HHS Public Access

Author manuscript

Artif Intell Med. Author manuscript; available in PMC 2018 May 07.

Published in final edited form as:

Artif Intell Med. 2017 September ; 81: 12–32. doi:10.1016/j.artmed.2017.03.003.

Inter-Labeler and Intra-Labeler Variability of Condition Severity Classification Models Using Active and Passive Learning Methods

Nir Nissim[1,2], **Yuval Shahar**[1], **Mary Regina Boland**[3,6], **Nicholas P Tatonetti**[3,4,5,6], **Yuval Elovici**[1,2], **George Hripcsak**[3,6], and **Robert Moskovitch**[3,4,5,6]

[1]Department of Information Systems Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel

[2]Malware Lab, Cyber Security Research Center, Ben-Gurion University of the Negev, Beer-Sheva, Israel

[3]Department of Biomedical Informatics, Ben-Gurion University of the Negev, Beer-Sheva, Israel

[4]Department of Systems Biology, Ben-Gurion University of the Negev, Beer-Sheva, Israel

[5]Department of Medicine, Ben-Gurion University of the Negev, Beer-Sheva, Israel

[6]Observational Health Data Sciences and Informatics, Columbia University, New York, New York, USA

Abstract

Background and Objectives—Labeling instances by domain experts for classification is often time consuming and expensive. To reduce such labeling efforts, we had proposed the application of active learning (AL) methods, introduced our CAESAR-ALE framework for classifying the severity of clinical conditions, and shown its significant reduction of labeling efforts. The use of any of three AL methods (one well known [SVM-Margin], and two that we introduced [Exploitation and Combination_XA]) significantly reduced (by 48% to 64%) condition labeling efforts, compared to standard passive (random instance-selection) SVM learning. Furthermore, our new AL methods achieved maximal accuracy using 12% fewer labeled cases than the SVM-Margin AL method.

However, because labelers have varying levels of expertise, a major issue associated with learning methods, and AL methods in particular, is how to best to use the labeling provided by a committee of labelers. First, we wanted to know, based on the labelers' learning curves, whether using AL methods (versus standard passive learning methods) has an effect on the *Intra*-labeler variability (*within* the learning curve of each labeler) and *inter*-labeler variability (*among* the learning curves of different labelers). Then, we wanted to examine the effect of learning (either passively or actively) from the labels created by the majority consensus of a group of labelers.

Corresponding Authors: Nir Nissim, **Address**: Ben-Gurion University of the Negev, P.O.B 653, Beer-Sheva, Israel 84105, **Phone No.:** +972 086428121, nirni@post.bgu.ac.il.

Methods—We used our CAESAR-ALE framework for classifying the severity of clinical conditions, the three AL methods and the passive learning method, as mentioned above, to induce the classifications models. We used a dataset of 516 clinical conditions and their severity labeling, represented by features aggregated from the medical records of 1.9 million patients treated at Columbia University Medical Center. We analyzed the variance of the classification performance within (*intra*-labeler), and especially among (*inter*-labeler) the classification models that were induced by using the labels provided by seven labelers. We also compared the performance of the passive and active learning models when using the consensus label.

Results—The AL methods produced, for the models induced from each labeler, smoother *Intra*-labeler learning curves during the training phase, compared to the models produced when using the passive learning method. The mean standard deviation of the learning curves of the three AL methods over all labelers (mean: 0.0379; range: [0.0182 to 0.0496]), was significantly lower (p = 0.049) than the *Intra*-labeler standard deviation when using the passive learning method (mean: 0.0484; range: [0.0275 to 0.0724).

Using the AL methods resulted in a lower mean *Inter*-labeler AUC standard deviation among the AUC values of the labelers' different models during the training phase, compared to the variance of the induced models' AUC values when using passive learning. The *Inter*-labeler AUC standard deviation, using the passive learning method (0.039), was almost twice as high as the *Inter*-labeler standard deviation using our two new AL methods (0.02 and 0.019, respectively). The SVM-Margin AL method resulted in an *Inter*-labeler standard deviation (0.029) that was higher by almost 50% than that of our two AL methods. The difference in the *inter*-labeler standard deviation between the passive learning method and the SVM-Margin learning method was significant (p = 0.042). The difference between the SVM-Margin and Exploitation method was insignificant (p = 0.29), as was the difference between the Combination_XA and Exploitation methods (p = 0.67).

Finally, using the consensus label led to a learning curve that had a higher mean intra-labeler variance, but resulted eventually in an AUC that was at least as high as the AUC achieved using the gold standard label and that was always higher than the expected mean AUC of a randomly selected labeler, regardless of the choice of learning method (including a passive learning method). Using a paired t-test, the difference between the *intra*-labeler AUC standard deviation when using the consensus label, versus that value when using the other two labeling strategies, was significant only when using the passive learning method (p = 0.014), but not when using any of the three AL methods.

Conclusions—The use of AL methods, (a) reduces *intra-labeler* variability in the performance of the induced models during the training phase, and thus reduces the risk of halting the process at a local minimum that is significantly different in performance from the rest of the learned models; and (b) reduces *Inter*-labeler performance variance, and thus reduces the dependence on the use of a particular labeler. In addition, the use of a consensus label, agreed upon by a rather uneven group of labelers, might be at least as good as using the gold standard labeler, who might not be available, and certainly better than randomly selecting one of the group's individual labelers. Finally, using the AL methods when provided by the consensus label reduced the intra-labeler AUC variance during the learning phase, compared to using passive learning.

Keywords

Active Learning; Electronic Health Records; Phenotyping; Condition; Severity; Variance; Labeling

1. INTRODUCTION

Active learning (AL), a form of machine learning in which the learning method actively requires labels for specific instances in which knowing the label seems most beneficial to the learning process, has been at the focus of a substantial amount of research over the last decades. AL has been shown to be successful in decreasing the amount of labeling requirements, compared to a traditional passive learning method, in many domains including the cyber security (25–27, 37–41, 68–71) and biomedical domains (30–33, 53–54). While labeling and learning with an active learner is often much more efficient and achieves higher classification accuracy with a smaller labeled training set, the learning curve may vary greatly according to the labeler's expertise in the domain. The clinical domain is an excellent example of a domain in which there is a large number of potential experts with varying levels of expertise, depending on their training and experience. However, physicians, and particularly experts, are often very busy, and their time is expensive. Thus, the focus of our current study is to examine the use of labelers with varying levels of clinical training and experience.

We have previously examined the effect of various learning methods on the specific task of determining the *severity level* of medical conditions. The severity level is an important aspect of each medical condition, which is expected to be useful for discriminating between sets of conditions or phenotypes. For the purposes of our research, we define severe conditions as those that are life threatening or permanently disabling. Such conditions would be considered as high priority in terms of the need to generate phenotype definitions for tasks such as pharmacovigilance (44, 45, and 47). Condition level severity classification can distinguish acne (mild condition) from myocardial infarction (severe condition). The bulk of the literature focuses on *patient level* severity, which generally requires individual condition metrics (8, 9, 10, 11), although whole-body methods exist (11, 12, 13).

Severity level is also useful for prioritizing conditions that are important for specialized phenotyping algorithms. Although several consortiums and partnerships, including the Observational Medical Outcomes Partnership (1) and the Electronic Medical Records and Genomics Network (2, 3), have developed methods for extracting conditions and their related characteristics from Electronic Health Records (EHRs), only a little more than 100 conditions/phenotypes have been successfully defined. Unfortunately, this represents just a small fraction of the approximately 401,200[1] conditions recorded in EHRs. Hurdles faced by experts when defining phenotype-extraction algorithms include overcoming definition discrepancies (4), data sparseness, data quality (5), bias (6), and healthcare process effects

[1]The number of SNOMED-CT codes as of September 9, 2014. Accessed via: http://bioportal.bioontology.org/ontologies/SNOMEDCT

(7). Condition severity can be one way of identifying conditions worthy of developing a specialized phenotype-extraction algorithm.

In our previous work, we developed an algorithm that we refer to as **C**lassification **A**pproach for **E**xtracting **S**everity **A**utomatically from Electronic Health **R**ecords (CAESAR) (13, 47), which uses standard machine learning (also referred to as *passive learning*) to classify condition severity based on metrics extracted from EHRs (13) and requires medical experts to manually review and assign a severity status to each condition (i.e., severe or mild) independently from EHR metrics. We have recently developed and assessed an *Active Learning Enhancement* version of *CAESAR*, called *CAESAR-ALE*, which was initially published as a preliminary study (49), and was then extended into a more detailed paper (76). Using three different AL methods, including two new AL methods that we developed, we demonstrated that the labeling burden on medical experts can be significantly reduced. All three AL methods decreased the labelers' efforts, compared to the passive learning methods applied by the original CAESER framework in which the classifier was trained on the entire set of conditions; depending on the AL strategy used in that study (13), the reduction ranged from 48% to 64%, which can result in significant savings, both in time and money.

Several labelers participated in our original study, and a separate learning curve was created for each labeler, depicting the classification model induced by using the labels provided by each labeler. The variance between the learning curves observed might be a result of the varying levels of clinical training and experience of the labelers. In the current study, we delve deeply into the *Intra*-labeler and *Inter*-labeler variance in the labelers' learning curves and investigate this variance in detail, including the effect of using passive or active learning methods on that variance. We also examine the effect of using the consensus of the whole group of labelers as a single label. The implications of such an investigation are quite important, especially in clinical domains, since the optimal use of the labels created by individual labelers is at the core of any attempt to create effective classifiers based on a committee of expert labelers. It is therefore crucial to understand how that performance varies and what steps can be taken to reduce such variance, and in particular, the effect of using AL methods, consensus labeling, or both.

A previous study we performed (17) involving the task of determining the eventual severity level of patients, given their initial emergency room record encompassing their triage process, before being seen by a physician, demonstrates the motivation for our in depth investigation of *inter*-labeler variability. The results showed that inducing a classification model using several standard passive (Random selection) learning methods resulted in a reasonable performance compared to each of the clinicians. However, a model induced only by the cases in which there was an agreement (a consensus) regarding the severity level between two physicians subsequently examining the records to assign a severity level to each record, was preferable to learning a model from either of the physicians' separate judgments. Nevertheless, the learning process in our previous study ignored cases in which there was a disagreement between the two labelers; a true committee tasked with coming to a consensus and making a majority decision on each case might have resulted in an even better performance.

In the current study, we examine the implications of using various AL methods on the *Intra*-labeler variability, i.e., within the learning curve of each labeler, and on the *Inter*-labeler variability, i.e., between the learning curves of different labelers, when training a classifier using a group of seven labelers. (The *intra*-labeler variability can be viewed as the *volatility* of the learning curve). In both cases, we compute the variability among the performance of the models induced using the provided labels, measured by the *Area Under the ROC Curve* (AUC). Our previous results, combined with the intriguing issue of *Inter*-labeler and *Intra*-labeler variability, had raised the following three research questions: (a) what is the *Intra*-labeler and *Inter*-labeler variance of the seven labelers' learning curves for each learning method, and in particular, for the AL methods versus for the passive learning method? (b) Does the difference in the *Inter*-labeler variance between the AL methods, and the baseline passive (Random selection) learning method, change during the learning process? How? and (c) what is the classification performance of the model induced when using the labelers as a *committee* and using the label having the majority votes (i.e., the *consensus* label), compared to the option of using the gold standard label, or even a random individual labeler picked from the potential labelers' group?

Note that on one hand, reducing *intra-labeler* variability in the performance of the models induced during a training process, which might not have any clear-cut stopping point, reduces the risk of halting the process at a local minimum that is significantly different in performance from the rest of the learned models. On the other hand, reducing *Inter*-labeler performance variance reduces the dependence on the use of a particular labeler. Finally, it would be highly useful if using a *consensus* label that is agreed upon by a rather uneven group of labelers, might be a reasonable alternative to using a gold standard expert labeler, who might not be available.

The rest of the paper is structured as follows. In Section 2 we provide background and related work to this study. In Section 3 we describe our new methods; this is followed by Section 4, in which we present the evaluation methods and the experimental design aimed at proving or disproving our hypotheses. In Section 5 we present a very brief summary of our main results from the previous study, which serves as the baseline for the analysis, and the detailed results of our new experiments. We discuss our new results and present our conclusions in Sections 6 and 7, respectively.

2. BACKGROUND

As the current study deals with *Inter*-Labeler variability of condition severity classification models using active and passive learning methods, we start by providing some important background information regarding active learning: what AL is, current approaches to AL, and existing studies that apply AL learning methods and algorithms to biomedical tasks (subsection 2.1 Active Learning Applications in Biomedical Data). The active learning algorithms are based on data mining and machine learning algorithms which are presented in subsection 2.2: Mining Electronic Health Records. Finally, since this study is about classifying the severity of conditions from ICD-9 and SNOMED-CT, we also introduce these ontologies/vocabularies (subsection 2.3: Biomedical Ontologies and Classification).

2.1 Active Learning Applications in Biomedical Data

Labeled examples, crucial for classification, are generally expensive to acquire, since they require medical experts for annotation. *Active learning* (*AL*) approaches are useful for selecting (for labeling) the most discriminative and informative conditions from a dataset during the learning process. This selection is expected to decrease the number of conditions that experts need to manually review and label. Studies in several domains have successfully applied AL to reduce the resources (i.e., time and money) required for labeling examples (25, 26, 27,68,69,70,71,81,82, 83). AL is divided roughly into two major approaches: 1) *membership queries* (28) in which examples are artificially generated from the problem space; and 2) *selective-sampling* (29) in which examples are selected from a pool, which is the focus of this paper. AL algorithms have been relatively widely utilized in multiple domains, although applications in the biomedical domain, including text (30, 32, 33), drug discovery (31), cell image pathology (53), and cell assay image classification (54), remain limited. Liu described a method similar to relevance feedback for cancer classification (30). Warmuth et al. used a similar approach to separate compounds for drug discovery (31). More recently, AL was applied in biomedicine for text (32) and radiology report classification (33).

2.2 Mining Electronic Health Records

Secondary use of EHRs through data mining (57) has become a trendy research topic in biomedical informatics (58, 80) and data mining literature (59, 60, 67). Learning predictive models in clinical medicine through data mining is an important and developing field (58, 72, 79). Ng et al. (61) introduced a distributed platform for healthcare analytics for EHR data which is based on MapReduce principles and parallels the entire process of cohort construction, feature construction, and selection and classification in a cross-validation fashion. Sun et al. (62) used this framework to predict hypertension transition points in EHR data without temporal representation. Until recently, very little temporal analysis has been introduced regarding various predictive events (63). Rana et al. (64) introduced a useful framework that models the change in interventions over time in order to predict outcome, and their framework considers the temporal evolution of the events. To handle temporal data, a comprehensive framework was introduced that enables learning patients' behavior over time, including the discovery of frequent temporal patterns (60), learning classification models (65), and acquiring cutoffs to discretize the variables into states to increase classification performance (66).

2.3 Biomedical Ontologies and Classification

In this study we used the SNOMED-CT ontology (14, 15). Each coded clinical event is considered a "condition" or "phenotype," with the knowledge that this is a broad definition (4). In biomedicine, condition classification follows two main approaches: 1) manual approaches in which experts manually assign labels to conditions; and 2) passive classification approaches that require a labeled training set. Manual approaches include the development of a Chronic Condition Indicator (CCI) (18) involving expert assignment of chronicity categories (acute versus chronic) to *International Classification of Diseases, Ninth Revision* (ICD-9) codes. Other researchers employed standard learning approaches in the

biomedical domain, including a classification approach that leveraged the ICD-9 hierarchy for improved performance (21). Another study classified conditions into chronicity categories (22). Other machine learning approaches have been used in biomedicine for text classification into condition hierarchies (48) to better enable subsequent retrieval (74, 55), as well as in full-text context-sensitive search engines (56). Torii et al. (23) showed that performance improved when trained on a dataset based on multiple data sources and noted that having more documents available during training improved performance (23). Nguyen et al. built an algorithm for classifying lung cancer stages using pathology reports and SNOMED-CT (24).

3. METHODS

This paper is based on the CAESAR-ALE framework which we developed in our recent study (49, 76). In this study, we aimed at comparing our AL methods to existing AL methods. Because we also wanted the CASAER-ALE framework to be practical for actual online use, it was important that the AL methods also be quite fast. Although general (i.e., not based on an SVM classifier) and high performing AL methods do exist, their effectiveness is limited; for example, the “Error-Reduction” method (75) is reportedly not practical enough, and many alternatives which are lighter-weight variants (yet not optimal) have been proposed in order to make it more useful.

