the modeling pipeline (e.g., the expert added a batch normalization network then retrained, lowering model error by 9%) to more broadly construed future actions (e.g., the expert suggested they would use the system in their future model development process).

We coded dependencies for each insight, including the *Human Knowledge Required* to reach that insight (e.g., existing domain-specific or domain-general knowledge), the forms of *Feedback via Model Signals* and *Dimensionality Reduction and Other Approximations* such as preprocessing steps or other use of algorithms to transform the data on which the visualizations or interactions depend. We also coded the specific *Dataset* and *Model* associated with the insight (e.g., CIFAR-10 with a 10+2 layer CNN with cross-entropy loss and ReLu [68]).

3.3 Coding and Analysis Procedure

Both authors independently coded a distinct set of 27 and 25 papers (including re-coding the initial five papers). Each paper took between 2 − 3 hours to code in which the authors engaged in a *deep reading* of the paper to extract and map the individual information onto the codebook in separate Google Spreadsheets. This included tracing each

example insight presented in the paper, identifying the human knowledge that went into generating the insight, specific configurations, and encodings of the visualization, and interaction parameters. Throughout the coding process, the authors also made notes about salient observations about VIS4ML contributions. After coding all the papers, the authors collaboratively analyzed the coded data within each category of codes. This included weekly hour-long discussions and using digital affinity diagramming on Microsoft Whiteboard to cluster the data within each column. The authors also participated in two 3-hour long co-located discussion sessions to synthesize findings about gaps in generalizability and brainstormed recommendations for addressing them.

Naturally, our analysis is influenced by our research backgrounds. Both researchers have extensive experience in visualization research, which between them includes prior work covering aspects of collaborative design and development of human-centered AI, evaluation practices in visualization (including ML), research transparency, and statistical decision theory. Neither author identifies as an ML researcher.

4 FINDINGS

We organize our findings based on how the 52 papers in our sample characterize humans involved in VIS4ML, the scope of the ML components and pipelines, the intended HITL tasks supported with VIS4ML, and the approach to implementing VIS4ML tools given existing workflows and evaluation of the overall system and critical abstractions.

4.1 Characterizing Humans in VIS4ML

Ideally, to innovate VIS4ML tools, researchers should determine the specific nature of human expertise, prior knowledge, and skills *representative* of real-world workflows. While a majority of papers (76.9%) directly engaged with experts – individuals who have the necessary expertise to intervene in the ML modeling process – to identify requirements for VIS4ML, many lacked rationale for why and how those experts were sampled and the nature of their expertise in supporting HITL tasks.

To involve stakeholders, researchers employed various need-finding methods, including interviews (10; 19.2%), regular meetings with experts (14; 26.9%), iterative design and evaluation (11, 21.1%), and participatory design (3; 5.7%). Across these approaches, participants comprised ML experts [MLE] (26; 50%), domain experts [DoE] (5; 9.6%), or data analysts [DAE] (2; 3.8%). Further, only five papers (9.6%) included both ML and domain experts as study participants. Considering that many of the VIS4ML systems require prior knowledge or expertise spanning across ML and specific domains such as health, multi-stakeholder involvement is not prevalent. The number of participants in formative studies ranged from 2 - 20, and in 3 papers,

Table 1: List of VIS4ML papers and key paper-level columns considered in our analysis.

#	System	Experts			Prior Knowledge							HITL Task							Action		ML		Evaluation						
		MLE	DoE	DAE	Dt	Do	ML	C	SR	Dg	T	Tr	I	E	FE	MS	MD	D	O	H	Model	Data	VDAR	UP	AP	UE	C-QRI	I-QRI	CTV
1	DataDebugger [88]	■	■		■												■		■			MNIST	■		■	■			
2	AdViCE [22]	■			■			■					■								SVM	HELOC*	■						
3	iForest [93]							■	■				■								RF	Titanic, GC	■		■		■		
4	INFUSE [37]			■	■										■					■	LR,DT,kNN	Diabetes	■						
5	ActiVis [34]	■	■		■								■							■	CNN	TREC	■			■			
6	REMAP [7]	■					■	■	■	■	■						■		■		DNN	CIFAR-10	■			■			
7	RetainVis [38]	■	■		■	■		■								■			■	■	RNN	HIRA-NPS	■		■	■		■	
8	RuleMatrix [50]					■							■						■		NN	PIMA	■	■	■	■			
9	DeepVID [79]	■											■								CNN	MNIST	■			■		■	
10	BaobabView [76]					■							■	■					■		DT	Oncology	■						
11	Perturber [66]				■				■				■								CNN	ImageNet	■			■			
12	VisLRPDesigner [28]												■								CNN	ImageNet	■			■			
13	TopoAct [61]												■								DNN	ImageNet	■		■		■		
14	Visevol [8]						■	■						■	■				■			Bio*	■		■		■		
15	Boxer [21]			■										■					■			IMDB	■						
16	DeepEyes [60]	■														■			■		CNN	MNIST	■					■	
17	Blocks [3]												■						■	■	CNN	ImageNet	■			■			
18	DQNViz [78]	■					■					■							■	■	DQN		■			■			
19	ConceptExplainer [27]												■								CNN	ImageNet	■			■			
20	SliceTeller [92]	■			■	■					■		■						■	■	DNN		■			■			
21	NAS-Navigator [74]									■						■			■		CNN	CIFAR-10*	■		■	■			
22	FSLDiagnotor [89]	■														■			■		CNN	ImageNet	■			■			
23	FeatureEnVi [9]				■	■	■								■				■		XGBoost	UCI	■			■			
24	GNNLens [33]							■					■							■	GNN	Cora-ML	■			■			
25	HetVis [82]	■					■		■				■							■	CNN	Face Mask	■			■			
26	DECE [13]				■		■	■	■				■							■	NN	Pima, GC	■		■	■			
27	NeuroCartography [59]												■								CNN	ImageNet	■	■		■			
28	What-If [85]						■						■						■		LR	UCI	■				■		
29	Errudite [87]						■						■								BiDAF	SQuAD	■	■					
30	HardVis [10]				■	■							■					■	■		kNN	Cancer*	■			■			
31	VATUN [58]	■					■			■			■						■		CNN	CIFAR-10	■			■			
32	CNN Explainer [83]	■											■						■		CNN	CIFAR-10	■			■			■
33	LSTM Vis [73]	■	■				■						■						■		LSTM	Synthetic	■			■			
34	SEQ2SEQ-VIS [72]	■						■					■						■		German-Eng	IWSLT'14	■						
35	SUMMIT [25]												■						■		CNN	ImageNet	■	■					
36	Confusion Wheel [2]												■						■			UCI,MNIST	■			■	■		
37	TensorFlow Graph [86]	■	■										■						■		CNN	CIFAR-10*	■						
38	Semantic Navigator [32]				■	■							■						■		DNN		■		■	■			
39	CNNVis [43]	■								■			■						■		CNN	CIFAR-10	■			■			
40	VA Workspace [17]						■	■	■				■						■				■		■	■	■		
41	DGMTracker [42]	■					■					■								■	GAN	CIFAR-10	■			■			
42	AEVis [41]	■			■				■				■								CNN	ImageNet	■			■			
43	RNNVis [49]	■			■		■						■						■		RNN	Yelp	■			■		■	
44	TNNVis [56]	■			■		■						■							■	PoS		■					■	
45	SCS [16]												■						■		Topic Model		■	■		■			
46	BOOSTVis [44]	■					■					■							■		GBDT		■					■	
47	SCANViz [80]	■											■						■	■	SCANViz*		■			■			
48	DRLIVE [81]	■											■						■	■	DRL	Atari	■			■		■	
49	ProtoSteer [51]	■	■		■			■					■						■	■	ProSeNet	Yelp	■				■		
50	GAN Lab [35]									■			■						■		GAN		■					■	
51	Beames [14]												■						■		LR	Housing	■						
52	DRLViz [31]	■											■						■	■	DRL		■			■			

