

Is Complete Labeling Necessary? Understanding Active Learning in Longitudinal Medical Imaging

Siteng Ma^{1,2}, Honghui Du^{1,2}, Prateek Mathur^{1,2}, Brendan S. Kelly^{2,5},
Ronan P. Killeen^{3,4}, Aonghus Lawlor^{1,2}, Ruihai Dong^{1,2}

¹School of Computer Science, University College Dublin, Ireland

²Insight Centre for Data Analytics, Dublin, Ireland

³School of Medicine, University College Dublin, Dublin, Ireland

⁴Department of Radiology, St Vincent's University Hospital, Dublin, Ireland

⁵Great Ormond Street Hospital for Children, London, United Kingdom

Emails: {siteng.ma,honghui.du,prateek.mathur,aonghus.lawlor,ruihai.dong}@insight-centre.org,
brendan.kelly@insight-centre.org, ronan.killeen@st-vincent.ie

Abstract—Detecting changes in longitudinal medical imaging using deep learning requires a substantial amount of accurately labeled data. However, labeling these images is notably more costly and time-consuming than labeling other image types, as it requires labeling across various time points, where new lesions can be minor, and subtle changes are easily missed. Deep Active Learning (DAL) has shown promise in minimizing labeling costs by selectively querying the most informative samples, but existing studies have primarily focused on static tasks like classification and segmentation. Consequently, the conventional DAL approach cannot be directly applied to change detection tasks, which involve identifying subtle differences across multiple images. In this study, we propose a novel DAL framework, named Longitudinal Medical Imaging Active Learning (LMI-AL), tailored specifically for longitudinal medical imaging. By pairing and differencing all 2D slices from baseline and follow-up 3D images, LMI-AL iteratively selects the most informative pairs for labeling using DAL, training a deep learning model with minimal manual annotation. Experimental results demonstrate that, with less than 8% of the data labeled, LMI-AL can achieve performance comparable to models trained on fully labeled datasets. We also provide a detailed analysis of the method's performance, as guidance for future research. The code is publicly available at https://github.com/HelenMa9998/Longitudinal_AL.

Index Terms—Active learning, longitudinal medical image, Multiple sclerosis.

I. INTRODUCTION

A longitudinal study in medical image analysis involves tracking and analysing a series of medical images over time to observe and measure subject-specific changes in a patient's condition. This plays a pivotal role in managing chronic and progressive diseases (e.g., tracking tumor growth and shrinkage, assessing the progression of Multiple Sclerosis) [1].

Recently, supervised deep learning (DL) approaches that train the deep model using a fully labeled dataset have become well known for effectively identifying and extracting meaningful information in computer vision and medical image analysis [2]. However, their effectiveness is heavily dependent on the availability of extensive datasets with accurate labels [3], which poses a major challenge in the context of longitudinal medical imaging [1]. Labeling longitudinal medical images

is notably more laborious, costly, and time-consuming [4] compared to other scenarios. It requires experienced experts to label the data across multiple time points [5]. Moreover, given that the changes in the images can be minute (3-10 voxels in size on MRI images) and may present with subtle appearances [6], it is easy to fail in annotating minor abnormalities or changes, leading to inaccuracies and inconsistencies in labeling [6], [7]. Therefore, this demands experts label the data very carefully and sometimes repeatedly to ensure all possible changes are annotated, making the process of labeling large longitudinal medical image datasets impractical.

Deep Active Learning (DAL) is a machine learning approach where the algorithm selectively queries the most valuable data examples for training deep learning models. Oracles (e.g., radiologists) label these examples to minimize the overall labeling effort and maintain or enhance the model's predictive performance [8]. Recent research [9]–[12] has demonstrated that DAL can effectively reduce labeling costs in 2D/3D medical image analysis.

As noted, labeling and training with longitudinal medical images can be much more challenging than with standard medical images. However, no studies have been conducted to explore the effectiveness of DAL with longitudinal medical imaging data. Consequently, it remains unclear what an effective query strategy would be for selecting images in pairs to train a DL model for change detection effectively. Simply applying existing DAL methods to longitudinal medical images may result in poor performance. Therefore, this paper aims to answer the following research question: *How can DAL be used to reduce the labeling effort in longitudinal medical imaging while maintaining the model's performance?*

To answer this question, we propose a novel approach named Longitudinal Medical Imaging Active Learning (LMI-AL). LMI-AL is the first framework specifically designed to adapt DAL approaches to change detection in longitudinal medical imaging, significantly reducing labeling efforts by enhancing the efficiency of DL model training. Initially, LMI-AL transforms 3D images into 2D slices. Each pair of

slices (baseline and follow-up scans) and their differences are grouped together. To ensure the model learns from all potential variations, LMI-AL creates the initial slice pool using all possible pairwise combinations of slices. LMI-AL then iteratively selects slice pairs and their corresponding differences based on the selected AL query strategy for labeling and trains the DL model accordingly. Through extensive experiments, we demonstrate that LMI-AL achieves comparable performance to fully supervised methods while requiring less than 8% of labels, significantly decreasing the annotated image slices. We also provide detailed analysis based on the result to give guidance for future researchers.

The rest of the paper is organized as follows: we present a survey of related research in Section II, the proposed LMI-AL framework is detailed in Section III. Then Section IV and Section V present the experiment setup, results, and corresponding analysis of our empirical studies. The conclusion can be found in Section VI.

II. RELATED WORK

In this section, we present an overview of DL change detection models and methods in longitudinal medical settings. Then Deep Active Learning (DAL) methods in general scenarios are given.

A. Change Detection in Longitudinal Medical Imaging

A majority of literatures [2], [13], [14] have focused on leveraging fully supervised learning for solving problems in longitudinal medical image analysis.

Initially, Recurrent Neural Network (RNN)-based models (e.g., LSTM [15] and GRU [16]) are extensively used to capture temporal relationships in longitudinal medical imaging data. These approaches typically involve using convolutional neural networks (CNNs) to extract features from images, which are then processed by RNN-based models to learn temporal dependencies [2], [17]. However, RNN-based models are difficult to train and prone to the exploding gradient problem [18]. Recent studies have integrated LSTM layers into CNNs to better model temporal images [13], [19].

