Medical Image Analysis 93 (2024) 103075

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

GANDALF: Graph-based transformer and Data Augmentation Active Learning Framework with interpretable features for multi-label chest Xray classification

Dwarikanath Mahapatra [a,d,*], Behzad Bozorgtabar [b,c], Zongyuan Ge [d], Mauricio Reyes [e]

[a] Inception Institute of AI, Abu Dhabi, United Arab Emirates
[b] École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[c] Lausanne University Hospital (CHUV), Lausanne, Switzerland
[d] Faculty of IT, Monash University, Melbourne, Australia
[e] ARTORG Center for Biomedical Engineering Research, University of Bern, Bern, Switzerland

ARTICLE INFO

Keywords:
Multi-label
Informative samples
Active learning
Data augmentation

ABSTRACT

Informative sample selection in an active learning (AL) setting helps a machine learning system attain optimum performance with minimum labeled samples, thus reducing annotation costs and boosting performance of computer-aided diagnosis systems in the presence of limited labeled data. Another effective technique to enlarge datasets in a small labeled data regime is data augmentation. An intuitive active learning approach thus consists of combining informative sample selection and data augmentation to leverage their respective advantages and improve the performance of AL systems. In this paper, we propose a novel approach called GANDALF (**G**raph-based Tr**AN**sformer and **D**ata **A**ugmentation Active **L**earning **F**ramework) to combine sample selection and data augmentation in a *multi-label setting*. Conventional sample selection approaches in AL have mostly focused on the single-label setting where a sample has only one disease label. These approaches do not perform optimally when a sample can have multiple disease labels (e.g., in chest X-ray images). We improve upon state-of-the-art multi-label active learning techniques by representing disease labels as graph nodes and use graph attention transformers (GAT) to learn more effective inter-label relationships. We identify the most informative samples by aggregating GAT representations. Subsequently, we generate transformations of these informative samples by sampling from a learned latent space. From these generated samples, we identify informative samples via a novel multi-label informativeness score, which beyond the state of the art, ensures that (i) generated samples are not redundant with respect to the training data and (ii) make important contributions to the training stage. We apply our method to two public chest X-ray datasets, as well as breast, dermatology, retina and kidney tissue microscopy MedMNIST datasets, and report improved results over state-of-the-art multi-label AL techniques in terms of model performance, learning rates, and robustness.

1. Introduction

While annotated medical image datasets are instrumental for deep learning (DL) methods to obtain state-of-the-art (SOTA) performance, annotating medical data is challenging due to the high expertise and costs involved. Active Learning (AL) methods enable an expert to select informative samples and allow a model to obtain high performance with minimal labeled samples (i.e., high learning rates). This is particularly suitable for medical image analysis tasks where AL methods must adapt to varying conditions like device vendor, imaging protocol, machine learning model, etc.

Data augmentation is also an effective approach where models are provided with synthetic samples generated from the training dataset (Perez and Wang, 2017). While conventional data augmentation (e.g., flipping, rotating, etc.) increases the dataset size, it does not ensure that informative samples are added to the training set. Recent works have used neural network approaches for data augmentation with Generative Adversarial Networks (GAN) (Goodfellow et al., 2014), Spatial Transform Networks (STN) (Jaderberg et al., 2015), and Variational Autoencoder (VAE) (Kingma and Welling, 2013). To enable greater exploration of the feature space, Mixup (Zhang et al., 2017) and its variants interpolate features and use the corresponding labels to enhance the dataset.

Combining data augmentation and active learning can help leverage both approaches' advantages. Tran et al. (2019) use a pipelined

* Corresponding author at: Inception Institute of AI, Abu Dhabi, United Arab Emirates.
E-mail address: dwarikanath.mahapatra@inceptioniai.org (D. Mahapatra).

https://doi.org/10.1016/j.media.2023.103075
Received 16 February 2023; Received in revised form 26 November 2023; Accepted 29 December 2023
Available online 6 January 2024
1361-8415/

approach to select informative samples by an acquisition function and generate augmented samples from them. Such a sequential approach, does not use the mutual interactions between the two steps. i.e., the acquisition function does not evaluate the potential information gain from the augmented samples. Consequently, the generated data does not guarantee informativeness. Kim et al. in Kim et al. (2021) propose an integrated approach, Look Ahead Data Augmentation (LADA), that evaluates the informativeness of potential augmentations and generates informative samples. Thus, augmented samples provide qualitatively different information from the base samples. However, as shown in our experiments, the LADA method is not equally effective for the multi-label setting.

Current AL and data augmentation methods have been developed for a single-label setting. In contrast, there are scenarios where an image has multiple disease labels, such as chest X-ray images. Informative sample selection in a multi-label setting is more challenging since one needs to consider the mutual influence and similarity of all potential class labels, as well as the different levels of class complexity (i.e., some diseases are more easily detectable than others). Additionally, augmenting such informative samples should, in turn, ensure appropriate informativeness of the new samples. In this paper we propose a novel method for multi-label active learning combined with a novel approach for informative sample data augmentation.

2. Prior work

2.1. Active sample selection

Efficient sample selection methodologies are essential to obtain optimal system performance with minimal expert interventions. This is particularly important to the time-pressured workflow of many clinical settings. Different informative sample selection approaches have been investigated for deep learning based medical image analysis, including sample entropy (Zhu and Bento, 2017), model uncertainty (Mahapatra et al., 2018; Gal et al., 2017), Fisher information (Sourati et al., 2019), visual saliency (Mahapatra and Buhmann, 2016) and clustering-based sample selection (Zheng et al., 2019).

Sample entropy quantifies a sample's difficulty in classification, with higher entropy characterizing higher sample informativeness. Wang et al. (2017a) use sample entropy, a least-confidence component, and margin sampling to select informative samples. The work in Zhou et al. (2016) uses GANs to synthesize samples close to the decision boundary, which are then annotated by human experts. Mayer and Timofte (2018) use a GAN model to generate high entropy samples, which are used as a proxy to find the most similar samples from a pool of real sample candidates to be annotated by experts.

Uncertainty-based methods identify informative samples for which a model is most uncertain. Yang et al. (2017) propose a two-step sample selection approach based on uncertainty estimation, followed by a second selection step based on a maximum set coverage similarity metric. Test-time Monte-Carlo dropout (Gal et al., 2017) has been used to estimate sample uncertainty, and consequently select the most informative ones for label annotation (Gal et al., 2017; Bozorgtabar et al., 2019).

The state-of-the-art in active learning is mostly dominated by methods relying on uncertainty estimations. However, the reliability of uncertainty estimations has been questioned for deep neural networks used in computer vision and medical imaging applications due to model calibration issues (Abdar et al., 2021; Guo et al., 2017; Jungo et al., 2020) These results reinforce our suggestion to examine other methods for selecting informative samples in active learning. Additionally, as highlighted in a recent survey paper on active learning and human-in-the-loop learning by Budd et al. (2021), interpretability is essential for medical imaging scenarios where experts and AI systems collaborate.

In Mahapatra et al. (2021), we proposed an interpretability-guided sample selection approach featuring state-of-the-art performance for classification and segmentation tasks. The approach ranks informative samples via an auxiliary machine learning model ranking samples based on the saliency map information of the class predicted by the primary model. Overall, the approaches mentioned have not been designed to work on multi-label classification problems where a sample can have multiple labels assigned. The next section describes specifically the state-of-the-art in multi-label active learning.

2.2. Multi-label active learning

A comprehensive survey on multi-label active learning can be found in Wu et al. (2020) and Yang et al. (2015). Wu et al. (2014) propose "Example-label"-based AL (LEMAL) where samples are selected based on maximum uncertainty across all label classes. Reyes et al. (2018) propose two uncertainty measures based on the base classifier predictions and uses a measure of the inconsistency of a predicted label set to select the most informative samples. Li and Guo (2013) propose a max-margin prediction uncertainty strategy and a label cardinality inconsistency strategy to measure the unified informativeness of unlabeled instances. Different from the above works, in Mahapatra et al. (2022a) we focus on the multi-label setting using different aggregation strategies of information derived from all potential class labels. The approach distills information from saliency maps computed for all class labels – referred to as intra-sample saliency maps – and builds a graph representation describing the similarity of intra-sample saliency maps. Informative samples are then characterized by graphs describing high similarity across intra-sample saliency maps. In Mahapatra et al. (2022a), we used simple graph aggregation strategies, such as weighted summation of edge weights and nodes, to identify the most informative samples. In the present work, we propose improved mechanisms to aggregate this information via graph transformers to better represent intra-sample saliency maps and a sample informativeness data augmentation strategy. Also, in our previous work Mahapatra et al. (2022a), no data augmentation strategy was used, which features class-label preservation and redundancy avoidance, as proposed in the current study. The next section describes the state-of-the-art in active learning with data augmentation.

2.3. Active learning with data augmentation

Prior work on leveraging data augmentation for active learning includes Bayesian Generative Active Deep Learning (BGADL), which combines acquisition and augmentation steps in a pipelined approach (Tran et al., 2019). However, a large number of labeled instances are needed to train the generative model, and BGADL does not measure the potential information gain from data augmentation. Consistency-based Active Learning (CAL) algorithms consider data augmentation in the acquisition process, by replacing the uncertainty with an augmentation-based CAL inconsistency term. The work of Bozorgtabar et al. (2019) builds on Mahapatra et al. (2018) and proposes a Class-Aware Generative Adversarial Network (CAGAN) to incorporate a class-balancing component to ensure that synthetically-generated samples have the same class label of the reference sample they are conditioned on. Kim et al. (2021) propose an approach termed look-ahead-data-augmentation (LADA) for selecting informative samples where the informativeness of generated samples is also considered. However, as show in our experiments, the LADA method does not perform accurately in a multi-label setting.

We also discuss other works that uses graphs or attention mechanisms for medical image analysis. Xiong et al. (2023) propose a transformer based Hierarchical Attention-Guided Multiple Instance Learning (HAG-MIL) for Whole Slide Image Classification (WSI). They use an Integrated Attention Transformer (IAT) exploiting multiple resolutions of the WSIs. Xiao et al. (2023) describe a method to perform domain adaptation using graphs that uses a novel representation learner. Lian et al. (2022) use GCN for lung lobe segmentation by using average pooling

to get a single value representation of features. Our method uses Graph attention transformers, which is more robust since it uses multi-head attention for quantifying node interactions. Velickovic et al. (2018) propose a graph attention network, which uses attention networks on graphs to learn better representations.

2.4. Contributions

In Mahapatra et al. (2022b), we demonstrate that saliency maps can be seen as fingerprints of a model's response to an input, and can be used as an inductive bias of the training process. In a follow-up work Mahapatra et al. (2022a), we propose an interpretability-guided sample informativeness selection approach, where intra-sample saliency maps are used as nodes of a graph structure to characterize the distinctiveness of class-specific saliency maps. In Mahapatra et al. (2022a) we use simple graph aggregation strategies, such as weighted summation of edge weights and nodes, to identify the most informative samples. In this present work, we build on these findings and propose to use graph attention transformers (GATs) that enable learning more accurate representations than simple averaging of graph edge information.

GATs identify most important nodes through a self-attention mechanism and better learn global relationships among different graph nodes Additionally, we also propose a novel approach that learns transformations (or augmentations) of a sample to generate informative synthetic ones used to boost model training. We call this *informative augmentation*. Previous state-of-the-art approaches do not ensure that newly generated synthetic samples are informative, and hence lead to redundancy of large numbers of generated samples. In contrast, with our proposed informative augmentation scheme, we have an improved and more efficient informative sample augmentation approach. GATs combined with informative augmentation results in a novel approach termed GANDALF (**G**raph-based Tr**AN**sformer and **D**ata **A**ugmentation Active **L**earning **F**ramework) that outperforms previous work on multi-label active learning approach for chest X-ray classification This paper makes the following contributions:

1. We use graph attention transformers to incorporate the importance of different nodes and better quantify graph informativeness. Previously used simple aggregation, such as the sum and mean of graph edge weights, do not emphasize nodes with information that can more effectively contribute to the task at hand.
2. We propose a novel score termed as multi-label informativeness score, that quantifies the importance of each sample based on multi-label interactions. This multi-label informativeness score is derived from the novel use of graph attention transformers.
3. A novel data augmentation approach to generate novel transformations such that new synthetic images are also informative compared to their base image while enforcing class label preservation and redundancy avoidance.
4. Based on benchmarking the proposed GANDALF method and nine other SOTA methods, on two publicly available chest X-ray datasets and four other multi-class MedMNIST datasets, we demonstrate that by combining data augmentation with multi-label active learning, the proposed approach outperforms SOTA methods for multi-label classification.

