Problem: We have limited labeled data distributed across multiple clients (e.g., organizations or devices) in a decentralized setup. There is a large amount of unlabeled data that needs to be annotated or classified. Data privacy must be preserved (data cannot leave the clients).
Solution: Use a shared latent space representation trained on labeled data from all clients to annotate unlabeled images in a decentralized, privacy-preserving manner.

Camaraderie: Content-based Knowledge Transfer for Medical Image Labelling using Supervised Autoencoders in a Decentralized Setting

Advait Padhye*
200100014@iitb.ac.in
Indian Institute of Technology, Bombay
Mumbai, India

Shreeja Bhakat*
shreejabhakat20@gmail.com
Indian Statistical Institute, Kolkata
Kolkata, India

Humaira Firdowse
humairafirdowse@gmail.com
Indian Institute of Technology, Bombay
Mumbai, India

Atharv Savarkar
avsavarkar@gmail.com
Indian Institute of Technology, Bombay
Mumbai, India

Ganesh Ramakrishnan
ganesh@cse.iitb.ac.in
Indian Institute of Technology, Bombay
Mumbai, India

Kshitij S. Jadhav
kshitij.jadhav@iitb.ac.in
Indian Institute of Technology, Bombay
Mumbai, India

ABSTRACT

Deep neural networks for medical imaging require large high-quality labelled data, a huge bottleneck for resource poor settings. Given the privacy requirements of medical data, institutes are unwilling to share data, causing an hindrance in resource poor settings. In the present paper, (**C**amaraderie: **C**ontent-b**a**sed Knowledge Transfer for **M**edical Image L**a**belling using Supe**r**vised **A**utoencoders in a **De**cent**r**al**i**zed S**e**tting) we propose to use Discrete Classifier Supervised Autoencoder (DC-SAE) to generate low-dimensional representations of a few annotated images at the **Donor** client and transfer both the DC-SAE's encoder part and the latent space representations to the **Recipient** client without sharing raw data. We then pass the unlabelled images of the Recipient Client through this encoder to obtain their latent space representation. In a supervised setting, using latent space representation of Donor client's labelled images, we accurately annotate images of Recipient client. Camaraderie demonstrates that DC-SAE outperforms Recipient end label accuracy beyond classical VAE based classification and anomaly detection based VAE. Thus, given a limited amount of labelled data in a decentralized privacy preserving scenario, one can transfer latent space representation across clients to annotate large number of unlabelled images with high accuracy.

A Variational Autoencoder (VAE) is a type of generative model that combines principles from probabilistic inference and deep learning. It is a subclass of autoencoders designed to learn latent representations of input data and generate new data samples similar to the training data. Unlike traditional autoencoders, VAEs impose a probabilistic structure on the latent space.

CCS CONCEPTS

• **Security and privacy** → **Privacy protections**; • **Applied computing** → **Health informatics**.

KEYWORDS

Supervised Autoencoders, Classification, Decentralized Setting

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3615216

ACM Reference Format:
Advait Padhye*, Shreeja Bhakat*, Humaira Firdowse, Atharv Savarkar, Ganesh Ramakrishnan, and Kshitij S. Jadhav. 2023. Camaraderie: Content-based Knowledge Transfer for Medical Image Labelling using Supervised Autoencoders in a Decentralized Setting . In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3583780.3615216

* Equal Contribution.

1 INTRODUCTION

Deep Learning is a powerful and versatile approach due to its ability to learn meaningful and complex representations from raw data leading to significant breakthroughs in tasks like image classification, object detection etc [5]. Large-scale labelled datasets allow models to learn generalizable representations, capturing an inherent data variability[8]. However, implementation of these models in resource poor setting is impeded due to the need for large amount of high-quality labelled data. Further, the privacy requirements associated with medical data is a challenge, as institutions are reluctant to share their data due to risks associated with data breaches[7].

Variational Auto-Encoders (VAEs) could address these aforementioned challenges. VAEs are powerful generative models that offer a unique framework for unsupervised learning, allowing for the discovery of underlying patterns and representations in complex datasets [4]. However, their utilization extends beyond this and the latent space of a trained VAE serves as an optimal feature representation, capturing essential characteristics of the input data [9] representing it in a low dimension, thus simplifying the classification process [10]. Also, VAEs can respect data privacy requirements by operating directly on the data in its encoded form. All this makes VAEs a suitable choice for resource-poor settings, where the confidentiality of data is paramount [2]. In this paper Camaraderie, we demonstrate that in a distributed data setting, one can transfer latent space representations from a donor to a recipient client to label large number of unlabelled images with higher accuracy using a supervised version of VAE.

