Graph Node Based Interpretability Guided Sample Selection for Active Learning

Dwarikanath Mahapatra*, Alexander Poellinger, Mauricio Reyes

***Abstract*— While supervised learning techniques have demonstrated state-of-the-art performance in many medical image analysis tasks, the role of sample selection is important. Selecting the most informative samples contributes to the system attaining optimum performance with minimum labeled samples, which translates to fewer expert interventions and cost. Active Learning (AL) methods for informative sample selection are effective in boosting performance of computer aided diagnosis systems when limited labels are available. Conventional approaches to AL have mostly focused on the single label setting where a sample has only one disease label from the set of possible labels. These approaches do not perform optimally in the multi-label setting where a sample can have multiple disease labels (e.g. in chest X-ray images). In this paper we propose a novel sample selection approach based on graph analysis to identify informative samples in a multi-label setting. For every analyzed sample, each class label is denoted as a separate node of a graph. Building on findings from interpretability of deep learning models, edge interactions in this graph characterize similarity between corresponding interpretability saliency map model encodings. We explore different types of graph aggregation to identify informative samples for active learning. We apply our method to public chest X-ray and medical image datasets, and report improved results over state-of-the-art AL techniques in terms of model performance, learning rates, and robustness.**

***Index Terms*— Interpretability, Graphs Multi-label, Sample Selection Lung disease classification.**

I. Introduction

Although supervised Deep Learning (DL) approaches trained on large datasets have shown state-of-the-art performance on medical image analysis tasks, obtaining large labeled datasets is challenging due to the high levels of required data curation time and expertise. In this sense, Active Learning (AL) methods enable a progressive learning that is suited for clinical setups where improvements over time, based on expert-feedback, is desired. In AL, given a deployed model and a pool of test (i.e. unlabeled) samples, the most informative ones are selected, expert-queried for labels, and used to further train the model. In this regard, central to AL is the identification of informative samples that enable a model to obtain high performance with minimal labeled samples (i.e. high learning rates). Such characteristic is particularly important when AL-based technologies are required to swiftly adapt to potential changes of the imaging protocol, vendor type, model, etc.

Current AL-based methods have been developed for a single-label setting where a given sample is assumed to have only one class label. However, multi-label settings are prevalent, especially in chest X-ray images, where an input image can present multiple conditions. The multi-label setting poses additional challenges since an expert has to analyze the image for the presence of multiple pathologies and account for their interactions. In this multi-label setting, informative sample selection 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). Hence it is imperative to develop informative sample selection techniques for multi-label settings.

In this paper we propose an AL-based approach for informative sample selection in multi-label settings. Using multi-label classification of chest X-rays as use case, we describe an approach to model intra-sample class label similarities through a graph model, where graph nodes describe class-specific latent representations of corresponding interpretability saliency map, and edges in this graph describe their similarity. In order to identify informative samples for active learning in this multi-label setting, we describe and report results using three different types of graph aggregation strategies to combine the information contained on each sample's graph directly into a sample selection ranking metric.

In the next section we summarize the prior work on active sample selection, and further motivate our proposition, rationale and hypothesis for a graph node based interpretability-guided sample selection for active learning.

II. Prior Work on Active Sample Selection

In AL a deployed model progressively improves over time as new training samples are made available through expert annotation. This enables a progressive learning capability based on expert-feedback. Due to the time limitations in clinical settings, sample selection methodologies are fundamental to attain optimal system performance with minimal expert interactions. In deep-learning for medical image computing applications, different informative sample selection approaches have been investigated, including sample entropy [1]–[3], model uncertainty [4]–[7], Fisher information [8] , visual saliency [9]–[11] and clustering [12].

D. Mahapatra is with the Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates. (email: dwarikanath.mahapatra@inceptioniai.org)

A. Poellinger is with the Department of Diagnostic, Interventional and Pediatric Radiology, Inselspital, Bern University Hospital, Bern, Switzerland, and University of Bern, Switzerland.

M. Reyes is with the ARTORG Center for Biomedical Engineering Research, University of Bern, Bern, Switzerland

Manuscript received *********; revised **********.

Sample entropy reflects the difficulty of the model to classify a sample, with higher entropy characterizing higher sample informativeness. In [2] sample entropy along with least confidence, and margin sampling metrics are used to select informative samples while [13] uses Generative Adversarial Networks (GAN) to synthesizes samples close to the decision boundary, which are then annotated by human experts. [14] use a GAN model to generate high entropy samples, which are used as proxy to find most similar samples, from a pool of real sample candidates, to be annotated by experts.

Uncertainty-based methods identify most informative samples as those for which a model is most uncertain. [7] 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 [6] has been used to estimate uncertainty of samples, and consequently select most informative ones for label annotation [4]–[6], with the approaches in [5] and [4] differing from [6] as they incorporate a conditional GAN-based data augmentation to synthesize similar samples to those selected by the uncertainty criteria, in order to further boost the learning rate of the model. The work of [6] was followed up in [15] to solve redundancy issues where many similar samples are selected. Instead of using MC dropout to yield uncertainty estimates, the authors in [16] proposed a more computationally intensive solution via model ensembling to characterize sample informativeness in a query-by-consensus scheme, where higher disagreement is associated to higher sample uncertainty.

Based on the Fisher information metric, the authors in [8] proposed to select samples based on an efficient low-dimensional approximation of the Fisher information metric targeting Convolutional Neural Networks. The approach, however, relies on a pre-selection step based on sample uncertainty estimation, and its performance hence depends on the sensitivity level of such uncertainty-based pre-selection. Other approaches include the work of [12] where sample selection is based on a representativeness approach wherein image patches are first projected into a latent space (e.g. via a Variational Autoencoder), clustered in the latent space and sorted by their representativeness using a cosine distance maximum set coverage metric, as done in [7].

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 [17]–[26].

These findings further motivate our proposition to investigate alternative approaches to select informative samples for active learning. Furthermore, as pointed out recently in their survey paper on active learning and human-in-the-loop learning [27], interpretability mechanisms are crucial for medical imaging applications where experts and AI systems interact.

In [28], we propose an interpretability-guided sample selection approach featuring state-of-the-art performance for classification and segmentation tasks, where latent representations of saliency maps from pool samples, are used by an additional machine learning classification model to rank samples by their informativeness level. Improving over this work, here we avoid the need of training a second machine learning model needed to rank samples by their informativeness level, and propose a direct approach to rank them. In addition, the proposed approach explicitly considers class distinctiveness across all class labels for each pool sample, instead of only using the information from the saliency maps calculated for the predicted class, as done in [28].

A. Multi-Label Active Learning

There are quite a few works that deal with multi-label based active learning, and a comprehensive survey can be found in [29], [30]. In [31] a Multi-Class AL approach is introduced that performs graph based label propagation in a transductive manner. However, this work is not of a multi-label setting. Wu et. al. [32] propose "Example-label"-based AL (LEMAL) where the samples are selected on the basis of maximum uncertainty across all label classes. Reyes et. al. [33] propose two uncertainty measures based on the base classifier predictions and use the inconsistency of a predicted label set to select the most informative example. [34] use correntropy calculated across all labeled and queried samples to select the most informative sample. Li et. al. [35] propose a max-margin prediction uncertainty strategy and a label cardinality inconsistency strategy to measure the unified informativeness of unlabeled instances.

The above works do not explicitly address the multi-label scenario where a sample can have multiple labels. Some works focus on the multi-class scenario where a sample has only one possible label from a set of multiple labels. Different from the above works we focus on the multi-label setting and our graph based approach learns a more accurate relationship between different labels.

In this paper we propose a novel approach for selecting informative samples in a multi-label setting. Different from prior works, here we propose to use interpretability saliency maps and graph analysis to identify most informative samples. The motivation for our proposition and hypothesis are founded on the following rationale. As a classification model is trained, class-specific features are learned to improve class distinctiveness. In state-of-the-art active learning approaches, such class distinctiveness is enforced by selecting samples through metrics that overall reflect the higher complexity of a given sample to be classified (e.g. high uncertainty, high model's loss function values, etc.). In interpretability of deep learning models, saliency maps can in general terms be seen as fingerprints of model's response to an input. These saliency maps result from back-propagated gradients, calculated from a user-specified output class label to the input. Hence, for a given sample, saliency maps for *all* potential output class labels can be calculated. We hypothesized that this set of information can be used to characterize the level of class distinctiveness of a given sample *across all* potential class labels, and therefore be used to derive a sample selection metric for active learning, where samples featuring an overall[1] low class distinctiveness

[1]hence the name *Gestalt: whole is more important than the sum of its component*

would be prioritized for active learning sample selection.