3. Methods

3.1. Outline of the proposed method

In our current work, we propose a model to identify multi-label informative samples by jointly considering the mutual influence of all potential class labels. We also synthesize new samples from identified informative real ones such that newly synthesized samples are warranted to be informative and provide new information to the training set. Fig. 1 depicts the different stages of our proposed workflow.

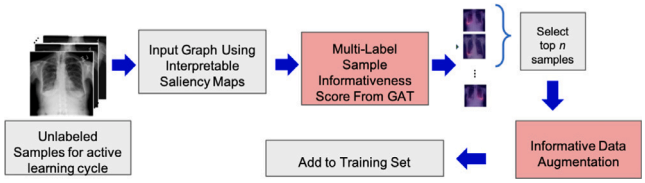

Fig. 1. Workflow of proposed GANDALF method. Given unlabeled pool samples in an active learning cycle, an input graph is constructed using interpretable saliency maps. They serve as input to a graph-multiset transformer (Fig. 2) which outputs a informativeness score to rank informative samples. Selected samples are then used to further synthesize informative and non-redundant samples. Selected samples and their synthetic derivatives are added to the training dataset for the next active learning cycle.

We divide the description of the proposed active learning approach into its two main components: (1) Sample selection based on multi-label informativeness scoring, and (2) Synthetic generation of informative and non-redundant samples. The general strategy is as follows: For the synthetic sample generation step, we take the initial batch of training images and train a variational autoencoder (VAE) to generate images. In parallel, an initial classification model is trained with a baseline number of randomly chosen training samples (which can be the same data used to train the VAE model, or a different one) to act as a baseline model for the active learning cycles. During test time, informative samples are identified by first deriving interpretability saliency maps for each unlabeled sample (i.e, pool sample). Following Mahapatra et al. (2022a,b), we calculate saliency maps for all potential class labels (intra-sample saliency maps), from which a graph is constructed with nodes corresponding to the latent representation of each class-specific saliency map, and edge weights characterizing the similarity between nodes. Different from Mahapatra et al. (2022a,b), in Section 3.2.1 we propose an improved mechanism to rank pool samples by their level of informativeness. Selected informative samples are then used as base images, and their mean and variance vectors are obtained from the encoder part of the previously trained VAE. New synthetic samples are generated from the base images by sampling their respective distributions and feeding them to the VAE's decoder. The level of informativeness of these generated samples is further evaluated, and samples that are found to be sufficiently informative are added to the training set. The classifier model is then updated with the set of new samples. This set of steps is repeated till no new informative samples are found. In the next sections, we describe in detail the two main components of our method. 1. VAE training and Augmentation using sampling. 2. Multi-label sample informativeness

To choose the initial set of training samples we randomly select equal number of samples from each label. In practice this amounts to $3-5\%$ of the training data. The selection of the initial sample set is termed as cold start active learning and it is a recent area of active research (Nath et al., 2022). Warm start active learning refers to selection of initial samples using different techniques other than random selection. Identifying novel methods for initial sample selection is an active area of research and prior work includes clustering techniques (Ash et al., 2020; Yuan et al., 2020), semi supervised learning for classification (Siméoni et al., 2021), and based on 2D segmentation (Wang et al., 2020; Lai et al., 2021). For comparison we adopt a cluster-based warm start, where we cluster the samples into the number of diseased labels and identify their centroid. Thereafter, we choose samples closest to the centroid in feature space and use them as our initial batch of training samples. As shown in the results, we notice an improvement in performance using this approach (See Table 1 (GANDALF_{Warm})).

While the warm start approach improves performance, in the case of unbalanced labels our approach has been to select an equal number

Table 1
AUC values for different baselines and proposed approach along with ablation studies. The $p-$values are with respect to GANDALF. DT: DeepTaylor; GC: GradCAM; Warm: Warm AL start; red: GANDALF with only redundancy avoidance; label: GANDALF with only label preservation; con-DA: Conventional Data Augmentation; GESTALT: GESTALT for sample informativeness, and GANDALF for informative augmentation; pooling: Conventional pooling; no-GAT: No Graph Attention; no-SA: No Self-Attention.

	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	p-
FSL	90.23	90.23	90.23	90.23	90.23	90.23	90.23	90.23	90.23	90.23	<0.001
Random	41.69	47.5	52.6	58.1	64.07	69.34	75.72	81.49	85.64	90.23	<0.001
Entropy	60.0	64.2	68.7	75.6	81.1	85.4	87.7	89.2	90.8	91.4	<0.001
Unc	62.15	66.56	72.41	80.16	85.80	88.12	90.34	90.72	90.84	91.03	<0.001
LEMAL (Wu et al., 2014)	64.70	69.17	76.08	81.25	88.32	89.03	91.09	91.40	91.74	92.08	<0.001
CVIRS (Reyes et al., 2018)	66.05	70.88	76.05	84.12	89.09	91.02	91.48	91.52	92.03	92.54	0.001
AlphaMix (Parvaneh et al., 2022)	68.30	72.57	79.11	87.22	91.43	93.12	93.47	93.87	94.09	95.01	0.005
GAL (Long et al., 2008)	68.91	73.01	79.61	87.99	91.74	92.82	93.30	93.91	94.24	94.56	0.007
LADA (Kim et al., 2021)	69.15	73.98	80.82.	89.08	92.93	93.64	94.04	94.72	95.13	95.73	0.001
GESTALT (Mahapatra et al., 2022a)	70.82	74.64	80.82	89.27	93.28	94.13	94.92	95.24	95.88	96.71	0.02
Info-Max (Xiao et al., 2023)	70.12	73.21	78.34	84.61	87.19	89.78	92.84	94.22	94.83	95.32	0.02
IAT (Xiong et al., 2023)	70.76	73.91	79.48	85.82	88.91	90.18	94.04	94.62	95.73	96.82	0.02
GANDALF_{DT}	72.24	76.18	82.47	91.02	94.90	95.88	96.42	96.92	97.43	98.21	–
$\text{GANDALF}_{DT-NoWeight}$	69.7	73.0	77.2	80.7	83.6	86.8	89.7	91.8	93.6	94.9	0.001
GANDALF_{GC}	71.63	75.84	81.16	90.35	94.53	95.11	95.79	96.07	96.7	97.57	0.07
$\text{GANDALF}_{GC-NoWeight}$	69.2	72.4	76.3	79.5	82.4	85.4	88.5	90.7	92.3	93.5	0.001
GANDALF_{Warm}	73.8	78.3	84.6	92.89	95.60	96.41	97.1	97.72	98.3	98.94	0.041
$\text{GANDALF}_{Warm-NoWeight}$	71.1	75.1	80.2	83.2	85.3	88.16	90.2	92.5	93.6	94.8	0.001
Ablation studies											
GANDALF_{red}	68.22	71.83	78.23	84.68	87.12	90.34	91.97	92.47	93.47	94.01	0.03
GANDALF_{label}	65.61	68.25	74.37	79.51	82.02	85.48	87.15	88.02	89.42	90.14	0.03
GANDALF_{con-DA}	71.23	75.32	81.83	89.93	94.08	94.45	95.62	95.89	96.24	97.41	0.034
$\text{GANDALF}_{GESTALT}$	71.41	75.37	81.93	90.18	94.09	94.81	95.75	96.61	97.01	97.51	0.032
$\text{GANDALF}_{pooling}$	64.13	67.24	71.34	76.23	79.21	83.02	86.25	87.12	87.64	88.52	0.032
GANDALF_{no-GAT}	62.21	65.52	69.71	73.95	77.12	80.94	83.84	85.36	86.02	86.78	0.023
$\text{GANDALF}_{no-MHSA}$	61.35	64.93	68.62	73.02	76.21	79.41	82.29	84.62	85.11	85.77	0.011

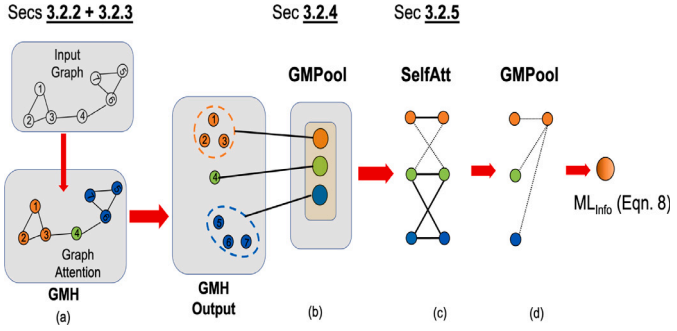

Fig. 2. Graph Multiset Transformer: A graph with n nodes depicting multi-label information of a sample is passed through several message-passing layers (a) and an attention-based pooling block (GMPool_k) (b) to get $k(<n)$ nodes. A self-attention block (SelfAtt) (c) encodes the relationship between k nodes, and passes through GMPool_1 (d), to obtain a single node value. Different node colors indicate different classes and the edge length denotes node similarity.

of samples from all classes, and also use a weighted loss function (weights of each class being inversely proportional to the number of samples). When all classes have equal samples the weights are equal. When the classes exhibit an imbalance then the weighted loss function addresses that to a certain extent. Without using a weighted loss the performance certainly drops off (See Table 1 $\text{GANDALF}_{Warm-NoWeight}$, $\text{GANDALF}_{GC-NoWeight}$, $\text{GANDALF}_{DT-NoWeight}$).

3.2. Multi-label sample informativeness

We use a graph-based approach for calculating the informativeness score of each image, and adopt the following steps.

1. We represent each image sample as a separate graph.
2. Nodes of the graph correspond to the potential class labels representing a disease or condition (i.e, the number of nodes in the graph equals the number of represented classes).
3. Following our findings in Mahapatra et al. (2022b,a), at each node, a class label is represented as the latent representation of the corresponding class-specific interpretability saliency map, which can be obtained through any available interpretability approach as done in Alber et al. (2019).
4. Edge weights in the graph represent the similarity between corresponding nodes. The similarity between labels (or nodes) is calculated by determining the cosine similarity between the latent representations of their class-specific saliency maps. As shown in our previous works (Mahapatra et al., 2022b,a), using the latent representations of saliency maps leads to improved results over using image features.

Assuming there are K nodes in each graph (i.e., K classes), each node has $K-1$ edge weights connecting to all the other nodes. Let us denote the edge weight between nodes i, j as w_{ij}, which is defined as

$$w_{ij} = cosine\_similarity(z_{S_{I,i}}, z_{S_{I,j}}) = \langle z_{S_{I,i}}, z_{S_{I,j}} \rangle, \quad (1)$$

where $z_{S_{I,i}}$ and $z_{S_{I,j}}$ are the latent feature vectors derived from saliency maps S of sample image I for class labels i and j, respectively. The latent feature vectors are obtained by forward-passing the saliency maps until a selected layer (e.g., second-to-last layer). Cosine similarity is a commonly used metric employed to compare latent representations. Since its range of values is bounded, the cosine similarity is also a good option for its inclusion in a loss, but we note that other similarity metrics could be used (e.g. L2 distance metric).

In conventional approaches, informative samples are determined based on the assumption that each sample has one label. However, to identify the most informative samples for multi-label settings, we propose to incorporate class-label interactions using a graph-based ranking metric scheme that leverages graph transformers, and a pooling mechanism to learn better global relationships among graph nodes. Below, we describe this in more detail.

