Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm Performance

HONG SHEN, Carnegie Mellon University, USA
HAOJIAN JIN, Carnegie Mellon University, USA
ÁNGEL ALEXANDER CABRERA, Carnegie Mellon University, USA
ADAM PERER, Carnegie Mellon University, USA
HAIYI ZHU, Carnegie Mellon University, USA
JASON I. HONG, Carnegie Mellon University, USA

Ensuring effective public understanding of algorithmic decisions that are powered by machine learning techniques has become an urgent task with the increasing deployment of AI systems into our society. In this work, we present a concrete step toward this goal by redesigning confusion matrices for binary classification to support non-experts in understanding the performance of machine learning models. Through interviews (n=7) and a survey (n=102), we mapped out two major sets of challenges lay people have in understanding standard confusion matrices: the general terminologies and the matrix design. We further identified three sub-challenges regarding the matrix design, namely, confusion about the direction of reading the data, layered relations and quantities involved. We then conducted an online experiment with 483 participants to evaluate how effective a series of alternative representations target each of those challenges in the context of an algorithm for making recidivism predictions. We developed three levels of questions to evaluate users' objective understanding. We assessed the effectiveness of our alternatives for accuracy in answering those questions, completion time, and subjective understanding. Our results suggest that (1) only by contextualizing terminologies can we significantly improve users' understanding and (2) flow charts, which help point out the direction of reading the data, were most useful in improving objective understanding. Our findings set the stage for developing more intuitive and generally understandable representations of the performance of machine learning models.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**.

Additional Key Words and Phrases: Algorithmic decision-making; Machine Learning; Explainable AI; Human-centered AI; Confusion Matrix

ACM Reference Format:
Hong Shen, Haojian Jin, Ángel Alexander Cabrera, Adam Perer, Haiyi Zhu, and Jason I. Hong. 2020. Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm Performance. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 153 (October 2020), 22 pages. https://doi.org/10.1145/3415224

Authors' addresses: Hong Shen, hongs@cs.cmu.edu, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, 15213, USA; Haojian Jin, haojian@cs.cmu.edu, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, 15213, USA; Ángel Alexander Cabrera, cabrera@cmu.edu, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, 15213, USA; Adam Perer, adamperer@cmu.edu, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, 15213, USA; Haiyi Zhu, haiyi@cs.cmu.edu, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, 15213, USA; Jason I. Hong, jasonh@cs.cmu.edu, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, 15213, USA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2020 Note: This is a pre-camera-ready version, not for public distribution.
2573-0142/2020/10-ART153 $15.00
https://doi.org/10.1145/3415224

1 INTRODUCTION

Imagine you are a journalist who needs to write a story covering issues of differential treatment of two demographic groups by a recidivism prediction algorithm [4]. This task involves explaining the performance of a machine learning classifier on two demographic groups to facilitate informed public debate. Effectively communicating the machine performance to your readers, the vast majority of whom have little or no knowledge of computer science or machine learning, is not easy, as most representations of these results are designed for machine learning experts. Lay people often lack the technical background required to understand relatively complex concepts and terms (e.g., true positive, false negative) as well as the performance of the underlying machine learning models.

The need for explaining and presenting the performance of algorithmic decision-making systems powered by machine learning techniques to the non-expert public is a widely shared challenge across many domains. With the increasing deployment of algorithmic decision-making systems in many critical parts of our society, such as college admissions [42], loan decisions [55], and child maltreatment prediction [12], many social and ethical concerns are being raised (e.g., [4, 15, 17, 43]) and urgently need public evaluation and debates. Effective public understanding of algorithmic decisions – in particular, the performance of the underlying machine learning models – serves as one way to support such public debates.

To date, the majority of past work on explaining and visualizing AI systems has primarily focused on supporting *expert* decision-making ([13, 16, 30, 31, 46]). With the growing impacts of algorithmic decision-making in our society, a growing body of work in HCI has called for more efforts to explore how to better explain and present algorithmic decisions to multiple stakeholders, especially to lay people without technical knowledge [58]. On the other hand, although researchers in Explainable AI (XAI) have made progress in simplifying and explaining complicated machine learning models, this line of work was critiqued for lacking a deep understanding of or evaluation with actual users or stakeholders [39]. As a result, recent studies in HCI (e.g., [13, 56]) have started to adopt a human-centered approach to both create and evaluate interactive user interfaces in this specific domain.

