The Machine Learning Model as a Guide: Pointing Users to Interesting Instances for Labeling through Visual Cues

Benedikt Grimmeisen
Aalen University of Applied Sciences
73430 Aalen, Germany
https://orcid.org/0000-0001-6853-2087

Andreas Theissler
Aalen University of Applied Sciences
73430 Aalen, Germany
https://orcid.org/0000-0003-0746-0424

ABSTRACT

The labeling of datasets is an important task for supervised and semi-supervised machine learning that can be addressed with visual analytics. With model-based active learning and user-based interactive labeling, there are two complementary strategies for this task. We present an approach that combines the strengths of both areas and aims to guide users through model-based recommendations and highlighting in one interface when selecting and labeling instances. For this purpose, an active learning strategy is used to recommend useful instances in addition to the user-based selection of instances. We have implemented the approach and conducted a user survey to research the effects guidance by visual cues has on the users' selection strategies. The proposed approach combines both perspectives in a single interactive visualization to support the user with different degrees of guidance in the selection of instances. Our results of the user survey suggest that user guidance has a positive influence on the users' perceived confidence and difficulty in selecting instances, on their orientation, and on their perceived impression of the models' performance. A video of the approach is available: https://youtu.be/TYPWG85Akn0.

CCS CONCEPTS

• **Human-centered computing → Visualization systems and tools**; • **Computing methodologies → Active learning settings**; • **Applied computing → Annotation**.

KEYWORDS

Visual-Interactive Labeling, Active Learning, Visual Analytics, Machine Learning, Classification

1 INTRODUCTION

The machine and deep learning approaches, which are currently perceived as successful [1, 30, 32], are in most cases based on supervised learning. To train such models, large amounts of labeled data are required. While the acquisition of the data in today's information age through connected sensors and Big Data is often not a problem, the provision of the corresponding labels is a major challenge. This often prevents the application of ML in those productive environments where labels are not available. To label the data, experts' knowledge is needed. The labeling process is therefore associated with a considerable amount of time and cost, as the experts have to carefully examine each instance.

Active learning (AL) [24] is one approach to reduce the number of instances to be labeled by users. Using various query strategies, the model queries the user for labels of instances that are informative for the model. While in AL, users are constrained to label suggested instances, a complementary approach is to use interactive visualizations. In those approaches, data points can be freely labeled by the user leveraging methods of Visual Analytics. Bernard et al. [3] compared AL and user strategies and derived 10 strategies, which users followed to select instances for labeling. In [3], the authors showed that pure user-driven selection of instances can lead to a biased and sub-optimal selection. With visual-interactive labeling (VIAL) Bernard et al. [5] proposed an approach combining the complementary strengths of user-based selection and model-based suggestions to label the best possible candidates.

Our work bases on the aforementioned research, in particular on [5]. We focus on answering the two research questions:

- **RQ$_1$**: How can users be guided in the selection of instances to be labeled to point them towards informative instances?
- **RQ$_2$**: How does guiding users in the selection process affect their selection and labeling strategy?

We present an approach that calculates the utility of each instance for training a classifier and presents it together with recommendations of particularly useful instances in an interactive user interface. The user is thereby guided by the model by the means of visual cues, however, for the selection of instances and labeling the users can freely decide where to follow the guidance and where to make their own decisions. Iteratively, a machine learning (ML) model is trained on the subset of user-labeled data points and enhances the labels by classifying the unlabeled data points (label propagation). The user may decide to refine the labeling or to confirm the manual labels together with the label propagation conducted by the ML model.

This work makes the following contributions:

(1) Adaptation of a model- and data-based active learning strategy to calculate the utility of unlabeled instances for training a classifier and to recommend instances to the user.
(2) Concept and design for guiding users in the selection of instances for the training process of a classifier through visual cues in a visual-interactive interface.
(3) Comparative analysis of the influence of user guidance on the users' selection of instances with a user survey.

2 RELATED WORK

Related work originates from different research areas dealing with interactive labeling of instances: active learning, interactive visualization, and the combination of both. Furthermore, there is related work focusing on user guidance.

In active learning (AL) [24, 25, 28] an iterative training process is used. The active learner queries a human oracle for labels of informative data points. Informative instances are those that will most likely improve the model's accuracy. By selecting the most effective instances for training, the aim is to minimize the number of user interactions. Strategies for the selection of instances can be grouped into different categories depending on the metrics and sources [3, 4, 13, 31]. In general, these strategies can either use information from the model or the data. Recently, approaches were presented that combine model and data-based strategies to avoid redundancies and outliers in the selection process [13] by considering correlations between features or labels. He et al. [14] introduced an approach that not only takes uncertainty and representativeness into account but also the diversity of the data. While AL approaches may produce good models in a fast manner, according to [8] users often feel frustrated having to answer long sequences of queries.

While in AL the model controls the workflow, visualization methods can be used to let the user freely select and label potentially informative instances. Visual Analytics enhances the cognitive abilities of humans by methodically preparing and presenting data in interfaces with interactive visualization [12]. Thereby, people can gain deeper insights into heterogeneous and complex data through explorative knowledge discovery [12]. Like the querying strategies in AL, users also follow certain strategies when selecting interesting instances. Bernard et al. [3] compared AL and user strategies in a study. Based on the observations of the subjects, 10 strategies were identified, which users followed to select instances for labeling. In subsequent work [4], formal building blocks were derived from these strategies, systematically describing recurring thinking patterns and concepts. The resulting user strategies can be categorized into data and model-based strategies based on the information used. In the data-based approaches, only information from the data set, the instances, and their projections are used for the decision. For the model-based strategies – in addition to the data-based information – the class labels predicted by a model are also considered.