3.2.1. Graph transformers

Graph pooling enables the calculation of lower-dimensional informative representations of high-dimensional graphs. The two main approaches to graph pooling are: (1) Node-drop methods (Zhang et al.,

2018; Lee et al., 2019a), which drops nodes with low scores based on information from graph convolutional layers, and (2) Node clustering methods (Ying et al., 2018; Bianchi et al., 2019), which merge similar nodes into a single one by exploiting their hierarchical structure.

Both pooling approaches have prominent drawbacks. Node-drop methods tend to drop informative nodes at every pooling step. Node-clustering methods compute a dense cluster assignment matrix that prevents them from exploiting sparsity in the graph topology. This leads to an excessively high computational complexity (Lee et al., 2019a). To accurately represent the graph, the obtained representation should be as powerful as the Weisfeiler–Lehman (WL) graph-isomorphism test (Weisfeiler and Leman, 1968), such that two different graphs are mapped to two distinct embeddings.

To overcome the limitations of existing methods, we adopt a graph-structured attention unit based on Graph Multiset Transformer (GMT) (Baek et al., 2021). GMTs are a pooling mechanism that condenses the given graph into a set of representative nodes, and then further uses attention to encode relationships among them to enhance the representation power of the graph. A GMT architecture can accurately represent an entire graph, given a multiset of node features. We first describe the multiset encoding scheme that enables the embedding of two different graphs into distinct embeddings. This condition is essential to ensure that different images (or samples) have unique representations. We then describe the graph multi-head attention mechanism that projects the graph topology in the attention-based multiset encoding.

3.2.2. Multiset encoding

A graph pooling function takes the set of graph nodes as input, which forms a multiset (i.e., allowing for repeating elements) since different nodes can have identical feature vectors. A graph pooling function that is as powerful as the WL test needs to satisfy the permutation invariance and injectiveness criteria over the multiset. Two non-isomorphic graphs should be encoded differently through the injective function. A simple sum pooling, as done in our previous work (Mahapatra et al., 2022a), satisfies the injectiveness condition (Xu et al., 2019) but does not consider each node's relevance to the task and treats them equally. This limitation is addressed using an attention mechanism on the multiset pooling function to capture structural dependencies among nodes within a graph, described below.

3.2.3. Graph multi-head attention (GMH)

Assume that we have n node vectors. The input of the attention function (Att) consists of query $\mathbf{Q} \in \mathscr{R}^{n_q \times d_k}$, key $\mathbf{K} \in \mathscr{R}^{n \times d_k}$ and value $\mathbf{V} \in \mathscr{R}^{n \times d_v}$, where n_q is the number of query vectors, n is the number of input nodes, d_k is the dimensionality of the key vector, and d_v is the dimensionality of the value vector. We compute the dot product of the query with all keys to assign higher importance (i.e., higher weight) to the relevant v nodes, as follows:

$$Att(\mathbf{Q},\mathbf{K},\mathbf{V}) = Ac(\mathbf{Q}\mathbf{K}^T)\mathbf{V}, \tag{2}$$

where Ac is a softmax activation function. Following the work of Vaswani et al. (2017), instead of computing a single attention head, one can use multi-head attention by linearly projecting the query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ h times, to yield h different representation subspaces. To facilitate the descriptions, we first describe multi-head attention below, to then motivate the improved approach we employed.

The output of the multi-head attention function (MH) is denoted as $MH(\mathbf{Q},\mathbf{K},\mathbf{V})$.

$$MH(\mathbf{Q},\mathbf{K},\mathbf{V}) = [O_1,\dots,O_h]\mathbf{W}^O \tag{3}$$
$$\text{with } O_i = Att(QW_i^Q, KW_i^K, VW_i^V)\ \mathrm{i} = \{1,\dots,h\}$$

The parameter matrices are $\mathbf{W}_i^Q \in \mathbb{R}^{d_k \times d_k}$, $\mathbf{W}_i^K \in \mathbb{R}^{d_k \times d_k}$, and $\mathbf{W}_i^V \in \mathbb{R}^{d_v \times d_v}$. The output projection matrix is $\mathbf{W}_i^O \in \mathbb{R}^{hd_v \times d_{model}}$, where d_{model} is the output dimensionality of the multi-head attention (MH) function.

Multi-head attention is superior to trivial pooling methods such as *sum* or *mean* as it considers global dependencies among nodes. However, the MH function suboptimally generates the key $\mathbf{K}$ and value $\mathbf{V}$ since it linearly projects the obtained node features $\mathbf{H}$. To tackle this limitation, we define a graph multi-head attention block (GMH). Given node features $\mathbf{H}$ with their adjacency matrix $\mathbf{A}$, key and value pairs, $\mathbf{K}$, $\mathbf{V}$ are constructed using Graph neural networks (GNNs) to explicitly leverage the graph structure as follows :

$$GMH(\mathbf{Q},\mathbf{H},\mathbf{A}) = [O_1,\dots,O_h]\mathbf{W}^O;$$
$$\text{with } O_i = Att(QW_i^Q, GNN_i^K(\mathbf{H},\mathbf{A}), GNN_i(\mathbf{H},\mathbf{A}))\ \mathrm{i} = \{1,\dots,h\}, \tag{4}$$

3.2.4. Graph Multiset Pooling (GMPool) with graph multi-head attention

Given node features $\mathbf{H}$ from GNNs, we define a Graph Multiset Pooling (GMPool), which is inspired by Transformers (Vaswani et al., 2017; Lee et al., 2019b), to compress the n nodes into k nodes ($k < n$), with a parameterized seed matrix S.

$$GMPool_k(\mathbf{H},\mathbf{A}) = LN(Z + rFF(Z));$$
$$\text{with } Z = LN(S + GMH(\mathbf{S},\mathbf{H},\mathbf{A})), \tag{5}$$

where rFF is any row-wise feedforward layer that processes each individual row independently and identically, and LN stands for layer normalization (Ba et al., 2016). The GMH function considers interactions between k seed vectors (queries) in S and n nodes (keys) in $\mathbf{H}$, to compress n nodes into k clusters with their attention similarities between queries and keys.

3.2.5. Self-Attention for inter-node relationship (SelfAtt)

To consider the interactions among different graph nodes, we use a Self-Attention function (*SelfAtt*), inspired by the Transformer architecture (Vaswani et al., 2017; Lee et al., 2019b):

$$SelfAtt(\mathbf{H}) = LN(Z + rFF(Z));$$
$$\text{with } Z = LN(\mathbf{H} + MH(\mathbf{H},\mathbf{H},\mathbf{H})), \tag{6}$$

where rFF is any row-wise feedforward layer that processes each individual row independently and identically, and LN is a layer normalization (Ba et al., 2016) The *SelfAtt* function captures inter-relationships among n nodes by placing node embeddings $\mathbf{H}$ on both query and key locations in MH (Eq. (3)).

3.2.6. Overall architecture

The overall architecture is shown in Fig. 2. For a graph G with node features $\mathbf{X}$ and an adjacency matrix $\mathbf{A}$, the encoder is denoted as:

$$Encoder(X,A) = GNN_2(GNN_1(\mathbf{X},\mathbf{A}),\mathbf{A}), \tag{7}$$

where we stack two GNNs to construct the deep structures of the Graph Multiset Transformer. In practice, one can have multiple GNNs, but in our experiments we noticed that going beyond two GNNs did not increase performance. After obtaining a set of node features $\mathbf{H}$ from an encoder, the pooling layer aggregates the features into a single vector. Finally, we get the reduced representation of the entire graph by using GMPool with $k = 1$ as follows:

$$ML_{info} = Pooling(H,A) =$$
$$GMPool_1(SelfAtt(GMPool_k(H,A)),A'), \tag{8}$$

where $\mathbf{A}'$ is the identity or coarsened adjacency matrix since adjacency information should be adjusted after compressing the nodes from n to k with $GMPool_k$.

The above step condenses the entire graph and represents it as a single value, ML_{info}, proposed here as a novel measure of graph informativeness. We hypothesize that higher ML_{info} values describe higher values of a graph's informativeness and hence suggest using ML_{info} as a direct metric of sample informativeness. To verify whether the value of ML_{info} truly reflects sample informativeness, we selected 500 images at random and calculated their ML_{info} and uncertainty values. Since uncertainty is a widely used metric to quantify the informativeness of a sample (Yang et al., 2017; Gal et al., 2017), we

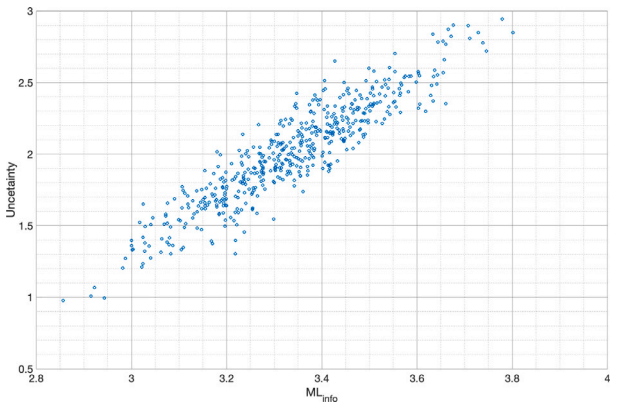

Fig. 3. Scatter plot between ML_{info} and uncertainty values of 500 samples. The points are concentrated around a line with a high correlation coefficient (0.92), suggesting that ML_{info} is a good measure of sample informativeness since uncertainty estimates have been used previously as a measure of sample informativeness.

hypothesized that for ML_{info} to be considered as a suitable measure of sample informativeness, a good level of correlation between ML_{info} and uncertainty should exist. Fig. 3 shows a scatter plot between the two sets of values. The majority of the samples are along a linear line, indicating that the two values are highly correlated, with a correlation coefficient of 0.92 indicating that higher values of ML_{info} indicate greater uncertainty and hence greater sample informativeness. This, provides initial evidence that ML_{info} is a good measure of sample informativeness.

Although our approach to quantifying multi-label informativeness is based on the paper by Baek et al. (2021) we incorporate the following novelties: (1) use of class specific saliency maps to represent each node; (2) we also validate the relevance of the graph pooling score as measure of multi-label informativeness.

3.3. Variational auto encoder training

Variational Autoencoders (VAEs) are helpful for generative modeling since their latent spaces are well-suited for random sampling and interpolation. The encoder of VAEs outputs two vectors of the same size: a vector of means, $\vec{\mu}$, and a vector of standard deviations, $\vec{\sigma}$. Given an encoding of an image, one can generate variations of it by sampling from the parameters $\vec{\mu}$ and $\vec{\sigma}$. For smooth interpolation amongst encodings and construction of new samples, a cost function defined as a combination of the reconstruction error term and the Kullback–Leibler (KL) divergence is commonly used:

$$\mathscr{L}_{VAE} = \sum \|x - \hat{x}\|^2 + KL\left[\mathcal{N}(\mu_x, \sigma_x), \mathcal{N}(0,1)\right] \tag{9}$$

For VAEs, the KL loss is equivalent to the sum of all KL divergences between the component $X \sim N(\mu_x, \sigma_x^2)$ and the standard normal, and is minimized when $\mu_x = 0, \sigma_x = 1$.

3.3.1. Informative synthetic image generation

After learning a VAE from the initial training data, we describe below the steps for our novel approach of generating variations of a given input image. We first identify informative samples from a given pool using the ML_{info} metric. Given informative image I, we pass it through the encoder to obtain latent distribution parameters μ_I, σ_I. By sampling from this distribution, we generate different transformations of the original image I, which we denote as I^n.

In the image generation step we aim to synthesize images that are variations of the base informative image I and add those images to the training set that have novel information enriching the training set. Hence, we propose the following novel criteria to generate new images:

1. *Class label preservation*: Image I and all generated I^n should have the same class labels.
2. *Redundancy avoidance*: I^n's semantic content should be sufficiently different from I to ensure that I^n are identified as informative.

We propose a novel scoring function that unifies and quantifies the degree to which the generated images meet the above criteria. The final score for each generated image is used to rank the images and select the most informative ones.

