EnsembleLens: Ensemble-based Visual Exploration of Anomaly Detection Algorithms with Multidimensional Data

Ke Xu, Meng Xia, Xing Mu, Yun Wang, and Nan Cao

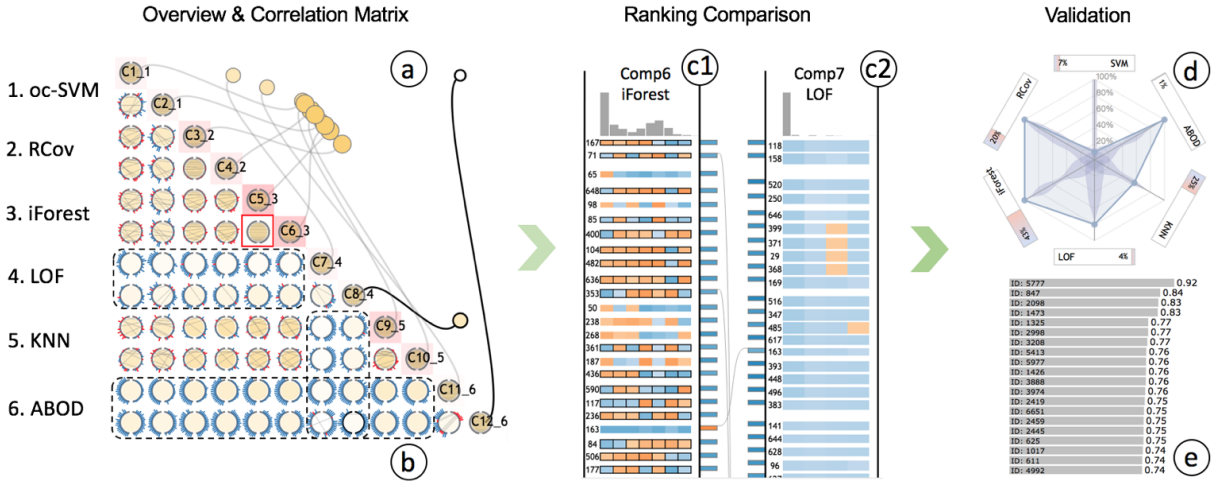

Fig. 1. EnsembleLens facilitates the exploration of anomaly detection algorithms via three levels of analysis, namely, (a) the overview as the macro level, (b) the correlation matrix view as the meso level, and (c) the ranking view as the micro level. (d) is for the validation. The figure showcases the analytic process based on the Wisconsin-Breast Cancer (Original) dataset with 699 instances and 10 attributes. First, (a) and (b) indicate that LOF (C7, C8) and ABOD (C11, C12) have little correlation with the others because the ensemble components generated by them are far from the major cluster, and the correlation matrix also shows their low correlation with the others (i.e., few crossing lines, many blue bars around the glyph). (c1) illustrates that the iForest algorithm has better performance for this dataset than LOF (c2), as most top-ranked points in (c1) are proved to be anomalous. By contrast, the top-ranked points in LOF (c2) are normal and consistent. (e) is the combination rank list of all algorithms based on their weights.

Abstract—The results of anomaly detection are sensitive to the choice of detection algorithms as they are specialized for different properties of data, especially for multidimensional data. Thus, it is vital to select the algorithm appropriately. To systematically select the algorithms, ensemble analysis techniques have been developed to support the assembly and comparison of heterogeneous algorithms. However, challenges remain due to the absence of the ground truth, interpretation, or evaluation of these anomaly detectors. In this paper, we present a visual analytics system named EnsembleLens that evaluates anomaly detection algorithms based on the ensemble analysis process. The system visualizes the ensemble processes and results by a set of novel visual designs and multiple coordinated contextual views to meet the requirements of correlation analysis, assessment and reasoning of anomaly detection algorithms. We also introduce an interactive analysis workflow that dynamically produces contextualized and interpretable data summaries that allow further refinements of exploration results based on user feedback. We demonstrate the effectiveness of EnsembleLens through a quantitative evaluation, three case studies with real-world data and interviews with two domain experts.

Index Terms—Algorithm Evaluation, Ensemble Analysis, Anomaly Detection, Visual Analysis, Multidimensional Data

1 Introduction

Anomaly detection is the identification of data points that do not conform to the expected patterns in a dataset. It is applied in a wide range of domains, such as intrusion detection in cyber-security systems, fraud detection in financial transactions and disease detection in public health [15]. Many anomaly detection algorithms, including supervised [70] and unsupervised algorithms [23], have been proposed since the 19*th* century [2]. Most of them are for multidimensional data. However, different anomaly analysis methods performed on multidimensional data are specialized for different properties of the data, which means anomaly detection results are sensitive to the choice of the algorithms and feature subspaces. This situation raises an open but widely ignored question. How can data mining experts or algorithm engineers properly evaluate, compare and select existing methods for a given dataset, which can either generate a more effective detection result or enumerate the prior algorithms for further development? Ensemble analysis has been introduced to solve this problem. *Anomaly ensembles* select and combine heterogeneous anomaly detection results to obtain a more robust set of outliers rather than some aspects of "the whole truth", which in turn can be applied to evaluate the performance of anomaly detection algorithms [1].

Further, automated anomaly ensemble techniques have been recently developed to support a more systematic selection of anomaly detection algorithms [26, 48, 62]. However, the effectiveness of these techniques

- *Ke Xu, Meng Xia, and Yun Wang are with the Hong Kong University of Science and Technology. E-mail: {kxuak, iris.xia, ywangch}@connect.ust.hk*
- *Xing Mu is with the Hong Kong University of Science and Technology. Email: hecate.l.mu@gmail.com*
- *Nan Cao is with the iDV^x Lab, Tongji University and is the corresponding author. Email: nan.cao@tongji.edu.cn*

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

is hindered by two inherent obstacles of anomaly detection. First, rigid definition of anomalous points usually does not exist. Second, high-quality labeled data for training the estimation model of anomaly detection is usually unavailable or time-consuming to obtain. Even when the labels are available, the performance metrics used to evaluate the models are based only on accuracy (e.g., true positive rate) or algorithm cost, instead of correlation, interpretation and reliable insights of the ensembles. Hence, human judgment, a flexible and principled exploration of the model behavior and the analysis results, is naturally required to evaluate the candidate algorithms.

By contrast, recent advances in ensemble data visualization have shown great promise towards understanding the relationships among different models that construct the ensemble, as well as the connections between input data and output ensemble from multidimensional datasets [17, 47, 75]. However, none of them was developed as an expert tool to help algorithm developers explore anomaly detection algorithms. It is of great importance to select the proper algorithms for different anomaly detection tasks, where visual ensemble analysis has a great potential to meet these requirements. After a comprehensive investigation of these preliminary designs, we conclude three visual analytics challenges in ensemble anomaly detection with multidimensional data. (1) Comparison: difficulties are encountered in designing a scalable visualization to compare multiple *ensemble components* based on multiple criteria and along with the raw data context. Here we define an ensemble *component* as one anomaly detection model generated by one algorithm/detector with a specific parameter setting and a sampled feature subspace of data. (2) Interpretation: designs that could visually represent the model behavior or reveal the semantic meaning behind the results by exhibiting the relationship between the choice of algorithms and the ensemble result, as well as the pairwise correlations of ensemble components themselves, are lacking. (3) Interaction: the needs of supporting ensembles investigation, incorporating human judgment and feedback, as well as iteratively guiding the system to produce a better evaluation of anomaly detection algorithms have not been addressed.

To address these challenges, we introduce EnsembleLens, a novel integrated visual analytics system for interactively exploring, analyzing and selecting anomaly detection algorithms for different multidimensional datasets. EnsembleLens employs an integrated model based on ensemble analysis to formulate the unsupervised process of algorithm evaluation, which incorporates a variety of baseline detection algorithms, feature bagging and ensemble combination functions. Multiple coordinated views are provided in EnsembleLens to visually represent the analysis results from different ensemble components, supporting analytical tasks including summarization, reasoning, assessment and correlation analysis. As its primary approach to implement ensemble-based analysis of anomaly detectors, the system uses the *level of analysis* (macro-meso-micro) exploration technique to display the anomaly ensembles with different scales and semantics. Moreover, EnsembleLens is closely linked with coordinated side views through rich interaction to help understand the relationship between results and users' selection of anomaly detectors and features. Specifically, this work makes the following contributions.

- **System.** We introduce an integrated visual analytics system that provides a user-guided evaluation of anomaly detection algorithms with multidimensional data based on ensemble analysis. This system visualizes the correlation between different anomaly detectors and illustrates the importance of these detectors by their weights accounting for the optimized ensemble.
- **Interactive Exploration.** We adopt an interactive ensemble approach that supports the construction of anomaly ensembles through three steps: algorithm setting, feature bagging and ensemble combination. To facilitate the exploration of anomaly detection algorithms based on anomaly ensembles, we implement the *level of analysis* (macro-meso-micro) visual analytics methods to provide a fine-grained evaluation of different detection algorithms based on user feedback.
- **Visualization Designs.** We propose a set of visualization designs, as well as layout algorithms for efficiently summarizing and evaluating the ensemble components generated by various anomaly detectors. In particular, we propose a novel layout algorithm to demonstrate the overview of different ensemble components, a matrix view with a customized glyph to display the correlation between each pair of ensemble components and a scalable ranking list view with a "barcode" metaphor for comparing the detailed outlier scores of different ensemble components.
- **Evaluation.** The effectiveness of EnsembleLens is demonstrated in multiple forms of evaluation. We describe how EnsembleLens works through three case studies with real-world data from the UCI Machine Learning repository. Each case study is followed by a quantitative study to evaluate the performance of ensemble analysis result. We also conduct the expert interviews with two researchers in the data mining domain.