2 OUR CONTRIBUTION

We demonstrate how knowledge can be transferred in a privacy preserving manner from a **Donor** (with few annotated images) to a **Recipient** (with large number of unannotated images) assisting in annotating images at the latter (*c.f.* Figure 1). One could employ contrastive learning to tackle this task. However, these methods generally rely on a centralized dataset which doesn't respect the privacy requirements. Traditional VAE based methods showed poor performance on this task since traditional VAEs are not trained to discriminate between different classes. These issues can be overcome by training VAEs on other kinds of reconstruction losses that encourage discriminative data representations of input images. Therefore, we use class specific latent space representations that were obtained by modifying VAE and termed as Discrete Classifier Supervised Autoencoder (DC-SAE) (*c.f.* Section 4.3),developed by introducing an additional **repulsion** loss term to the loss function of a standard VAE. This encourages greater separation between the latent space representations corresponding to different classes (*c.f.* Section 4.2). Then, both the encoder of the DC-SAE and latent space representation of the Donor site images were transferred to the Recipient site (*c.f.* Figure 1). This transferred encoder was used for feature extraction of unlabelled images of the Recipient site (*c.f.* Section 4.5). We trained a Support Vector Machine (SVM) on the latent space representations of the labelled images from Donor site to establish a decision boundary and (*c.f.* Section 4.5) the latent space representations of the Recipient site unlabelled images were then used by inputting them to the aforementioned SVM to predict their labels. We demonstrate that our method provides higher predictive labelling accuracy compared to two baseline settings, namely using traditional VAE and anomaly detection based VAE (*c.f.* Figure 3).

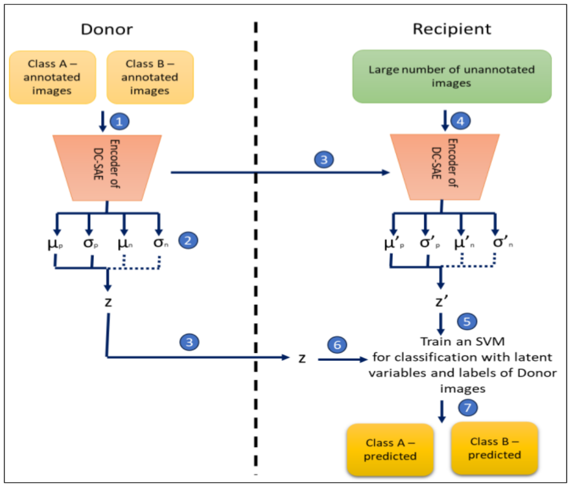
Donor
Recipient
Class A – annotated images
Class B – annotated images
Large number of unannotated images
1
Encoder of DC-SAE
3
4
Encoder of DC-SAE
μp σp μn σn
2
μ'p σ'p μ'n σ'n
z
z'
5
3
z
6
Train an SVM for classification with latent variables and labels of Donor images
7
Class A – predicted
Class B – predicted

Figure 1: Visual representation of the proposed algorithm. The numbers represent the steps in our algorithm

3 PROBLEM SETTING

Let us consider that there are two clients **A** and **B**. Client **A** (Donor) has a small labelled dataset $\mathbf{D}_a = \{(\mathbf{X}_i^{(a)}, \mathbf{y}_i^{(a)})\}_{i=1}^{\mathbf{n}}$ where $\mathbf{X}_i^{(a)}$ is the i$^{\text{th}}$ image present in the dataset of size **n** and $\mathbf{y}_i^{(a)}$ is the label corresponding to this image. On the other hand, Client **B** (Recipient) has a large unlabelled dataset $\mathbf{D}_b = \{\mathbf{X}_i^{(b)}\}_{i=1}^{\mathrm{N}}$ where $\mathbf{X}_i^{(b)}$ is the i$^{\text{th}}$ image present in the dataset of size N. For this problem, we will assume that $\mathbf{D}_a$ and $\mathbf{D}_b$ consist of only 2 classes i.e $\mathbf{y}_i^{(a)} = \{0, 1\}$. Our objective is to label $\mathbf{D}_b$ using the features that can be learned from $\mathbf{D}_a$ albeit without sharing $\mathbf{D}_a$ with Client **B**.