It has been experimentally shown [3] that purely user-driven selection and labeling of instances can lead to a biased and sub-optimally labeled data set. While leaving the user in control, supportive guidance appears to be beneficial. Ceneda et al. [9] describe guidance in terms of visual analytics as "[…] a computer-assisted process that aims to actively resolve a knowledge gap encountered by users during an interactive visual analytics session". Guidance aims at supporting the user in fulfilling a task through a dynamic process. In their guidance model, Ceneda et al. [9] characterize user guidance through a knowledge gap, input and output, and guidance degree. Three guidance degrees are distinguished: (1) *orienting* which is the lowest degree and supports users to build a mental map with hints on possible targets or paths towards the solution. For this purpose, visual cues in visualizations are often used. (2) *directing* which emphasizes a possible course towards a solution. Alternative options that lead to the desired result are offered. Users can accept or ignore these suggestions. (3) *prescribing* which is the highest degree of guidance. Decisions on the next steps are made in an automated process leading towards the desired result.

Bernard et al. [5] proposed visual interactive labeling (VIAL), a unified process for the interactive labeling of instances with the user in the loop. VIAL aims to combine the mutual strengths of AL and interactive visualizations. In the iterative approach, there are two complementary alternatives for identifying instances for labeling: interactive visualization and candidate suggestion. Whereas the users are queried by the model with automated suggestions in the AL perspective, they take a more active role in the interactive visualization perspective by exploring and selecting instances. The VIAL process proposes to either include AL-based guidance concepts included in visual interfaces, or visual-interactive interfaces for the analysis and steering of AL strategies [5]. Kucher et al. [17] implemented VIAL for the classification of social media text using an AL approach for model-based selection of instances. Users can either label the suggested instance or ignore it. Ritter et al. [21] introduced a visual-interactive approach for the classification of music genres. In addition to suggesting instances determined by an AL strategy, users can select their own instances in a visualization. In [27], Theissler et al. discuss the challenges of visual interactive labeling for anomaly detection. In [10] mVis is proposed where interactive visualizations are used for the selection of instances. These visualizations are enhanced with the predicted or user-assigned labels. To guide the user in the selection, different ML methods like clustering, classification, and AL are used, which the user can parameterize. In [2], Beil et al. propose an approach for interactive labeling of instances combining density-based clustering with anomaly detection. Identified groups of similar instances are iteratively cleaned from outliers and afterwards labeled as batches.

To the best of our knowledge, no approach has been proposed that implements VIAL in such a way that both, the user-driven selection of instances and the model-based suggestions, are unified in one visualization with visual cues. Such an approach would, however, be beneficial since users would not have to decide on one of the two degrees of guidance *orienting* and *prescribing* but rather have the combined spectrum with *directing* at their disposal.

3 APPROACH: VIAL WITH USER GUIDANCE

We present a hybrid approach that combines user- and model-based selection for interactive instance labeling (see Fig. 1). Starting with unlabeled data, users select and label instances. The user is guided from the very beginning of the process by suggesting useful instances and showing them with visual cues. By calculating suggestions in the first iteration using the representativeness of
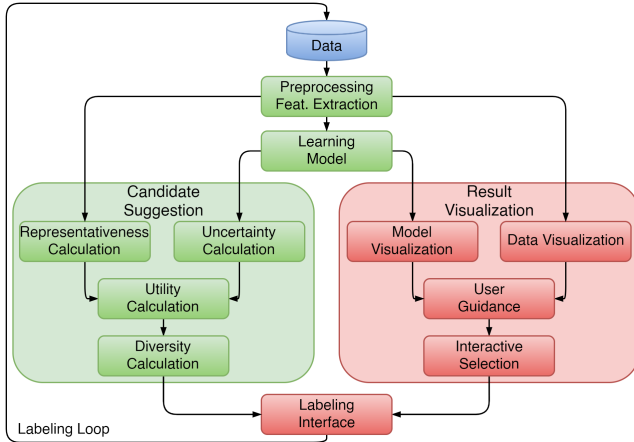
Figure 1: Overview of the proposed approach based on the VIAL process [5]. Our iterative approach includes algorithmic models (green) and visual interfaces (red). Both of the two complementary main components, *Candidate Suggestion* and *Result Visualization*, consist of individual building blocks for the calculation and presentation of information.

the unlabeled instances instead of an untrained model solely, we overcome the cold start problem known from Active Learning. An interplay between the user and a classifier takes place intending to collaboratively label the entire data set with a low number of user interactions. Therefore, the classifier is iteratively trained on the current set of manually labeled data points to propagate the labels by tentatively assigning predictions to unlabeled data points. The approach implements the VIAL process [5] with the two complementary steps *Candidate Suggestion*, and *Result Visualization* to identify candidates to be labeled. The iterative workflow consists of three steps. In the first step, similar to Norman's *Stages of Action* [20], the user evaluates the data and the state of the model by comparing it with his mental model and initial goals. The user is provided with the calculated utility of each instance and with recommended instances. Especially at the beginning of the process, when the user has assigned only a few or no labels and therefore the model is insufficiently trained, the calculated utility and the suggestions are essential for the user to identify influential and interesting instances. Following that, in the second step, the user assigns the corresponding labels to the selected instances. In the final step, the labeled instances are added to the training set and the model is retrained. Simultaneously, the model calculates new recommendations. This process is iteratively repeated, starting with the first step displaying the updated model and the new recommendations. The user repeats this procedure until satisfied. Then the model's predicted classes are accepted as labels for the remaining instances that have not been manually labeled.

The approach consists of the two main components *Candidate Suggestion* and *Result Visualization*. While the first is model-based, the latter is an interactive visualization showing the data, the model state, and the visual cues to guide the user.