In this section we present the methods and techniques upon which our framework is based. We aim to provide a solution to an existing challenge in the area of efficient condition severity classification, and our framework is based on a combination of methods and techniques derived from previous research (ours and research conducted by others) that we believe will be most appropriate for achieving the goals of this study.

One of the leading AL methods, fast enough for practical use and considered very stable, is the SVM-Simple-Margin method. This method is based on the SVM classification principles and is known for its agility, efficiency, and high performance. It is thus a highly suitable methodology to use, and in our previous study was a very strong baseline to which we could compare our enhancements. Additionally, the CASEAR-ALE framework is not restricted to any particular AL method; it is a modular framework for practical use and research, and the results of using one AL method can be easily generalized to other AL methods. Thus, in the current study we continued to work with the AL methods and SVM classifier that we had previously integrated in the CAESAR-ALE framework. Finally, this strategy proved to be quite efficient, since the goal of the current study was to analyze the *Inter*-labeler variability of condition severity classification, given a proven set of AL methods, rather than to compare different AL methods. Such a comparison was partially performed in our previous work, in which we compared three AL methods and a passive learning method.

3.1 Random Selection

The passive (Random selection) learning method is not an active learning method, but it is the “lower bound” of the selection methods that will be discussed. Based on our knowledge, no biomedical machine learning based solution has used an active learning method to reduce the labeling efforts of medical doctors in the task of condition severity classification. The

passive (Random selection) learning method doesn't have a sophisticated selection strategy; consequently, we expect that all of the AL methods we examine will perform better than a selection process based on the random selection of conditions. In the context of our framework, the passive (Random selection) learning method will feed the SVM classifier with conditions that were randomly selected from the pool of unlabeled conditions. In our experiments we refer to this method as Random selection or just Random.

3.2 The SVM-Simple-Margin AL Method (SVM-Margin)

The SVM-Simple-Margin method (35) (or SVM-Margin) is based on SVM classifier principles, and the most significant elements of this classifier are presented in order to properly describe the AL methods which are based on it.

The support vector machines (SVM) classifier is a binary classifier that finds a linear hyperplane that separates given examples into two specific classes, yet is also capable of handling multiclass classification (50). As Joachims (51) demonstrated, the SVM is widely known for its ability to handle a large amount of features, a capability which is useful in the textual domain.

Given a training set in which an example is a vector $x_i = <f_1, f_2 \ldots f_m>$, in which f_i is a feature labeled by $y_i = \{-1, +1\}$, the SVM attempts to specify a linear hyperplane with the maximal margin defined by the maximal (perpendicular) distance between the examples of the two classes. Figure 1 illustrates a two-dimensional space where the examples are positioned according to their features. The hyperplane splits them based on their labels.

The examples lying closest to the hyperplane are the "supporting vectors." W, the Normal of the hyperplane, is a linear combination of the most important examples (supporting vectors) multiplied by LaGrange multipliers (α), as can be seen in Equation 3. Since the dataset in the original space cannot always be linearly separated, a kernel function K is used. SVM actually projects the examples into a higher dimensional space in order to create a linear separation of the examples. Note that when the kernel function satisfies Mercer's condition, as Burges (52) explained, K can be presented using Equation 1, where Φ is a function that maps the example from the original feature space into a higher dimensional space, while K relies on the "inner product" between the mappings of examples x_1, x_2. For the general case, the SVM classifier will be in the form shown in Equation 2, where n is the number of examples in the training set, K is the kernel function, α is the LaGrange multiplier that defines the linear combination of the Normal W, and y_i is the class label of support vector X_i.

$$K(x_1, x_2) = \Phi(x_1)\Phi(x_2) \quad (1)$$

$$f(x) = \text{sign}(w \cdot \Phi(x)) = \text{sign}\left(\sum_{1}^{n} \alpha_i y_i K(x_i x)\right) \quad (2)$$

$$w = \sum_{1}^{n} \alpha_i y_i \Phi(x_i) \quad (3)$$

Two commonly used kernel functions are utilized: 1) the polynomial kernel, as shown in Equation 4, creates polynomial values of degree p, where the output depends on the direction of the two vectors, examples x_1, x_2, in the original problem space (note that a private case of a polynomial kernel, where $p=1$, is actually the linear kernel), and 2) the radial basis function (RBF), as shown in Equation 5, in which a Gaussian function is used as the RBF, and the output of the kernel depends on the Euclidean distance of examples x_1, x_2.

$$K(x_1, x_2) = (x_1 \cdot x_2 + 1)^P \quad (4)$$

$$K(x_1, x_2) = \mathbf{e}\left(-\frac{\|x_1 - x_2\|^2}{2\sigma^2}\right) \quad (5)$$

After the principles of SVM are clearer, we now delve into our AL method itself. Using a kernel function (such as the polynomial or RBF), the SVM implicitly projects the training examples into a different (usually higher dimensional) feature space. In this space there is a set of hypotheses that are consistent with the training set, creating a linear separation of the training set. The SVM identifies the best hypothesis with the maximal margin from among the consistent hypotheses (referred to as the version space [*VS*]). To achieve a situation in which the VS contains the most accurate and consistent hypothesis, the SVM-Margin AL method, selects examples from a pool of unlabeled examples, sends them to a human expert for accurate labeling, and then adds them to the training set. Adding additional labeled examples to the training set creates further constraints for the SVM separating hyperplane, thus thereby ultimately reducing the number of hypotheses, since a smaller number of hypotheses are now consistent with all of the labeled examples. Note that adding informative examples to the training set will result in a reduction in the number of hypotheses for any SVM classifier, yet SVM-Margin is aimed at selecting the most informative ones, so that the maximal number of hypotheses will be achieved.

The SVM-Margin method selects examples according to their distance from the separating hyperplane to explore and acquire informative conditions, disregarding their labels. Examples that lie closest to the separating hyperplane (see Figure 2 in which the selected examples from both classes are colored in red and lie inside the margin) are more likely to be informative (may improve the classifier's capabilities) and therefore are acquired and labeled. SVM-Margin is fast; yet, as its authors indicate (35), this agility is achieved because of its rough approximation and assumptions that the VS is fairly symmetric and the hyperplane's Normal, W (Equation 3), is centrally placed. These assumptions have been

shown to fail significantly (36), because SVM-Margin may query instances whose hyperplane does not intersect with the VS and therefore may not be informative.

3.4 The CAESAR-ALE Framework

The purpose of our enhanced method, CAESAR-ALE, is to decrease the experts' labeling efforts using AL methods. CAESAR-ALE does this by only asking experts to label informative conditions. Figure 3 illustrates the process of labeling and acquiring new conditions by maintaining the updatability of the classification model within CAESAR-ALE. Conditions are introduced to the classification model, which is induced by an SVM algorithm on which AL methods are based. The classification model scrutinizes conditions and provides two values for each condition: a classification decision using the SVM classification algorithm and a calculation of the distance from the separating hyperplane. Informative conditions are defined as conditions that are expected to improve the classification model's predictive capabilities when added to the training set. A condition that is identified as potentially informative by the AL method is sent to a human expert for labeling. In this manner, most potentially informative conditions are labeled and added to the training set so that a new and updated model will be induced.

The CAESAR-ALE framework includes several AL methods that use different strategies for considering informative conditions. One of the strategies that will be described further consists of acquiring conditions that are located deep within the positive (severe conditions) sub-space of the SVM classifier's separating hyperplane, i.e., as far as possible from the hyperplane on the positive side.

By selecting the most informative conditions, the use of an AL method leads to a theoretical decrease in the labeling efforts, as compared to learning from the entire set of conditions. Using the AL approach, we can maintain an accurate model while decreasing the labeling efforts, since the induction method requires fewer examples, i.e., conditions, because the input instances are more informative. Accordingly, in our context, there are two types of conditions that may be considered informative. The first type includes conditions that the classifier has identified with a low level of confidence. Here, the probability of being mild is close to the probability of being severe. The calculation of probability in based on the distance of the example from the separating hyperplane of the SVM classifier – thus a maximal distance from the separating hyperplane represents a high level of confidence, while minimal distance from the separating hyperplane represents low confidence. Equation 6 represents the distance of example x from the separating hyperplane of the SVM classifier (note that Equation 2 makes use of this distance and provides a classification decision regarding the sign of the distance, in which a positive sign means a positive class classification, while a negative sign means a negative class classification).

$$f(x) = w \cdot \Phi(x) = \sum_{1}^{n} \alpha_i y_i K(x_i x) \quad (6)$$

In order to calculate this probability using the distance of example *x* from the separating hyperplane according to the given classifier's knowledge, we use a transformation function that converts distance value into probability (42), see Equation 7:

$$P(y|x) = \frac{1}{1 + e^{-y \cdot f(x)}} \quad (7)$$

Where *y* is an optional label of example *x*. {+1,-1}, h(x) is the decision value provided by Equation 6. An illustration can be seen in Figure 4, which shows two examples for which the SVM produced classification decision values.

For instance: P (y=-1 | X2) = 0.8 means that the classifier is quite confident that x2 belongs to class (-1). While P (y= +1 | X2) = 0.2 means that the classifier is quite confident that X2 does not belong to class (+1); if P is (y=-1 | X2) = (y= +1 | X2) = 0.5, the classifier has a severe lack of confidence regarding the class of X2. A graphical analysis of Equation 7 can be seen in Figure 5.

The second type of informative condition includes conditions that are at a maximal distance from the separating hyperplane; these conditions are deep within the *severe* instances sub-space of the SVM's separating hyperplane. Nevertheless, some *mild* conditions may still exist within this space of otherwise severe conditions (although, of course, their being *mild* is unknown to the algorithm, until they are selected and labeled). Consequentially, presenting these mild conditions to the induction algorithm is expected to greatly inform and improve the resulting classification model.

The overall CAESAR-ALE framework includes a repetition of two main phases: training and classification/updating, which includes the selection of potentially informative examples (i.e., conditions), labeling them, and then training the model with the new labeled conditions.

Training—The model is trained using an initial pool of severe and mild conditions. The model is evaluated against a test set consisting of conditions that were not used during training to estimate the classification accuracy.

Classification and Updating—The AL method estimates and ranks how potentially informative each condition is within the pool of unlabeled conditions left, based on the classification model's prediction. Only the most informative are selected and labeled by the expert. These conditions are added to the training set and removed from the pool. The model is then retrained using the updated training set, and this process repeats iteratively until a sufficient level of accuracy is reached, or alternatively, until the entire pool of conditions have been acquired.

We employed the SVM classification algorithm using the radial basis function (RBF) kernel in a supervised learning approach. This combination has been shown to be very efficient when combined with AL methods (26, 27). We use the Lib-SVM implementation (34) and modify it to implement our AL methods.

Although our focus in this prior study is on reducing the condition labeling efforts while maintaining similar or enhanced classification performance, the detection of *severe* conditions - even during the learning phase (as opposed to the detection of *mild* conditions) - has some advantages, due to their value for training and insurance reporting purposes.

3.5 CAESAR-ALE's Active Learning Methods

CAESAR-ALE includes two AL methods (Exploitation and Combination_XA), which we now describe in detail.

3.5.1 Exploitation—One of the AL methods implemented in CAESAR-ALE is called Exploitation, referred to as such because it exploits the current separating hyperplane to find condition instances that are most likely to be severe. Exploitation has demonstrated efficiency at detecting unknown malicious code content, files (37–40), and documents (41). Exploitation is based on SVM classifier principles and selects examples more likely to be severe, those lying further from the separating hyperplane, as can be seen in Figure 2. Thus, this method aims at boosting the classification capabilities of the model through the acquisition of as many new severe conditions as possible. For every condition x, Exploitation measures its distance from the separating hyperplane using Equation 8, based on the Normal (W) of the separating hyperplane of the SVM classifier. The separating hyperplane of the SVM is represented by W, which is a linear combination of the most important examples (supporting vectors), multiplied by LaGrange multipliers (α) and by the kernel function K that assists in achieving linear separation in higher dimensions. Accordingly, the distance in Equation 8 is calculated between example x and the Normal (W) presented in Equation 3. The distance calculation required for each instance in Exploitation is equal to the time it takes to classify an instance using SVM-Margin.

$$\mathrm{Dist}(X) = \left(\sum_{1}^{n} \alpha_i y_i K(x_i x) \right) \quad (8)$$

Acquiring several severe conditions that are highly similar to each other (i.e., which have similar values for all of the meaningful features, and of course, belong to the same target class) would waste labeling resources, while not contributing much to the future classification capabilities (generality) of the induced classifier; therefore, acquiring one representative condition from this set is preferable. In the Exploitation method, conditions are acquired if they are classified as severe and have maximal distance from the separating hyperplane (marked with a red circle in Figure 6.1).

To enhance the training set as much as possible, we also check the similarity among selected conditions using the kernel farthest-first (KFF) method suggested by Baram et al. (42), enabling us to avoid acquiring similar conditions. Consequently, only potentially informative conditions likely to be labeled as severe are selected. In Figure 6.1, it can be seen that there are several sets of highly similar conditions, based on their distance in the kernel space. However, only representative conditions that are more likely to be severe are acquired.

Contrary to SVM-Margin, Exploitation explores the "severe space" to discover potentially more informative severe conditions, a process which enables further detection of severe conditions. Figure 6.1 also illustrates an additional ability of Exploitation, as it sometimes discovers conditions located far inside the severe side (i.e., class) that were ultimately labeled by the expert as mild. Finding such a surprise is highly useful - this type of confusing condition will become a new support vector of the SVM classifier and update the classification model with the new discovery and knowledge; thus, these "surprises" play an important role in increasing the accuracy of the resultant classifier.

3.5.2 Combination_XA: A Combined Active Learning Method—The "Combination_XA" method is a hybrid of SVM-Margin and Exploitation. It conducts a cross acquisition (XA) of potentially informative conditions. That means that during the first trial (and all odd-numbered trials) it acquires conditions according to the SVM-Margin method's criteria, while during the next trial (and all even-numbered trials) it selects conditions using the Exploitation method's criteria. This strategy alternates between the exploration phases (conditions acquired using SVM-Margin) and the exploitation phase (conditions acquired using Exploitation) to select the most informative conditions, both mild and severe, while boosting the classification model with severe conditions or very informative mild conditions that lie deep inside the severe side of the SVM's hyperplane.

4. EVALUATION

In this section we describe our dataset and the label creation process that serves as the basis of our experimental design, which we present in detail later.

4.1 Dataset

We obtained a dataset of 516 conditions, along with six severity-related metrics associated with each condition. These metrics or features consist of: *average number of comorbidities, average number of procedures, average number of medications, average cost of procedures, average treatment time*, and a *proportion term* (47). Each of the severity-related metrics was computed using an underlying dataset of over 1.9 million patients. Each of the 516 conditions was labeled as either severe or benign by the gold standard labeler as was defined in our previous work (13).

Our dataset, called the "CAESAR dataset," was created from the medical records of 1.9 million patients treated at Columbia University Medical Center (CUMC). A proportion term was calculated to normalize each severity metric using the entire corpus and to combine all five metrics into one weighted term. Each condition's proportion term was calculated previously as part of CAESAR's construction, additional details are found in that study (47). The method for calculating the proportion term is as follows. To calculate the proportion term we first calculate a proportion for each of the five measures. We then sum these individual proportions and divide by the total (i.e., five). It is easiest to illustrate this with an example. Let us assume the condition "myocardial infarction" has an *average procedure cost* of $10,000, an *average treatment length* of 30 days, an *average number of medications* of ten, an *average number of procedures* of six, and an *average number of comorbidities* of three. Each of these values would be divided by their respective maximums. Therefore, the

proportions are as follows: *average procedure cost* - $10,000/$50,000; *average treatment length* - 30/1406 days; *average number of medications* - 10/25; *average number of procedures* - 6/15; and *average number of comorbidities* - 3/100. Each of these proportions are then summed:

$$\frac{\left(\frac{10,000}{50,000}+\frac{30}{1406}+\frac{10}{25}+\frac{6}{15}+\frac{3}{100}\right)}{5}=\frac{1.051}{5}=0.210$$

This yields the proportion index term for this condition.

The proportion terms are the features used by the various learning methods we explored.

Of the 516 conditions included in the above mentioned dataset, 100 conditions were also labeled by seven additional labelers (three of the labelers were medical doctors who had completed their residency training, and the remaining four were informatics experts with at least a master's degree). The 100 training conditions were presented to the seven labelers in the same order. Each of the labelers provided a label for each of the cases, and their labeling time was not limited. Typically, they returned the labels within a week or two. These 100 conditions served as the basis for the three main experiments in the current study in which we examined the various implications of using AL methods on the learning curve and its *Intra*-labeler (during the training phase) and *Inter*-labeler variance (between labelers) when training the classifier using a group of different labelers.

