Supplementary Material of EnsembleLens

I. Evaluation with Expert Interview: Analysis of Hong Kong Air Quality Dataset

We enhance the evaluation of EnsembleLens through an additional expert interview with the domain expert's data. This section describes the procedure, dataset, and summary of the in-depth interview.

The expert is a project manager who has lengthy research experience and many publications on theoretical and empirical analyses of models and algorithms. He runs a project about a personalized air quality and health management visualization system, and intends to deploy an advanced model to detect and predict air pollution in his city. Therefore, he used our system to compare, evaluate and select some alternative anomaly detection algorithms with the air quality data at hand.

Procedure. We conducted the interview in the form of a brief case study with a real-world dataset from the domain expert we interviewed. The interview included three sessions: (1) introduction (20 minutes): the expert was introduced about the purpose of EnsembleLens and the functions of each view, followed by a tutorial with the Breast Cancer dataset, (2) discovery (20 minutes): the expert was asked to use EnsembleLens to explore prior algorithms for the air quality data from himself, and (3) comment (20 minutes): the expert reported on their findings and commented on the system capabilities. During the interview, a moderator was responsible for demonstrating the system, answering the questions and promopting topics for discussion, like "Which are the algorithms suitable for the dataset after exploration?" "What views are important or easy to use when discovering?" "What are your suggestions on the system?" "How would the system fit into your work?". The interview lasted about 1 hour, during which notes were taken and the entire procedure was recorded.

Dataset. In the case study, we used the data that contains the records of Hong Kong air quality from June 2016 to May 2017. The air quality records have seven attributes, namely, PM2.5 (integer), PM10 (integer), CO (real), SO_2 (real), NO_2 (real), NOX (real), O_3 (real). Moreover, the records have a time step of one hour.

Summary of the interview study. The expert was impressed by the system's capability in both the algorithm evaluation and the ensemble anomaly detection. The most common refrain from the expert was "Cool!" and "Powerful!". He confirmed the utility of our system and commended, "I have never imagined that multiple [anomaly detection] results of my data can be compared in such an easy way!" immediately when he loaded the data into the system. The expert also appreciated the inspection view due to its in-depth comparison and informative visualization design. For example, he was surprised to find ABOD has a good performance because "This design [correlation glyph] shows it has many lines [connected] with other methods... I thought it was bad before the experiment." As shown in Fig.1(a1)&(a2), although oc-SVM (a1) and ABOD (a2) both have a low average correlation with other algorithms, the ABOD is different from oc-SVM as there are many crossings inside the correlation glyphs, which means that many of its top ranked outliers are consistent with other algorithms. Then he clicked some glyphs related with ABOD and inspected them in the ranking view. Many of the results in ranking view verify the findings in the inspection view. For example, Fig.1(b1) shows that the results from ABOD and LOF are highly correlated with the top ranked data points. The expert regarded the design of ranking view efficient for fast data navigation. He noted, "The raw data information like PM2.5 value encoded in the bar and tooltip assists me in judging whether a sample is an outlier or not." Fig.1(b2) shows an example that the dark red can give a direct information of the anomalous high value of O_3 and NO_2. Finally, he got an ensemble anomaly detection result as shown in Fig.1(c), where most top ranked outliers are around June 12 and June 25 of 2016 (he specifically inspected the data of June 2016). After checking the historical weather data of Hong Kong, he found there was a strong monsoon signal on June 12 and very hot weather warning around June 25. "The system can also be effective in anomaly detection!" he commended. He finished his exploration with the weight distribution: oc-SVM(1), RCov(21), iForest(22), LOF(19), LNN(17) and ABOD(19).

The expert also provided valuable suggestions. Although he thought the inspection view is comprehensive and informative, he felt that more automated recommendations could be provided to direct the exploration. Otherwise, the user might "have no idea to select the [ensemble] components to inspect or label at the beginning." In addition, he also pointed out that "The system could be improved to support the comparison among multiple [ensemble] components." By saying this, he meant that the current correlation glyph could only compare two components.

II. Case Study III: Analysis of Biodegradation Dataset

We provide the original visualization results (Figure 8 in original paper) of case study III in our paper in Fig.2.