3.1 Candidate Suggestion

The Candidate Suggestion calculates the potential benefit of each unlabeled instance and, based on this, automatically selects meaningful instances that are suggested to the user. We denote the estimated benefit of an instance as *utility*, inspired by prior research [13]. To calculate the utility we use model- and data-based information, supported by Fu et al. [13] and He et al. [14] who consider not only the uncertainty of the model but also the representativeness of the instance when calculating the utility. The use of representativeness compensates for the fact that the uncertainty of the model considers the instances isolated from each other. Thus, potential outliers that are not representative for the data set might get recommended. The representativeness prevents this by incorporating similarities between the data points. We define the utility of an unlabeled instance x_i, inspired by [13, 14], as:

$$\text{Utility}(x_i) := \text{Uncertainty}(x_i) \times \text{Representativeness}(x_i)^{\beta} \quad (1)$$

where β controls the trade-off. For each instance, the weighted product of the model uncertainty and the similarity to all other instances in the data set is calculated.

3.1.1 Uncertainty Calculation. We use a probabilistic model for the calculation of the uncertainty, which represents the predicted result of a classifier as a vector containing the posterior probability of the classes learned by the model for each instance. Following the three user strategies from [4] (class borders refinement, class outlier labeling, and class intersection minimization), we use the margin approach to minimize the error rate in classification. There is criticism that this approach only considers the two most probable classes and neglects additional information. However, similar to user strategies [4], our user-based approach seeks to minimize classification errors by distinguishing clearly between specific classes – the key feature of the margin approach. The originally proposed metric [13] is adopted in the following way to obtain the difference between the posterior probabilities of the two most probable classes for each unlabeled instance x_i:

$$f_{\text{margin}}(x_i) = P_{\Theta}(\hat{y}_1|x_i) - P_{\Theta}(\hat{y}_2|x_i) \quad (2)$$

where $\hat{y}_1$ and $\hat{y}_2$ are the first and second probable class labels. With the traditional least margin approach, the model tries to separate the two classes with the highest probability as distinct from each other as possible, which is indicated by the largest possible margin between the probabilities. Instances with a small margin are therefore particularly informative for the model since in these cases the model is unsure. The complement of the margin (eq. (2)) is used as the measure of uncertainty:

$$\text{Uncertainty}(x_i) := f_{\text{uncertainty}}(x_i) = 1 - f_{margin}(x_i) \quad (3)$$

By this conversion, a small margin expresses a high degree of uncertainty, since in that case confusion can easily occur between the first two classes.

To determine the uncertainty, a trained model is necessary. However, at the beginning of the labeling process, no labeled instances are available to train a model. Therefore, we initially assign the highest possible uncertainty of 1 to each instance. Hence, an initial utility can be calculated, and instances can be recommended.

3.1.2 Representativeness Calculation.
To calculate how representative an instance x_i is, our approach uses the information density I. The higher the information density of an instance, the more similar the instance is to the rest of the data, expressed by:

$$\text{Representativeness}(x_i) := I(x_i) = \frac{1}{U} \sum_{u=1}^{U} sim(x_i, x_u) \qquad (4)$$

where U denotes the size of the unlabeled dataset. There are several ways to calculate the similarity. We use the Euclidean similarity, a reformulation of the Euclidean distance:

$$sim_{\text{eucl}}(x_i, x_u) = \frac{1}{1 + \sqrt{\sum_{j=1}^{n} (x_{i_j} - x_{u_j})^2}} \qquad (5)$$

Calculating the information density using Euclidean similarity prefers the centers of clusters. We assume that these clusters correspond to actual classes in the data with the identified centers being the most representative instances of these classes. Based on the data-based user strategy ideal labels first [4], which focuses on the most characteristic instances of a class, these archetypes can then be generalized and used for the remaining instances of the cluster.

3.1.3 Utility Calculation.
Incorporating eq. (3) and eq. (4) into eq. (1), yields the equation to determine the utility of an unlabeled instance x_i:

$$\text{Utility}(x_i) = \left(1 - f_{margin}(x_i)\right) \times \left(\frac{1}{U} \sum_{u=1}^{U} sim_{\text{eucl}}(x_i, x_u)\right)^{\beta} \qquad (6)$$

3.1.4 Diversity Calculation.
Based on the calculated $\text{Utility}(x_i)$, particularly useful instances for training are determined and suggested to the user. A naive approach would be to suggest the top n instances with the highest utility. However, since these can be located very close to each other, superfluous instances may be labeled. To distribute the recommendations meaningfully over the entire data set, we use kernel k-means clustering [11] as proposed by [14]. However, we do not use all unlabeled instances for clustering, but rather the subset with a high utility. We choose the instances closest to each of the k calculated cluster centroids. Thus k instances are recommended which are both useful and diverse.

3.1.5 Learning Model.
We use a random forest classifier [7] for the prediction of class labels. An advantage of random forests is that they directly yield the probability of each class for each instance, which is necessary for the uncertainty calculation. Besides, the classifier is fast in training and prediction, which is important for seamless user interaction and allows the approach to scale well for higher-dimensional data. The respective results must be made available to the users as quickly as possible, otherwise, they will lose interest and attention due to long waiting times. In the implementation, the standard configuration of the scikit-learn[1] package was used. Each random forest consists of 100 decision trees generated using Giny impurity and bootstrapping without restricting the depth of the trees. Besides determining the uncertainty for the utility calculation, the classifier is used for the automated labeling

of all unlabeled instances. These labels are tentative, users can confirm them manually. During the training process, predicted labels may change as the classifier is retrained.