4.2 Experimental Setup

4.2.1 The Basic Experiment—It is important to understand the basic experiment that served as the core of our previous study in order to appreciate the motivation and contribution of the current study; thus, we briefly explain our original experimental setup. This basic experiment was *aimed* at evaluating and comparing the selection (learning) methods in the task of efficiently creating an accurate severity classification model, while reducing the labeling efforts of medical experts.

We used a repository of 516 conditions (the CAESAR dataset) consisting of 372 mild and 144 severe conditions. Ten randomly selected datasets were created in order to perform 10-fold cross-validation evaluation. Each fold contains three elements: 1) an initial set of six conditions that are used to induce the initial classification model, 2) a test set of 200 conditions on which the induced classifier is tested and evaluated in each active learning iteration, and 3) a pool of 310 unlabeled conditions from which the conditions are selected to be labeled by each of the examined selection methods. The process is repeated, throughout the iterative acquisition steps, until the entire pool of training conditions is acquired. The performance of the classification models was averaged over the ten runs of the 10-fold cross-validation.

The experiment's steps follow, along with Figure 6.2 illustrating the workflow based on these steps.

1) Induce the initial classification model from the initial training set containing six seed conditions.

2) Evaluate the classification model using the 200 conditions test set to measure its initial performance.

3) Introduce the pool of 310 unlabeled conditions to the sampling methods. During every trial, the five most informative conditions are selected according to the selection method's preferences, and their labels are revealed by the single gold standard labeler used in the original CAESAR system (in a non-experimental system, the selected conditions will be labeled by an expert when needed, but in our dataset all of the conditions are already labeled). We used a low acquisition amount of five conditions per trial, because our primary goal was to minimize the number of conditions sent to medical experts for manual labeling and thereby reduce costs.

4) Add the acquired conditions to the training set and remove them from the pool.

5) Induce an updated classification model using the updated training set and apply the updated model to the conditions remaining in the pool.

6) This process (stages 2–6) iterates until the entire pool of training conditions is acquired.

We now present the experimental design of the three core experiments in the current study. We have designed and conducted specific experiments (Experiments 1, 2, and 3) for each of the new research questions (A, B, and C) which were previously mentioned.

4.2.2 Experiment 1 – Assessing the Intra-Labeler and Inter-Labeler Learning Curve Variance—This experiment attempts to provide insight regarding the variance within and among the learning curves induced as follows: 1) when learning from the labels provided by each of the seven labelers, and 2) when using different instance selection methods, i.e., active and passive learning. Thus, this experiment is aimed at assessing the differences between the various learning methods, by examining the variance within the learning curves of each labeler for all learning methods, as well as the *Inter*-labeler variance across all labelers. The conditions to be labeled are selected by either the three AL methods or by the Random selection (passive) learning method for each of the seven different labelers, in order to better compare the AL and random methods for different labelers. Thus, here we use the part of the dataset containing the 100 conditions labeled by seven different labelers as our pool. We follow the same steps outlined in the basic experiment of our previous study (having a seed of six randomly selected conditions, and iteratively acquiring a chunk of conditions out of the one hundred left to label); the initial set of six seed conditions was again labeled by the gold standard labeler. After each acquisition step we evaluated the performance of each of the labelers using the remaining 410 conditions labeled by the gold standard labeler, according to the following calculation:

(516 conditions – 6 seed conditions) – (100 conditions labeled by all seven labelers) = 410 remaining [test set] conditions labeled by the gold standard labeler. In addition, in each iteration we compare the variance among the different labelers after each acquisition step.

We performed this experiment using a 10-fold cross-validation process. During the training phase of each of the active learning algorithms, the labels for the 100 condition instances were queried by the algorithm in a different order for each set of instances labeled by each of the labelers. This is due to the nature of active learning, depending on the particular model being induced dynamically, and based on the labels provided by that labeler.

4.2.3 Experiment 2 – The Labelers' Learning Curves' Inter-Labeler Variance—

This experiment was designed to address the issue of the difference, during the training process, between the *Inter*-labeler variance when using classifiers induced by the different AL based learning models, and the variance when using the passive learning (Random selection) method.

For each selection (learning) method (the Random selection and an additional AL method from the three: Combination_XA, Exploitation, SVM-Margin), we measured the performance for each labeler across the various selection methods, and the variance among the labelers at each point in the training phase. We then calculated the difference in the variance of the models induced by the seven labelers, focusing on the difference between the AL methods and the passive (Random selection) learning method.

In addition, we examined the same AL-passive learning difference in *Inter*-labeler variance when using the gold standard label and the single consensus label assigned by a majority of the labelers.

4.2.4 Experiment 3 – Examining the Effect of Learning Only from the Consensus Label—

This experiment aimed to address the question of how the performance (AUC) of the classifier changes over time during the training phase, when learning is performed using all three AL methods and the passive learning method, using two different labels: (1) the gold standard label (originally defined by the CAESER framework) (13), versus (2) the consensus label (the label determined by the majority of the seven labelers). Note that the consensus label was created only retrospectively, after all of the instances were labeled by each of the seven labelers. Thus, each labeler labeled the conditions independently and the consensus was not achieved as a committee, but rather based on the majority of their retrospective labels.

In addition to comparing the performance of the two models induced using either of these two labels, we compared them to the performance of a third model induced by a randomly selected labeler, which was represented by the mean AUC of the labelers. (This situation might occur when a labeler is selected at random from a group of labelers, without prior notion of any skill differences among labelers). We represented the expectation of the AUC of the random labeler model by the mean AUC of the seven models induced by using the labels provided by the seven different labelers.

5. RESULTS

To better understand our new *Inter*-labeler variability experiments with the CAESER-ALE framework, we first summarize the results of our original *basic experiment*, which are

presented in greater detail in our previous papers (49, 76) and have been significantly expanded in the current work. We present a summary of the results for the accuracy, although in our previous paper additional evaluation measures were used including: TPR, FPR, and AUC, as well as the number of new severe conditions discovered and acquired into the training set during each trial (49, 76).

We now briefly present the results (in figures 7,8 and 9) of our basic experiment from the previous study. Figure 7 presents the accuracy levels and their trends in the 62 trials. In most trials, all of the AL methods outperformed the Random selection method, when provided with the same number of labeled instances. Summarizing the results for the accuracy rate, the performance of the model induced using Combination_XA represents a reduction of 48% in labeling efforts compared to the use of the passive (Random-selection) method, achieving an accuracy rate of 0.95 after 23 trials compared to the 44 trials acquired by the passive (Random selection) method. Considering the overall learning phase, after 35 trials, the model induced using the Combination_XA method performed slightly better than the model induced using the Exploitation method, indicating that the cross-acquisition strategy is superior over the long run. Additional details regarding these experiments, including other aspects, such as the rate of acquisition of severe conditions, are presented in our previous papers (49, 76).

Figure 8 shows the TPR levels (*severe* is the positive class) and their trends over 62 trials. Both Exploitation and Combination_XA outperformed the other selection methods, achieving a TPR rate of 0.85 after only 17 trials (85/310 conditions), while the passive (Random selection) method achieved the same TPR rate after 47 trials. Summarizing the results for the TPR rates, the performance of Exploitation and Combination_XA represents a reduction of 64% in labeling efforts compared to the passive learning method.

Figure 9 shows that by the fifth trial, CAESAR-ALE's methods, Exploitation and Combination_XA, outperformed the other selection methods with respect to the rate of acquiring severe condition instances (both lines overlap in Figure 9). After 23 trials (115 conditions), both of CAESAR-ALE's methods acquired 73 out of the 82 severe conditions in the pool, compared to 42 trials (210 conditions) for the SVM-Margin method and 60 trials (300 conditions) for the passive (Random selection) method. This represents a 46% reduction in the number of trials required to acquire the same number of *severe* condition instances compared to the SVM-Margin method, and a 62% reduction in the labeling efforts compared to the passive learning method.

Note that Exploitation and Combination_XA overlap only in the number of severe conditions acquired, and not in the number of mild conditions acquired. This observation is supported by Figure 7, in which the accuracy rate achieved by each of the methods is different. Due to the strategy of SVM-Margin, which acquires instances located within the SVM margin, using the SVM-Margin method will not contribute to the acceleration of the learning process when larger numbers of mild conditions exist. The reason for this is that most of the severe conditions are not located inside the margin, but rather are located deep inside the positive side of the SVM – which is the reason for the effectiveness of the proposed AL methods. Note that the Combination_XA method is slightly more stable than

the Exploitation method with regard to accuracy, yet their results are comparable; however the comparison of different AL methods, and especially these two in particular, is not the focus of the current study.

We now present the results of the three core experiments of the current study.

5.1 Experiment 1 – Examining the Variance in the Intra-Labeler and Inter-Labelers' Learning Curves

Figure 10 (A–D) displays the results for each of the four learning methods. Each figure presents, the seven learning curves for one of the learning methods induced by the labels provided by the seven labelers, measured by AUC. The model induced by using the gold standard label is shown as well. As can be seen, several of the labelers consistently performed poorly, both in terms of their internal AUC variance over time (i.e., their *intra-labeler* variance), as well as their overall classification performance. The differences might stem from some of the labelers being less experienced, or less knowledgeable, in the task of condition severity classification. Using the gold standard label led to consistent, smooth, high performance curves in all four cases, although not necessarily the highest performance curve. However, it should be noted that such a situation is not uncommon in medical domains and reducing less desirable implications is one of our objectives. Thus, in the three experiments, we wanted to learn which learning method and learning strategy result in less learning curve performance variance, and thus reduce the dependence on using a specific labeler, or on stopping the learning process at a particular, arbitrary point in time.

Figure 11 displays the standard deviation among the models induced by each of the labelers, for each acquisition trial, for the four learning methods (11.A), and the mean standard deviation over all of the acquisition trials for these methods (11.B).

In figure 11(A), the standard deviation of the seven labelers' learning curves for each method is presented for each trial acquisition. The standard deviation increases slightly with the acquisition of additional conditions for the Combination_XA and Exploitation AL methods, but is larger for the SVM-Margin in the beginning and decreases as more labels are acquired; the random method follows the same pattern, but in this case the standard deviation was even larger. Note that after 20 trials the standard deviation of all of the methods is identical, as expected.

Figure 11(B) shows a box-plot visualization of the distribution of the *Inter*-labeler standard deviation values, for each of the four selection methods, among the seven different labelers, across all 20 trials. Combination_XA resulted in classification models that had the lowest *Inter*-labeler standard deviation (0.0197), followed closely by Exploitation, which resulted in models that had an inter-labeler standard deviation of 0.0209. At the same time, the *Inter*-labeler standard deviation of the models induced by SVM-Margin was the highest among the AL methods (0.029, almost 50% more than the standard deviation of the models induced by Combination_XA). However, the passive (Random selection) learning method resulted in an *Inter*-labeler standard deviation of 0.039, almost twice as much as the *Inter*-labeler standard deviation of the models induced using our advanced AL methods.

These results suggest that it is indeed possible to reduce the variance among the labelers' learning curves by using AL methods, and that the Exploitation and Combination_XA AL methods seem to be even less sensitive to the quality of the labeling than the established SVM-Margin AL method. We also examined the statistical differences between the learning curves of the active and passive learning methods. A single factor ANOVA statistical test on the standard *Inter*-labeler deviation for the Random selection method versus the SVM-Margin learning method resulted in a statistically significant difference ($p = 0.042$). A similar statistical test comparing the SVM-Margin and Exploitation methods resulted in a statistically insignificant difference ($p = 0.29$). The difference between Combination_XA and Exploitation was similarly insignificant ($p = 0.67$). To sum up, it seems that active learning methods are much more robust to the differences in clinical training and/or experience among the labelers, than the standard, passive learning method.

Figure 12-A presents the learning curves induced by the four selection methods from eight series of labels: from the labels provided by the seven labelers (L1, L2….L7), as well as the gold standard label. Note that in general, the AL methods demonstrated visibly smoother learning curves, compared to those produced when using the passive (Random selection) learning method. The curves induced when using the gold standard label also seem smoother, even when using the Random selection (passive) learning method. We examine these curves quantitatively in the next figure.

Figure 12-B displays the mean *intra-labeler* variance, over the 20 acquisition trials of the training phase, of the performance of the models induced from the labels provided by each labeler, for the seven labelers, as well as for the gold standard label, comparing the variance of the models induced using the passive learning method to the mean variance of the models induced using the three active learning methods.

Note that in cases in which the variance of the learning curves by the Random selection method was lower than the variance of the learning curves of the active learning, the differences are quite small; however when the variance of the AL method was lower than that of the Random selection method, the differences often manifested themselves as a variance that was greater by 30% to 100% than the variance of the models induced when using AL methods.

Our next figure focuses more explicitly on the difference in *intra*-labeler variance when using the three active learning methods and the passive learning method.

Figure 12-C presents a box-plot visualization of the *Intra*-labeler learning curves variance when using the passive (Random selection) learning method, and when using each of the three active learning methods. The box-plot integrates the models induced from the labels provided by the seven labelers and the gold standard label. Note that the *Intra*-labeler standard deviation of the Random selection learning method (mean: 0.0484; range: 0.0341to 0.0724) is much higher than the standard deviation values of the models induced when using the three AL methods (means: 0.0373, 0.0381, and 0.0383; range: 0.0128 to 0.0502).

A single factor ANOVA statistical test comparing the *Intra*-labeler standard deviation of the passive (Random selection) learning method and the three AL methods provided *p*-values of

10% and 6.9% for Exploitation and SVM-Margin, respectively, which was not statistically significant at the 5% alpha value we have set, despite the relatively high absolute differences between the standard deviations of the *Intra*-labeler AUC values. The difference in the *Intra*-labeler AUC standard deviation between the Combination_XA AL method and the Random selection method was, however, statistically significant ($p = 0.047$).

Finally, Figure 12-D displays explicitly, using a box-plot visualization, the distribution of the *intra*-labeler variance for all of the three AL methods (mean: 0.0379; range: [0.0182 to 0.0496]), compared to the variance when using the passive (Random selection) learning method (mean: 0.0484; range: [0.0275 to 0.0724]). Clearly, the *Intra*-labeler variance in the models' performance when using the AL methods is greatly decreased, compared to the variance when using passive learning.

A single factor ANOVA statistical test on the standard *Inter*-labeler standard deviation between the Random selection method and the AL methods demonstrated a statistically significant difference ($p = 0.049$).

5.2 Experiment 2 – Comparing the Differences in Inter-Labeler Variance between the Active and Passive Learning Methods during the Training Process

Figure 13 shows that the absolute differences in the Inter-labeler AUC standard deviation, between the AL methods and the passive (Random selection) learning method, increase and then decrease during the label acquisition process. These results support our original expectation, namely, that the difference (in standard deviation) between the AL methods and the passive (Random selection) learning method will initially grow. This might be expected, since AL methods converge faster towards the "real" classification model, so that the difference in the inter-labeler standard deviation of the AL methods compared to the passive learning method is expected to increase; then, however, the AL methods-passive learning method difference in inter-labeler variance should gradually decrease, as more and more data instances are acquired by both types of methods. Finally, both method types (AL and passive learning) converge when the model is fully knowledgeable, for all of the methods. Thus, the results demonstrate a U-shaped curve, in which the maximal difference occurs in the middle of the learning phase, and the gap eventually narrows to zero when all of the training instances have been acquired by each of the learning methods.

Understanding this behavior might be beneficial when trying to optimize the reduction in variance among models induced by different labelers.

5.3 Experiment 3 – Assessing the Effect of Using the Labelers' Group Consensus Label

Figure 14(A-D) shows the results of the three labeling strategies, namely using the gold standard label, versus the consensus (majority) label, to train the various learning methods, versus the expected AUC for a model induced by picking at random one of the labelers (as might happen when we have no prior notion of any differences among labelers, and decide to use one of them at random).

It is worthwhile to note that typically, using the gold standard labeler to train the model initially results, for all learning methods, in a model that has the highest AUC. However,

using the *consensus* (majority) label during the training phase eventually *resulted in the highest AUC*, regardless of the choice of learning method (including the standard passive learning method). This was followed by the AUC of the models induced by using the gold standard label, and finally, by the mean AUC of the seven labelers, representing the expected AUC of a model induced using a randomly selected single labeler.

Thus, it seems that ultimately, in the later stages of learning, the learning methods leveraged the consensus of the seven labelers and resulted in models whose AUC value was even slightly higher than that of the models learned from the gold standard labeler, and certainly always higher than the mean AUC of the classifiers induced using the seven labelers. Note that the passive (Random selection) learning method exploited the consensus of the labelers for a better performance only at the very end of the training phase. In the early training phase, the best learning curve was always obtained using the gold standard labeler, while the next best AUC usually resulted from using the consensus of the labelers.

