Exploring Active Learning based on Representativeness and Uncertainty for Biomedical Data Classification

Rafael S. Bressan, Guilherme Camargo, Pedro H. Bugatti, Priscila T. M. Saito

***Abstract*—Nowadays, there is an abundance of biomedical data, such as images, genetic sequences, among others. However, there is a lack of annotation to such volume of data, due to the high costs involved to perform this task. Thus, it is mandatory to develop techniques to ease the burden of human annotation. To reach such goal active learning strategies can be applied. However, the state-of-the-art active learning methods, generally, are not feasible to lead with real-world datasets. Another important issue, that is generally neglected by these methods, is related to the conception that the classifier tends to learn more and more at each iteration. Their adopted selection criteria do not properly exploit the knowledge of the classifier. Therefore, in the present paper, we propose the use of an active learning approach, in order to leverage the learning process, including the proposal of a novel active learning strategy. The main difference of our proposed strategy is related to the participation of the classifier in an extremely active way in its learning process. So, we can better maximize and prioritize the knowledge that is obtained by the classifier at each iteration, making use of this knowledge in a more appropriate and useful way when selecting more informative samples. To do so, in our selection criteria, we give significant importance to the classifications suggested by the classifier. In addition, jointly with the participation and the knowledge of the classifier, we consider both uncertainty and representativeness criteria through a fine-grained analysis of the samples. Experimental results show that our novel active learning approach outperforms state-of-the-art active learning methods, considering several supervised classifiers. Hence, dealing with real dataset problems in a better way, equalizing the trade-off between annotation task and higher accuracy rates.**

***Index Terms*—biomedical datasets, healthcare, medical diagnosis, data analysis, knowledge discovery, machine learning.**

I. Introduction

A large amount of biomedical data (e.g. images, videos, among others) are generated and it has been increasing continuously and rapidly in different contexts, such as applications in medicine and biology [1]. Thus, automating the analysis and classification of these datasets becomes an extremely important task. In general, the classification and diagnosis based on the analysis of biomedical data requires an accurate annotation, in order to represent the content of interest from the data and to train a classification model. However, in real problems, it is common to have to deal with databases that present just a few labeled samples. This occurs due to the high effort and costs involved in the manual annotation process, since experts are required to perform such task. Another drawback is that inconsistencies may occur in the annotations given by the experts, because it is a tiresome process that leads to the fatigue generated by the analysis of the large amount of samples [2].

In order to tackle these problems, several supervised machine learning techniques [3], [4], [5], [6], [7] have been developed. Although, well-know and widely used in different contexts and applications [8], [9], [10], many of them are not suited for these datasets, because there is a lack of annotated samples. Besides, they do not deal with some constraints regarding to efficacy and efficiency required for real-world applications. Despite some efforts [11], [12] to reach such constraints, in order to provide a faster and more effective classifier, they still require a large amount of annotated data to train a robust classifier. Then, there is a new trade-off between the need for a considerable amount of annotated data to obtain higher accuracies and the effort and the cost of annotation accomplished by an expert. In this case, active learning strategies can be explored in order to deal with this trade-off.

Active learning techniques allow a small set of unlabeled learning samples to be selected and displayed, iteratively, for expert annotations, and then the annotated set is used for training the classifier. During the process, the classifier actively participates in its learning, selecting the most informative samples [13], [14]. The purpose of active learning techniques is to minimize expert involvement without losing control over the classifier's learning process. The effort and cost of expert annotation is reduced, since: i-) a small subset of more informative samples requires annotation; ii-) the expert needs only to correct the misclassified labels because the classifier previously provides a label for the samples. In addition, iii-) as the classifier learns, fewer misclassifications occur, and so the expert performs fewer corrections.

In the present paper, we seek to take advantage to the maximum the synergy between both machine and human experts. The human knowledge is indispensable for the success of the machine learning process, as well as for a greater acceptability of the use of computational tools in the expert routine. These tools can assist experts, avoiding misannotations caused by the exhaustive work or by lack of knowledge from less experienced professionals. On the other hand, experts time and effort are precious resources. Therefore, we

This work has been supported by CNPq (grants #431668/2016-7, #422811/2016-5); CAPES; Fundação Araucária; SETI; and UTFPR.

Rafael S. Bressan, Guilherme Camargo, Pedro H. Bugatti and Priscila T. M. Saito are from the Department of Computing, Federal University of Technology - Paraná, PR, Brazil. Priscila T. M. Saito is also with the Institute of Computing, University of Campinas, SP, Brazil. E-mails: {rafaelbressan, gcamargo}@alunos.utfpr.edu.br), {pbugatti, psaito}@utfpr.edu.br

propose a new machine learning algorithm in order to select the most informative samples for training classifiers more effectively and with minimal human intervention, as well as apply it, specifically, in different biomedical applications. In this context, one of the main objectives is also be specially careful for considering the challenges and constraints in digital healthcare. In this case, it is fundamental the analysis of the trade-off between the complexity and the effectiveness of the considered algorithms in order to develop feasible solutions to the biomedical context. Thus, we can enable valuable contributions and greater acceptability of such advances in biomedical and health informatics.

Contributions. This paper extends the conference paper [12]. Our contributions here are twofold. First, we bind active learning methods to the learning process, where, while learning, the classifier selects and suggests the annotation of samples that are more difficult to classify, and ease the burden of human annotation of all learning dataset required in the traditional supervised learning approach. This proposal enables not only a better gain in accuracies, but also a better efficiency. Secondly, we introduce a novel active learning method capable to better explore the most informative samples to the classifier learning through the proposed selection criteria based on representativeness and uncertainty of samples. The selection strategy is crucial to achieve higher accuracies as fast as possible (i.e. at the first learning iterations). We evaluated our method on several public biomedical datasets and the experiments showed that our approach, in many cases, outperforms state-of-the-art active learning methods, considering several supervised classifiers.