Our work contributes to this emerging line of research. In this work, we explore how to create **non-expert-oriented representations of confusion matrices for binary classifications**. Confusion matrices are a widely used tool in machine learning to communicate the performance of a classifier on a set of training or test data where ground truth is known [2]. They help explain the fundamental performance metrics in binary classifications, including true positives, false positives, true negatives, and false negatives, which are often used as the basis of other evaluation metrics.

More recently, with the surging interest in explaining and visualizing AI, confusion matrices have also been used as one of the basic explanation components and core visual elements in many existing XAI toolkits and tutorials, including the ones explaining fairness metrics that have attracted increasing attention among ML and HCI researchers (e.g., [40, 48, 56]). For example, in a widely circulated tutorial, Narayanan used binary confusion matrices to explain a series of fairness metrics developed by the ML community [40]. Despite such wide applications, however, there exists very limited understanding in terms of whether confusion matrices are a useful and effective way to communicate ML performance to the non-expert public. Previous studies have also shown that lay people in general face great challenges in probabilistic reasoning (e.g., [22]), which might impact their ability in effectively interpreting confusion matrices. However, there lacks a systematic investigation in terms of whether lay people are able to interpret information contained in confusion matrices without additional help. In this paper, we set out to tackle this increasingly critical yet under-explored problem.

Our research questions are: RQ1: What are the main challenges non-expert lay people have in understanding the information contained in standard confusion matrices? RQ2: How effective different alternative representations perform in addressing those challenges, in terms of objective understanding, time costs and subjective understanding?

Our work explores how to create non-expert-oriented alternative representations of standard confusion matrices for binary classification, as a step towards facilitating public understanding of the performance of machine-learning-powered algorithmic decision-making systems. Towards this end, we first conducted interviews (n=7) and a survey (n=102) to identify the major challenges lay people have in understanding standard confusion matrices. Based on these initial studies, we conducted an online experiment where 483 participants used different alternative information representations of confusion matrices to understand the performance of an algorithm for making recidivism predictions. We developed three levels of questions to assess participants' understanding of the performance of the algorithm, and assessed these representations for accuracy in answering those questions, completion time, and understanding, which refers to self-reported perceived understandability. Our contributions are three-fold:

First, our work identified major challenges lay people have in understanding standard confusion matrices, a widely used method in machine learning to present the performance of a classifier on a set of test data where ground truth is known.
Second, we designed a few alternative representations to tackle those challenges: Contextualized confusion matrices that map all the terminologies (e.g., false positive and false negative) back in specific problem domains, Tree diagrams, Bar charts, and Flow charts.
Third, we evaluated our designs for their effectiveness. Our evaluations suggested several design implications that might serve the foundation for further exploration in this space: (1) contextualizing the terminologies can significantly improve users' objective and subjective understanding; (2) flow charts, which help point out the direction of reading the data, were most useful in improving objective understanding.

2 RELATED WORK

In this section, we outline relevant past work in two areas. First, we survey the related work in explainable AI and visualizing AI, and describe how our work is positioned in this space. Next, we present an overview of existing work on human perceptions of algorithmic decisions, a quickly growing field that has received increasing attention in CSCW and describe how our work contributes to this emerging line of research.

2.1 Explaining and Visualizing AI

While interpreting intelligent systems has a long history in AI and HCI, the recent widespread deployment of machine learning enabled algorithmic systems in many critical and complex social environments has lead to a renewed attention to interpretability. Generally labeled as explainable AI (XAI), this growing body of work aims at providing a human-understandable explanation for decisions made by black box machine learning models in order to support decision-making, improve transparency, and increase trust (see review articles [1, 10, 57]).