CNN models have proven effective in image analysis [20], are frequently used for change detection [6], [14], [21]. For instance, UNet [22], a popular CNN-based segmentation architecture, concatenates baseline and follow-up images as 2-channel inputs to enhance model performance [6], [14], [21], where [6] further involves the difference between baseline and follow-up images to help the model concentrate on the potential change areas. Moreover, UNet remains the core model and employs multiple views, such as coronal, axial, and sagittal planes in 3D medical imaging (e.g., MRI or CT scans) by concatenating adjacent temporal sequences as inputs and averaging the outputs, which can further improve the results [23]. Recent innovations have incorporated attention mechanisms into UNet to better focus the model on target regions [24], maintaining the basic framework described above.

Due to the typically small size of new lesions, cascaded CNNs [25] (e.g., a series of convolutional neural networks

are connected sequentially, with the output of one used as the input for the next) are introduced. Researchers use an initial network to identify lesion centers from baseline images, extract patches from both baseline and follow-up images, and then feed these patches into the next model to enhance change detection [26]. To handle hard-to-learn cases, [27] adapts Online Hard Example Mining strategy that uses two networks, where the first identifies potential new lesion voxels, and the second reduces misclassified voxels from the first network's output. Siamese networks [28] are proposed to compare two inputs and determine their similarity, such as face verification and signature verification. Fenneteau et al. [29] utilize this approach to segment each time slot, and then a refinement block concatenates segmentation maps produced for each FLAIR image to predict the final segmentation of new lesions. Vision Transformers (ViTs) [30] have gained popularity due to their ability to learn inherent image features and temporal relationships. A recent study [31] adopts ViTs to enhance learning in longitudinal medical imaging.

Although all these methods have achieved excellent results, they are fully supervised and require large amounts of well-labeled data to train the model, which is impractical in longitudinal medical imaging.

B. Deep Active Learning

Unlike traditional supervised learning, DAL reduces labeling costs by selectively querying an oracle to label the most informative data points, achieving similar or better performance with fewer labels [8], [32]. Every DAL scenario involves determining the information contained in unlabeled instances, which is defined as a query strategy. Uncertainty sampling is the most common query strategy in DAL, where the model selects the most uncertain data points for labeling (e.g., the model's prediction confidence on an instance is less than 0.5) [8]. By focusing on labeling uncertain examples, the model reduces overall uncertainty, which improves its ability to make accurate predictions and enhances learning efficiency [33], [34]. Using the softmax output of the deep model [33] and information entropy [8], [35] are two common methods for measuring the uncertainty of an instance. Then, as a way to evaluate the uncertainty of the model, [36] introduced the use of Bayesian Convolutional Neural Networks for Active Learning (BALD) to mitigate the issue of the DL model exhibiting excessive confidence in its output results. Monte Carlo (MC) dropout [36] is proposed as an easier way to evaluate Bayesian uncertainty.

Although the uncertainty-based method is simple in form and has low computational complexity, they consider each sample independently and therefore ignore the correlation between samples [32]. Diversity-based query strategy is another major approach in DAL. It selects a diverse set of data points for labeling to cover a broad range of the input space, improving the model's generalization and reducing redundancy in the training data [37]. The Coreset technique is proposed as an effective representation learning method to select samples [37], by identifying and selecting the subset of

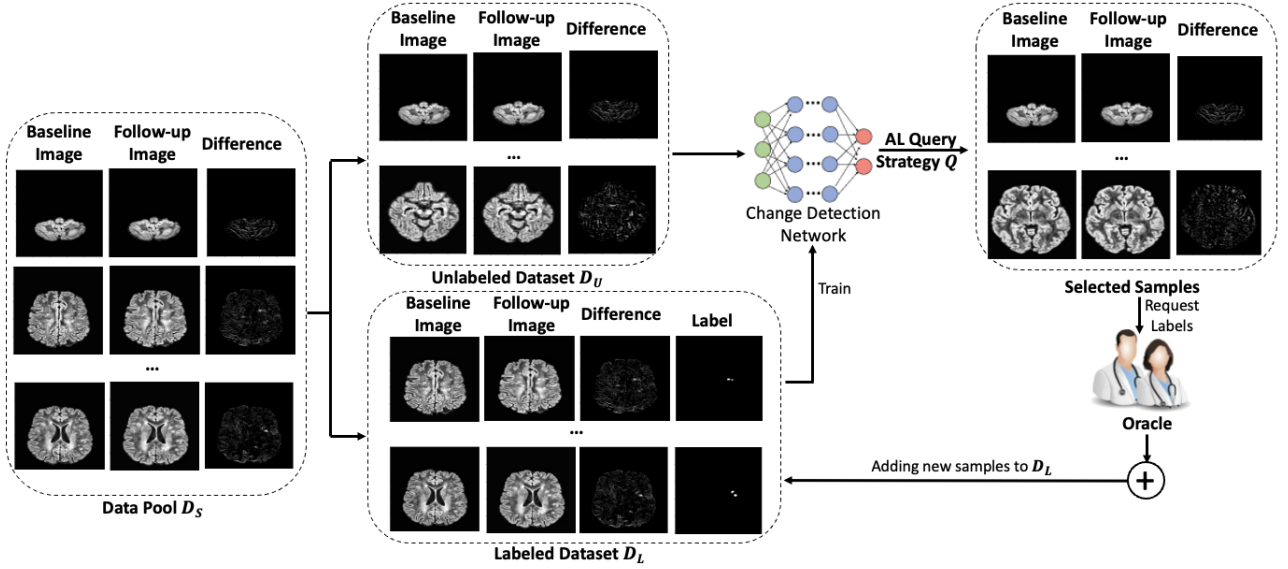


Fig. 1: The overall process of applying active learning in longitudinal medical image analysis.

data that exhibits substantial deviations from the distribution of the already labeled data. To leverage the advantages of both uncertainty and diversity methods, hybrid methods combine different query strategies to optimize the selection of data points for labeling [38], [39].