3.2 Result Visualization

The second component is focused on conveying the data as well as information of the model to the user. We use visualization since this enhances the cognitive abilities of humans by methodically preparing and presenting data in interfaces with interactive visualization [12]. Thereby, the user can gain deeper insights into heterogeneous and complex data through explorative knowledge discovery [12].

3.2.1 Data Visualization.
When displaying the instances to be labeled, the high dimensionality of most data sets can cause difficulties in the visualization. Although there are forms of representation such as parallel coordinates [16] or radar-plot [22] which can represent many dimensions, they can become cluttered when there is too much data, or they cannot reveal the relationships and clusters in the data. Hence, projection methods are used to reduce the dimensionality and display the data with a 2D scatter plot. Sedlmair et al. [23] showed that a scatter plot of projected data can clearly display clusters in the data.

For the projection of the data, our visualization approach uses t-distributed stochastic neighbor embedding (t-SNE). This algorithm is widely used in ML and visual interactive labeling due to its "ability to create compelling two-dimensional maps from data with hundreds or even thousands of dimensions" [29] effectively. The method is occasionally criticized for creating visualizations that lead to wrong conclusions, as they distort the data set or present nonexistent clusters. However, such scenarios can be avoided by correctly configuring the main hyperparameters [15, 29]. The methodical configuration of these hyperparameters was not part of this work and remains an open challenge for future work. For our prototype, we experimentally figured out *perplexity* = 20, *maximum iterations* = 5000, and used the default configuration of the scikit-learn[1] function for the remaining parameters. This visualization allows users to apply data-based strategies [4] to identify interesting instances based on the location and relation of the instances.

3.2.2 Model Visualization.
To enable users to apply model-based strategies, the visualization of the data is extended by additional information from the model. Regarding the recognized model-based strategies for instance selection, users always fall back on the model's predicted labels to evaluate the current state of the model. Thus, we extend the visualization so that for each data point the predicted or manually assigned class label is indicated by color.

Bernard et al. [5] raised the concern whether the predicted labels should be displayed, as they might lead to possible bias among users. We argue that this information is necessary for users to evaluate the current state of the model. Furthermore, Bernard refers the criticism primarily to collaborative scenarios where labels of one analyst might influence the decisions of another analyst, while our approach aims at the use of one user at a time.

To distinguish labeled from unlabeled data points, they are shown by different icons in the scatter plot, showing the coverage of labeled data points. Moreover, this provides users with a visual indication of their impact on the system, since not only the icons of the instances

---

[1]https://scikit-learn.org/stable/

manually labeled by the users change but also the color, size, and icons of the other instances. Visualizing the model's improvement achieved by the users should increase their perceived relevance in the process and motivate them [19]. In addition to shape and color, a further visual cue is offered by encoding the size of the icons w.r.t. the calculated utility of each instance (see eq. (1)).

3.2.3 User Guidance. The approach supports users by the first two degrees of guidance [9], i.e. *orienting* and *directing*, in the selection of instances. With the help of visualization, users can compare their mental models containing ideas about an ideal model with the current state of the model. They can evaluate the current model and derive steps to label and add instances to the training set to improve the model. To guide users to identify useful instances, the measures introduced in the *Candidate Suggestion* step are used (see Section 3.1). To involve the model in the instance selection, the visualization is used for the suggestion of instances. A highly distinct icon of increased size is used with a signal color not used for any other class. Thus, the model's recommendation clearly stands out and can be easily noticed by the users. Inspired by findings on the perception of interactivity, we exploit visual cues such as shape, color, and size to direct users' attention to specific areas in the visualization. Especially the star icon, as an "*abstract icon*" [6], is predestined to convey the information that the respective instance is informative. Many modern interfaces use the star icon as symbolism for favorites or featured items.

3.2.4 Interactive Selection. The presented visualization is an interactive interface that allows users to interact with the visualization, the data, and thus indirectly with the model. Besides simple interactions following the information-seeking mantra [26], like changing the displayed area of the visualization by moving, zooming, or filtering, users can also perform semantic interactions. These include the selection of instances and the assignment of labels. Users can select and label several instances at once in one interaction to make the process more efficient. The selection of the instances to be labeled is done exclusively by the user, guided by the discussed visual cues.

3.3 Design and Implementation

We implemented a prototype to evaluate the effectiveness of our approach. The logic behind the utility calculation, recommendations, and projection were implemented using scikit-learn[1] and modAL[2] in Python. For the user interface a javascript-based web application was implemented, using Plotly[3] for interactive visualizations. The user interface consists of four main components (Fig. 2):

3.3.1 Instance Overview and Selection. The *Instance Overview and Selection* area is the central element of the workspace (see the center of Fig. 2). It shows the projected instances with the corresponding visual mappings for the status, class, and utility. A legend shows the color assigned to each class, which is used consistently throughout the tool to identify that class. Labeled instances are represented by an outlined diamond with a dot, to distinguish them easily from unlabeled circles. Suggested instances, indicated by the red stars, visibly stand out from the surrounding instances. Using the toolbar, users can move the displayed area and zoom in. By default, the

lasso tool is active so that users can select interesting instances by clicking and circling around them. This view allows users to select instances from potentially interesting areas for review and labeling.

3.3.2 Detail View and Filtering. All instances selected in *Instance Overview and Selection* are displayed in descending order according to their utility in the view *Detail View and Filtering* (see bottom right in Fig. 2). In this list view, the user gets additional details about each selected instance such as a visual representation along with the manually assigned or predicted label. The corresponding symbol of each instance is displayed so that instances in the list view can be associated with data points in the visualization. Moreover, the instance that the user hovers over in the list is highlighted in the visualization. The *utility* (see eq. (1)) of the instance is shown by a utility bar so that it can be better perceived and compared than a plain number. By clicking on the corresponding instance in the list view, users can filter out non-interesting or unwanted instances from the selection. When working with image data, small preview images are included in the list view. For more complex, multi-modal data, an appropriate visualization should be embedded. As the focus of this work is on guiding users in their selection, this challenge is beyond the scope of this work and can be addressed in further work. With this view, selected instances from the overview can be examined more closely and filtered if necessary.