Preserving Label Similarity: To preserve the labels between I and I^n we focus on the class probability values of each image. The hypothesis here is that class label preservation is enforced if the class labels of synthetic images I^n are similar to the set of probabilities calculated for their base image I For a given base image I, we calculate class probabilities for each label using the current classification model.

Let us denote the probability of class k for image I as p_I^k. For generated image I^n, the corresponding probability is denoted as $p_{I^n}^k$. The idea is to ensure that if $p_I^k > 0.5$, then we would expect that $p_{I^n}^k > 0.5$. Alternatively if $p_I^k < 0.5$, we would expect $p_{I^n}^k < 0.5$. Additionally, the change in probability values should not be too high to avoid introducing spurious information. Consequently, we define the label score as a function of the relative difference between class probabilities computed for the base image I and their corresponding synthetic samples I^n, as follows:

$$\text{Score}_{label} = \begin{cases} \frac{\left|p_{I^n}^k - p_I^k\right|}{p_I^k}, & \text{if } \frac{\left|p_{I^n}^k - p_I^k\right|}{p_I^k} \leq \eta_1 \\ -\gamma_1, & \text{otherwise} \end{cases} \tag{10}$$

In our experiments, we set $\eta_1 = 0.3$ and $\gamma_1 = 0.1$, and we also put a condition that $p_I^k - 0.5$ and $p_{I^n}^k - 0.5$ always have the same sign. If the signs are different, then the labels are different; hence, the generated sample is not considered for informativeness evaluation. As per the above formulation if the difference of probability values is less than equal to 30% the score is same as this value. If the difference is greater than 30% it indicates a significant change of the probability distributions. Such a high change may bias the probability to large or low values, which indicate that the generated image is not informative since the classifier is very confident in predicting its label. Such an approach ensures that generated images with high distorted content are not included in the training set.

Redundancy avoidance: To ensure that the generated images have new information compared to the original image, we calculate an informativeness score for I^n and I in a similar manner as done to enforce label similarity. This score is called the multi-label informativeness score (MIS), whose calculation is based on Eq. (8) using the Graph Multiset Transformer strategy. We put a condition that the relative difference between their scores is above a threshold. We define the redundancy avoidance score as:

$$\text{Score}_{red} = \begin{cases} \frac{\left|MIS^I - MIS^{I^n}\right|}{MIS^I}, & \text{if } \eta_2 \leq \frac{\left|MIS^I - MIS^{I^n}\right|}{MIS^I} \leq \eta_3 \\ -\gamma_2, & otherwise \end{cases} \tag{11}$$

In our experiments, we set $\eta_2 = 0.05, \eta_3 = 0.25$, and $\gamma_2 = 0.15$ to define the range of thresholds within which the MIS score of the new sample can vary. The total informativeness of the generated sample is then calculated as,

$$\text{Score}_{sample} = \lambda_1 \text{Score}_{label} + \lambda_2 \text{Score}_{red}. \tag{12}$$

The higher the value of Score_{sample}, the higher its overall informativeness. While we want the labels of the original base image to be preserved in the generated images, we also aim to have diversity in information content to avoid redundancy. This is achieved using the Score_{Sample} term. The idea is to ensure a balance such that the generated images are not very different from the base image (to avoid

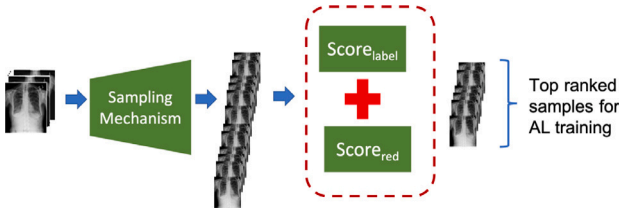

Fig. 4. Steps for augmenting and choosing informative samples from a base informative one. Given a few base informative samples we generate additional ones by sampling from a variational autoencoder (trained with the current training set). However, to ensure that only informative generated samples are added to the training set, we calculate two scores — the label score Score$_{label}$ and the redundancy avoidance score Score$_{red}$. The final informativeness of a sample is the weighted sum of these two scores. The top-ranked informative samples are added to the training set for further classifier training.

having unrealistic images) and at the same time introduce diversity. We rank the generated samples based on Score$_{sample}$ and select the top-n informative samples to add to the training set. λ_1, λ_2 are weights to determine the relative contribution of each term (see Fig. 4).

4. Baseline, implementation and dataset details

4.1. Baseline methods

In this section, we describe the baseline methods used for comparison purposes.

Random Sample Selection: As the first baseline, we considered a framework where no sample selection is considered, and pool samples are randomly selected for querying and active learning training. In clinical practice, the number of samples reflects the amount of user interaction needed to incorporate new samples into the next cycle of active learning. Hence, it needs to be kept as low as possible. In the results section, we refer to this approach as *Random*.

Fully supervised Learning: The fully supervised learning (FSL) baseline consists of a fully supervised approach trained on the designated training sets. It provides a performance reference obtained when training a model with all available data. We use a DenseNet-121 classifier (Huang et al., 2016) for the CheXpert dataset.

Uncertainty-driven sample selection: Uncertainty estimation can be used as a metric of sample informativeness for active learning, as proposed in Mahapatra et al. (2018). Given a deep learning model M used for disease classification, mapping an input image I, to a unary output $\hat{y} \in R$, the predictive uncertainty for pixel y is approximated using:

$$Var(y) \approx \frac{1}{T}\sum_{t=1}^{T}\hat{y}_t^2 - \left(\frac{1}{T}\sum_{t=1}^{T}\hat{y}_t\right)^2 + \frac{1}{T}\sum_{t=1}^{T}\hat{\sigma}_t^2 \tag{13}$$

$\hat{\sigma}_t^2$ is the model's output for the predicted variance for pixel y_t, and $\hat{y}_t, \hat{\sigma}_t^2{}_{t=1}^{T}$ being a set of T sampled outputs.

The obtained uncertainty estimates are sorted from high to low uncertainty, and the *top-n* samples are chosen for label querying and added to the next active learning cycle. In the results section, we refer to this approach as *Uncertainty*.

Competing Methods For Graph Informativeness

In our previous work GESTALT (Mahapatra et al., 2022a), we propose three different graph aggregation strategies, namely GESTALT$_{\text{Node}}$, GESTALT$_{\text{Link}}$ and GESTALT$_{\text{Weighted-Node}}$, using the sum and rank of different nodes and weights. GESTALT$_{\text{Weighted-Node}}$ was the best-performing version, we briefly describe below. For all subsequent results, we report results of GESTALT$_{\text{Weighted-Node}}$ as GESTALT.

Given N graphs (corresponding to a sample pool of N images to be ranked by their informativeness), each one with K nodes, we denote matrix $\mathbf{W} = (w_k^n) \in \mathbb{R}^{N\times K}$, with entry w_k^n describing the aggregated weight for the kth node of the nth graph, defined as:

$$w_k = \sum_{j\in(1,\ldots,K), j\neq k} w_{kj}. \tag{14}$$

For each kth node, aggregated weight nodes are compared across the set of N graphs by ranking them across columns of matrix $\mathbf{W}$. We denote matrix $\mathbf{R} = (r_k^n) \in \mathbb{R}^{N\times K}$, with column element $\mathbf{r}_k = rank(w_k^1, \ldots, w_k^N)^\intercal$.

To derive a final rank for sample n, we aggregate ranks by calculating the mean rank of the individual node ranks:

$$r_{mean}^n = \frac{1}{K}\sum_{k=1}^{K}\mathbf{R}(n,k). \tag{15}$$

Samples are then directly ranked, and the top-most informative ones are selected for active learning.

Given M training images (i.e., including those added, as the model is trained with newly queried samples), we redefine Eq. (14) as:

$$w_k = \sum_{j\in(1,\ldots,K), j\neq i} \alpha_{kj} w_{kj}, \tag{16}$$

where,

$$\alpha_{kj} = \frac{w_{kj} - \overline{w}_k}{\sigma_k} \tag{17}$$

$$\overline{w}_k = \frac{\sum_{m=1}^{M}\mathbf{W}(m,k)}{M} \tag{18}$$

$$\sigma_k = \sqrt{\frac{\sum_{m=1}^{M}\mathbf{W}(m,k) - \overline{w}_k}{M}} \tag{19}$$

The summary statistics, $\overline{w}_k$ and σ_k are calculated for each class label k of the aggregated node weight matrix $\mathbf{W} = (w_k^m) \in \mathbb{R}^{M\times K}$, and are used to construct z-scores α weights based on summary statistics extracted from the current training set. The motivation behind this variant is to incorporate a prior on the distribution of intra-sample similarities, modeled via aggregated node weights per class label. For further details, we refer the reader to Mahapatra et al. (2022a).

4.2. Implementation details

In implementing our method we first randomly choose the initial set of training samples (simulating an AL setup, with approximately 3%–5% training samples) and train a DenseNet-121 (Huang et al., 2016) model on the NIH ChestXray14 dataset (Wang et al., 2017b), as it is a common architecture used for the task of lung disease classification. Thereafter we select a batch of 128 pool images from the unlabeled dataset. For each image we generate saliency maps for each class, set up the graph nodes, calculate the feature vector of each node and the similarity between nodes, as explained in Eq. (1). We then calculate the multi-label informativeness score of each sample according to Eq. (8) and select the top 3 samples of each class as informative samples. Using these informative samples as base samples, we generate more informative samples using the steps described in Section 3.3. We rank the synthetic samples using the sample informativeness score of Eq. (12). The top informative samples are added along with the base image to the training set and the classifier is updated. We then start another round of finding informative base images and their variations, and continue till there are no more informative samples, or the entire training set has been used.

Our method is implemented in TensorFlow. We used Adam (Kingma and Ba, 2014) with $\beta_1 = 0.93$, $\beta_2 = 0.999$, batch normalization, binary cross-entropy loss, learning rate $1e{-}4$, 10^5 update iterations and early stopping based on the validation accuracy. The architecture and trained parameters were kept constant across compared approaches. Training and testing were performed on an NVIDIA Titan X GPU having 12 GB RAM.

Images are fed into the network with size 320 × 320 pixels. We employed 4-fold data augmentation (i.e., each sample augmented four times) using simple random combinations of rotations ($[-25, 25]°$), translations ($[-10, 10]$ pixels in horizontal and vertical directions), and isotropic scaling ($[0.95, 1.05]$ scaling factors). For the generation of interpretability saliency maps, we used default parameters of the iNNvestigate implementation of Deep Taylor (Alber et al., 2019), as well as for GradCAM (Selvaraju et al., 2017), used to assess model performance for a different interpretability approach. We selected these two representative methods based on their popularity and the good results we have obtained (Mahapatra et al., 2022b; Silva et al., 2020; Mahapatra et al., 2022a). For uncertainty estimation, we used a total of $T = 20$ dropout samples with dropout distributed across all layers (Kendall and Gal, 2017).

4.3. Dataset description

We used the CheXpert dataset (Irvin et al., 2019) consisting of 224,316 chest radiographs of 65,240 patients labeled for the presence of common chest conditions. The training set has 223,414 images, while validation and test sets have 200 and 500 images, respectively. The validation ground truth is obtained using majority voting from annotations of 3 board-certified radiologists. The consensus of 5 board-certified radiologists label test images. The test set evaluation protocol is based on 5 disease labels: *Atelectasis*, *Cardiomegaly*, *Consolidation*, *Edema*, and *Pleural Effusion*.

Considering the fact that most publicly available medical image datasets fall in the multi-class setting, the multi-label setting is restricted to chest X-ray datasets. We also run our model on the NIH ChestXray14 dataset (Wang et al., 2017b) having 112,120 expert-annotated frontal-view X-rays from 30,805 unique patients.