However, DAL has been typically used for single-image tasks (e.g., classification, segmentation). There is no existing approach that can be directly applied to the medical change detection task.

III. PROPOSED METHOD

Let $\mathcal{D} = \{I_n^t \mid n \in \{1, \dots, N\}, t \in \{1, \dots, T_n\}\}$ denotes a longitudinal medical image dataset that includes images from N patients, where $I_n^t \in \mathbb{R}^{h \times w \times c}$ is a 3D image for patient n scanned at a specific time-step t ; h, w, c represent the height, width, and number of slices of the image, respectively. $T_n \geq 2$ is the total number of images for patient n , T_n can vary across patients. Given two unseen images $I_{n'}^t$ as the baseline and $I_{n'}^{t'}$, $t' > t$ as the follow-up of the same patient n' , the goal of change detection in longitudinal medical imaging is to learn a model $f(\cdot)$ from $\mathcal{D}$ that can identify all the new lesions in $I_{n'}^{t'}$ compared to $I_{n'}^t$. Formally, this can be defined as follows:

$$f(I_{n'}^t, I_{n'}^{t'}) = m; m : \mathbb{R}^{h \times w \times c} \rightarrow \{0, 1\} \quad (1)$$

where m is the binary change map for voxels. The definition can be simplified to 2D settings when $c = 1$, so that $I_n^t \in \mathbb{R}^{h \times w}$. In this paper, to better illustrate the proposed method, all I are considered as 3D images.

Annotating 3D images directly can be complex and overwhelming due to the high dimension. It is common to divide 3D images into 2D slices to make the annotation task more manageable [40]. Consider that each I_n^t is composed of a set of 2D slices $\mathcal{X}_n^t = \{\mathbf{x}_{n,k}^t\}_{k=1}^c$, where $\mathbf{x}_{n,k}^t \in \mathbb{R}^{h \times w}$ is the k th slice of I_n^t . The dataset $\mathcal{D}$ can be denoted as $\mathcal{D} = \{\mathbf{x}_{n,k}^t \mid n \in \{1, \dots, N\}, t \in \{1, \dots, T_n\}, k \in \{1, \dots, c\}\}$.

Algorithm 1 Learning Procedure of LMI-AL

Require: Dataset $\mathcal{D}$, Budget B , Deep learning model f , Base Active learning approach Q , The number of initial query slice pairs q_0 , The number of querying slice pairs at each iteration q .

- 1: Initialize $\mathcal{D}_S$
- 2: Initialize $\mathcal{D}_L$ by randomly querying q_0 slice pairs from $\mathcal{D}_S$
- 3: **while** $|\mathcal{D}_L| < B$ **do**
- 4: $\mathcal{D}_U \leftarrow \mathcal{D}_S - \mathcal{D}_L$
- 5: Train model f on $\mathcal{D}_L$
- 6: $\mathcal{D}_Q \leftarrow Q(\mathcal{D}_U, q, f)$
- 7: $\mathcal{D}_L \leftarrow \mathcal{D}_L \cup \mathcal{D}_Q$
- 8: **end while**

Let $\mathbf{y}_{n,k}^{t,t'} \in \{0, 1\}^{h \times w}$ represents the corresponding 2D change pixel label between a pair of slices $\mathbf{x}_{n,k}^t$ and $\mathbf{x}_{n,k}^{t'}$, where $t' > t$. The objective of the proposed method is to minimize the number of labeled slice pairs needed to train $f(\cdot)$ through DAL while maintaining performance similar to that of supervised learning.

The learning pipeline and pseudo-code for LMI-AL are shown in Figure 1 and Algorithm 1, respectively. LMI-AL begins dividing the 3D images into 2D slices. Unlike other single-image tasks (e.g., segmentation, classification), change detection requires a pair of images as input (e.g., one baseline and one follow-up slice). To ensure the model can learn from all potential variations and differences between slices and maximize available training information, LMI-AL creates the initial pairwise slice pool $\mathcal{D}_S$ using all possible pairwise combinations of slices. Additionally, due to variations in imaging conditions (e.g., physiological changes, patient movement, and positioning differences), slices of the same location at different

times always exhibit slight differences. To better focus on changes, reduce noise, and improve learning efficiency, LMI-AL uses the difference between two slices as an additional input. The initial data pool $\mathcal{D}_S$ can be formally written as $\mathcal{D}_S = \{(\mathbf{x}_{n,k}^t, \mathbf{x}_{n,k}^{t'}, \mathbf{x}_{n,k}^{t'} - \mathbf{x}_{n,k}^t) \mid n \in \{1, \dots, N\}, t \in \{1, \dots, T_n - 1\}, t' \in \{2, \dots, T_n\}, t' > t, k \in \{1, \dots, c\}\}$ (line 1).

A small pre-defined number q_0 of slice pairs from $\mathcal{D}_S$ are randomly selected and labeled to form the initial labeled pool $\mathcal{D}_L = \{((\mathbf{x}_{n,k}^t, \mathbf{x}_{n,k}^{t'}, \mathbf{x}_{n,k}^{t'} - \mathbf{x}_{n,k}^t), \mathbf{y}_{n,k}^{t,t'}) \mid n \in \{1, \dots, N\}, t \in \{1, \dots, T_n - 1\}, t' \in \{2, \dots, T_n\}, t' > t, k \in \{1, \dots, c\}\}$ (line 2). Subsequently, LMI-AL commences a loop continues until the size of $\mathcal{D}_L$ reaches a predefined budget B (line 3). In each iteration, the unlabeled set $\mathcal{D}_U$ is updated by excluding the labeled set $\mathcal{D}_L$ from the initial pairwise slice pool $\mathcal{D}_S$ (line 4). The DL model f is then trained on the current labeled set $\mathcal{D}_L$ (line 5) and then used to identify the most informative slice pairs. Based on the DAL strategy Q , the model selects q new slice pairs from $\mathcal{D}_U$ to be labeled by an oracle (line 6). Any DAL query strategy can be used as Q , e.g., uncertainty-based, diversity-based, or hybrid AL methods. These selected pairs are then incorporated into $\mathcal{D}_L$, enriching it with new data for subsequent training iterations (line 7).