3.3.3 Labeling Interface. Users can assign class labels to the remaining selected instances in the *Labeling Interface* (top right in Fig. 2). A pre-defined class alphabet was used for the prototype. However, for scenarios with an unknown number of classes, our approach can be extended to add new classes as required. In addition to the explicit assignment of labels, users can accept labels predicted by the model. This is especially useful in an advanced state of the model, where predictions are usually correct. Thus, at the end of the interactive labeling process, all remaining unlabeled instances can be automatically labeled by the model with one button click.

3.3.4 Support Visualizations. On the left of the workspace (Fig. 2) additional information about the model's current state and the distribution of the labels are shown. For the prototype, the calculated accuracy of the model was displayed for orientation. This information is not available in a real application but can be replaced by information such as hyperparameters of the model. Also, data on the distribution of the labels is shown. Since the class alphabet used for the prototype was known, the user can be informed about classes that have not been labeled yet. This is an attempt to prevent classes from being overlooked. A bar plot shows the distribution of the assigned labels over the respective classes encouraging users to distribute labels evenly to prevent bias in the model. For data sets with imbalanced classes, this plot can be extended to include predicted labels allowing users to see if the model reflects the actual known distribution of labels. If the distribution is unknown, this view can help to develop a mental model of the data set.

4 EVALUATION: USER SURVEY

While RQ_1 was addressed by the design of the approach itself, RQ_2 of how guiding users affects their selection of instances to be labeled, was evaluated with a user survey. 19 master students familiar with machine learning and visualization were recruited.
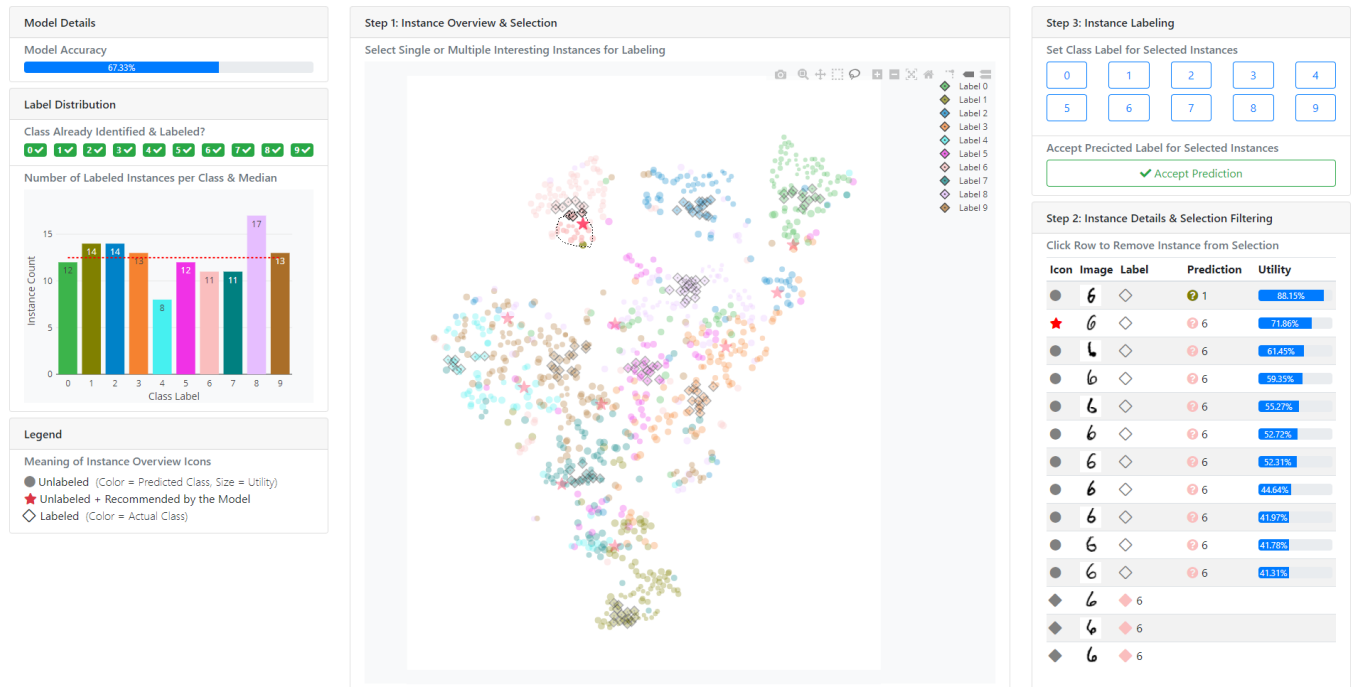

Figure 2: Workspace of the tool used for the study with the areas *Instance Overview and Selection* (center), *Detail View and Filtering* (bottom right), *Labeling Interface* (top right), and *Support Visualizations* (left). The selected instances from *Instance Overview and Selection* area in the *Detail View and Filtering* view are ordered by their utility, which is visualized by a bar.

4.1 Study Setup

The participants were asked to label a subset of 1000 instances from the MNIST data set [18] which contains 10 classes with the handwritten digits 0...9.