II. Background

Several strategies can be used to select a small set of samples, which must be annotated by an expert, constituting the labeled training set. We can simply randomly select a number of samples. However, it is well known in the literature that the active learning approach outperforms a naive baseline that simply selects random samples. In the literature, several studies [14], [15], [16], [17], [18], [19], [20], [21], [22] have already demonstrated the effectiveness of active learning approaches in relation to the passive learning ones (without the selection of the most informative samples) in different contexts and applications. The experimental results presented by these works have shown that the active learning approach can significantly reduce the amount of labeled training samples.

In this sense, the selection criteria based on representativeness, diversity and uncertainty are widely used, in different contexts, for selecting the most informative training samples in an active learning process. The representativeness and diversity in unlabeled samples have been explored through clustering methods [21], [22], [23]. Representativeness criterion can be obtained for choosing samples closest to the cluster centers. Diversity criterion aims to select samples from all classes faster. For example, [21] proposed the *Heuristic Clu* (HClu) method, in which the learning samples are grouped into distinct clusters. At each learning iteration, samples are randomly selected from each cluster, trying to get (diverse) samples from different classes. However, their performance may directly depends on the clustering algorithm and the data distribution.

Selection of samples based on uncertainty is regarded to the most likely samples to be misclassified by the classifier. Typical approaches include query-by-committee and based on boundary. In query-by-committee approaches several classifiers are considered for selecting the samples with the largest disagreements in the labels predicted by these classifiers [24], [25], [26]. Approaches based on boundary consists of selecting samples according to the maximum uncertainty through the distances to the classification boundaries [27], [28], [16]. For instance, [16] propose an active learning method with support vector machines (AL_SVM) for text classification problems. The method selects the samples which are closest to the classification boundary of the SVM classifier. In [17], the authors introduced the active learning method with SVM and applied it to gene expression profiles of colon, lung, and prostate cancer samples. In addition, based on SVM active learning, some efforts have been proposed. Although they are more robust, outperforming SVM active learning [16], they need to optimize the objective function, resulting in high computation complexity of $\mathcal{O}(n^3)$.

There are also some efforts that try to reduce the dataset and work with a small subset in the learning process, in order to avoid the re-training of all samples from the dataset. However, in real problems and considering biomedical contexts, this results in a major challenge. This type of strategies, that consider a reduction process, is usually not so well accepted by this specific area community, since relevant biomedical data can be discarded from the learning process. In this case, it is important to be careful because some crucial samples can be discarded and, it is interesting to search for other robust alternatives. Then, [29] and [30] have explored the combination of some selection criteria, as well as with different classifiers. In [22], the learning samples are also grouped into distinct clusters, and criteria based on diversity and uncertainty are considered to select the most informative samples. The works of [31] and [30] also proposed an active learning framework combining uncertainty and representativeness. However, these works consider a semi-supervised learning approach, which requires that the input data should satisfy the semi-supervised assumption, in order to guarantee the performance [29].

The major problem of these combined approaches is related to the complexity. Besides that, considering the uncertainty criterion, a large number of samples need to be obtained before the optimal decision boundary is found. In this context, despite efforts and somehow successful results, there are still several aspects to be improved. To the best of our knowledge, the active learning works found in the literature are not suitable for the biomedical applications, since effectiveness and efficiency constraints should be considered, due to this type of application often requires dealing with large datasets, interactive response times, and minimal expert intervention in the learning process.

The literature works basically perform an initial preprocessing, for example, by grouping the samples with a specific clustering technique. Then, they select the most informative samples based on the initial organization performed by the

clustering technique. One problem is that according to the technique used, it can compromise the learning of the classifier. The technique does not necessarily separate different classes into different clusters. Although other works propose some refinements, performing the clustering of samples at each iteration, it is usually time costly for large learning sets. Another problem is that the adopted selection criteria is poor, without properly exploiting the classifier, and without taking into account that the classifier improves at each iteration.

III. Proposed Method

We aim to propose solutions that speed up the learning process and that are feasible to be applied in health care. Unlike previous works, our main concern is related to the trade-off between complexity of the algorithms and satisfactory results. So, initially, we consider the clustering only once to obtain samples from each cluster and compose the first instance of the classifier. Clustering enables the identification of samples from all (or most) classes which improves the initial classifier. Then, throughout the learning iterations, we explore to the maximum the knowledge of the classifier, which is continuously acquired and improved at each iteration.

The idea is to consider a straightforward and meaningful manner of selecting the most informative samples, achieving high accuracies more quickly and minimizing the expert annotation effort. Therefore, in our selection criteria, we give significant importance to the classifications suggested by the classifier. Besides that, jointly with the participation of the classifier, we explore both representativeness and uncertainty criteria through a fine-grained analysis of the samples. We take into account not only samples closer to the centers of mass from each *class*, but also those at the boundaries between possible different *classes*. It is worth mentioning that we observe the *classes* provided by the current instance of the classifier, unlike the literature approaches [21], [22] that observe the *clusters* obtained from the initial clustering.