2 Related Work

This section provides an overview of research that is most related to our work, which generally includes: (1) algorithms for anomaly detection and their evaluation, (2) techniques for visual ensemble analysis, and (3) visual exploration of anomaly detection algorithms.

2.1 Anomaly Detection Algorithms and Evaluations

Various anomaly detection related methods have been developed in diverse research areas and application domains over the past decades, including the traditional anomaly detection algorithms, the ensemble approaches and the evaluation methods.

Anomaly Detection Algorithms. Generally, the objective of anomaly detection is to find special patterns in data that appear to be inconsistent with well-defined behavior [15, 37, 66]. Existing techniques mainly fall into four categories based on how they model and detect anomalies, including classification-based algorithms (usually supervised [35, 51, 76]), neighbor-based or distance-based algorithms [8,12,34,42], statistics-based algorithms [7,77] and tensor-based algorithms [13]. Various approaches have been taken to address the problem with multidimensional data [18, 80]. For example, spectral-based algorithms approximate the raw data to a lower dimensional subspace, which includes multidimensional scaling (MDS) [45], principal component analysis (PCA) [68] and compact matrix decomposition (CMD) [71]. Angle-based outlier detection (ABOD) evaluates the variance between an abnormal candidate and all other pairs of points from the perspective of angles [44]. Scatterplot matrices and parallel coordinates represent data values across multiple dimensions [38]. All of the techniques discussed are not comprehensive but represent different approaches. Our system adopts some of the most representative algorithms from different categories as the baseline algorithms.

Anomaly Ensembles. Most of the aforementioned techniques are specific to different observational features and fit only to parts of "the whole truth" of anomaly detection. A general way to reduce imbalance is the ensemble-based approach, which selects and combines the anomaly detection results from a set of algorithms to obtain robust anomaly scores. Ensemble analysis [21] has been widely studied and long proven effective for classification [32, 61, 63] and clustering [24, 28, 31, 74] problems. Existing attempts at anomaly ensembles reveal their particular effectiveness in multidimensional anomaly detection [3, 40, 48]. For example, Nguyen et al. [53] and Fu et al. [26] constructed their ensembles by matching feature subspaces and anomaly detectors within a time limit. The nature of ensemble anomaly detection, comparing and selecting results from different algorithms according to the data characteristics, inspired our research on evaluating different algorithms for a given dataset. Yet, instead of using the anomaly ensembles to generate a more robust detection result, we take advantage of them to assist in exploring the baseline algorithms, thereby enumerating the most effective detectors for different datasets.

Anomaly Detection Algorithm Evaluation. A number of methods have been proposed to evaluate the anomaly detection algorithms with different evaluation metrics. The traditional method uses the true positive rate or the receiver operating characteristic (ROC) curve which merely considers the rank of outlier scores [3, 4, 8, 30, 39, 48, 56]. Such evaluation ways have been applied in many domains like maritime navigation [5], video surveillance [6] and intrusion detection [36]. However, these methods have limited effectiveness for unsupervised anomaly detection, because the ground truth for calculating the precision is lacking. A wider perspective on evaluating anomaly detection results considers both the outlier scores and the ranking [27, 43]. For example, Muller et al. [52] introduced an "outlier ranking coefficient"

for each instance based on an adaptive degree of deviation in different subspaces. Goix et al. [29] proposed two novel criteria based on existing Excess-Mass (EM) and Mass-Volume (MV) curves. When comparing multiple detection results, the similarity or correlation of the outlier scores is commonly used, such as Pearson's r [57], Spearmans ρ [69] and Kendall's τ [41]. Advanced correlation coefficients incorporate weights to different data points to provide more reasonable evaluation [10, 46, 67]. Most relevantly to our work, Schubert et al. [65] proposed a generalized view of evaluation methods to compare different anomaly detection methods based on similarity of rankings and anomaly scores. Although these methods are developed to evaluate anomaly ensembles from different detection algorithms, they failed to reveal the reasoning of anomaly ensemble comparison. Thus, assessment and interpretation of the results are very difficult if not impossible. The method proposed in our work offers a more fine-grained and interpretative result, with the overall clustering of ensemble components, the top-ranked items' correlation and detailed explanatory information about the outlier score ranking lists.

2.2 Visual Ensemble Analysis

Ensemble visualization is a non-trivial research problem [55] as the ensemble data are usually large and multidimensional. Numerous visualization approaches have been introduced to support ensemble analysis for various purposes, for example, serving for weather forecast [47] or climate model evaluation like SimEnvVis [54], Ensemble-vis [60], Noodle [64] and Albero [20]. Also, ensemble data has been studied for uncertainty visualization [17, 25, 59], as well as biomedical applications [22, 78]. These systems are designed specifically for their applications based on the domain-specific assumptions and requirements. Compared with these systems, a more general work without specific scenarios is EnsembleMatrix, which combines various classifiers according to the merits reflected by confusion matrices [72]. The methods to visualize and evaluate ensembles of these systems can be broadly categorized into three types: (1) glyph-based visualizations [11, 33] that encode different aspects of data into distinct visual representations, which possess low extendability to general multidimensional data; (2) parallel coordinated views that compare different ensemble components based on certain attributes, but fail to reveal the detailed information involved in data [58, 75]; and (3) multi-charts or matrix-based visualizations [9, 19] that effectively combine ensemble comparison and detailed information.

However, none of these approaches are designed for anomaly detection applications. To the best of our knowledge, EnsembleLens is the first system that applies visualization techniques to anomaly ensembles and supports the exploration of anomaly detectors. Furthermore, the visualization designs of existing systems face the issue of visual clutter when the data size is large. Inspired by these early design guidelines, we provide a *level of analysis* workflow that allows both general evaluation of ensembles and detailed comparison of rankings. Moreover, the visualization is dynamic such that users can explore the changes caused by different combinations of ensemble components and support the evaluation of algorithms that produce the ensemble.

2.3 Visual Exploration of Anomaly Detection Algorithms

Although, a variety of works are related to visual anomaly detection and algorithm evaluation [14, 49, 73, 79], they were developed to explore anomalous instances in specific application domains. Anomaly detection evaluation possesses two inherent challenges: (1) no rigid definition of which cases are exactly normal or abnormal, and (2) the lack of labeled data for training and verifying models. These challenges are further aggravated by the increase of data dimensions, where different features have different semantic interpretations. Some ensemble anomaly detection studies have used the heatmap-based visualization techniques to display the results from different algorithms. For example, remix in [26] allows users to navigate the results with a heatmap matrix that groups similar detectors. Schubert et al. [65] also used the similarity matrix to compare various detection algorithms. All these visualization designs are too simple to enable a comprehensive exploration of different anomaly detection algorithms. By contrast, we introduce a visual interactive framework based on ensemble analysis that evaluates anomaly detectors with novel visual representations, rich contextualized information and visual reasoning that assist users in examining and refining their exploration based on their feedback.

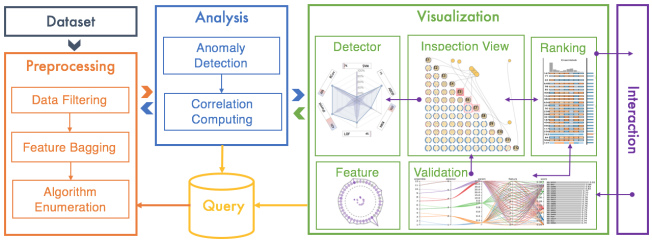

Fig. 2. System overview and data processing pipeline.

To the best of our knowledge, EnsembleLens is the first visual analytics system that compares and evaluates different anomaly detection algorithms for a given dataset in an unsupervised manner. Specifically, we provide a scalable visual design of multi-attributes ranks to improve the evaluation of anomaly detection results derived by anomaly detectors which are diverse and only accurate to a certain extent.

3 System Overview

Our system was designed to meet the real-world requirements for selection of appropriate anomaly detection algorithms when the given data is multidimensional or heterogeneous. We held regular research discussion meetings with domain experts over a four-month period. One domain expert is a project manager in the field of data mining and data visualization. The expert has many years of research experience and publications in analyzing urban data and anomaly detection. He wanted to develop a specific anomaly detection model to monitor air pollution. The other expert is a professor with domain knowledge in visualization and anomaly detection. During these meetings, a variety of design requirements were specified and preliminary designs were assessed. Below, we list the most critical requirements (R1–R4) that guided the design in our system.

- **R1** **The ensemble generation.** Build effective ensembles based on heterogeneous anomaly detection algorithms to facilitate the evaluation of algorithms in terms of the parameter settings or the feature subspaces for a given multidimensional dataset.
- **R2** **Multifaceted comparison of anomaly detection algorithms.** Compare anomaly ensembles in different scales, from the summarization (macro) to the correlation (meso) to the outlier score ranks (micro), thereby allowing a comprehensive understanding of the algorithms.
- **R3** **Interpreting exploration results in context.** Create useful visual designs to help users in comparing alternative anomaly detection algorithms and understanding "when and why some algorithms are good or not" with detailed sub-level information of ensemble components such as pairwise correlation, feature subspaces and algorithm settings.
- **R4** **Human-in-the-loop ensemble analysis.** Due to the lack of the ground truth, users should be able to label the data during the exploration, so that the system can conduct a refined evaluation of anomaly detection algorithms based on user feedback.