The study was subdivided into two subsequent setups using the same prototype (depicted in Fig. 2), without the guidance by visual cues in setup 1 (S_1) and with guidance in setup 2 (S_2). In both setups, the interactive visualization contains the separation of classes by color and a distinction between labeled and unlabeled instances. S_2 additionally includes guiding users through the visualization of the instances' utility (see Section 3.1) by the size of the icons and highlighting recommended instances with colored stars. After an initial demonstration of the prototype and the possibility to get familiar with the tool using a different data set, the two setups were subsequently conducted with a 10 minutes time limit for each. Two different subsets of the MNIST data set were used.

After completing the two tasks, the participants were asked to complete a digital questionnaire (see Table 1). The questionnaire included questions on subjective perceptions of the difficulty and familiarity of identifying and selecting instances under the two setups, preferences, and informal feedback. The answer options were formulated as 5-point Likert scales. Further, questions with plain text were posed to allow for open feedback. The questionnaire was designed to determine how users assess and perceive the influence of guidance on their selection. Furthermore, it should be evaluated which particular factors influence the users' selection of instances to better understand the guidance of users for labeling data.

Table 1: Questions on user self-assessment to identify interesting instances for labeling and the influence of guidance on their selection using 5-point Likert scales.

| Question Text	Answer options [1...5]		
Q_1: How confident were you in identifying interesting instances WITHOUT recommendation stars and different sized icons?	insecure	neutral	confident
Q_2: How difficult did you find the identification of interesting instances WITHOUT support of the stars and different sized icons?	hard	neutral	easy
Q_3: How confident were you in identifying interesting instances INCLUDING recommendation stars and different sized icons?	insecure	neutral	confident
Q_4: How difficult did you find the identification of interesting instances INCLUDING support of the stars and different sized icons?	hard	neutral	easy
Q_5: How much did the recommendation stars and the icons INCLUDING different sizes influenced your decision, which instances are interesting?	weak	neutral	strong
Q_6: In which way did the recommendation stars and the icons INCLUDING different sizes influence your decision, which instances are interesting?	confusing	neutral	strengthening
Q_7: How did you experience the recommendation stars and Icons INCLUDING different sizes when identifying interesting instances compared to identification WITHOUT additional support?	hindering	neutral	helpful
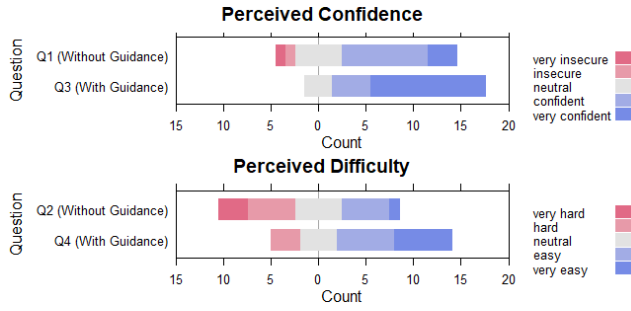
Figure 3: Perceived confidence of users (top) when selecting instances with user guidance (Q_1) and without (Q_3). No user stated being insecure in S_2. Perceived difficulty of users (bottom) in identifying instances with user guidance (Q_2) and without (Q_4). The number of users having difficulties identifying instances for labeling decreased by more than half compared to S_1 without guidance.

Figure 4: Perceived strength (top), perceived type (middle), and perceived usefulness (bottom) of the guidance's influence by the users.

4.2 Study Results

The answer distributions regarding perceived confidence (Fig. 3 top), difficulty (Fig. 3 bottom), and the influence of guidance (Fig. 4) were analyzed and compared. To analyze further impressions and opinions of the users about the tool and its two versions, we conducted an explorative analysis of the answers to the open questions.

4.2.1 Perceived Confidence. Users' self-assessment of the perceived confidence in identifying interesting instances (Fig. 3 top) reveals that the majority of users felt confident in finding interesting instances regardless of the guidance. None of the users stated that they were insecure in their self-assessment of the version including user guidance. 16 of 19 users felt confident in the guided selection.

The users' perception of the unguided version shows that the majority of users found it difficult to identify interesting instances for labeling (Fig. 3 bottom). In the guided version, $\frac{3}{4}$ of the users reported that it was easy for them to identify instances. Only 3 users found it "very hard". Compared to the version without guidance, the number of users who found it easy to identify interesting instances more than doubled in the version including user guidance.

4.2.2 Perceived Influence. A majority of users stated that they were strongly influenced in their selection by the guidance through the emphasis and recommendations and that this influence was additionally helpful. Half of the users said that the highlights and recommendations encouraged them in their choice. Only two resp. three of the 19 users reported negative experiences that the guidance had hindered or confused them. Part of them stated that they did not make use of the offered guidance and did neither pay attention to the icon size nor the recommendations. One participant claimed that the recommendations introduced a bias towards certain instances and classes, thus compromising his/her intuition suggesting that guidance by icon size only would be sufficient.

4.2.3 Feedback on the Tool. The participants provided plain text feedback on the two versions of the tool. Users liked the version without user guidance that they were not influenced or distracted by the recommendations. In S_1 they did not have to search for the stars.
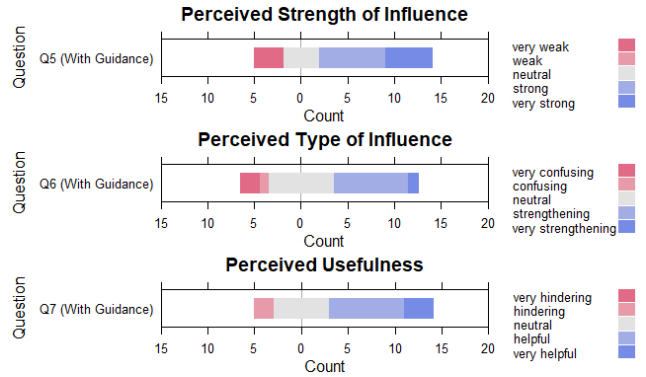
The users stated that they recognized the distribution and cluster of the instances via the density of the data points well. However, the majority of users reported in the version without guidance they often had no orientation on what instances should be selected and labeled next. In general, it was difficult to find interesting instances due to the unclear presentation of the data points.

