Deep Regression for Predicting Annotation Variability in Skin Lesion Segmentation

Rodrigo Bravo Iriarte, Sophia Tung, Timmy Tsai, Hooyar Foroughizadeh, and Khiem Dinh Can

CMPT419 Spring 2025, Prof. Hamarneh

Abstract. Automated prediction of inter-annotator disagreement in medical image segmentation tasks gives valuable insights for improving annotation consistency and model reliability. In this study, we investigate deep learning-based regression to predict disagreement levels (quantified as values between 0 and 1) directly from dermatological images. We present a framework that employs stratified dataset splitting in order to preserve label distribution between training, validation, and test sets and fine-tunes ImageNet-pre-trained models (ResNet, EfficientNet, Vision Transformer) for regression. Models are trained to minimize mean squared error (MSE) and tested across both MSE and mean absolute error (MAE) metrics. Our results demonstrate the utility of transfer learning on this task. Our approach addresses the main challenge of quantifying annotation variation in medical image processing pipelines with a foundation for quality control for crowd-sourced segmentation procedures.

Keywords: Medical image analysis · Inter-annotator disagreement prediction · Deep learning regression · Transfer learning · Model fine-tuning · Predicting annotation variability · Skin lesion segmentation

1 Introduction

1.1 Application and Problem Description

Medical image segmentation is vital for diagnostics, enabling precise contouring of structures. Yet, it faces inter-annotator variability, especially in dermatological imaging with ambiguous lesion borders [1]. Variability is measured from 0 (no overlap) to 1 (full overlap) using the Dice-Sorensen coefficient [2]. Traditional post-hoc quality control slows crowd-sourced annotation pipelines. Predicting disagreement directly from images offers real-time detection of disputed cases.

1.2 Motivation

Crowd-sourced medical annotations demand robust quality control [3]. Retrospective agreement calculations lack real-time feedback. A prediction system could: (1) prioritize expert review, (2) guide annotator training via disagreement patterns, and (3) enhance uncertainty estimation for analysis. This is key to scaling reliable medical AI.

1.3 Background

Inter-annotator disagreement is common in medical imaging, notably in skin lesion segmentation due to irregular boundaries [1]. Fleiss' kappa and Dice scores are standard measures [4,2], but recent deep learning efforts focus on classification, not regression of disagreement scores. Vision transformers and advanced CNNs offer new potential for detecting annotation ambiguity [5].

1.4 Related Work

Prior studies vary in tackling annotation variability. Ronneberger et al. showed U-Net excels in segmentation, not uncertainty [6]. Jungo et al. used Monte Carlo dropout for mask uncertainty [7], unlike our annotator-focused regression. We leverage transfer learning, building on Sharma et al.'s texture feature extraction [8].

1.5 Report Overview

This report proceeds: Section 2 details our 2,261-image dataset and metrics. Section 3 covers methodology—stratified splitting, regression modifications, and training of ResNet, EfficientNet, and Vision Transformer. Section 4 evaluates performance with MSE/MAE, statistical tests, and correlation analysis. Section 5 discusses challenges and learnings, Section 6 lists contributions, and Sections 7-8 offer conclusions and future directions.

2 Materials

In order to train and test our deep learning model, we used an archive of 2261 2D RGB images of varying skin lesions, all resized to a resolution of 224x224. In addition, we used a CSV file which provided the names of each image, the count for how many experts segmented each image, the mean Dice-Sorensen coefficient (Dice score), and Fleiss' kappa value. However, for the purposes of our project, we did not use the Fleiss' kappa value.

To guide our model design and training pipeline, we referenced course-provided lecture notes and materials. For model architecture, we used pre-trained computer vision models—ResNet-18, EfficientNet-B0, and ViT-B/32—all initialized with ImageNet weights to leverage transfer learning. These models were accessed through the torchvision.models module in PyTorch.

Additionally, we used a variety of Python-based tools and libraries including gdown for downloading datasets from Google Drive, pandas for tabular data processing, scikit-learn for preprocessing and evaluation utilities, and torch and matplotlib for model training and result visualization. All software dependencies were managed and installed through the Python Package Index (PyPI).