Based on these requirements, we have designed EnsembleLens to visually represent anomaly ensembles and yield improved evaluation of anomaly detection algorithms. Fig. 2 illustrates the system architecture and the ensemble-based exploration pipeline. The system consists of four major modules: (1) prepossessing, (2) ensemble analysis, (3) visualization, and (4) interaction modules. The prepossessing module transforms raw data into a multidimensional format. Data filtering, such as removing the attributes that are categorical or show strong discretization effects, is also conducted at this stage. In addition, we determine various feature subspaces for the baseline anomaly detection algorithms (**R1**). The analysis module runs anomaly detection based on ensemble analysis, which not only generates varying outputs of anomaly ranks but also calculates the pairwise correlation and the overall distribution of ensemble components, thus supporting the *level of analysis* comparison (**R2**). The visualization module uses multiple coordinated views to support a comprehensive visual interpretation and reasoning of the ensemble in a multifaceted context (**R3**). The interaction module provides an online, responsive interface to dynamically

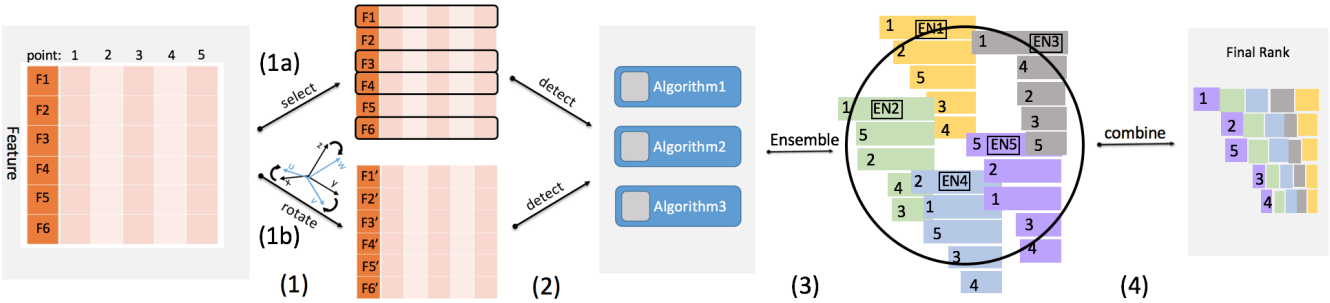

Fig. 3. Anomaly ensemble pipeline: (1) feature selection, (2) algorithm enumeration, (3) ensemble generation, and (4) ensemble combination.

evaluate the anomaly detection algorithms. By using this module, users can label the anomalous points and assign different weights to baseline algorithms in real-time by incorporating their judgment (**R4**).

4 Ensemble Anomaly Detection

In this section, we first introduce the model used in EnsembleLens to achieve the ensemble, which consists of three parts: feature bagging algorithms, baseline anomaly detection algorithms and combination algorithms. Then, we describe a novel method to evaluate different ensembles components in a fully unsupervised fashion.

4.1 Feature Bagging

The first step in ensemble analysis is to construct a feature subspace of multidimensional data, as illustrated in Fig. 3(1). For the table in Fig. 3, each column represents a data point, and each row represents the feature value for each data point. Feature bagging is a popular technique in ensemble learning that samples different features in multivariate data to reduce the variance between anomaly detectors [3]. EnsembleLens implements three feature bagging methods based on two classic data-based ensemble methods [3, 48] and an extended method considering the correlation of features [26].

Random Feature Bagging. This algorithm has two steps, illustrated in 3(1): (1) randomly select one number s from $\lfloor d/2 \rfloor$ to $d-1$, where d is the number of feature dimensions in a given dataset; and (2) randomly select s features from the dataset. This algorithm will have a deteriorated performance when the selected features are highly related.

Non-Redundant Feature Bagging. This method also samples features, but it constructs a feature subspace with less correlation by the following four steps: (1) create a set $\mathscr{F}_p$ of features by calculating the correlation between all pairs of the features and removes the one with max correlation; (2) select the top-l features in $\mathscr{F}_p$ ranked by their Laplacian scores; (3) sample a randomized family of subspace $\mathscr{F}_r$ to maximize the coverage and diversity of the feature subspace; and (4) obtain the non-redundant features $\mathscr{F}_{nr}$ that are the union of $\mathscr{F}_p$ and $\mathscr{F}_r$. The problem of this algorithm is the deteriorated bias characteristics.

Rotated Bagging. This method (Fig. 3(1b)) can reduce variance without compromising bias too much by projecting data to a rotated axis system before conventional feature bagging [48]. The overall algorithm works as follows: (1) determine a randomly rotated axis system in the data; and (2) randomly select $r = 2 + \lceil \sqrt{d}/2 \rceil$ directions from the rotated axis system and project data along these r directions.

4.2 Baseline Anomaly Detection Algorithms

The second step is to enumerate baseline algorithms for the sampled feature subspace (Fig. 3(2)). In order to decide which anomaly detection algorithms should be integrated, two principles are applied: (1) cover typical anomaly detection techniques; and (2) control the whole number of algorithms to maintain system efficiency. By testing existing algorithms and surveying model-based ensemble paper, we choose six representative anomaly detection algorithms from five categories.

One-Class Support Vector Machine (oc-SVM), from classification-based algorithms, uses a hyperplane to distinguish two classes [16]. *RBF* (radial basis function) kernel is used in our system to deal with high-dimensional data, and the kernel coefficient γ is chosen as the adjustable parameter in our system.

K^{th}-Nearest Neighbor (KNN) is a neighbor-based anomaly detection technique that assigns anomaly scores to each data instance based on the k^{th} nearest neighbor [15]. K is the parameter that can be tuned in our system.

Local Outlier Factor(LOF) is also one of the neighbor-based analysis methods, but it is density-based [12]. It determines an outlier instance a by comparing a's k-neighborhood density to the k-neighborhood density of a's k-neighbors. We select K as the parameter.

Angle-Based Outlier Detection (ABOD) is a spectral-based algorithm which identifies anomalous points a by calculating the angle-based outlier factor (ABOF), which is the variances of a point's difference vectors with its k nearest neighbors. K is the parameter. [44].

Robust Covariance Estimation (RCov) is a statistic-based algorithm that assumes the data follow a known distribution (e.g., Gaussian distribution). We use the Mahabolis distances to determine the outlyingness of a point from the known distribution. We set the proportion of points to be included in the support of the raw MCD (minimum covariance determinant) estimate as the parameter for this algorithm.

Isolation Forest (iForest) is a model-based method, which randomly selects a feature and a split value in the range of the selected feature. Then, it isolates data recursively by this step. The shorter the split path is, the more likely the data is an anomaly [50]. The fraction of the samples drawn from data to train each base estimator is the input parameter for this detector in our system.

4.3 Combination Algorithms

Given the outlier scores from different ensemble components (Fig. 3(3)), the last step we need to do is to combine the scores and form a final result, as displayed in Fig. 3(4). There exist many combination methods, such as using the average or maximization scores from different ensemble components. After trials in pilot experiments, we found that the results or importance are imbalanced for different algorithms with different datasets. Therefore, we choose the weighted averaging where the anomaly scores are normalized to [0, 1] before combination as the combination method in our system. For a given point r, its final outlier score is calculated by:

$$\overline{O(r)} = \sum_l w(l) O(r)_l, \sum w(l) = 1,$$

where $w(l)$ is the weight assigned to each ensemble component or detector l, and the $O(r)_l$ is the outlier score for point r in l.

4.4 Anomaly Ensembles Assessment

Considering the unlabelled nature of real-world data, we need reasonable evaluation metrics to assess the anomaly ensembles and their components. We define the following metrics for assessment.

Average Correlation. Judging the similarity or correlation among different rankings of anomaly scores is an important way to compare and evaluate detection results [65]. Three most common types of correlations are Pearson correlation, Kendall rank correlation, and Spearman correlation. However, the Spearman correlation is only correct for a total ranking with no ties, while the Pearson correlation assumes that both variables should be normally distributed. These requirements may not be met by all anomaly ensembles. Thereby, we choose the Kendall rank correlation as the basic method. We formulate the first evaluation score avg_τ of a given anomaly ensemble $\mathbf{r}$ as:

$$avg_\tau(\mathbf{r}) = \frac{\sum_{(\mathbf{j} \neq \mathbf{r}, \mathbf{j} \in \mathscr{E})} tau(\mathbf{r}, \mathbf{j})}{|\mathscr{E}|}, tau(\mathbf{r}, \mathbf{j}) = \frac{p-q}{\sqrt{(p+q+t)(p+q+u)}},$$