In Figure 15 we can see another view of the results presented in Figure 14, in which the four selection methods were compared for each of the different labeler setups. This view directly compares the various learning methods, for each label type. We can see that when using labels provided by the gold standard labeler, Exploitation achieved the highest (by a small margin) AUC rate (92.6%) after 14 trials (60 conditions). For this label type, the passive (Random selection) learning method provided surprisingly good results compared with the AL methods. When using the consensus label, and when using a random labeler represented by the mean AUC of the seven labelers, we can see that the classifier induced by the passive (Random selection) learning method suffered from inconsistency, while the AL method induced models were characterized by a more consistent performance.

Figure 16-A shows the mean variance, for the four learning methods, of the model induced from the labels provided by the three labeling strategies (gold standard, consensus label, and a “random” labeler represented by the mean AUC of the seven labelers). Using the consensus label induced models with a much higher mean standard deviation (0.0499) of the AUC during the training phase, compared to the standard deviation when using the gold standard label (0.0388) and the standard deviation of the mean AUC of the seven labelers (0.0377).

The distribution of the standard deviation, when compared among the three labeling strategies for each of the four learning methods, demonstrates a similar pattern (Figure 16-B): For each learning method, and in particular, the Random selection (passive) method, the variance when using the consensus label is the highest, compared to the other two labeling strategies.

Using a paired t-test, the difference between the *intra*-labeler AUC standard deviation of the consensus labeling strategy and that of the other two labeling strategies was significant only when using the Random selection method ($p = 0.014$), but was insignificant when using any of the three AL methods.

However, much of the high intra-labeler variance in the performance of the models induced by using the consensus label seen in Figure 16A, might perhaps be due to the rather high

variance produced when inducing, from that label, a model by using the passive learning method, as shown in Figure 16-C. There was, in fact, a significant difference, found only when using the consensus label, between the AUC standard deviation of the Random selection method and that of the SVM-Margin AL method, using a t-test ($p = 0.039$).

6. DISCUSSION

In the current study, we focused on the issue of reducing the variance among the models induced using the various learning methods – both active and passive – and in particular, reducing the variance among those induced using the labels provided by different labelers for all of the instances used during the learning processes.

The use of AL in combination with the use of multiple labelers has recently become the focus of several preliminary investigations. In a recent study (77) performed by Zhang and Chaudhuri, the authors presented interesting ideas based on their comparison of using weak and strong labelers for medical images, in which they used strong but costly labelers, and also attempted to exploit weak labelers that occasionally misclassify images. Yan et al. (78) investigated the use of multiple labelers in cases in which an oracle is not available. Based on their strategy, which included several experts, the authors were able to consider which data sample should be labeled next, and which annotator should be queried to most benefit a learning model. Note, however, that in the current study we focused on the investigation of the integration of labels provided by several *labelers*, unlike the use of several *classifiers* that form *a committee of classifiers*, as used by the *Query by Committee* (*QBC*) methodology (78), which is a different AL approach.

Our current study focused on the rigorous analysis of one passive and several active learning processes, all based on the labels provided by seven different labelers, an investigation that, as far as we know, has not previously been performed. We analyzed the variance of the learning curves induced by the three AL methods and by the passive learning process, for the seven different labelers. In addition to the established *SVM-Margin* AL method, we again used the two novel AL methods that we developed as part of the *CAESAR-ALE* framework, and we used the standard passive learning method (*Random selection*) as a baseline.

We focused on the analysis of the *Intra*-labeler and Inter-labeler AUC variance, when using different learning methods and different types of labels.

The use of the AL methods, as was demonstrated in Experiment 1, resulted, as we conjectured, as follows (a) significantly reduced *intra-labeler* variance in the performance of the induced models during the training phase, as measured by their AUC, reduces the dependence on stopping at the right point in time during a lengthy learning process, since it reduces the concern about halting the process at a local minimum that is significantly different in performance from the rest of the models; and (b) significantly reduced *Inter*-labeler performance variance reduces the dependence on the use of a particular labeler when an expert labeler is selected at random. Both results might stem from the fact that, as shown

in our previous experiments, classification models induced using AL methods converge faster to the “true” classification model that has the maximal AUC.

The results of Experiment 2 supported our original conjecture, namely, that the difference between the *inter*-labeler variance of the AL methods and the *inter*-labeler variance of the passive learning method during the training phase will initially grow, since AL methods converge faster towards the “true” classification model; but the difference in variance will start to decrease, as more and more data is shown to both types of methods. Ultimately, of course, the variance values of both the active and passive learning methods converge, when all of the data is known for all types of learning methods and labelers. Thus, the results demonstrated an inverse U-shaped curve, in which the maximal gap occurs in the middle of the learning phase, and eventually narrows to zero when all of the training instances have been acquired by all learning methods.

Finally, using the consensus (majority) label during the training phase produced a learning curve with higher mean variance across all four learning methods, especially when using the passive (Random selection) learning method, but eventually resulted (during the training phase) in the highest AUC, regardless of the choice of learning method (including the standard passive learning method). This value was at least as high as the AUC of the classifier that resulted from using (during the training phase) the gold standard label, and was certainly higher than the mean AUC of the seven labelers. One could consider the mean AUC of the labelers as representing the *expectation* of the AUC of a model induced by a labeler who is *randomly selected* from the labelers’ group. (It might be worthwhile to view this expectation as representing the common case of using one expert, whose labeling performance is unknown).

This somewhat surprising “crowdsourcing” result when using a consensus label, might mean that although the different labelers had varying levels of medical training and experience with labeling conditions, their majority consensus label was eventually at least as useful, and possibly even slightly better, than the gold standard label, for the practical purpose of classifying the severity of a set of new medical conditions.

It might be surprising that using a consensus label can result in a model that can classify new, unseen instances (whose label is determined by the gold standard labeler), even slightly better than the model learned by only using the gold standard label. However, the difference in our experiments was very slight; and it might well be due to a few borderline training cases in which using the consensus labels during the learning phase (instead of the actual “true” label) actually had a “smoothing” effect on the resulting induced model, which slightly enhanced its accuracy on the testing set.

In any case, the effectiveness of using the consensus label during the training phase is highly encouraging, since it might mean (a) that the use of the consensus label agreed upon by a rather uneven group of labelers might be at least as good as using a gold standard labeler, who might not be available, and (b) that using the consensus label will certainly be better than randomly selecting one of the group’s individual labelers (when nothing is known about

the individual labeler's quality) whose expected performance was represented by the mean AUC of the labeling committee.

The difference between the intra-labeler AUC standard deviation of the consensus labeling strategy and that of the other two labeling strategies was significant only when using the Random selection method, but was insignificant when using any of the three AL methods. Furthermore, a significant difference between the AUC standard deviation of the Random selection method and that of the SVM-Margin AL method was found only when using the consensus label and not the other two labeling strategies. This result suggests that the use of AL methods can reduce the variance between models induced using different labeling strategies, and in particular, between models induced using the consensus label and models induced using the gold standard label. It might also provide more flexibility in the common practical situation, in which a gold standard labeler, such as the established medical expert in some domain, is unavailable, and using the consensus of a committee of labelers of an uneven composition is preferable.

Finally, the effectiveness of using the consensus label during the learning phase also raises an intriguing option for handling the common situation in which several training instances that display the same set of features are labeled differently at different times during the learning phase, even by the *same labeler*. Instead of discarding such instances as manifesting a human error, or using a probabilistic interpretation of their label, one could consider the single labeler to have represented, at different points of time, several differing domain expert opinions, perhaps representing several different schools of thought, and thus use his/her "consensus" (majority) label, possibly with the same useful effect.

7. CONCLUSIONS AND FUTURE WORK

The CAESAR-ALE framework, which uses active learning methods, and more particularly, our two new AL methods, were found to be more efficient in reducing the *Intra*-labeler and *Inter*-labeler variance among the models induced by using the labels of different experts, compared to the use of an existing AL method, and in particular, to the use of a passive (Random selection) learning method.

Our experiments also suggest that although it leads to greater volatility of the performance of the learned model, using the consensus of a highly uneven labeling committee eventually results in a classifier with the highest AUC, regardless of the choice of learning method (including the standard passive learning method), compared to the use of the gold standard label and the mean AUC of the seven labelers. Thus, using the consensus label to induce a model during the training phase proved, in our experiments, to always be better than using a randomly selected individual labeler, and was at least as effective for this purpose as using the gold standard labeler.

In our future work, we would like to examine the effect of using a majority consensus to perform the actual *classification* of new cases, as opposed to the *labeling*, by using the multiple models induced by different labelers in a manner similar to ensemble classification. We also intend to apply our CAESAR-ALE multiple labeler framework to additional

important medical domains, in order to reduce the labeling efforts of medical experts and to better leverage the efforts of a group of experts to improve the classification performance of the induced classifier. To assess the full value of this methodology we would also like to use our framework for larger datasets and with a larger number of labelers.

Acknowledgments

We thank the Malware Lab at the Ben-Gurion University's Cyber Security Research Center (CSRC) and the Medical Informatics Research Center for their support of this research. We also thank all of the labelers at Columbia University who patiently labeled condition severity status for our dataset.

This research was paritally supported by the National Cyber Bureau of the Israeli Ministry of Science, Technology and Space. Support for portions of this research was provided by **R01 LM006910** (GH) and **R01 GM107145** (NPT). MRB is supported by the National Library of Medicine training grant **T15 LM00707**.

List of Abbreviations

CAESAR	Classification **A**pproach for **E**xtracting **S**everity **A**utomatically from Electronic Health **R**ecords
CAESAR-ALE	Classification **A**pproach for **E**xtracting **S**everity **A**utomatically from Electronic Health **R**ecords – Active Learning Enhancement
EHR	Electronic Health Record
AL	Active Learning
SVM	Support Vector Machines
VS	Version Space
SNOMED-CT	Systemized Nomenclature of Medicine-Clinical Terms
ICD-9	International Classification of Diseases - Version 9
SVM-Margin	Support Vector Machines-Margin Method - An existing AL method oriented towards acquiring informative conditions that lie closest to the separating hyperplane (inside the margin)
Exploitation	An AL method included in the CAESAR-ALE framework that is oriented towards acquisition of severe conditions
Combination_XA	An AL method included in the CAESAR-ALE framework that combines elements of the Exploitation method and the SVM-Margin method, so that it applies a hybrid acquisition strategy for enhanced improvement of the CAESER method

References

1. Stang PE, Ryan PB, Racoosin JA, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Ann Intern Med. 2010 Nov 2; 153(9): 600–6. [PubMed: 21041580]
2. Kho AN, Pacheco JA, Peissig PL, et al. Electronic medical records for genetic research: results of the eMERGE consortium. Science translational medicine. 2011 Apr 20.3(79):79re1.
3. Denny JC, Ritchie MD, Basford MA, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. Bioinformatics. 2010 May 1; 26(9):1205–10. [PubMed: 20335276]
4. Boland MR, Hripcsak G, Shen Y, Chung WK, Weng C. Defining a comprehensive verotype using electronic health records for personalized medicine. J Am Med Inform Assoc. 2013 Dec 1; 20(e2):e232–e8. [PubMed: 24001516]
5. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013; 20(1):144–51. [PubMed: 22733976]
6. Hripcsak G, Knirsch C, Zhou L, Wilcox A, Melton GB. Bias associated with mining electronic health records. Journal of biomedical discovery and collaboration. 2011:6–48.
7. Hripcsak G, Albers DJ. Correlating electronic health record concepts with healthcare process events. J Am Med Inform Assoc. 2013 Dec 1; 20(e2):e311–e8. [PubMed: 23975625]
8. Rich P, Scher RK. Nail psoriasis severity index: a useful tool for evaluation of nail psoriasis. Journal of the American Academy of Dermatology. 2003; 49(2):206–12. [PubMed: 12894066]
9. Bastien CH, Vallières A, Morin CM. Validation of the Insomnia Severity Index as an outcome measure for insomnia research. Sleep Medicine. 2001; 2(4):297–307. [PubMed: 11438246]
10. McLellan AT, Kushner H, Metzger D, et al. The fifth edition of the Addiction Severity Index. Journal of substance abuse treatment. 1992; 9(3):199–213. [PubMed: 1334156]
11. Rockwood TH, Church JM, Fleshman JW, et al. Patient and surgeon ranking of the severity of symptoms associated with fecal incontinence. Diseases of the colon & rectum. 1999; 42(12):1525–31. [PubMed: 10613469]
12. Horn SD, Horn RA. Reliability and validity of the severity of illness index. Medical care. 1986; 24(2):159–78. [PubMed: 3080648]
13. Boland, MR., Tatonetti, N., Hripcsak, G. Intelligent Systems for Molecular Biology Phenotype Day. Boston, MA: 2014. CAESAR: a Classification Approach for Extracting Severity Automatically from Electronic Health Records; p. 1-8.In Press
14. Elkin, PL., Brown, SH., Husser, CS., et al. Mayo Clinic Proceedings. Elsevier; 2006. Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists; p. 741-8.2006
15. Stearns, MQ., Price, C., Spackman, KA., Wang, AY. SNOMED clinical terms: overview of the development process and project status; Proceedings of the AMIA Symposium; 2001: American Medical Informatics Association; 2001. p. 662
16. Elhanan G, Perl Y, Geller J. A survey of SNOMED CT direct users, 2010: impressions and preferences regarding content and quality. Journal of the American Medical Informatics Association. 2011 Dec 1; 18(Suppl 1):i36–i44. 2011. [PubMed: 21836159]
17. Zmiri, Dror, Yuval, Shahar, Meirav, Taieb-Maimon. Classification of patients by severity grades during triage in the emergency department using data mining methods. Journal of evaluation in clinical practice 18.2. 2012:378–388.
18. HCUP Chronic Condition Indicator for ICD-9-CM. Healthcare Cost and Utilization Project (HCUP). 2011. http://www.hcup-us.ahrq.gov/toolssoftware/chronic/chronic.jsp Accessed on February 25 2014
19. Hwang W, Weller W, Ireys H, Anderson G. Out-Of-Pocket Medical Spending For Care Of Chronic Conditions. Health Affairs. 2001 Nov 1; 20(6):267–78. 2001.
20. Chi, M-j, Lee, C-y, Wu, S-C. The prevalence of chronic conditions and medical expenditures of the elderly by chronic condition indicator (CCI). Archives of Gerontology and Geriatrics. 2011; 52(3): 284–9. [PubMed: 20452688]

21. Perotte A, Pivovarov R, Natarajan K, Weiskopf N, Wood F, Elhadad N. Diagnosis code assignment: models and evaluation metrics. Journal of the American Medical Informatics Association. 2014 Mar 1; 21(2):231–7. 2014. [PubMed: 24296907]
22. Perotte A, Hripcsak G. Temporal properties of diagnosis code time series in aggregate. IEEE journal of biomedical and health informatics. 2013 Mar; 17(2):477–83. [PubMed: 24235118]
23. Torii M, Wagholikar K, Liu H. Using machine learning for concept extraction on clinical documents from multiple data sources. Journal of the American Medical Informatics Association. 2011 Jun 27. 2011.
24. Nguyen AN, Lawley MJ, Hansen DP, et al. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. Journal of the American Medical Informatics Association. 2010 Jul 1; 17(4):440–5. 2010. [PubMed: 20595312]
25. Nissim N, Moskovitch R, Rokach L, Elovici Y. Novel active learning methods for enhanced PC malware detection in windows OS. Expert Systems with Applications. 2014 Oct 1; 41(13):5843–57.
26. Nissim N, Moskovitch R, Rokach L, Elovici Y. Detecting Unknown Computer Worm Activity via Support Vector Machines and Active Learning. Pattern Analysis and Applications. 2012; 15:459–75.
27. Nissim N, Cohen A, Glezer C, Elovici Y. Detection of malicious PDF files and directions for enhancements: A state-of-the art survey. Computers & Security. 2015; 48:246–66.
28. Angluin D. Queries and concept learning. Machine Learning. 1988; 2:319–42.
29. Lewis, D., Gale, W. A sequential algorithm for training text classifiers; Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval; Springer-Verlag. 1994. p. 3-12.
30. Liu Y. Active learning with support vector machine applied to gene expression data for cancer classification. Journal of chemical information and computer sciences. 2004; 44(6):1936–41. [PubMed: 15554662]
31. Warmuth MK, Liao J, Rätsch G, Mathieson M, Putta S, Lemmen C. Active learning with support vector machines in the drug discovery process. Journal of chemical information and computer sciences. 2003; 43(2):667–73. [PubMed: 12653536]
32. Figueroa RL, Zeng-Treitler Q, Ngo LH, Goryachev S, Wiechmann EP. Active learning for clinical text classification: is it better than random sampling? Journal of the American Medical Informatics Association. 2012 amiajnl-2011–000648.
33. Nguyen DH, Patrick JD. Supervised machine learning and active learning in classification of radiology reports. Journal of the American Medical Informatics Association. 2014 amiajnl-2013–002516.
34. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST). 2011; 2(3):27.
35. Tong S, Koller D. Support vector machine active learning with applications to text classification. Journal of Machine Learning Research. 2000–2001; 2:45–66.
36. Ralf H, Graepel T, Campbell C. Bayes point machines. The Journal of Machine Learning Research. 2001; 1:245–79.
37. Nissim N, Moskovitch R, Rokach L, Elovici Y. Novel Active Learning Methods for Enhanced PC Malware Detection in Windows OS. Expert Systems With Applications. 2014; 41(13)
38. Nissim N, Moskovitch R, Rokach LYE. Detecting unknown computer worm activity via support vector machines and active learning. Pattern Analysis and Applications. 2012; 15(4):459–75.
39. Moskovitch, R., Nissim, N., Elovici, Y. Malicious code detection using active learning; ACM SIGKDD Workshop in Privacy, Security and Trust in KDD; Las Vegas. 2008.
40. Moskovitch R, Stopel D, Feher C, Nissim N, Japkowicz N, Elovici Y. Unknown Malcode Detection and The Imbalance Problem. Journal in Computer Virology. 2009; 5(4)
41. Nissim N, Cohen A, Moskovitch R, et al. ALPD: Active Learning Framework for Enhancing the Detection of Malicious PDF Files Aimed at Organizations. Proceedings of JISIC. 2014
42. Baram Y, El-Yaniv R, Luz K. Online choice of active learning algorithms. Journal of Machine Learning Research. 2004; 5:255–91.