For the second version of the tool including recommendations and varying sizes of the icons, users said that it gave them clues as to where it would make sense to label. More information was available to find instances. Users felt that the model's accuracy increased quicker and that they had more confidence in their selection.

13 of 19 participants preferred S_2 including user guidance. According to them, the version is clearer, faster, and more effective to work with owing to the guidance. They liked that the tool allows for a more confident selection of instances, as the tool reinforces their decision. While it might seem obvious that the guided version is better for navigating users towards interesting instances, we argue that any general navigation solution might also interfere with the users' intuition and decision-making. Our study showed that some, yet only very few, participants preferred the unguided version.

In general, users enjoyed the clarity and easy handling of both versions. Apparently the tool has made a rather exhaustive and repetitive task enjoyable. For future versions, users demand more convenience through features like hotkeys, the ability to filter labeled data or to toggle the guidance on and off.

5 CONCLUSION

We presented an approach to interactively label instances that implements the VIAL process, providing a combination of user-based selection and model-based suggestions to identify candidates for labeling (Section 3). Based on an AL query strategy, the estimated benefit for all unlabeled instances as well as recommendations of particularly useful instances are calculated and then presented to the user via visual cues in an interactive visualization. The approach was evaluated with a user survey (Section 4) to determine the effects of user guidance. The use of visual cues in an interactive visualization offers a promising strategy for guiding users to focus their attention on specific regions and instances. However, the

influence of the individual cues on an accurate representation of the data set must be taken into account, otherwise, distortions may occur. The results of the survey show that guidance influences the selection and labeling strategy and thus, according to the majority of participants, makes labeling easier and more reliable.

Possible limitations of our work result from the use of the simple data set MNIST. It contains only a few, well distinguishable classes, which are also easy to represent visually. It is possible that the colors used in the individual classes may not be as visually distinguishable as their number increases. It is also open how more complex data can be visually represented in the detail view. Besides, we have performed our experiment on a subset of the data set. A larger number of instances could lead to a confusing visualization that users can no longer work with. On the other hand, strategies for a reasonable selection of instances can be used, so that, as in our case, only a manageable subset is used, and the labels are transferred to the entire data set via the model.

In future work, we want to conduct a user study measuring the progress of labeling in addition to the reported self-assessment. As part of this, we want to use an additional data set to demonstrate the capabilities of our tool and how the user interface can handle different scenarios. Furthermore, we will analyze how the system processes instances whose class labels are ambiguous or are changed frequently by the users throughout the labeling process. Further research could be conducted on how the degree and form of guidance can be adapted to the users' individual needs so that they can either determine the degree of guidance themselves to receive recommendations or precise instructions for their selection. Besides, it would be useful to know whether a model or algorithm can automatically determine what kind of guidance a user needs.

REFERENCES

[1] H Baumgartl, J Tomas, R Buettner, and M Merkel. 2019. A novel deep-learning approach for automated non-destructive testing in quality assurance based on convolutional neural networks. *ACEX-2019 Proceedings* (2019).

[2] David Beil and Andreas Theissler. 2020. Cluster-Clean-Label: An interactive Machine Learning approach for labeling high-dimensional data. In *Proceedings of the 13th International Symposium on Visual Information Communication and Interaction - VINCI'2020*. ACM Press. https://doi.org/10.1145/3430036.3430060

[3] Jurgen Bernard, Marco Hutter, Matthias Zeppelzauer, Dieter Fellner, and Michael Sedlmair. 2018. Comparing Visual-Interactive Labeling with Active Learning: An Experimental Study. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (Jan. 2018), 298–308. https://doi.org/10.1109/tvcg.2017.2744818

[4] Jürgen Bernard, Matthias Zeppelzauer, Markus Lehmann, Martin Müller, and Michael Sedlmair. 2018. Towards User-Centered Active Learning Algorithms. *Computer Graphics Forum* 37, 3 (June 2018), 121–132. https://doi.org/10.1111/cgf.13406

[5] Jürgen Bernard, Matthias Zeppelzauer, Michael Sedlmair, and Wolfgang Aigner. 2018. VIAL: a unified process for visual interactive labeling. *The Visual Computer* 34, 9 (March 2018), 1189–1207. https://doi.org/10.1007/s00371-018-1500-3

[6] Jeremy Boy, Louis Eveillard, Francoise Detienne, and Jean-Daniel Fekete. 2016. Suggested Interactivity: Seeking Perceived Affordances for Information Visualization. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (Jan. 2016), 639–648. https://doi.org/10.1109/tvcg.2015.2467201

[7] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32. https://doi.org/10.1023/a:1010933404324

[8] Maya Cakmak, Crystal Chao, and Andrea L Thomaz. 2010. Designing Interactions for Robot Active Learners. *IEEE Transactions on Autonomous Mental Development* 2, 2 (June 2010), 108–118. https://doi.org/10.1109/tamd.2010.2051030

[9] Davide Ceneda, Theresia Gschwandtner, Thorsten May, Silvia Miksch, Hans-Jorg Schulz, Marc Streit, and Christian Tominski. 2017. Characterizing Guidance in Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (Jan. 2017), 111–120. https://doi.org/10.1109/tvcg.2016.2598468

[10] Mohammad Chegini, Jürgen Bernard, Philip Berger, Alexei Sourin, Keith Andrews, and Tobias Schreck. 2019. Interactive labelling of a multivariate dataset for supervised machine learning using linked visualisations, clustering, and active

[11] Inderjit S. Dhillon, Yuqiang Guan, and Brian Kulis. 2004. Kernel k-means. In *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*. ACM Press. https://doi.org/10.1145/1014052.1014118

[12] Brian Fisher. 2005. *Illuminating the Path: An R&D Agenda for Visual Analytics*. IEEE Computer Society Press.