Table 1: List of VIS4ML papers and key paper-level columns considered in our analysis. **Expert** engagement includes Machine Learning Experts (**MLE**), Domain Experts (**DoE**), and Data Analysts (**DAE**). Assumed expert **prior knowledge** includes: Domain Knowledge (**Do**), Data Knowledge (**Dt**), Machine Learning Knowledge (**ML**), Tacit Knowledge (**T**), Scientific Reasoning ability (**SR**), Choice Assessment (**C**), and Diagnostics skills (**Dg**). The **human-in-the-loop tasks** are Interpreting and Assessing models (**I**), Model Selection/Choice (**MS**), Debug and Fix Errors (**E**), Model Design (**MD**), Model training (**Tr**), Feature Engineering (**FE**), and Examining/Preparing Data (**D**). **Insight-informed actions** include Observed Actions (**O**) and Hypothetical Actions (**H**). **Evaluation** taxonomy is based on [30]

the authors themselves were expert participants. While papers often implied that engaging with experts was critical to ensure the validity of their work, very few systematically reported details about the study protocol, including descriptions of specific expertise or expert knowledge, recruitment and study design, or the nature of design probes and feedback, so as to enable reproducing the methods.

4.1.1 Assumptions about Prior Knowledge and Skills

Almost all papers lacked explicit descriptions of prior knowledge and skills required of human experts to inform VIS4ML designs. However, authors' descriptions of design goals based on formative studies surfaced *implicit* assumptions and prerequisites about the expertise and skills required to be in the loop. Of the 52 papers, 11 (21.1%) mention the need for **machine learning knowledge [ML]**, including the conceptual understanding of specific models and modeling techniques, practical (hands-on) experience with training models, and the ability to comprehend model statistics and performance. For example, *LSTM Vis* [73] assumes that *"architects are deeply knowledgeable about machine learning, neural networks, and the internal structure of the system"* Further, 14 papers (26.9%) emphasize **data knowledge [Dt]** or expertise required to use VIS4ML systems. Data knowledge includes familiarity with specific datasets, contextual understanding of data and sub-groups, ground truth labels, and the ability to judge the relative importance of data instances and classes. *HardVis* [10] assumes that users are *"competent in judging the influence of a suggestion on the whole data set"* when exploring automated sampling suggestions.

In 8 papers (15.3%), authors indicated the need for **domain knowledge [Do]**, including domain-general knowledge (e.g., how language works), the ability to contextualize and comprehend model decisions, the ability to recognize good and bad model behavior, and the ability to notice errors and foresee domain consequences of bad model behavior. In describing the task of analyzing feature transformation, *FeatureEnVi* [9] requires that *"A user should be competent in judging the influence of feature transformations before applying them."* Lastly, 4 papers (7.7%) based their design choices on experiential or **tacit knowledge [T]** about modeling pipelines, including prior knowledge about performant network architecture, known constraints to search the architecture space, past experience revising the space of hyperparameters, and knowledge about critical data examples to assess model behavior.

Papers also made assumptions about the specific skill sets required to interact with VIS4ML systems. From our analysis, we identified three types of skill sets that are a combination of domain, data, and ML knowledge. Ten papers (19.2%) assume that users are able to perform **diagnostic analysis [Dg]** of the modeling pipeline, including running ablation studies, analyzing model behavior using adversarial examples, and root cause identification through exploration and inspection. Nine papers (17.3%) require that users are able to engage in **scientific reasoning [SR]** through hypothesis generation and testing, counterfactual analysis, and case-based reasoning. Finally, the design paradigms for four papers (7%) are based on choice architecture, i.e., the ability to **assess choices [C]** and make modeling decisions (e.g., comparing models, selecting from a list of automated recommendations, etc.).

4.2 Scope of ML Components and Modeling Pipeline

Based on need-finding studies with experts, papers defined concrete task requirements (36; 69.2%), identified design challenges (7; 13.4%), and derived design goals (20; 38.4%) for VIS4ML systems. By analyzing these requirements across papers, we identified researchers' *aspirations* about the scope of VIS4ML contributions within the modeling pipeline. Papers aspired to address critical challenges of **scale, generalizability, high dimensionality, perceptibility, and varied data types** in the ML pipeline. For instance, *ConceptExplainer* [27] aims to support multi-scale concept visualization (e.g., ImageNet dataset with 1.2 M images for 1000 classes), *REMAP* [7] aims to lower time and resource cost of finding performant model architectures, and *ActiVis* [34] seeks to solve model exploration for multiple types of data. Other systems aim to help analysts perceive and discover salient patterns and insights in high-dimensional data and complex model architectures (e.g., [25]).

4.2.1 Datasets and Model Types in VIS4ML Implementation

All but two papers in our sample report on specific data and models used to construct examples in the paper or run an evaluation. Most papers relied on established benchmark datasets to demonstrate tools, including but not limited to ImageNet (9; 17%), CIFAR-10 (7; 13%),

MNIST (6; 11.5%), Yelp restaurant reviews (3; 6%), and other examples from the UCI ML repository (6; 11.5%). Several demonstrated the tool using synthetic data (3; 6%). We observed some disparity in how papers valued using simple examples and well-known datasets. Many implied that using well-known datasets or benchmarks made their work stronger, but some papers commented on the need for tools to support and be evaluated on data that experts care about, for example, because using popular data like MNIST for better verification led to little insight about diagnosis and model refinement due to its simplicity [43]. Others constructed systems, for example, for pedagogical purposes, that were acknowledged to be unrealistic (e.g., simple 2D data) to avoid dependencies on approximations like dimensionality reduction [35] in surfacing the results for users.

In terms of target model types, many systems are designed for convolutional neural nets (CNNs: 20; 38.5%). Others focused on recurrent neural nets (RNNs: 8; 15.4%), generic deep neural networks (DNNs: 3, 5.8%), decision tree-based approaches (7, 13.5%), k-nearest-neighbors (3; 5.8%), GANs and deep reinforcement learning (2 each), and zero-shot models, graph neural networks, and deep Q networks (1 each). Many systems were framed as intended for large models or datasets; however, scalability in terms of classes or concepts was often a stated limitation. Scalability constraints were described in terms of numbers of classes (e.g., up to 20, 100, 1000, etc.), concepts (up to 40, 100), features, instances, feature maps, and nodes in a max-pooling layer, as well as dimensionality (of both datasets and hidden states). Some systems were described as intended for small models (e.g., [7, 83]). Others were highly specific (e.g., GANs that can run in a browser with 2D data samples [35]). Papers also occasionally described expected performance under other properties of inputs or outputs, like imbalanced data [2] or non-orthogonal concepts [80].

The papers often commented on the scalability of their contributions to broader classes or data conditions. A few systems were described as directly applicable to varying architectures (e.g., [80]), data modalities like images or text (e.g., [11]), model types (e.g., [2, 61]), and encoder types [72]). However, other times papers made generalizability claims implying that the general combination of representations or interactivity would be adaptable to other cases given further engineering.

4.3 Human-in-the-loop Tasks

All papers motivated the need for incorporating human knowledge in the modeling pipeline. By coding the descriptions in the introduction section, we identified seven categories of *human-ML interchange* in the modeling pipeline. While several papers ascribe human roles in multiple stages of the pipeline, 25 papers (48.1%) specifically emphasize human involvement to **interpret and assess [I]** model behavior. Further, 15 papers (28.8%) motivate humans' role in **debugging and fixing model errors [E]**. In targeting the earlier stages of the modeling pipeline, two papers (3.8%) mention human inputs in **examining and preparing data [D]**, three papers (5.8%) focus on **feature engineering [FE]**, five papers (9.6%) highlight the need for human expertise in **choosing the right modeling techniques [MS]**, two (3.8%) on **model design and configuration [MD]**, and three (5.8%) motivate humans' role in **monitoring and managing the model training [Tr]** process.