A large body of past work has explored different explanation styles to probe the inner working of such black box models. For example, Datta et al. [16] offered a solution to help users measure the influence of a series of features in inputs on system outputs. Nugent et al. [41] selected similar cases from the training data to offer explanation to the decision in question. Lakkaraju et al. [32] aimed at generating short rule-based explanation. Despite making greater progress, this line of work was critiqued for lacking a deep understanding of or evaluation with actual users or stakeholders [36,

Meanwhile, although previous non-expert-oriented visualizations in (e.g., [28, 38]) used simple diagrams to target speci c probability problems such as Bayesian Reasoning, the recent surge of XAI has led to more complex visual techniques, systems, and toolkits to facilitate expertusers' evaluation of model performance (e.g. see [2, 21]). For example, Google's People + AI Research group (PAIR) released the open-source What-if tool [52] to help people who are not formally trained in machine learning visualize the e ects of bias metrics. Prospector [30] enabled data scientists to understand how a given feature in uences algorithmic prediction and support for manipulating feature values for response. Similarly, studies have also been conducted to create alternative visualizations to confusion matrices, to help machine learning practitioners as well as non-ML-experts adjust their models. For example, Squares [46] introduced a novel visual system to help machine learning practitioners evaluate the performance of a multi-class classi er. Classee [7] o ered a simpli ed barchart to help end-users with no expertise in machine learning choose and tune parameters and understand classi cation errors. While these systems provided valuable methods to explore machine learning data and model, they are designed primarily for expertusers with relatively advanced technical literacy including both technical experts like machine learning experts and data scientists as well as domain experts like doctors.

Recently, researchers in HCI seek to expand this line of research by taking a human-centered design approach, with an emphasis on both improving usability of these explainable tools and performing empirical tests for evaluation. For example, Cheng et al. [18] conducted non-expert-oriented experiments to test the performance and trade-o s among interactive vs. static explanation as well as white-box vs. black-block explanations. Their ndings suggested that while both interactive and white-box explanations can improve non-expert users' comprehension, interactive approach is more time consuming.

Our work contributes to this emerging line of research in HCI by focusing on confusion matrices, a basic and widely used representation in the evaluation of algorithm performance. It helps set the stage for understanding and developing laypeople-oriented representations of machine learning model performance.

2.2 Human Perceptions towards Algorithmic Decisions

With the increasing penetration of algorithmic decision-making systems in our society, there is a widely shared concern that ML-enabled algorithmic systems may produce socially and ethically questionable decisions that are not aligned with human values (e.g. see [4, 6, 18, 19]).

In response to these concerns, an emerging line of work in the HCI community has started to look at those decisions from a human-centered perspective. This body of research has used interviews, surveys, and experiments to empirically probe people's perceptions towards algorithmic decisions. Examples include studies on how humans perceive algorithmic decisions versus human decisions in managerial contexts [34], whether people perceive certain features (such as criminal history or neighborhood safety) as fair to be used to predict criminal risk [35, 50], how explanation styles might matter in shaping people's justice perceptions [9], how members of traditionally marginalized communities feel about algorithm (un)fairness [54], how a ected communities feel about algorithmic decisions in the context of a child welfare system [12], how the general public perceives online behavioral advertising that used demographic factors (e.g., race) as targeting variables [44], which statistical de nitions of fairness people perceive to be the fairest in the context of loan decisions [47], as well as how humans use AI systems to make decisions [23, 24]. This past body of work mostly used storyboards or text to present several algorithmic scenarios to their study participants, often without tackling the results and performance of the underlying machine learning models. This is understandable, as lay people often lack the required technical knowledge to fully understand and evaluate highly specialized algorithmic results.

As a result, researchers in HCI are calling for more efforts to explore how to better explain and present algorithmic decisions to multiple stakeholders, especially to non-experts with limited technical literacy [58]. Recent studies (e.g., [13, 35, 50, 56]) have started to use visualizations or create user interfaces to communicate algorithmic decisions to their study participants.

Our research contributes to this line of work by offering a new and complementary angle, looking at making the performance of machine learning model more comprehensible. By creating alternative representations of standard confusion matrices, we aim at developing more easily understandable methods to represent the performance of machine learning models for the non-expert public. Our work has the potential to be used to better solicit public perception of algorithmic decisions (e.g., help lay users evaluate fairness) for future research in this domain.

3 OVERVIEW OF THE STUDIES

We conducted a series of studies, both qualitative and quantitative, to explore how to better help non-expert users interpret a very common output of machine learning classifiers: confusion matrices. We chose binary confusion matrices as a starting point to improve model literacy for the non-expert public for several reasons. First, they are commonly used representations of the performance of machine learning models. Second, they work for all kinds of classifiers regardless of the underlying algorithm and so have general applicability. Third, they help convey some important aspects of how well a machine learning model is (or is not) working. And fourth, they help explain the fundamental performance metrics in binary classifications (i.e., true positives, false positives, true negatives, and false negatives), which can be further used as the basis of other evaluation metrics.