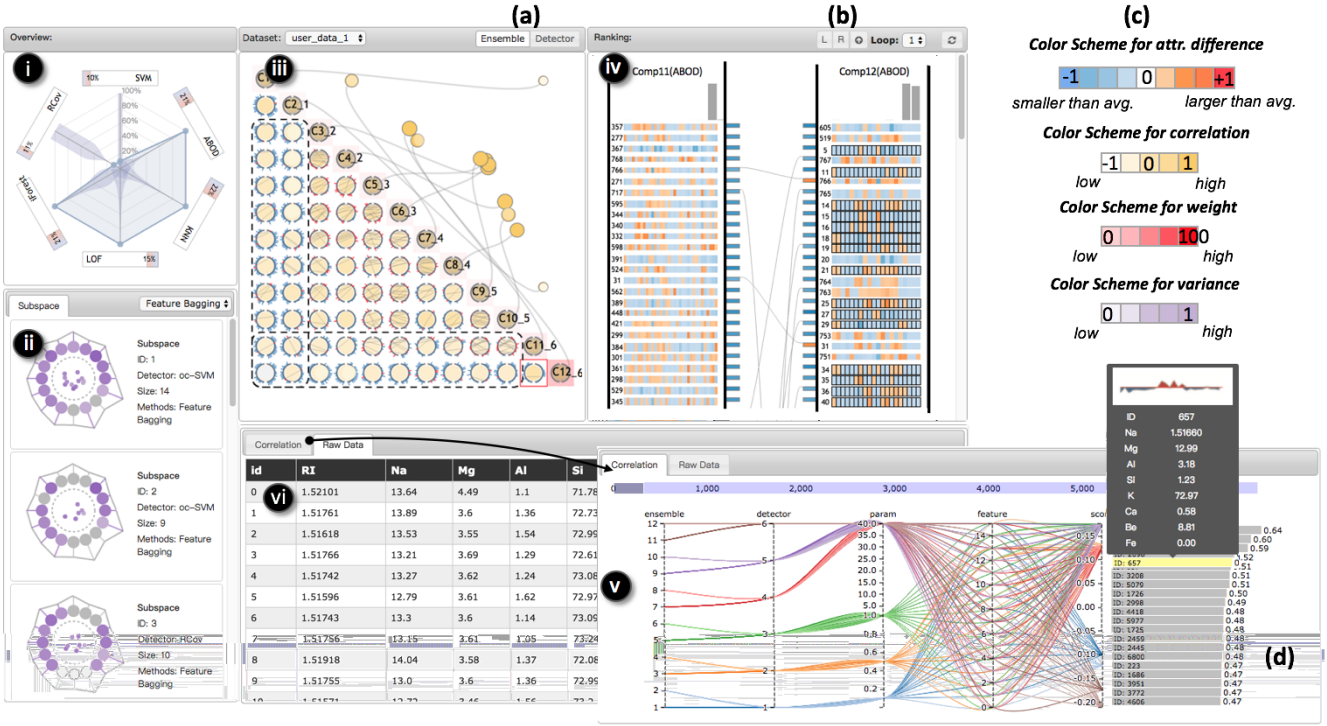

Fig. 4. The EnsembleLens system contains six interactively coordinated views: (i) a detector view, (ii) a feature subspace view, (iii) an inspection view (global inspection view & correlation matrix view), (iv) a ranking view, (v) a validation view and (vi) a raw data table. Users can change the detection mode in (a) and provide their feedback by using (b) after they label the detected anomalous points. The progress of exploration can be reflected by (d) the real-time combination result. Raw data description can be obtained via informative tooltips. (c) is the color schemes used in different views.

where $\mathscr{E}$ is the set of ensemble components, p is the number of concordant pairs, q is the number of discordant pairs, t is the number of ties only in $\mathbf{r}$, and u is the number of ties only in $\mathbf{j}$. If a tie occurs for the same pair in both $\mathbf{r}$ and $\mathbf{j}$, it is not added to either t or u. A reliable ensemble tends to have a higher avg_τ, which shows a strong relationship with other ranks. Moreover, $tau(\mathbf{r},\mathbf{j})$ is also used as the correlation between two ensemble components, $\mathbf{r}$ and $\mathbf{j}$, in our system.

5 Visualization

In this section, we present the design tasks derived from the requirements of visual ensemble anomaly detection and the specific visualizations motivated by these tasks.

5.1 Design Tasks

A list of design tasks were settled to guide the visualization designs based on the requirements outlined in **R1–R4**. In general, a desired visualization system should help users efficiently explore, analyze, and select anomaly detection algorithms with contextualized visual interpretation, comparison and reasoning of ensemble components. In addition, the system should allow experts to update the ensemble components and refine the detection results based on their feedback. To fulfill these requirements, we consolidate the list of visualization design tasks as follows.

- **T1 Show the ensemble overview.** The visual design should clearly reveal the the visual summary of ensemble components generated by the baseline anomaly detection algorithms and their correlations with others to provide explainable overviews of the patterns.
- **T2 Interpret anomaly ensembles in multi-attribute contexts.** In addition to the anomaly scores and ranks, the visualization should be able to present the multidimensional data in the context of corresponding feature values and their statistical information. Such a schematic representation of anomaly patterns helps users develop a mental model to understand various types of normal and abnormal cases, which improves the capacity of users to judge the performance of baseline algorithms and adjust their weights.
- **T3 Facilitate detector comparisons and correlations via ensembles.** The visualization designs should support a comparison among different ensemble components or detectors in both explicit (showing the rank changes) and implicit (showing the statistical value, such as correlation coefficient or similarity) ways.
- **T4 Enhance the visual reasoning of ensemble anomaly detection.** The visualization design and coordinated views in the system should help users possess visual reasoning during their inspection and judgment of anomaly ensembles, in the perspective of either anomaly detector or feature subspace.
- **T5 Allow flexible selection and setting of anomaly detectors.** Users should be able to conveniently select and set the algorithms and their parameters with rich interactions and visual cues indicating the effect of user behaviors.
- **T6 Update evaluation results based on human judgment.** To offer the users a reliable result, the visualization should allow incorporating human judgment. Thus, users will feel satisfied and confident about their detection results.
- **T7 Provide easy access to raw data.** In spite of the significance of the anomaly scores and ranks, the raw data that contain the feature values are also essential for the determination and validation whether a data point is an anomaly of interests.

5.2 User Interface

The aforementioned tasks guided our design of visualization and interaction modules. As shown in Fig. 4, the user interface consists of six major views, which are consistent with the macro-meso-micro exploration workflow, and display the ensemble results with various scales and semantics (**T1, T2, T3**). To support the macro and meso level exploration, a primary inspection view (Fig. 4(iii)) displays an overview of the generated ensemble components (macro) and the correlation between different pairs of ensemble components (meso). At the micro level, a ranking view (Fig. 4(iv)) further compares the detailed ranks between a pair of ensemble components. A validation view (Fig. 4(v)) takes the parallel coordinates design to record each combination result in terms of algorithms and their parameters, used features, and final combined top detected anomalies. After the validation, a novel detector view (Fig. 4(i)) reflects the importance of each baseline algorithm and facilitates the re-setting of these algorithms to update

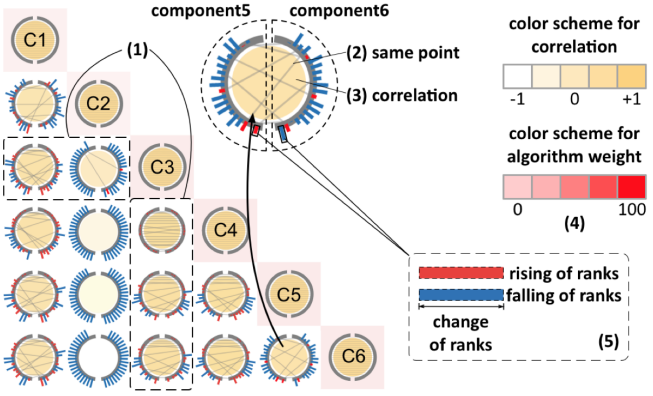

Fig. 5. Visual encoding of the correlation matrix view. (1) The glyph in the matrix is composed of two halves, with each representing an ensemble component in a row or column. The top 20 detected anomaly data points for each component are encoded around the semicircular arc. (2) A line is drawn to link the same point in both components. (3) The background color of the inner circle indicates the correlation of the two paired components. (4) The background color of ensemble component indicated the evaluated weight. (5) The bars perpendicular to the circular arc represent the top 20 detected outliers of the corresponding component, whose color and length are determined by the change of the rank compared with the opposite component.

the exploration result. In addition, there are some coordinated side views. A feature subspace view (Fig. 4(ii)) lists the feature subspaces for each ensemble component based on three feature bagging methods. And a raw data table (Fig. 4(vi)) allows easy access to the raw input dataset. Different color schemes are designed to illustrate the different information (Fig. 4(d)). Basically, all the color schemes are in linear scale, which corresponds to the linear value changes in the statistical information, like correlation and weight. More design details of each view are introduced in the following sections.

Usage Scenario. To understand how these views facilitate the exploration of baseline algorithms, let us consider the following scenario. Suppose Alex is an algorithm engineer who is required to develop a fraud detection algorithm for a bank. He uses EnsembleLens to do some preliminary experiments for a set of alternative algorithms with his financial transactions data. After loading the generated anomaly ensemble into the system, he first observes the inspection view to compare all the ensemble components in general. He finds an outlier, oc-SVM, in the clustering, which also has little correlation with other components. Then he clicks the pair of components of oc-SVM and compares their top-ranked outliers in the ranking view. The dichotomous color scheme used in this view reveals that one ranks from oc-SVM has many anomalous attribute values. Thus Alex checks the raw data of the rank, labels many data points that are indeed outliers and feeds back to the system. The detector view then automatically updates the weights for algorithms. Alex also inspects the feature subspace view to make sure that the selected features for oc-SVM's components are not significantly different. He continuously conducts the exploration until the weights distribution tends to be stable. All the generated ensemble components and their detailed settings are recorded in the validation view during the exploration, and can be later retrieved for more in-depth analysis.

5.3 Inspection View

The inspection view (Fig. 4 (iii)) contains two sub-views: (1) the global inspection view displays the overall clustering of different ensemble components at the top right (**T1**); (2) the correlation matrix view depicts the correlation between different pairs of ensemble components at the bottom left (**T3**). They support the macro and meso level exploration of anomaly detection algorithms, respectively.

The global inspection view and the correlation matrix were divided by the circles on the diagonal. Each circle on the diagonal represents one ensemble component. The background color of the circle (ranging from light to dark red) encodes the weight of the component based on its importance for the explored dataset (Fig. 5(4)). Here, light red indicates a low weight and dark red indicates the opposite. To eliminate the bias of different anomaly detection algorithms, we provide two outlier scores by using two different feature subspaces for each anomaly detection algorithm (e.g., $C1$, $C2$ are two ensemble components for oc-SVM). A total of twelve ensemble components ($C1$–$C12$) corresponding to six anomaly detection algorithms are demonstrated by default.

5.3.1 Global Inspection View

At the macro level, we use a novel layout algorithm to show the overview of multiple ensemble components by clustering analysis (**T1**). Each circle in the top right area (Fig. 4(iii)) represents one ensemble component and is linked with the circle on the diagonal, the size of which represents the average correlation with other ensemble components (ranging from -1 to 1). Also, the average correlation is indicated by the filling color ranging from white to yellow, with -1 shown in white and 1 shown in yellow (Fig. 4(c)). The distance between two components shows the similarity. From this view, we can clearly find the overall distribution of different ensemble components.