III. Contributions

In this paper we make the following contributions:

1) We propose a novel concept called Graph Nod*E* Ba*S*ed In*T*erpret*A*bility Guided Samp*L*e Selec*T*ion approach (GESTALT), where interpretability information across all potential class labels is combined to derive a novel sample selection metric, leading to improved model performance, learning rates and model generalization. In comparison to the state-of-the-art methods, the proposed GESTALT approach is designed to consider a better handling of sample selection for multi-class classification problems, and in comparison to other interpretability-based state-of-the-art method does not require additional machine learning models to perform ranking of informative samples. We report results using three different graph aggregation variants of GESTALT. The best variant includes a novel mechanism to consider prior information of the distribution of intra-sample class distinctiveness, as seen by the model during learning. We also include a second mechanism to consider the different levels of complexity in detecting diseases.
2) We demonstrate the added value of the proposed interpretability-driven active sample selection approach by means of comparison to an standard active learning (i.e. no sample selection involved), and state-of-the-art uncertainty-driven and interpretability-driven active learning approaches on two public lung-disease classification databases, and four other datasets of retinal fundus, pathology, dermatology and breast ultrasound images.

IV. Methods

A. Motivation

In a previous work [28] we presented an approach that employs interpretability saliency maps to identify the most informative image samples in a pool for active learning, outperforming current uncertainty-based sample selection methods. In [28] we also show during training of a model via active learning, its corresponding saliency maps also evolve during the training process, and the information contained in their latent representations can be used in a self supervised learning setup, to train a random forest classifier that ranks the informativeness of samples. Two limitations of this method are: 1) it involved a two-stage process where the self supervised learning stage requires a separate training step; 2) The method did not explicitly take into account the multi-label setting and solely relies on a saliency map calculated on the predicted class. In our current work we address the above shortcomings by proposing an end-to-end trainable model to identify multi-label informative samples by jointly considering the mutual influence of all potential class labels.

B. Main Components Of Proposed Method

Figure 1 shows the general workflow of our proposed method. Given a set of unlabeled test samples (i.e. sample pool), and an associated deep learning classification model (e.g. DenseNet) trained iteratively during active learning, an interpretability saliency map generator is used to produce saliency maps (or the class activation maps) for each potential class label. The saliency maps are inputted to the GESTALT module, which ranks the most informative samples. Selected informative samples are then queried for labels, and added to retrain and update the classification model. We describe each component in detail and in relation to the clinical problem of automating lung disease classification, as well as baseline methods used to benchmark the proposed approach.

1) Classification model: Although the classification model is not a contribution of our method we include it to facilitate the presentations and descriptions of the data workflow. Any robust classification model can be used as the approach is not restricted to particular architectures. For lung disease classification from X-ray images, we experimented with 3 different models namely, DenseNet-121 [36],ResNet-50 [37] and VGG16 [38], and found the DenseNet-121 architecture to perform the best. Below, we denote as M, the DenseNet-121 model, and point the reader to section V-E, for further implementation details of the trained model.

2) Interpretability Saliency Map Generator: Saliency maps (also called heatmaps) operate under the basic principle of highlighting areas of an image that drive the prediction of a model. Among several approaches,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 (i.e., selected class label) to the input image, or by ablation mechanism analyzing the effect of a pixel (or region) to the output when that pixel (or region) is perturbed. In this work we focus on gradient-based saliency maps, building on our observations from [28], on the ability of latent representations of saliency maps to encode information regarding sample informativeness.

To generate interpretability saliency maps we use the iNNvestigate library [39] [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 [40].

Figure 2 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.

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

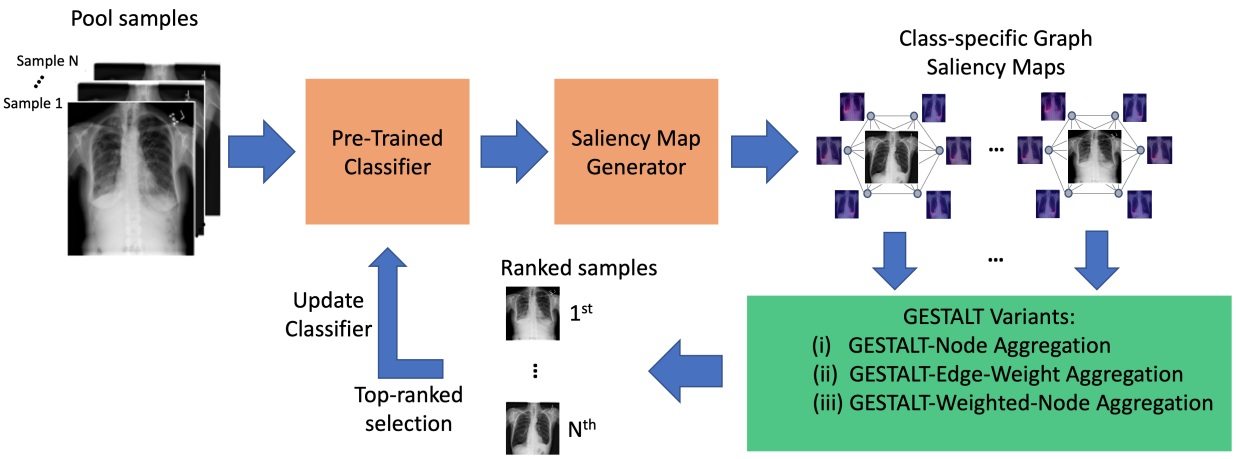
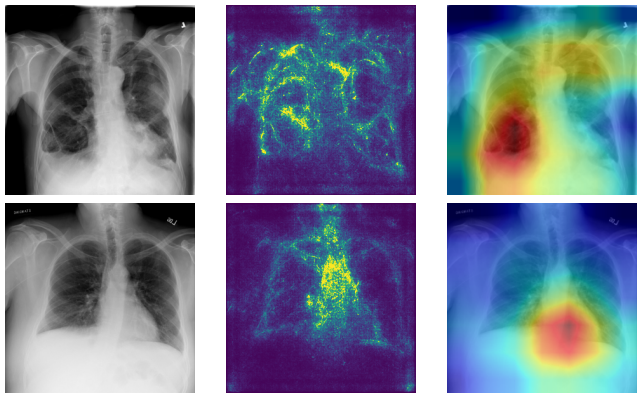


Fig. 1: Workflow of proposed *GESTALT* concept for *Graph NodE BaSed InTerpretAbility Guided SampLe SelecTion* approach. Given pool samples and a trained classifier, interpretability saliency maps for all class labels (illustrated here with five classes) are generated for each testing sample, yielding class-specific graph representations. The GESTALT module then ranks the most informative sample based on multi-label setting. Three different variants to rank samples are presented (described in section IV-D). Orange-colored blocks are standard and interchangeable components in the pipeline.



(a) (b) (c)

Fig. 2: 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.

This justifies our choice of using Deep Taylor approach for generating saliency maps, which is supported by a radiologist with more than 15 years of experience. Overall, in this study we selected DeepTaylor because of its greater accuracy in highlighting important regions.

C. Multi-Label Sample Informativeness

Figure 3 illustrates the concept behind GESTALT and shows the different aggregation strategies used in this paper. We also show the intuition behind different strategies. Our graph is arranged in the following manner:

1) We represent each image sample as a separate graph.
2) Within a graph each of the class labels (representing a disease or condition) is represented by a node.
3) At each node, a class label is represented as the latent representation of the corresponding class-specific saliency map.
4) Edge weights in the graph represent the similarity between corresponding nodes using latent representations.

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

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

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. 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.

In conventional approaches informative samples are determined based on the assumption that each sample has one label. However in order to identify most informative sample for multi-class settings, we propose to incorporate class labels interactions using a graph node based ranking metric scheme.