Table 1. Utilized Materials

Category	Name / Description	Source
Dataset	`ISIC_Archive` — RGB skin lesion image dataset used for training and evaluation.	Course-provided
Tabular Data	`metrics.csv` — Contains image IDs and associated annotator disagreement values.	Generated from ISIC_Archive annotations (Course-provided)
Methodology	Lecture slides and written notes — Used to guide the approach and model design.	Course-provided
Pre-trained Models	ResNet-18, EfficientNet-B0, ViT-B/32 — ImageNet-weighted models for transfer learning.	PyTorch (`torchvision.models`)
Tools	`gdown`, `pandas`, `scikit-learn`, `torch`, `matplotlib`, etc. — Used for data loading, modeling, evaluation, and plotting.	Python packages (via PyPI)

3 Methods

3.1 Proposed Method

Overview To predict annotation variability from images of skin lesions, we followed the following approach and considerations, including the preprocessing of the acquired dataset of dermatological images, followed by the training of multiple ImageNet-based machine learning models with a focus on regression-based deep learning prediction. In addition, heatmap images are generated to improve the interpretability of model predictions.

Annotation Variability and Dice Score The variable of interest in this paper is annotation variability, which refers to the disagreement or inconsistencies found in how annotators (certified dermatologists for the purpose of this study) label the same images. The content to be analyzed from a csv file each contains the associated dice score mean. The dice score measures the similarity, or overlap, between two datasets. From two annotations, it gets the ratio of their intersection with respect to the contribution in size from both areas. For the present study, the dataset contains the average dice score per image for more than 2 annotators.

3.2 Algorithmic Details

Model Selection and Training Three distinct ImageNet architectures were utilized They each followed a standardized training pipeline, considering transformations, data augmentation, loss function, optimizer, and early stopping

strategies. Trained on the same image datasets. All models are initialized with pre-trained weights, a particular characteristic for ImageNet weights, and the final layers are modified to output a single regression value, which is the prediction of interest.

ResNet-18 A convolutional network with residual connections

EfficientNet-B0 More recent architecture/design based on a compound scaling method which is expected to give it more efficiency and balance.

Vit-B/32 Instead of using convolusions, this particular model uses self-attention mechanisms, a distinct approach that may affect the overall performance.

Data Preprocessing In order to prepare the raw data for the training of the model, the images underwent a series of transformations expected to improve model performance

Resizing and Tensor transformation. The images were resized to 224 x 224 pixels and then converted to tensor, this was done using the transforms functions from torchvision. It is important to note that both of these transformations were implemented for the training and validation datasets.

Data Augmentation. In addition, random horizontal flipping and random rotations of 10 degrees were applied, also from torchvision. However, this was only done for the training dataset.

Training & Evaluation

Split From the collected dataset of images, we are splitting them into 75% for the test set, 15% for the test set, and 15% for the validation set. To do so, the process undergoes 10 epochs, and will include early stopping with a patience of 3 epochs to prevent overfitting.

Stratified Training With the objective to increase the robustness of the machine learning model's training, stratified training was used to ensure a proper distribution of the dice-scores across all bins.

Mean Absolute Error To evaluate how far off the model's predictions are during testing and validation, the mean absolute error is employed for an easier to digest visualization of its accuracy. For reference, its formula is as follows:

$$(MAE) = 1/N \sum_{i=1}^{N} |(y_i - y_i')| \tag{1}$$

3.3 Loss Function and Optimization

For the purpose of this paper, the **Mean Squared Error (MSE)** is being employed as the loss function.

$$(MSE) = 1/N \sum_{i=1}^{N} (y_i - y_i')^2 \tag{2}$$

In the case of optimization, it has been performed using the **Adam Optimizer** with weight decay and a learning rate of 10^{-4}. Additionally, a learning rate scheduler is implemented with the purpose of reducing the learning rate if the validation loss plateaus, done by a factor of 0.1.

3.4 Model Interpretability

With the objective to increase the models reliability, and interpretability, Grad-Cam was applied. The Gradient-Weighted Class Activation Mapping generates heatmaps highlighting the relevant regions that are influencing the model's prediction. While this technique is ready for all three architectures, the focus of analysis will be on the best performing model exclusively. The resulting heatmaps will be shown and assessed for potential biases and identifying important image features that have been key for the model's decisions.