When LMI-AL needs to perform change detection on an unseen pair of 3D images $I_{n',k}^t$ and $I_{n',k}^{t'}$, it first slices both 3D images into 2D sections. For each corresponding pair of 2D slices from the baseline and follow-up images, LMI-AL computes a difference image, with special concern for whitened areas (tumors are white). The original slices and their difference images are then concatenated to form the model input. Each input consists of the baseline slice, the follow-up slice, and the difference image. The model then identifies areas $\hat{\mathbf{y}}_{n',k}^{t,t'}$ where lesions or other changes have occurred between the two time points on each slice. Finally, the model outputs a binary change map $m = \{\hat{\mathbf{y}}_{n',k}^{t,t'} \mid k \in \{1, \dots, c\}\}$ indicating the detected changes between the 3D images, allowing for detailed analysis of disease progression or treatment response.

IV. EXPERIMENT SETUP

A. Datasets

Multiple Sclerosis (MS) is an autoimmune neuro-degenerative disease that causes inflammation, damaging the myelin sheath around nerves and resulting in evolving lesions and progressive brain atrophy. In clinical practice, this progression is typically monitored through longitudinal brain MRI analysis [4]. Therefore, in this study, we use two MS longitudinal MRI datasets to evaluate our proposed method.

- **MSSEG-2 Dataset** [41]: MSSEG-2 is a publicly available dataset consisting of 200 MRI images from 100 MS patients, acquired from 15 different MRI scanners across multiple centers. Each patient was scanned at two time points, typically 1 to 3 years apart. Since not all the data in this dataset are labeled, we use only the MRI images from the 40 patients with available labels in this study.

- **MS Longitudinal MRI Dataset (SVUH)** [31]¹: This internal dataset was collected from St. Vincent's University Hospital in Ireland between 2019 and 2022, including only patients who had new lesions. It contains 2467 MRI images from 170 patients, with each patient undergoing up to 5 scanning sessions using 2 different scanners over a period ranging from 16 to 885 days. All data were labeled by two certified radiologists in their first year post-examination using ITK Snap. The labels were validated by a third expert radiologist specializing in neuroradiology.

Each dataset contains multiple sequences. Since change detection is similar to segmentation, and the FLAIR sequence is particularly well-suited for visualizing pathology, making it the most commonly used modality for segmentation tasks. Therefore, all the experiments are conducted using the FLAIR sequence. For MSSEG-2, the data are split into 28 patients for training (5,520 slice pairs), 6 for validation (767 slice pairs), and 6 for testing (1,074 slice pairs). For SVUH, the splits are 110 patients for training (29,638 slice pairs), 30 for validation (5,792 slice pairs), and 30 for testing (7,050 slice pairs). To ensure consistency and reliability of the model input, all slices are cropped to the largest brain cross-section, then normalized and resized to 256x256. Data augmentation techniques, including Gaussian blur, flipping, and rotation, are applied to each image in the training dataset.

B. Backbone Model and Active Learning Methods

We employ the UNet for biomedical image segmentation [22] from glass library² as our backbone model. Given the significant imbalance between the few target pixels and the many background pixels in each slice, focal loss [42] ($\alpha = 1, \gamma = 2$) is used to minimize the impact of this imbalance. The UNet is trained for up to 100 epochs with early stopping after 5 epochs and uses the Adam optimizer [43]. To evaluate how effectively DAL reduces labeling effort while maintaining change detection model performance, we compare the model under both DAL and supervised settings. The hyperparameters are tuned separately under supervised and DAL settings on each dataset to obtain optimal performance results. For the supervised setting, the learning rate and batch size are set to 0.0005 and 8, respectively, for both datasets. For AL setting, the learning rate is set to 0.0001 for both datasets, while the batch size is set to 8 for the MSSEG-2 dataset and to 4 for the SVUH dataset, respectively. The number of initial query slice pairs q_0 is set to 100 for MSSEG-2 and 500 for SVUH dataset. The number of query slice pairs in each subsequent iteration q is set to 50 for MSSEG-2 and 200 for SVUH.

We select eight well-known and widely used DAL methods, covering uncertainty-based (**LeastConfidence** [33], **Margin** [34], **Entropy** [8], **BALD** [36]), diversity-based (**Coreset** [44]), hybrid (**ClusterMargin** [38], **Hybrid Sampling** [39]), and the common baseline **Random** as the base AL query

¹<https://zenodo.org/records/8146426>

²<https://github.com/FrancescoSaverioZuppicchini/glasses/tree/master>

strategies Q in our proposed framework to evaluate their performance on longitudinal imaging.

1) *Least Confidence* [33]: selects samples for which the model has the lowest confidence in its most likely prediction. This is measured by the probability of the most probable label as predicted by the model.

2) *Margin Sampling* [34]: focuses on the difference between the first and second most probable predictions made by the model. Samples are chosen based on the smallest margins, indicating closely competing predictions.

3) *Entropy Sampling* [8]: is based on the uncertainty principle, where the selection criterion is the information entropy of the predictive probability distributions of the models. Samples that yield the highest entropy are considered the most uncertain and are thus prioritized for labeling.

4) *BALD-Dropout* [36]: exploits model uncertainty in a Bayesian framework. By utilizing dropout during inference as an approximation of Bayesian posterior, this method selects samples that maximize the mutual information between predictions and model parameters. In our experiment, Dropout is employed with the hyper-parameter n_drop set to 10.

5) *K-Center Greedy* [44]: is a core-set approach where the goal is to select samples that are representative of the entire dataset. It chooses samples that are farthest from the current set of labeled samples, effectively covering the input space more uniformly.

