Koita Center for Digital Health

Indian Institute of Technology, Bombay

Dual Degree Project 2 (DH 594)

Report

Data Efficient Machine Learning for Labelling Medical Images

Akshit Srivastava (180110008)

Date: 18-06-2023

Guide: Prof. Ganesh Ramakrishnan (CSE)

Co-Guide: Prof. Kshitij Jadhav (KCDH)

Approval Sheet

This thesis entitled **Data Efficient Machine Learning for Labelling Medical Images** by **Akshit Srivastava** is approved for the degree of **Dual Degree (B.Tech.+M.Tech.)**.

Examiners

Digital Signature
Ranjith Padinhateeri (i09078)
28-Jul-23 09:12:45 PM

Dr. Ranjith Padinhateeri

Supervisors

Digital Signature
Kshitij Jadhav (10002036)
28-Jul-23 09:06:49 AM

Dr. Kshitij Jadhav

Digital Signature
Ganesh Ramakrishnan (i09037)
28-Jul-23 09:09:39 AM

Dr. Ganesh Ramakrishnan

Date: 27th June, 2023

Place: IIT Bombay

https://ams.iitb.ac.in/d/115909-S2CE5WG08NXXVVRL

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Digital Signature
Akshit Srivastava (180110008)
28-Jul-23 12:42:25 AM

Akshit Srivastava

180110008

Date: June 27, 2023

https://ams.iitb.ac.in/d/115908-FMZPZ2YVD7UGBKX7

Abstract

The necessity of large amounts of labelled data to train deep learning models, especially in the medical imaging field creates a bottleneck for their implementation in resource-constrained environments. In this paper, we present **INSITE** (labell**IN**g medical image**S** us**I**ng submodular func**T**ions and s**E**mi-supervised data programming) where we apply representation-based subset selection to identify a small number of most representative images from a huge pool of unlabelled data which are then labelled by a medical domain expert. These newly labelled images are then used as exemplars to develop several data programming-driven labelling functions each of which output a label and a similarity score when given an unlabelled image as an input. A consensus is brought about by using a label aggregator function to assign the final label to each unlabelled data point. Labelling subsequent to representation-based subset selection was higher in accuracy compared to randomly selected subsets. Further, labelling accuracy following Continuous and Quality-Guided Labelling Function aggregator was higher than state-of-the-art RESNET-18 Convolutional Neural Network when the budget for labelling was extremely small. This demonstrates that representation-based subset selection followed by semi-supervised data programming approaches in a resource-poor setting performs better than data-hungry deep models.

List of Contents

INSITE: Labelling Medical Images Using Submodular Functions and Semi-Supervised Data Programming

1. Introduction

In today's rapidly evolving healthcare industry, an abundant amount of data is being generated, particularly in the realm of medical imaging. This data surge can be attributed to the remarkable progress in data acquisition techniques like magnetic resonance imaging (MRI), computed tomography (CT) scans, and whole slide images obtained from pathology sections [13] [22] [12]. As a result, the medical domain is widely recognized as the next frontier for artificial intelligence (AI) and machine learning (ML), with Deep Neural Networks (DNNs) anticipated to play a pivotal role. DNNs have already revolutionized various domains including image recognition, video recognition, natural language processing, recommendation systems, finance, gaming, robotics, climate change, and transportation, among others [2].

This anticipation is well-founded considering that DNNs excel at learning intricate non-linear relationships between input variables and outputs. This is made possible through automated feature extraction, which reduces the dependence on manual feature engineering and significantly enhances the efficiency of predictive analysis [12]. Moreover, DNNs exhibit remarkable capabilities in operating within high dimensional spaces, while also leveraging transfer learning techniques to train for one task and adapt to another [6]. However, notwithstanding the impressive advancements made by Deep Neural Networks (DNNs), there exist several notable disadvantages that have significantly impeded their widespread adoption, particularly in resource-constrained settings with limited access to advanced technologies and computational resources. One of the primary concerns lies in the substantial computational power demanded by these networks, especially during the resource-intensive training phase, which not only makes them prohibitively expensive but also considerably slows down the overall learning process [7]. Another significant limitation that hampers the effective utilization of DNNs is their insatiable hunger for extensive amounts of data for accurate training. This entails the imperative need for equal representation across various datasets, thus necessitating the availability of vast quantities of meticulously labeled data. Regrettably, this requirement is often overlooked and underappreciated, leading to challenges and setbacks in the practical implementation of DNNs [21]. A particularly noteworthy predicament arises when considering the labeling process itself, where human annotators play a crucial role in providing the "ground truth" necessary for supervised machine learning tasks. In the context of DNNs, the accurate labeling of data becomes an arduous and time-consuming task, significantly contributing to the high costs associated with their adoption, especially in resource-constrained settings and low- and middle-income countries [14]. It is important to acknowledge that data labeling, typically performed by skilled individuals, is a laborious affair that requires meticulous review and precise annotation of each data point. The magnitude of this challenge is further amplified in the medical

domain, where the need for expert-level annotation and consensus among multiple medical professionals adds an additional layer of complexity and cost [25].

To exemplify the substantial efforts required in the medical domain, one can consider the ImageNet Large-scale Visual Recognition Challenge (ILSVRC) [20], a renowned competition that focuses on object classification, localization, and detection in everyday images. In this competition, approximately 1 million labeled images are made available through crowd-sourced annotations provided by non-experts. However, when juxtaposed with computational biology competitions [15] [24], which typically offer only a few hundreds to thousands of labeled images due to the scarcity of expert annotators, the challenges and limitations become glaringly evident. It is crucial to underscore the indispensability of human-labeled data, as it integrates nuanced human knowledge and expertise into machine learning models. This significance is amplified manifold in the medical domain, where the confluence of medical expertise and knowledge accumulated over several years is distilled and consolidated to achieve a consensus among multiple experts. Consequently, the labeling of medical domain datasets becomes an expensive and resource-intensive endeavor, exacerbating the already substantial costs associated with the adoption of DNNs, particularly in resource-constrained settings and low- and middle-income countries [25].