Additionally, we also show results on the *multi-class* MedMNIST dataset (Yang et al., 2021) due to its balanced and standardized datasets spanning across various modalities. We select subsets of the collection appropriate for multi-class disease classification, namely, (1) BreastMNIST (Al-Dhabyani et al., 2020) having 546/78/156 breast ultrasound images (consisting of 2 classes) in the training/validation/test split for malignancy detection; (2) DermaMNIST (Tschandl et al., 2018; Codella et al., 2019) having 7007/1003/2005 training/validation/test dermatoscope images for lesion classification (consisting of 7 classes), (3) RetinaMNIST (Liu et al., 2022) having 1080/120/400 training/validation/test fundus images for diabetic retinopathy severity grading (consisting of 5 classes), and (4) TissueMNIST having 165,466/23,640/47,280 training/validation/test Kidney Cortex Microscope images for multiple disease classification (consisting of 8 classes). Another important reason for choosing these datasets is the fact that other datasets show high AUC values for the benchmark methods, while these datasets provide the scope for demonstrating the advantages of informative sample selection.

4.4. Interpretability saliency map generator

Image-specific saliency maps operate under the basic principle of highlighting areas of an image that drive the prediction of a model. The importance of these areas can be obtained by investigating the flow of the gradients of a DL model calculated from the model's output to the input image, or by analyzing the effect of a pixel (or region) to the output when that pixel (or region) is perturbed. This type of visualization facilitates interpretability of a model but also serves as a confirmatory tool to check that algorithm-based decisions align with common domain knowledge (Reyes et al., 2020). To generate interpretability saliency maps we use the iNNvestigate library (Alber et al., 2019),[1] which implements several known interpretability approaches. We employ Deep Taylor, a known interpretability approach to generate saliency maps, due to its ability to highlight informative regions while yielding minimal importance to other regions. Deep Taylor operates similarly as other interpretability approaches by decomposing back-propagation gradients, of the studied model, into layer-wise relevance maps of individual cell activations, as a function of a queried input sample and class label (e.g. disease class) (Montavon et al., 2017). GradCAM maps are obtained using the method of Selvaraju et al. (2017). In Section 3.2 we have already described the approach for graph construction.

[1] https://github.com/albermax/innvestigate.

4.5. Comparison and ablation methods

In this section, we present the main results obtained by our proposed method GANDALF (**G**raph-based Tr**AN**sformer and **D**ata **A**ugmentation Active **L**earning **F**ramework) and compare it with other AL methods such as (1) 'GAL'- Graph-Based Active Learning (GAL) approach of Long et al. (2008); (2) 'LEMAL': the "example-label based sampling strategy (LEMAL)" approach by Wu et al. (2014); (3) 'CVIRS'- the Uncertainty sampling based on "Category Vector Inconsistency and Ranking of Scores (CVIRS)" approach of Reyes et al. (2018); (4) 'AlphaMix' - the "Active Learning by Feature Mixing (Alpha-Mix)" method of Parvaneh et al. (2022). We also compare with (5) the LADA method (Kim et al., 2021) that combines data augmentation and active learning but is designed for single-label cases, and with (6) our previous graph-based multi-label active learning method called 'GESTALT' (Mahapatra et al., 2022a). Our current method, GANDALF, is different from GESTALT since it combines data augmentation with Graph Multiset Transformer-based multi-label informative sample selection. (7) 'Info-Max' - the Graph Infomax method using Conditional Adversarial Networks in Xiao et al. (2023). We use the graph info-max method in place of our graph multiset transformers; (8) 'IAT'- the integrated attention transformer method of Xiong et al. (2023) where we integrated the attention transformer (IAT) instead of our graph attention transformer.

5. Results and discussion

5.1. Comparative results for CheXpert dataset

For each baseline method, we measured the Area Under the Curve (AUC) for every 10% increment of training data to simulate an active learning scenario. Table 1 shows the performance of different methods at different percentages of the training data. Except for the random-based sample selection method, all the AL-based methods outperform the fully-supervised learning model (FSL), confirming the benefits of selecting samples based on their informativeness. Among the other AL methods, the uncertainty-based approach required 70% of the training data to surpass FSL. This finding aligns with other similar works , where it has been observed AL methods outperforming the FSL baseline with lower percentages of data, Mayer and Timofte (2018), Yang et al. (2017) and Sourati et al. (2019) indicating the capability of AL methods to boost the learning rate of trained models further. This behavior has not been fully explained in the literature, but we provide a possible reason below. In any given dataset there are many samples with noisy labels (inaccurate labels) or noisy images. When a classifier is fed all samples at random it encounters samples with different levels of informativeness and quality. As a result the classifier's learning rate and final performance can be negatively impacted with respect to the situation where a model only uses highly informative samples that do not have ambiguous labels, leading to improved performance and faster learning rates than a fully-supervised model (FSL).

Amongst other methods, AlphaMix and GAL are more competitive and show similar results. LEMAL and CVIRS perform slightly better than a vanilla uncertainty approach since they are based on uncertainty calculation. Our previously proposed method, GESTALT, does better than these methods, while LADA is slightly worse than GESTALT. We

Table 2
Ranking Loss (lower is better) and label ranking average precision (LRAP, higher is better) values for different methods in the multi-label $K = 5$ setting.

	Baseline methods			GANDALF variants		
	DenseNet-121	Pham (Pham et al., 2020)	GESTALT (Mahapatra et al., 2022a)	GANDALF	GANDALF $_{GAT}$	GANDALF $_{GESTALT}$
Rank-Loss ↓	0.21	0.18	0.12	0.08	0.10	0.11
LRAP ↑	0.8786	0.8934	0.9211	0.9381	0.9322	0.9288

show results for two versions of GANDALF — GANDALF$_{DT}$, when using saliency maps obtained with the Deep Taylor method, and GANDALF$_{GC}$, when using GradCAM saliency maps. Of the two, GANDALF$_{DT}$ shows better results and we refer to it as GANDALF in subsequent discussions.

As evidenced by the results, GANDALF yielded significant improvements over GESTALT and LADA by integrating data augmentation with multi-label AL. GESTALT outperforms FSL at 45% of labeled data, whereas GANDALF outperforms FSL at 37% of training data. This proves that the proposed approach requires significantly less labeled data to attain better performance. Although LADA combines data augmentation with active learning, it does not consider the multi-label scenario. The improved performance of GANDALF is attributed to the fact that: (a) it uses a better method to select most informative samples based on multi-label interactions, and (2) it leverages data augmentation to generate more informative synthetic samples based on multi-label interactions. We demonstrate this through a series of ablation experiments, presented below.

5.2. Ablation studies

We attribute the improved performance of GANDALF to the following factors: (1) integration of data augmentation with multi-label active learning; (2) use of graph attention transformers that improve the representation capacity of graphs and do a better job than GESTALT in identifying informative samples in a multi-label setting. We conducted ablation studies on GANDALF$_{DT}$ to quantify their individual contributions. The results are summarized in Table 1.

In the first set of experiments, we remove the data augmentation part from the pipeline and use only the initially identified informative samples and use very basic augmentation strategies like rotation, translation, and scaling instead of our proposed informative augmentation. We refer to this method as GANDALF$_{con-DA}$ (i.e., GANDALF$_{DT}$ using conventional data augmentation). This setting can be seen as GESTALT with a different method to identify multi-label informative samples. The fact that GANDALF$_{con-DA}$ does better than GESTALT shows the merit of our newly proposed method for multi-label sample selection.

A second variant of GANDALF$_{DT}$ consists of using GESTALT for sample informativeness, and combining it with the informative augmentation of GANDALF, which we term as GANDALF$_{GESTALT}$. This method does better than GESTALT due to the added informative augmentation but fares worse than GANDALF. We argue this is due to the use of graph transformer networks, doing a better job of learning the global dependencies, and being more accurate in identifying informative samples.

The final AUC values (i.e., at 100% data) were derived from an average of 10 runs to reduce stochasticity effects, and the statistical significance concerning GANDALF's results was calculated using a paired $t-$test, with $p-$values shown in Table 1.

5.3. Importance of graph multi set pooling

To quantify the importance of graph multi-set pooling we perform a set of experiments where we replace the multi-set pooling with conventional pooling and report the results in Table 1 as GANDALF$_{pooling}$. We observe a significant drop in performance compared to the original method (GANDALF$_{DT}$). This clearly shows that the multiset pooling has a significant role in the improved performance of our method as it leads to better feature learning. In further ablation studies we completely remove the graph attention, and self attention, and show the results in Table 1 as GANDALF$_{no-GAT}$, and GANDALF$_{no-SA}$, respectively. The individual results clearly show the importance of each of the components of the graph pooling stage.

We also investigate the role of the parameters k and n. The parameter n is the number of input nodes, which corresponds to the number of labels for the input image. In our experiments we use $n = 5$ since the test set of CheXpert dataset evaluates on 5 diseased labels. If we decrease the value of n we observe a decrease in performance since fewer nodes result in inferior learning of features. On the other hand by increasing n we obtain improved performance since more nodes (disease labels) improve the feature learning ability of the network. This comes at increased computation cost. For the NIH dataset using $n = 14$ gives an AUC of almost 0.95 (Fig. 6). This drops to AUC = 0.92 when using $n = 5$, but takes 15% less time in computing the multilabel informativeness score. This illustrates the tradeoff between method efficiency and accuracy.

The parameter k denotes the dimensions of the Graph Multiset Pooling stage. In our experiments we set $k = 3$ (k_{int}) for the intermediate layers while $k = 1$ for the final layer (k_{fin}). We observe that $k_{int} = 3$ is the optimal value. For $k_{int} = 4$ there is no significant change in performance but the computation cost increased by 5%. For $k_{int} = 2$, although the computation cost reduces by 8% the performance drops by 10%. Hence $k_{int} = 3$ is the optimal value for intermediate layers found for this dataset. We set $k_{fin} = 1$ for easy interpretation of the final value of ML_{Info}. In such a situation $k_{fin} = 2$ does not make much sense.

5.4. Influence of multiple labels on learning

We analyzed the performance levels of different benchmarked models trained with samples having co-occurring disease labels. Since there are relatively few samples with a large number of co-occurring labels, we tested performance for different number, K, of co-occurring labels. As evaluation metrics, we evaluated two different metrics suggested for multi-class scenarios and available within scikit-learn (Pedregosa et al., 2011): (1) Label Ranking Average Precision (LRAP), which measures the label rankings of each sample, where ranking is based on the model's prediction scores. LRAP values are bounded [0,1], with 1 being the perfect score. (2) Ranking Loss (Tsoumakas et al., 2009), which averages over the samples the number of label pairs incorrectly ordered, with a perfect score at zero.

In Table 2 we show results for $K = 5$ and observe that GANDALF gives the best performance as per the lowest Ranking loss and highest LRAP values. There are clear improvements over the baseline DenseNet-121, and the results of GESTALT (Mahapatra et al., 2022a; Pham et al., 2020), which is the second-best ranked method for the CheXpert dataset. This demonstrates that our approach of using graph transformers identifies more informative multi-label samples because of the self-attention module of transformers.

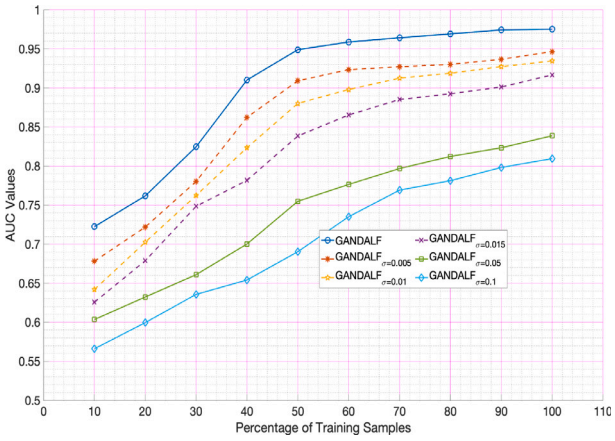

Fig. 5. AUC measures for different features for added Gaussian noise of $\mu = 0$ and different σ.