In identifying these human roles, a majority of papers present motivating claims that fall under one of five categories, including (1) current limitations of machine learning use in real-world, (2) requirements for machine learning applications to succeed, (3) modeling complexity, (4) limitations of current approaches for HITL, and (5) lack of support for incorporating human knowledge in the modeling pipeline. Domain criticality of ML models and the need for trust (e.g., *"knowing how entire classes are represented inside of a model is important for trusting a model's predictions…"* [25]), human effort and experience (e.g., *"…can't blindly trust automated methods (e.g., in a medical setting, doctors will want explanations of predictions),"* [13]), and scalability are prominent themes across motivation claims, though formal definitions of these goals are not given.

While the majority of reported VIS4ML task requirements are concrete, such as making *comparisons* between classes or models, in a

few cases, tasks are defined only in the abstract and lack clear descriptions of scope or task resolution (e.g., 'exploring' details or 'understanding' model behavior). Further, while many papers ground their design goals in measures of model performance, effort, effectiveness, scale, and heterogeneity of modeling characteristics, their definitions of these measures and whether or how the properties emerged from need-finding studies or were chosen a priori are not always clearly specified.

To understand ways in which VIS4ML systems support the HITL tasks described above, we coded insight examples provided in usage scenarios or case studies and actions they purportedly inspire.

4.3.1 Forms of Human Generated Insights

We identified six categories of insights across visualizations of training data, fitted models, and model representations and configuration. Given that interpretability is a central topic (nearly half of the papers in our sample), VIS4ML systems are meant to support insight generation about **inference mechanisms** (i.e., what the model learns and how it makes inferences). Examples and descriptions suggested analysts could generate causal hypotheses about the influence of structure and features on prediction results, how different models learn features, mapping between layers, classes, and concepts, what the model has learned from data, agent strategies in reinforcement learning, and how data attributes such as pixels in images contribute to classification. For instance, at the data level, in [28], the analyst attempts to learn through pixel-flipping (setting selected pixel values to zero) the relevance of water surface pixels contributing to the 'boat house' class. Or, by observing cross-class links of concepts for different vehicle classes (e.g., car windows), the analyst might conclude that the neural network has learned common features across different cars [27].

Further, through exploratory analysis, analysts might gather **evidence of erroneous behavior** and **root causes of errors** (i.e., finding and fixing errors). VIS4ML solutions were aimed at surfacing inconsistencies in model decisions, clusters with low accuracy, evidence of concept incoherent topics, lack of clear class separation, edge cases and hard-to-distinguish classes, model vulnerabilities to adversarial perturbations, etc. As an example, in the scatterplot visualization in *DataDebugger* [88], the analyst sees that the classes 'knitwear' and 'sweater' were heavily mixed. In DeepEyes [60], the analyst sees that activation of a "digit-5" associated filter also showed strong activation on digit-3, indicating perfect class separation is impossible in that layer. Visualizations also help in debugging and identifying root causes of errors, including denigrated or oversized filters, limitations with neuron cluster composition, concept entanglement, problematic layers in the network, sub-optimal agent strategies, and errors originating in different model components. For example, in DQNViz [78], the analyst is said to observe that the agent moving the paddle left and right (a strategy in the Breakout game) comprised 31% and 47% of 25,000 steps in the epoch but did not contribute to achieving rewards.

In addition to debugging model and data errors, VIS4ML systems aim to support **validation** or assertions of intended model behavior. Across our set of papers, analysts were described as being able to evaluate hypotheses about known characteristics of good training (e.g., important classifiers centered at the beginning) and similarities between neural networks and human decision rationales and confirm consistency in the model's decision-making. For instance, in *SUMMIT* [25], the analyst verifies that, similar to humans, the model classifies black bear and brown bear based on color. Further, VIS4ML systems aim to help analysts understand the **space of data and modeling choices** for subsequent decisions and actions. Concretely, analysts are thought to gain insights into the strengths and limitations of different feature selection algorithms, feature importance and which features to exclude, comparative differences between models, architecture choices such as which layer to remove in a neural network, which models to include in an ensemble, and how to slice datasets.

In the process of deducing these different insights, analysts also need to make inferences about **model performance**. To facilitate this understanding, VA tools present information about the model depth and classification accuracy, model accuracy for different classes, which model performs best, data slices and performance, data quality and ground truth impact on performance, model convergence time, memory, and compute time during training, etc.

4.3.2 Model Signals for Insights

To support users' inference processes, designers of VIS4ML systems must identify what forms of feedback on model quality, interpretability, or performance to surface. Most papers surfaced some form of metric to provide feedback. Of the 52 papers, 23 (44%) provided information about model performance using confusion matrices, plain text, or line charts showing loss and accuracy curves. In addition, 7 papers (13.4%) provided information about class probabilities using text or color-coded bar charts. Lastly, 11 papers (21.1%) provided feedback about modeling tasks such as training time, number of parameters, number of data items corrected through active learning, delta-changes or improvement to model performance, and feature importance. The level of motivation for the specific model signals that papers used varied considerably, with many papers providing a very brief motivation and a few providing more rigorous motivation of why the chosen signals were good estimators for improving model performance.

4.3.3 Insight-informed Actions

Given the prevalence of claims that integrating human knowledge via VIS4ML leads to improved ML use, we wanted to see what sorts of actions—from concrete operations applied to the modeling pipeline to more broadly construed future actions—papers described resulting from the insight gains of a system. For each human-generated insight, we coded any actions described as taken based on that insight.

We noted whether the described actions occurred in the context of *author-provided* examples, such as fictional case studies or running examples used throughout the paper, or *expert case studies*. We also noted whether the actions were **observed [O]** (e.g., actually applied to a modeling pipeline), such as when papers described actions an expert took based on using the tool or presented results from re-training a model after implementing a change, versus actions that were **hypothetical [H]**, i.e., that could be taken or were referred to as possible future steps. Finally, we noted whether specific before and after performance comparisons were made (e.g., accuracy comparisons) to validate the utility of observed actions in the larger ML pipeline or research endeavor.

Observed actions: Overall, the majority of papers (42; 82.7%) described some action as the result of an insight gained from the VIS4ML system. Of these, 27 (about half of the total 52 papers) described at least one observed action. About a third of the total papers (15 of 52) described at least one action taken by an expert. The remainder (23.1% of 52) were actions by the authors as part of the case studies they designed. Examples of actions that authors or system users took based on insights include *changes to the training data*, such as sampling to deal with class imbalances or modifying training labels; *changes to the features*, like switching input images to grayscale [3]; *changes to a model representation*, like adding or removing rules from rule-based classifiers [2] or moving neurons between clusters [43]; *changes to model parameters*, like reducing the number of latent dimensions [80]; changes to the model architecture, like adjusting layers or filters of deep NNs [7, 60]; and changes to the training process, like changing the learning rate [35] or variance sampler [42] of deep generative models.

Slightly more than one third of the total papers (19 of 52) described or quantified the effect of an action on the modeling pipeline. Most cited numeric changes in model accuracy or error. Several papers presented accuracy, precision, and recall statistics (e.g., [9]) to acknowledge trade-offs or changes in training speed (e.g., [60]). A few other papers reported on trade-offs between accuracy and interpretability, such as how accuracy remained similar after human-driven adjustments while the number of nodes in the latent representation significantly reduced [76]. Hence, over 60% of the papers in our sample provided no concrete evidence of the impacts of the VIS4ML system on the ML pipeline.