Figure 1 and Algorithm 1 present each step of the proposed active learning approach. In step 1, the method is fed with a given dataset $\mathcal{Z}$, which is then normalized by a given technique (e.g. min-max, z-score, etc). Afterwards, the dataset $\mathcal{Z}$ is divided into two subsets, a learning set $\mathcal{Z}_2$, and a test set $\mathcal{Z}_3$. The test set $\mathcal{Z}_3$ is used only to evaluate the classifier performance in an unknown dataset during its learning process. Besides, initially, two empty sets are also created $\mathcal{Z}_1$ and $\mathcal{Z}'_1$. The set $\mathcal{Z}_1$ will be used to train the classifier incrementally (through iterations) and the set $\mathcal{Z}'_1$ will be used as a temporary training set.

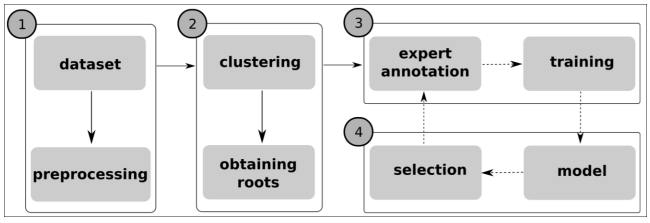


Fig. 1. Pipeline of the proposed active learning approach.

Algorithm 1: Proposed Approach

input : dataset $\mathcal{Z}$, number $\mathcal{N}$ of samples selected at each iteration for user validation, number k of clusters desired for the clustering.
output : final learning model $\mathcal{M}^{\Omega}$.
auxiliaries: learning set $\mathcal{Z}_2$, root set R, temporary training set $\mathcal{Z}'_1$, training set $\mathcal{Z}_1$, and a learning model instance $\mathcal{M}$.

1 $\mathcal{Z}_2 \leftarrow$ preprocessing($\mathcal{Z}$);
2 clustering($\mathcal{Z}_2$, k);
3 $\mathcal{R} \leftarrow$ getRoots($\mathcal{Z}_2$, k);
4 $\mathcal{Z}'_1 \leftarrow \mathcal{R}$;
5 $\mathcal{Z}_1 \leftarrow$ oracleAnnotateSamples($\mathcal{Z}'_1$);
6 $\mathcal{M} \leftarrow$ trainSamples($\mathcal{Z}_1$);
7 **repeat**
8 $\quad \mathcal{Z}'_1 \leftarrow$ selectSamples($\mathcal{Z}_2 \backslash \mathcal{Z}_1$, $\mathcal{N}$, $\mathcal{M}$);
9 $\quad \mathcal{Z}_1 \leftarrow \mathcal{Z}_1 \cup$ oracleAnnotateSamples($\mathcal{Z}'_1$);
10 $\quad \mathcal{M} \leftarrow$ reTrainSamples($\mathcal{Z}_1$, $\mathcal{M}$);
11 **until** *satisfied*;
12 $\mathcal{M}_{\Omega} \leftarrow \mathcal{M}$;

In step 2, after the preprocessing (i.e. normalization), the samples of $\mathcal{Z}_2$ are clustered with a given clustering approach. Then, a list $\mathcal{R}$ is created and it receives the k closest samples to each centroid (root samples) from $\mathcal{Z}_2$. As an auxiliary training set, $\mathcal{Z}'_1$ receives the samples contained in $\mathcal{R}$.

In step 3, each sample of $\mathcal{Z}'_1$ is presented to the oracle to be annotated. The annotated samples constitute the training set $\mathcal{Z}_1$. After the training process, in Step 4, one instance of the learning model $\mathcal{M}$ is created, and then this model actively participates in the process of selection of the most informative samples. The selection criteria based on representativeness and uncertainty of samples is implemented, and the set $\mathcal{Z}'_1$ receives such $\mathcal{N}$ samples, according to the current learning model $\mathcal{M}$, and the remaining samples in $\mathcal{Z}_2$ ($\mathcal{Z}_2 \cap \mathcal{Z}_1 = \emptyset$ or in a simplified notation $\mathcal{Z}_2 \backslash \mathcal{Z}_1$).

The selected sample set $\mathcal{Z}'_1$ is displayed to the oracle. From the first learning iteration, this set of selected samples are already previously labeled by the current instance of the classifier. So, the oracle needs only to correct the labels of misclassified samples. The samples confirmed and corrected properly by the oracle are added to the previous training set $\mathcal{Z}_1$. The training is performed again and a new instance of the classifier $\mathcal{M}$ is generated. Steps 3 and 4 are repeated until it is satisfied with the learning process.

The proposed selection criteria is deeply described by Algorithm 2. Initially, a candidate sample set $\mathcal{SC}_i$ ($0 < i \leq nc$) is created to receive the candidate samples to be displayed for the oracle, where i denotes the i-th class and nc the number of classes. As aforementioned, our idea is to make the best use of the classifier knowledge. Then, the learning set $\mathcal{Z}_2 \backslash \mathcal{Z}_1$ is classified by the current learning model ($\mathcal{M}$), in order to evaluate which ones are more informative for its own learning.

Samples are separated according to the class labels provided by the model, so each sample $s \in \mathcal{Z}_2 \backslash \mathcal{Z}_1$ is stored in a list

Algorithm 2: Selection Strategy

input : learning set $\mathcal{Z}_2 \backslash \mathcal{Z}_1$, number $\mathcal{N}$ of samples selected at each iteration for user annotation, current learning model instance $\mathcal{M}$.
output : selected sample set $\mathcal{SS}$.
auxiliaries: candidate sample set $\mathcal{SC}_i$, learning lists organized based on labels $\mathcal{L}_i$, nearest neighbor sample adjs, and a sample list corresponding to the center of mass of each label $\mathcal{L}com_i$.