6) *Cluster Margin Sampling* [38]: combines Hierarchical Agglomerative Clustering (HAC) and margin sampling to enhance the selection process. The dataset is first clustered into several groups, and then margin sampling is applied within each cluster. This method ensures that the samples selected are not only uncertain but representative of the various data clusters, leading to a more balanced dataset for training. It is noteworthy that while this method has been employed for natural image classification tasks, HAC tends to be computationally inefficient when applied to pixel-level segmentation. Consequently, we have replaced HAC with the more conventional KMeans clustering to better suit our scenario. The hyper-parameter $n_clusters$ and $random_state$ for KMeans are set to 20 and 42, respectively. Additionally, the first round of selection focuses on diverse samples, choosing $n = \min(k*10, \text{len}(\text{unlabeled_idxs}))$, where k is the number of samples selected per round, followed by the uncertainty selection.

7) *Hybrid Sampling* [39]: integrate a Bayesian framework with a hybrid selection method that incorporates density (measured by cosine similarity) and diversity (quantified through mutual information) to select samples that exhibit both high uncertainty, density and diversity. The Bayesian is based on MC Dropout with n_drop set to 10. Similarly, for the initial step of uncertainty-based method selection, the top 500 most uncertain samples are chosen for subsequent selection based on diversity and density. The influence of diversity compared to density is in a ratio of 2:1.

8) *Random Selection*: involves selecting samples randomly without any heuristic or model-driven strategy. This method

serves as a control to assess the efficiency and effectiveness of the other more sophisticated active learning strategies.

C. Performance Metrics

Due to the similarity between segmentation and change detection, we use common segmentation evaluation metrics (e.g., precision, recall and dice coefficient) to assess the performance of change detection. Precision is the ratio of correctly identified changed voxels to all voxels predicted as changed, indicating the accuracy of the model in detecting true changes. Recall measures the ratio of correctly identified changed voxels to the total actual changed voxels, reflecting the model's ability to detect all true changes in the 3D data. The dice coefficient evaluates the overlap between the predicted changed voxels and the actual changed voxels, calculated as twice the intersection divided by the sum of the total predicted and actual changed voxels, reflecting the accuracy and completeness of the detection. All experiments are conducted three times, and the average performance across these runs are reported.

V. EXPERIMENT RESULTS

A. Fully Supervised Learning Result and Analysis

TABLE I: Fully Supervised Learning Performance Evaluation on MSSEG-2 and SVUH Datasets

Dataset	Data Used	dice	Rec	Prec
MSSEG-2	All slices	0.4973	0.6097	0.4467
	Slices with target	0.4379	0.6674	0.3685
SVUH	All slices	0.5907	0.1049	0.4109
	Slices with target	0.6473	0.2654	0.4596

Table I shows the overall supervised learning result for the two datasets. Our experimental evaluation demonstrates distinct performance characteristics when comparing the use of the entirety of available samples versus a focus solely on slices with targets in the MSSEG-2 and SVUH datasets for fully supervised learning.

For the MSSEG-2 dataset, training with all slices results in better overall metrics compared to using only slices with targets. This can be primarily attributed to the relatively small size of the training set, which necessitates a larger pool of samples to effectively capture the comprehensive distribution of the data. By utilizing all available slices, the model achieves an increase in the overall dice score of 0.4973 compared to 0.4379 with only target-contained slices, and precision increased by 0.08. However, using all slices creates an imbalanced dataset, as many of them do not include the target regions. This imbalance reduces the model's confidence in detecting target areas, as indicated by a drop in recall from 0.6674 to 0.6097.

Conversely, the larger SVUH dataset with more changed cases shows an advantage when focusing solely on slices with targets, which are large enough for model training, substantially reducing the degree of imbalance. This not only boosts the model's confidence in segmenting the target areas but also enhances the recall by 0.16 (0.1049 to 0.2654) and both the precision and dice by approximately 0.05.

B. Active Learning Result

Table II presents the experimental results for the MSSEG-2 and SVUH datasets with varying percentages of labeled slice pairs. LMI-AL achieves comparable performance to a DL trained on a fully labeled dataset with only 8% (e.g., LMI-AL(Random)) of labeled slice pairs for MSSEG-2 and 5% (e.g., LMI-AL(Coreset) and LMI-AL(ClusterMargin)) for SVUH. The results demonstrate that even with a small percentage of labeled data, LMI-AL strategies can achieve performance close to that of fully supervised models. The Highest dice scores achieved in the process show that all AL methods perform similarly (dice values within 0.06 of the best value on MSSEG-2 and 0.03 on SVUH dataset) except for Least Confidence and Margin. This could be because medical imaging datasets often suffer from class imbalance, with a large number of negative (no change) slice pairs compared to positive (change) slice pairs. Both of these methods use simple approaches to measure uncertainty and do not inherently address class imbalance, which may result in sub-optimal selection of positive samples for labeling.

To further evaluate the effectiveness of the proposed method, we compared the two best-performing DAL strategies for each dataset (e.g., Random and Coreset for MSSEG-2, ClusterMargin and Hybrid for SVUH) with and without integration into the LMI-AL framework. This comparison helps assess the actual improvements achieved by LMI-AL in change detection for longitudinal images. On MSSEG-2, both Random and Coreset strategies, when integrated with LMI-AL, show improved performance across different percentages of labeled slice pairs, with the highest increase observed at 8% labeled slice pairs (Random shows a 0.06 increase and Coreset shows a 0.07 increase in dice). Similarly, in SVUH, ClusterMargin and Hybrid also show performance gains with LMI-AL across different portions of labeled data.

C. Analysis and Discussion

Although supervised learning results across datasets are inconsistent (shown in Table I), the outcomes in the DAL settings reveal similar trends (shown in Table II). Here in this section, we provide corresponding analysis.