Hypothetical actions: Of the 25 papers where no concrete action was observed, 15 (28.5%) described a hypothetical action. Six papers (11.5%) referred to hypothetical actions proposed by an expert in a case study; the remainder were proposed by authors as part of examples they developed. Hypothetical actions could be well defined, such as when papers mentioned specific actions on a pipeline that could be taken upon reaching some specific insight, e.g., reparameterizing a deep RL system after observing that a lower dimensional representation appears to have explanatory power [81]. Other hypothetical actions on a pipeline were referred to in less specific terms, such as informing data collection or experiment design [3, 51, 73], allowing fine-tuning of a model [2, 81, 85], or supporting bug finding [85].

Other hypothetical actions concerned changes to experts' processes, such as when experts said that they would incorporate the system in their future model development processes [43], or their insights were thought to inform subsequent analysis or theoretical investigations [49, 73, 78], or future research or other "endeavors" to improve such models [13, 33].

Of the 10 papers where neither observed nor hypothetical actions were described, one described a system developed for educating non-experts [83], noting that they chose this goal because supporting an interactive training process would be unrealistic. Other papers in this group focused on interpretation goals without necessarily providing reasons why actions were not considered.

4.4 VIS4ML Implementation and Evaluation

Based on need-finding studies, papers identified implementation desiderata for supporting human understanding, design, and improvement of the model. Specifically, the papers aimed for their systems to **align with existing modeling practices and workflows.** Through their interactions with experts, authors either captured existing task workflows or defined new workflows for VA tasks. For instance, *FeatureEnVi* [9] defines a unified workflow for feature engineering by "fusing stepwise selection and semi-automatic extraction approaches." A few papers emphasized **human-knowledge integration and guided discovery** of model and data characteristics. In rationalizing system design considerations, authors made connections to human knowledge and comprehension support needs. For instance, in designing *NAS-Navigator* [75], authors intended that users be able to design and edit template models based on their experience. Additionally, some authors hoped to **support varied work environments** and overcome ML deployment challenges.

Prior taxonomies broadly describe the specific visualizations and interaction formats we observed across papers. Hence we focus on papers' reliance on abstraction techniques for model interpretability and approaches to VIS4ML evaluation.

4.4.1 Post-Hoc Interpretability Methods

We observed frequent use of dynamic algorithms and approximating representations in the visualizations employed in VIS4ML tools. Many papers used dimensionality reduction techniques (e.g., PCA, UMAP, MDS; 12 papers; 23.1%) and/or projection-based visualization techniques like t-SNE (16; 30.8%) to visualize high dimensional data in 2D. While t-SNE generated layouts are guaranteed to recover structure in high dimensional data under certain conditions [4], without proper tuning, they can produce artifacts that mislead users to perceive structure that doesn't exist [84]. Two papers seemed to acknowledge such limitations, noting that they had intentionally opted not to use projections like t-SNE, for example, because the authors "found that if the tool loses its inherent connection to the data, results were less interpretable to the user" [73].

Similarly, the machine learning interpretability literature has contributed a number of post-hoc explanation methods designed to provide intuition into how a model reaches a decision or what it has learned. The VIS4ML systems we surveyed made frequent use of feature attribution approaches such as feature visualization through partial dependency plots and other graphics in feature space (13; 25%) or deriving of importance scores for ranking or recommendation (7; 13.5%).

Pixel-based saliency maps and other forms of activation heatmaps were also used (5; 9.6%). In a few cases, papers referred to limitations of these approaches, such as by discussing how maps of salient image patches activating a neuron were not appropriate for explaining the activity in neurons of very deep CNNs, where activations are influenced by very large patches [41].

For VIS4ML tools to be implemented in practice requires transparency on how hyperparameters used in critical approximating representations are set, including any tuning processes used. Papers varied considerably in the extent to which they specified such information for dynamic visualization algorithms like t-SNE, clustering approaches, or other optimizations. Interactive hyperparameter tuning, such as the ability to set a target k for clustering algorithms, was occasionally made available to end users as a strategy to avoid dependence on authors' decisions. In cases where it was not but a parameterized technique was used, some authors described only the values they set, while others described the values and the tuning process they used, and others provided suggestions for how those adapting the system should use or not use the specific values or process they used. Occasionally papers did not commit to any single technique, instead mentioning various projection methods that could be used (e.g., [72]).

4.4.2 VIS4ML Evaluation

To understand VIS4ML evaluation as a subset of visualization evaluation more broadly, we used Isenberg et al.'s [30] adaption of Lam et al.'s taxonomy [39] to characterize forms of visualization evaluation and the metrics or outcome variables associated with them. We observed instances of all evaluation styles except *Evaluating Collaborative Data Analysis*. We report on formative studies aimed at *Understanding Work Practices* described above.

Most notably, every paper in our sample exemplified validation through **Visual Data Analysis and Reasoning [VDAR]**. VDAR captures case-study style evaluations of how a visualization tool supports analysis and reasoning about data and allows domain experts to derive knowledge. Case studies could be based on collaboration or observation of expert users or describe hypothetical users' analysis processes (Section 4.3.3). Sometimes experts were paired with authors in case studies, though similar to Sperlle et al.'s [70] observation of missing details in evaluation, it was not always clear whether authors were involved. In reporting these scenarios, however, papers narrated VDAR processes by describing how observations at various points could be interpreted as evidence of certain facts about the data, model, or pipeline.

The implied evaluation metric of a VDAR-style evaluation is the validity of the insights that are generated. As discussed above, only about one-third of the papers provided evidence of improved model performance as a result of insight-informed actions. Hence, most of these evaluations inherit the ambiguity of what makes an insight meaningful or important, which is frequently discussed in visualization research [57, 90]. In a couple of papers, authors performed robustness testing to establish the validity of insights, such as testing to verify that a user's insights generalize beyond the specific input data and/or model architecture used [66] or using PCA to validate structures identified using a contributed topological summary approach for exploring DNN activations [61].

More commonly, however, expert judgment was implied to establish that insights were useful. Authors often signaled why a process was meaningful by interspersing quotes from the experts' thinking aloud as they used the tool. For example, some papers included quotes from experts describing how they were able to use a system to confirm prior knowledge, such as confirming evidence of under- or overfitting in comparing CNN models that varied in complexity [43]. In other cases, quotes summarized how the insights or actions they were able to achieve with the system represented solutions to problems they often face (e.g., [43]). Others described how insights stimulated further consideration on the part of users.

A few papers implied that an insight gained from a VDAR process was valid because it corroborated an observation or a hunch identified in prior work. For example, papers describe how an observation "confirms earlier work that demonstrates simple context-free models in

RNNs and LSTMs" [72, 73] or how the absence of positive values in a DNN's training process was "observed, but not explained" by prior work [42].

VDAR evaluations were commonly paired with subjective feedback and opinions from users (**User Experience [UE]**; 65.4% of total papers). Many papers used think-aloud protocol and/or semi-structured interviews to gather qualitative feedback from VDAR participants. Others elicited Likert-style responses on the usefulness and usability of a tool. Less commonly, surveys were used to elicit experts' self-assessments of their findings (e.g., [87]). This was rare, however; in most cases, it was implicit that because an expert made an observation, they must have faith in that observation. User Experience observations were most frequently used to establish that experts found the system useful and to describe potential improvements.

Sixteen papers (30.8%) included what prior surveys of visualization evaluations have called **isolated Qualitative Results Inspection [I-QRI]**, referring to validation of techniques by inspecting a technique's results in isolation, along with a description of how the technique achieves some result. All isolated QRI examples we observed were used to validate the visualization techniques used. For example, papers might point to how a certain pattern exemplified in a visualization was indicative of some fact about the latent representation, training process, or other parts of the pipeline: "it is evident in Fig. 3b that the digits 1, 2 and 7 are often confused with each other, as their shapes are similar to some degree [2]. On the other hand, five papers (9.6%) used **comparative (QRI) [C-QRI]** to compare visual outputs to other state-of-the-art techniques to suggest that an adopted approach is superior (e.g., "Without Fourier basis parametrization, the differences between the models are more visually distinct" [66]).