$\mathcal{SC}_i \leftarrow \emptyset$, i = 1, 2, ..., nc;
classifySamples($\mathcal{Z}_2 \backslash \mathcal{Z}_1$, $\mathcal{M}$);
for *each* $s \in \mathcal{Z}_2 \backslash \mathcal{Z}_1$ **do**
  $\mathcal{L}_i \leftarrow \mathcal{L}_i \cup \{s\}$, i = s.labelid;
  adjs $\leftarrow$ get1NN(s);
  if *s.labelid* $\neq$ *adjs.labelid* **then**
    $\mathcal{SC}_i \leftarrow \mathcal{SC}_i \cup \{s\}$, i = s.labelid;
    $\mathcal{SC}_i \leftarrow \mathcal{SC}_i \cup \{adjs\}$, i = adjs.labelid;
  end
end
if $\mathcal{SC}_i \neq \emptyset$ **then**
  $\mathcal{L}com_i \leftarrow$ computeCenterOfMass($\mathcal{SC}_i$), i=1, 2, ..., nc;
  $\mathcal{SC}_i \leftarrow$ organizeSamples($\mathcal{SC}_i, \mathcal{L}com_i$), i=1, 2, ..., nc;
end
else
  $\mathcal{L}com_i \leftarrow$ computeCenterOfMass($\mathcal{L}_i$), i=1, 2, ..., nc;
  $\mathcal{SC}_i \leftarrow$ organizeSamples($\mathcal{L}_i, \mathcal{L}com_i$), i=1, 2, ..., nc;
end
while $i \leq \mathcal{N}$, *i=1, 2, ... nc* **do**
  $\mathcal{SS} \leftarrow \mathcal{SS} \cup \{s\}$, $s \in \mathcal{SC}_i$, following the order given by $\mathcal{SC}_i$;
end

of labels (L_i) corresponding to their respective class label i ($i = s.labelid$). Then, the closest sample (1-NN) to the sample s, named here as adjacent sample ($adjs$), is evaluated. If the labels from the samples s and $adjs$ are different (i.e $s.labelid \neq adjs.labelid$), these samples constitute informative samples, since they can really be off distinct classes. So, these samples are stored in a list of candidates $\mathcal{SC}_i$, corresponding to their respective class label i.

Since there are samples stored in the candidate lists $\mathcal{SC}_i$, the centers of mass com_i from the feature space of the samples of each class i are located and stored in their corresponding list of centers of mass $\mathcal{L}com_i$. To pinpoint each com_i, a simple computation is performed considering the average of the samples from the i-th class. Afterwards, the candidate samples are organized in their respective list $\mathcal{SC}_i$ in a descending order according to the distances between them and the com_i from $\mathcal{L}com_i$. However, due to the feature space distribution, the aforecited policy may not be satisfied, i.e. $\mathcal{SC}_i = \emptyset$, given there is no $s.labelid \neq adjs.labelid$. To solve this scenario, it is calculated the com_i from each learning sample list L_i, and then, they are stored in their respective lists of centers of mass $\mathcal{L}com_i$. With this information, the candidate samples $\mathcal{SC}_i$ are organized in a descending order in the same way as aforementioned, but now focusing on samples from L_i.

Finally, our selection strategy returns the selected sample set $\mathcal{SS}$, composed of $\mathcal{N}$ most informative (representative and uncertainty) samples, corresponding to the first samples obtained from each candidate sample set $\mathcal{SC}_i$. The selection criteria are related to the centers of mass from different class labels (representativeness), as well as they take into account the nearest samples that may be of different classes (uncertainty). Through this selection strategy, we were able to extract meaningful information, since we select them based on the classifications (i.e. continuously accumulated and enhanced knowledges) given by the model.

IV. Experiments

A. Datasets

In order to demonstrate the effectiveness of the proposed method, we used 6 public biomedical datasets. Each dataset presents different difficulty levels with varied sizes and dimensions. Information on these datasets can be analyzed below.

- **Breast Cancer Dataset** [32]: consisting of 97 samples, 2 classes and $24,481$ features. Samples represent patients who had developed distance metastases within 5 years (labeled as "relapse") and patients who remained healthy from the disease after their initial diagnosis for interval of at least 5 years (labeled as "non-relapse").
- **Central Nervous System Dataset** [33]: composed of 60 samples, 2 classes and $7,129$ features. It is used to analyse the outcome of the treatment for central nervous system embryonal tumor. Samples are distributed in 21 survivors (patients who are alive after treatment) and 39 failures (patients who succumbed to their disease).
- **Colon Tumor Dataset** [34]: comprised of 62 samples, 2 classes and $2,000$ features. Samples were collected from colon-cancer patients, wherein 40 tumor biopsies are from tumors (labeled as "negative") and 22 normal (labeled as "positive") biopsies are from healthy parts of the colons of the same patients.
- **Lung Cancer Dataset** [35]: consisting of 181 samples, 2 classes and $12,533$ features. It is divided into 31 samples of malignant pleural mesothelioma (MPM) of the lung and 150 samples of adenocarcinoma (ADCA) of the lung.
- **MLL Leukemia Dataset** [36]: composed of 72 samples, 3 classes and $12,582$ features. Among them, 24 samples are acute lymphoblastic (ALL), 20 mixed-lineage leukemia (MLL) and 28 acute myelogenous leukemia (AML).
- **Subtypes of Acute Lymphoblastic Leukemia Dataset** [37]: composed of 327 samples, 7 classes and $12,558$ features. It contains all the known acute lymphoblastic leukemia subtypes, including T-cell (T-ALL), E2APBX1, TEL-AML1, MLL, BCR-ABL, and hyperdiploid (Hyperdip > 50), as well as one class that contains diagnostic samples that did not fit into any one of the mentioned classes (labeled as "Others").