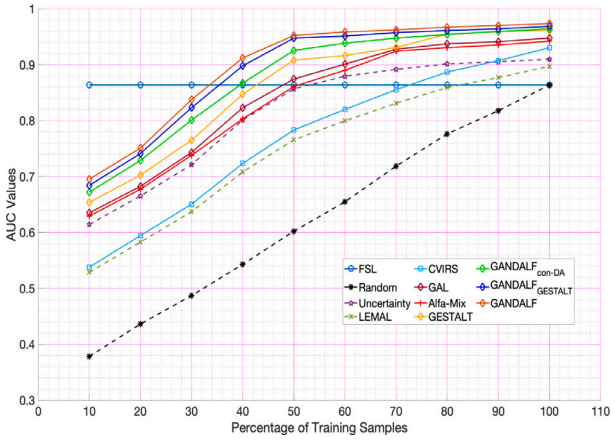

Fig. 6. Additional dataset — NIH DataSet: AUC measures at different percentage levels of training percentage for baselines (indicated with dotted lines) and the proposed GANDALF approach, including three investigated variants. As a reference, the AUC of a fully-supervised model (FSL) is also included as a horizontal line. Improved learning rates and model performance is observed for the proposed GANDALF approach.

5.5. Robustness and generalization

To test the robustness of the proposed approach, we added simulated noise of $\mu = 0$ and different $\sigma \in \{0.005, 0.01, 0.015, 0.05, 0.1\}$. Fig. 5 shows the AUC values for the baseline performance of GANDALF and different σ. The results are close to GANDALF for $\sigma = 0.005, 0.01$, but start to degrade significantly for noise levels above $\sigma = 0.01$, which we term as noise threshold. As expected, the performance of all methods degrades as more noise is added. However, the proposed GANDALF approach performs better than others and is more robust.

5.6. Performance on additional datasets

In Fig. 6 we show result obtained for the NIH dataset across different methods. The trend is similar to the results shown in Table 1 wherein different active learning methods outperform the fully supervised learning method at a lower training data percentage, and the proposed GANDALF approach outperform other methods.

Fig. 7 shows the learning plots for the MedMNIST datasets using GANDALF and other baseline methods for comparison. Results of the NIH and MedMNIST datasets clearly demonstrate the performance improvement of our proposed multi-label approach is relevant across different chest X-ray datasets and also generalizes well to the multi-class setting.

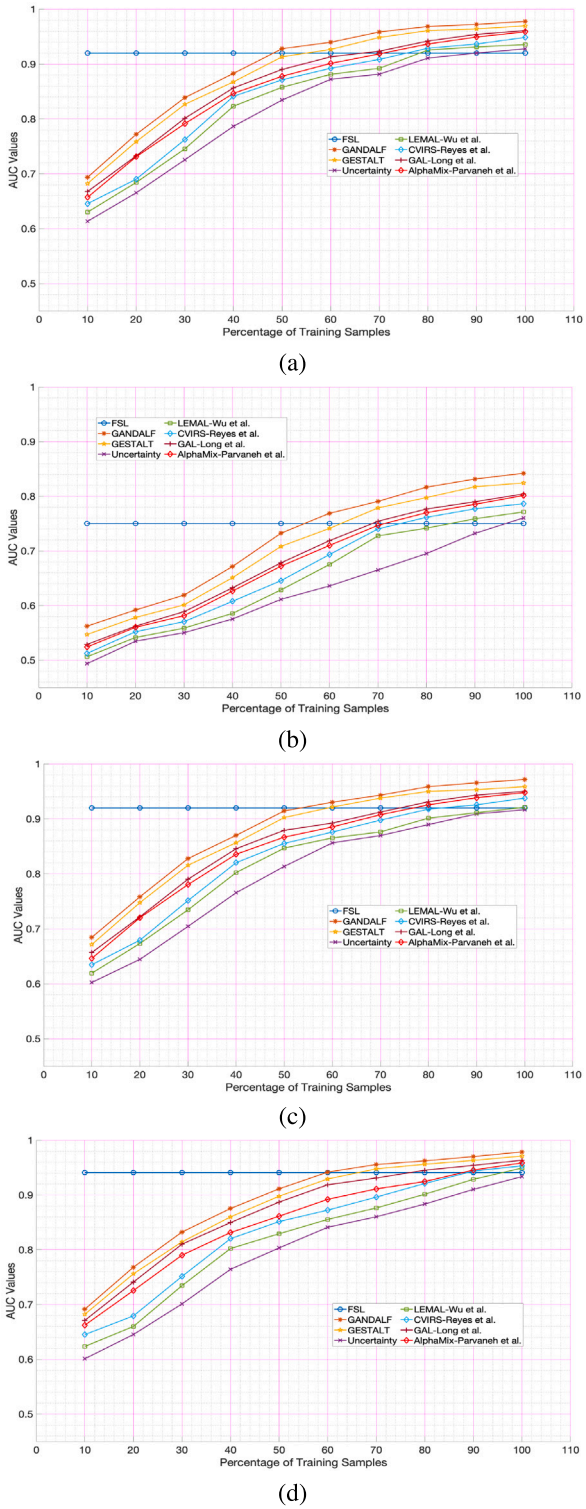

Fig. 7. MedMNIST DataSets: AUC measures at different percentage levels of training percentage for the proposed GANDALF approach and four multi-label Al learning approaches. As reference, AUC of a fully-supervised model (FSL) is also included as an horizontal line. Improved learning rates and model performance is observed for our proposed GANDALF approach which outperforms other state of the art methods. (a) Dermatology; (b) Retinal fundus images; (c) Breast dataset; (d) Tissue dataset.

Table 3
AUC values for GANDALF for different values of the parameters η_1, η_2, η_3, γ_1, γ_2.

	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
η_1	93.6	94.1	94.7	**95.3**	94.8	94.0	93.5	92.8	92.1	91.7	91.2
η_2	95.5	**96.4**	95.9	95.2	94.8	94.3	93.5	92.9	92.3	91.7	91.1
η_3	94.3	95.2	96.1	**97.5**	96.7	96.0	95.4	94.8	94.1	93.6	92.9
γ_1	96.7	**98.1**	96.8	96.0	95.6	94.9	94.2	92.7	92.1	91.8	91.2
γ_2	96.8	**98.2**	96.9	96.2	95.7	94.8	94.1	92.5	91.8	91.3	90.6

5.7. Hyperparameter settings

For hyperparameter selection we adopt the following steps to set the value of η_1, η_2, η_3. For η_1 we varied the values from $[0, 1]$ in steps of 0.05, keeping $\eta_2 = 0.05, \eta_3 = 0.35$. The best results were obtained for $\eta_1 = 0.3$, which was our final value. Following similar steps, we set $\eta_2 = 0.1$, while keeping $\eta_1 = 0.3, \eta_3 = 0.35$. Thereafter we fix $\eta_1 = 0.3, \eta_2 = 0.1$, and vary the values for η_3 and get the best results were obtained for $\eta_3 = 0.3$. While obtaining the optimal values of η_1, η_2, η_3 we keep fixed $\gamma_1 = 0.2$ and $\gamma_2 = 0.2$. After setting the values of η_1, η_2, η_3 we vary γ_1, γ_2 and obtain the best values for $\gamma_1 = 0.1$ and $\gamma_2 = 0.1$. The sensitivity of the different parameters is shown in Table 3.

The threshold η_1 is used to ensure that the probability values of the generated image do not change significantly as to make it uninformative. For example, if p^k_{fn} is close to 0.9 then the generated image is not very informative as the classifier is very confident about the prediction. In such a case the score function $Score_{label}$ is assigned a negative value of 0.1 (γ_1). This ensures that this particular sample's informativeness is reduced and is given less importance in selecting informative synthesized samples. We observe that large high negative values for γ_1 set a disproportionate importance to the probability score and will unfairly reduce the score of the sample despite a high value for $Score_{red}$.

The threshold parameters η_2, η_3 control the degree of redundancy that may be allowed for the generated images. We want that the generated images should have a minimum degree of novelty which is controlled by $\eta_2 = 0.05$. Quantitatively, this may be interpreted as the multilabel informativeness score changes by at least 5%. On the other hand a large change of the informativeness score indicates major distortions to the image, which can be attributed to an 'outlier' image. Hence the upper threshold $\eta_3 = 0.25$ indicates a limit change of 25% of the image's informativeness score. This ensures that the transformed images are not too different from the base image. The penalty γ_2's optimal value is 0.1 since large high values give disproportionate importance to the redundancy score.

5.8. Importance of score values

We investigate the importance of each of the scoring terms in Eq. (12). Table 1 show the performance measures when using only Score_{red} (GANDALF_{red}) and only Score_{label} (GANDALF_{label}). The results clearly show that discarding either of the terms degrades the performance. Excluding Score_{red} leads to worse performance than excluding Score_{label}. This may be explained by the fact that the redundancy score uses the multilabel informativeness score ML_{info} to determine informative samples.

In another set of experiments we take 50% of the labeled training set for which we get an AUC value of 94.90 as shown in Table 1. Thereafter we vary the values of λ_1 and λ_2 between $[0.2, 1.5]$ in steps of 0.05. The AUC values are shown as a heat map in Fig. 8. We observe that the highest AUC is obtained when $0.9 \leq \lambda_1 \leq 1.1$ and $0.8 \leq \lambda_2 \leq 1.05$. This makes $\lambda_1 = \lambda_2 = 1$ a good choice for robustness across different datasets.

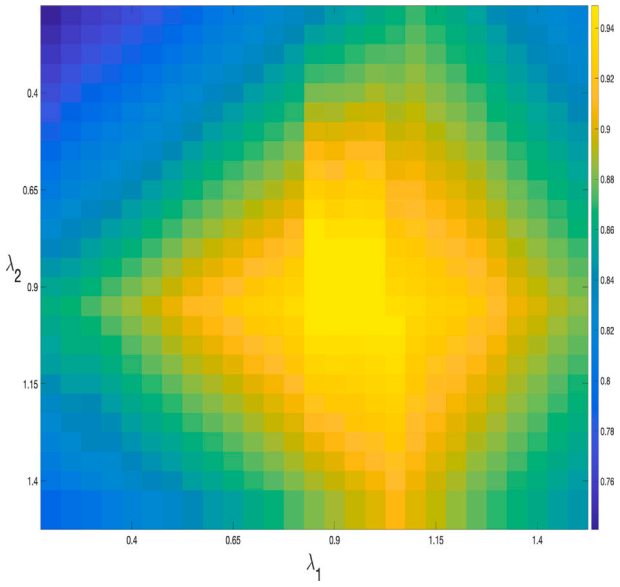

Fig. 8. Heat map showing AUC values at 50% labeled training data for different values of λ_1, λ_2.

5.9. Computation time

For a training dataset of 100,000 images of size 320 × 320, the training time for different methods on an NVIDIA Titan X GPU having 12 GB RAM is summarized in Table 4. Compared to GESTALT, our proposed GANDALF method has an 8% higher training time. This is due to the extra computations involved in the informative augmentation and graph transformer attention which is an integral part of the process. However, the resulting performance improvement justifies the added complexity of our method. GESTLAT also shows a 12% higher training time than the other conventional approaches due to the additional computations and graph construction involved. The inference time for a single image is also summarized in Table 4 for different methods.

5.10. Image visualizations

In this section we show different visualizations that highlight the role of saliency maps. To generate interpretability saliency maps we use the iNNvestigate library (Alber et al., 2019),[2] which implements several known interpretability approaches. We employ Deep Taylor, a known interpretability approach to generate saliency maps, due to its ability to highlight informative regions while yielding minimal importance to other regions. Deep Taylor operates similarly as other interpretability approaches by decomposing back-propagation gradients, of the studied model, into layer-wise relevance maps of individual cell activations, as a function of a queried input sample and class label (e.g. disease class). Each neuron of a deep network is viewed as a function that can be expanded and decomposed on its input variables. The decompositions of multiple neurons are then aggregated or propagated backwards, resulting in a saliency map (Montavon et al., 2017).

Fig. 9 shows the saliency map visualizations using Deep Taylor and Grad-CAM for two images. The image in the top row has similar regions highlighted by both approaches. However, for the bottom row image the localized regions are quite different. Deep Taylor method highlights regions near the lung but the Grad-CAM method tends to localize an area beyond the lung region where there is no anatomy of interest. This justifies our choice of using Deep Taylor approach for generating saliency maps. Overall, in this study we selected DeepTaylor because of its greater accuracy in highlighting important regions.