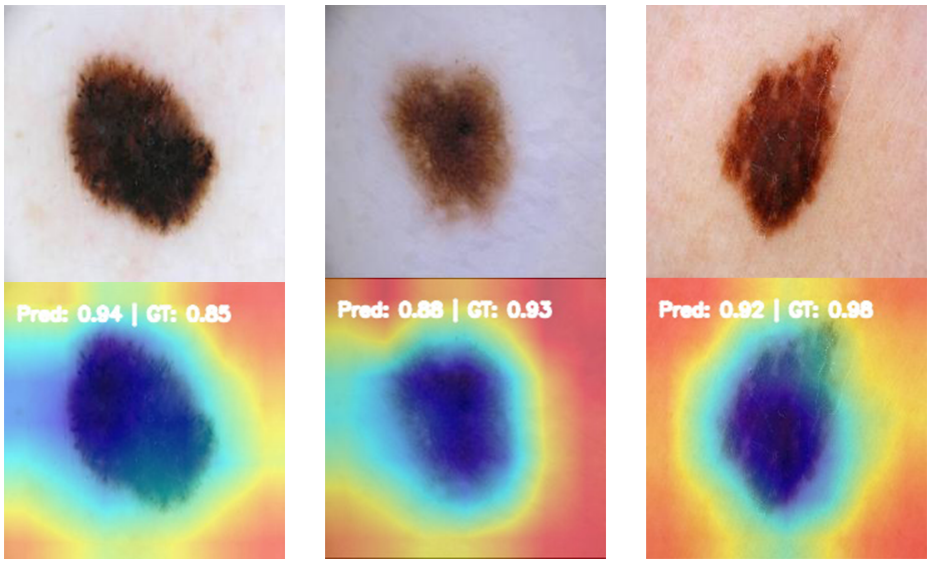

Fig. 1. Example of original images (top) compared to the corresponding heatmap generation images (bottom).

4 Results

We evaluated the effectiveness of the three different pre-trained architectures, ResNet-18, EfficientNet-B0, and ViT-B/32. The quantitative values that we focused on in our comparisons were the MSE and MAE.

Table 2. Test run performance comparison across different model architectures with standard deviation.

Model	**Test MSE**	**Test MAE**
ResNet-18	0.0340 ± 0.0603	0.1380 ± 0.1223
EfficientNet-B0	0.0382 ± 0.0654	0.1498 ± 0.1254
ViT-B/32	0.0418 ± 0.0996	0.1391 ± 0.1499

Table 2 shows a sample run with and it's important quantitative data. Through numerous other test runs, we noticed the same pattern, ResNet-18 and EfficientNet-B0 usually had a lower MSE and MAE value than ViT-B/32. However, in order to conclude whether certain models are more efficient, we must prove that the differences in the MSE and MAE are statistically significant.

We originally planned on using t-tests to determine the statistical significance of the difference in the MSE and MAE, however, we were unable to use the t-tests because the data did not satisfy the normality assumption, as shown in Figure 2.

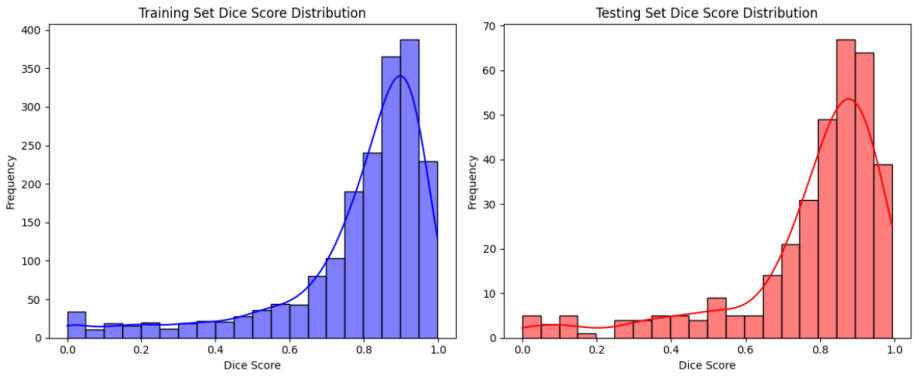

Fig. 2. Dice score distributions

4.1 Experiment

We conducted an experiment to compare the performance of ResNet-18, EfficientNet-B0, and ViT-B/32 for predicting skin annotation variability.