D. Graph Based Sample Ranking Methods

To aggregate the information contained on each graph to derive a sample informativeness ranking metric, we experiment with three different approaches, described as follows:

1) Node Ranking Across Graphs: Given a sample pool of N images to be ranked by their informativeness, corresponding N graphs are constructed, each composed of K nodes connecting to other $K-1$ nodes with edge weights calculated using Eqn.1. Per graph, a node-based aggregation can then be calculated as:

$$w_k = \sum_{j \in (1,\dots K), j \neq k} w_{kj}, \tag{2}$$

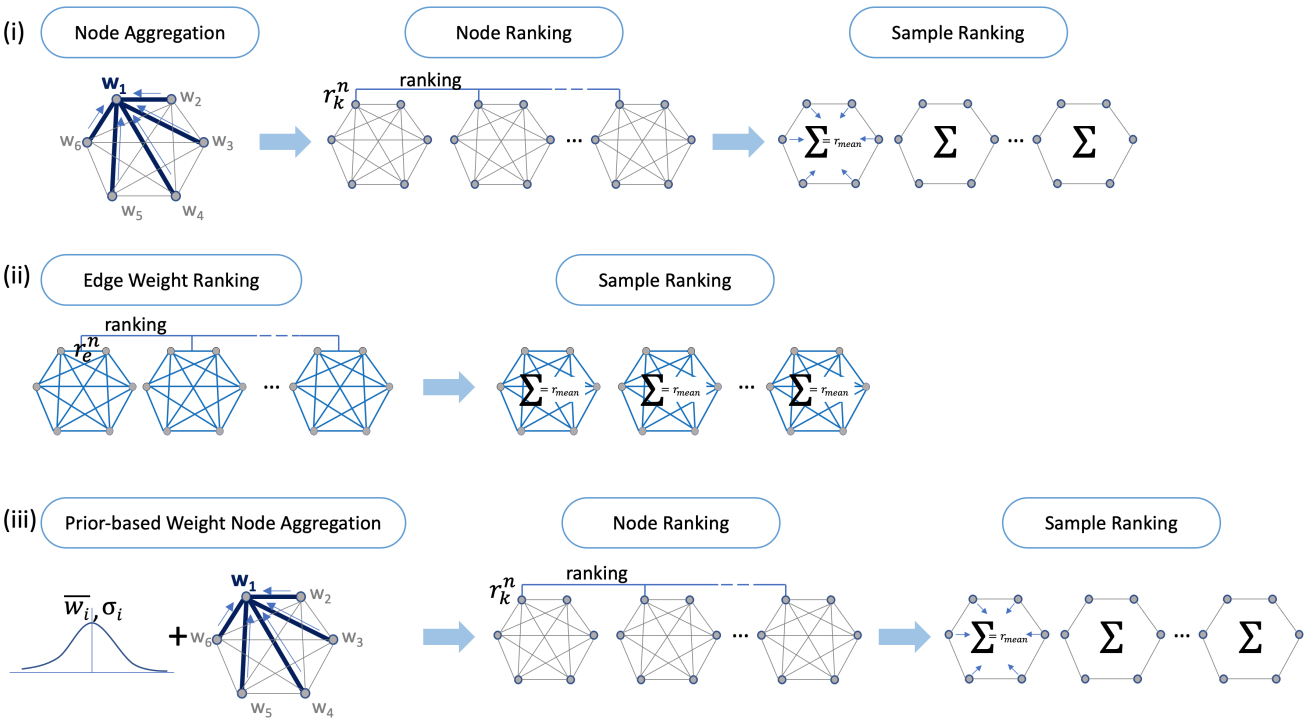


Fig. 3: Graph node based sample informativeness ranking. Three different variants (i), (ii), and (iii) to rank samples in a pool. Variant (i): Edge-weights are aggregated per node, and then ranked for each class label and across all samples. Final sample ranking consists of averaging ranks per sample. Variant (ii): Edge-weights are directly ranked across all samples. Final sample ranking consists of averaging edge ranks per sample. Variant (iii): Edge-weights are aggregated per node in a weighted manner, with weights enabling comparisons of edge-weights of pool samples to prior edge-weight information acquired during model training. Aggregated edge-weights are then ranked for each class label and across all samples. Final sample ranking consists of averaging ranks per sample. Note: diagram using a fix number of six classes for illustrative purposes only.

where w_k denotes the aggregated weight for node k ($k \in (1, \ldots, K)$).

Given N graphs, 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 k-th node of the n-th graph.

For each k-th 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 $\mathbf{r}_k = rank(w_k^1, \ldots, w_k^N)^\intercal$.

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

$$r_{mean}^n = \frac{1}{K} \sum_{k=1}^{K} \mathbf{R}(n,k). \tag{3}$$

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

We note that ranking of nodes for the same class label is performed in order to account for the different levels of complexity in distinguishing each class (i.e., some diseases are harder to differentiate than others, and hence their distinctiveness is lower). This design allows us to handle different levels of distinctiveness typically seen in clinical scenarios, as the ranking values, and hence the sample selection process, will not be biased towards classes of intrinsically higher complexity. As comparison, a less optimal alternative would be for example, to first average all edge weights per sample, and then rank samples by these average values. Such strategy is less optimal since the averaging does not account for the fact that class labels have different levels of similarities (Eq. 1). To further illustrate this rationale, we make the parallel to grand challenges in medical image analysis, where instead of calculating an average metric across a testing set, to then rank by these average values (i.e., ranking of average values), it is preferred to rank per case and then average the ranking values (i.e., average of ranking values), as such approach considers the different levels of complexity of cases in a population. Since this approach is based upon the ranking of nodes, we denote this ranking method as $\text{GESTALT}_{\text{Node}}$.

2) Edge-weight Ranking Across Graphs: In this second variant we determine the most informative samples using ranks of the edge weights. Contrary to the first variant where we first aggregate edge weights per node to then rank them, here we invert the process and rank edge weights and then aggregate. Given a sample pool of N images to be ranked by their informativeness, corresponding N graphs are constructed, each composed of K nodes and comprising a total of $K \times (K-1)/2$ edge weights (since we have an undirected graph), calculated using Eqn.1.

We denote matrix $\mathbf{W} = (w_e^n) \in \mathbb{R}^{N \times K \times \frac{(K-1)}{2}}$, with entry w_e^n describing the edge weight for the e-th edge weight of the n-th graph. Each edge weight on a graph is compared

and ranked with the corresponding ones across all N graphs. We denote matrix $\mathbf{R} = (r_e^n) \in \mathbb{R}^{N \times K \times \frac{(K-1)}{2}}$, with column element $\mathbf{r}_e = rank(w_e^1, \ldots, w_e^N)^\intercal$. In order to derive a final rank for sample n, we aggregate ranks by calculating the mean rank of the individual edge ranks:

$$r_{mean}^n = \frac{2}{K \times (K-1)} \sum_{e=1}^{K \times (K-1)/2} \mathbf{R}(n, e). \quad (4)$$

Samples are then directly ranked and the top-most informative ones can be selected for active learning. Since this approach is based upon the ranking of edge weights, we referred to this ranking variant as GESTALT$_{\text{Edge}}$.

3) Weighted Node Ranking Across Graphs: In the third variant we extend the node-based ranking (variant #1) and incorporate knowledge from the evolving training set during active learning. Given M training images (i.e., including those added as the model is trained with newly queried samples), we redefine equation 2 as:

$$w_k = \sum_{j \in (1, \ldots K), j \neq i} \alpha_{kj} w_{kj}, \quad (5)$$

where,

$$\alpha_{kj} = \frac{w_{kj} - \overline{w}_k}{\sigma_k} \quad (6)$$

$$\overline{w}_k = \frac{\sum_{m=1}^{M} \mathbf{W}(m, k)}{M} \quad (7)$$

$$\sigma_k = \sqrt{\frac{\sum_{m=1}^{M} \mathbf{W}(m, k) - \overline{w}_k}{M}} \quad (8)$$

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 used to construct z-scores α weights based on summary statistics extracted from the 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. In this manner, as queried samples are analyzed their levels of informativeness, and therefore their ranking, can be weighted in relation to previously observed levels of intra-sample node similarities.

The rest of the procedure follows the same steps as for the first variant to derive the final ranking per sample. When the most informative images in a batch are identified, queried for annotations and added to the training set, the summary statistics $\overline{w}_k$, and σ_k, are updated for the next active learning cycle.