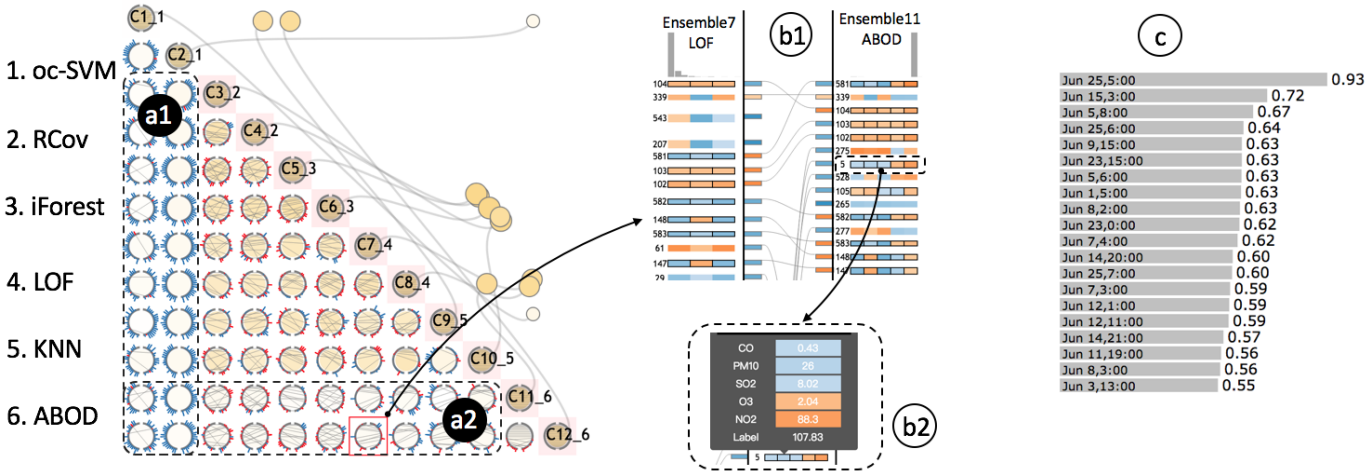

Fig. 1. Analysis results of the domain expert's dataset that contains the air quality records of Hong Kong from June 2016 to May 2017. (a1) the correlation glyphs between the ensemble components from oc-SVM and others; (a2) the correlation glyphs between the ensemble components from ABOD and others; (b1) the ranking view showcases the different performance of two ensemble components from LOF and ABOD; (b2) the raw data information and tooltip of a data point. (c) the rank of outliers scores based on the combination result of different ensemble components.

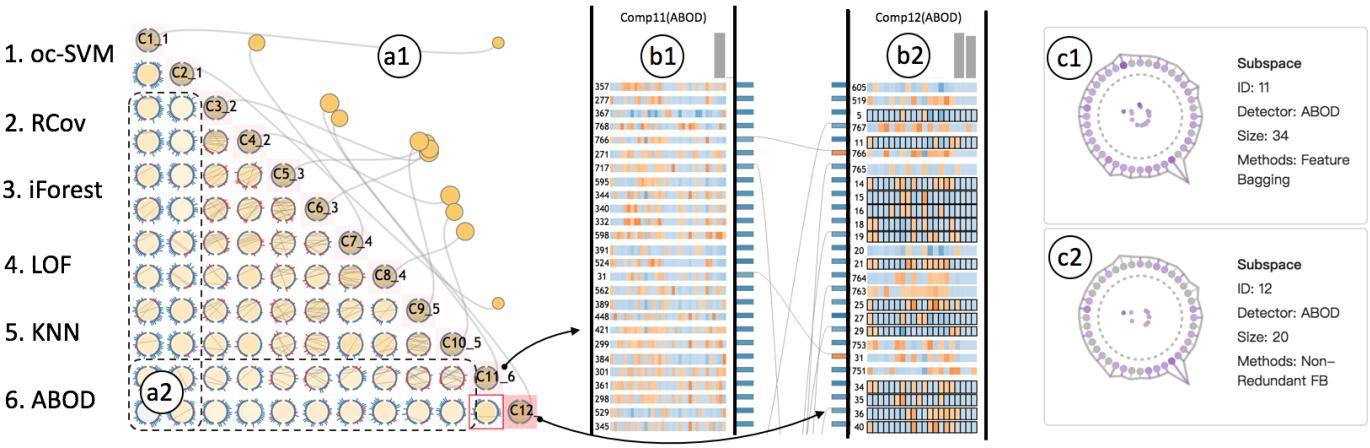

Fig. 2. Analysis results of the Biodegradation dataset that has 41 features. (a1) The ensemble components have a sparse distribution; (a2) the ensemble components from oc-SVM and ABOD have little correlation with others; (b1) and (b2) showcase the different performance of two ensemble components both from ABOD; (c1) and (c2) display the feature subspace corresponding to component (b1) and (b2), where (c2) contains fewer number of features than (c1).