1) *Minimal Sample Sufficiency*: Both datasets demonstrate that a small portion of labeled data can achieve comparable performance compared to the supervised learning baseline (less than 8% of the data was required for the MSSEG-2 dataset and less than 5% for the SVUH dataset). **This efficiency may be attributed to the high redundancy inherent in the datasets.** To further demonstrate this, we visualized the distribution of slice pair selections made by the top-performing uncertainty- (LMI-AL(Entropy)), diversity- (LMI-AL(Random) and LMI-AL(Coreset)), and hybrid-based AL (LMI-AL(ClusterMargin)) methods on the MSSEG-2 dataset. Figure 2 shows representative results, where the number of target-containing pixels in each slice is depicted in shades of blue, with darker shades indicating more targets. Due to the different principles behind each query strategy, the distribution of selected slice pairs varies greatly. However, it

is interesting to note that their performance does not show significant differences (see Table II). We suppose this is due to the redundancy of longitudinal medical datasets (e.g., many slices with new lesions that are very similar in shape, size, and location), demonstrating the importance of DAL in this setting.

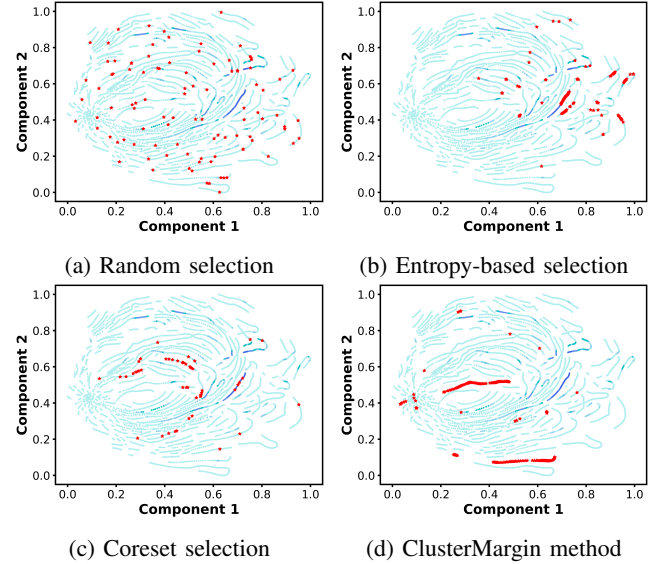


Fig. 2: Visualizations of sample selections by different active learning methods. Red points indicate the selected samples up to the point where performance matches the fully supervised baseline, while blue points represent all available unlabeled images in the pool.

2) *Suitability of Diversity-based and Hybrid Methods*: As shown in table II, the diversity and hybrid query strategies perform better in both two dataset. **The success of these methods likely relates to the limited budget of labeled samples necessary for training.** Since the required number of samples is relatively low, the models benefit significantly from learning a general distribution of the data [12]. Hybrid-based method that dynamically adjust between diversity and uncertainty-based methods could potentially harness the strengths of both strategies. While uncertainty-based methods typically become more relevant with increased labeling budgets, **our analysis suggests that in the context of MS longitudinal datasets, extensive budgets may not be necessary. Therefore, diversity-based and hybrid methods should be applied when a new longitudinal dataset is given.**

3) *Comparison between Active learning and Fully supervised learning*: In the MSSEG2 dataset, AL results in table II show that recall is comparable to that of supervised learning, while precision is on average higher. This can be attributed to the limited sample amount of MSSEG2, where supervised learning needs the full dataset to yield better performance. Despite achieving higher precision compared to using only target-containing samples, the dataset remains highly imbalanced (the area of target is quite small compared to background). However, **DAL methods can mitigate the imbalance problem.** For example, Coreset, as shown in Figure 2(c), even focusing

TABLE II: Model performance with varied label percentages for the MSSEG-2 (upper section) and SVUH (lower section) datasets. For MSSEG-2, supervised learning achieves 0.4973 (dice), 0.6097 (recall), and 0.4467 (precision). For SVUH, supervised learning results are 0.6473 (dice), 0.2654 (recall), and 0.4596 (precision). The best approach combined with LMI-AL in each column is highlighted in red, while the second-ranked approaches are in bold with a grey background. The methods in bold are the two best methods. The highest dice score is the highest performance achieved during the process, starting after the initial query to 8% for MSSEG-2 and 5% for SVUH.

Percentage of $\mathcal{D}_L$ in MSSEG-2	4%			6%			8%			Highest dice
Metrics	dice	recall	precision	dice	recall	precision	dice	recall	precision	
LMI-AL(Margin)	0.2637	0.3048	0.4561	0.3483	0.4355	0.3981	0.3575	0.3928	0.4093	0.3575
LMI-AL(LeastConfidence)	0.3471	0.4114	0.4261	0.3829	0.5533	0.3710	0.3707	0.4986	0.3682	0.3829
LMI-AL(Entropy)	0.4537	0.5571	0.4535	0.4209	0.4317	0.4897	0.4357	0.4797	0.4744	0.4537
LMI-AL(BALD)	0.3411	0.3620	0.4233	0.4205	0.5633	0.3892	0.4577	0.6015	0.3908	0.4205
LMI-AL(ClusterMargin)	0.4132	0.4338	0.5166	0.3886	0.4695	0.4079	0.4418	0.5264	0.4699	0.4446
LMI-AL(Hybrid)	0.4027	0.4735	0.4889	0.4215	0.5344	0.4465	0.3963	0.6015	0.4202	0.4620
LMI-AL(Random)	0.4016	0.4608	0.4663	0.4279	0.5410	0.4418	0.4849	0.6592	0.4269	0.4849
LMI-AL(Coreset)	0.4021	0.5165	0.4339	0.4206	0.5582	0.4636	0.4655	0.5339	0.4656	0.4655
Random	0.4165	0.4792	0.5277	0.4226	0.5969	0.4038	0.4241	0.4362	0.4740	0.4241
Coreset	0.4019	0.5128	0.4680	0.4544	0.5419	0.4590	0.3951	0.4572	0.5014	0.4544
Percentage of $\mathcal{D}_L$ in SVUH	3%			4%			5%			Highest dice
Metrics	dice	recall	precision	dice	recall	precision	dice	recall	precision	
LMI-AL(Random)	0.6272	0.1849	0.4222	0.5776	0.0814	0.4136	0.6152	0.1660	0.4401	0.6067
LMI-AL(Margin)	0.6114	0.1533	0.3410	0.5744	0.0746	0.4080	0.5839	0.0996	0.3830	0.5899
LMI-AL(LeastConfidence)	0.5572	0.0508	0.3935	0.5968	0.1239	0.4054	0.6111	0.1469	0.4054	0.5884
LMI-AL(Entropy)	0.6099	0.1464	0.3916	0.5996	0.1204	0.4344	0.6215	0.1953	0.4700	0.6103
LMI-AL(BALD)	0.6073	0.1261	0.4350	0.6275	0.1913	0.4203	0.6163	0.1525	0.4268	0.6170
LMI-AL(Coreset)	0.5927	0.1028	0.4256	0.6236	0.1817	0.4026	0.6322	0.1962	0.4412	0.6162
LMI-AL(ClusterMargin)	0.6180	0.1484	0.4346	0.6574	0.3003	0.4083	0.6288	0.2552	0.3969	0.6347
LMI-AL(Hybrid)	0.6125	0.1438	0.4395	0.6512	0.2207	0.4506	0.6094	0.1293	0.4406	0.6244
ClusterMargin	0.6016	0.1815	0.3263	0.6485	0.2483	0.4337	0.6149	0.2183	0.3253	0.6217
Hybrid	0.5927	0.1358	0.3618	0.6229	0.1552	0.4699	0.6080	0.1746	0.3425	0.6079