V. Baseline Methods For Comparison

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

A. 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 trained a model with all available data. We use a DenseNet-121 classifier [41] for the CheXpert dataset, described below.

B. Standard Active Learning

As first baseline we considered a standard active learning framework where no sample selection is considered, and pool samples are randomly selected for querying and active learning training. It is worth noting that 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, and hence it needs to be kept as low as possible. In the results section we refer to this approach as *Random*.

C. Uncertainty-driven sample selection

This corresponds to our second baseline. As proposed in [4], [5], uncertainty estimation can be used as a metric of sample informativeness for active learning. Given the deep learning model M used for disease classification, mapping an input image I, to a unary output $\widehat{y} \in R$, the predictive uncertainty for pixel y is approximated using:

$$Var(y) \approx \frac{1}{T} \sum_{t=1}^{T} \widehat{y}_t^2 - \left(\frac{1}{T} \sum_{t=1}^{T} \widehat{y}_t \right)^2 + \frac{1}{T} \sum_{t=1}^{T} \widehat{\sigma}_t^2 \quad (9)$$

$\widehat{\sigma}_t^2$ is the model's output for the predicted variance for pixel y_t, and $\widehat{y}_t, \widehat{\sigma}_{t\ t=1}^{2\ T}$ 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*.

D. IDEAL Method

As third comparison approach we consider our previously proposed method in [28] which is denoted as IDEAL (**I**nterpretability-**D**riv**E**n s**A**mple se**L**ection). Below we provide a brief description of IDEAL and refer the reader to [28] for details.

In IDEAL, sample selection is based on a two-step approach. Given a set of pool samples, model predictions are yielded for each pool sample, and corresponding saliency maps are calculated for the predicted class. In a second step, the objective is to associate a level of informativeness to each pool sample based on the information contained in the corresponding saliency map and rank the samples accordingly. To this end, IDEAL employs an additional model trained to classify the latent representation of obtained saliency maps into a predefined set of clusters, each associated to a ranking level (i.e. ordinal clustering), based on positive changes to AUC observed on a separate validation set. The proposed GESTALT approach also builds on information contained on saliency maps, however it does not require an additional model for sample ranking, as it directly ranks pool samples. Furthermore, it explicitly considers the interactions across multiple classes, instead of relying solely on the saliency map for the predicted class, as it is the case of IDEAL.

E. Implementation details

Our method was implemented in TensorFlow. We trained using DenseNet-121 [41] on NIH ChestXray14 dataset [42],

as it is a common architecture used for the task of lung disease classification, and also matching the benchmarked state-of-the-art model in [28]. We used Adam [43] 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 test was performed on a 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 4 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 generation of interpretability saliency maps, we used default parameters of the iNNvestigate implementation of Deep Taylor [39], as well as for GradCAM [44], used to assess model performance for a different interpretability approach. For uncertainty estimation we used a total of $T = 20$ dropout samples with dropout distributed across all layers [45]. During active learning the batch size for our experiments was set to 16.

VI. Results and Discussion

A. Dataset Description

We used the CheXpert dataset [46] 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 set have 200 and 500 images respectively. The validation ground-truth is obtained using majority voting from annotations of 3 board-certified radiologists. Test images are labeled by consensus of 5 board-certified radiologists. The test set evaluation protocol is based on 5 disease labels: *Atelectasis*, *Cardiomegaly*, *Consolidation*, *Edema*, and *Pleural Effusion*, which were selected in order to compare to the IDEAL method [28].

For each task, the dataset was split into training (70%), validation (10%) and test (20%), at the patient level such that all images from one patient are in a single fold.

B. Comparative Results For CheXpert Dataset

In this section we present the main results obtained by the proposed GESTALT method and comparisons with the other benchmarked baseline methods presented in [28], and described in section V. As evaluation metrics we adopted the Area Under the Curve (AUC) evaluated for different methods at every 10% increment of training data, to simulate an active learning scenario. We show comparative results for the following methods: 1) 'GAL-Long et al.'- Graph-Based Active Learning (GAL) approach of [31]; 2) 'LEMAL-Wu et al.': the "example-label based sampling strategy (LEMAL)" approach by [32]; 3) 'CVIRS-Reyes et al.'- the Uncertainty sampling based on "Category Vector Inconsistency and Ranking of Scores (CVIRS)". approach of [33]; 4) 'AlphaMix-Parvaneh et al.' - the "Active Learning by Feature Mixing (Alpha-Mix)" method of [47].

In Figure 4 we show the performance of the three proposed variants of GESTALT using different ranking strategies, and the other benchmarked methods. As reference, results obtained with a fully-supervised model (FSL, AUC=0.902)) are also included, and seen as an horizontal line on each plot.

From Figure 4 we observe that 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, which was surpassed at a much lower value for IDEAL, at 53%. This finding aligns with other similar reports indicating the capability of AL methods to further boost the learning rate of trained models. This behavior aligns with previous works where the same pattern is observed [7], [8], [14], but we note that its exploration goes beyond the scope of this study.

Amongst the other methods, AlphaMix and GAL are the best performing and show similar results. LEMAL and CVIRS perform slightly better than a vanilla uncertainty approach since they are based on uncertainty calculation.

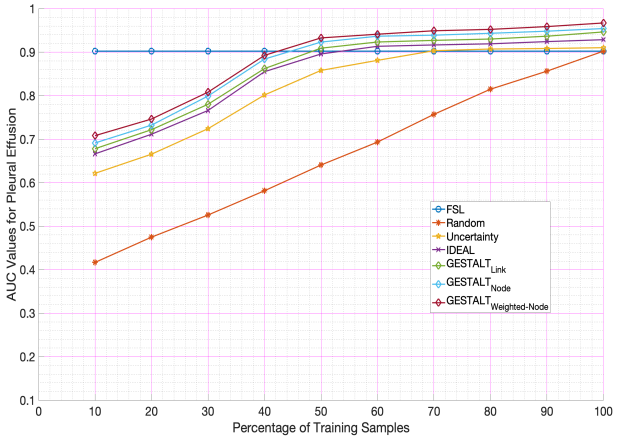
As shown in Figure 4, amongst the proposed GESTALT variants, the third variant, GESTALT$_{Weighted-Node}$, based on a weighted node ranking approach yielded the best results in terms of learning rate and final attainable AUC values. In relation to IDEAL, GESTALT$_{Weighted-Node}$ yield a statistically significant ($p = 0.002$) improved AUC of 0.0384 (0.9287 vs. 0.9671). The two other proposed approaches, GESTALT$_{Node}$ (AUC=0.9543) and GESTALT$_{Link}$ (AUC=0.9465), also outperform IDEAL. GESTALT$_{Link}$ shows slightly inferior performance than AlphaMix and GAL, although GESTALT$_{Node}$ consistently outperforms the two methods, and GESTALT$_{Weighted-Node}$ significantly outperforms AlphaMix and GAL.

We attribute the improved performance of all GESTALT variants due to its explicit characterization of mutual influence across different labels. Within the three variants we observe that GESTALT$_{Node}$ outperforms GESTALT$_{Link}$, which seems to be attributed to the better characterization of node-to-node (or class-to-class) influencing, compared to the link-to-link characterization that includes pairs of class (effectively four classes per link-to-link comparison). The improved results obtained by the node-based aggregation strategy are further emphasized by the third GESTALT variant, where a prior-based weighted node aggregation was investigated. The weighted node based ranking method, GESTALT$_{Weighted-Node}$, yielded the best results by incorporating prior information on the changing distribution of the training data as the model is continuously trained. This allows us to rank samples by their informativeness, not only in relation to other samples in the pool, but also in relation to previously observed samples. We note this is a novel feature other methods have not explored before.