Now, presence of large reservoirs of unlabelled data could be considered as an impediment to applications of DNNs. However, recent advances in data programming techniques can be utilized to label unlabelled data since it combines human knowledge with weak supervision to generate labels for unlabelled data [18]. This process involves the creation of labelling functions or rules to as- sign labels driven by heuristics or expert domain knowledge. While labelling functions may not be accurate by themselves, they identify some signal in the datapoint of interest and then the outcome of several such labelling functions are aggregated using a generative model to assign the final label [1] [18]. Thus, Data programming could facilitate the labelling of large unlabelled datasets especially in the medical domain since medical expert driven labelling rapidly becomes expensive and untenable. Data programming driven annotation could show addition improvement if we introduce a human- in-the-loop system for annotating a small amount of data which could then work as exemplars for developing nuanced labelling functions.

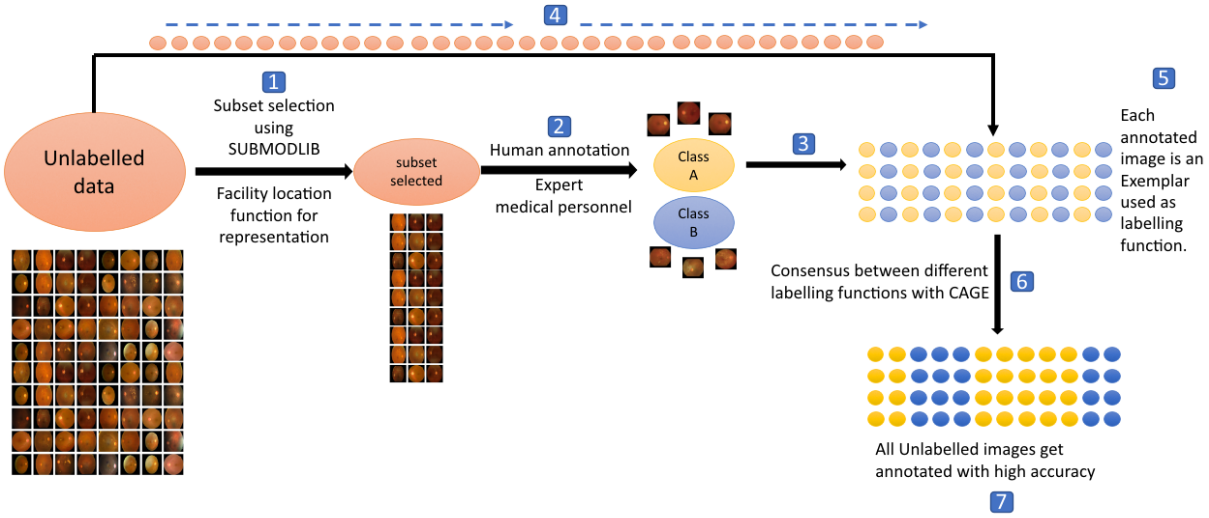

Figure 1: Labelling with Insite: We use Facility Location for subset selection, similarity scores with exemplars as labelling functions and CAGE [3] for label aggregation

2. Our Contributions

A large amount of unannotated medical data is often available in low resource settings which if accurately labelled can contribute to generating novel insights in the therapeutics and diagnostic domain. However, given the high cost of labelling which is further escalated in the medical field, labelling of the entire unlabelled dataset in re- source poor settings is untenable. Moreover, deep learning models like CNNs, which are being used extensively in the field of medical image computing, have millions of trainable parameters, which means that they require a significant amount of computation to learn. This can be a major drawback, especially when working with limited computational resources. For instance, ResNet-18 has 11.689 million trainable parameters even though it is a relatively simple model using residual connections [8]. Here, we address these problems by demonstrating that intelligent representation driven subset selection even with a small budget for annotation by medical experts can assist in accurately labelling the larger pool of unlabelled data (c.f. Figure1). To this end we applied representation- based subset selection to identify the most representative subset of images from a large pool of unlabelled medical image dataset compared to random subset selection (c.f. Section 2). This was done using submodular functions which follow the principle of diminishing returns with increasing sample sizes (c.f. Section 4.1). Thus, by applying representation-based subset selection we identify the smallest budget which can get representation of all available classes and these are then annotated by a domain medical expert. We then use these expert-annotated images as exemplars to derive similarity metric based labelling functions (c.f. Section 4.2) which output a label (same as the gold label) and a similarity score (c.f. Section 4.4). Further, we aggregated the outputs of these labelling functions using Continuous and Quality-Guided Labelling Function aggregator [3] (c.f. Section 4.3). Our approach, Insite demonstrates that the label accuracy compared to the gold truth was higher than when the same annotated images are trained on state of the art CNNs such as RESNET18 (c.f. Figures 4, 6 and 8). Thus, we effectively demonstrate that even with a small budget for labelling which is a realistic problem in the medical field especially in

resource-constrained settings Insite can help label large numbers of unlabelled images with a high degree of accuracy.

3. Related Work

Peikari et al. [18] propose a semi-supervised learning method for labelling medical images. The method first analyzes groups of labeled and unlabeled points in multidimensional feature space to identify areas of high density. The learning method is then guided to place decision boundaries through the regions with low density. This technique was applied to the analysis of digital pathology images of breast cancer.

Liu et al. [16] present their medical image classification framework PLAB-GAN (Pseudo-labelling GAN). The framework first clusters the unlabeled samples to the cluster centers of the labeled images. Pseudo labels are then estimated based on the CNN features extracted from the samples using a pre-trained ResNet-20 network. From each cluster, a small number of labeled data and a greater number of unlabeled data are selected to train the discriminator/classifier. A new class is further added to the discriminator output for the synthetic data so that the synthetic images can be classified into the K+1 category (K classes for the real data and 1 pseudo-class for the synthetic data). The trained discriminator is then used for medical image classification.

