Homework-1
Paper-1: Learning Active Learning from Data
https://infoscience.epfl.ch/server/api/core/bitstreams/02776ded-4ae4-4d2f-b708-be0b8fa9458c/content

1. What problem does this paper try to solve? i.e., its motivation

The Paper is trying to address the challenge of efficiently selecting which instances to label in the context of Active Learning. They are trying to do it because many modern machine learning models require large amount of labeled data to perform well and sometimes especially in specialized domains it can be expensive and time consuming, as it may need experts annotating the data.
The Paper proposes a data-driven approach called Learning Active Learning (LAL) to counteract the limitations of existing approaches, i.e. their success heavily depend on the ability to estimate classification performance from scarcely annotated data. LAL method formulates the problem as a regression task, predicting generalization error that can be expected by labeling specific samples.

2. How does it solve the problem?

First, LAL tries to generalize the choice problem in query selection using a Regression Problem. It trains a regressor that predicts the expected reduction in generalization error given the current learning state. Which, according to them, allows the method to choose samples that are expected to provide the most benefit when labelled. It learns from the history of active learning results instead of depending only on heuristics.

The regressor can be trained on synthetic datasets, which enables a wide variety of scenarios to learn from. It can also be trained on a small subset of domain-specific data. LAL can thus focus on samples that provide the most significant benefit by estimating how much labelling a certain sample will improve the model performance. This paper shows that LAL outperforms most of the baseline methods, which include traditional heuristics and other meta-learning methods in real-world applications of biomedical imaging, economics, and molecular biology.

3. A list of novelties/contributions

The paper makes several new contributions and improvements upon state-of-the-art active learning strategies, including

- LAL departs from this traditional approach of combining available active learning heuristics into a regression problem, such as uncertainty sampling and query-by-committee. The goal is to enable the learning from data itself without any pre-defined strategies in order to adapt to a wide range of datasets and tasks more compared to methods which just combine the heuristics in question.
- It does this by automatically adapting the sample selection to take into consideration the relative prevalence of classes in the dataset. This adaptability helps to handle one common limitation for many active learning strategies: they are not able to handle class imbalance, which could render LAL highly effective in performing with much variation in real-world scenarios.
- It is demonstrated that this methodology works both with Random Forest and Gaussian Process classifiers. A fair interpretation could be that it may have wider applicability compared to classifier-specific active learning methods.
- It turns out that the LAL-iterative-2D method reaches the highest ranking of uncertainty sampling 2.2-5 times faster on complex real-life tasks. For the Striatum dataset, LAL-iterative-

2D provides for an IOU of 0.6 with about 300 labeled points, while uncertainty sampling requires more than 500. For an MRI dataset, LAL-iterative-2D reaches a dice score of 0.7 with about 150 labeled points compared to 250 by uncertainty sampling.

4. Would you say there are downsides to this work?

One clear drawback of the LAL framework is that it is quite computationally expensive in its implementation. Run time data is given for a variety of strategies and tasks; while LAL does produce competitive results, doing so comes at considerable computational cost. For instance, for the Checkerboard dataset, LAL was recorded to run at 0.54 seconds per iteration against 13.12 seconds of the ALBE strategy, proving that indeed LAL is more efficient compared to some of its alternatives on some simpler datasets. In contrast, on the most complex datasets, such as the Striatum, LAL took 19.50 seconds, far over the 2.05 seconds for the US strategy. With this in mind, LAL could be effective but sometimes not perfectly so, especially in high-dimensional data settings where computational resources are at a premium. Performance intrinsically depends on the quality of the data from which the regression model is learned. If the original dataset to which LAL is applied is noisy, unrepresentative, or biased, then the predictions obtained with the LAL framework may turn out to be flawed. This alone would be a serious bottleneck in applications where obtaining clean and labeled data is difficult.

Paper-2: AI-Driven Tools for Coronavirus Outbreak: Need of Active Learning and Cross-Population Train/Test Models on Multitudinal/Multimodal Data
https://link.springer.com/article/10.1007/s10916-020-01562-1

1. What problem does this paper try to solve? i.e., its motivation

The paper tries to address the challenge of effectively analyzing and forecasting the detection of COVID-19, particularly at its early stages when the data availability is small. Specifically, the paper talks about traditional AI models, which can usually require large amounts of training data that usually would not be present during the early stages of any disease outbreak. It also addresses to the point that the world needs AI models that they can be trained on one population but then can be effectively applied to detect outbreaks in other geographical area. Lastly, it also says that using a single data type cannot cover the total information about the outbreak. The above problems take into consideration an approach that can be tentatively applied to the future pandemic situation, for the betterment of the overall preparedness with respect to global health issues.