Layout Algorithm. We place the ensemble components in a 2D triangle space. We first construct a vector for each component based on their outlier scores. Then, we compute the similarity between each pair of components based on the Euclidean distance of the ensemble vectors. For the three components that have the least average similarity with others, we fix their positions in the right triangular vertices, with the leg length proportional to their correlation value between each other. Next, for the remaining ensemble components, like $\mathbf{e}$, we calculate their position based on the Barycentric coordinate system:

$$\mathbf{e} = \alpha_1 \mathbf{e_1} + \alpha_2 \mathbf{e_2} + \alpha_3 \mathbf{e_3}, \alpha_1 + \alpha_2 + \alpha_3 = 1,$$

where $\mathbf{e_1}, \mathbf{e_2}$, and $\mathbf{e_3}$ are three points in the vertices, and α_k is the similarity between $\mathbf{e}$ and $\mathbf{e_k}$. Then the position of each ensemble can be obtained by the Barycentric coordinates $\mathbf{e_1}, \mathbf{e_2}$, and $\mathbf{e_3}$.

5.3.2 Correlation Matrix View

At the meso-level, to further explore the differences among ensemble components in performance, we design a matrix view on the bottom left to show the correlation between different paired components (Fig. 5)).

Correlation Glyph. Each circular glyph $G(i, j)$ (Fig. 5(1)) in i^{th} row and j^{th} column reveals the comparison of C_{i+1} with C_j in terms of their correlation values and the top 20 detected anomaly data points (**T3**). The correlation values are calculated by Kendall's tau (See Section.4.4) correlation and encoded by the filling color of the circle (Fig. 5(3)). The color scheme for correlation is consistent with the global inspection view (Fig. 4(c)). To display the relationship between two components intuitively, the glyph is evenly divided into two parts with the left accounting for C_j and right accounting for C_{i+1}. On each side of the circular glyph, 20 bars are perpendicular to the circular arc, representing the top 20 detected outliers of the corresponding ensemble component. As shown in Fig. 5(5), two colors are used in this glyph to show the direction of a data point's rank change compared with the opposite side; red indicates a rising one and blue indicates a falling one. The length of the bar is determined by the amplitude of rank change. If one data point is detected among the top 20 anomalies by both ensembles, a line that links its position on both sides will be drawn (Fig. 5(2)). The reason to select the top 20 is that the top detected outliers usually represent the most significant ones to evaluate an anomaly detection algorithm [65]. From this circular glyph, users can clearly understand the multi-level correlations between different ensemble components.

In addition, we provide an extra "Detector" mode to help find prior algorithms more directly (Fig. 4(a)). The outlier scores of each component (i.e., a detector) in this mode are computed by averaging all the outlier scores from this detector. And the result detected in this mode is consistent with the "Ensemble" mode.

To alleviate the workload of one-by-one comparison, our system automatically highlights the recommended glyphs with black strokes based on the correlation between the corresponding paired components.

5.4 Ranking View

At the micro level, the ranking view (Fig. 4(iv)), mainly serves for **T3** and **T5**, is designed for users to explore the detailed relationships between two ensemble components from the inspection view. This view, following a "barcode" metaphor, is implemented to show a variety of

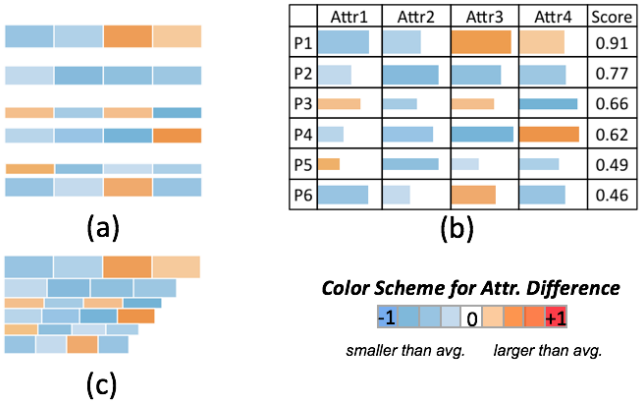

Fig. 6. Ranking list in our system and two alternatives. Each row represents a data point, and the rectangles in each row represent the attributes that use the same color scheme. Outlier scores are encoded differently. Our design (a) uses the gap between two rows to encode the score difference; (b) uses the text to record the outlier score; and (c) uses the row width to indicate the outlier score.

related information to help users judge which ensemble component is more reasonable in the perspective of the outlier scores and rankings. Five types of information are encoded in this view. **(1) Outlier score rank:** two lists of outlier scores that represent the ensemble components (C_j & C_{i+1}) will be generated in the descending order when the corresponding glyph $G(i, j)$ is clicked in the inspection view. Each row stands for one data point with the id in the front, and the width of the gap between two bars encodes the outlier score difference (the wider the gap goes, the larger the score difference is). **(2) Data feature:** within each row, we use different columns to depict different features of raw data. The feature value is indicated by the dichotomous color scheme ranging from blue, to white to red (Fig. 4(d)), with red/blue encoding a value higher/lower than the average (white). The darker it goes, the higher/lower the value is [14]. **(3) Rank stability:** the height of each row is encoded to express the rank variance of each data point among all ensemble components. The larger the more unstable. **(4) Outlier score distribution:** on the top of each ranking list, a histogram is drawn to show the outlier score distributions of the corresponding ensemble component. **(5) Raw data:** the raw data value of each feature is provided by tooltips (Fig.4(e)). **(6) Others:** In each ranking list, we display the top 100 ranked data points due to the space limitation. The same point in the two paired ranking lists is linked by a line in case no corresponding data point is found in the opposite list. An additional rectangle beside each row is set to show the rank change, with red encoding increasing and blue decreasing in ranks. These visual design, together with rich interactions (described in Section 5.7), helps users compare two ensemble components by their ranks.

Ranking List Alternatives. We consider several design alternatives as shown in Fig. 6. Our first design alternative is to use the length of each row to encode the entire outlier score (Fig. 6(c). However, some data points have a relatively low outlier score, making the their feature information cannot be clearly shown because the row width is too small. Our second design alternative uses tabular design (Fig. 6(b)), in which each column represents one feature, and the bar width encodes the value of the feature. In this design, the last column is used to show the outlier score. However, some datasets with many features cannot be distinctly shown if we assign each column a narrow width. The rank order and the relative score difference are essential for rank comparison. In summary, both alternatives have the scalability problem.

5.5 Validation View

A parallel coordinates view in Fig. 4(v) is designed to describe the general relationships among the generated ensemble components, which helps in determining the final combination results. This view records all the ensemble components and their corresponding algorithms, parameters, and feature subspaces. Each line records one ensemble component with a categorical color. The ranking list (Fig. 4(f)) displays the anomalous points based on a weighted combination result of these components (**T2, T4**). During the exploration, Fig. 4(f) will change accordingly. The data labeled by humans are also highlighted in this ranking list to help decide whether they can finish the exploration by judging whether the combined ranking list is stable.

5.6 Additional Contextual Views

Several additional views are developed to assist users in exploring and selecting anomaly detection algorithms.

Detector View. This view (Fig. 4(i)) uses a novel hexagon to show the weight of the six baseline algorithms (SVM, RCov, iFores, LOF, KNN and ABOD) during the exploration process. It is also a parameter tuning tool for the algorithms (**T5**). Users can understand the importance of each algorithm with references to this calculated weight, which is displayed on each side of the hexagon. In addition, users can tune the parameter of each algorithm by dragging the nodes in the hexagon. All the views are initialized with a value computed on the back-end and will be updated based on the new parameters. The shaded areas along each diagonal reveal the sensitivity intervals of each parameter. Specifically, the wider the area is, the more sensitive the algorithm is when tuning the parameter around this value.

Feature Subspace View. The feature subspace view (Fig. 4(ii)) shows the selected features for each ensemble component to provide a visual reasoning for the ensemble analysis results (**T4**). Each sub-view shows the features selected and method used towards the component. In each polygon, the nodes representing specific features will be colored if they are selected in this feature subspace, and the opacity of color encodes its variance, ranging from dark purple (highest) to light purple (lowest). Unselected features are encoded by the grey dots. In addition, the relationships among all the selected features in each ensemble are indicated by MDS and displayed in the middle of the polygon.

Raw Data Table. We also integrated raw data into a table for users to refer to during the exploration and inspection (Fig. 4(vi)).

5.7 Interactions

We adopt the following interactions to support efficient exploration and selection of anomaly algorithms among different views.

Querying and Filtering. Users can load different datasets via the query box in the top of the inspection view and query newly generated ensemble components by tuning the parameter in the detector view. In the validation view, the parallel coordinate view supports the exploration of specific relationships between input setting and output ensemble via brushing and filtering the axes. The rank slider (shown in the top of Fig. 4(v)) can be used to focus the combination result on selected ranks of interest.

Tooltips and Highlights. When users hover over the bar on the ranking view or validation view, the associated raw data information will be shown on the tooltip (Fig. 4(d), **T7**). Moreover, by hovering on one glyph in the correlation matrix view, users can see the emphasized links connecting the corresponding components in the overview. The counter-wise highlighting also stands. Similarly, in the feature subspace view, the nodes representing different features will be highlighted together in the polygon and inner MDS plot.

Data Labeling. In the ranking view, users can label the detected anomalous points by clicking the corresponding bars (**T6**). Then the bar will be highlighted, and these data points will be stored and highlighted in the other ranking lists whenever they appear. The system will immediately update the weights of each baseline algorithms based on the labeling results via clicking the button in Fig. 4(b).