Beyond qualitative evaluations of user experience, we observed a smaller number of papers (8; 15.4%) using a **User Performance [UP]** style evaluation, where the performance of users with the system (or independently with the visualizations) was recorded to isolate the effects of specific features. For example, [88] used logged data as part of an expert case study to compare improvements in label accuracy after iterations of an algorithm. Other studies recorded task completion time and/or accuracy on predetermined tasks given to users to evaluate effectiveness [50, 93], or scored (via expert ratings, or automatically) the results of an analysis session [16, 17]. A few papers [72, 85] reported usage statistics such as page views from open-source releases.

Ten papers (19.2%) described **Algorithm Performance [AP]** evaluations, consisting of quantitative studies of the performance or quality of visualization algorithms or algorithms on which the visualized information depends. A few focused on improving ML model understanding among novices [35, 83], using an *Evaluating Communication Through Visualization (CTV)* approach focused on how successfully novice end users grasped important concepts as communicated by the tool.

5 DISCUSSION

Our findings characterize the current scope of VIS4ML research within the broader requirements of human-in-the-loop in ML production. Many of the papers we reviewed make bespoke visualization contributions demonstrated by specific use cases and evaluate them by taking study participants' insights taken at face value. This is not necessarily problematic if the goals of VIS4ML research are exploratory, driven by the purpose of identifying new engineering solutions for surfacing certain information in ML pipelines and reporting lessons learned in the process. This kind of design study methodology has become an established mode for visualization research [47]. However, we observed that papers often used their case studies to draw general conclusions about the utility of certain representations or interaction designs for improving the performance of the model or pipeline, such as by making claims about how a system demonstrates the value of integrating human knowledge. This suggests an overlooked distinction between the nature of a design study versus a controlled empirical comparison [47], and between engineering artifacts and scientific knowledge [29]. Here we discuss specific threats to the generalizability of VIS4ML research grounded in our analysis and propose near-term and future-looking recommendations for improving the alignment between claims made in VIS4ML research and the procedures used to validate them.

5.1 Threats to Generalizability

Ambiguous characterization of human expertise: Similar to prior work by Sperrle et al. [70], we also observe strong yet often implicit knowledge assumptions of users of VIS4ML systems. At times, papers' claims about the accessibility of a tool seemed to contradict the knowledge needed to use the system confidently; e.g., one paper presented a hypothetical scenario involving a domain expert who was "not conversant with complex modeling techniques," using a tool that required interpreting residuals and parameter weights. Describing required expertise and reporting sources of confusion and the amount of training needed for experts to use the system successfully (as a few papers did [2, 43, 44]) is vital for transparency around dependencies, and will benefit practitioners who try to adopt the work.

Overfitting to specific use cases: The predominance of hypothetical usage scenarios and case studies suggests that success is often interpreted as a paper's ability to demonstrate a few (usually 1-3) instances in which a VIS4ML system appears to provide value. A risk of this form of validation is that researchers will inadvertently build tools to confirm their a priori knowledge about some dataset or modeling pipeline but which are overfit to those specific insights. We observed papers often using common datasets that have been thoroughly explored in the ML community, and validating systems by showing that previously identified patterns were also visible using the new system. The problem is that being able to present at least one usage scenario in which a tool is perceived as helpful is different from showing that, on average, using that tool improves some aspect of the ML pipeline. The former takes the form of an existence proof, whereas the latter requires studying a greater number of cases relative to a baseline representing what experts would do without a tool.

Constrained evaluation practices: Overfitting can also arise from heavy reliance on a few experts as consultants in the design process, especially when those same experts are consulted to evaluate a system, as we observed in multiple papers. Based on our analysis, papers consistently rely on certain forms of evaluative evidence—namely, a few usage scenarios or case studies with small numbers of experts in which the experts make what they perceive as discoveries or confirm their prior expectations—to validate the work. While these practices are not necessarily unreasonable for a burgeoning research area in which researchers invest considerable effort in making tools to support ML practice they may not be intimately familiar with, failure to acknowledge such "validation gaps" threatens the transparency and generalizability of VIS4ML research.

Postulating broader utility from exploratory evidence: Helping users draw causal inferences to improve a learning pipeline is a common implicit goal in VIS4ML research. However, post-hoc visualizations, even when interactive, are limited in their ability to validate many types of causal hypotheses [54]. Sometimes papers acknowledged these limitations, such as by calling observations "speculative" [61], noting that it is difficult to provide specific reasoning about what led to a model decision [72], and noting that whether an action could effectively address a perceived problem requires further investigations [81] or statistical evidence [31]. However, across the papers in our sample such statements were sparse.

Only about one-third of papers reported performance statistics for a usage scenario as evidence that a model had been improved through human involvement. Others relied on experts' statements that a tool was useful through user experience questionnaires or interviews or simply the authors' ability to construct a hypothetical use scenario. While users may be able to confirm propositions about the specific model or pipeline at hand with a VIS4ML tool (e.g., whether removing subsets of data or changing certain hyperparameters affects model output for a certain validation set), hypotheses about the *general* behaviors of an approach cannot be verified without testing under new conditions. These include, for example, hypotheses that a given class of

models learns in a particular way or is subject to certain blindspots. A few papers acknowledged the potential for overfitting when users' explanations couldn't be tested on new data [2, 14]. One paper even reported how some expert users of a tool disapproved of the idea of using post-hoc rule extraction to interpret a model's decisions based on potential overfitting and described conditions to avoid this, such as the collection of new data [2]. However, such acknowledgments were rare.

If the goal is to investigate whether human knowledge can *ever* be helpful and how it might be integrated, then papers' motivations should reflect this, for example, by posing questions rather than as-sertions. However, many papers motivate the work by referring to the demonstrated power of VIS4ML to overcome the shortcomings of fully automated ML pipelines. For example, papers cite prior VIS4ML work in describing how "As shown by many recent works..., inter-preting DNNs with visual analytics has achieved great success" [81], or "These visualizations have achieved great success in understanding and analyzing those deep learning models", referring to visualizations to facilitate developing CNNs, RNNs, GANs, and DQNs [33]. While not necessarily false, such "success" is narrowly defined.

Underspecified effort to insights: How much difficulty researchers faced in identifying "successful" use cases for a tool is likely to be a useful signal of how reproducibly a tool can improve ML practice. However, papers generally did not describe the selection of specific forms of insights they reported in VDAR-style evaluations, leaving it unclear how unique the scenarios they demonstrated might be. An ex-ception to this is a paper that explicitly noted how the authors explored a large space of possible VDAR processes that the system afforded before deciding on the selected scenarios [73]. However, it remains unclear whether such exploration was needed to identify successful examples.

5.2 Recommendations for Action

Below we summarize near-term and forward-looking aims to address the generalizability gaps listed above. Our recommendations are in-formed by proposed solutions to larger problems of a lack of rigor and robustness in social science research and may be applicable to the broader area of visual analytics application design. However, our suggestions should be taken as tentative, as it is beyond the scope of our paper to rigorously evaluate these reform proposals as should be expected if reforms are to be effective [15].

5.2.1 Documenting Constraints on Generalizability

Some of the gaps we highlight can be addressed simply by improving documentation practices about known dependencies in VIS4ML re-search. Conventional reporting styles in VIS4ML research do not ade-quately describe the constraints on generality: What dependencies the success of the system may have on the specific configuration of com-ponents, users, and other specifics of the setting that was studied. Our results suggest extending prior reporting guidelines for VIS4ML sys-tems [43,70] to encourage reporting of hidden dependencies, including (1) all datasets used in examples and evaluations, (2) all scalability and memory-related constraints, (3) expectations about how much time users will need to learn the system (and reporting of learning time for expert case studies), and (4) observed sources of confusion and failures among users. In addition, to help "close the loop" researchers should (5) provide model, pipeline, and deployment summary stats before and after changes inspired by using the system. For reproducibility, they should (6) describe parameters and values applied in examples for any parametrized pre-processing steps as well as how those values were reached. Finally, given that VIS4ML claims are demonstrated through specific examples and insights, it is important that the researcher (7) report the nature of their own exploration and insight generation (i.e., time and effort in identifying example insights). Future work should investigate documentation standards and reporting guidelines by tak-ing inspiration from other disciplines such as the social sciences [67].