4 METHODOLOGY

We propose a new framework, **Discrete Classifier Supervised Autoencoder (DC-SAE)**, for feature extraction by modifying Variational Autoencoder architecture. After feature extraction, **Donor (A)** transfers the set of features $\mathbf{f}_a$ as well as a portion of DC-SAE to **Recipient (B)** Client. With the help of the encoder of DC-SAE, **Recipient** extracts features $\mathbf{f}_b$ from $\mathbf{D}_b$ which will be compared with $\mathbf{f}_a$. For the classification task, we used supervised machine learning based Support Vector classifier technique (*c.f.* Algo 1).

One of the key advantages of Algorithm 1 is that it allows clients possessing poor computational capabilities to annotate their unlabelled datasets without training some large image classification model. Moreover, this method preserves privacy since either clients do not share their data.

4.1 Preliminaries : Variational Autoencoder

VAE consists of an Encoder and a Decoder. Encoder being parameterized by weights θ, approximates posterior distribution $q(z|x)$ and maps the input $\mathbf{x}$ to a lower dimensional latent space. Decoder is similarly parameterized by weights ϕ and yields an approximate likelihood distribution $p(x|z)$. The latent variables $\mathbf{z}$, sampled from the approximate posterior $q_\theta(z|x)$, are fed into the Decoder which attempts to reconstruct the original input. During the training of VAE, we maximise the Evidence Lower Bound (ELBO) given by

$$L(\theta, \phi; x) = -\mathrm{D}_{\mathrm{KL}}(q_\theta(z|x) \parallel p(z)) + E_{\sim q_\theta(z|x)}[p_\phi(x|z)] \quad (1)$$

where the first term represents the Kullback-Leibler Divergence between the approximate posterior $q_\theta(z|x)$ and the latent prior $p(z)$. The second term is the reconstruction loss, often taken as the Mean Square Error between the input $\mathbf{x}$ and the output $\hat{\mathbf{x}}$ of the decoder. To obtain a closed form of the loss function, we choose $p(z)$ to follow a Standard Normal Distribution and $q_\theta(z|x)$ to follow a Gaussian Distribution of mean μ and variance σ^2 which are the outputs of the Encoder. To enable backpropagation during the training of VAE, we use the reparameterization process as demonstrated below instead of actually sampling from $\mathcal{N}(\mu, \sigma^2)$

$$\mathbf{z} = \mu + \epsilon \odot \sigma \qquad \text{where } \epsilon \sim \mathcal{N}(0, \mathrm{I}) \quad (2)$$

Finally, the closed form loss function used for training a standard VAE on input images x_i, $i = 1, 2, \ldots, \mathrm{N}$ in a training set is given by

$$L = \frac{1}{2}\sum_{i=1}^{\mathrm{N}}(\mu_i^2 + \sigma_i^2 - 1 - \log(\sigma_i^2)) + \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} \parallel x_i - \hat{x_i} \parallel_2^2 \quad (3)$$

4.2 Limitations of Standard VAE

The features $\mathbf{f}_a$ used for training the classifier model, are latent variables $\mathbf{z}$ sampled from the learned posterior $\mathcal{N}(\mu, \sigma^2)$ since they capture the input data distribution. However, due to the unsupervised learning process used for training a standard VAE, input data belonging to the different classes usually have similar representations in the latent space (*c.f.* Figure 2). Therefore, using the latent space representations of the standard Variational Autoencoders for classification task results in poor performance.

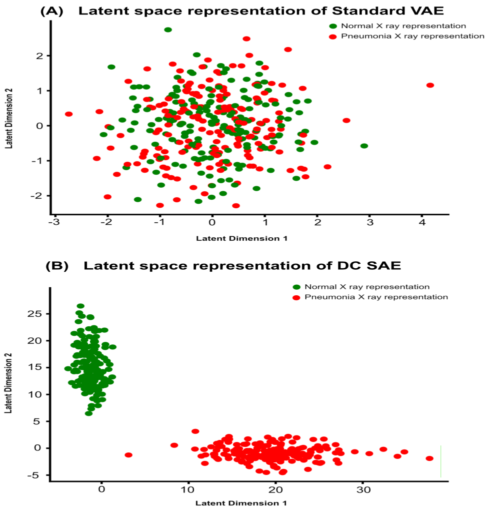

Figure 2: Latent Space Comparison between Standard VAE and DC-SAE on PneumoniaMNIST Dataset