Zooming and Scaling. The inspection view supports zooming and panning to obtain a clearer view of the correlation glyphs. The ranking list can be scaled by dragging the line between two lists to see the whole features, especially when there are too many features.

6 Evaluation

We evaluate the usefulness and usability of our system in multiple methods. First, we describe three case studies, where EnsembleLens is applied to three real-world datasets from the UCI Machine Learning repository. These datasets have distinct application domains: (1) the Wisconsin-Breast Cancer (Original) dataset, (2) the Glass Identification dataset, and (3) the QSAR Biodegradation dataset. The three case studies have different data types, and use feature bagging methods. For each case study, a quantitative evaluation is also used to further evaluate

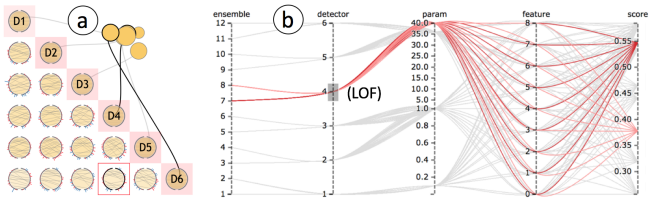

Fig. 7. Exploration result of the Glass dataset. (a) shows the similar correlation among all the detectors. (b) displays that the results of LOF are sensitive to its parameter setting.

the effectiveness of our exploration results. Finally, we gather feedback on our system from two experts in data mining domain.

6.1 Case Study I: Analysis of Breast Cancer Dataset

In the first case study, the dataset contains the inspection results of breast cancer such as clump thickness or uniformity of cell size. The dataset contains 699 instances with 9 attributes, and all the attributes are integers ranging from 1 to 10. Moreover, the dataset has been classified into two classes with a 34% outlier percentage, where we consider the *malignant* class as outliers and the *benign* as inliers. To reduce bias, we use the random feature bagging methods to extract two feature subspaces for each anomaly detection algorithm. Thus, we get an ensemble with 12 components in total.

Visual exploration and results. Following the macro-meso-micro analytics workflow, the first feature that attracts our attention is the overall distribution of the ensemble components, most of which are clustered into one large category except two from LOF and ABOD (Fig. 1(a)). The finding is also verified by the correlation matrix view. The paired correlation glyphs from LOF and ABOD have many blue bars surrounding the circle (Fig. 1(b)), and the middle circle in these glyphs is a light yellow, indicating that LOF and ABOD generally have little correlation with other algorithms. Moreover, we inspect the detailed ranks generated by these two algorithms in the ranking view. Obvious patterns are also discovered. As shown in Fig. 1(c2), most of the top-ranked instances detected by LOF have a consistent distribution of attribute values, which indicates they are very likely to be the normal data points. Nevertheless, they have a large bar height, which means these points possess a high variance in ranking among different ensemble components, and the rank is not stable. As revealed by our visual cues, we find most detected outliers from these two ensemble components are incorrect after checking the raw data via the tooltip. Thereby, we mark the findings and then the weight (importance) of LOF is reduced accordingly. ABOD has a similar situation.

Furthermore, we notice a special case in the correlation matrix view (Fig. 1(b)), where the two ensemble components ($C5\_3$ & $C6\_3$) from iForest have the same outlier score ranking. We click the glyph and find that most of the top-ranked instances in these two components are actually outliers (Fig. 1(c1)). We label the data and update the findings again, and the weight of iForest increases drastically. After several additional loops, the weight of each detector tends to be stable. Finally, we complete the adjustment as the combined results in Fig. 1(e) are the expected outliers. The weights of each algorithm are: oc-SVM (7), RCov (20), iForest (43), LOF (4), KNN (35) and ABOD (1) (Fig. 1(d)). The reason that iForest performs best is not only because of its high average accuracy or correlation with other algorithms, but also that the top-ranked outliers from its ensemble components are highly correct, consistent and stable during the exploration. In addition, we inspect the feature subspace view and find that each algorithm has diverse feature subspaces. Therefore, for RCov, iForest, LOF and KNN, which have few differences when their parameters are changed, their performances are thought to be mainly determined by the data characteristics and integer attribute value. For the remaining algorithms, their performances may also be affected by the parameter setting.

Quantitative validation. We validate our exploration results through a comparison of the ROC curves with the six baseline algorithms. We also provide an ensemble method (denoted as "EN") based on our exploration result, which is a combination of the baseline algorithms with an assigned weight for each detector. The assigned weights and the feature bagging methods (i.e., random feature bagging) are consistent with our case study. As shown in Fig. 8(a), the ROC plot validates the exploration result in our case study, where iForest and EN have a higher true positive rate than the others when the false positive rates remain low (below 0.1). KNN and RCov have a medium result while LOF, ABOD and oc-SVM perform worst.

6.2 Case Study II: Analysis of Glass Dataset

In our second case study, we apply EnsembleLens to the Glass dataset containing 10 attributes (the volumes of chemical composition such as"Na", "Mg" and "Al") with real values. The glasses are classified into seven classes based on their types. We choose class 5 as the outlier class, thereby the outlier percentage is 6.1% (13/214). We apply the rotated feature bagging to build the feature subspace because all the chemical compositions are important for judging the class type.

Visual exploration and results. To get all the attributes involved, we choose the 'Detector' mode and explore the Glass dataset directly in the perspective of anomaly detection algorithms. Both the clustering view and correlation matrix view show that the results from different detectors are similar to each other (Fig. 7(a)). Specifically, there are many crossing lines in each glyph in the matrix view, and rank changing bars around glyphs are relatively low, which indicates that the outlier score rank of a given instance is not excessively changed in different detectors. Then, we inspect the parallel coordinates view (Fig.4(5)), in which we find an interesting case for LOF. As shown in Fig. 7(b), only the average correlation of LOF was affected largely by its parameter, whereas the others have similar rank scores across different parameter settings. Therefore, we tune the parameter of LOF for several times and check the outliers in ranking view. Although the rank calculated by LOF changes significantly itself, the general positive truth rate had little fluctuation. Finally, we stop our exploration when the result tends to be stable. The inspection view shows that all the algorithms have similar and stable performances in terms of both their correlation with others and their top-ranked data points from the same detector. Therefore, we suspect that the small size of the data is the main reason for the result.

Quantitative validation. We validate our exploration in a similar way in case study I. We compare the six baseline results and an ensemble method based on the weight for each detector obtained by our system and the same rotated feature bagging method. As shown in Fig. 8(b), the result is consistent with our exploration result. All the algorithms, including our ensemble method, have a close ROC curve.

6.3 Case Study III: Analysis of Biodegradation Dataset

In our third case study, we analyze a dataset that has forty-one molecular attributes to study the relationships between chemical structure and biodegradation of the molecule. Originally, it contains ready (356) and not ready (699) biodegradable molecules. We further down-sample the ready class to 71 and consider them as outliers [3]. The outlier percentage is 9%. Considering the high dimensional attributes, we adopt the non-redundant feature bagging method which can extract the most unrelated features for anomaly detection. Still, we generate two feature subspaces for each detector to reduce bias.

Visual exploration and results. After loading the data, we immediately find that the ensemble components have a sparse distribution (Fig. 4(iii)). Then, we inspect the correlation matrix and distinguish four components that have little correlation with the others, which are generated by oc-SVM and ABOD. Most interestingly, although three of them are barely correct, the top-ranked outlier points detected by one component ($C12\_6$) from ABOD have a very high true positive rate (Fig.4(iv)). Hence, we label the correctly detected points in the ranking view and the weight of ABOD increases. After several loops, we obtain a stable result for the assignment of weights for baseline algorithms, where ABOD is the best detector for high dimensional data, whereas the others except oc-SVM have a similar but moderate performance. The final result is: oc-SVM (8), RCov (12), iForest (12), LOF (11), KNN (12) and ABOD (45). Although ABOD has low correlation with the others, it is still the best. The result is reasonable because ABOD usually has a better performance when the data dimension is high. We further analyze the reason why the two components from ABOD have totally different results. We observe the feature subspace view and find an obvious contrast. The feature subspace for the component with high accuracy ($C12\_6$) contains fewer features than the others. This indicates that some features might worsen the anomaly detection result, and sometimes fewer subspaces are better for detecting outliers.

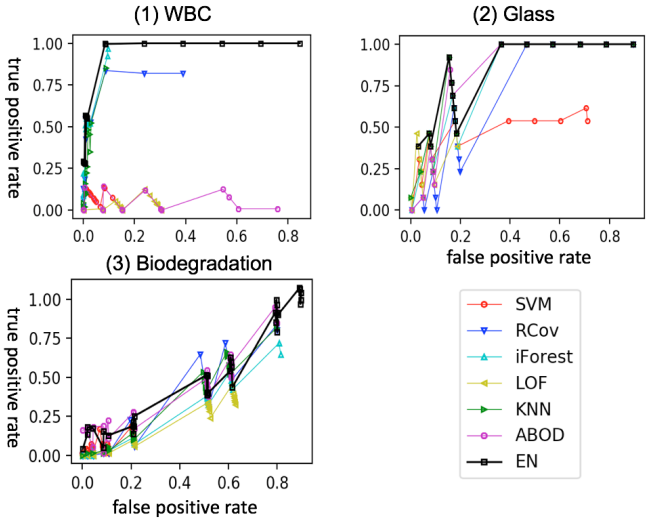

Fig. 8. Quantitative validation of three exploration results based on EnsembleLens with datasets from three different domains.

Quantitative validation. Fig. 8(c) shows that the performance for all algorithms except ABOD is not good, especially when the false negative rate is low, which validates our exploration in EnsembleLens.