Our team for this project was composed of researchers with diverse backgrounds, including machine learning, software engineering, visualization, and social science. We began with a session for rapid ideation to identify our task domain.

3.1 Choice of Task Domain and Dataset

3.1.1 Task Domain. Recidivism prediction is an important public issue where algorithms have been increasingly deployed to help judges assess the risk of re-offending among defendants. This topic has also generated one of the most controversial cases so far in the debates around fairness and public algorithmic decision-making [4]. Instead of looking at issues around fairness, in this work, we used it as our task domain to test and facilitate public understanding of algorithmic outputs and performance. Here, we chose to focus on a specific domain, instead of abstract algorithmic results, because past studies in psychology (e.g., Wason's selection task [54]) suggest people tend to perform better with their reasoning powers on realistic cases than with abstract formulations.

3.1.2 Dataset. We used the dataset gathered by ProPublica in its 2016 report, which examined the performance of COMPAS, a widely used recidivism algorithm in the US. This dataset contains two years worth of COMPAS scores from Broward County, Florida in 2013 and 2014 [3]. COMPAS assigns risk scores to criminals to determine their likelihood of re-offending, with 1-4 constituting "low risk", 5-7 "medium risk", and 8-10 "high risk". Following the study by ProPublica [3], which used 4 as the threshold, we similarly present this as a binary classification task with score at and below 4 as negative prediction ("Labeled as low risk"), and above 4 as positive prediction ("Labeled as high risk"). Ground-truth labels "Re-offended" and "Didn't re-offend" correspond to whether a defendant released on bail was arrested for another crime within 2 years of their release.

3.2 Study 1

To understand what challenges non-expert lay people have in understanding standard confusion matrices, we conducted a qualitative study comprised of interviews (n=7) and an Amazon MTurk

survey (n=102). In both studies, we asked our study participants to answer a series of factual questions about information contained in confusion matrices as well as what were the challenges they have in reading the table and answering those questions. We identi ed two sets of main challenges in Study 1.

3.3 Study 2

Building on the challenges identi ed in Study 1, in Study 2 we developed and evaluated alternative representations of confusion matrices. We conducted a between-subjects online experiment (n=483) on Amazon Mturk to assess the relative e ectiveness of our alternatives in addressing those challenges,in terms of comprehension, time costs and subjective preference.

3.4 Evaluation Framework of Objective Understanding

Over the years, the machine learning community has developed di erent error metrics to assess the performance of classi cation algorithms (e.g., see [27]). However, these metrics are designed primarily for expert users.

In this work, we developed three levels of questions comprehension, comparison and simulation to assess and facilitate non-experts' understanding of the statistical results and performance of machine learning classi ers. Our evaluation questions were developed and adapted from previous frameworks in learning science [20], which identi ed three levels of graph understanding of statistical results for lay people: Level 1 focuses on extracting or locating information; Level 2 focuses on nding relationships; and Level 3 focuses on analyzing implicit relations contained in the graph. In total, we created 11 questions, including 4 comprehension questions, 4 comparison questions and 3 simulation questions.

Note that the framework developed in this study o ers a general guideline and all the questions can be adapted to di erent domains and problem statements in future research (see Table 1).

3.4.1 estion Level 1: Comprehension In this level, we created questions to measure how accurately users can identify and locate certain information in the information representation.

Questions we created and used in this level include:

> In Group A, how many defendants were labeled high risk and re-o ended?
> In Group A, how many defendants were labeled high risk but did not re-o end?
> In Group B, how many defendants were labeled low risk but re-o ended?
> In Group B, how many defendants were labeled low risk and did not re-o end?

3.4.2 estion Level 2: Comparison In this level, we measured how accurately users can compare information and performance of the machine learning classi er on two demographic groups, as well as how well they can perform simple analysis. Questions in this level also served as a bridge to lead users to the next level.

Questions we created and used in this level include:

> Between Group A and B, which one has more defendants who were labeled as high risk and re-o ended?
> Between Group A and B, which one has more defendants who were labeled as low risk but re-o ended?
> In Group A, among people who did not re-o end, what percentage were labeled as high risk?
> In Group A, among people who were labeled as high risk, what percentage did not re-o end?