TABLE I
MEAN ACCURACIES OBTAINED BY EACH CLASSIFIER IN A TRADITIONAL SUPERVISED LEARNING.

Dataset	k-NN	NB	OPF	RF	SVM
Breast Cancer	61.05 ± 10.87	52.63 ± 2.48	58.42 ± 8.75	**69.47±16.42**	67.89 ± 13.69
Central Nervous Syst.	57.50 ± 12.08	63.33 ± 12.27	53.33 ± 11.91	65.83 ± 6.14	**67.50±8.29**
Colon Tumor	76.67 ± 10.24	58.33 ± 11.11	74.17 ± 13.86	**86.67±8.05**	85.00 ± 7.66
Lung Cancer	94.86 ± 3.23	97.57 ± 3.47	94.86 ± 3.70	97.84 ± 2.48	***99.19±1.31**
MLL Leukemia	80.00 ± 14.60	95.71 ± 4.99	80.00 ± 14.21	94.29 ± 4.52	**97.14±3.69**
Subtypes of Acute Lymphoblastic Leukemia	74.09 ± 7.01	51.67 ± 3.94	74.39 ± 7.47	83.33 ± 3.19	**91.21±1.72**

B. Scenarios

In order to validate the proposed method, each dataset was divided into two subsets, a learning set containing 80% of the samples and a test set containing 20% of the samples. This division procedure (i.e. holdout) was performed 10 times in a stratified way. It was generated 10 different training and test sets with empty intersection between them, and divided proportionally according to the number of samples in each class. For all the experiments the same split of each dataset was maintained, allowing a fair comparison between the methods.

Inittialy, we perform experiments considering the traditional supervised learning, in order to previously analyze the most appropriate classifier for each dataset. We considered the classifiers: *k-Nearest Neighbor* - k-NN [3], *Naive Bayes* - NB [4], *Optimum-Path Forests* - OPF [5], *Random Forest* - RF [7] and *Support Vector Machine* - SVM [6]. Then, to evaluate our proposed method, named here as ourAL, we also perform experiments against two the state-of-the-art active learning methods HClu [21] and AL_SVM [16].

For all methods, we set the $\mathcal{N}$ equals to the number of classes (nc) in a given dataset. Moreover, different clustering and classification techniques can be applied. In our experiments, we considered the k-means clustering technique for all methods that require the clustering of the learning samples, where we defined $k = nc$. For classification, we consider the SVM classifier for all methods, given that it presented the best results in the traditional supervised learning.

We also analyzed different supervised classification strategies, in order to evaluate and compare the performance of our method using each of them. We considered the same (k-NN, NB, OPF, RF and SVM) classifiers.

In order to summarize the results it was used the mean accuracy obtained by each classifier per iteration of the AL methods. Moreover, we presented the total of annotated samples and the learning times obtained by each method.

C. Results

Considering the performance obtained by different classifiers in a traditional supervised learning, in general, for all datasets, the SVM classifier presented the best accuracies, as can be seen in Table I. So, the SVM classifier was used in the experiments (Figure 2) with the active learning approaches.

Analyzing the results obtained by our active learning approach (Figure 2) against the supervised paradigm (Table I), we can observe a considerable gain. Considering the MLL Leukemia dataset, using the SVM classifier, our AL approach reached 97% of accuracy with just 24 samples (8th iteration) from the dataset, while the traditional supervised learning requires the annotation of all samples (58) from the dataset. Thus, our approach were capable to reduce in up to 41.37% the number of annotated samples, while reaching a similar accuracy. The same behavior was observed considering the other datasets. For instance, on the Subtypes of Acute Lymphoblastic Leukemia dataset, with the SVM classifier, we obtained 91% of accuracy with only 140 samples (20th iteration). This means that our proposed approach not only reduced in up to 53.63% the annotation process when compared with the supervised paradigm, but also achieved, approximately, the same accuracy.

TABLE II
TOTAL OF ANNOTATED SAMPLES OBTAINED BY EACH APPROACH.