4.3 Discrete Classifier SAE

Similar to the standard VAE, our proposed Discrete Classifier SAE (DC-SAE) consists of an Encoder Network and a Decoder Network. Unlike the standard VAE, the DC-SAE generates two distributions in the latent space by learning two sets of mean and variance { (μ_p, σ^2_p), (μ_n, σ^2_n) } (since we are testing in a binary setting). The latent variable **z**, which is given as an input to the Decoder, is sampled from one of the set of mean and variance as per the class label. DC-SAE's Decoder functions in the same way as the Decoder in the standard VAE (*c.f.* Figure 1).

```
Algorithm 1: DC-SAE Algorithm
Input: $\mathrm{D}_a = \{(\mathrm{X}_i^{(a)}, \mathrm{y}_i^{(a)})\}_{i=1}^{\mathrm{n}}$, $\mathrm{D}_b = \{\mathrm{X}_i^{(b)}\}_{i=1}^{\mathrm{N}}$
Output: $\{\mathrm{y}_i^{(b)}\}_{i=1}^{\mathrm{N}}$
Model: DC-SAE Encoder $f_\theta$, DC-SAE Decoder $g_\phi$
Donor Site:
    $\theta, \phi \leftarrow$ Train DC-SAE made of $f_\theta$ and $g_\phi$ using $\mathbf{D}_a$
    for $i = 1$ to $n$ do
        $(\mu_{l_i}, \sigma^2_{l_i}, \mu_{\sim l_i}, \sigma^2_{\sim l_i}) = f_\theta(\mathrm{X}_i^{(a)})$
        $\epsilon \sim \mathcal{N}(0, I)$, $\mathbf{f}_i^{(a)} = \mu_{l_i} + \epsilon \odot \sigma_{l_i}$
    Transfer $\{\mathbf{f}_i^{(a)}\}_{i=1}^{n}$ and $f_\theta$ to Recipient Client
Recipient Site:
    for $i = 1$ to $N$ do
        $(\mu_p, \sigma^2_p, \mu_n, \sigma^2_n) = f_\theta(\mathrm{X}_i^{(b)})$
        $\epsilon \sim \mathcal{N}(0, I)$, $\mathbf{f}_i^{(b)} = \mu_p + \epsilon \odot \sigma_p$
    C $\leftarrow$ SVM Binary Classifier trained on $\mathbf{f}_i^{(a)}$
    $\mathbf{y}^{(b)} = C(\mathbf{f}^{(b)})$
```

4.4 Training and Latent Space

We propose to modify the loss function used for training the DC-SAE as follows:

$$\mathbf{L} = \mathbf{L}_{\mathrm{kl}} + \alpha \cdot \mathbf{L}_{\mathrm{mse}} + \beta \cdot \mathbf{L}_{\mathrm{rep}} \tag{4}$$

$$\text{where, } \mathbf{L}_{\mathrm{mse}} = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \| x_i - \hat{x}_i \|_2^2$$

$$\mathbf{L}_{\mathrm{kl}} = \frac{1}{2} \sum_{i=1}^{\mathrm{N}} (\mu^2_{\sim l_i} + \sigma^2_{\sim l_i} - 1 - \log(\sigma^2_{\sim l_i}))$$

$$\mathbf{L}_{\mathrm{rep}} = \frac{1}{\rho} \sum_{i=1}^{\mathrm{N}} \max(0, \rho - \| \mu_{l_i} - \mu_{\sim l_i} \|_2^2)^2$$

Here, l_i represents the class label of the i$^{\mathrm{th}}$ input image. For example, an input image $\mathbf{x}_i$ belonging to Positive Class i.e $\mathbf{y}_i = 1$ will have μ_{l_i} as μ_p and $\mu_{\sim l_i}$ as μ_n.

The unsupervised KL Divergence Loss has been replaced by a supervised KL Divergence Loss term $\mathbf{L}_{\mathrm{kl}}$ which forces the Gaussian Distribution parameterized by $(\mu_{\sim l_i}, \sigma^2_{\sim l_i})$ to be closer to a standard normal distribution. While, the repulsion loss term $\mathbf{L}_{\mathrm{rep}}$ forces the means $(\mu_{\sim l_i}, \mu_{l_i})$ to be a minimum distance of ρ away from each other. The combined effect of these two loss terms forces the Gaussian distribution defined by $(\mu_{l_i}, \sigma^2_{l_i})$ to be far away from the origin and thus separate the two class clusters as seen in Figure 2. This results in better discriminative representations of the input images which are suitable for classification. At the same time, the reconstruction loss term $\mathbf{L}_{\mathrm{mse}}$ aims to preserve adequate input information in the latent space in order to reconstruct the input image using the Decoder Network.