Level	Purpose	Question type (X may equal to Y)	Examples
1	Comprehension	In Group A, how many people were classi ed as X but/and belonged to Y?	In Group A, how many defendants were labeled high risk but did not re-o end?
2	Comparison	Between Group A and B, which one has more people who were classi ed as X but/and belonged to Y?	Between Group A and B, which one has more defendants who were labeled as high risk but did not re-o end?
		In Group A/B, among people who belonged to X, what percentage were classi ed as Y?	In Group A, among people who did not re-o end, what percentage were labeled as high risk?
		In Group A/B, among people who were classi ed as Y, what percentage belonged to X?	In Group A, among people who were labeled as high risk, what percentage didn't re-o end?
3	Simulation	For a new defendant in Group A/B, who we already know belongs to X, how likely will he be labeled as Y?	For a new defendant in Group A, who we already know won't re-o end, how likely will he be labeled as high risk?
		For a new defendant, who we already know belongs to X, in which group he/she is more likely to be classi ed as Y?	For a new defendant, who we already know won't re-o end, in which group he/she is more likely to be labeled as high risk?

Table 1. Our evaluation framework used three levels of questions to assess non-experts' objective understanding of the statistical outputs of binary classification. Note that the framework o ers a general guideline and all the questions can be adapted for di erent domains and problem statements in future research.

3.4.3 estion Level 3: Simulation. In this level, we wanted to measure how well users can perform simulations based on the presented information, e.g., can identify implicit information contained in the information representation.

Questions we used in this level include:

For a new defendant in Group A, who we already know will not re-o end, how likely will he/she be labeled as high risk?

For a new defendant in Group B, who we already know will not re-o end, how likely will he/she be labeled as high risk?

For a new defendant, who we already know will not re-o end, in which group he/she is more likely to be labeled as high risk?

To ensure that our participants could interpret and understand the questions above, we tested all questions using cognitive interviews [45, 53]. We asked our participants to think aloud as they answered the questions, and we identi ed and addressed any concerns and questions that emerged from the process. We conducted cognitive interviews with four volunteer participants who had not seen or heard of confusion matrices before. Two main changes were made to modify the questions

based on their feedback: (1) we adjusted the wording of the level 1 and 2 questions by adding conjunctions like and and but ; and (2) we also adjusted the order of the clauses (e.g., who we already know won't re-o end) for level 3 questions to better facilitate understanding. We iteratively re ned our questions based on participants' feedback until no new concerns or questions were identi ed.

4 STUDY 1: AN EXPLORATORY INVESTIGATION OF THE CHALLENGES LAY PEOPLE HAVING IN UNDERSTANDING STANDARD CONFUSION MATRICES

In Study 1, we began by interviewing non-expert lay people to understand what challenges they might have in reading standard confusion matrices with original terminologies, which was followed by a survey. In both interviews and the survey, we used the three levels of questions developed for measuring objective understanding (see section 3.4) to test their comprehension and then asked open-ended questions to probe in greater detail about the challenges.

4.1 Method

4.1.1 Interview (n=7). We rst conducted semi-structured interviews with seven participants (three males and four females) to learn about the challenges they are having. We recruited our participants in a university located around a northeast US metropolitan area from non-technical majors. None of them had seen or heard of a confusion matrix before. During our interviews, we showed them a standard confusion matrix of the ProPublica data, along with factual questions regarding the information contained in the confusion matrix. We asked them to think aloud as they worked through the process. We then probed in detail what made them struggle in answering the questions. Each interview lasted for 20-30 minutes and participants received $10 for time compensation. All interviews were audio-recorded and transcribed.

4.1.2 Survey (n=102). In addition to qualitative interviews, we also conducted a quantitative survey (n=102) on Amazon Mechanical Turk in August, 2019. The survey consisted of two sets of questions. The rst part included ll-in-the-blank and single choice questions. We asked factual questions about information contained in confusion matrices, using the three levels of questions developed for measuring objective understanding (see section 3.4). The second part included two open-ended questions and probed in greater detail about users' challenges in understanding. The average time for completing the survey was 7 minutes and each participant received $1.