Gu et al. [5] identify that the performance bottleneck of SemiSupervised Random Forest (SS-RF) under insufficient data is the biased information gain calculation when selecting an optimal splitting parameter. They modify the training procedure of SS-RF to relieve this bias by replacing the original information gain with graph-embedded entropy. Graph-embedded entropy exploits the data structure of unlabeled data by using both labeled and unlabeled data to construct a graph whose weights measure local similarity among data. The loss function is then minimized by summing the supervised loss over labeled data and a graph Laplacian regularization term.

Liu et al. [15] propose the anti-curriculum pseudo-labelling (ACPL) method. ACPL first selects the most informative unlabeled images using the proposed cross-distribution sample informativeness. A pseudo-labelling mechanism is then suggested that combines the model classification with a K-nearest neighbor (KNN) classification guided by sample informativeness. The most informative pseudolabelled samples are then selected to be included in the labeled anchor set to improve the pseudo-labelling accuracy of the KNN classifier.

4. Methods

We use submodular functions to identify a small number of most representative images from a huge pool of unlabelled data which are then labelled by a medical domain expert. These newly labelled images are then used as exemplars to develop several data programming-driven labelling functions each of which output a label and a similarity score when given an unlabelled image as an input. A consensus is brought about by using a label aggregator function to assign the final label to each unlabelled data point.

5.1. Submodular functions

Submodular functions are a mathematical concept that describes functions with diminishing returns or decreasing marginal gains. These functions are defined over sets and satisfy a property called submodularity, which states that the marginal gain of adding an element to a smaller set is greater than or equal to the marginal gain of adding the same element to a larger set. This property captures the idea that the value obtained from adding an item to a set diminishes as the set grows larger.

Submodular functions have diverse applications in optimization, set cover problems, and machine learning. They are used to model problems where there is a limited benefit to adding additional elements to a set once certain elements have already been included. This property allows for efficient optimization algorithms and enables the use of greedy algorithms, which iteratively select elements that provide the maximum marginal gain.

In the field of machine learning, submodular functions have proven useful for tasks such as feature selection, data summarization, active learning, and recommendation systems. They capture notions of diversity, coverage, and information gain, making them valuable for solving problems that involve selecting a representative subset or making efficient choices among a large set of options. Overall, submodular functions provide a mathematical framework for understanding and solving problems with diminishing returns. Their applications span various domains, offering efficient optimization techniques and enabling intelligent decision-making in complex settings.

To select a subset from a dataset we used submodular functions, each of which has different properties. Examples of submodular functions include facility location, graph cut, log-determinants, etc. These submodular functions are part of the submodlib library. (https://github.com/decile-team/submodlib) [18]. After having tested several diversity based submodular functions we zeroed on the Log-determinant function to select the diverse set of images from the given dataset given its higher efficacy in our algorithm. The assumption is that diversity based subset selection will be able to capture all the selective but distinct subsets from the data.

5.2. Data Programming

Let X denote the space of input instances and $Y = \{0, 1, \ldots, K\}$ denote the space of labels and $P(X, Y)$ denote their joint distribution. The goal is to learn a model to associate a label y with an example $x \in X$. Let the sample of n unlabelled instances be $x_1, x_2, \ldots, x_n$. Instead of the true y's we are provided a set of n labelling functions (LFs) $\lambda_1, \lambda_2, \ldots, \lambda_n$ such that each LF λ_j can be either discrete or continuous. Each LF λ_j is attached with a class k_j and on an instance x_i outputs a discrete label $\tau_{ij} = l_j$. If λ_j is continuous, it also outputs a score $s_{ij} \in (0, 1)$. To learn the true label y, we use a generative model to aggregate heuristic labels by creating a consensus among outputs of LFs. The generative model imposes a joint distribution between true label y and the values l_{ij}, s_{ij} returned by any LF λ_j on any data sample Xi drawn from the hidden distribution $P(X, Y)$. [3] define the hidden distribution as -

$$P(y, l_{ij}, s_{ij}) = \frac{1}{Z_\theta} \prod_{j=0}^{k-1} \psi_\theta(l_{ij}, y)\psi_\pi(l_{ij}, s_{ij}, y) \qquad (1)$$

where θ, π denote the parameters used in defining the potentials ψ_θ , ψ_π coupling the discrete and continuous variables respectively. All LFs are independent of each other since each similarity score is calculated separately. For designing the potentials, [3] suggest that:

$$\psi_\theta(l_{ij}, y) = \begin{cases} \exp(\theta_{jy}) & \text{if } l_{ij} \neq 0 \\ 1 & \text{otherwise} \end{cases} \qquad (2)$$

$$\psi_\pi(l_{ij}, s_{ij}, y) = \begin{cases} Beta(s_{ij}; \alpha_a, \beta_a) & \text{if } k_j = y \;\&\; l_{ij} \neq 0, \\ Beta(s_{ij}; \alpha_d, \beta_d) & \text{if } k_j \neq y \;\&\; l_{ij} \neq 0, \\ 1 & \text{otherwise} \end{cases} \qquad (3)$$

where *Beta* density is expressed in terms of $\alpha, \beta > 0$ as $P(s|\alpha, \beta) \propto s^{\alpha-1}(1 - s)^{\beta-1}$. [3] replace α, β with alternative parameters such that $\alpha_a = q^c_j \pi_{jy}$ and $\beta_a = (1 - q^c_j)\pi_{jy}$ are parameters of the agreement distribution, and $\alpha_d = (1 - q^c_j)\pi_{jy}$ and $\beta_d = q^c_j \pi_{jy}$ are parameters of the disagreement distribution. With these potentials, normalizer Z_θ of our joint distribution (Eqn 1) can be calculated as -

$$Z_\theta = \sum_{y \in Y} \prod_j (1 + \exp(\theta_{jy})) \qquad (4)$$

For training the aggregation model, we maximise likelihood on the observed τ_i and s_i values of the training sample $D = x_1, \ldots, x_m$ after marginalising out the true y. The training objective can be expressed as -

$$\max_{\theta,\pi} LL(\theta, \pi|D) \tag{5}$$

where,

$$LL(\theta, \pi|D) = \sum_{i=1}^{m} \log \sum_{y \in Y} P_{\theta,\pi}(\tau_i, s_i, y)$$

$$= \sum_{i=1}^{m} \log \sum_{y \in Y} \sum_{j=1}^{n} \psi_j(\tau_{ij}, y)\psi_j(s_{ij}, \tau_{ij}, y)\text{cont}(\lambda_j) - m \log Z_\theta \tag{6}$$