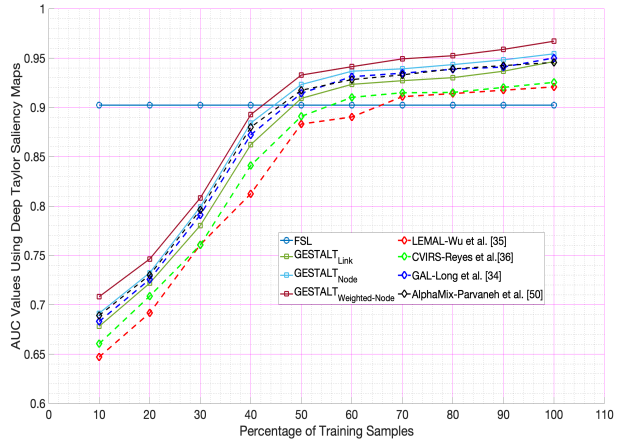
The final AUC values (i.e., at 100% data) were derived from an average of 10 runs and the statistical significance with respect to GESTALT$_{Weighted-Node}$ was calculated using a paired $t-$test. Table I summarizes these results for the different evaluated methods.

Baselines				GESTALT Variants		
FSL	Random	Uncertainty	IDEAL	Node ranking	Edge weight ranking	Weighted node ranking
0.9023 (0.001)	0.9023 (0.0009)	0.9103 (0.008)	0.9287 (0.01)	0.9543 (0.03)	0.9465 (0.005)	0.9671

TABLE I: AUC values for evaluated baselines and proposed GESTALT approach from a 10-fold validation CheXpert dataset. Values between parentheses correspond to p-values with respect to the best performing model, GESTALT$_{Weighted-Node}$



(a)



(b)

Fig. 4: **CheXpert Dataset**: AUC measures at different percentage levels of training percentage for baselines (indicated with dotted lines) and proposed approach, including three investigated variants. (b) Shows results for three 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 the proposed GESTALT approach.

C. Ablation Studies

In this section we include two different ablation experiments in order to (i) analyze the effect of choosing the least informative samples (instead of the most informative) on the learning curves, (ii) utilize the input images, instead of the interpretability saliency maps for feature extraction; (iii) use fixed values of α in the weighting parameter for GESTALT$_{Weighted-Node}$. The first experiment aims at assessing the impact of the sample informativeness ranking, while the second experiment aims at showing the benefits of employing the saliency maps, instead of the input images, to rank samples by their informativeness.

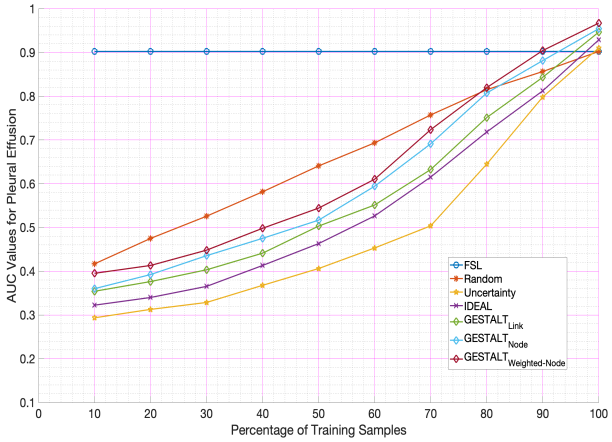


Fig. 5: Ablation studies on the lung classification task to analyze the effect of choosing least informative samples (instead of the most informative) on the learning curves. Results confirm the importance of selecting informative samples.

Figures 5 shows the classification performance of models when the sample informativeness ranking is inverted, and hence, least informative samples per batch are selected. In comparison to Figure 4, we observe in this ablation an overall decrease of the learning rate, lower than the random-based sample selection baseline. In the initial stages since the selected samples are not very informative the increase on learning rate is very low. However, As the number of training samples reaches approximately 60%, the learning rate increases, which we attribute to fact that the remaining samples are more informative and the models benefit more from their integration into the training set. 'Random' continues to have the same behaviour because it does not involve any sample selection. These results confirm the importance of selecting informative samples.

1) Using Latent Features From Images: Figure 6 shows the AUC curves for the second ablation experiment, where we test the benefit of employing the saliency maps as source of sample informativeness vs. the one obtained directly from the input image. As shown in Figure 6 by using image features we obtain a maximum AUC of 0.9231, compared to an AUC of 0.9671 when using saliency maps for a model based on weighted node aggregation. These results also align with those reported in our previous work [28], suggesting that saliency maps highlight information regarding the pathology, whereas latent representations of the entire image also encode other information (e.g. overall anatomy) of much lower relevance for the trained model. Consequently, using deep features from the input image leads to inferior performance.

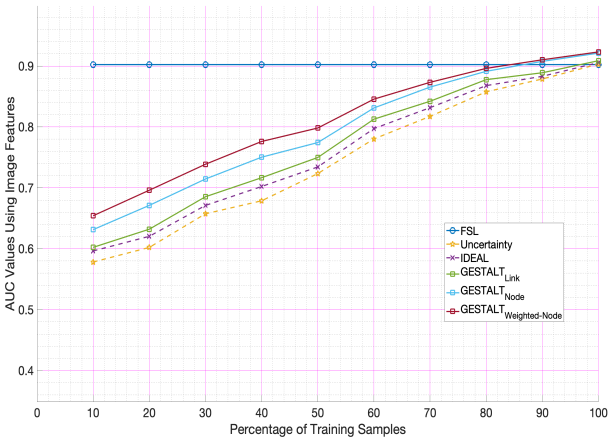


Fig. 6: Ablation studies on the lung classification task to analyze the effect of choosing image features instead of saliency map features.

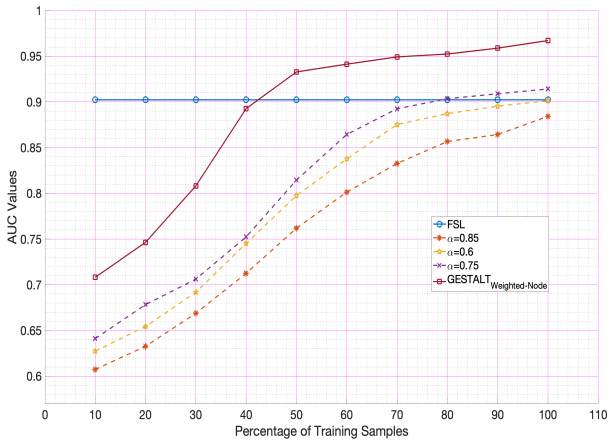


Fig. 7: AUC plot for fixed values of α in Eqn. 5 for GESTALT$_{Weighted-Node}$.

2) Fixed Values of α in Eqn. 5: In Eqn. 5 the value of α changes dynamically with the change of the training samples. This incorporates a dynamic prior knowledge into the node ranking process. We also investigate the effect of using a fixed α on the performance by varying α between $[0.05, 1]$ in steps of 0.05 and found the best performance for $\alpha = 0.75$. However this level of performance is lower than using a dynamic value of α. Figure 7 shows the learning curves for the top 3 values of α. As expected, our originally proposed variant of GESTALT$_{Weighted-Node}$ does much better than using a fixed α.

D. Influence of Saliency Map Computation

In this section we show results when using a different saliency map extraction method such as Grad-CAM [44] in order to assess the sensitivity of the approach when using a different interpretability approach. As alternative approach we employed Grad-CAM [44], due to its popularity.

Figure 8 shows AUC plots using Grad-CAM generated saliency maps. Overall, we observed similar findings as to those reported above using Deep Taylor saliency maps,

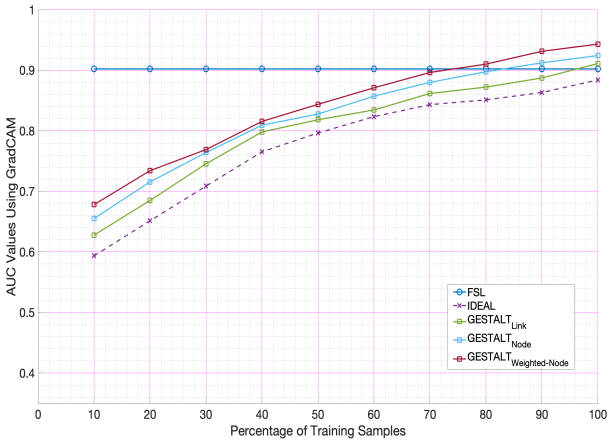


Fig. 8: AUC measures using Grad-CAM saliency maps at different percentage levels of training percentage for baselines and proposed GESTALT approach. Plots are shown for IDEAL and GESTALT variants. As reference, AUC of a fully-supervised model (FSL) is also included as an horizontal line.

showing the superiority of the proposed approach over the baselines. However in comparison with Deep Taylor, model performance yielded via Grad-CAM is lower for each of the corresponding feature extraction methods (see Figure 4). These results then suggest that while improved performance is expected with the proposed GESTALT using different interpretability approaches, selection of the interpretability approach is still needed. This aligns with previous reports and recommendations, regarding differences among interpretability methods proposed in the literature [48], [49]. Nonetheless, the modularity of the proposed pipeline facilitates testing and selection of its different components.