5.2.2 Tightening Logical Derivation Chains

VIS4ML papers should also carefully consider and communicate the deductive logic behind the choices that are made, from motivating the research to defining the conditions of the study. Papers should avoid *loose derivation chains*, to borrow a term applied by Paul Meehl to underspecification in experimental research [46]: a lack of evidence of rigorous deductions in moving from theoretical premises to predic-tions or choices made regarding observed relations. The exploratory nature of VIS4ML research may mean that achieving tight derivation chains is not realistic for many projects. However, papers could be improved by striving to document where choices were made more ar-bitrarily (e.g., out of convenience) in deciding how to instantiate hy-potheses or aspirations in systems. Along these lines, future work might take inspiration from the design study literature, such as in the design activity framework [45], to develop similar methods specific to VIS4ML decisions for helping designers reason about and identify best practices in connecting design goals, methods, and outcomes. Fu-ture research could extend visualization design and evaluation frame-works (e.g., the nested model [53]) to emphasize human expertise and prior knowledge captured in VIS4ML design hypotheses and represen-tativeness of said expertise in real-world practice.

5.2.3 Bridging from Design Studies to VIS4ML in Practice

Despite the strength of claims that we observed in the VIS4ML liter-ature, the nature of many of the studies we analyzed appears closer to a design study pattern, in which "researchers analyze a specific real-world problem faced by domain experts, design a visualization system that supports solving this problem, validate the design, and reflect about lessons learned in order to refine visualization design guidelines" [65]. Notions of rigor and validation look different in such studies [47, 48, 53], and expecting replicability and generalizability of VIS4ML research may be premature. While Meyer and Dykes sug-gest that design researchers "aim to produce explicit and appropriately scoped expressions of knowledge claims," our analysis suggests that properly scoping expressions of knowledge claims may require more concrete guidance to achieve.

In the longer term, VIS4ML research could benefit from forging stronger partnerships with adjacent ML and HCI communities. As VIS4ML brings visualization research towards the center of data sci-ence, ML, and human-centered AI, VIS4ML research should look beyond 'insightism'–the superficial reliance on apparent insights pro-duced by use–into pragmatism (usefulness) and cognitivism (impacts on individual and social cognition) to really put the human in the loop [12]. Bridging research to ML practice requires exploring ways to negotiate responsibilities, building deeper research collaborations between visualization and ML researchers (e.g., CARE-ful partner-ships [1]), defining boundary objects for knowledge transfer, and ad-dressing the cost and effort of replicating findings in VIS4ML re-search.

6 CONCLUSION

We contribute a focused analysis of 52 VIS4ML papers representing design hypotheses about how integrating human knowledge can help "close the loop" in ML practice. We observe a general optimism about the potential for human integration to transform ML practice and re-search and heavy reliance on collaborations with experts who such tools might help. However, these aspirations are not always accom-panied by evaluations demonstrating success in these goals. Our find-ings show gaps in the generalizability of VIS4ML research contribu-tions, indicating that we are only closing a narrow instantiation of the human-in-the-loop model. We make recommendations for action that include transparent reporting, tightening logical derivation chains in VIS4ML research practices, and extending current design study ap-proaches to ML practices by exploring partnerships with the broader human-centered AI research community.

7 ACKNOWLEDGMENTS

REFERENCES

[1] D. Akbaba, D. Lange, M. Correll, A. Lex, and M. Meyer. Troubling collaboration: Matters of care for visualization design study. 2023. 9

[2] B. Alsallakh, A. Hanbury, H. Hauser, S. Miksch, and A. Rauber. Visual methods for analyzing probabilistic classification data. *IEEE transactions on visualization and computer graphics*, 20(12):1703–1712, 2014. 4, 5, 6, 7, 8, 9

[3] B. Alsallakh, A. Jourabloo, M. Ye, X. Liu, and L. Ren. Do convolutional neural networks learn class hierarchy? *IEEE transactions on visualization and computer graphics*, 24(1):152–162, 2017. 4, 6, 7

[4] S. Arora, W. Hu, and P. K. Kothari. An analysis of the t-sne algorithm for data visualization. In *Conference On Learning Theory*, pp. 1455–1462. PMLR, 2018. 7

[5] N. Boukhelifa, A. Bezerianos, R. Chang, C. Collins, S. Drucker, A. Endert, J. Hullman, C. North, and M. Sedlmair. Challenges in evaluating interactive visual machine learning systems. *IEEE Computer Graphics and Applications*, 40(6):88–96, 2020. 2

[6] E. T. Brown, A. Endert, and R. Chang. Human-machine-learner interaction: The best of both worlds. In *Proceedings of the CHI Workshop on Human Centred Machine Learning (HCML)*, 2016. 1

[7] D. Cashman, A. Perer, R. Chang, and H. Strobelt. Ablate, variate, and contemplate: Visual analytics for discovering neural architectures. *IEEE transactions on visualization and computer graphics*, 26(1):863–873, 2019. 4, 5, 6

[8] A. Chatzimparmpas, R. M. Martins, K. Kucher, and A. Kerren. Visevol: Visual analytics to support hyperparameter search through evolutionary optimization. In *Computer Graphics Forum*, vol. 40, pp. 201–214. Wiley Online Library, 2021. 4

[9] A. Chatzimparmpas, R. M. Martins, K. Kucher, and A. Kerren. Featureenvi: Visual analytics for feature engineering using stepwise selection and semi-automatic extraction approaches. *IEEE Transactions on Visualization and Computer Graphics*, 28(4):1773–1791, 2022. 4, 5, 6, 7

[10] A. Chatzimparmpas, F. V. Paulovich, and A. Kerren. Hardvis: Visual analytics to handle instance hardness using undersampling and oversampling techniques. *arXiv preprint arXiv:2203.15753*, 2022. 4, 5

[11] S. Chaudhuri, V. Ganti, and R. Kaushik. Data debugger: An operator-centric approach for data quality solutions. *IEEE Data Eng. Bull.*, 29(2):60–66, 2006. 5

[12] M. Chen and D. J. Edwards. "isms" in visualization. *Foundations of Data Visualization*, pp. 225–241, 2020. 9

[13] F. Cheng, Y. Ming, and H. Qu. Dece: Decision explorer with counterfactual explanations for machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1438–1447, 2020. 4, 5, 7

[14] S. Das, D. Cashman, R. Chang, and A. Endert. Beames: Interactive multimodel steering, selection, and inspection for regression tasks. *IEEE computer graphics and applications*, 39(5):20–32, 2019. 4, 9

[15] B. Devezer, D. J. Navarro, J. Vandekerckhove, and E. O. Buzbas. The case for formal methodology in scientific reform. 2020. doi: 10.1101/2020.04.26.048306 9

[16] M. El-Assady, R. Kehlbeck, C. Collins, D. Keim, and O. Deussen. Semantic concept spaces: Guided topic model refinement using word-embedding projections. *IEEE transactions on visualization and computer graphics*, 26(1):1001–1011, 2019. 4, 8

[17] 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, 2018. 4, 8

[18] A. Endert, M. S. Hossain, N. Ramakrishnan, C. North, P. Fiaux, and C. Andrews. The human is the loop: new directions for visual analytics. *Journal of intelligent information systems*, 43(3):411–435, 2014. 2

[19] A. Endert, W. Ribarsky, C. Turkay, B. W. Wong, I. Nabney, I. D. Blanco, and F. Rossi. The state of the art in integrating machine learning into visual analytics. In *Computer Graphics Forum*, vol. 36, pp. 458–486. Wiley Online Library, 2017. 1