2. How does it solve the problem?

The solution aims to use an active learning mechanism that allows the model to become dynamic and to improve continually with the availability of newer data, thus bypassing the limitation that is scarcity of datasets. This is further enhanced by the use of a cross-population train/test framework that allows learning the model on data from one region but testing of this model to be applied to others, which addresses the further issues of geographical disparity of data. It also makes use of incremental learning as a part of an active learning framework to maintain updated models towards new information without losing the history of previously acquired knowledge.

It suggests, for greater stability within analysis, fusion of multitudinal and multimodal data sources that combine information across the classes of medical records, such as EHRs, CRRS, CT scans, and evolved genetic sequencing. The trace mechanisms of anomaly detection can then indicate patterns or trends that have varied high variance, suggesting a change in the behavior and spread of the virus.

3. A list of novelties/contributions

New Contributions as a part of these paper is active learning of models such that learning is incremental without the need for a prior dataset. This is an extremely important capability when one considers the COVID-19 scenario, where time for data collection about the disease is not available. On the other hand, in most cases, traditional methods of machine learning must rely on big annotated datasets which, in such fast-evolving crises, may be too time-consuming or even unrealistic.

Additionally, the concept of cross-population train/test models is emphasized, enabling AI tools to adapt to different populations and contexts. This helps to increase the robustness and applicability to different scenarios—a pitfall of previous techniques that may not generalize so well across geographies or demographics. The combination with multimodal and multitudinal data provides an even finer step towards understanding the diffusion of viruses and allows making better-informed decision processes compared to methods using only one kind of data. The paper also emphasizes that such data should be analyzed and used in real time to make the corresponding decisions versus conventional approaches that relied on the collection of such data from historical times. Embedding methods of anomaly detection into an active learning framework finally empowers a model to highlight those events or observations that seldom appear and are far from typical patterns, thus being responsive to unexpected developments in data.

4. Would you say there are downsides to this work?

The main limitation is that this paper does not have concrete implementation detail nor results that would justify the propounded approaches. It's predominantly theoretical, thereby making it difficult to gauge the effectiveness of the paper to the problem. For one, the paper does not discuss possible challenges when applying models developed on data from one population to another with completely different demographics and a healthcare system.

Also, another limitation is inadequately discussing the possible biases associated with AI-based anomaly analysis is another con. This indicates bias in the output when the models are applied to underrepresented groups, since they are trained with data gathered mainly from specific population groups—an important point in healthcare this paper neglects to indicate. Last but not least, the problems of poor data quality and standardization are not fully discussed by the paper in cases with multiple sources of data, as these should be critical regarding the approach of multi-modal data integration proposed.

Paper-3: Batch Active Learning at Scale
https://arxiv.org/abs/2107.14263
1. What problem does this paper try to solve? i.e., its motivation

The objective of the paper is to tackle the problem of efficient data sampling in active learning, especially when dealing with very large batch sizes. It tries to solve several key issues, i.e. scalability of active learning methods, as they usually cannot handle large datasets and batch sizes much bigger than those considered in the literature. Most of the current algorithms have been designed, in most cases, for small batch sizes, and the vast majority of them do not scale well. Further, this sets up inefficiencies both at the data labeling and model training stage. Moreover, label efficiency is one of the critical challenges since it is generally expensive and time-consuming to get labeled data. The balance between informativeness and diversity in sample selection is another important aspect that this paper dealt with. That is, active learning methods have to make an effective trade-off between the selection of the samples which the model is uncertain about and representative selection about different aspects of the data distribution. Many of the previous approaches usually fail to handle this balance and hence yield a suboptimal sampling strategy.

2. How does it solve the problem?

This paper proposes the Cluster-Margin algorithm for the problem of batch active learning, which is well adapted to the large-scale dataset scenario where most traditional methods fail in efficiency and effectiveness. However, in these situations where the datasets are very large-in the millions-scalable and efficient active learning methods are truly important. In this respect, the Cluster-Margin algorithm merges hierarchical clustering with margin-based sampling, enabling it to handle batch sizes that are orders of magnitude larger than those usually considered in the literature.

Three steps form the base of the Cluster-Margin algorithm: initialization, clustering, and sampling. The initialization step selects a seed set P of examples uniformly at random from the unlabeled pool X. A common size of this is ~ 10,000 examples. Train a neural network on that seed set and use the penultimate layer to embed all the examples in X. The embedding step is a basic step in changing the raw data into their representations embedded with the hidden structure of the data, which any clustering algorithm needs to perform well.