E. 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 at $K = 5$ number of labels. As evaluation metrics we evaluated two different metrics suggested for multi-class scenarios, and available within scikit-learn [50]: 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 [51], which averages over the samples the number of label pairs that are incorrectly ordered, and with perfect score at zero.

In Table II we observe that for $K = 5$, GESTALT$_{Weighted-Node}$ 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 Pham et al. [52], which is the second best ranked method for the CheXpert dataset. This demonstrates that our graph based approach identifies more informative multi-label samples because of the inherently better approach of graph based methods to learn multi-label interactions.

	Baseline Methods		GESTALT Variants		
	DenseNet-121	Pham [52]	Link	Node	Weighted-Node
Rank-Loss ↓	0.21	0.18	0.15	0.14	0.12
LRAP ↑	0.8786	0.8934	0.9013	0.9124	0.9211

TABLE II: 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.

F. Robustness and Generalization

In order to test the robustness of the proposed approach we added simulated noise of $\mu = 0$ and different $\sigma \in \{0.005, 0.01, 0.05, 0.1\}$. Figure 9 shows the AUC values for the baseline performance of GESTALT$_{Link}$ and different σ. The results are close to GESTALT$_{Link}$ for $\sigma = 0.005, 0.01$, but starts to degrade significantly for noise levels above $\sigma = 0.01$, which we term as noise threshold. The noise threshold for GESTALT$_{Node}$ is $\sigma = 0.015$ and for GESTALT$_{Weighted-Node}$ it is $\sigma = 0.019$, indicating higher robustness to noise.

Figure 10 shows the change of the loss function for multiple methods on the validation set. All the methods converge after sufficient number of iterations. However, the error for the baseline methods of DenseNet-121 and Pham is higher than the proposed GESTALT based approaches. Our GESTALT based approaches also converge earlier than the baseline methods, thus demonstrating an added advantage. Amongst the three variants of GESTALT, GESTALT$_{Weighted-Node}$ shows the lowest error and earliest convergence thus highlighting the benefits of including additional information from unlabeled data.

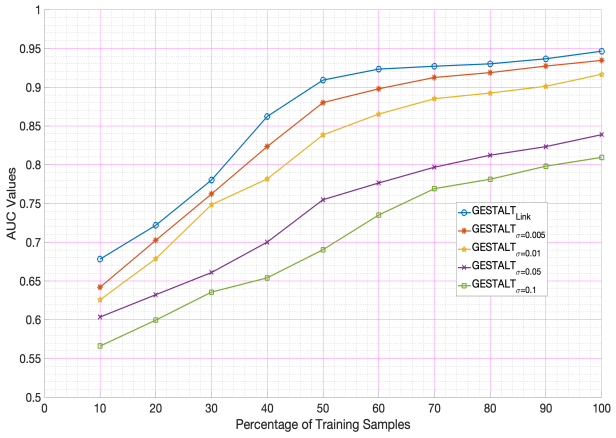


Fig. 9: AUC measures for different features for added Gaussian noise of $\mu = 0$ and different σ. The values are shown in reference to GESTALT$_{Link}$.

G. Performance on Additional Datasets

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 run our model on the NIH ChestXray14 dataset [42] having $112,120$ expert-annotated frontal-view X-rays from $30,805$ unique patients.

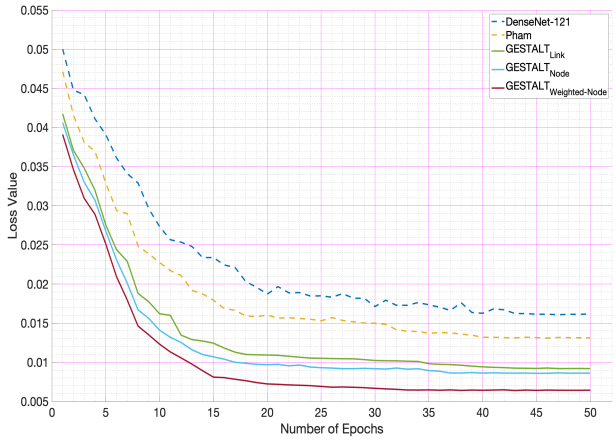


Fig. 10: Loss curves for proposed GESTALT variants and benchmarked approaches.

The set of labels is the same as the CheXpert dataset. We show the learning plots for different methods in Figure 11 which show a similar phenomenon for the CheXpert learning rates in Figure 4.

We also show results on the *multi-class* MedMNIST dataset [53] due to its balanced and standardized datasets spanning across various modalities. We select subsets of the collection appropriate for multi-class disease classification, namely, BreastMNIST [54] having $546/78/156$ breast ultrasound images in the training/validation/test split for malignancy detection, DermaMNIST [55], [56] having $7007/1003/2005$ training/validation/test dermatoscope images for lesion classification, RetinaMNIST [57] having $1080/120/400$ training/validation/test fundus images for diabetic retinopathy severity grading, and TissueMNIST having $165,466/23,640/47,280$ training/validation/test Kidney Cortex Microscope images for multiple disease classification. 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. Figure 12 shows the learning plots for the datasets using GESTALT$_{Weighted-Node}$ 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.

H. Computation Time

For a training dataset of $100,000$ images of size 320×320 the training time for different methods on a NVIDIA Titan X GPU having 12 GB RAM is summarized in Table III. The 12% higher training time for GESTALT$_{Weighted-Node}$ is due to the additional computations and graph construction involved. However, the resulting performance improvement justifies the added complexity of our method. The inference time for a single image is also summarized for different methods, as computation times depend upon the method's complexity.

Training Phase - Time in Hours										
DenseNet-121	Random	Unc	IDEAL	GESTALT $_{Link}$	GESTALT $_{Node}$	GESTALT $_{Weighted-Node}$	GAL [31]	LEMAL [32]	CVIRS [33]	AlfaMix [47]
18(0.72T)	18.5(0.74T)	19.5(0.78T)	22(0.8T)	23.6(0.94T)	24(0.96T)	25(T)	23.5(0.94T)	20(0.8T)	21.5(0.86T)	24(0.96T)
Test/Inference Phase - Time in Seconds										
0.18	0.19	0.2	0.25	0.3	0.3	0.32	0.28	0.22	0.24	0.3

TABLE III: Training and Inference time for different methods.

VII. Conclusions

In this work we present a novel sample informativeness selection approach for active learning, referred to as GESTALT for Graph Node Based Interpretability Guided Sample Selection. The basic idea behind GESTALT builds on our previous findings [28], where we observed that information contained in saliency maps can be used to derive rankings of sample informativeness outperforming state of the art uncertainty-based active learning methods. In this work we propose to leverage the multi-class information derived from class-specific saliency maps, describing hypothetical or pseudo-counterfactual saliency maps across different potential predicted class labels, by encoding this information in the form of a graph, and associate levels of intra-graph node similarities to sample informativeness. In relation to the state of the art, the proposed GESTALT approach directly establishes a ranking of pool samples, without the need of additional trained models as required in [28]. We report results on three GESTALT variants outperforming other compared approaches. In these regards, while we only tested three different variants to aggregate information for sample ranking, we highlight the diversity of options available from the set of information contained in the proposed graph representation, and expect that other advanced aggregation mechanisms can be proposed to further boost model performance. Among the three proposed variants, the weighted-node ranking variant yielded the best results. The proposed weighted-noded ranking approach incorporates a novel mechanism to include collected prior-information, on intra-sample node similarities, into the sample informativeness ranking. Interestingly, this feature can be further extended to include other desired characteristics of ranked samples for active learning. Notably, the approach can be extended to leverage approaches having the ability to select both informative and non-redundant samples, where an inter-sample nodes similarity metric can be created to further penalize redundant samples from being selected. The proposed GESTALT approach yields improved learning rates, which we highlight in light of the importance of minimizing expert annotations in the clinical routine while targeting high accuracy levels. Similarly, in scenarios where imaging protocols or vendors change or evolve, we esteem to be important to develop methods featuring high learning rates. The modularity of the approach enables direct integration of advances in model architecture (i.e. it does not require architectural changes), and related progress in the fast-evolving area of interpretability (e.g., [58], [59]).