on diversity-based selection, the sample selection tends to focus on areas with more target pixels, reducing the level of imbalance and resulting in a precision of 0.4656, compared to 0.4467 in supervised learning.

On the other hand, using only target-containing samples in supervised learning demonstrates higher recall, reaching 0.6674 compared to 0.6097. AL achieves similar results using only 8% of the samples, where Hybrid and Random sampling reach 0.6959 and 0.6592, respectively. **This suggests that even a small number of blank and target-containing samples can perform comparably to using all target samples, likely due to the broader distribution of blank samples.**

The SVUH dataset yields similar conclusions. Even when using all samples with target for supervised learning, AL results achieve comparable or even superior outcomes, significantly surpassing the results obtained when using all slices (0.5907, 0.1049, 0.4109). This indicates that the SVUH data contains a substantial amount of redundancy, and that blank samples are beneficial to the model's learning process.

VI. CONCLUSION

In this paper, we investigate an important and previously unexplored problem: training a model for change detection in longitudinal medical imaging using limited labeled data. We propose a novel framework LMI-AL that adapts active learning for change detection in longitudinal imaging. The experimental results show that using less than 8% of labeled data allows the deep learning model to achieve performance similar to training with a fully labeled dataset, which indicates that labeling all

the data in longitudinal medical imaging is unnecessary. We also provide detailed analysis and give guidance for future related research. Considering the difficulties and high cost of labeling longitudinal medical images, our approach offers a practical solution to efficiently train models with minimal labeling effort, potentially reducing both time and resources in clinical applications. Future work could focus on analyzing the spatial distribution of the selected slices through visualization, revealing which regions are more frequently sampled. This may provide further insights into the model's learning behavior and patterns.

VII. ACKNOWLEDGMENT

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number [12/RC/2289_P2].

REFERENCES

- [1] Q. Zhao, Z. Liu, E. Adeli, and K. M. Pohl, "Longitudinal self-supervised learning," *Medical image analysis*, vol. 71, p. 102051, 2021.
- [2] R. Cui, M. Liu, A. D. N. Initiative *et al.*, "Rnn-based longitudinal analysis for diagnosis of alzheimer's disease," *Computerized Medical Imaging and Graphics*, vol. 73, pp. 1–10, 2019.
- [3] V. Cheplygina, M. De Bruijne, and J. P. Pluim, "Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Medical image analysis*, vol. 54, pp. 280–296, 2019.
- [4] E. Dufresne, D. Fortun, B. Kumar, S. Kremer, and V. Noblet, "Joint registration and change detection in longitudinal brain mri," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020, pp. 104–108.