[20] P. Federico, M. Wagner, A. Rind, A. Amor-Amorós, S. Miksch, and W. Aigner. The role of explicit knowledge: A conceptual model of knowledge-assisted visual analytics. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 92–103. IEEE, 2017. doi: 10.1109/VAST.2017.8585498 2

[21] M. Gleicher, A. Barve, X. Yu, and F. Heimerl. Boxer: Interactive comparison of classifier results. In *Computer Graphics Forum*, vol. 39, pp. 181–193. Wiley Online Library, 2020. 4

[22] O. Gomez, S. Holter, J. Yuan, and E. Bertini. Advice: Aggregated visual counterfactual explanations for machine learning model validation. In *2021 IEEE Visualization Conference (VIS)*, pp. 31–35. IEEE, 2021. 4

[23] B. Haibe-Kains, G. A. Adam, A. Hosny, F. Khodakarami, L. Waldron, B. Wang, C. McIntosh, A. Goldenberg, A. Kundaje, C. S. Greene, et al. Transparency and reproducibility in artificial intelligence. *Nature*, 586(7829):E14–E16, 2020. 2

[24] F. Hohman, M. Kahng, R. Pienta, and D. H. Chau. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE transactions on visualization and computer graphics*, 25(8):2674–2693, 2018. 1, 2

[25] F. Hohman, H. Park, C. Robinson, and D. H. P. Chau. Summit: Scaling deep learning interpretability by visualizing activation and attribution summarizations. *IEEE transactions on visualization and computer graphics*, 26(1):1096–1106, 2019. 4, 5, 6

[26] S. R. Hong, J. Hullman, and E. Bertini. Human factors in model interpretability: Industry practices, challenges, and needs. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–26, 2020. 2

[27] J. Huang, A. Mishra, B. C. Kwon, and C. Bryan. Conceptexplainer: Interactive explanation for deep neural networks from a concept perspective. *IEEE Transactions on Visualization and Computer Graphics*, 2022. 1, 4, 5, 6

[28] X. Huang, S. Jamonnak, Y. Zhao, T. H. Wu, and W. Xu. A visual designer of layer-wise relevance propagation models. In *Computer Graphics Forum*, vol. 40, pp. 227–238. Wiley Online Library, 2021. 4, 6

[29] J. Hullman, S. Kapoor, P. Nanayakkara, A. Gelman, and A. Narayanan. The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning. *arXiv preprint arXiv:2203.06498*, 2022. 2, 8

[30] 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. 3, 4, 7

[31] T. Jaunet, R. Vuillemot, and C. Wolf. Drlviz: Understanding decisions and memory in deep reinforcement learning. In *Computer Graphics Forum*, vol. 39, pp. 49–61. Wiley Online Library, 2020. 4, 8

[32] S. Jia, Z. Li, N. Chen, and J. Zhang. Towards visual explainable active learning for zero-shot classification. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):791–801, 2021. 4

[33] Z. Jin, Y. Wang, Q. Wang, Y. Ming, T. Ma, and H. Qu. Gnnlens: A visual analytics approach for prediction error diagnosis of graph neural networks. *IEEE Transactions on Visualization and Computer Graphics*, 2022. 4, 7, 9

[34] M. Kahng, P. Y. Andrews, A. Kalro, and D. H. Chau. A cti v is: Visual exploration of industry-scale deep neural network models. *IEEE transactions on visualization and computer graphics*, 24(1):88–97, 2017. 4, 5

[35] M. Kahng, N. Thorat, D. H. Chau, F. B. Viégas, and M. Wattenberg. Gan lab: Understanding complex deep generative models using interactive visual experimentation. *IEEE transactions on visualization and computer graphics*, 25(1):310–320, 2018. 4, 5, 6, 8

[36] D. Kerrigan, J. Hullman, and E. Bertini. A survey of domain knowledge elicitation in applied machine learning. *Multimodal Technologies and Interaction*, 5(12):73, 2021. 2

[37] J. Krause, A. Perer, and E. Bertini. Infuse: interactive feature selection for predictive modeling of high dimensional data. *IEEE transactions on visualization and computer graphics*, 20(12):1614–1623, 2014. 1, 4

[38] B. C. Kwon, M.-J. Choi, J. T. Kim, E. Choi, Y. B. Kim, S. Kwon, J. Sun, and J. Choo. Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. *IEEE transactions on visualization and computer graphics*, 25(1):299–309, 2018. 4

[39] 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, 2011. 7

[40] T. Liao, R. Taori, I. D. Raji, and L. Schmidt. Are we learning yet? a meta review of evaluation failures across machine learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 2

[41] M. Liu, S. Liu, H. Su, K. Cao, and J. Zhu. Analyzing the noise robustness of deep neural networks. In *2018 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 60–71. IEEE, 2018. 4, 7

[42] M. Liu, J. Shi, K. Cao, J. Zhu, and S. Liu. Analyzing the training processes of deep generative models. *IEEE transactions on visualization and*

computer graphics, 24(1):77–87, 2017. 4, 6, 8

[43] M. Liu, J. Shi, Z. Li, C. Li, J. Zhu, and S. Liu. Towards better analysis of deep convolutional neural networks. *IEEE transactions on visualization and computer graphics*, 23(1):91–100, 2016. 3, 4, 5, 6, 7, 8, 9

[44] S. Liu, J. Xiao, J. Liu, X. Wang, J. Wu, and J. Zhu. Visual diagnosis of tree boosting methods. *IEEE transactions on visualization and computer graphics*, 24(1):163–173, 2017. 4, 8

[45] S. McKenna, D. Mazur, J. Agutter, and M. Meyer. Design activity framework for visualization design. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2191–2200, 2014. 9

[46] P. E. Meehl. Why summaries of research on psychological theories are often uninterpretable. *Psychological reports*, 66(1):195–244, 1990. 9

[47] M. Meyer and J. Dykes. Criteria for rigor in visualization design study. *IEEE transactions on visualization and computer graphics*, 26(1):87–97, 2019. 8, 9

[48] M. Meyer, M. Sedlmair, P. S. Quinan, and T. Munzner. The nested blocks and guidelines model. *Information Visualization*, 14(3):234–249, 2015. 9

[49] Y. Ming, S. Cao, R. Zhang, Z. Li, Y. Chen, Y. Song, and H. Qu. Understanding hidden memories of recurrent neural networks. In *2017 IEEE conference on visual analytics science and technology (VAST)*, pp. 13–24. IEEE, 2017. 4, 7

[50] 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, 2018. 4, 8

[51] Y. Ming, P. Xu, F. Cheng, H. Qu, and L. Ren. Protosteer: Steering deep sequence model with prototypes. *IEEE transactions on visualization and computer graphics*, 26(1):238–248, 2019. 4, 7

[52] S. Mohseni, N. Zarei, and E. D. Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3-4):1–45, 2021. 2

[53] T. Munzner. A nested model for visualization design and validation. *IEEE transactions on visualization and computer graphics*, 15(6):921–928, 2009. 9

[54] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019. 8

[55] F. Neutatz, B. Chen, Z. Abedjan, and E. Wu. From cleaning before ml to cleaning for ml. *IEEE Data Eng. Bull.*, 44(1):24–41, 2021. 2

[56] S. Nie, C. Healey, K. Padia, S. Leeman-Munk, J. Benson, D. Caira, S. Sethi, and R. Devarajan. Visualizing deep neural networks for text analytics. In *2018 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 180–189. IEEE, 2018. 4

[57] C. North. Toward measuring visualization insight. *IEEE Computer Graphics and Applications*, 26(3):6–9, 2006. doi: 10.1109/MCG.2006.70 7