43. Herman, R. 72 Statistics on Hourly Physician Compensation. http://www.beckershospitalreview.com/compensation-issues/72-statistics-on-hourly-physician-compensation.html 2013; Accessed in January 2015
44. Boland, MR., Tatonetti, NP. Are All Vaccines Created Equal? Using Electronic Health Records to Discover Vaccines Associated With Clinician-Coded Adverse Events; AMIA Summits on Translational Science Proceedings 2015; San Francisco, CA, USA. 2015. p. 196-200.
45. Boland, MR., Tatonetti, NP., Hripcsak, G. Intelligent Systems for Molecular Biology Phenotype Day. Boston, MA: 2014. CAESAR: a Classification Approach for Extracting Severity Automatically from Electronic Health Records.
46. Vapnik, V. Statistical learning theory. New York: Springer; 1998.
47. Boland MR, Tatonetti NP, Hripcsak G. Development and Validation of a Classification Approach for Extracting Severity Automatically from Electronic Health Records. Journal of Biomedical Semantics. 2015; 6(14)
48. Moskovitch R, Cohen-Kashi S, Dror U, Levy I, Maimon A, Shahar Y. Multiple hierarchical classification of free-text clinical guidelines. Artificial Intelligence in Medicine. 2006; 37:177–190. [PubMed: 16730962]
49. Nissim, N., Boland, MR., Moskovitch, R., Tatonetti, NP., Elovici, Y., Shahar, Y., Hripcsak, G. Artificial Intelligence in Medicine. Springer International Publishing; 2015. An Active Learning Framework for Efficient Condition Severity Classification; p. 13-24.(AIME-15)
50. Vapnik, VN. Estimation of dependences based on empirical data. Vol. 41. New York: Springer-Verlag; 1982.
51. Joachims T. Making large scale SVM learning practical. 1999
52. Burges CJ. A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery. 1998; 2(2):121–167.
53. Berthold, Michael. The Fog of Data: Data Exploration in the Life Sciences; Invited talk at 11th AIME conference 2007 In Artificial Intelligence in Medicine;
54. Nicolas, Cebron, Berthold, Michael R. Active learning for object classification: from exploration to exploitation. Data Mining and Knowledge Discovery April 2009. Jul 27; 2008 18(2):283–299.
55. Moskovitch, Robert, Hessing, Alon, Shahar, Yuval. Vaidurya - A Concept-Based; Context-Sensitive Search Engine For Clinical Guidelines, MedInfo 2004; San Francisco, USA. 2004.
56. Moskovitch, Robert, Martins, Suzana, Behiri, Eytan, Weiss, Aviram, Shahar, Yuval. A Comparative Evaluation of a Full-text, Concept Based, and Context Sensitive Search Engine. Journal Of American Medical Informatics Association. 2007; 14:164–174.
57. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics. 2012; 13(6)
58. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. International Journal of Medical Informatics. 2008; 77(2)
59. Batal, I., Fradkin, D., Harrison, J., Moerchen, F., Hauskrecht, M. Mining Recent Temporal Patterns for Event Detection in Multivariate Time Series Data; Proceedings of Knowledge Discovery and Data Mining (KDD); Beijing, China. 2012.
60. Moskovitch R, Shahar Y. Fast Time Intervals Mining Using Transitivity of Temporal Relations. Knowledge and Information Systems. 2015a; 42(1)
61. Ng K, Ghoting A, Steinhubl SR, Stewart WF, Malin B, Sun J. PARAMO: A PARAllel predictive MOdeling Platform for Healthcare Analytic Research Using Electronic Health Records. Journal of Biomedical Informatics. 2014; 48:160–170. [PubMed: 24370496]
62. Sun J, McNaughton CD, Zhang P, Perer A, Gkoulalas-Divanis A, Denny JC, Kirby J, Lasko T, Saip A, Malin BA. Predicting Changes in Hypertension Control Using Electronic Health Records from a Chronic Disease Management Program. Journal of the American Medical Informatics Association. 2014; 21:337–344. [PubMed: 24045907]
63. Hripcsak, G. Physics of the Medical Record: Handling Time in Health Record Studies; Artificial Intelligence in Medicine; Pavia, Italy. 2015.
64. Rana S, Gupta S, Phung D, Venkatesh S. A predictive framework for modeling healthcare data with evolving clinical interventions. Statistical Analysis and Data Mining. In Press.

65. Moskovitch R, Shahar Y. Classification of Multivariate Time Series via Temporal Abstraction and Time Intervals Mining. Knowledge and Information Systems. 2015b; 45(1):35–74.
66. Moskovitch R, Shahar Y. Classification Driven Temporal Discretization of Multivariate Time Series. Data Mining and Knowledge Discovery. 2015c; 29(4):871–913.
67. Huang Z, Dong W, Bath P, Ji L, Duan H. On Mining Latent Treatment Patterns from Electronic Medical Records. Data Mining and Knowledge Discovery. 2015; 29:914–949.
68. Nissim, Nir, Moskovitch, Robert, BarAd, Oren, Rokach, Lior, Elovici, Yuval. ALDROID: Efficient Update of Android Anti-Virus Software Using Designated Active Learning Methods. Knowledge and Information System. In Press.
69. Moskovitch, R., Nissim, N., Elovici, Y. Malicious Code Detection and Acquisition Using Active Learning; IEEE International Conference on Intelligence and Security Informatics (IEEE ISI-2007); Rutgers University, New Jersey, USA. 2007.
70. Nissim, Nir, Cohen, Aviad, Moskovitch, Robert, Barad, Oren, Edry, Mattan, Shabatai, Assaf, Elovici, Yuval. ALPD: Active Learning Framework for Enhancing the Detection of Malicious PDF Files; Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint; 2014. p. 91-98.
71. Moskovitch, R., Nissim, N., Englert, R., Elovici, Y. Detection of Unknown Computer Worms Activity using Active Learning; The 11th International Conference on Information Fusion; Cologne, Germany. 2008.
72. Moskovitch, Robert, Choi, Hyunmi, Hripsack, George, Tatonetti, Nicholas. Prognosis of Clinical Procedures with Temporal Patterns and One Class Feature Selection; ACM/IEEE Transactions on Computational Biology and Bioinformatics; In Press
73. Zheng, Yaling, Scott, Stephen, Deng, Kun. Active learning from multiple noisy labelers with varied costs; Data Mining (ICDM), 2010 IEEE 10th International Conference on. IEEE; 2010.
74. Moskovitch R, Shahar Y. Vaidurya: a multiple-ontology, concept-based, context-sensitive clinical-guideline search engine. Journal of biomedical informatics. 2009 Feb; 42(1):11–21. [PubMed: 18721900]
75. Roy, Nicholas, McCallum, Andrew. Toward optimal active learning through Monte Carlo estimation of error reduction; Proceedings of the 18th International Conference on Machine Learning, (ICML 2001); Williamstown. 2001. p. 441-448.
76. Nissim, Nir, Boland, Mary Regina, Tatonetti, Nicholas P., Elovici, Yuval, Hripcsak, George, Shahar, Yuval, Moskovitch, Robert. Improving condition severity classification with an efficient active learning based framework. Journal of Biomedical Informatics. Jun.2016 61:44–54. ISSN 1532–0464. [PubMed: 27016383]
77. Zhang, C., Chaudhuri, K. Active Learning from Weak and Strong Labelers; Proceedings of Advances in Neural Information Processing Systems 28 (NIPS 2015); 2015. p. 703-711.
78. Yan, Y., Fung, GM., Rosales, R., Dy, JG. Active learning from crowds; Proceedings of the 28th international conference on machine learning, (ICML-2011); p. 1161-1168.
79. Moskovitch, Robert, Wang, Fei, Walsh, Colin, Hripcsak, George, Tatonetti, Nicholas. Prediction of Outcome Events via Time Intervals Mining; IEEE International Conference on Data Mining (ICDM); Atlantic City, USA. 2015.
80. Boland, Mary Regina, Jacunski, Alexandra, Lorberbaum, Tal, Romano, Joseph, Moskovitch, Robert, Tatonetti, Nicholas P. Systems Biology Approaches for Identifying Adverse Drug Reactions and Elucidating Their Underlying Biological Mechanisms. Wiley Interdisciplinary Reviews: Systems Biology and Medicine. In Press.
81. Moskovitch, Robert, Nissim, Nir, Elovici, Yuval. Acquisition of malicious code using active learning. Proc. 2nd Int'l Workshop on Privacy, Security, & Trust in KDD; 2008.
82. Nissim N, Cohen A, Moskovitch R, Shabtai A, Edri M, BarAd O, Elovici Y. Keeping pace with the creation of new malicious PDF files using an active-learning based detection framework. Security Informatics. 2016; 5(1):1.
83. Nissim, Nir, Cohen, Aviad, Elovici, Yuval. ALDOCX: Detection of Unknown Malicious Microsoft Office Documents Using Designated Active Learning Methods Based on New Structural Feature Extraction Methodology. IEEE Transactions on Information Forensics and Security. Mar; 2017 12(3):631–646. DOI: 10.1109/TIFS.2016.2631905

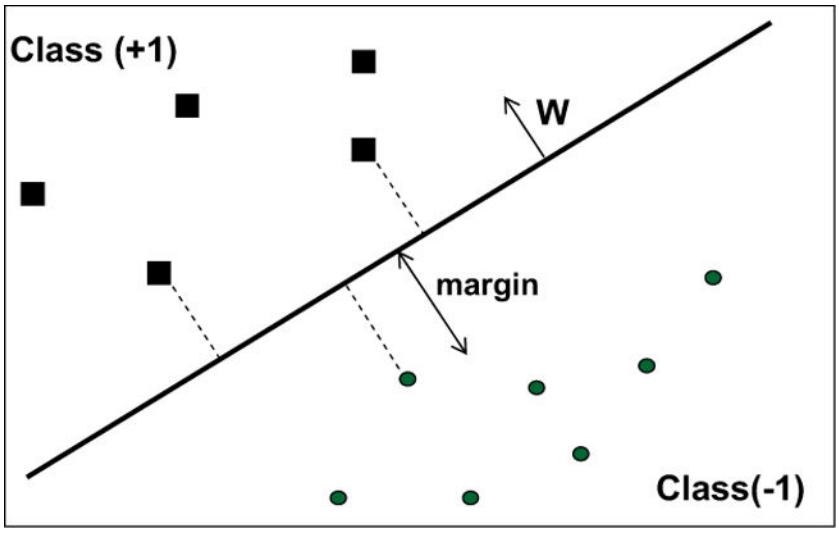

Figure 1.
An SVM with a maximal margin which separates the training set into two classes in a two-dimensional space (two features).

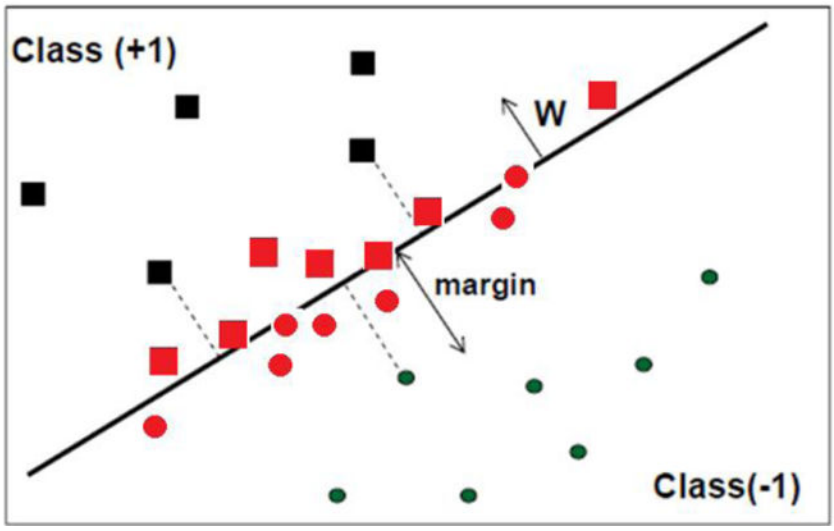

Figure 2.
The examples (colored in red) that will be selected according to the SVM-Margin AL method's criteria.

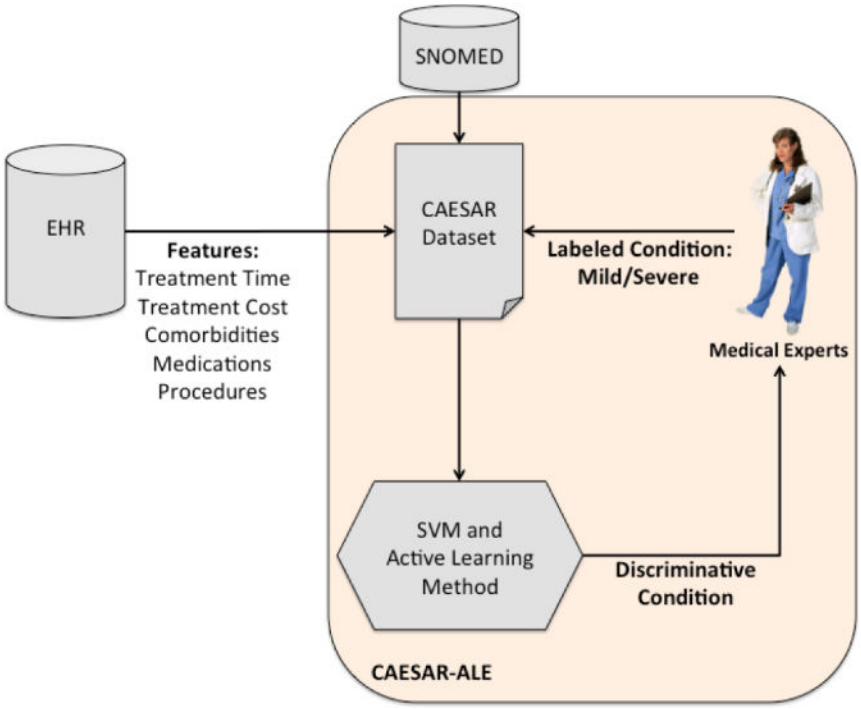

Figure 3.

The process of using AL methods to detect discriminative conditions requiring medical expert labeling.