There are few limitations of the study worth mentioning. In this study we tested the proposed approach on the problem of lung disease multi-class classification. Further research is needed to verify that the same findings apply for other multi-class classification tasks. In this study we did not test other architectures and focused on a single model architecture (DenseNet-121). Although this architecture is commonly used and representative of current model performance, we think that other architectures would be worth exploring in conjunction with the proposed sample informativeness selection approach.

We anticipate several extensions of the proposed graph-based interpretability guided sample representation to other deep learning tasks. For example, in medical image retrieval, the proposed graph representation can be naturally extended to provide a similarity metric between samples. Similarly, the proposed graph representation can be used in the context of inductive bias and model training, where learned features of a trained model are enforced to yield desired metrics of intra-sample multi-class distinctiveness.

VIII. Acknowledgements

We highly appreciate the help given by Behzad Bozorgtabar in implementing some of the comparison algorithms. This work was supported by the Swiss National Foundation grant number 198388, and Innosuisse grant number 31274.1.

References

[1] J.-J. Zhu and J. Bento, "Generative adversarial active learning," *arXiv preprint arXiv:1702.07956*, 2017.

[2] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin., "Cost-effective active learning for deep image classification." *IEEE Trans. CSVT.*, vol. 27, no. 12, pp. 2591–2600, 2017.

[3] D. Mahapatra, P. Schüffler, J. Tielbeek, F. Vos, and J. Buhmann, "Semi-supervised and active learning for automatic segmentation of crohn's disease," in *Proc. MICCAI, Part 2*, 2013, pp. 214–221.

[4] B. Bozorgtabar, D. Mahapatra, H. von Teng, A. Pollinger, L. Ebner, J.-P. Thiran, and M. Reyes, "Informative sample generation using class aware generative adversarial networks for classification of chest xrays," *Computer Vision and Image Understanding*, vol. 184, pp. 57–65, 2019.

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

[6] Y. Gal, R. Islam, and Z. Ghahramani, "Deep Bayesian Active Learning with Image Data," in *Proc. International Conference on Machine Learning*, 2017.

[7] L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Chen, "Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation," in *Proc. MICCAI*, 2017, pp. 399–407.

[8] J. Sourati, A. Gholipour, J. G. Dy, X. Tomas-Fernandez, S. Kurugol, and S. K. Warfield, "Intelligent labeling based on fisher information for medical image segmentation using deep learning," *IEEE transactions on medical imaging*, vol. 38, no. 11, pp. 2642–2653, 2019.

[9] D. Mahapatra and J. Buhmann, "Visual saliency-based active learning for prostate magnetic resonance imaging segmentation." *SPIE Journal of Medical Imaging*, vol. 3, no. 1, p. 014003, 2016.

[10] ——, "Visual saliency based active learning for prostate mri segmentation," in *In Proc. MLMI*, 2015, pp. 9–16.

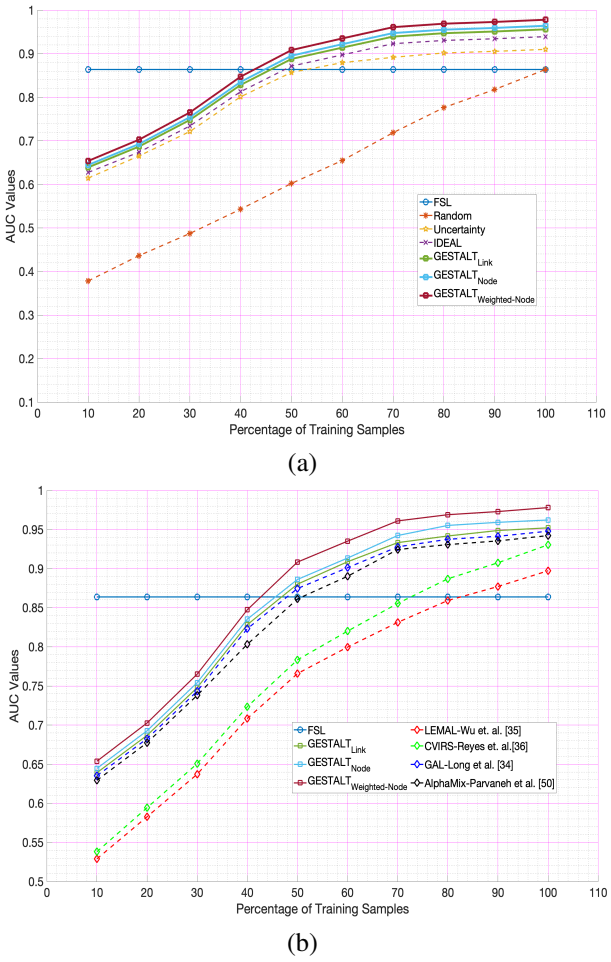


Fig. 11: **NIH DataSet:** AUC measures at different percentage levels of training percentage for baselines (indicated with dotted lines) and proposed approach, including three investigated variants. (b) Shows results for three multi-label Al learning approaches. As reference, AUC of a fully-supervised model (FSL) is also included as an horizontal line. Improved learning rates and model performance is observed for the proposed GESTALT approach.

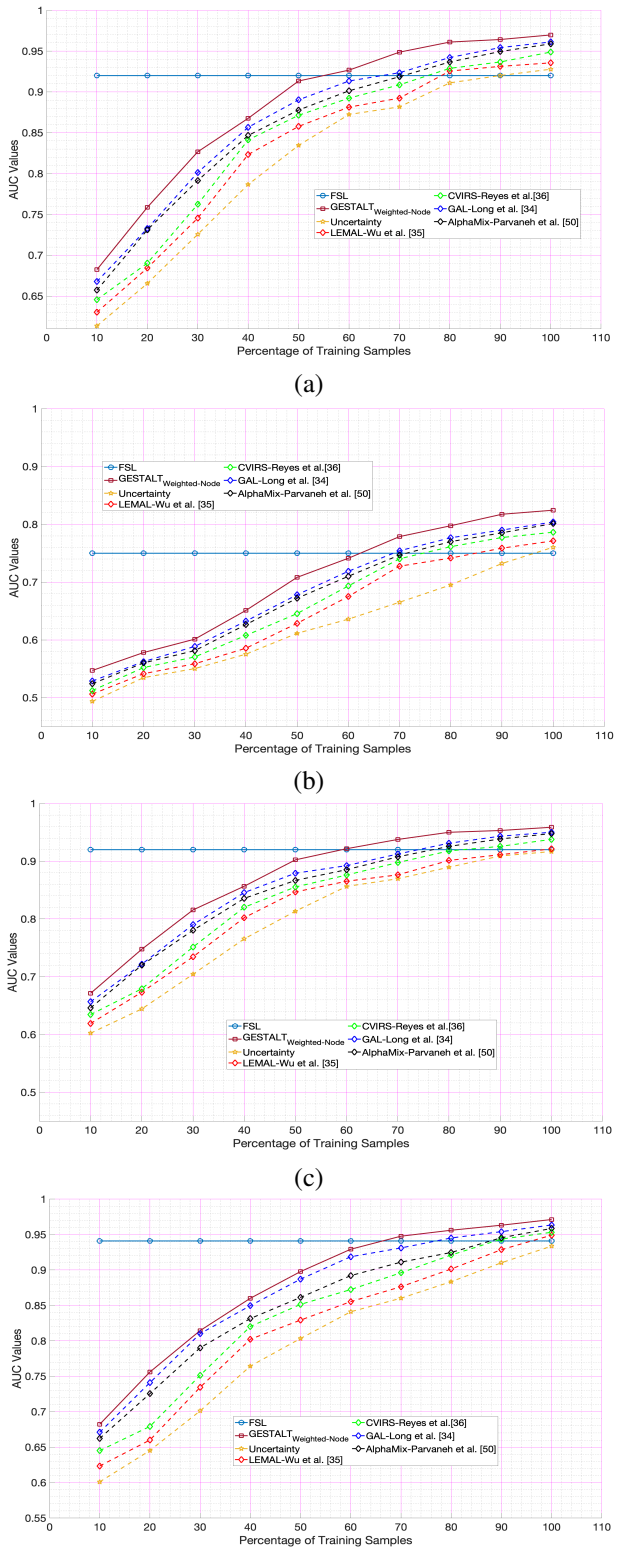


Fig. 12: **MedMNIST DataSets:** AUC measures at different percentage levels of training percentage for proposed approach and four multi-label Al learning approaches. As reference, AUC of a fully-supervised model (FSL) is also included as an horizontal line. Improved learning rates and model performance is observed for the proposed GESTALT approach. (a) Dermatology; (b) Retinal fundus images; (c) Breast dataset; (d) Tissue dataset

[11] D. Mahapatra, J.Tielbeek, J. Makanyanga, J. Stoker, S. Taylor, F. Vos, and J. Buhmann, "Active learning based segmentation of crohn's disease using principles of visual saliency," in *Proc. IEEE ISBI*, 2014, pp. 226–229.