5.3. Our Approach

Let X represent the set of input images, and $Y = 0, 1, \ldots, K$ represent the set of labels in a classification task. With a limited budget b, we employ the facility location approach to identify a subset of size b that best represents the dataset. This subset is carefully labelled by a medical domain expert and denoted as $L = x1, x2, \ldots, xb$. The remaining unlabeled images form the set U, such that $X = U \cup L$. To facilitate the labelling process, b labelling functions (LFs) are created, denoted as $\lambda 1, \lambda 2, \ldots, \lambda b$. Each LF λi is associated with a corresponding image $xi \in L$. Given an input image xi , the LF λi provides two outputs: the label li associated with xi , and the cosine similarity score $si \in (0, 1)$ between the features extracted by the RESNET-18 model for images xi and xl . For each image $xu \in U$, we obtain b labels ($lu1$,$lu2$, . . . ,lub) and similarity scores ($su1$, $su2$, . . . , sub) from the corresponding LFs. To infer the true label yu for an image xu, we employ a generative model that aggregates heuristic labels by achieving a consensus among the outputs of the LFs. This consensus-based approach aims to improve the accuracy of label assignment for the unlabeled images in U.

5.4. Algorithms for our method and baselines

Algorithm 1 Insite: Facility location-based subset selection, labelling using CAGE

Require: Initial Image Set X, Budget: B, Labelled Set: $L = \emptyset$

1: Facility Location function $f(L) = \sum_{i \in X} \sum_{j \in L} d_{ij}$, where d_{ij} is the distance between facilities i and j
2: $L = \text{argmax}_{L \subseteq X, |L| \leq B} \sum_{x \in L} f(x)$
3: $U = X \setminus L$
4: Domain Expert creates T, such that t_i is the label of $x_i \in L$
5: Extract features $\text{vec}(x)$ for each image $x \in L$ using ResNet-18
6: **for** LF creation round, $i = 1 : B$ **do**
7: $\lambda_i(u \in U)$ outputs 2 values, $s_{ui} = 1 - \frac{\text{vec}(x_i) \cdot \text{vec}(u)}{|\text{vec}(x_i)||\text{vec}(U)|}$ & $l_{ui} = t_i$, the label associated with i
8: **end for**
9: **for** Labelling round, $u = 1 : |U|$ **do**
10: After iterating through all LFs $(\lambda_1, \lambda_2, \ldots, \lambda_b)$, $l_u = \{l_{u1}, l_{u2}, \ldots, l_{ub}\}$, $s_u = \{s_{u1}, s_{u2}, \ldots, s_{ub}\}$
11: As shown in **??**, Assuming y is the true label, as the joint distribution is given as -

$$P(y, l_{uj}, s_{uj}) = \frac{1}{Z_\theta} \prod_{j=1}^{b} \psi_\theta(l_{uj}, y) \psi_\pi(l_{uj}, s_{uj}, y) \tag{7}$$

12: Train ψ_θ, ψ_π to achieve $\max_{\theta,\pi} LL(\theta, \pi | D)$
13: Return y
14: **end for**

Algorithm 2 `Baseline 1: Facility location-based subset selection, labelling using CNN`

Require: Initial Image Set X, Budget: B, Labelled Set: $L = \emptyset$
Facility Location function $f(L) = \sum_{i \in X} \sum_{j \in L} d_{ij}$, where d_{ij} is the distance between facilities i and j
$L = \text{argmax}_{L \subseteq X, |L| \leq B} \sum_{x \in L} f(x)$
$U = X \setminus L$
Domain Expert creates T, such that t_i is the label of $x_i \in L$
Load the pre-trained ResNet-18 model
Train the model using L as training data
Get final labels using U as test data

Algorithm 3 `Baseline 2: Random subset selection, labelling using CNN`

Require: Initial Image Set X, Budget: B, Labelled Set: $L = \emptyset$
for Selection round, $i = 1 : B$ **do**
Select random image from X and add to L
end for
$U = X \setminus L$
Domain Expert creates T, such that t_i is the label of $x_i \in L$
Load the pre-trained ResNet-18 model
Train the model using L as training data
Get final labels using U as test data

5. Results and Discussion

In this section, we discuss the results of our framework on several medical imaging datasets. The results demonstrate that Insite is able to achieve high accuracy even with 1% labelled data. We hypothesise that this is due to the fact that we select the most representative images driven by nuanced representation based subset selection. We then utilize these images which are first annotated by medical domain experts which are then used to train our label aggregation model that can generalise over the entire dataset. Our work has several implications for the field of image classification. First, it shows that it is possible to achieve high accuracy with a small number of labelled images. This is important for tasks such as medical image classification, where it can be difficult and expensive to obtain labelled data. Second, our work exhibits that facility location can be used to select informative images for labelling. This is a promising new approach for improving the performance of downstream image classification algorithms. For all the experiments, we use the same training setup, including the hyperparameters. For feature extraction, we use the RESNET-18 model, which extracts a feature vector of size 512. We use the facility-location submodular function for subset selection. We keep the budget B around 1-2 % of the original dataset. We then create similarity-based labelling functions whose labels are then aggregated by our aggregation model, that has a learning rate = 0.01, and trains for 100 epochs.