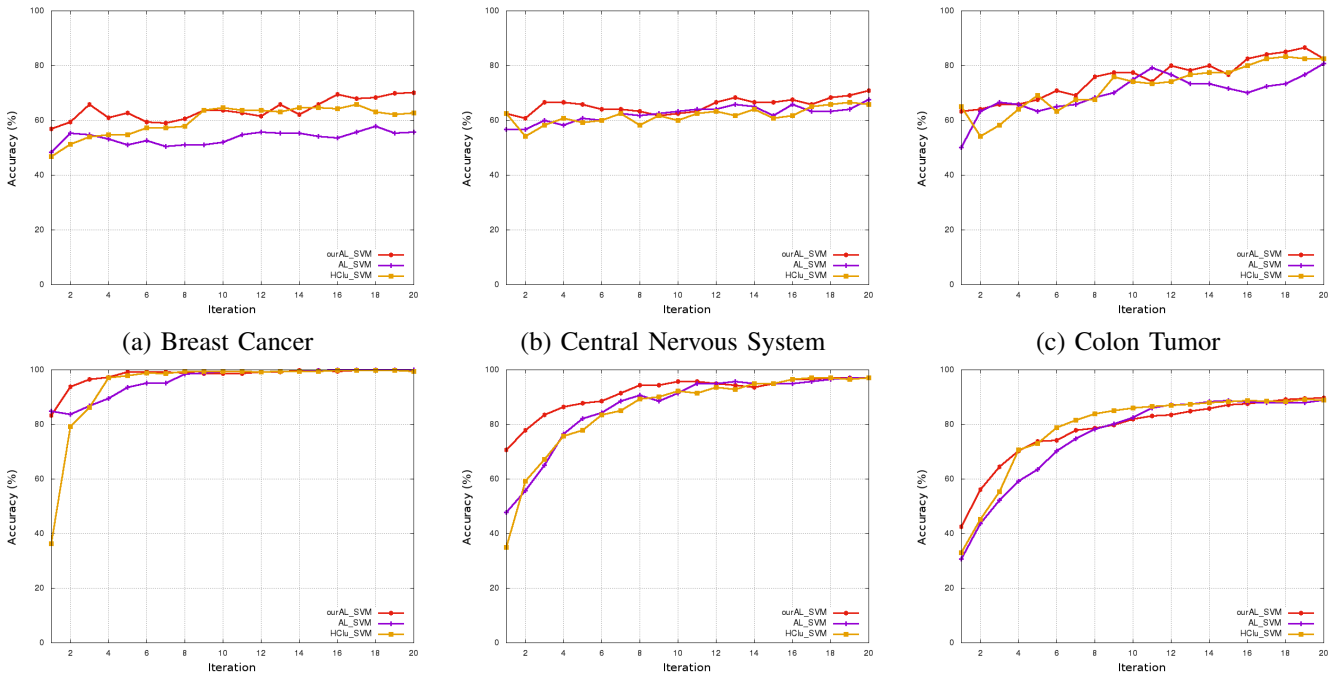
Dataset	ourAL_SVM	AL_SVM	HClu_SVM
Breast Cancer	18.70 ± 0.65	16.30 ± 0.63	40.00 ± 0.00
Central Nervous Syst.	16.20 ± 0.58	16.40 ± 0.63	40.00 ± 0.00
Colon Tumor	13.60 ± 0.54	11.70 ± 0.55	40.00 ± 0.00
Lung Cancer	6.00 ± 0.22	6.00 ± 0.19	40.00 ± 0.00
MLL Leukemia	$11,20 \pm 0.45$	10.90 ± 0.40	58.00 ± 0.00
Subtypes of Acute	38.20 ± 1.00	35.50 ± 0.96	134.22 ± 0.67

Furthermore, we can consider statistically significant gains presented by the proposed approach in comparison with AL_SVM and HClu_SVM (95% of confidence level using the student's t-distribution), especially in the first learning iterations (Figure 2). Throughout the iterations, active learning approaches are considered equivalent. Such a scenario is expected, since as known in the literature, during iterations all approaches tend to stabilize, and at the end of the learning with the entire dataset all the techniques will present the same performance.

We also presented the results obtained by our approach regarding different classifiers (k-NN, NB, OPF, RF and SVM). It is clear to note that, generally, the ourAL_SVM approach presented the best accuracies for almost all datasets (see Figure 3). Considering the Color Tumor dataset, the ourAL_k-NN and the ourAL_OPF (from 3rd to 10th iterations) also reached good results.

Table II shows the mean of samples annotated by each approach. It is clear to see that HClu requires the annotation of a much larger number of samples, when compared with ourAL_SVM and AL_SVM (HClu demands 6.6 times more samples to be annotated). In Table III it is illustrated the wall-clock time (seconds) required to execute each approach. We can note that ourAL_SVM and AL_SVM statistically ties. However, ourAL_SVM reaches better accuracies. For instance, considering the MLL Leukemia dataset, ourAL_SVM reached an accuracy of 70.71% with a time of 4.74s, while AL_SVM

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2018.2881155, IEEE Journal of Biomedical and Health Informatics



(a) Breast Cancer (b) Central Nervous System (c) Colon Tumor

(d) Lung Cancer (e) MLL Leukemia (f) Subtypes of Acute

Fig. 2. Mean accuracy per iteration obtained by ourAL_SVM, AL_SVM and HClu_SVM approaches.

TABLE III
LEARNING TIMES (SECONDS) OBTAINED BY EACH APPROACH.

Dataset	ourAL_SVM	AL_SVM	HClu_SVM
Breast Cancer	34.97 ± 9.61	32.48 ± 9.51	1.74 ± 0.37
Central Nervous Syst.	1.55 ± 0.59	1.33 ± 0.55	0.22 ± 0.08
Colon Tumor	0.17 ± 0.03	0.16 ± 0.03	0.07 ± 0.02
Lung Cancer	19.24 ± 4.43	18.33 ± 4.21	1.77 ± 0.19
MLL Leukemia	4.74 ± 2.62	3.91 ± 2.46	0.46 ± 0.27
Subtypes of Acute	31.12 ± 8.56	27.94 ± 7.66	3.58 ± 0.89

achieved an accuracy of 47.85% with a time of 3.91s.