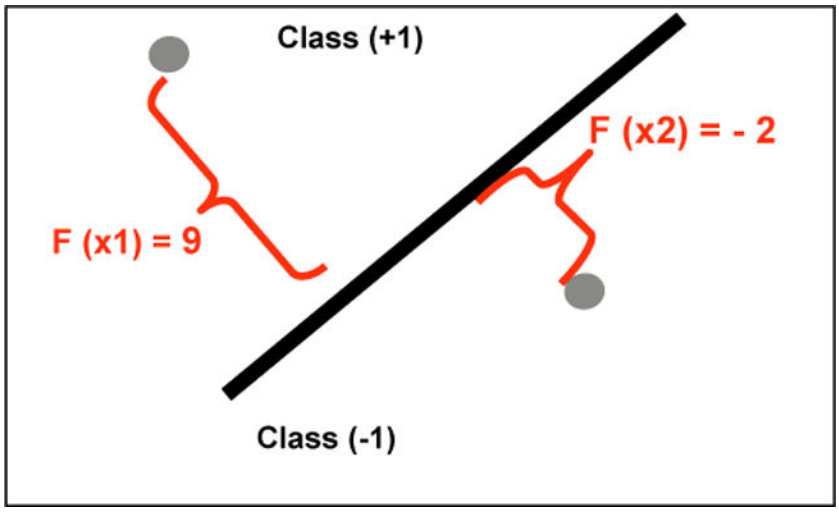

Figure 4.
Decision values given to two examples.

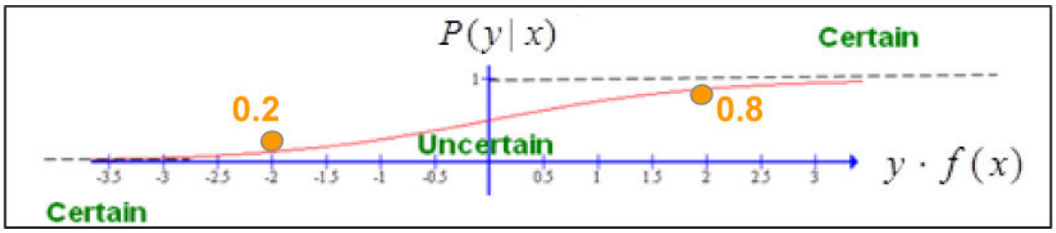

Figure 5.

Analysis of Equation 7 - the larger the distance the example is from the separating hyperplane, the higher the probability and the more confidence of the classifier.

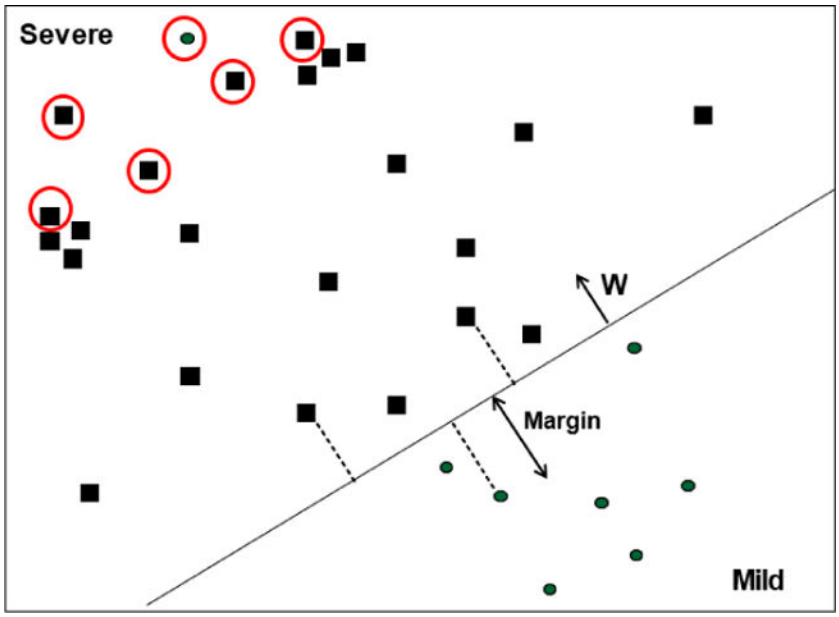

Figure 6.1.
An illustration showing the Exploitation method's criteria for acquiring new severe conditions.

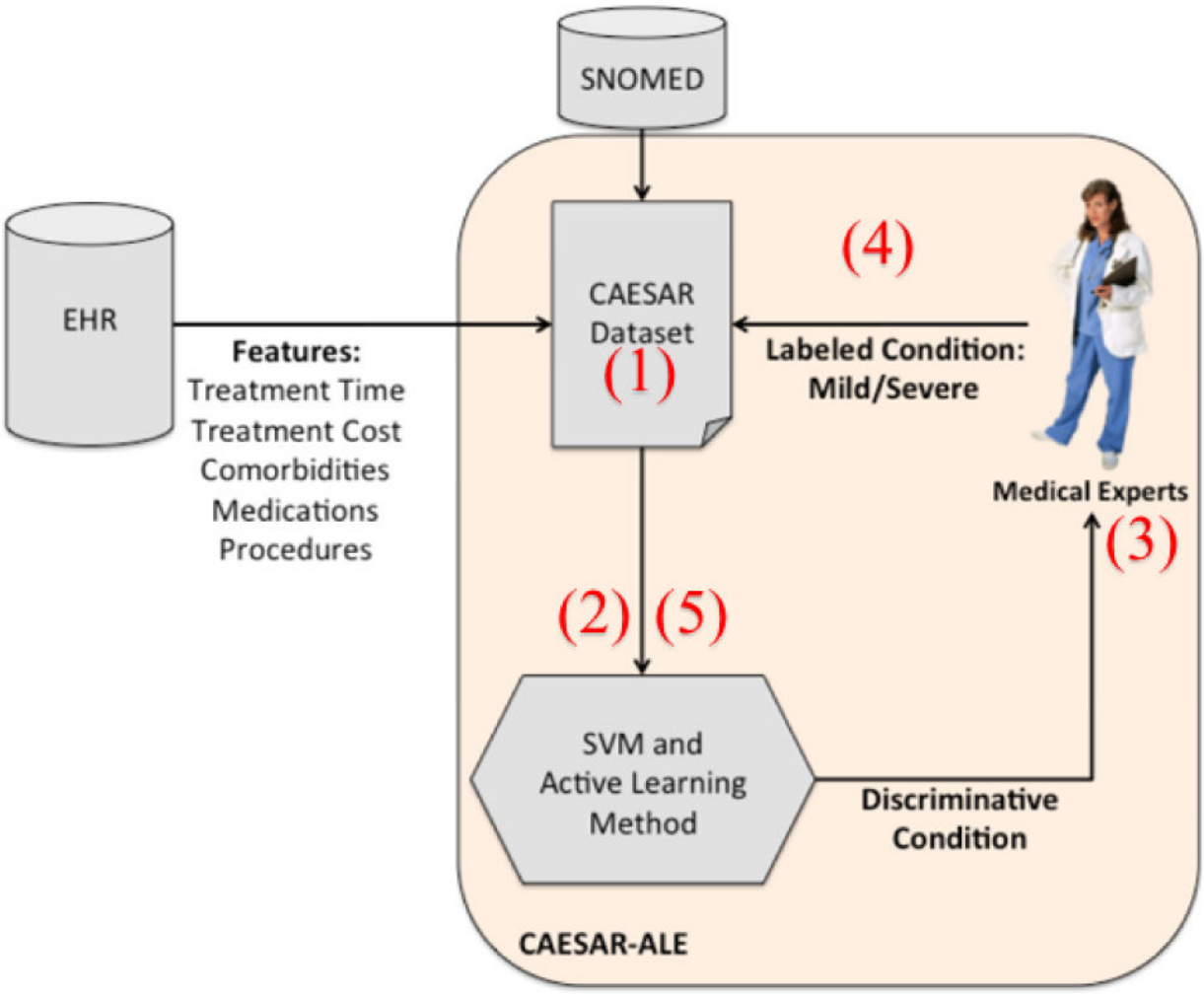

Figure 6.2.

The process and steps (1–5) of CAESAR-ALE - using AL methods to detect discriminative conditions requiring medical expert labeling.

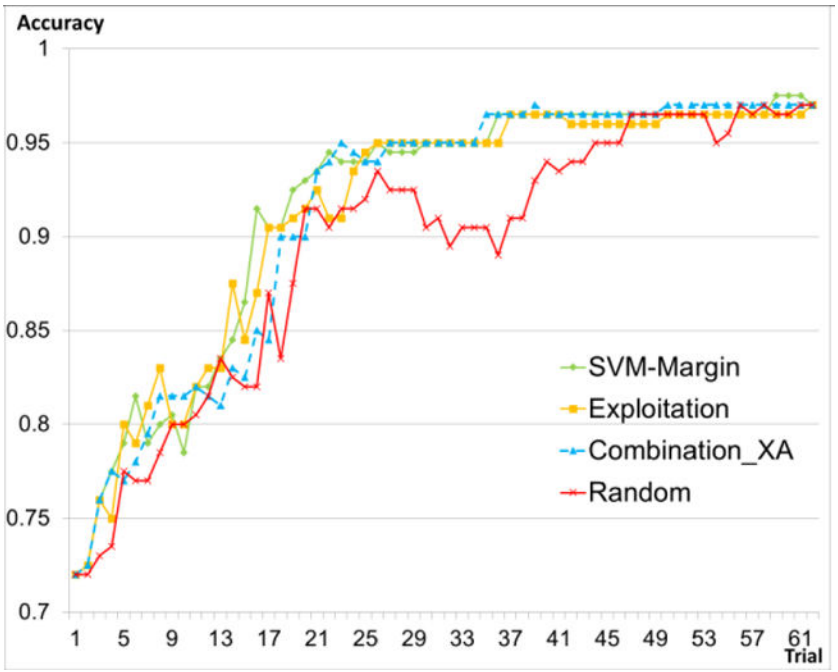

Figure 7.

The accuracy of the CAESAR-ALE models induced using the two new active learning methods versus the models induced using the SVM-Margin and the passive (Random selection) method, over 62 trials (five conditions acquired during each trial).

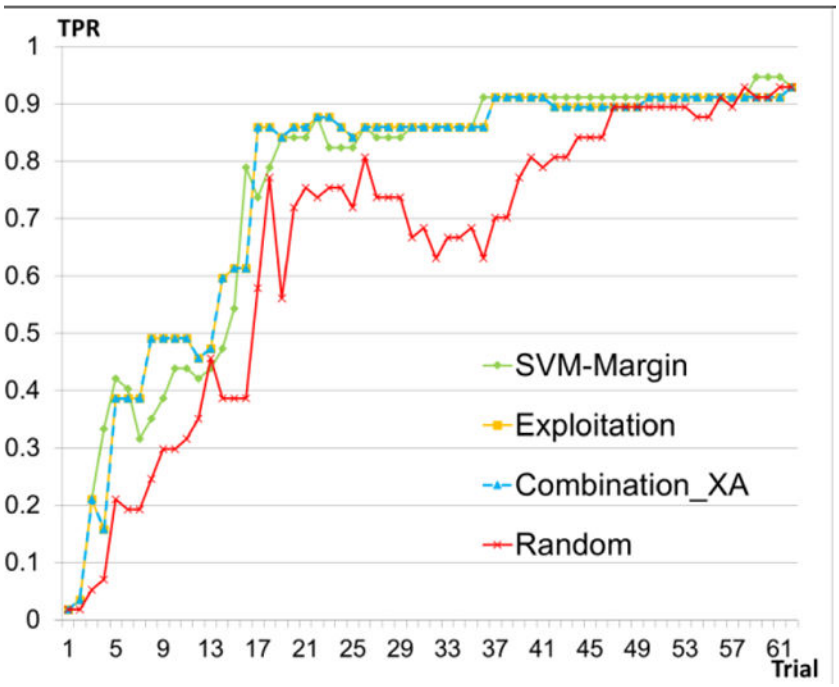

Figure 8.
TPR for active learning and random selection methods over 62 trials.

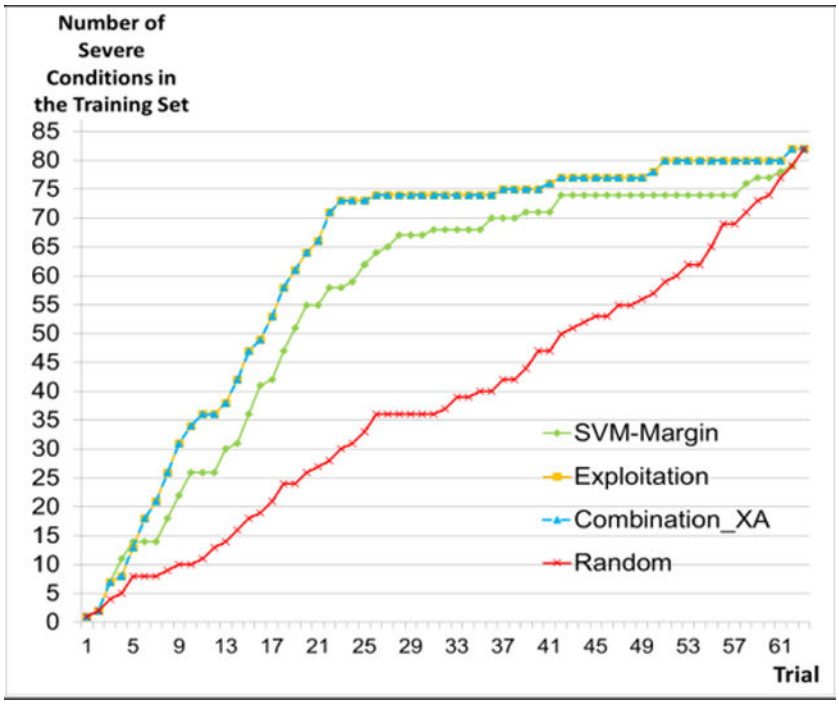

Figure 9.

The accumulated number of severe conditions acquired in the training set by each AL method over 62 trials.

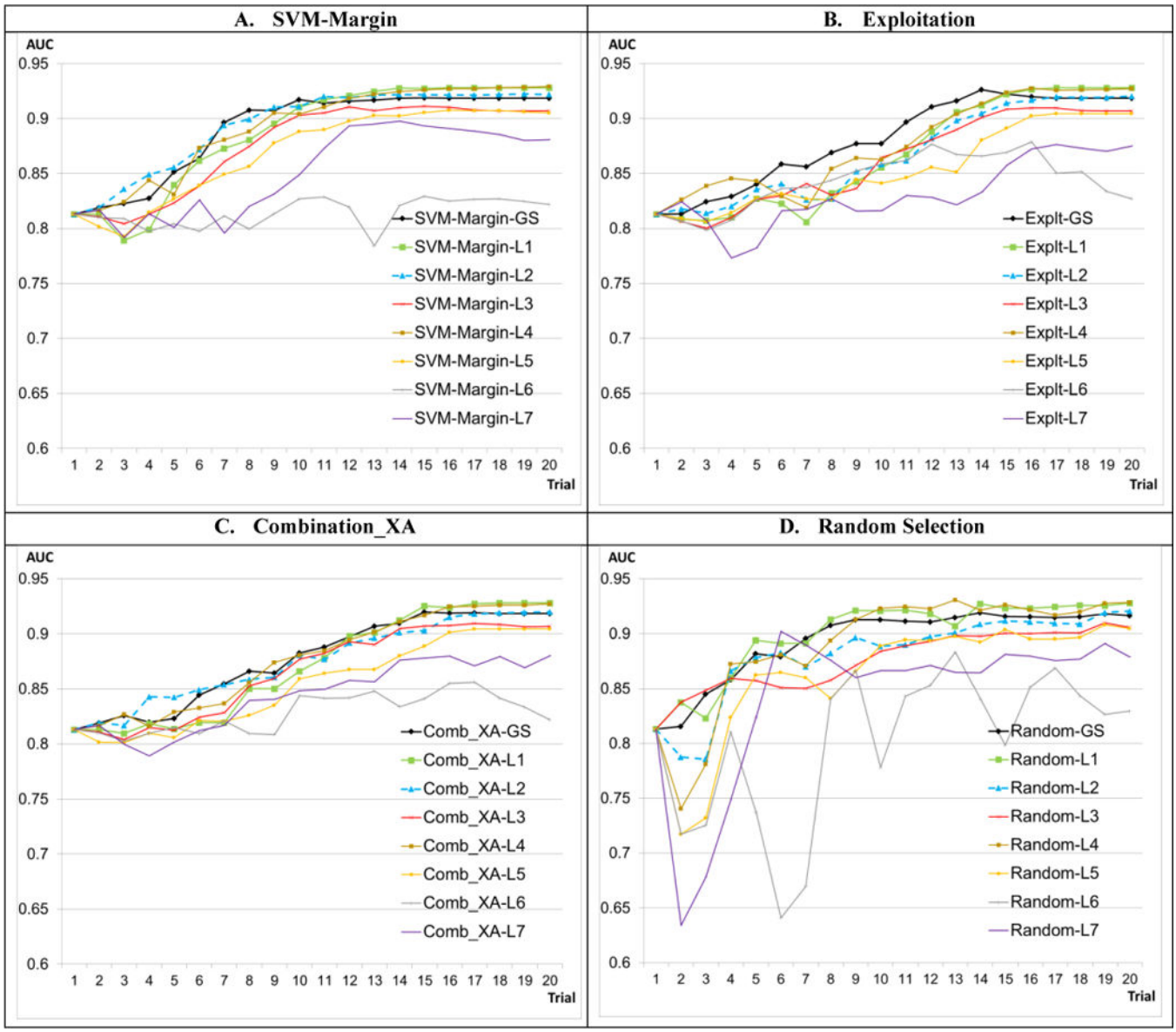

Figure 10.