6.1. Dataset: APTOS 2019 Diabetic Retinopathy

The APTOS 2019 Blindness Detection Challenge [9] included a Diabetic Retinopathy dataset, which was used as the benchmark dataset for this experiment. This dataset comprises retinal images captured using fundus photography. The images are categorised into five classes: normal (class 0: 1,805 images), mild (class 1: 370 images), moderate (class 2: 999 images), severe (class 3: 193 images), and proliferative (class 4: 295 images). For our experiment, we combined images from classes 1, 2, 3, and 4 to create a single diseased class (class 1: 1,857 images) of Diabetic Retinopathy. Figure 2 depicts the class distribution of the APTOS dataset for the purposes of our experiment. Figure 3 compares the accuracies of Insite (Facility location + CAGE) and baselines on the APTOS dataset on different budgets for subset selection. For the baselines, subsets are selected either randomly or using the facility location submodular function and then used as training data for the pre-trained RESNET-18. When the budget is 20-30 images in total, Insite outperforms all baselines. As the budget raises to 40, RESNET trained on images selected using facility location slightly outperforms Insite, while RESNET trained on a random subset of images always performs lower than Insite. This indicates that random subset selection before label aggregation does not yield good results. We postulate that as the budget of the number of images for labelling increases wider and more diverse feature representations become available that can be learned by the RESNET netowrk effectively compared to our method. However, smaller budgets cause an hindrance to RESNET learning all representations thereby not being able to generalize effectively over wider unlabelled data populations. This situation is not

observed if we implement Insitefor smaller budgets and our method is able to generalise our wider feature dsitributions as exemplified by better performance than RESNET.

Insite achieves an accuracy of 89% with a budget of 20 images, 92% with a budget of 30 images, and maintains its accuracy at 92% with a budget of 40 images. Facility location-based subset selection + CNN achieved an average accuracy of 88.33%, 91.33%, and 92.67%, respectively, while random subset selection + CNN achieved an average accuracy of 84.67%, 87.33%, and 78%, respectively. It is important to note that the accuracy of the CNN varied depending on the initialization of the model. However, Insite was insensitive to initialization and consistently achieved the same high accuracy. This also demonstrates the robustness and reliability of Insite over CNN especially scenarios of smaller budgets of labelling. Given a constrained environment where only 20-30 images (1% of the APTOS dataset) can be labelled by a domain expert, Insite is a more efficient algorithm for subset selection for labelling images. Even for higher budgets, when RESNET was trained on the subset of images selected by the facility location function being used by Insite, it consistently achieved a higher accuracy than when it was trained on a randomly selected subset of images. This suggests that selecting subsets of images in a programmatically intelligent fashion can improve the accuracy of any labelling framework, when compared to selecting subsets randomly. Thus, Insite is especially effective for medical imaging classification tasks where the spatial relationships between the images are important, as it compares cosine-similarities of images while labelling.

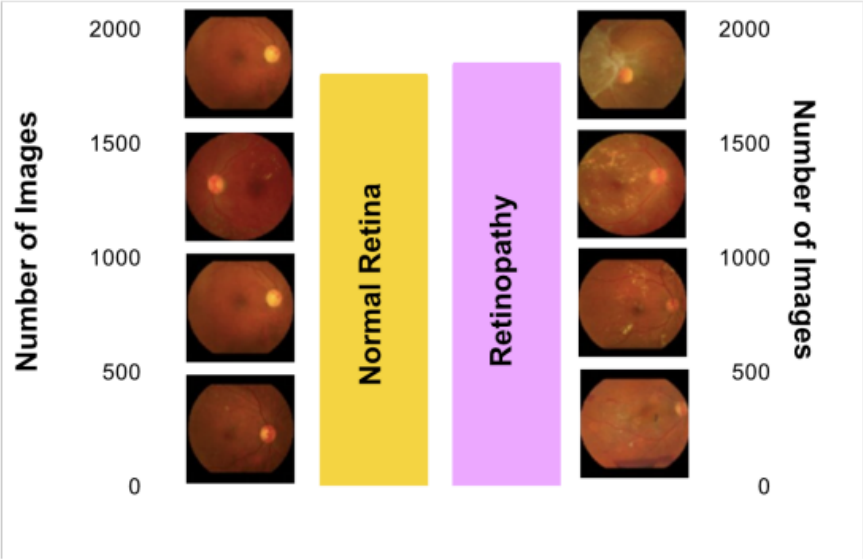

Figure 2: Class Distribution for APTOS Diabetic Retinopathy

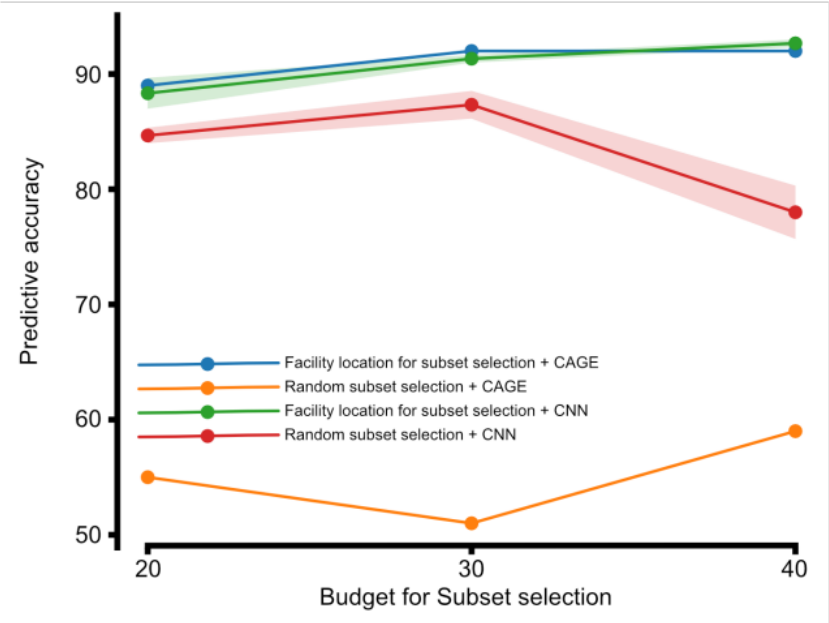

Figure 3: Accuracies of Insite and baseline algorithms for different budgets on APTOS dataset