4.1.3 Data analysis. We applied thematic analysis [11] to all interview data as well as all the open-ended questions in our survey. Two researchers read through the data set individually and held weekly discussions. The aim of the coding is to nd out comprehension challenges that the study participants had. With this idea in mind, the two researchers coded the data set independently, looking for as many basic codes as possible. After this step, they met again, discussing and re ning their initial codes and sorted the di erent codes into themes. Through iterative discussion and deductive and inductive thinking, they further reviewed, de ned and named themes until a satisfactory thematic map has been reached.

4.2 Results

Below we discuss our ndings from our interviews and survey. Interviewees are identi ed with a P and survey participants are identi ed with a S . Survey responses are also accompanied with percentages.

In general, users were frustrated with the information presented using standard confusion matrices in original terminologies. In some cases, this led to negative feelings toward the decision-making system as well as the people working behind it. For example, S35 commented: I think

Fig. 1. The standard confusion matrix shown to participants in Study 1.

that the individual who chose to present the data this way should be red. If someone working for me presented this in a meeting, I would let them go.

They also felt incompetent and dumb. S36 commented, "Can you dumb it down for us regular folks? Otherwise, those of us who are not so bright will not be able to pair the information appropriately."

4.2.1 Challenge 1: Confusion about terminology. The first set of challenges that emerged from our data is that users in general have difficulty interpreting the terminology used in confusion matrices (true positive, true negative, false positive, and false negative). They did not know what positive and negative meant and had a hard time mapping it to the original problem. This was a widely shared concern in both our interviews (7/7) and survey (98%).

P3 explained: "What does positive mean? If this is from my doctor's office, I know there might be something wrong with my test. But in this (recidivism predication) case, does this mean `high risk' or `re-offended'?

S2 felt the repeated use of positive and negative in the matrix increased the difficulty of the task, i.e., that the term "positive" and "negative" can refer to both predicted and actual classes: "Positive and negative make it too sciency. Also, what does true mean? Does it mean positive positive?"

S14 suggested replacing the terminologies and labeling the matrix appropriately according to the context: "Label the tables better with terms match what is asked in the questions. I had to guess to start which axis X and Y was represented in the question about the tables."

4.2.2 Challenge 2: Confusion about the underlying relationships between the four types of predictions. Apart from terminology issues, users also reported that they need help in clarifying the underlying structure of the confusion matrix. In general, they felt that the matrix design does not convey the underlying relationships between the four types of predictions. Three sub-challenges were further identified under challenge 2.

Sub-challenge 2.1: Confusion about the direction of reading the data. Users were confused about how the data flows from one category into other categories, i.e., direction of reading the data in confusion matrices, in particular, about "False Positive" and "False Negative". They made mistakes

in answering questions like In Group A, how many people were classi ed as X but belonged to Y? Four of seven interview participants and 54% of survey participants reported this challenge.

For example, P1 said: You are asking me how many re-o ending defendants were labeled as high risk, there is a `directional relationship' here like, re-o ending defendants were thrown into another group. But from this table, it is really hard to tell. There is no clear clue in terms of which direction the data go.

S55 asked: Can you better explain how the di erent groups went into the other groups? Like `in Group A, how many defendants were labeled high risk but did not re-o end?' I think understand that `positive' means 'labeled high risk,' but I'm confused after that.

P2 suggested to add direction metaphor to the representation: Maybe you can identify which one (class) is the `starting point' and which one (class) is the `end point'? So it can be read as a map I will have a better idea of how the data travels ...

Sub-challenge 2.2: Confusion about layered relations. Some users are confused about the layered relations embedded in confusion matrices. In particular, when answering questions like In Group A, how many people were classi ed as X and belonged to X, they tend to mistake Positive with True Positive (or Negative with True Negative). Four out of seven interview participants and 45% of our survey participants reported this challenge.

When asked why this is di cult to answer, P4 explained: There are too many layers in the table! You see, the table has a top and bottom row and when you ask `how many people were labeled as high risk and re-o ended?' I just added everything up here.

S13 suggested: Can you break down the graph into subsections so it can only have one set of headings right now there are too many things messed up here.

S27 said: While I understand the terminologies, I found the true vs. false in addition to positive and negative to be di cult to understand.

S29 reported: I just needed better organization. This setup with triple layers on each side wasn't easy to read.