The learning curves of the three active learning methods and of the passive (Random selection) learning method, by using the labels provided by the labelers and the gold standard (GS) label.

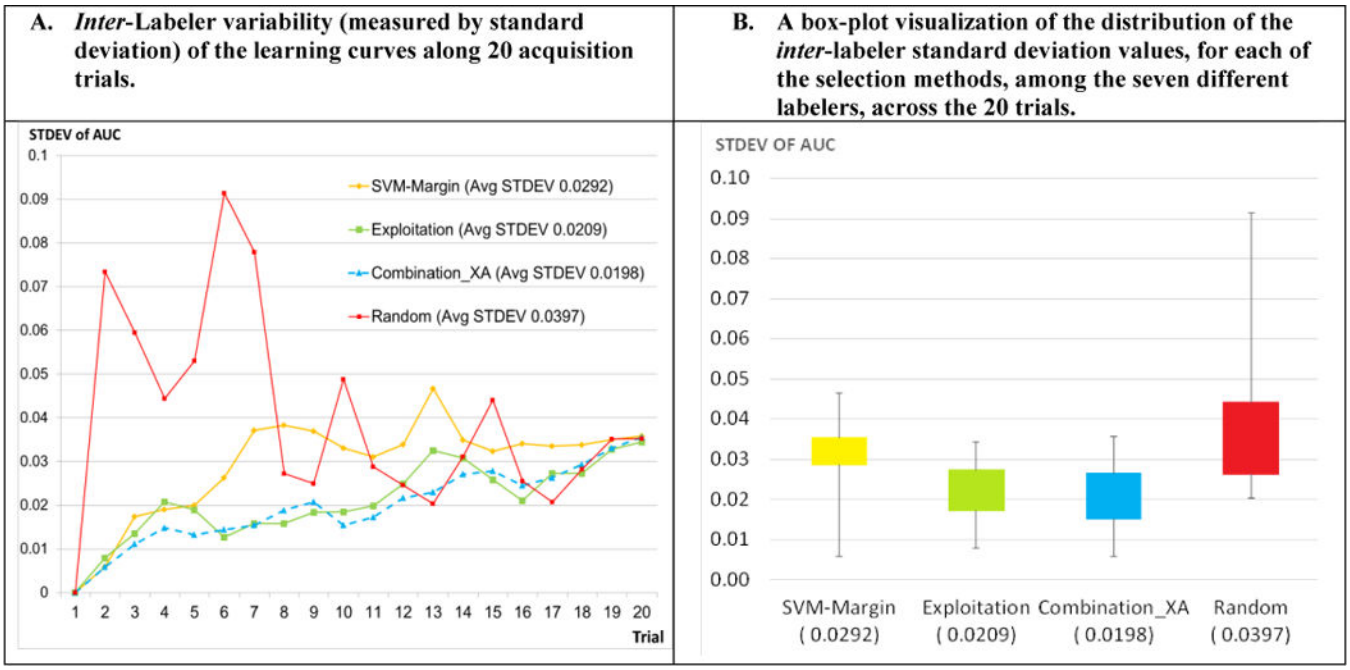

Figure 11.

Inter-labeler variability of the four learning methods. The standard deviation among the seven models induced by each of the seven labelers after each data acquisition trial is plotted across the 20 acquisition trials, for each of the four learning methods (11-A). A box-plot visualization displays the distribution of the standard deviation values, among the seven labelers, over the 20 acquisition trials for these methods (11-B). Each box's lower and upper boundaries denote the 25th and 75th percentiles; the whiskers denote the absolute minimal and maximal values. The mean *Inter*-labeler standard deviation value across the 20 trials, for each of the four methods, appears in parentheses below the name of each of the methods shown in the box-plot visualization.

Figure 12-A.

The learning curves, measured as the area under the curve (AUC) values, of the models induced from the labels provided by each of the seven labelers and gold standard label, for each selection method (three AL methods and the passive [Random selection] method) and the *Intra*-labelers' variance, represented by the mean standard deviation of the models induced from every labeler across each of the four selection methods, over his/her performance during the acquisition phase.

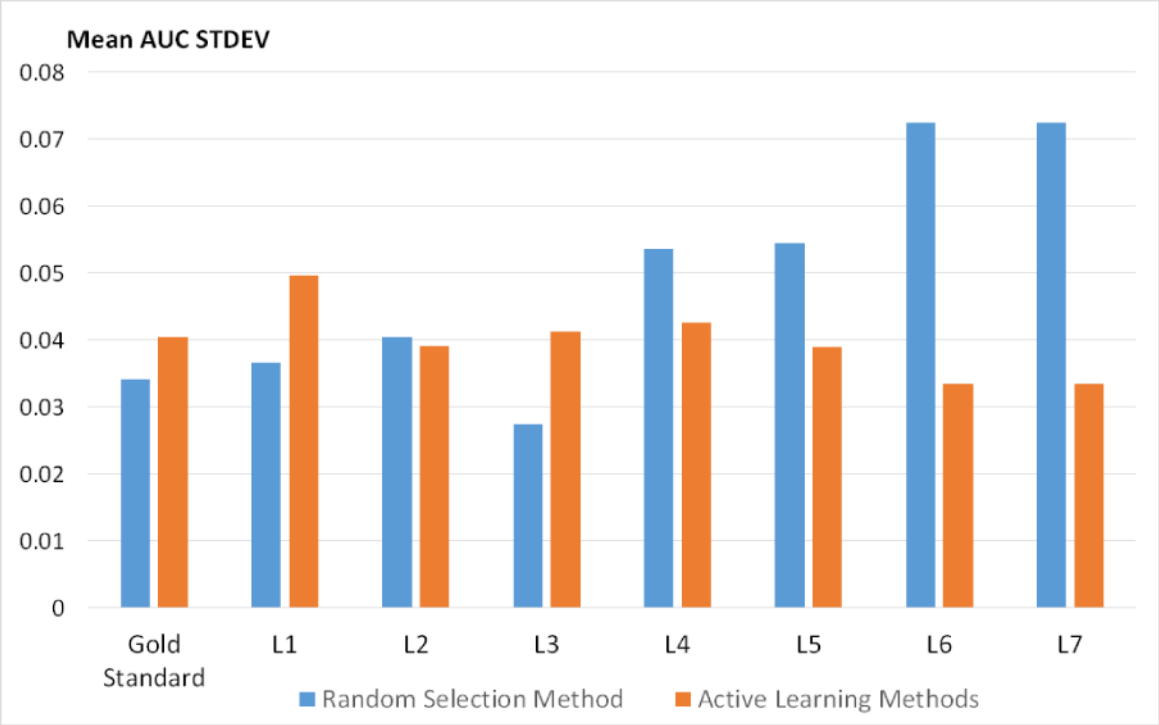

Figure 12-B.

The mean *intra-labeler* variance, for the 20 acquisition trials, in the performance of the models induced from the labels provided by each labeler, for the seven labelers and the gold standard label. For each labeler (and for the gold standard label), the mean variance over time of the models induced using the passive learning method is compared to the mean variance of all of the models induced over time using the three active learning methods.

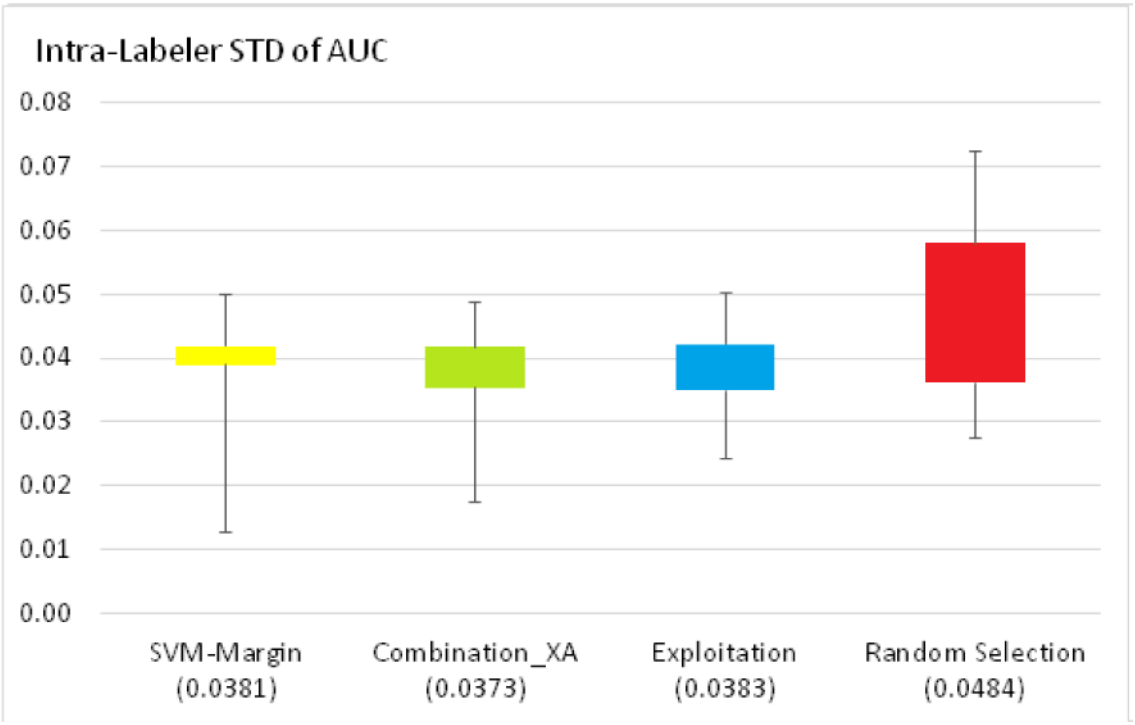

Figure 12-C.

The *intra-labeler* variance of the models induced using each of the three active learning methods and the passive (Random selection) method, across the models induced from the labels provided by the seven labelers and from the gold standard label. The mean values of the standard deviation, across all labelers, appear in parentheses under each method.

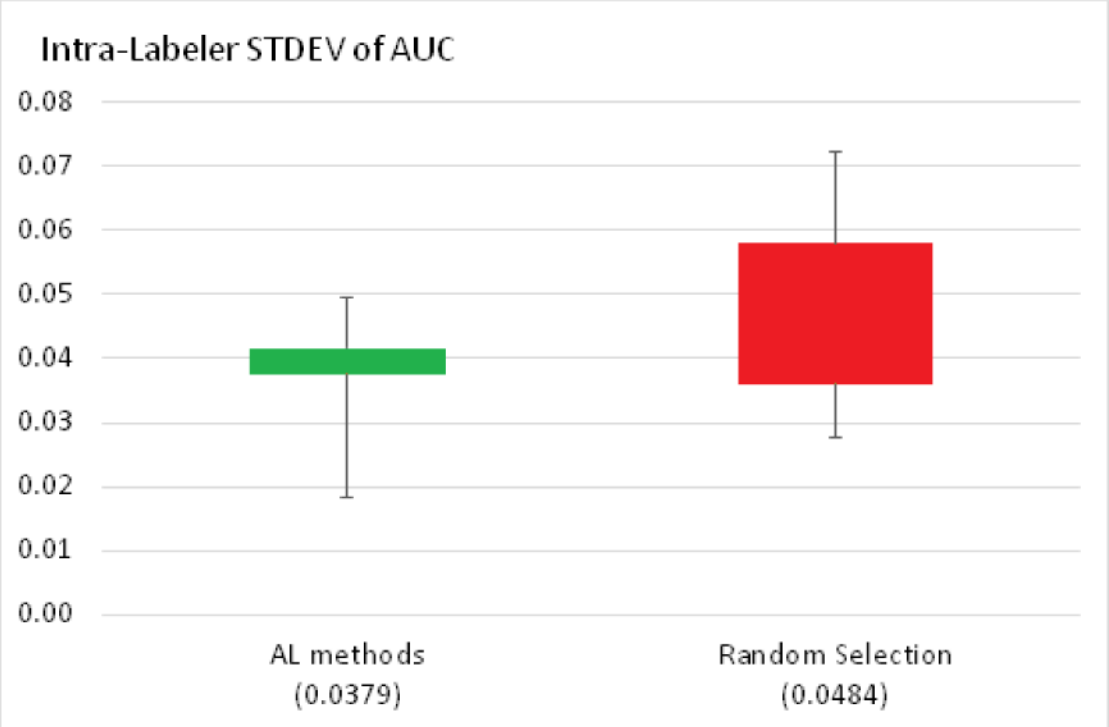

Figure 12-D.

The distribution of the *Intra-labeler* variance of the models induced using all of the three active learning methods, compared to the *Intra*-labeler variance of the models induced using the passive (Random selection) method, across the models induced from the labels provided by the seven labelers and from the gold standard label. The mean values of the standard deviation, across all labelers and learning methods, appear in parentheses under each method type.

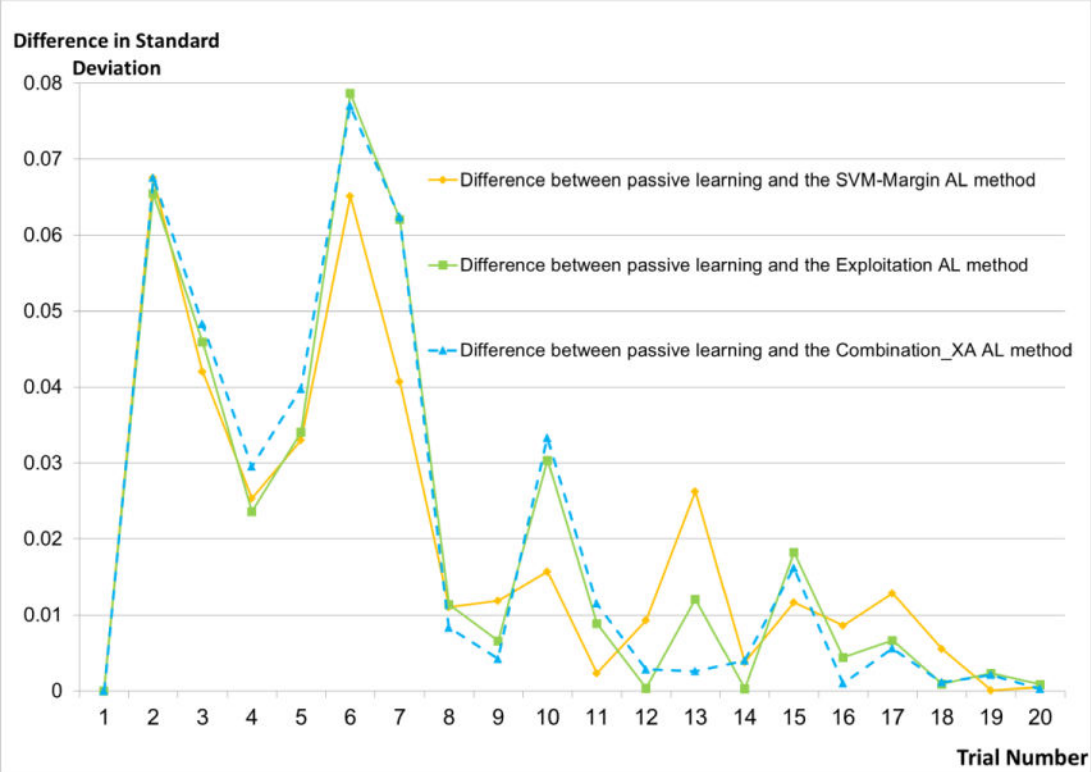

Figure 13.

The difference in standard deviation (absolute value) of the AUC among the classifiers induced by the AL methods and the passive (Random selection) learning method.