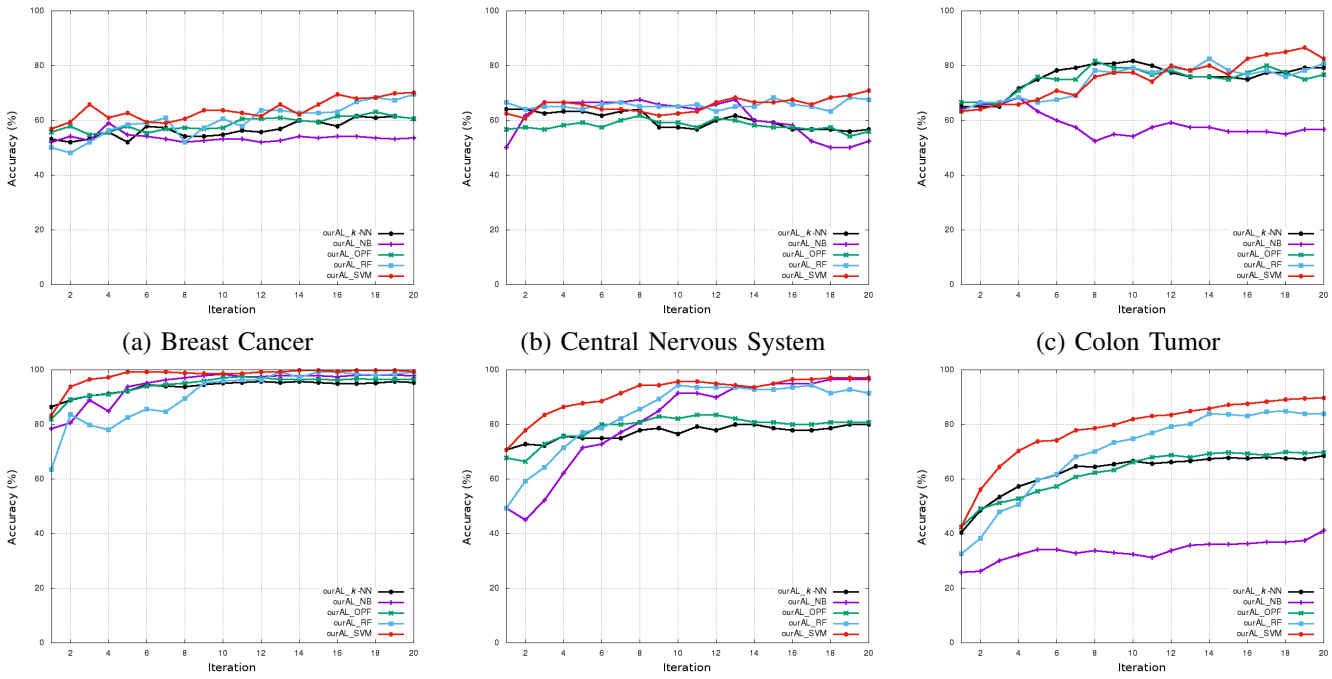
From the overall results, we can argue that applying a fine-grained analysis of uncertainty and representativeness of the samples in the feature space can considerably improve not only the learning process, but also enable to reach such accuracy in a faster way. Thus, leading to gains in efficacy and efficiency, and dealing with applicable and suitable strategies for classification of biomedical data.

V. CONCLUSION

We have introduced a novel active learning strategy to perform a fine-grained process, selecting the most informative samples, based on uncertainty and representativeness, and promoting higher accuracies to a given classifier, as fast as possible. This is a mandatory issue because supervision, as well known, demands a higher cost regarding the labeling process. Hence, we show that our method is able to reach higher accuracies already at the first iterations, in order to mitigate this main drawback. In addition, we demonstrate our method in 6 public biomedical datasets against state-of-the-art active learning approaches, and using different supervised classifiers. According to experiments, in the majority of cases, we reached notable accuracy improvements, and well-suited number of annotated samples and learning times. Thus, achieving a better trade-off between accuracy, annotated samples and learning times. Therefore, the proposed approach becomes a valuable contribution to extraction of meaningful information, representation of knowledge and personalisation of digital medical systems.

REFERENCES

[1] D. T. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*. Wiley-Interscience, 2004.
[2] E. Lughofer, R. Richter, U. Neissl, W. Heidl, C. Eitzinger, and T. Radauer, "Explaining classifier decisions linguistically for stimulating and improving operators labeling behavior," *Inf. Sci.*, vol. 420, pp. 16–36, 2017.
[3] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. on Information Theory*, vol. 13, no. 1, pp. 21–27, 2006.
[4] T. Bayes, "An essay towards solving a problem in the doctrine of chances," *Philos. Trans. R. Soc. Lond.*, vol. 53, pp. 370–418, 1763.
[5] J. P. Papa, A. X. Falcão, V. H. C. de Albuquerque, and J. a. M. R. S. Tavares, "Efficient supervised optimum-path forest classification for large datasets," *Pattern Recognition*, vol. 45, pp. 512–520, 2012.
[6] J. Joachims, "Making large-scale SVM learning practical., Advances in kernel methods-support vector learning, 169-184," 1999.
[7] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
[8] R. Prashanth, S. D. Roy, P. K. Mandal, and S. Ghosh, "High-accuracy classification of parkinson's disease through shape analysis and surface fitting in 123i-ioflupane spect imaging," *IEEE J Biomed and Health Inform*, vol. 21, no. 3, pp. 794–802, 2017.
[9] C. Ye, B. V. Kumar, and M. T. Coimbra, "An automatic subject-adaptable heartbeat classifier based on multiview learning," *IEEE J Biomed and Health Inform*, vol. 20, no. 6, pp. 1485–1492, 2016.
[10] D. R. Amancio, C. H. Comin, D. Casanova, G. Travieso, O. M. Bruno, F. A. Rodrigues, and L. da F. Costa, "A systematic comparison of supervised classifiers," *Plos One*, vol. 9, no. 4, pp. e94 137–1, 2014.
[11] P. T. M. Saito, R. Y. M. Nakamura, W. P. Amorim, J. P. Papa, P. J. de Rezende, and A. X. Falcão, "Choosing the most effective pattern classification model under learning-time constraint," in *Plos One*, vol. 10, 2015, p. e0129947.