Sub-challenge 2.3: Confusion about quantities involved. Finally, a small group of users (one out of seven of interviewees and 9.8% of survey participants) reported that they need support for comparing the quantities involved in the confusion matrix, especially when they needed to compare the performance of two classi ers. Even if they could locate the information correctly, it was hard for them to compare.

P7 said: I think everything is OK until you asked me to compare in which group a new defendant who won't re-o end is more likely to be labeled as high risk. You see, this involved two steps I need to gure out, rst, how likely a new defendant who won't re-o end is going to be labeled as high risk in Group A and Group B, and second, do comparison. It would be much easier if you can make everything proportionally so I don't have do the rough calculation I can just tell directly from the graph.

S6 suggested: It was di cult to tell percentages at a glance. If the di erent percentages could be represented with visual size di erences, it might be easier to compare groups at a glance.

4.3 Summary

We identi ed two major sets of challenges non-expert lay people have in understanding standard confusion matrices. First, we found that the majority of our participants had di culties understanding the terminology and thus had a hard time mapping them back to the scenario. Second, they also reported structural issues, i.e., that the matrix design does not convey the underlying relationships between the four types of predictions. We then further identi ed three sub-challenges under this structural issue: (1) some of our participants reported problems in understanding how the data ows into other categories, i.e., the direction of reading the data; (2) others were confused

about the multiple layers contained in the table and suggested us to unpack these layers; and (3) a few of them suggested to represent the data proportionally to facilitate easy comparisons.

5 STUDY 2: AN EXPERIMENT OF USERS' PERFORMANCE OF ALTERNATIVE REPRESENTATIONS

Building on the ndings of our rst study, we conducted a design workshop to develop alternative representations of confusion matrices that could address the challenges raised by our participants. We then conducted a crowd-sourced experiment on Mturk to test the e ectiveness of our alternatives in helping them understand the performance of a machine learning algorithm for making recidivism predictions.

5.1 Design Workshop

We invited one machine learning expert, one software engineer, one former journalist, two visualization experts, two UX designers (current graduate students at our university), and one psychologist to join the workshop. We introduced the goal of facilitating public understanding of the performance of machine learning classi ers using alternative representations of confusion matrices. We shared the challenges we identi ed from Study 1 and asked our participants to come up with multiple alternatives that could address those challenges. To avoid potential interaction e ects among di erent designs, in the workshop we asked participants to think about making each representation targeting one speci c challenge.

During our initial ideation and prototyping phase, we drew on previous studies that used simple non-expert-oriented diagrams to target probability problems like Bayesian reasoning (e.g., [28, 38]) and developed 20 initial alternatives. Participants were also encouraged to develop visual representations to tackle those challenges, minimizing the use of text descriptions.

We performed three steps to narrow down these initial prototypes. First, we discussed and assessed the initial alternatives for their feasibility in targeting the speci c structural challenges identi ed in Study 1, usability for our targeted audience, and generalizability. Second, we clustered similar ideas, identi ed common themes, and combined di erent options. Third, over the course of our design process, we also iteratively conducted informal user testing with 12 participants. In these tests, we showed our mock-ups and assessed them for understandability (e.g., we asked questions like Tell us what this table shows?), how well they helped the participants in answering a set of questions (e.g., How many people were correctly labeled as `high risk? '), as well as subjective preference.

At the end of the process, we settled on two sets of solutions. The rst set addresses the terminology issue by replacing general terminology with (a) contextualized confusion matrix that uses terminology that maps back to the speci c case in question, e.g., Reo ended and Labeled high risk instead of Positive . We opted for this design since confusion over terminology was a common challenge identi ed in Study 1.

In addition to changing terminology, we also developed a second set of representations, each of which addresses a speci c sub-challenge identi ed in Study 1. These representations include a (b) a tree diagram to unpack layered relations, (c) a ow chart to point out the direction of reading the data, and (d) a bar chart to clarify the quantities involved.

We made two additional choices in terms of visual encoding. First, in the design of the tree diagram and ow chat, we explored how the matrix can be read from prediction to true label. We followed the best practice in data visualization design to use lines (and arrows) to connect disparate regions to help readers follow the visualization, based on the Gestalt principle of connectedness [36]. Second, when choosing the color used in the design, we followed best practices in information visualization to use di erent color hues to represent di erent categorizations [26]. In particular, we