6.4 Expert Interview

We conducted semi-structured interviews with two experts to gather qualitative feedback on the usability and usefulness of our system. The first expert (**E1**) is a project manager who is leading a project about personalized air quality and health management visualization system. The second expert (**E2**) has highly relevant expertise in knowledge discovery in databases. Both of them had never used EnsembleLens before the interview. For each interview, we started with an introduction about the purpose of EnsembleLens and the functions of each view, followed by a tutorial with the Breast Cancer dataset. Then **E1** explored our system to search for appropriate anomaly detection algorithms using his own data (one-year of Hong Kong air pollution records), and **E2** was asked to use EnsembleLens with the Biodegradation dataset. More details about the interviews are provided in the supplementary material[1]. Each interview lasted about 1.5 hours, during which notes were taken, and we summarized them as follows.

System. Both experts regarded this system easy-to-use and powerful. "I have never imagined that multiple [anomaly detection] results of my data can be compared in such an easy way!" **E1** commended. **E2** confirmed the effectiveness of our system as ABOD is among the best algorithm for the Biodegradation dataset. He mentioned that both oc-SVM and ABOD are useful for high dimensional data, and explained why oc-SVM works worst in this situation, "oc-SVM usually performs well when the inliers have multiple modes [clusters]." However, the inliers of the Biodegradation dataset is unimodal with one big cluster of data. **Visualization.** The visualization designs have generally met the design tasks. **E1** appreciated the inspection view due to its informative visualization design. By observing this view in his exploration, he noted that KNN, LOF and ABOD often have high correlation and he believed "this is reasonable" as the aforementioned algorithms are all based on KNN. **E1** also liked the ranking view, where anomalous values of attributes, like O_3 and NO_2, are obviously revealed as outliers. Both experts felt the correlation glyph is novel and useful. **Adaption of Workflow.** The experts were able to utilize the capabilities of the system in their analytics workflow.. For example, following the method of macro-meso-micro visual analytics, **E1** detected that ABOD actually had a sound performance because many of its top-20 ranked outliers were consistent with those detected by the others. Previously, he regarded ABOD as "bad" due to its low-average correlation with regular algorithms like LOF or KNN. Furthermore, following the validation and reasoning step, **E2** could efficiently refine the evaluation results by tuning the parameters or inspecting the feature subspaces.

[1] `https://lukeluker.github.io/supp/vast18_ensemblelens.pdf`

7 Discussions

There are some advantages and limitations when EnsembleLens helps data mining experts to evaluate the anomaly detection algorithms prior to model deployment. Compared with accuracy-based evaluation, EnsembleLens can evaluate algorithms in an unsupervised fashion by using the weight to indicate the overall importance. The weight is evaluated by a comprehensive comparison of ensemble components, from the overview clustering and the pair correlation to the top-ranked outliers' inspection, and will be continuously refined based on user feedback during the exploration. Even when the accuracy is available with the labeled data, EnsembleLens can provide interpretation and reasoning for the results. For example, the case studies suggest that the performance of anomaly detection algorithms can be affected by three factors, namely, the algorithm setting, the data characteristics (application domain, dimension and attribute type), and the feature subspace. Specifically, case study III shows that ABOD tends to be more effective when the data dimension is high, and the feature subspace can also result in different performances for the same algorithm. Although we can provide reasoning from the feature subspace view, one limitation of our system is that users are unable to construct the feature subspace for the algorithm. Currently, the feature subspace can only be changed at the back end. One expert we interviewed mentioned another limitation that our system does not provide enough automated recommendations for the suspected ensemble components and data points during the exploration, which could save the efforts of users further.

EnsembleLens can also be extended to help joint human-machine decisions when the anomaly detection algorithms are deployed. Although this system focuses on the exploration of algorithms, the ensemble model and the analytics workflow we proposed can be widely applicable. Specifically, the construction of anomaly ensembles integrates the methods of feature bagging, algorithm enumeration and ensemble combination, which allow the ensembles to be customized or optimized at any stage. When generating the anomaly ensembles, the macro-meso-micro analytics workflow provides a comparison of ensemble components in different scales based on our visual representations, especially the correlations among different components, which assists users in selecting the most meaningful or suspected components to check their outlier ranking. Finally, our system will update the anomaly ensembles and generate refined results of outlier ranking with response to user feedback. In this sense, our system can be used for interactive anomaly detection. However, the main concern is that EnsembleLens is designed from the perspective of algorithm interpretation but not data interpretation. Therefore, there should be more visualization designs to reveal the inherent characteristics or patterns in the data.

8 Conclusion and Future Work

We have presented EnsembleLens, a novel visual exploration system designed to evaluate the performance of different anomaly detection algorithms based on ensemble analysis. EnsembleLens incorporates the macro-meso-micro analytics method to promote an in-depth inspection and comparison of anomaly ensembles with multidimensional data. We proposed multiple coordinated views with rich interactions to support the exploration. A novel visualization design is used to illustrate the correlation between paired ensemble components. The system can update the evaluation results iteratively based on human feedback. We evaluated EnsembleLens through three real-world datasets from three different domains and conducted interviews with experts in the data mining area. These results demonstrated that our design can be used to evaluate the importance of heterogeneous algorithms for a given dataset and assign proper weight to each of them. In future, we intend to supplement EnsembleLens with abilities that allow users to customize feature subspaces for more reasoning analysis. Moreover, we will also conduct additional experimental studies with domain experts' data to obtain more insightful evaluation of the usability of our system.

9 Acknowledgements

We would like to thank all the reviewers and domain experts for their comments. This work is part of the research supported by HSBC 150th Anniversary Charity Program, NFSC Grants-61602306, and the Fundamental Research Funds for the Central Universities.

REFERENCES

[1] C. C. Aggarwal. Outlier ensembles: position paper. *ACM SIGKDD Explorations Newsletter*, 14(2):49–58, 2013.

[2] C. C. Aggarwal. Outlier analysis. In *Data Mining: The Textbook*, pp. 237–263. Springer International Publishing, 2015.

[3] C. C. Aggarwal and S. Sathe. Theoretical foundations and algorithms for outlier ensembles. *ACM SIGKDD Explorations Newsletter*, 17(1):24–47, 2015.

[4] F. Angiulli and C. Pizzuti. Fast outlier detection in high dimensional spaces. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 15–27. Springer, 2002.

[5] M. Anneken, Y. Fischer, and J. Beyerer. Evaluation and comparison of anomaly detection algorithms in annotated datasets from the maritime domain. In *SAI Intelligent Systems Conference (IntelliSys), 2015*, pp. 169–178. IEEE, 2015.

[6] B. Auslander, K. M. Gupta, and D. W. Aha. A comparative evaluation of anomaly detection algorithms for maritime video surveillance. In *Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense X*, vol. 8019, p. 801907. International Society for Optics and Photonics, 2011.

[7] V. Barnett and T. Lewis. *Outliers in Statistical Data*. Wiley, 1974.

[8] S. D. Bay and M. Schwabacher. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In *Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 29–38. ACM, 2003.

[9] M. Behrisch, J. Davey, S. Simon, T. Schreck, D. Keim, and J. Kohlhammer. Visual comparison of orderings and rankings. In *EuroVis*, 2013.

[10] D. C. Blest. Theory & methods: Rank correlationan alternative measure. *Australian & New Zealand Journal of Statistics*, 42(1):101–111, 2000.

[11] A. Bock, A. Pembroke, M. L. Mays, L. Rastaetter, T. Ropinski, and A. Ynnerman. Visual verification of space weather ensemble simulations. In *Scientific Visualization Conference (SciVis), 2015 IEEE*, pp. 17–24. IEEE, 2015.

[12] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: identifying density-based local outliers. In *ACM Sigmod Record*, vol. 29, pp. 93–104. ACM, 2000.

[13] N. Cao, C. Lin, Q. Zhu, Y.-R. Lin, X. Teng, and X. Wen. Voila: Visual anomaly detection and monitoring with streaming spatiotemporal data. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):23–33, 2018.

[14] N. Cao, Y.-R. Lin, D. Gotz, and F. Du. Z-glyph: Visualizing outliers in multivariate data. *Information Visualization*, 17(1):22–40, 2018.

[15] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):15, 2009.

[16] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[17] H. Chen, S. Zhang, W. Chen, H. Mei, J. Zhang, A. Mercer, R. Liang, and H. Qu. Uncertainty-aware multidimensional ensemble data visualization and exploration. *IEEE Transactions on Visualization and Computer Graphics*, 21(9):1072–1086, 2015.

[18] D. Dasgupta and N. S. Majumdar. Anomaly detection in multidimensional data using negative selection algorithm. In *Evolutionary Computation, CEC'02. Proceedings of the 2002 Congress on*, vol. 2, pp. 1039–1044. IEEE, 2002.

[19] I. Demir, C. Dick, and R. Westermann. Multi-charts for comparative 3d ensemble visualization. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2694–2703, 2014.

[20] A. Diehl, L. Pelorosso, C. Delrieux, K. Matković, J. Ruiz, M. E. Gröller, and S. Bruckner. Albero: A visual analytics approach for probabilistic weather forecasting. In *Computer Graphics Forum*, vol. 36, pp. 135–144. Wiley Online Library, 2017.

[21] T. G. Dietterich. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pp. 1–15. Springer, 2000.

[22] B. Duffy, J. Thiyagalingam, S. Walton, D. J. Smith, A. Trefethen, J. C. Kirkman-Brown, E. A. Gaffney, and M. Chen. Glyph-based video visualization for semen analysis. *IEEE Transactions on Visualization and Computer Graphics*, 21(8):980–993, 2015.

[23] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo. A geometric framework for unsupervised anomaly detection. In *Applications of Data Mining in Computer Security*, pp. 77–101. Springer, 2002.

[24] X. Z. Fern and W. Lin. Cluster ensemble selection. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 1(3):128–141, 2008.