(a) Breast Cancer (b) Central Nervous System (c) Colon Tumor

(d) Lung Cancer (e) MLL Leukemia (f) Subtypes of Acute

Fig. 3. Mean accuracy per iteration obtained by our approach using different (k-NN, NB, OPF, RF and SVM) classifiers.

[12] G. Camargo, R. S. Bressan, P. H. Bugatti, and P. T. M. Saito, "Towards an effective and efficient learning for biomedical data classification," in *IEEE Intl Symp on Computer-Based Medical Systems*, 2017, pp. 13–18.

[13] E. Lughofer, "On-line active learning: A new paradigm to improve practical useability of data stream modeling methods," *Inf. Sci.*, vol. 415, pp. 356–376, 2017.

[14] B. Settles, *Active Learning*, ser. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2012.

[15] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," *Machine Learning*, vol. 15, no. 2, pp. 201–221, May 1994.

[16] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *JMLR*, vol. 2, pp. 45–66, 2002.

[17] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1, pp. 389–422, Jan 2002.

[18] A. Bordes, S. Ertekin, J. Weston, and L. Bottou, "Fast kernel classifiers with online and active learning," *JMLR*, vol. 6, pp. 1579–1619, 2005.

[19] S. Dasgupta, A. T. Kalai, and C. Monteleoni, "Analysis of perceptron-based active learning," *JMLR*, vol. 10, pp. 281–299, 2009.

[20] P. Donmez and J. G. Carbonell, "From active to proactive learning methods," in *Advances in Machine Learning I*, 2010, pp. 97–120.

[21] N. Alajlan, E. Pasolli, F. Melgani, and A. Franzoso, "Large-scale image classification using active learning," *IEEE Geoscience Remote Sensing Letters*, vol. 11, no. 1, pp. 259–263, 2014.

[22] P. T. Saito, C. T. Suzuki, J. F. Gomes, P. J. de Rezende, and A. X. Falcão, "Robust active learning for the diagnosis of parasites," *Pattern Recognition*, vol. 48, no. 11, pp. 3572–3583, 2015.

[23] E. Lughofer, "Hybrid active learning for reducing the annotation effort of operators in classification systems," *Pattern Recognition*, vol. 45, no. 2, pp. 884–896, 2012.

[24] D. Vasisht, A. Damianou, M. Varma, and A. Kapoor, "Active learning for sparse bayesian multilabel classification," in *ACM Intl. Conf. on Knowledge Discovery and Data Mining*. ACM, 2014, pp. 472–481.

[25] P. Melville and R. J. Mooney, "Diverse ensembles for active learning," in *Intl. Conf. on Mach. Learn.* ACM, 2004, pp. 74–81.

[26] I. Guyon, G. C. Cawley, G. Dror, and V. Lemaire, "Results of the active learning challenge," in *Mach. Learn. Res.*, 2011, pp. 19–45.

[27] M. Wang and X.-S. Hua, "Active learning in multimedia annotation and retrieval: A survey," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 2, pp. 10:1–10:21, 2011.

[28] Y. Chen and A. Krause, "Near-optimal batch mode active learning and adaptive submodular optimization," in *Intl. Conf. Mach. Learn.*, vol. 28. JMLR.org, 2013, pp. I–160–I–168.

[29] B. Du, Z. Wang, L. Zhang, L. Zhang, W. Liu, J. Shen, and D. Tao, "Exploring representativeness and informativeness for active learning," *IEEE Transactions on Cybernetics*, vol. 47, no. 1, pp. 14–26, 2017.

[30] Y. Yang, Z. Ma, F. Nie, X. Chang, and A. G. Hauptmann, "Multi-class active learning by uncertainty sampling with diversity maximization," *Int. J. Comput. Vis.*, vol. 113, no. 2, pp. 113–127, 2015.

[31] S. Huang, R. Jin, and Z. Zhou, "Active learning by querying informative and representative examples," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 10, pp. 1936–1949, 2014.

[32] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, pp. 530–536, 2001.

[33] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. H. Kim, L. C. Goumnerova, P. M. Black, C. Lau, J. C. Allen, D. Zagzag, J. M. Olson, T. Curran, C. Wetmore, J. A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D. N. Louis, J. P. Mesirov, E. S. Lander, and T. R. Golub, "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature*, vol. 415, pp. 436–442, 2002.

[34] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *National Academy of Sciences USA*, vol. 96, no. 12, pp. 6745–6750, 1999.

[35] G. J. Gordon, R. V. Jensen, L. L. Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswamy, W. G. Richards, D. J. Sugarbaker, and R. Bueno, "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma," *Cancer Research*, vol. 62, no. 17, pp. 4963–4967, 2002.

[36] S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer, "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia," *Nature Genetics*, vol. 30, no. 1, pp. 41–47, 2002.

[37] E. J. Yeoh, M. E. Ross, S. A. Shurtleff, W. K. Williams, D. Patel, R. Mahfouz, F. G. Behm, S. C. Raimondi, M. V. Relling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X. Zhou, J. Li, H. Liu, C. P. H., W. E. Evans, C. Naeve, L. Wong, and J. R. Downing, "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling," *Cancer Cell*, vol. 1, no. 2, pp. 133–143, 2002.