[2] https://github.com/albermax/innvestigate.

Table 4
Training and inference time for different methods.

Training phase — Time in hours										
DenseNet-121	Random	Unc	GESTALT (Mahapatra et al., 2022a)	GAL (Long et al., 2008)	LEMAL (Wu et al., 2014)	CVIRS (Reyes et al., 2018)	AlfaMix (Parvaneh et al., 2022)	GANDALF	GANDALF $_{GAT}$	GANDALF $_{GESTALT}$
18(0.67T)	18.5(0.69T)	19.5(0.72T)	25(0.93T)	23.5(0.87T)	20(0.74T)	21.5(0.8T)	24(0.89T)	27(T)	26.5(0.98T)	26.1(0.97T)
Test/Inference phase — Time in seconds										
0.18	0.19	0.2	0.32	0.28	0.22	0.24	0.3	0.4	0.4	0.4

Table 5
Results for low data scenarios: AUC values for different baselines and proposed approach along with ablation studies. The p−values are with respect to GANDALF. DT: DeepTaylor; GC: GradCAM; Warm: Warm AL start; red: GANDALF with only redundancy avoidance; label: GANDALF with only label preservation; con-DA: Conventional Data Augmentation; GESTALT: GESTALT for sample informativeness, and GANDALF for informative augmentation; pooling: Conventional pooling; no-GAT: No Graph Attention; no-SA: No Self-Attention.

	0.5%	1%	1.5%	2%	2.5%	3%	p-
Random	33.3	34.1	34.9	35.7	36.6	37.1	<0.001
Entropy	39.1	39.9	40.6	41.5	42.8	44.1	<0.001
Unc	42.8	43.7	44.9	45.8	46.9	48.1	<0.001
LEMAL (Wu et al., 2014)	44.7	46.1	47.6	48.7	50.1	51.2	<0.001
CVIRS (Reyes et al., 2018)	46.4	47.7	48.9	50.3	51.5	52.3	0.001
AlphaMix (Parvaneh et al., 2022)	48.7	49.8	50.3	51.5	52.6	53.8	0.005
GAL (Long et al., 2008)	48.9	49.9	51.1	52.4	53.6	54.8	0.007
LADA (Kim et al., 2021)	49.6	50.9	52.0	53.2	54.6	55.8	0.001
GESTALT (Mahapatra et al., 2022a)	50.5	51.6	52.9	53.9	55.0	56.4	0.02
Info-Max (Xiao et al., 2023)	50.1	51.2	52.4	53.5	54.6	55.8	0.02
IAT (Xiong et al., 2023)	50.7	51.5	52.6	53.7	54.8	56.1	0.02
GANDALF $_{DT}$	52.8	54.2	55.6	57.2	58.3	59.9	–
GANDALF $_{DT-NoWeight}$	49.9	51.2	52.3	53.4	54.9	56.0	0.001
GANDALF $_{GC}$	52.5	53.9	55.3	56.9	58.0	59.4	0.07
GANDALF $_{GC-NoWeight}$	49.5	50.9	52.0	53.1	54.5	55.5	0.001
GANDALF $_{Warm}$	54.1	55.8	57.1	58.7	59.8	61.3	0.041
GANDALF $_{Warm-NoWeight}$	52.1	53.6	55.0	56.5	57.7	59.1	0.001
Ablation studies							
GANDALF $_{red}$	48.4	49.5	50.0	51.1	52.2	53.5	0.03
GANDALF $_{label}$	46.1	47.4	48.5	49.9	51.2	52.0	0.03
GANDALF $_{con-DA}$	52.1	53.6	55.0	56.5	57.6	59.1	0.034
GANDALF $_{GESTALT}$	52.4	53.9	55.3	56.9	57.8	59.5	0.032
GANDALF $_{pooling}$	44.3	45.7	47.2	48.3	49.8	50.8	0.032
GANDALF $_{no-GAT}$	42.9	43.8	45.1	45.9	47.0	48.2	0.03
GANDALF $_{no-MHSA}$	42.6	43.5	44.8	45.7	46.8	48.0	0.011

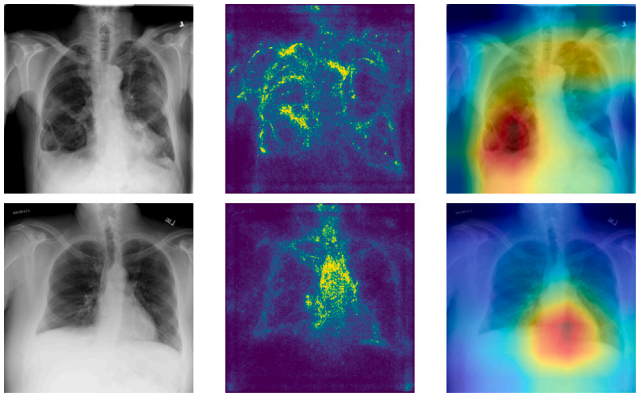

Fig. 9. Comparative visualization of GradCAM and Deep Taylor models. (a) original image; Saliency maps using (b) Deep Taylor method; (c) Grad-CAM method. Especially for the bottom row image, the Deep Taylor method gives a more accurate localization of informative regions than Grad-CAM.

5.10.1. Visualization of informative augmentation

In Fig. 10(a) we show base informative images with the diseased region annotated by a radiologist with over 15 years of experience and Fig. 10(b) shows the corresponding saliency maps. Subsequent columns show the saliency maps of different informative images generated by our method using our sampling approach. We show the saliency maps as changes in the informative image content can be easily discerned than looking at the original images. It is quite obvious that the saliency maps of generated images sufficiently differ from the base image to indicate informativeness and at the same time preserve the content of the base image.

5.10.2. Graph visualization

Fig. 11 shows examples of informative and uninformative images from the same batch. Saliency maps of each class is generated for the images. In this example case we show an example for 5 classes with same number of saliency maps. The resulting graph is with the edge weights indicated by the respective values. After one stage of applying GAT the map is reduced to a graph with 3 nodes and then finally a single value ML_{info} is obtained that quantifies the graph's informativeness. We note that the informative image yields ML_{info} = 67, whereas the uninformative images yields a lower value of ML_{info} = 32. The graph weights have been scaled to a range of $1-25$, while the ML_{info} is scaled to a range of $1-100$ for every batch of images.

5.11. Results for low data scenarios

We conduct experiments to test our method's efficacy in low-label scenarios by varying the datasets from 0.5% to 3% an d summarizing the results in Table 5. The results show that although our method obtains better results than competing methods, it still requires more data to reach the optimum performance. In order to perform well on

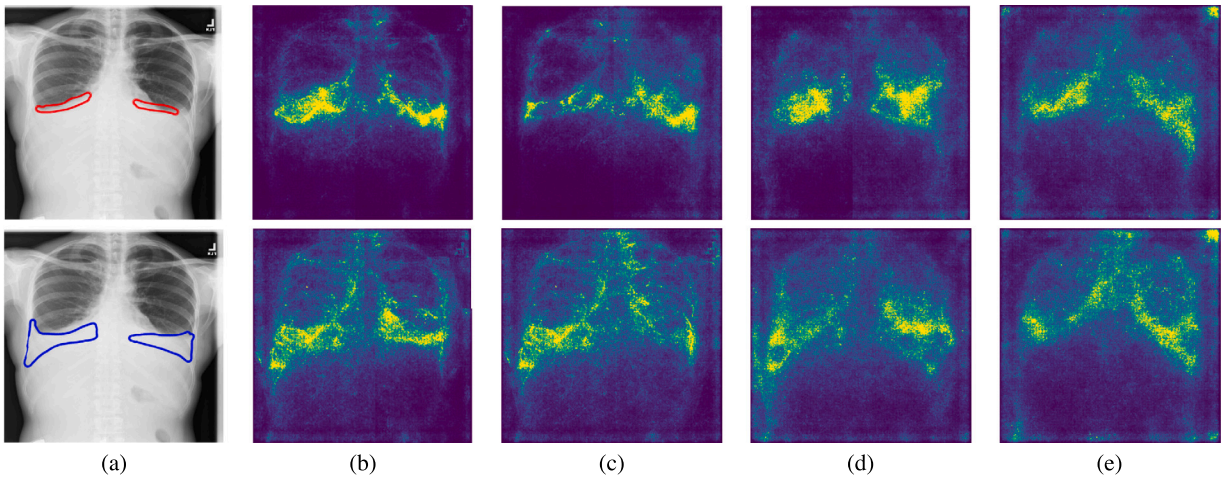

Fig. 10. (a) Base informative image with expert-annotated outlines of diagnosed conditions. Saliency maps for different methods: (b) base image; (c)–(e) Different informative images generated from the base image. Top row shows Pleural Effusion image and bottom row shows Atelactasis.

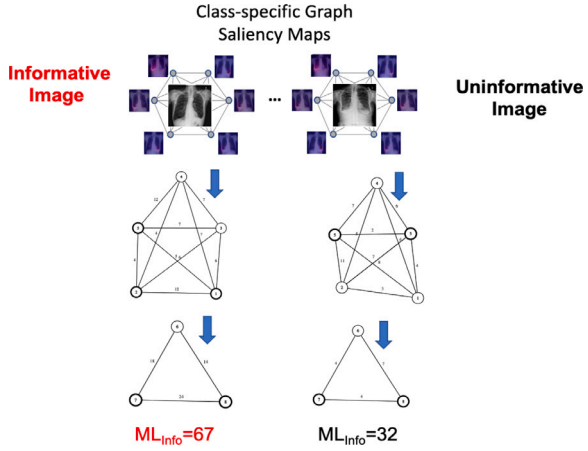

Fig. 11. Illustration of graph construction of informative and non-informative images.

very low data scenarios the method needs to be modified significantly with an entirely new approach.

6. Conclusions

In this paper, we present a novel approach for multi-label active learning that combines active learning with data augmentation. We refer to the proposed method as "GANDALF" (**G**raph-based Tr**AN**sformer and **D**ata **A**ugmentation Active **L**earning **F**ramework). The key motivation in combining active learning with data augmentation is to leverage their mutually complementary strengths and ensure that the data augmentation step also generates informative samples from an informative base sample.

Unlike most current works that deal with multi-class AL, we focus on multi-label AL, where a given sample can have more than one disease label. To learn the interaction between different disease labels, we use graphs to quantify the informativeness of each sample. Going beyond (Mahapatra et al., 2022a), which proposes simple aggregation strategies such as mean and the sum of node weights, we use graph attention transformers with graph neural networks to learn more discriminative graph aggregations.

Complementing the improved graph aggregation strategy is the informative augmentation step that takes a base informative image, generates augmented versions, and calculates a score based on label preservation and informativeness of the augmented images. The overall informativeness of the augmented samples is the sum of the two scores, and the most informative samples are added to the training set for further training.

Our proposed GANDALF method yields better results than our previous method, GESTALT (Mahapatra et al., 2022a), and other competing methods. Subsequent ablation studies also highlight the importance of the graph attention transformers and the informative augmentation step in the overall performance of GANDALF.

In future work, we aim to test our model on other multi-label medical image datasets. We also aim to test its robustness and generalizability to different classification architectures and segmentation methods. We also anticipate that transformers will play a greater role in graph-based interpretability and active learning tasks. Hence, our future focus will be on exploiting the properties of graph attention transformers to learn more powerful graph representations on multi-omics dataset combining imaging and non-imaging information. Current active learning methods are typically tested with initial dataset sizes in the order of 10% of the total available data, as also performed in this study. An interesting avenue of further research includes the combination of new approaches proposed to work under more extreme low-label regimes, such as Taher et al. (2023) and Chen et al. (2023a,b) for active learning scenarios.

CRediT authorship contribution statement

Dwarikanath Mahapatra: Conceptualization, Methodology, Validation, Writing – original draft, Writing – review & editing. **Behzad Bozorgtabar:** Writing – original draft, Writing – review & editing. **Zongyuan Ge:** Writing – original draft, Writing – review & editing. **Mauricio Reyes:** Conceptualization, Formal analysis, Methodology, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

NA.