6.2. Dataset: Chest X-Ray Images

Daniel S. et al [10] curated a large and diverse dataset of 5,856 Chest X-Ray images for diagnosis of pediatric pneumonia. The dataset is divided into 2 classes - Normal and Pneumonia. For our experiment we used 1341 images from normal class and 1341 images from Pneumonia class, giving a total of 2682 images. Figure 4 illustrates the class distribution of the modified Chest X-Ray dataset. Figure 5 compares the accuracies of Insite and baselines on the Chest X-Ray dataset on different budgets for subset selection. For the baselines, subsets are selected either randomly or using the facility location submodular function and then used as training data for the pre-trained RESNET-18.

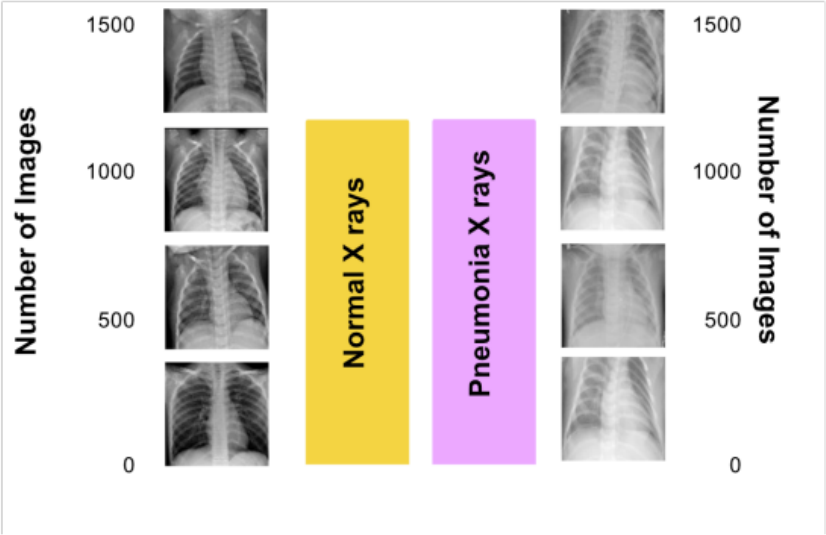

Figure 4: Class Distribution for Chest X-Ray dataset

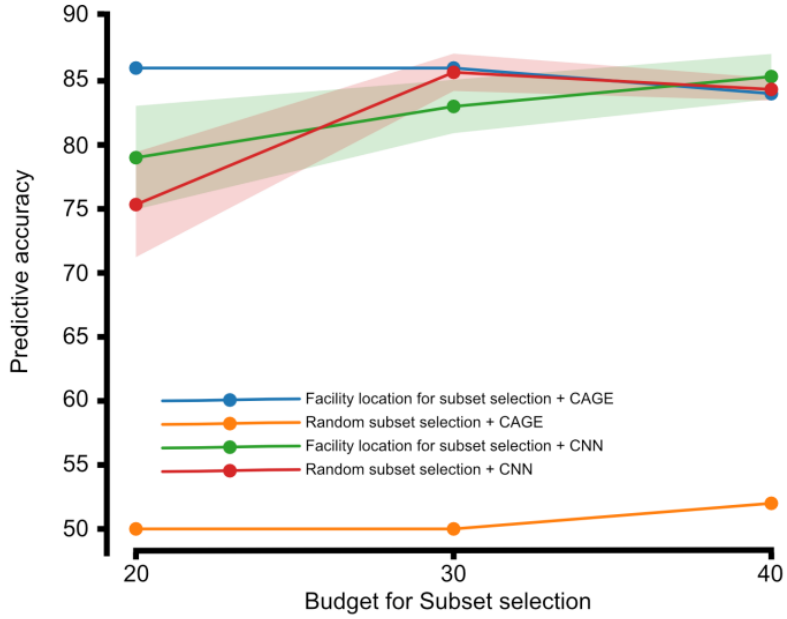

Figure 5: Accuracies of Insite and baseline algorithms for different budgets on Chest X-Ray dataset

For a budget of 20 images, Insite outperforms all baselines by a huge margin of >10%. RESNET trained on a random subset of images performs poor compared to when trained on facility location based selected subset. As the budget raises to 30, RESNET trained on images selected using facility location improves its performance and approaches the accuracy of Insite. For a budget of 40 images, RESNET slightly outperforms Insite, irrespective of the subset being selected randomly or using facility location.

Insite yields an accuracy of 86% with a budget of 20 images, maintains it at 86% when the budget increases to 30 images, but drops to 84% on a higher budget of 40 images. Facility location based subset selection + CNN achieved an average accuracy of 79%, 83%, and 85.33%, respectively, while random subset selection + CNN achieved an average accuracy of 75.33%, 85.67%, and 84.33%, respectively.

When only 20 images (<1% of the Chest X-Ray dataset) can be labelled by a domain expert, Insite performs much better than baselines. RESNET trained on the subset of images selected by facility location performs better than the one trained on a randomly selected subset of images as well. When the budget doubles to 40, both baselines marginally outperform Insite, signifying that although our framework performs better in an even more resource constrained scenario, given enough data points, RESNET will outperform Insite for higher budgets - irrespective of the images being selected randomly or via facility location.