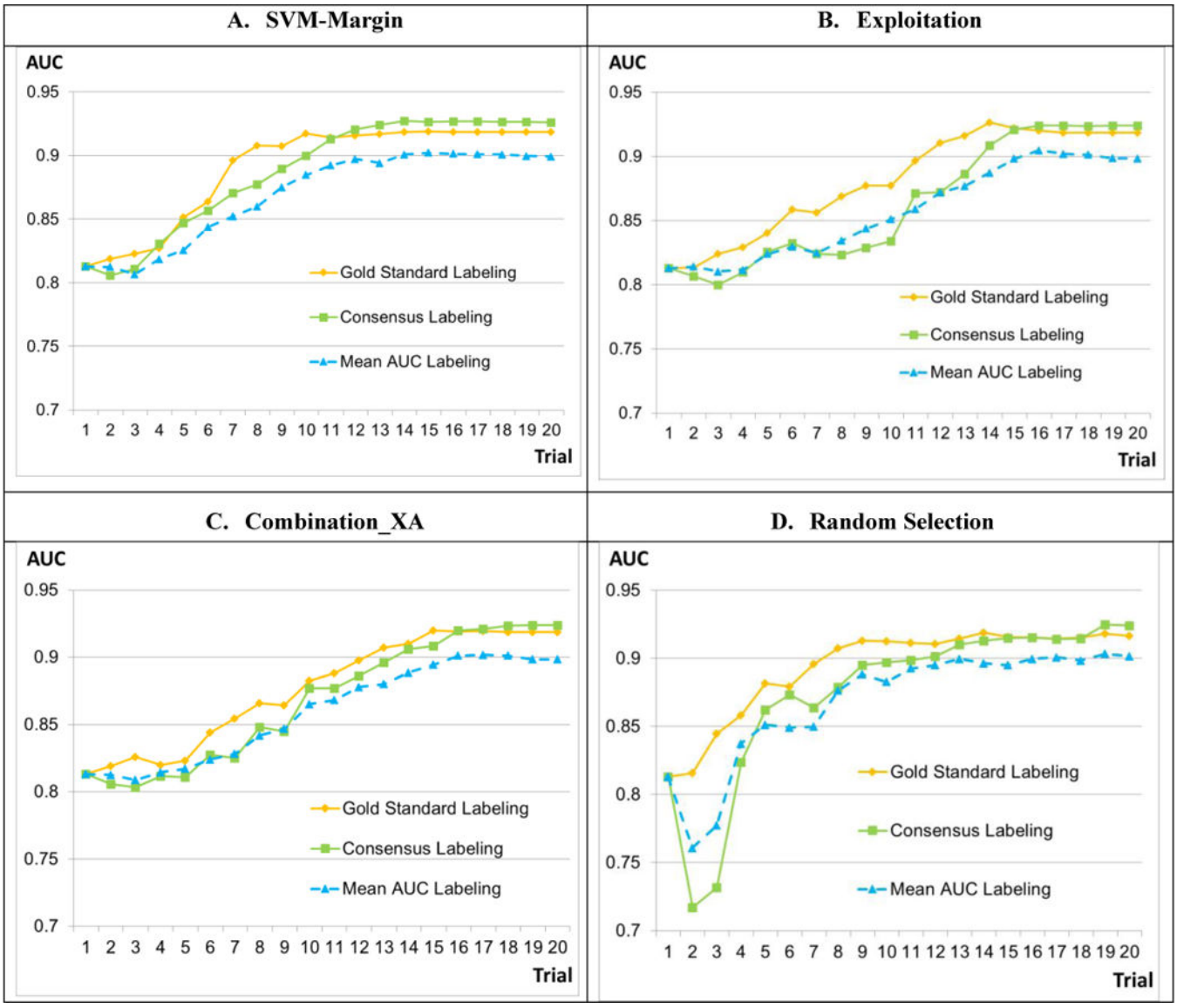

Figure 14.

The learning curves of the models induced by using the three AL methods and the passive (Random selection) learning method, for the three different labeling setups: gold standard labeler, consensus (majority) labeler, and mean AUC of the seven labelers, representing a randomly selected labeler.

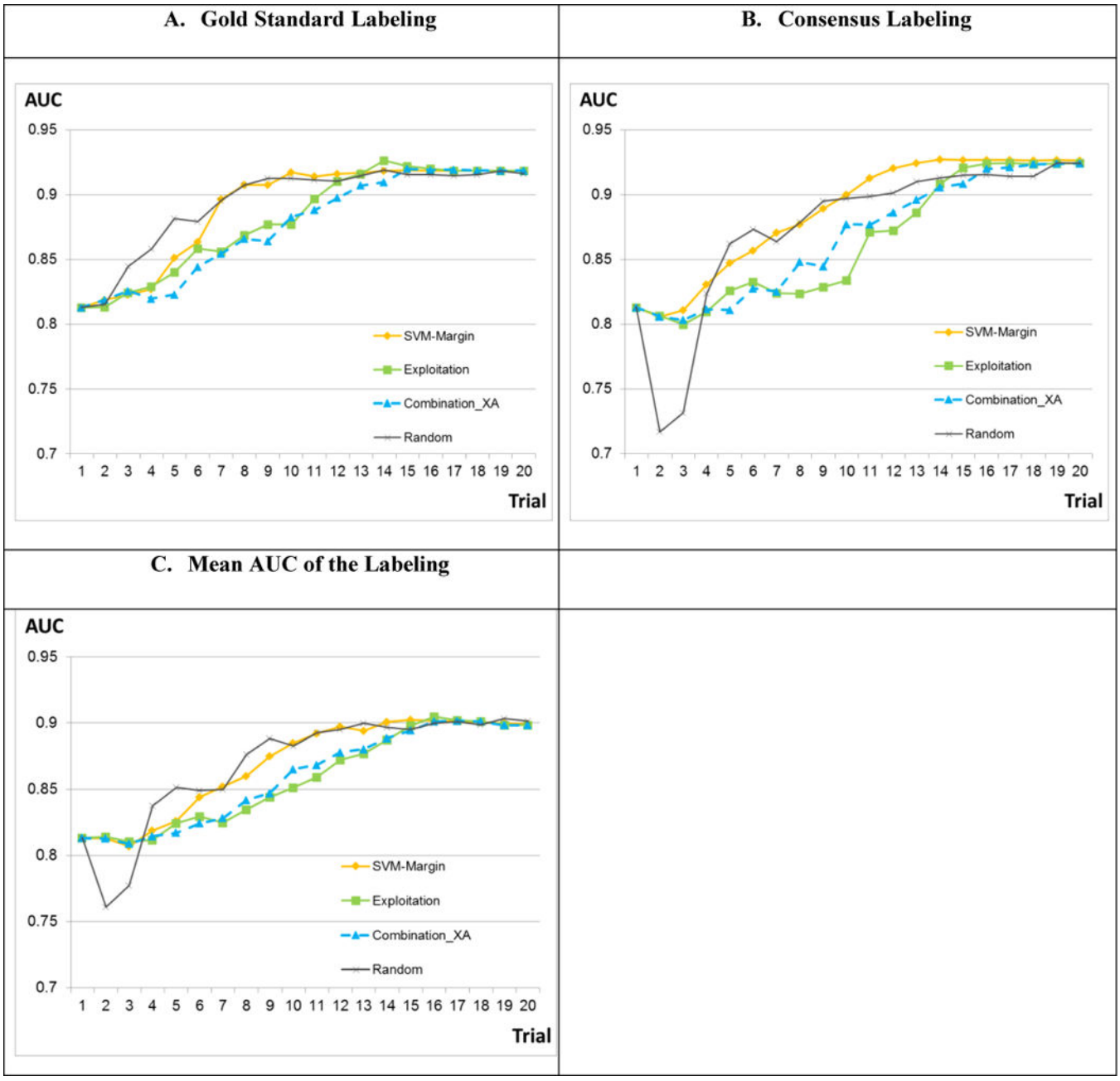

Figure 15.

The learning curves of the models induced from the labels provided by the three different labeling setups: gold standard labeler, consensus (majority) labeler, and mean AUC of the seven labelers, for each of the selection methods: the passive (Random selection) learning method and the three AL methods.

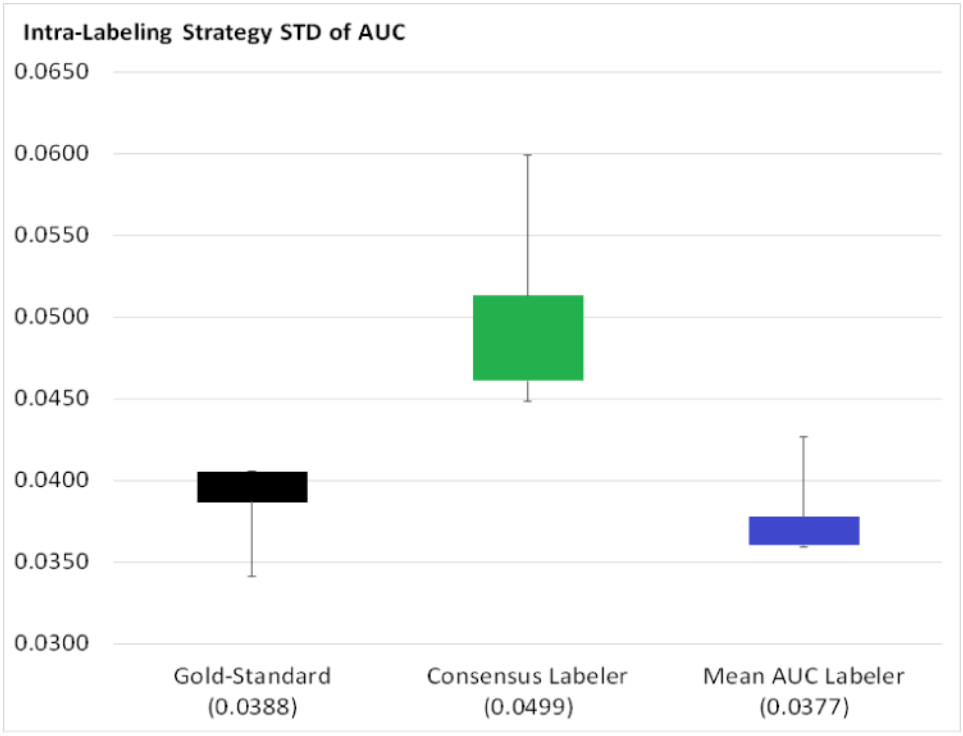

Figure 16-A.

A [25th, 75th percentile] box-plot of the mean *Intra*-labeler standard deviation of the AUC, and its minimal and maximal ranges, for the four learning methods, during the training phase, for the three labeling strategies. The mean values of the standard deviation, appear in parentheses under each method type.

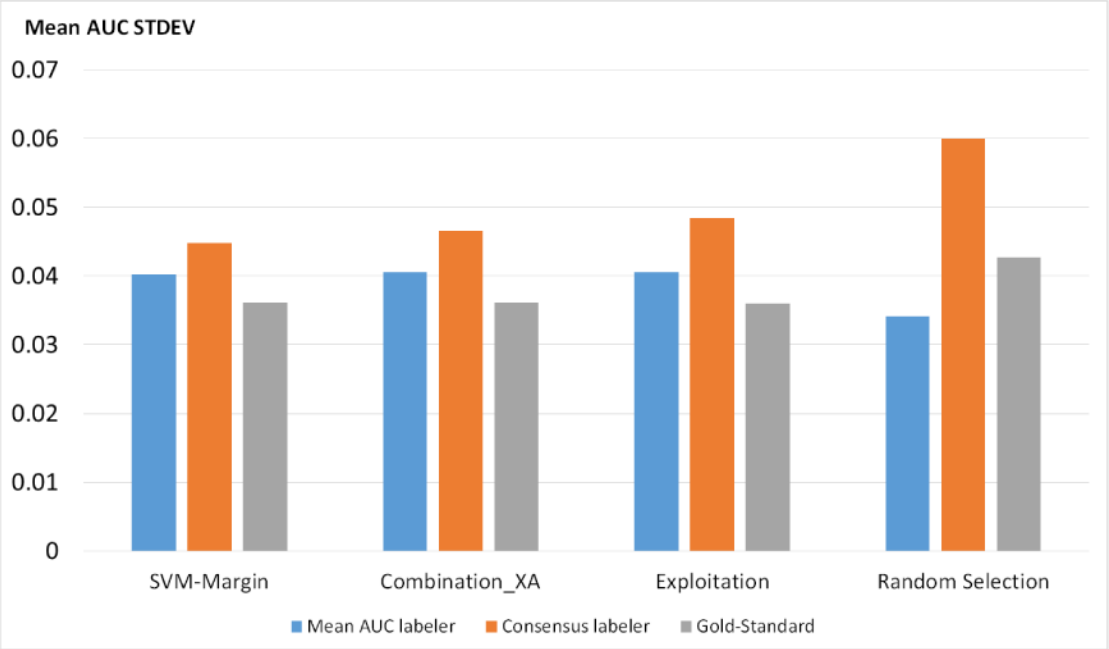

Figure 16-B.

The mean *Intra*-labeler standard deviation of the AUC, comparing, for each of the four selection methods, the three labeling strategies

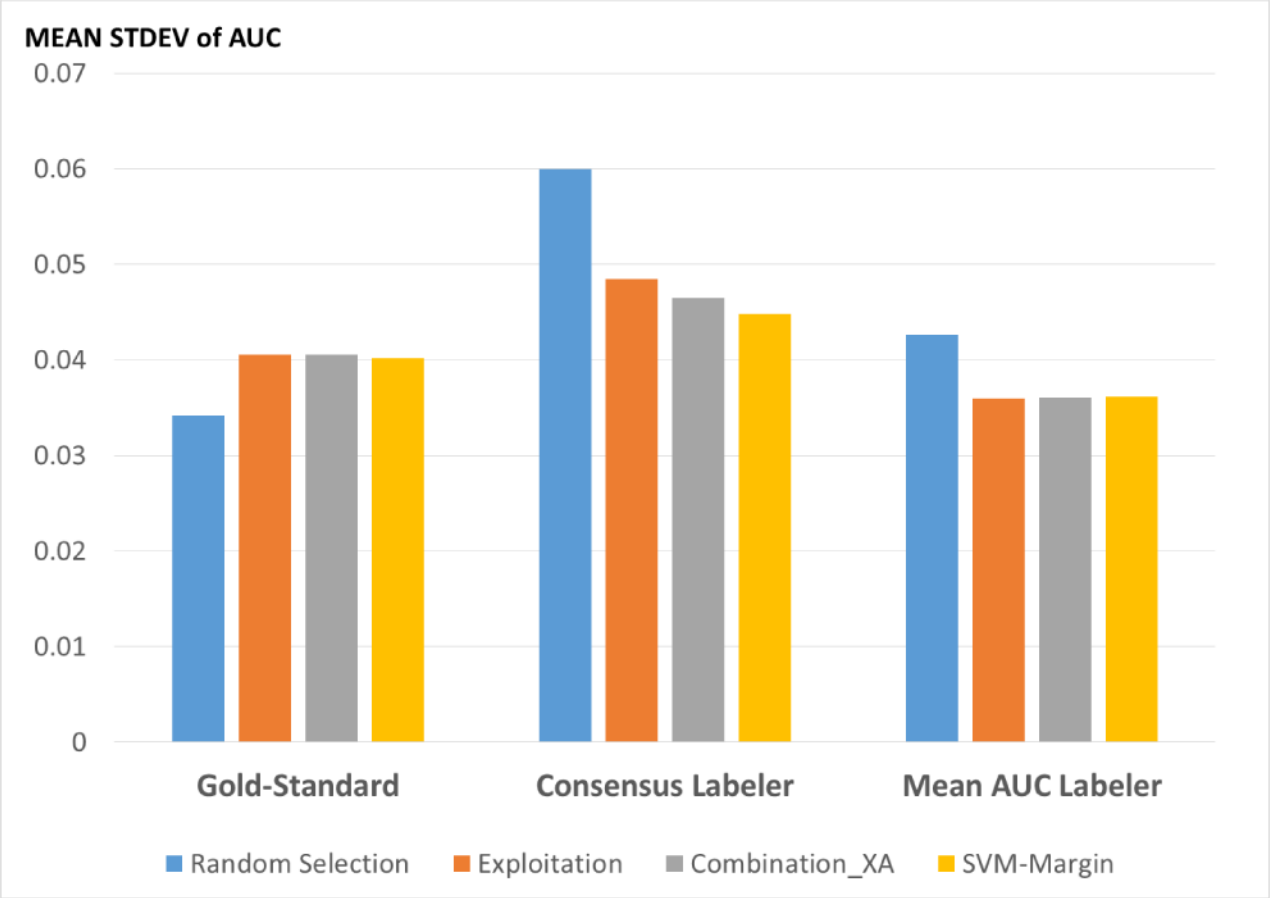

Figure 16-C.
A comparison of the mean standard deviation of the AUC among the four leaning methods, for each of the three labeling strategies.