- [5] D. Martin, E. Tong, B. Kelly, K. Yeom, and V. Yedavalli, "Current perspectives of artificial intelligence in pediatric neuroradiology: an overview," *Frontiers in radiology*, vol. 1, p. 713681, 2021.
- [6] N. M. Sepahvand, D. L. Arnold, and T. Arbel, "Cnn detection of new and enlarging multiple sclerosis lesions from longitudinal mri using subtraction images," in *2020 IEEE 17th international symposium on biomedical imaging (ISBI)*. IEEE, 2020, pp. 127–130.
- [7] G. Gerig, J. Fishbaugh, and N. Sadeghi, "Longitudinal modeling of appearance and shape and its potential for clinical use," pp. 114–121, 2016.
- [8] B. Settles, "Active learning literature survey," 2009.
- [9] F. Ozdemir, Z. Peng, P. Fuernstahl, C. Tanner, and O. Goksel, "Active learning for segmentation based on bayesian sample queries," *Knowledge-Based Systems*, vol. 214, p. 106531, 2021.
- [10] M. Shen, J. Y. Zhang, L. Chen, W. Yan, N. Jani, B. Sutton, and O. Koyejo, "Labeling cost sensitive batch active learning for brain tumor segmentation," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 1269–1273.
- [11] V. Nath, D. Yang, H. R. Roth, and D. Xu, "Warm start active learning with proxy labels and selection via semi-supervised fine-tuning," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2022, pp. 297–308.
- [12] S. Ma, H. Du, K. M. Curran, A. Lawlor, and R. Dong, "Adaptive curriculum query strategy for active learning in medical image classification [accepted]," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2024.
- [13] Y. He, A. Carass, Y. Liu, P. A. Calabresi, S. Saidha, and J. L. Prince, "Longitudinal deep network for consistent oct layer segmentation," *Biomedical optics express*, vol. 14, no. 5, pp. 1874–1893, 2023.
- [14] B. D. Basaran, P. M. Matthews, and W. Bai, "New lesion segmentation for multiple sclerosis brain images with imaging and lesion-aware augmentation," *Frontiers in neuroscience*, vol. 16, p. 1007453, 2022.
- [15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [16] C. Gulcehre, K. Cho, R. Pascanu, and Y. Bengio, "Learned-norm pooling for deep feedforward and recurrent neural networks," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I 14*. Springer, 2014, pp. 530–546.
- [17] Y. Xu, A. Hosny, R. Zeleznik, C. Parmar, T. Coroller, I. Franco, R. H. Mak, and H. J. Aerts, "Deep learning predicts lung cancer treatment response from serial medical imaging," *Clinical Cancer Research*, vol. 25, no. 11, pp. 3266–3275, 2019.
- [18] A. H. Ribeiro, K. Tiels, L. A. Aguirre, and T. Schön, "Beyond exploding and vanishing gradients: analysing rnn training using attractors and smoothness," in *International conference on artificial intelligence and statistics*. PMLR, 2020, pp. 2370–2380.
- [19] A. Dixit, J. Yohannan, and M. V. Boland, "Assessing glaucoma progression using machine learning trained on longitudinal visual field and clinical data," *Ophthalmology*, vol. 128, no. 7, pp. 1016–1026, 2021.
- [20] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.
- [21] P. Ashtari, B. Barile, S. Van Huffel, and D. Sappéy-Mariniér, "New multiple sclerosis lesion segmentation and detection using pre-activation u-net," *Frontiers in Neuroscience*, vol. 16, p. 975862, 2022.
- [22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.
- [23] S. Hitziger, W. X. Ling, T. Fritz, T. D'Albis, A. Lemke, and J. Grilo, "Triplanar u-net with lesion-wise voting for the segmentation of new lesions on longitudinal mri studies," *Frontiers in Neuroscience*, vol. 16, p. 964250, 2022.
- [24] B. Sarica and D. Z. Seker, "New ms lesion segmentation with deep residual attention gate u-net utilizing 2d slices of 3d mr images," *Frontiers in Neuroscience*, vol. 16, p. 912000, 2022.
- [25] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, "Face alignment in full pose range: A 3d total solution," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 1, pp. 78–92, 2017.
- [26] M. Salem, M. A. Ryan, A. Oliver, K. F. Hussain, and X. Lladó, "Improving the detection of new lesions in multiple sclerosis with a cascaded 3d fully convolutional neural network approach," *Frontiers in Neuroscience*, vol. 16, p. 1007619, 2022.
- [27] M. Schmidt-Mengin, T. Soulier, M. Hamzaoui, A. Yazdan-Panah, B. Bordini, N. Ayache, B. Stankoff, and O. Colliot, "Online hard example mining vs. fixed oversampling strategy for segmentation of new multiple sclerosis lesions from longitudinal flair mri," *Frontiers in Neuroscience*, vol. 16, p. 1004050, 2022.
- [28] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a siamese time delay neural network," *Advances in neural information processing systems*, vol. 6, 1993.
- [29] A. Fenneteau, P. Bourdon, D. Helbert, C. Fernandez-Maloigne, C. N. Habas, and R. Guillemin, "Siamese convolutional neural network for new multiple sclerosis lesion segmentation," in *MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure. MICCAI 2021*, 2021.
- [30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020.
- [31] B. S. Kelly, P. Mathur, R. P. Killeen, and A. Lawlor, "A siamese u-transformer for change detection on mri brain for multiple sclerosis, a model development and external validation study," *medRxiv*, pp. 2024–04, 2024.
- [32] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, B. B. Gupta, X. Chen, and X. Wang, "A survey of deep active learning," *ACM computing surveys (CSUR)*, vol. 54, no. 9, pp. 1–40, 2021.
- [33] D. D. Lewis, "A sequential algorithm for training text classifiers: Corrigendum and additional data," in *Acm Sigir Forum*, vol. 29, no. 2. ACM New York, NY, USA, 1995, pp. 13–19.
- [34] T. Scheffer, C. Decomain, and S. Wrobel, "Active hidden markov models for information extraction," in *International symposium on intelligent data analysis*. Springer, 2001, pp. 309–318.
- [35] Q. Qi, Y. Li, J. Wang, H. Zheng, Y. Huang, X. Ding, and G. K. Rohde, "Label-efficient breast cancer histopathological image classification," *IEEE Journal of biomedical and health informatics*, vol. 23, no. 5, pp. 2108–2116, 2018.
- [36] N. Houlsby, F. Huszár, Z. Ghahramani, and M. Lengyel, "Bayesian active learning for classification and preference learning," *stat*, vol. 1050, p. 24, 2011.
- [37] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in *International Conference on Learning Representations*, 2018.
- [38] G. Citovsky, G. DeSalvo, C. Gentile, L. Karydas, A. Rajagopalan, A. Rostamizadeh, and S. Kumar, "Batch active learning at scale," *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 933–11 944, 2021.
- [39] G. Li, Y. Otake, M. Soufi, M. Taniguchi, M. Yagi, N. Ichihashi, K. Uemura, M. Takao, N. Sugano, and Y. Sato, "Hybrid representation-enhanced sampling for bayesian active learning in musculoskeletal segmentation of lower extremities," *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–10, 2024.
- [40] G. Wang, M. A. Zuluaga, W. Li, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin *et al.*, "Deepigeos: a deep interactive geodesic framework for medical image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1559–1572, 2018.
- [41] O. Commowick, F. Cervenansky, F. Cotton, and M. Dojat, "Msseg-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure," in *MICCAI 2021-24th international conference on medical image computing and computer assisted intervention*, 2021, p. 126.
- [42] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.
- [43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [44] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in *International Conference on Learning Representations*, 2018.