6.3. Dataset: HAM10000

The HAM10000 ("Human Against Machine with 10000 training images") dataset [22] is a collection of 10015 dermatoscopic images from different populations, acquired and stored by different modalities. It contains a total of seven different classes namely actinic keratoses and intraepithelial carcinoma, basal cell carcinoma, benign keratosis-like lesions, dermatofibroma, melanoma, melanocytic nevi, and vascular lesions. We used images from 2 classes - melanoma (1113 images) and melanocytic nevi (1113 images). Figure 6 illustrates the class distribution of our modified dataset.

Figure 7 compares the accuracies of Insite and baselines on the HAM10000 dataset on different budgets for subset selection. In this experiment, for a budget of 20 images, Facility location-based subset selection + CNN performs better than Insite on average, but is still sensitive to initialisation, giving a standard error of 2%. When provided with 30 labelled images, Insite performs better than RESNET trained on images selected using facility location. When trained on a random subset, RESNET slightly outperforms Insite on average while still being prone to a standard error of 1.5%. For a budget of 40 images, RESNET with facility location slightly outperforms Insite while RESNET trained on random subset performs poorly in predicting the labeling of unlabelled data.

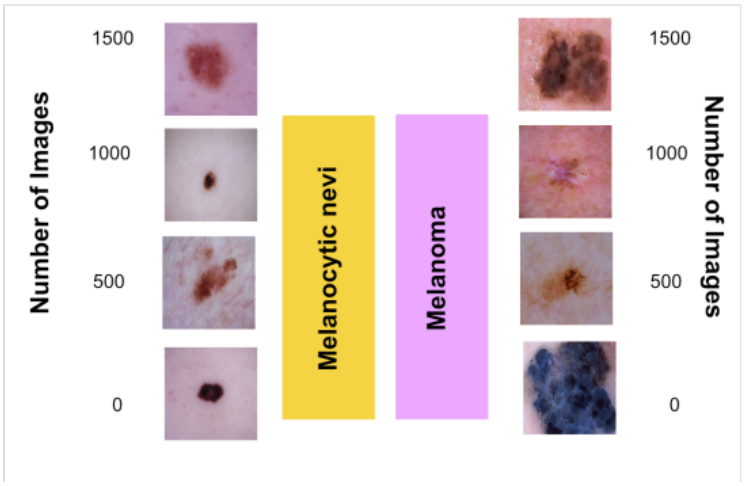

Figure 6: Class Distribution for DermaMNIST dataset

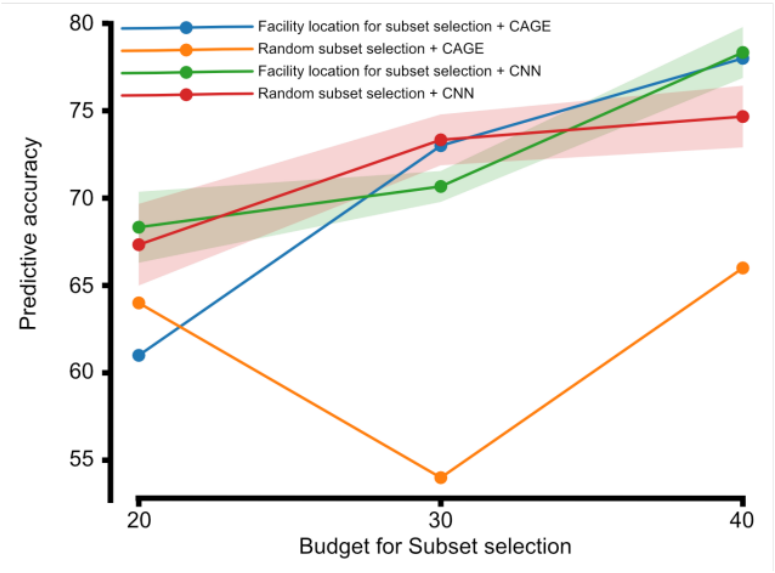

Figure 7: Accuracies of Insite and baseline algorithms for different budgets on Chest X-Ray dataset

Insite achieves an accuracy of 61% with a budget of 20 images, goes up to 73% when the budget increases to 30 images, and then further rises to 78% on a higher budget of 40 images. Facility location-based subset selection + CNN achieved an average accuracy of 68.33%, 70.67%, and 78.33%, respectively, while random subset selection + CNN achieved an average accuracy of 67.33%, 73.37%, and 74.67%, respectively. For budgets 20 and 40, Facility location + CNN performs the best. While for a budget of 30, Random + CNN performs slightly better than Insite.

6. Conclusion

A large amount of medical imaging data is being currently generated in low and middle income countries (LMICs) which due to paucity of medical experts and high cost of labelling continue to be under-utilized for generating machine learning driven insights. Moreover, training CNNs is computationally expensive since they require learning millions of parameters, requiring more time and While it can be argued that large labelled data is available in the western world and could be used across the globe, given the demographic differences and region specific variability especially in the medical domain, locally generated annotated data is necessary for developing specific insights in LMICs. Our framework Insite, inches towards addressing this problem of having large unannotated data in a resource poor setting. We demonstrate that, if there is a small budget to label the data, one can perform programmatically driven unsupervised subset selection to identify a representative sample of images from a pool of unlabelled dataset. These selected small number of images can then be labelled by domain medical experts even in resource constrained situations. These newly labelled images can now become exemplars and be used as labelling functions such that each unlabelled image will now be compared to these exemplars with a similarity score. These multitudes of similarity scores can then be aggregated by label aggregators and we demonstrate that with even 1% of the budget, Insite achieves higher accuracies compared to state of the art CNN models such as RESNET18. Importantly, in scenarios of very low budgets for labelling, random subset selection is not an efficient way of selecting images and would mainly cause wastage of resources since models developed on the top of these do not perform well for annotating unlabelled images. Thus, looking forward, intelligent representation based subset selection could be one of the ideal solutions to budget labelling and then annotating large pool of unlabelled images in resource poor settings.

