Identifying the Origins of Synthetic Biomedical Images: An Ensemble Approach with Pseudo-Labeling

Notebook for the ImageCLEF Lab at CLEF 2025

Amilcare Gentili[1,2,*,†]

[1]San Diego VA Health Care System
[2]University of California, San Diego

Abstract

The increasing use of synthetic data in biomedical image analysis necessitates robust methods to ensure data provenance and detect potential links to original training datasets is prevented. This working note presents the results of SDVAHCS/UCSD team in the ImageCLEFMedical 2025 GANs Task, specifically Subtask 2: "Identify Training Data Subsets". Our approach involved training an ensemble of EfficientNet models (b0, b1, and b2) using a multiclass classification framework. We explored a comprehensive hyperparameter space and employed a max voting ensembling strategy to improve prediction accuracy. Furthermore, we investigated the benefits of pseudo-labeling the unlabeled test data to augment our training set. To assess for overfitting on the validation data, we utilized a sequestered portion of the original training data to evaluate the reliability of our pseudolabeling process by comparing prediction accuracy on both datasets. Our final submissions demonstrated the effectiveness of this combined approach, with ensemble models leveraging pseudolabeled data achieving strong performance in identifying the origins of synthetic biomedical images. We discuss the implications of our findings and propose avenues for future research, including exploring alternative architectures and advanced ensembling techniques to further enhance the traceability and security of synthetic medical data.

Keywords

EfficientNet, Ensemble Learning, Pseudo-Labeling, Synthetic Biomedical Images, ImageCLEF

1. Introduction

The development of AI systems for medical image analysis, including disease prediction, detection, and classification, hinges on the availability of large and diverse training datasets. High-quality data enable these models to learn intricate patterns, enhancing their accuracy and reliability. However, the acquisition of real medical data is often restricted due to patient privacy concerns, limiting the data available for effective AI model training and hindering advancements in healthcare applications. Synthetic data, artificially generated to resemble real medical data without coming from actual patients, presents a potential solution to this challenge. Generative models, such as Generative Adversarial Networks (GANs), can produce these datasets, allowing researchers to develop and evaluate AI systems while safeguarding patient privacy and facilitating the collection of diverse training information. A critical concern with synthetic data is to ensure the absence of hidden links or "fingerprints" from the real data used for its generation. The potential traceability of synthetic data back to the original patient information poses a privacy risk. Therefore, guaranteeing that synthetic data are completely devoid of such imprints is paramount to maintain patient confidentiality while leveraging its benefits for AI-driven healthcare innovation. Previous iterations of this task at ImageCLEF[1] (2023 and 2024)[2] training strategies, resulting in four distinct submissions to the competition investigated the presence of these "fingerprints" in synthetic images generated by various models. The findings consistently revealed that generative models retain and embed features of their training data, highlighting significant implications

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
†Email: agentili@ucsd.edu
agentili@ucsd.edu (A. Gentili)
https://gentili.net (A. Gentili)
0000-0002-5623-7512 (A. Gentili)

for security and privacy. These results underscore the urgent need for effective methods to detect and mitigate these imprints to ensure the privacy and utility of synthetic medical images for research and development. The second edition further confirmed the existence of unique "fingerprints" that could be used to attribute synthetic images to specific generative models based on identifiable patterns and features. This working note presents the results of SDVAHCS/UCSD team in the ImageCLEFMedical 2025 GANs Task, specifically Subtask 2: "Identify Training Data Subsets". Our approach to identifying the training data subset for each synthetic biomedical image involved training deep learning models to perform a six-class classification (real vs. five generated subsets). We explored various model architectures and training strategies, resulting in four distinct submissions to the competition [3].

2. Methods

2.1. Data Description

The benchmarking dataset for this task comprises both real and synthetic biomedical images. The real images consist of axial slices from 3D CT scans of approximately 8,000 patients with lung tuberculosis. These slices exhibit variability in appearance, ranging from relatively normal to displaying distinct lung lesions, including severe cases. The real images are provided in an 8-bit per pixel PNG format with a standardized resolution of 256x256 pixels. The synthetic images, also sized at 256x256 pixels, were generated using various generative models, including Generative Adversarial Networks (GANs) and Diffusion Neural Networks. By providing both real and synthetic datasets, the task enables participants to analyze and compare the characteristics of synthetic images with their real counterparts, investigating potential "fingerprints" and patterns related to the training process. The training data set is organized into two main folders: "generated" and "real." The "generated" folder contains five subfolders, each holding synthetic images produced using a different training dataset. The "real" folder also contains five corresponding subfolders, each representing a specific training data set used to train the generative model. The real images within each of these subfolders were used to generate the synthetic images found in the correspondingly named subfolder within the "generated" directory. The specific mapping is as follows:

- Folder `t1` (real images) → Used to generate synthetic images in `gen_t1`
- Folder `t2` (real images) → Used to generate synthetic images in `gen_t2`
- Folder `t3` (real images) → Used to generate synthetic images in `gen_t3`
- Folder `t4` (real images) → Used to generate synthetic images in `gen_t4`
- Folder `t5` (real images) → Used to generate synthetic images in `gen_t5`

The test dataset includes 25,000 generated images, each derived from a real subgroup of images within the training dataset. The images were divided into 6 classes, (0 for real images, 1 through 5 for synthetic images generated from the 't1' through 't5' real subsets, respectively) See Table 1. For the first 2 submissions only the training images were utilized during training, using the classification obtained by this training pseudolabels were assigned to the test images, and the test images were also used for training but not validation. For the third submission, some of the training images were sequestered and not used for training or validation, but only to check the accuracy of the model. For the final submission, all images were used for training.