4.5 Feature Extraction and Classification

Upon training the DC-SAE model on the labelled dataset $\mathbf{D}_a$ at Client **A**, the Decoder is discarded and the labelled dataset $\mathbf{D}_a$ is passed through the Encoder to obtain latent space representations of the input images. The features $\mathbf{f}_a$ are sampled from the Gaussian Distribution $\mathcal{N}(\mu_{l_i}, \sigma^2_{l_i})$. These features and the Encoder is shared with Client **B**. Client **B** passes unlabelled dataset $\mathbf{D}_b$ through this shared Encoder to obtain 2 sets of mean and variance {(μ_p, σ^2_p), (μ_n, σ^2_n)}. The features sampled from Gaussian Distribution $\mathcal{N}(\mu_p, \sigma^2_p)$ prove to be a good predictor for positive class images while features sampled from Gaussian Distribution $\mathcal{N}(\mu_n, \sigma^2_n)$ prove to be a good predictor for negative class images. Finally, to perform classification, we train a Support Vector Machine on features $\mathbf{f}_a$ and predict the class labels on the basis of features $\mathbf{f}_b$.

client a, b -> training,testing set ?

5 EXPERIMENTS

5.1 Datasets

- **PneumoniaMNIST** [3] For our experiment we used 1341 images from each of the normal and pneumonia X-ray classes.
- **APTOS Dataset** [6] This consists of retinal images captured using fundus photography categorized into five classes namely normal(0), mild(1), moderate(2), severe(3); and proliferative(4). We merged classes 1, 2, 3, and 4 into a single class labeled as 1 (1857 samples). Class 0 remained separate with 1805 samples.

hard / soft / meta learner at integration step ?

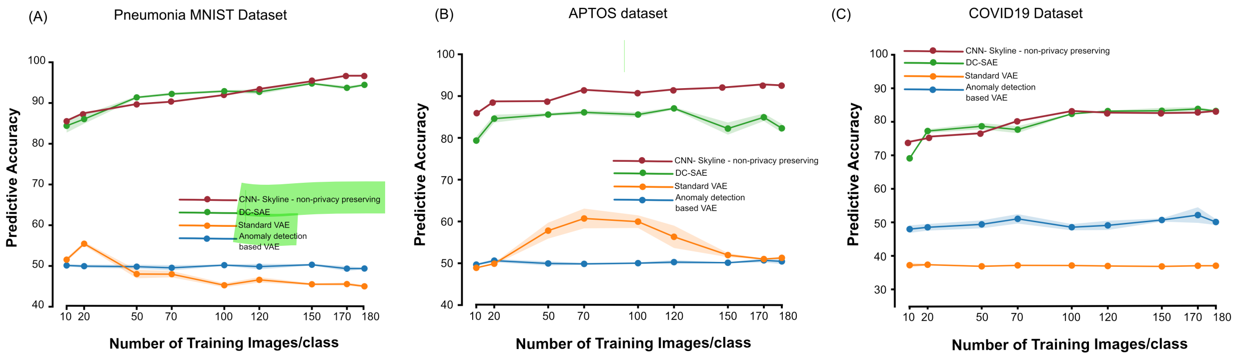

Figure 3: Predictive accuracy on unlabelled data at Recipient Client. Data represented as mean +/- SEM for 4 runs at each point

- **COVID-19 Dataset** [1] The dataset includes chest X-rays categorized into four classes: normal, pneumonia, lung opacity, and COVID. We exclusively utilized 3616 COVID X-rays for positive class and 6012 lung opacity images for negative classs.

5.2 Experimantal Setup

The encoder component of our DC-SAE model consists of convolutional layers and the fully connected layers to obtain the latent space parameters $\{(\mu_\mathrm{p}, \sigma_\mathrm{p}^2), (\mu_\mathrm{n}, \sigma_\mathrm{n}^2)\}$. The decoder component also has a convolutional and fully connected layers.

The latent space dimension is set to 2 and the learning rate is set to 0.0001. We maintained the configurations mentioned in Table 1 for the remaining hyperparameters. α and β represent the weightages of reconstruction loss term and repulsion loss term respectively in the total loss function. ρ represents the separation distance between clusters belonging to different classes.

We have used 9 different values for the size of training dataset $\mathbf{D}_a$ (in the range 10 to 180) at the Donor Site. This allows us to gain insights into the relationship between the number of images used for training at Donor Site and the accuracy of our system.