[58] C. Park, S. Yang, I. Na, S. Chung, S. Shin, B. C. Kwon, D. Park, and J. Choo. Vatun: Visual analytics for testing and understanding convolutional neural networks. In *Eurographics Conference on Visualization (EuroVis)-Short Papers*, 2021. 4

[59] H. Park, N. Das, R. Duggal, A. P. Wright, O. Shaikh, F. Hohman, and D. H. P. Chau. Neurocartography: Scalable automatic visual summarization of concepts in deep neural networks. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):813–823, 2021. 4

[60] N. Pezzotti, T. Höllt, J. Van Gemert, B. P. Lelieveldt, E. Eisemann, and A. Vilanova. Deepeyes: Progressive visual analytics for designing deep neural networks. *IEEE transactions on visualization and computer graphics*, 24(1):98–108, 2017. 4, 6

[61] A. Rathore, N. Chalapathi, S. Palande, and B. Wang. Topoact: Visually exploring the shape of activations in deep learning. In *Computer Graphics Forum*, vol. 40, pp. 382–397. Wiley Online Library, 2021. 4, 5, 7, 8

[62] D. Sacha, M. Kraus, D. A. Keim, and M. Chen. Vis4ml: An ontology for visual analytics assisted machine learning. *IEEE transactions on visualization and computer graphics*, 25(1):385–395, 2018. 1, 2

[63] D. Sacha, M. Sedlmair, L. Zhang, J. A. Lee, D. Weiskopf, S. North, and D. Keim. Human-centered machine learning through interactive visualization. ESANN, 2016. 1, 2

[64] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, and D. A. Keim. Knowledge generation model for visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1604–1613, 2014. doi: 10.1109/TVCG.2014.2346481 2

[65] 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. 9

[66] S. Sietzen, M. Lechner, J. Borowski, R. Hasani, and M. Waldner. Interactive analysis of cnn robustness. In *Computer Graphics Forum*, vol. 40, pp. 253–264. Wiley Online Library, 2021. 4, 7, 8

[67] D. J. Simons, Y. Shoda, and D. S. Lindsay. Constraints on generality (cog): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, 12(6):1123–1128, 2017. 9

[68] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[69] D. Smilkov, N. Thorat, C. Nicholson, E. Reif, F. B. Viégas, and M. Wattenberg. Embedding projector: Interactive visualization and interpretation of embeddings. *arXiv preprint arXiv:1611.05469*, 2016. 2

[70] F. Sperrle, M. El-Assady, G. Guo, R. Borgo, D. H. Chau, A. Endert, and D. Keim. A survey of human-centered evaluations in human-centered machine learning. In *Computer Graphics Forum*, vol. 40, pp. 543–568. Wiley Online Library, 2021. 1, 2, 7, 8, 9

[71] F. Sperrle, M. El-Assady, G. Guo, D. H. Chau, A. Endert, and D. Keim. Should we trust (x) ai? design dimensions for structured experimental evaluations. *arXiv preprint arXiv:2009.06433*, 2020. 1, 2

[72] H. Strobelt, S. Gehrmann, M. Behrisch, A. Perer, H. Pfister, and A. M. Rush. S eq 2s eq-v is: A visual debugging tool for sequence-to-sequence models. *IEEE transactions on visualization and computer graphics*, 25(1):353–363, 2018. 4, 5, 7, 8

[73] H. Strobelt, S. Gehrmann, H. Pfister, and A. M. Rush. Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. *IEEE transactions on visualization and computer graphics*, 24(1):667–676, 2017. 4, 5, 7, 8, 9

[74] A. Tyagi, C. Xie, and K. Mueller. Nas-navigator: Visual steering for explainable one-shot deep neural network synthesis. *IEEE Transactions on Visualization and Computer Graphics*, 1912. 4

[75] A. Tyagi, C. Xie, and K. Mueller. Nas-navigator: Visual steering for explainable one-shot deep neural network synthesis. *IEEE Transactions on Visualization & Computer Graphics*, (01):1–11, 2022. 7

[76] S. Van Den Elzen and J. J. Van Wijk. Baobabview: Interactive construction and analysis of decision trees. In *2011 IEEE conference on visual analytics science and technology (VAST)*, pp. 151–160. IEEE, 2011. 4, 6

[77] L. Von Rueden, S. Mayer, K. Beckh, B. Georgiev, S. Giesselbach, R. Heese, B. Kirsch, J. Pfrommer, A. Pick, R. Ramamurthy, et al. Informed machine learning–a taxonomy and survey of integrating knowledge into learning systems. *arXiv preprint arXiv:1903.12394*, 2019. 2

[78] J. Wang, L. Gou, H.-W. Shen, and H. Yang. Dqnviz: A visual analytics approach to understand deep q-networks. *IEEE transactions on visualization and computer graphics*, 25(1):288–298, 2018. 4, 6, 7

[79] J. Wang, L. Gou, W. Zhang, H. Yang, and H.-W. Shen. Deepvid: Deep visual interpretation and diagnosis for image classifiers via knowledge distillation. *IEEE transactions on visualization and computer graphics*, 25(6):2168–2180, 2019. 4

[80] J. Wang, W. Zhang, and H. Yang. Scanviz: Interpreting the symbol-concept association captured by deep neural networks through visual analytics. In *2020 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 51–60. IEEE, 2020. 4, 5, 6

[81] J. Wang, W. Zhang, H. Yang, C.-C. M. Yeh, and L. Wang. Visual analytics for rnn-based deep reinforcement learning. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):4141–4155, 2021. 4, 7, 8, 9

[82] X. Wang, W. Chen, J. Xia, Z. Wen, R. Zhu, and T. Schreck. Hetvis: A visual analysis approach for identifying data heterogeneity in horizontal federated learning. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):310–319, 2022. 4

[83] Z. J. Wang, R. Turko, O. Shaikh, H. Park, N. Das, F. Hohman, M. Kahng, and D. H. P. Chau. Cnn explainer: learning convolutional neural networks with interactive visualization. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1396–1406, 2020. 4, 5, 7, 8

[84] M. Wattenberg, F. Viégas, and I. Johnson. How to use t-sne effectively. *Distill*, 1(10):e2, 2016. 7

[85] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, 26(1):56–65, 2019. 4, 7, 8

[86] K. Wongsuphasawat, D. Smilkov, J. Wexler, J. Wilson, D. Mane, D. Fritz, D. Krishnan, F. B. Viégas, and M. Wattenberg. Visualizing dataflow

graphs of deep learning models in tensorflow. *IEEE transactions on visualization and computer graphics*, 24(1):1–12, 2017. 4

[87] T. Wu, M. T. Ribeiro, J. Heer, and D. S. Weld. Errudite: Scalable, reproducible, and testable error analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 747–763, 2019. 4, 8

[88] S. Xiang, X. Ye, J. Xia, J. Wu, Y. Chen, and S. Liu. Interactive correction of mislabeled training data. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 57–68. IEEE, 2019. 1, 4, 6, 8

[89] W. Yang, X. Ye, X. Zhang, L. Xiao, J. Xia, Z. Wang, J. Zhu, H. Pfister, and S. Liu. Diagnosing ensemble few-shot classifiers. *IEEE Transactions on Visualization and Computer Graphics*, 28(9):3292–3306, 2022. 4

[90] J. S. Yi, Y.-a. Kang, J. T. Stasko, and J. A. Jacko. Understanding and characterizing insights: how do people gain insights using information visualization? In *Proceedings of the 2008 Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization*, pp. 1–6, 2008. 7

[91] J. Yuan, C. Chen, W. Yang, M. Liu, J. Xia, and S. Liu. A survey of visual analytics techniques for machine learning. *Computational Visual Media*, 7(1):3–36, 2021. 1, 2

[92] X. Zhang, J. P. Ono, H. Song, L. Gou, K.-L. Ma, and L. Ren. Sliceteller: A data slice-driven approach for machine learning model validation. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):842–852, 2022. 4

[93] 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, 2018. 4, 8