2.2. Data Preparation and Loading

Standardized data loading and preprocessing were applied, including resizing images to the input size of the chosen EfficientNet variant [4], converting to tensors, and normalization.

2.3. Ensembling Strategy

For ensemble submissions, the final prediction for each test image was determined by using the max voting ensembling technique [5]

Table 1
Number of images per label and dataset split, with and without pseudo-labeling and sequestered data.

Dataset	Class	No pseudolabel	Pseudolabel and sequestered images	Pseudolabel no sequestered images
test	-	25000	25000	25000
train	0	8000	6692	8322
	1	1600	6258	6594
	2	1600	6344	6658
	3	1600	6334	6662
	4	1600	6351	6702
	5	1600	6328	6691
total train	-	16000	38307	41629
validation	0	2000	1684	1684
	1	400	341	341
	2	400	351	351
	3	400	340	340
	4	400	325	325
	5	400	330	330
total validation	-	4000	3371	3371
sequestered	0	0	1630	0
	1	0	336	0
	2	0	314	0
	3	0	328	0
	4	0	351	0
	5	0	363	0
total sequestered	-	0	3322	0

2.4. Pseudo-Labeling and Sequestered Data

Pseudolabeling was used to leverage information from the unlabeled test set. The use of a sequestered dataset (in submission 1782) helped to verify the generalization of the pseudolabeling process.

2.5. Model Architectures and Training Parameters

We primarily utilized the EfficientNet family of convolutional neural networks (specifically versions b0, b1, and b2), pretrained on ImageNet. We conducted extensive hyperparameter tuning, exploring batch sizes of 16, 32, 64, and 128, learning rates of 0.0001, 0.00005, and 0.0005, and training for 5, 10, 20, and 30 epochs.

2.6. Individual Submissions

- **Submission 1425 (Efficientnet1_lr_0.0001_bs_32_epoch.zip):** This submission represents the results of a single EfficientNet-b1 model. The specific model submitted was selected as the best performing model after training across the aforementioned hyperparameter grid.
- **Submission 1426 (EfficientNet0-2ensemble.zip):** This submission was an ensemble of models based on EfficientNet-b0, EfficientNet-b1, and EfficientNet-b2. These individual models were trained using the same hyperparameter search space as described for submission 1425. Only models that achieved at least 98% accuracy on the validation set were included in this ensemble. The final prediction was determined by a max voting ensembling of the individual model predictions.
- **Submission 1782 (EfficientNet1Pseudolabelval2Ens26Ac.zip):** This submission involved an ensemble of 26 models trained with EfficientNet-b0, EfficientNet-b1, and EfficientNet-b2. The

models in this ensemble were trained with batch sizes of 32 and 64, learning rates of 0.0001, 0.00005, and 0.0005, and trained for 20 and 30 epochs. A key aspect was the use of pseudo-labeled images generated from the test set, and a sequestered group of training data used to check for overfitting. Models were initially trained, and their predictions were tested on the sequestered group. The accuracy of these predictions on the sequestered data was compared to the accuracy on the validation set to ensure that the pseudolabeling process was not introducing overfitting on the validation data. Only models that achieved at least 99% accuracy on the validation set were included in this ensemble.

- **Submission 1871 (EfficientNet1PseuelAlldolab99.36.zip):** This submission utilized an ensemble of models based solely on EfficientNet-b1 trained on the entire original training dataset combined with the pseudo-labeled test images. No training data was sequestered in this case. The models in this ensemble were trained with batch sizes of 32 and 64, learning rates of 0.0001, 0.00005, and 0.0005, and trained for 20 and 30 epochs.

3. Results

The results of our submission are presented in Table 2.

Table 2
Performance Metrics for Different Runs

Run	Filename	accuracy	f1	precision	recall	specificity
1425	Efficientnet1_lr_0.0001_bs_32_epoch.zip	0.97076	0.971084	0.971588	0.97076	0.99306
1426	EfficientNet0-2ensemble.zip	0.98784	0.987952	0.988093	0.98784	0.996852
1782	EfficientNet1Pseudolabelval2Ens26Ac.zip	0.98796	0.988074	0.98821	0.98796	0.996893
1871	EfficientNet1PseuelAlldolab99.36.zip	0.98796	0.988074	0.98821	0.98796	0.996893

Run 1425 using a single model had the worst results, even if the model with the best accuracy on validation set was used. Ensembling multiple models, Run 1426 gave a significant improvement. Adding pseudolabeled test images, Run 1782, to the training set provided a small additional improvement. Training on the entire training set and pseudolabeled test dataset, Run 1871, did not provide any further improvement.