Clustering of the Embeddings: Clustering of these embeddings forms the next step. This is done by HAC (Hierarchical Agglomerative Clustering) wherein the algorithm at every step merges the closest cluster. The distance of average linkage defines the proximity of two clusters. The distance between two clusters, say A and B, can be given as

$$d(A, B) = 1/(|A||B|)\ \Sigma\_\{a \in A, b \in B\}\ d(a, b)$$

where $d(a, b)$ denotes the distance between two individual examples. This clustering step is performed only once on embeddings of the entire pool X thus greatly reducing the computations. In this approach, by using HAC the algorithm would be effectively able to group similar examples which becomes crucial for further sampling.

In the final step of sampling, the algorithm selects those examples in which the model is least confident. To do this, margin scores are calculated for every example. This can mathematically be defined as the difference between the highest and the second highest predicted scores from the model. Margin score of an example X can be written as:

$$margin(x) = f(x) - f\_2(x)$$

where $f(x)$ is the score of the most confident class and $f_2(x)$is the score of the second most confident class. Examples with the lowest margin scores are those that the model finds most ambiguous, hence those are prime candidates for labelling.

It will draw from M, the set of least confident examples, in order to retrieve the underlying clusters Cm, and will choose a representative subset through a round-robin sampling strategy. This round-robin approach ensures that the selected batch is not only focused on the low-confidence examples but also retains diversity across different clusters. By sorting the clusters according to size and then sampling each cluster in a balanced way, the algorithm reduces risk of redundancy so that selected examples can represent a wide spectrum of data.

Empirical verification also provides evidence for the Cluster-Margin algorithm. Extensive experiments are conducted on large-scale datasets, such as the Open Images Dataset, showing that Cluster-Margin significantly outperforms the existing methods in label efficiency. For example, Cluster-Margin uses only 40% of the labels of the next best method while achieving comparable performance. Such amazing efficiency is highly desirable in many real-world applications due to the high labelling cost.

3. A list of novelties/contributions

The Cluster-Margin algorithm contains a few major steps forward in the process of batch active learning, particularly for large-scale datasets. Its effectiveness for batch sizes up to 100K and even 1M distinguishes it from prior methods like BADGE and CoreSet, which could not handle more than 1,000 points. This is further reflected in its excellent label efficiency, where at the end of training with 1.3M examples, it achieves a final pooled average precision of more than 0.76, using only 40% of labels needed by the next best method.

One of the key novelties of Cluster-Margin is that it uses Hierarchical Agglomerative Clustering as a kind of preprocessing step. Since it performs HAC once on the pool of unlabeled data, it allows for efficient sampling in the subsequent iterations, which cuts computational costs compared to methods requiring repeated clustering. Indeed, this is the efficiency to be converted to performance-as one can see, Cluster-Margin outperformed CoreSet in empirical studies.

The algorithm's sampling mechanism combines the principles of diversity and margin-based selection, focusing on examples with the smallest margins between class probabilities, while ensuring that this choice is diverse via a round-robin scheme across clusters. Cluster-Margin achieved 85% accuracy on CIFAR10 and 78% on CIFAR100, considerably outperforming classical uncertainty sampling methods. More importantly, it required only 37% of the labels compared to Margin Sampling for a batch size of 1M to achieve comparable performance.

Extensive experiments on large-scale datasets further solidify Cluster-Margin's superiority. For instance, on Open Images, after the final pooled AP was over 76%, BADGE and CoreSet both remained under 70%. Its performance advantage is also consistently reproduced on various datasets like SVHN and CIFAR, where Cluster-Margin always stays in front in accuracy compared with other established methods like BADGE and Margin Sampling.

4. Would you say there are downsides to this work?

Some possible shortcomings are, nevertheless involved with the Cluster-Margin algorithm. First, although Cluster-Margin itself is an ingenious method, relying on HAC as preprocessing makes later sampling efficient but incorporates a computational bottleneck that may prevent its eventual use in dynamic, real-time situations. Such a preprocessing requirement contrasts with more flexible methods, such as BADGE and CoreSet, which can more easily adapt to a changing data environment. The

intrinsic implementational complexity of the algorithm, which includes clustering and managing various samplings, may prevent it from being adopted widely in practice. Simpler methods can, at times, give comparable results with less overhead for smaller data or less complicated tasks.

The Cluster-Margin heavily relies on the quality of the initial clustering. If the HAC method fails to capture the underlying data distribution due to a number of noisy examples or outliers, then further sampling may be performed sub optimally. It could be the point that gives sensitivity to initial clustering quality, which leads to potential performance degradation with high variance datasets. By being greedy with respect to the low-confidence examples of an established cluster, the algorithm could limit the search in the remainder of the data space and thus not retrieve any further informative examples from outside those identified clusters.