Dataset	α	β	ρ	Epochs
PneumoniaMNIST	10	10	5	50
Aptos	20	10	10	1000
Covid-19	20	10	10	100

Table 1: Hyperparameters for DC-SAE

We evaluated the accuracy of our model by comparing it with two baseline models: Standard VAE and Anomaly Detection based VAE.

The Standard VAE follows a classical VAE architecture, where the loss function consists of the KL-Divergence and Reconstruction loss. After training the model, we utilize the latent space representation of Donor Client images as features to train an SVM Classifier for image classification and labeling at the Recipient client.

The Anomaly Detection based VAE also adopts a standard VAE framework. However, in this approach, we exclusively train the model on positive class images at Donor client. Subsequently, we transfer the entire network to Recipient client where we label an image as a negative class if its reconstruction loss exceeds a predefined threshold; otherwise, it is labeled as a positive class.

Skyline : Assuming that preserving the privacy is not a concern, we train a RESNET18 model at the Recipient by directly transferring images from the Donor to classify unlabelled images at Recipient.

6 RESULTS

Through our approach Camaraderie, we demonstrate that the latent space representations generated by standard VAEs exhibit an overlap for both the classes in all three datasets, potentially, lacking any discernible separation between the two classes (*c.f.* Figure 2). Consequently, the accuracy achieved by the standard VAE model was approximately 50%, which is comparable to random labeling (*c.f.* Figure 3). This outcome underlines the VAE model's inability to effectively distinguish between the two classes in all three datasets. Similar results were observed for the Anomaly detection based VAE baseline (*c.f.* Figure 3).

In contrast, our proposed DC-SAE model (*c.f.* Figure 1) demonstrated superior performance in terms of predictive accuracy of image labels at the Recipient client in all three datasets as shown in Figure 3. It successfully discriminated between the two classes by establishing distinct boundaries in the latent space representations, leading to significantly improved accuracy. In the case of PneumoniaMNIST dataset, accuracy increased from 81% to 95% as the number of training images per class increased from 10 to 180. Similar trends were observed in the APTOS and COVID-19 datasets, with the DC-SAE model outperforming the baselines with accuracy ranging from 79% to 87% and 70% to 86%, respectively. This demonstrates that a larger labelled dataset contributed to enhanced predictive accuracy, likely due to the increased diversity of images, allowing the DC-SAE model to learn more comprehensive and representative features for accurate classification. Also, we see that the predictive accuracy is close to the skyline accuracy. This noteworthy improvement in performance highlights the effectiveness of our approach in accurately labeling medical images by successfully differentiating the two classes in the latent space.

7 CONCLUSION

Inability to share data due to privacy concerns along with presence of less amount of annotated data are limitations for different clients in resource constrained settings to collaborate together. Our approach Camaraderiedemonstrates that using a modified VAE, knowledge can be transferred between clients to annotate unlabelled data while still preserving privacy. This can be done even with less labelled data to begin with bringing in the possibility of collaboration between clients to annotate medical images with high accuracy which can then be utilized to develop deep models for predictive analysis.

REFERENCES

[1] Muhammad EH Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al Emadi, et al. 2020. Can AI help in screening viral and COVID-19 pneumonia? *Ieee Access* 8 (2020), 132665–132676.

[2] Xintao Duan, Jingjing Liu, and En Zhang. 2019. Efficient image encryption and compression based on a VAE generative model. *Journal of Real-Time Image Processing* 16 (2019), 765–773.

[3] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell* 172, 5 (2018), 1122–1131.

[4] Diederik P Kingma, Max Welling, et al. 2019. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning* 12, 4 (2019), 307–392.

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

[6] Ruhan Liu, Xiangning Wang, Qiang Wu, Ling Dai, Xi Fang, Tao Yan, Jaemin Son, Shiqi Tang, Jiang Li, Zijian Gao, et al. 2022. Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge. *Patterns* 3, 6 (2022), 100512.

[7] W Nicholson Price and I Glenn Cohen. 2019. Privacy in the age of medical big data. *Nature medicine* 25, 1 (2019), 37–43.

[8] Iqbal H Sarker. 2021. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN Computer Science* 2, 6 (2021), 420.

[9] Ramzi Snoussi and Habib Youssef. 2023. VAE-Based Latent Representations Learning for Botnet Detection in IoT Networks. *Journal of Network and Systems Management* 31, 1 (2023), 4.

[10] Yang Zhi-Han. 2022. Training Latent Variable Models with Auto-encoding Variational Bayes: A Tutorial. *arXiv preprint arXiv:2208.07818* (2022).