7. References

[1] Guttu Sai Abhishek, Harshad Ingole, Parth Laturia, Vineeth Dorna, Ayush Maheshwari, Rishabh Iyer, and Ganesh Ramakrishnan. 2021. SPEAR: Semisupervised Data Programming in Python. arXiv preprint arXiv:2108.00373 (2021).

[2] Richard Baraniuk, David Donoho, and Matan Gavish. 2020. The science of deep learning. Proceedings of the National Academy of Sciences 117, 48 (2020), 30029–30032.

[3] Oishik Chatterjee, Ganesh Ramakrishnan, and Sunita Sarawagi. 2019. Data Programming using Continuous and Quality-Guided Labeling Functions. CoRR abs/1911.09860 (2019). arXiv:1911.09860 http://arxiv.org/abs/1911.09860

[4] Satoru Fujishige. 2005. Submodular functions and optimization. Elsevier.

[5] Lin Gu, Xiaowei Zhang, Shaodi You, Shen Zhao, Zhenzhong Liu, and Tatsuya Harada. 2020. Semi-Supervised Learning in Medical Images Through GraphEmbedded Random Forest. Frontiers in Neuroinformatics 14 (nov 2020). https: //doi.org/10.3389/fninf.2020.601829

[6] Jaya Gupta, Sunil Pathak, and Gireesh Kumar. 2022. Deep Learning (CNN) and Transfer Learning: A Review. In Journal of Physics: Conference Series, Vol. 2273. IOP Publishing, 012029.

[7] Pedram Havaei, Maryam Zekri, Elham Mahmoudzadeh, and Hossein Rabbani. 2023. An efficient deep learning framework for P300 evoked related potential detection in EEG signal. Computer Methods and Programs in Biomedicine 229 (2023), 107324.

[8] Rishabh Krishnan Iyer. 2015. Submodular optimization and machine learning: Theoretical results, unifying and scalable algorithms, and applications. Ph. D. Dissertation.

[9] Kaggle. 2019. Aptos 2019 blindness detection. https://www.kaggle.com/c/aptos2019-blindness-detection/data

[10] Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C.S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, Justin Dong, Made K. Prasadha, Jacqueline Pei, Magdalene Y.L. Ting, Jie Zhu, Christina Li, Sierra Hewett, Jason Dong, Ian Ziyar, Alexander Shi, Runze Zhang, Lianghong Zheng, Rui Hou, William Shi, Xin Fu, Yaou Duan, Viet A.N. Huu, Cindy Wen, Edward D. Zhang, Charlotte L. Zhang, Oulan Li, Xiaobo Wang, Michael A. Singer, Xiaodong Sun, Jie Xu, Ali Tafreshi, M. Anthony Lewis, Huimin Xia, and Kang Zhang. 2018. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell 172, 5 (feb 2018), 1122–1131.e9. https: //doi.org/10.1016/j.cell.2018.02.010

[11] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. nature 521, 7553 (2015), 436–444.

[12] Hui Li, Yitan Zhu, Elizabeth S Burnside, Erich Huang, Karen Drukker, Katherine A Hoadley, Cheng Fan, Suzanne D Conzen, Margarita Zuley, Jose M Net, et al. 2016. Quantitative

MRI radiomics in the prediction of molecular classifications of breast cancer subtypes in the TCGA/TCIA data set. NPJ breast cancer 2 (2016), 16012.

[13] Trond Linjordet and Krisztian Balog. 2019. Impact of training dataset size on neural answer selection models. In Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part I 41. Springer, 828–835.

[14] Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob van de Loo, Rob Vogels, et al. 2018. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. GigaScience 7, 6 (2018), giy065.

[15] Fengbei Liu, Yu Tian, Yuanhong Chen, Yuyuan Liu, Vasileios Belagiannis, and Gustavo Carneiro. 2021. ACPL: Anti-curriculum Pseudo-labelling forSemi-supervised Medical Image Classification. CoRR abs/2111.12918 (2021). arXiv:2111.12918 https://arxiv.org/abs/2111.12918

[16] Kun Liu, Xiaolin Ning, and Sidong Liu. 2022. Medical Image Classification Based on Semi-Supervised Generative Adversarial Network and Pseudo-Labelling. Sensors 22, 24 (dec 2022), 9967. https://doi.org/10.3390/s22249967

[17] Ayush Maheshwari, Krishnateja Killamsetty, Ganesh Ramakrishnan, Rishabh Iyer, Marina Danilevsky, and Lucian Popa. 2021. Learning to Robustly Aggregate Labeling Functions for Semi-supervised Data Programming. arXiv preprint arXiv:2109.11410 (2021).

[18] Mohammad Peikari, Sherine Salama, Sharon Nofech-Mozes, and Anne L. Martel. 2018. A Cluster-then-label Semi-supervised Learning Approach for Pathology Image Classification. Scientific Reports 8, 1 (may 2018). https://doi.org/10.1038/ s41598-018-24876-0

[19] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. International journal of computer vision 115 (2015), 211–252.

[20] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In Proceedings of the IEEE international conference on computer vision. 843–852.

[21] Roger Sun, Elaine Johanna Limkin, Maria Vakalopoulou, Laurent Dercle, Stéphane Champiat, Shan Rong Han, Loïc Verlingue, David Brandao, Andrea Lancia, Samy Ammari, et al. 2018. A radiomics approach to assess tumourinfiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study. The Lancet Oncology 19, 9 (2018), 1180–1191.

[22] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data 5, 1 (aug 2018). https://doi.org/10.1038/ sdata.2018.161

[23] Ruchika Verma, Neeraj Kumar, Abhijeet Patil, Nikhil Cherian Kurian, Swapnil Rane, and Amit Sethi. 2020. Multi-organ nuclei segmentation and classification challenge 2020. IEEE transactions on medical imaging 39, 1380-1391 (2020), 8.

[24] Fuyong Xing, Tell Bennett, and Debashis Ghosh. 2019. Adversarial domain adaptation and pseudo-labeling for cross-modality microscopy image quantification. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22. Springer, 740–749.