[25] F. Ferstl, M. Kanzler, M. Rautenhaus, and R. Westermann. Visual analysis of spatial variability and global correlations in ensembles of iso-contours. In *Computer Graphics Forum*, vol. 35, pp. 221–230. Wiley Online Library, 2016.

[26] Y. Fu, C. Aggarwal, S. Parthasarathy, D. S. Turaga, and H. Xiong. Remix: Automated exploration for interactive outlier detection. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 827–835. ACM, 2017.

[27] J. Gao and P.-N. Tan. Converting output scores from outlier detection algorithms into probability estimates. In *Sixth International Conference on Data Mining, 2006*, pp. 212–221. IEEE, 2006.

[28] J. Ghosh and A. Acharya. Cluster ensembles: Theory and applications. pp. 551–570. Citeseer, 2013.

[29] N. Goix. How to evaluate the quality of unsupervised anomaly detection algorithms? *arXiv preprint arXiv:1607.01152*, 2016.

[30] M. Goldstein and S. Uchida. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS one*, 11(4):e0152173, 2016.

[31] S. T. Hadjitodorov, L. I. Kuncheva, and L. P. Todorova. Moderate diversity for better cluster ensembles. *Information Fusion*, 7(3):264–275, 2006.

[32] L. K. Hansen and P. Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001, 1990.

[33] L. Hao, C. G. Healey, and S. A. Bass. Effective visualization of temporal ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):787–796, 2016.

[34] V. Hautamaki, I. Karkkainen, and P. Franti. Outlier detection using k-nearest neighbour graph. In *ICPR 2004. Proceedings of the 17th International Conference on Pattern Recognition, 2004*, vol. 3, pp. 430–433. IEEE, 2004.

[35] S. Hawkins, H. He, G. Williams, and R. Baxter. Outlier detection using replicator neural networks. In *International Conference on Data Warehousing and Knowledge Discovery*, pp. 170–180. Springer, 2002.

[36] L. He, Z. Li, and C. Shen. Performance evaluation of anomaly-detection algorithm for keystroke-typing based insider detection. In *Proceedings of the ACM Turing 50th Celebration Conference-China*, p. 32. ACM, 2017.

[37] V. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004.

[38] A. Inselberg and B. Dimsdale. Parallel coordinates for visualizing multi-dimensional geometry. In *Computer Graphics 1987*, pp. 25–44. Springer, 1987.

[39] W. Jin, A. K. Tung, and J. Han. Mining top-n local outliers in large databases. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 293–298. ACM, 2001.

[40] F. Keller, E. Muller, and K. Bohm. Hics: high contrast subspaces for density-based outlier ranking. In *IEEE 28th International Conference on Data Engineering, 2002*, pp. 1037–1048. IEEE, 2012.

[41] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[42] E. M. Knorr, R. T. Ng, and V. Tucakov. Distance-based outliers: algorithms and applications. *The VLDB JournalThe International Journal on Very Large Data Bases*, 8(3-4):237–253, 2000.

[43] H.-P. Kriegel, P. Kroger, E. Schubert, and A. Zimek. Interpreting and unifying outlier scores. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, pp. 13–24. SIAM, 2011.

[44] H.-P. Kriegel, M. S hubert, and A. Zimek. Angle-based outlier detection in high-dimensional data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 444–452. ACM, 2008.

[45] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[46] R. Kumar and S. Vassilvitskii. Generalized distances between rankings. In *Proceedings of the 19th International Conference on World Wide Web*, pp. 571–580. ACM, 2010.

[47] A. Kumpf, B. Tost, M. Baumgart, M. Riemer, R. Westermann, and M. Rautenhaus. Visualizing confidence in cluster-based ensemble weather forecast analyses. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):109–119, 2018.

[48] A. Lazarevic and V. Kumar. Feature bagging for outlier detection. In *Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pp. 157–166. ACM, 2005.

[49] H. Lin, S. Gao, D. Gotz, F. Du, J. He, and N. Cao. Rclens: Interactive rare category exploration and identification. *IEEE Transactions on Visualization and Computer Graphics*, 2017.

[50] F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation forest. In *Eighth IEEE International Conference on Data Mining, 2008*, pp. 413–422. IEEE, 2008.

[51] M. V. Mahoney and P. K. Chan. Learning rules for anomaly detection of hostile network traffic. In *Third IEEE International Conference on Data*

Mining, 2003., pp. 601–604. IEEE, 2003.

[52] E. Müller, M. Schiffer, and T. Seidl. Statistical selection of relevant subspace projections for outlier ranking. In *IEEE 27th International Conference on Data Engineering, 2011*, pp. 434–445. IEEE, 2011.

[53] H. V. Nguyen, H. H. Ang, and V. Gopalkrishnan. Mining outliers with ensemble of heterogeneous detectors on random subspaces. In *International Conference on Database Systems for Advanced Applications*, pp. 368–383. Springer, 2010.

[54] T. Nocke, M. Flechsig, and U. Böhm. Visual exploration and evaluation of climate-related simulation data. In *Proceedings of the 39th Conference on Winter Simulation: 40 years! The Best is yet to Come*, pp. 703–711. IEEE Press, 2007.

[55] H. Obermaier and K. I. Joy. Future challenges for ensemble visualization. *IEEE Computer Graphics and Applications*, 34(3):8–11, 2014.

[56] G. H. Orair, C. H. Teixeira, W. Meira Jr, Y. Wang, and S. Parthasarathy. Distance-based outlier detection: consolidation and renewed bearing. *Proceedings of the VLDB Endowment*, 3(1-2):1469–1480, 2010.

[57] K. Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895.

[58] J. Poco, A. Dasgupta, Y. Wei, W. Hargrove, C. Schwalm, R. Cook, E. Bertini, and C. Silva. Similarityexplorer: A visual inter-comparison tool for multifaceted climate data. In *Computer Graphics Forum*, vol. 33, pp. 341–350. Wiley Online Library, 2014.

[59] K. Potter, P. Rosen, and C. R. Johnson. From quantification to visualization: A taxonomy of uncertainty visualization approaches. In *Uncertainty Quantification in Scientific Computing*, pp. 226–249. Springer, 2012.

[60] K. Potter, A. Wilson, P.-T. Bremer, D. Williams, C. Doutriaux, V. Pascucci, and C. R. Johnson. Ensemble-vis: A framework for the statistical visualization of ensemble data. In *IEEE International Conference on Data Mining Workshops, 2009*, pp. 233–240. IEEE, 2009.

[61] C. Preisach and L. Schmidt-Thieme. Ensembles of relational classifiers. *Knowledge and Information Systems*, 14(3):249–272, 2008.

[62] S. Rayana and L. Akoglu. Less is more: building selective anomaly ensembles. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 10(4):42, 2016.

[63] L. Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2):1–39, 2010.

[64] J. Sanyal, S. Zhang, J. Dyer, A. Mercer, P. Amburn, and R. Moorhead. Noodles: A tool for visualization of numerical weather model ensemble uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1421–1430, 2010.

[65] E. Schubert, R. Wojdanowski, A. Zimek, and H.-P. Kriegel. On evaluation of outlier rankings and outlier scores. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, pp. 1047–1058. SIAM, 2012.

[66] E. Schubert, A. Zimek, and H.-P. Kriegel. Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. *Data Mining and Knowledge Discovery*, 28(1):190–237, 2014.

[67] G. S. Shieh. A weighted kendall's tau statistic. *Statistics & Probability Letters*, 39(1):17–24, 1998.

[68] M.-L. Shyu, S.-C. Chen, K. Sarinnapakorn, and L. Chang. A novel anomaly detection scheme based on principal component classifier. Technical report, Miami University Department of Electrical and Computer Engineering, 2003.

[69] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

[70] I. Steinwart, D. Hush, and C. Scovel. A classification framework for anomaly detection. *Journal of Machine Learning Research*, 6(Feb):211–232, 2005.

[71] J. Sun, Y. Xie, H. Zhang, and C. Faloutsos. Less is more: Compact matrix decomposition for large sparse graphs. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pp. 366–377. SIAM, 2007.

[72] J. Talbot, B. Lee, A. Kapoor, and D. S. Tan. Ensemblematrix: interactive visualization to support machine learning with multiple classifiers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1283–1292. ACM, 2009.

[73] D. Thom, H. Bosch, S. Koch, M. Wörner, and T. Ertl. Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages. In *Visualization Symposium (PacificVis), 2012 IEEE Pacific*, pp. 41–48. IEEE, 2012.

[74] A. Topchy, A. K. Jain, and W. Punch. Clustering ensembles: Models of consensus and weak partitions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1866–1881, 2005.

[75] J. Wang, X. Liu, H.-W. Shen, and G. Lin. Multi-resolution climate ensemble parameter analysis with nested parallel coordinates plots. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):81–90, 2017.

[76] W.-K. Wong, A. W. Moore, G. F. Cooper, and M. M. Wagner. Bayesian network anomaly pattern detection for disease outbreaks. In *Proceedings of the 20th International Conference on Machine Learning, 2003*, pp. 808–815, 2003.

[77] K. Yamanishi, J.-I. Takeuchi, G. Williams, and P. Milne. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. *Data Mining and Knowledge Discovery*, 8(3):275–300, 2004.

[78] C. Zhang, M. Caan, T. Höllt, E. Eisemann, and A. Vilanova. Overview+ detail visualization for ensembles of diffusion tensors. In *Computer Graphics Forum*, vol. 36, pp. 121–132. Wiley Online Library, 2017.

[79] J. Zhao, N. Cao, Z. Wen, Y. Song, Y.-R. Lin, and C. Collins. # fluxflow: Visual analysis of anomalous information spreading on social media. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1773–1782, 2014.

[80] A. Zimek, E. Schubert, and H.-P. Kriegel. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(5):363–387, 2012.