4. Discussion

Our experiments demonstrated the effectiveness of using pretrained EfficientNet models for the task of identifying the training data subsets of synthetic biomedical images. The high validation accuracies achieved by individual models, particularly those included in our ensembles, suggest that these architectures are well-suited for capturing subtle differences between real and synthetically generated images from different origins.

The use of ensemble learning consistently improved performance over single models, highlighting the benefit of combining the strengths of different model instances. The mode-based ensembling strategy proved effective in aggregating the predictions.

Pseudolabeling provided a valuable mechanism to leverage the information present in the unlabeled test data. By incorporating these pseudo-labels into the training set, we were able to further refine our models. The strategy of using a sequestered dataset to generate and validate the quality of the pseudo-labels (as in submission 1782) was crucial in ensuring that this process contributed positively to generalization and did not lead to overfitting on the validation set. The comparison of prediction accuracy on the sequestered data with that on the validation set provided a useful metric for assessing the reliability of the generated pseudolabels.

The fact that adding sequestered data did not change the submission classification for the final submission (submission 1871) suggests that the performance gains from increasing training data can

plateau if the models are already capturing the underlying data patterns effectively.

5. Ideas for Future Work

Several avenues for future research and improvement could be explored:

- **Exploring other state-of-the-art Vision Transformers (ViTs):** While EfficientNet models proved effective, investigating the performance of ViTs [6], which have shown remarkable success in various computer vision tasks, could yield further improvements. Their attention mechanisms might be particularly adept at identifying subtle "fingerprints."
- **Advanced Ensembling Techniques:** Instead of simple vote-based ensembling, more sophisticated techniques like weighted averaging based on validation performance, or even using a meta-learner to combine the predictions of individual models could be investigated [7].
- **Larger and More Diverse Datasets:** Training on larger and more diverse datasets of both real and synthetic biomedical images could improve the generalizability of the models.

6. Conclusion

Our participation in this task demonstrated a successful approach to identifying the training data subsets of synthetic biomedical images using deep learning techniques. The combination of pre-trained EfficientNet models, careful hyperparameter tuning, effective ensemble strategies, and a principled approach to pseudolabeling allowed us to achieve competitive results. The comparison of performance on a sequestered dataset with the validation set provided a crucial step in ensuring the reliability of our methods. Future work focusing on exploring alternative architectures, advanced ensembling techniques, holds the potential for further advancements in this important area of research.

Declaration on Generative AI

During the preparation of this work, the author(s) used Gemini and Perplexity to convert the Word document to LaTeX format to follow the publication formatting guideline. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] B. Ionescu, H. Müller, D.-C. Stanciu, A.-G. Andrei, A. Radzhabov, Y. Prokopchuk, Ştefan, Liviu-Daniel, M.-G. Constantin, M. Dogariu, V. Kovalev, H. Damm, J. Rückert, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. G. Pakull, B. Bracke, O. Pelka, B. Eryilmaz, H. Becker, W.-W. Yim, N. Codella, R. A. Novoa, J. Malvehy, D. Dimitrov, R. J. Das, Z. Xie, H. M. Shan, P. Nakov, I. Koychev, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, D. Fabre, C. Macaire, B. Lecouteux, D. Schwab, M. Potthast, M. Heinrich, J. Kiesel, M. Wolter, B. Stein, Overview of imageclef 2025: Multimedia retrieval in medical, social media and content recommendation applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 16th International Conference of the CLEF Association (CLEF 2025), Springer Lecture Notes in Computer Science LNCS, Madrid, Spain, 2025.

[2] A. Andrei, A. Radzhabov, D. Karpenka, Y. Prokopchuk, V. Kovalev, B. Ionescu, H. M uller, Overview of 2024 imageclefmedical gans task, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 15th International Conference of the CLEF Association (CLEF 2024), Springer Lecture Notes in Computer Science LNCS, Grenoble, France, 2024.

[3] A.-G. Andrei, M. G. Constantin, M. Dogariu, A. Radzhabov, L.-D. Ştefan, Y. Prokopchuk, V. Kovalev, H. Müller, B. Ionescu, Overview of ImageCLEFmedical 2025 – GANs Task, in: CLEF2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.

[4] M. Tan, Q. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 6105–6114. URL: https://proceedings.mlr.press/v97/tan19a.html.

[5] B. Hopkinson, A. King, D. Owen, M. Johnson-Roberson, M. Long, S. Bhandarkar, Automated classification of three-dimensional reconstructions of coral reefs using convolutional neural networks, PLoS One 15 (2020) e0230671.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, CoRR abs/2010.11929 (2020). URL: https://arxiv.org/abs/2010.11929. arXiv:2010.11929.

[7] S. Imran, P. N, A review on ensemble machine and deep learning techniques used in the classification of computed tomography medical images, International Journal of Health Sciences and Research 14 (2024) 201–213. URL: https://doi.org/10.52403/ijhsr.20240124. doi:10.52403/ijhsr.20240124.