Acknowledgments

This work was supported by the Swiss National Foundation grant number 212939, and Innosuisse grant number 31274.1.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., Makarenkov, V., Nahavandi, S., 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Inf. Fusion 76, 243–297. http://dx.doi.org/10.1016/j.inffus.2021.05.008.

Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A., 2020. Dataset of breast ultrasound images. Data Brief 28, http://dx.doi.org/10.1016/J.Dib.2019.104863.

Alber, M., Lapuschkin, S., Seegerer, P., Hagele, M., Schutt, K.T., Montavon, G., Samek, W., Muller, K.-R., Dahne, S., Kindermans, P.-J., 2019. iNNvestigate neural networks. J. Mach. Learn. Res. 20 (93), 1–8.

Ash, J.T., Zhang, C., Krishnamurthy, A., Langford, J., Agarwal, A., 2020. Deep batch active learning by diverse, uncertain gradient lower bounds.

Ba, L.J., Kiros, J.R., Hinton., G.E., 2016. Layer normalization. arXiv preprint, arXiv:1607.06450.

Baek, J., Kang, M., Hwang, S.J., 2021. Accurate learning of graph representations with graph multiset pooling. In: International Conference on Learning Representations.

Bianchi, F.M., Grattarola, D., Alippi, C., 2019. Spectral clustering with graph neural networks for graph pooling. arXiv preprint arXiv:1907.00481.

Bozorgtabar, B., Mahapatra, D., von Teng, H., Pollinger, A., Ebner, L., Thiran, J.-P., Reyes, M., 2019. Informative sample generation using class aware generative adversarial networks for classification of chest Xrays. Comput. Vis. Image Underst. 184, 57–65.

Budd, S., Robinson, E.C., Kainz, B., 2021. A survey on active learning and human-in-the-loop deep learning for medical image analysis. Med. Image Anal. 71, 102062. http://dx.doi.org/10.1016/J.MEDIA.2021.102062.

Chen, L., Bai, Y., Huang, S., Lu, Y., Wen, B., Yuille, A., Zhou, Z., 2023a. Making your first choice: To address cold start problem in medical active learning. In: Medical Imaging with Deep Learning. URL: https://openreview.net/forum?id=5iSBMWm3ln.

Chen, Y., Liu, L., Li, J., Jiang, H., Ding, C., Zhou, Z., 2023b. MetaLR: Meta-tuning of learning rates for transfer learning in medical imaging. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. Springer Nature Switzerland, Cham, pp. 706–716.

Codella, N., et al., 2019. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368.

Gal, Y., Islam, R., Ghahramani, Z., 2017. Deep Bayesian active learning with image data. In: Proc. International Conference on Machine Learning.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Proc. NIPS. pp. 2672–2680.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks. In: International Conference on Machine Learning. PMLR, pp. 1321–1330.

Huang, G., Liu, Z., van der Maaten, L., Weinberger, K., 2016. Densely connected convolutional networks. https://arxiv.org/abs/1608.06993.

Irvin, J., Rajpurkar, P., et al., 2019. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. arXiv preprint arXiv:1901.07031.

Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K., 2015. Spatial transformer networks. In: NIPS.

Jungo, A., Balsiger, F., Reyes, M., 2020. Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. Front. Neurosci. 14, 282.

Kendall, A., Gal, Y., 2017. What uncertainties do we need in Bayesian deep learning for computer vision? In: Advances in Neural Information Processing Systems.

Kim, Y.-Y., Song, K., Jang, J., Moon, I.-c., 2021. LADA: Look-ahead data acquisition via augmentation for deep active learning. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (Eds.), Advances in Neural Information Processing Systems, Vol. 34. Curran Associates, Inc., pp. 22919–22930.

Kingma, D., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kingma, D.P., Welling, M., 2013. Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114.

Lai, Z., Wang, C., Oliveira, L.C., Dugger, B.N., Cheung, S.-C., Chuah, C.-N., 2021. Joint semi-supervised and active learning for segmentation of gigapixel pathology images with cost-effective labeling. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 591–600.

Lee, J., Lee, I., Kang, J., 2019a. Self-attention graph pooling. In: ICML. pp. 3734–3743.

Lee, J., Lee, Y., Kim, J., Kosiorek, A.R., Choi, S., Teh., Y.W., 2019b. Set transformer: A framework for attention-based permutation-invariant neural networks. In: ICML. pp. 3744–3753.

Li, X., Guo, Y., 2013. Active learning with multi-label SVM classification. In: IJCAI '13. pp. 1479–1485.

Lian, J., Long, Y., Huang, F., Ng, K.-S., Lee, F.M.Y., Lam, D.C.L., Fang, B.X.L., Dou, Q., Vardhanabhuti, V., 2022. Imaging-based deep graph neural networks for survival analysis in early stage lung cancer using CT: A multicenter study. Front. Oncol. 12, http://dx.doi.org/10.3389/fonc.2022.868186.

Liu, R., Wang, X., et al., 2022. DeepDRiD: Diabetic retinopathy—Grading and image quality estimation challenge. Patterns 3 (6), 100512.

Long, J., Yin, J., Zhao, W., Zhu, E., 2008. Graph-based active learning based on label propagation. In: Torra, V., Narukawa, Y. (Eds.), Modeling Decisions for Artificial Intelligence. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 179–190.

Mahapatra, D., Bozorgtabar, B., Thiran, J.-P., Reyes, M., 2018. Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network. In: Proc. MICCAI. pp. 580–588.

Mahapatra, D., Buhmann, J., 2016. Visual saliency-based active learning for prostate magnetic resonance imaging segmentation. SPIE J. Med. Imaging 3 (1), 014003.

Mahapatra, D., Poellinger, A., Reyes, M., 2022a. Graph node based interpretability guided sample selection for active learning. IEEE Trans. Med. Imaging 1. http://dx.doi.org/10.1109/TMI.2022.3215017.

Mahapatra, D., Poellinger, A., Reyes, M., 2022b. Interpretability-guided inductive bias for deep learning based medical image. Med. Image Anal. 81, 102551.

Mahapatra, D., Poellinger, A., Shao, L., Reyes, M., 2021. Interpretability-driven sample selection using self supervised learning for disease classification and segmentation. IEEE Trans. Med. Imaging 40 (10), 2548–2562.

Mayer, C., Timofte, R., 2018. Adversarial sampling for active learning. arXiv preprint arXiv:1808.06671.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Müller, K.-R., 2017. Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recognit. 65 (1), 211–222.

Nath, V., Yang, D., Roth, H.R., Xu, D., 2022. Warm start active learning with proxy labels and selection via semi-supervised fine-tuning. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. pp. 297–308.

Parvaneh, A., Abbasnejad, E., Teney, D., Haffari, G.R., van den Hengel, A., Shi, J.Q., 2022. Active learning by feature mixing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12237–12246.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Perez, L., Wang, J., 2017. The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621.

Pham, H.H., Le, T.T., Tran, D.Q., Ngo, D.T., Nguyen, H.Q., 2020. Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels. arXiv preprint arXiv:1911.06475.

Reyes, M., Meier, R., Pereira, S., Silva, C.A., Dahlweid, F.-M., Tengg-Kobligk, H.v., Summers, R.M., Wiest, R., 2020. On the interpretability of artificial intelligence in radiology: Challenges and opportunities. Radiol.: Artif. Intell. 2 (3), e190043. http://dx.doi.org/10.1148/ryai.2020190043, arXiv:https://doi.org/10.1148/ryai.2020190043.

Reyes, O., Morell, C., Ventura, S., 2018. Effective active learning strategy for multi-label learning. Neurocomputing 273, 494–508.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proc. ICCV. pp. 618–626.

Silva, W., Poellinger, A., Cardoso, J.S., Reyes, M., 2020. Interpretability-guided content-based medical image retrieval. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 305–314.

Siméoni, O., Budnik, M., Avrithis, Y., Gravier, G., 2021. Rethinking deep active learning: Using unlabeled data at model training. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 1220–1227.

Sourati, J., Gholipour, A., Dy, J.G., Tomas-Fernandez, X., Kurugol, S., Warfield, S.K., 2019. Intelligent labeling based on fisher information for medical image segmentation using deep learning. IEEE Trans. Med. Imaging 38 (11), 2642–2653.

Taher, M.R.H., Gotway, M.B., Liang, J., 2023. Towards foundation models learned from anatomy in medical imaging via self-supervision. arXiv:2309.15358.

Tran, T., Do, T.-T., Reid, I., Carneiro, G., 2019. Bayesian generative active deep learning. arXiv preprint arXiv:1904.11643.

Tschandl, P., Rosendahl, C., Kittler, H., 2018. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data 5.

Tsoumakas, G., Katakis, I., Vlahavas, I., 2009. Mining multi-label data. In: Data Mining and Knowledge Discovery Handbook. Springer, pp. 667–685.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. In: NeurIPS. pp. 5998–6008.

Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y., 2018. Graph attention networks.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R., 2017b. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proc. CVPR.

Wang, J., Wen, S., Chen, K., Yu, J., Zhou, X., Gao, P., Li, C., Xie, G., 2020. Semi-supervised active learning for instance segmentation via scoring predictions.

Wang, K., Zhang, D., Li, Y., Zhang, R., Lin., L., 2017a. Cost-effective active learning for deep image classification. IEEE Trans. Circuits Syst. Video Technol. 27 (12), 2591–2600.

Weisfeiler, B.Y., Leman, A.A., 1968. A reduction of a graph to a canonical form and an algebra arising during this reduction. Nauchno-Tech. Informatsia.

Wu, J., Sheng, V.S., Zhang, J., Li, H., Dadakova, T., Swisher, C.L., Cui, Z., Zhao, P., 2020. Multi-label active learning algorithms for image classification: Overview and future promise. ACM Comput. Surv. 53 (2).

Wu, J., Sheng, V.S., Zhang, J., Zhao, P., Cui, Z., 2014. Multi-label active learning for image classification. In: 2014 IEEE International Conference on Image Processing (ICIP). pp. 5227–5231. http://dx.doi.org/10.1109/ICIP.2014.7026058.

Xiao, J., Dai, Q., Xie, X., Dou, Q., Kwok, K.-W., Lam, J., 2023. Domain adaptive graph infomax via conditional adversarial networks. IEEE Trans. Netw. Sci. Eng. 10 (1), 35–52.

Xiong, C., Chen, H., Sung, J., King, I., 2023. Diagnose like a pathologist: Transformer-enabled hierarchical attention-guided multiple instance learning for whole slide image classification.

Xu, K., Hu, W., Leskovec, J., Jegelka, S., 2019. How powerful are graph neural networks? In: International Conference on Learning Representations.

Yang, Y., Ma, Z., Nie, F., et al., 2015. Multi-class active learning by uncertainty sampling with diversity maximization. Int. J. Comput. Vis. 113, 113–127.

Yang, J., Shi, R., Ni, B., 2021. MedMNIST classification decathlon: A lightweight AutoML benchmark for medical image analysis. In: ISBI.

Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D., 2017. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In: Proc. MICCAI. pp. 399–407.

Ying, Z., You, J., Morris, C., Ren, X., Hamilton, W.L., Leskovec, J., 2018. Hierarchical graph representation learning with differentiable pooling. In: NeurIPS. pp. 4805–4815.

Yuan, M., Lin, H.-T., Boyd-Graber, J., 2020. Cold-start active learning through self-supervised language modeling.

Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2017. Mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412.

Zhang, M., Cui, Z., Neumann, M., Chen, Y., 2018. An end-to-end deep learning architecture for graph classification. In: AAAI. pp. 4438–4445.

Zheng, H., Yang, L., Chen, J., Han, J., Zhang, Y., Liang, P., Zhao, Z., Wang, C., Chen, D.Z., 2019. Biomedical image segmentation via representative annotation. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 5901–5908.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization. In: Proc. CVPR. pp. 2921–2929.

Zhu, J.-J., Bento, J., 2017. Generative adversarial active learning. arXiv preprint arXiv:1702.07956.