[13] Yifan Fu, Xingquan Zhu, and Bin Li. 2012. A survey on instance selection for active learning. *Knowledge and Information Systems* 35, 2 (June 2012), 249–283. https://doi.org/10.1007/s10115-012-0507-8

[14] Tianxu He, Shukui Zhang, Jie Xin, Pengpeng Zhao, Jian Wu, Xuefeng Xian, Chunhua Li, and Zhiming Cui. 2014. An Active Learning Approach with Uncertainty, Representativeness, and Diversity. *The Scientific World Journal* 2014 (2014), 1–6. https://doi.org/10.1155/2014/827586

[15] Qiwen Hu and Casey S. Greene. 2018. Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell RNA transcriptomics. In *Biocomputing 2019*. WORLD SCIENTIFIC.

[16] A. Inselberg and B. Dimsdale. [n.d.]. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In *Proceedings of the First IEEE Conference on Visualization: Visualization 90*. IEEE Comput. Soc. Press. https://doi.org/10.1109/visual.1990.146402

[17] Kostiantyn Kucher, Carita Paradis, Magnus Sahlgren, and Andreas Kerren. 2017. Active Learning and Visual Analytics for Stance Classification with ALVA. *ACM Transactions on Interactive Intelligent Systems* 7, 3 (Oct. 2017), 1–31. https://doi.org/10.1145/3132169

[18] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324. https://doi.org/10.1109/5.726791

[19] Mario Nadj, Merlin Knaeble, Maximilian Xiling Li, and Alexander Maedche. 2020. Power to the Oracle? Design Principles for Interactive Labeling Systems in Machine Learning. *KI - Künstliche Intelligenz* 34, 2 (Jan. 2020), 131–142. https://doi.org/10.1007/s13218-020-00634-1

[20] Don Norman. 2008. Design of Everyday Things. (2008).

[21] Christian Ritter, Christian Altenhofen, Matthias Zeppelzauer, Arjan Kuijper, Tobias Schreck, and Jürgen Bernard. 2018. Personalized Visual-Interactive Music Classification. In *EuroVis Workshop on Visual Analytics (EuroVA)*, Christian Tominski and Tatiana von Landesberger (Eds.). The Eurographics Association. https://doi.org/10.2312/eurova.20181109

[22] M. Joan Saary. 2008. Radar plots: a useful way for presenting multivariate health care data. *Journal of Clinical Epidemiology* 61, 4 (April 2008), 311–317. https://doi.org/10.1016/j.jclinepi.2007.04.021

[23] Michael Sedlmair, Tamara Munzner, and Melanie Tory. 2013. Empirical Guidance on Scatterplot and Dimension Reduction Technique Choices. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (Dec. 2013), 2634–2643. https://doi.org/10.1109/tvcg.2013.153

[24] Burr Settles. 2009. *Active learning literature survey*. Technical Report. University of Wisconsin-Madison Department of Computer Sciences.

[25] Burr Settles. 2011. From Theories to Queries: Active Learning in Practice. In *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010 (Proceedings of Machine Learning Research, Vol. 16)*, Isabelle Guyon, Gavin Cawley, Gideon Dror, Vincent Lemaire, and Alexander Statnikov (Eds.). PMLR, Sardinia, Italy, 1–18.

[26] B. Shneiderman. 1996. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*. IEEE Comput. Soc. Press, 336–343. https://doi.org/10.1109/vl.1996.545307

[27] Andreas Theissler, Anna-Lena Kraft, Max Rudeck, and Fabian Erlenbusch. 2020. VIAL-AD: Visual Interactive Labelling for Anomaly Detection – An approach and open research questions. In *4th International Workshop on Interactive Adaptive Learning (IAL2020)* (Ghent, Belgium). CEUR-WS.

[28] Devis Tuia, Michele Volpi, Loris Copa, Mikhail Kanevski, and Jordi Munoz-Mari. 2011. A Survey of Active Learning Algorithms for Supervised Remote Sensing Image Classification. *IEEE Journal of Selected Topics in Signal Processing* 5, 3 (June 2011), 606–617. https://doi.org/10.1109/jstsp.2011.2139193

[29] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. 2016. How to Use t-SNE Effectively. *Distill* (Jan. 2016). https://doi.org/10.23915/distill.00002

[30] Jenna Wiens and Erica S Shenoy. 2017. Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology. *Clinical Infectious Diseases* 66, 1 (Aug. 2017), 149–153. https://doi.org/10.1093/cid/cix731

[31] Liping Yang, Alan MacEachren, Prasenjit Mitra, and Teresa Onorati. 2018. Visually-Enabled Active Deep Learning for (Geo) Text and Image Classification: A Review. *ISPRS International Journal of Geo-Information* 7, 2 (Feb. 2018), 65. https://doi.org/10.3390/ijgi7020065

[32] You Zhu, Li Zhou, Chi Xie, Gang-Jin Wang, and Truong V. Nguyen. 2019. Forecasting SMEs' credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach. *International Journal of Production Economics* 211 (May 2019), 22–33. https://doi.org/10.1016/j.ijpe.2019.01.032