[12] H. Zheng, L. Yang, J. Chen, J. Han, Y. Zhang, P. Liang, Z. Zhao, C. Wang, and D. Z. Chen, "Biomedical image segmentation via representative annotation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 5901–5908.

[13] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. CVPR*, 2016, pp. 2921–2929.

[14] C. Mayer and R. Timofte, "Adversarial sampling for active learning," in *arXiv preprint arXiv:1808.06671*, 2018.

[15] A. Kirsch, J. Van Amersfoort, and Y. Gal, "Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning," *Advances in neural information processing systems*, vol. 32, pp. 7026–7037, 2019.

[16] W. H. Beluch, T. Genewein, A. Nürnberger, and J. M. Köhler, "The Power of Ensembles for Active Learning in Image Classification," in *Computer Vision and Pattern Recognition*, 2018, pp. 9368–9377.

[17] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges,"

Information Fusion, vol. 76, pp. 243–297, dec 2021.

[18] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1321–1330.

[19] D. Mahapatra, B. Bozorgtabar, and L. Shao, "Pathological retinal region segmentation from oct images using geometric relation based augmentation," in *In Proc. IEEE CVPR*, 2020, pp. 9611–9620.

[20] A. Jungo and M. Reyes, "Assessing reliability and challenges of uncertainty estimations for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 48–56.

[21] D. Mahapatra, A. Poellinger, and M. Reyes, "Interpretability-guided inductive bias for deep learning based medical image classification and segmentation," *Medical Image Analysis*, p. 102551, 2022.

[22] A. Mehrtash, W. M. Wells, C. M. Tempany, P. Abolmaesumi, and T. Kapur, "Confidence calibration and predictive uncertainty estimation for deep medical image segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 12, pp. 3868–3878, 2020.

[23] M. Minderer, J. Djolonga, R. Romijnders, F. Hubis, X. Zhai, N. Houlsby, D. Tran, and M. Lucic, "Revisiting the Calibration of Modern Neural Networks," *arXiv preprint arXiv:2106.07998*, 2021.

[24] J. Moon, J. Kim, Y. Shin, and S. Hwang, "Confidence-aware learning for deep neural networks," in *international conference on machine learning*. PMLR, 2020, pp. 7034–7044.

[25] D. Mahapatra, Z. Ge, and M. Reyes, "Self-supervised generalized zero shot learning for medical image classification using novel interpretable saliency maps," *IEEE Transactions on Medical Imaging*, vol. 41, no. 9, pp. 2443–2456, 2022.

[26] J. Zhang, B. Kailkhura, and T. Y.-J. Han, "Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 11 117–11 128.

[27] S. Budd, E. C. Robinson, and B. Kainz, "A survey on active learning and human-in-the-loop deep learning for medical image analysis," *Medical Image Analysis*, vol. 71, p. 102062, jul 2021.

[28] D. Mahapatra, A. Poellinger, L. Shao, and M. Reyes, "Interpretability-driven sample selection using self supervised learning for disease classification and segmentation," *IEEE TMI*, vol. 40, no. 10, pp. 2548–2562, 2021.

[29] J. Wu, V. S. Sheng, J. Zhang, H. Li, T. Dadakova, C. L. Swisher, Z. Cui, and P. Zhao, "Multi-label active learning algorithms for image classification: Overview and future promise," *ACM Comput. Surv.*, vol. 53, no. 2, 2020.

[30] Y. Yang, Z. Ma, F. Nie, and et al, "Multi-class active learning by uncertainty sampling with diversity maximization," *Int J Comput Vis*, vol. 113, pp. 113–127, 2015.

[31] J. Long, J. Yin, W. Zhao, and E. Zhu, "Graph-based active learning based on label propagation," in *Modeling Decisions for Artificial Intelligence*, V. Torra and Y. Narukawa, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 179–190.

[32] J. Wu, V. S. Sheng, J. Zhang, P. Zhao, and Z. Cui, "Multi-label active learning for image classification," in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 5227–5231.

[33] O. Reyes, C. Morell, and S. Ventura, "Effective active learning strategy for multi-label learning," *Neurocomputing*, vol. 273, pp. 494–508, 2018.

[34] B. Du, Z. Wang, L. Zhang, L. Zhang, and D. Tao, "Robust and discriminative labeling for multi-label active learning based on maximum correntropy criterion," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1694–1707, 2017.

[35] X. Li and Y. Guo, "Active learning with multi-label svm classification," in *IJCAI '13*, 2013, p. 1479–1485.

[36] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Ng, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," in *arXiv preprint arXiv:1711.05225,*, 2017.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *In Proc. CVPR*, 2016.

[38] K. Simonyan and A. Zisserman., "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[39] M. Alber, S. Lapuschkin, P. Seegerer, M. Hägele, K. T. Schütt, G. Montavon, W. Samek, K.-R. Müller, S. Dähne, and P.-J. Kindermans, "innvestigate neural networks," *Journal of Machine Learning Research*, vol. 20, no. 93, pp. 1–8, 2019.

[40] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern Recognition*, vol. 65, pp. 211–222, 2017.

[41] G. Huang, Z. Liu, L. van der Maaten, and K. Weinberger, "Densely connected convolutional networks," in *https://arxiv.org/abs/1608.06993,*, 2016.

[42] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *In Proc. CVPR*, 2017.

[43] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *arXiv preprint arXiv:1412.6980,*, 2014.

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

[45] A. Kendall and Y. Gal, "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?" in *Advances in Neural Information Processing Systems.*, 2017.

[46] J. Irvin, P. Rajpurkar, and et al., "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *arXiv preprint arXiv:1901.07031*, 2019.

[47] A. Parvaneh, E. Abbasnejad, D. Teney, G. R. Haffari, A. van den Hengel, and J. Q. Shi, "Active learning by feature mixing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 237–12 246.

[48] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017.

[49] M. Reyes, R. Meier, S. Pereira, C. A. Silva, F.-M. Dahlweid, H. v. Tengg-Kobligk, R. M. Summers, and R. Wiest, "On the interpretability of artificial intelligence in radiology: Challenges and opportunities," *Radiology: Artificial Intelligence*, vol. 2, no. 3, p. e190043, 2020. [Online]. Available: https://doi.org/10.1148/ryai.2020190043

[50] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[51] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining multi-label data," in *Data mining and knowledge discovery handbook*. Springer, 2009, pp. 667–685.

[52] H. H. Pham, T. T. Le, D. Q. Tran, D. T. Ngo, and H. Q. Nguyen, "Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels," in *arXiv preprint arXiv:1911.06475,*, 2020.

[53] J. Yang, R. Shi, and B. Ni, "Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis," in *ISBI*, 2021.

[54] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images." in *Data Brief 28, https://doi.org/10.1016/j.dib.2019.104863*, 2020.

[55] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," in *Sci. data 5*, 2018.

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

[57] "Deepdrid. the 2nd diabetic retinopathy – grading and image quality estimation challenge," in *https://isbi.deepdr.org/data.html*, 2020.

[58] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 782–791.

[59] B. Pan, R. Panda, Y. Jiang, Z. Wang, R. Feris, and A. Oliva, "Ia-red2: Interpretability-aware redundancy reduction for vision transformers," *Advances in Neural Information Processing Systems*, vol. 34, 2021.