Null hypothesis: There is no significant difference in the predictive performance (measured by MAE and MSE) among ResNet-18, EfficientNet-B0, and ViT-B/32 for estimating annotation variability in skin lesions.

Alternative hypothesis: At least one of the model architectures (ResNet-18, EfficientNet-B0, or ViT-B/32) is significantly better for estimating annotation variability in skin lesions.

The target variable used in the analysis is `dice_mean`, which represents the mean Dice score for each image.

4.2 Kruskal-Wallis Test

To determine whether differences in the predictive performance of the models were statistically significant, we performed the Kruskal-Wallis test, a non-parametric method for comparing multiple groups. The test was conducted on the results of 10 runs for each model, using both MSE and MAE as metrics. We used the significance level 0.05.

However, in order to perform the Kruskal-Wallis test, the distribution of the data must be similar, which it is not. In order to make it similar, we applied z-score normalization to the data, as shown in Figure 3 and Figure 4.

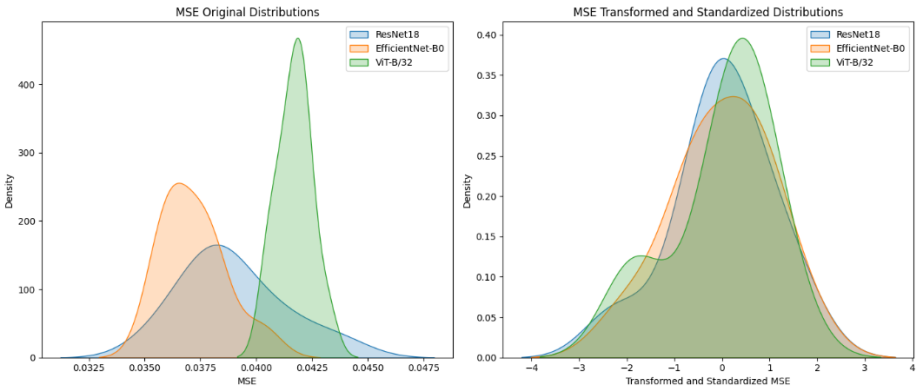

Fig. 3. MSE distributions

We also ensured our models are using the same images as each other for the training and testing for each test run, to ensure that the difference in performance was not due to a variance in the data set.

As shown in Table 3, the p-values for both metrics are greater than the significance level (0.05).Thus, the null hypothesis can not be rejected, indicating no significant differences in predictive performance among the models.

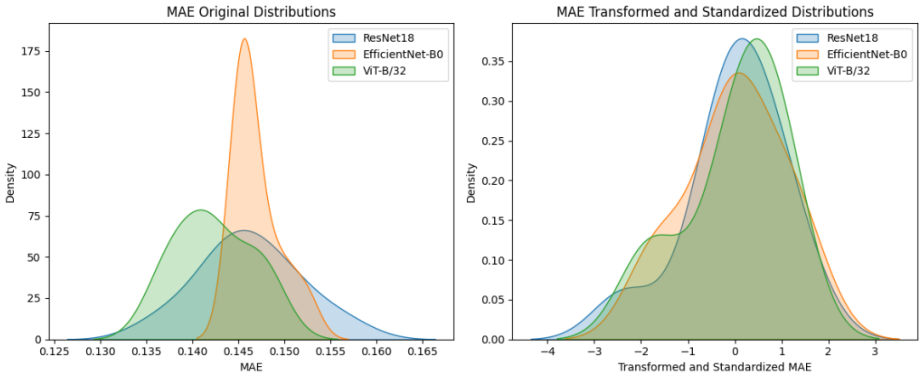

Fig. 4. MAE distributions

Table 3. Kruskal-Wallis test results for MSE and MAE.

Metric	Test Statistic	p-value
MSE	0.09319	0.95447
MAE	0.11104	0.94599

4.3 Diagnosis Correlation Analysis

To investigate potential confounding factors in annotation disagreement, we examined the relationship between lesion diagnosis (benign vs. malignant) and inter-annotator consensus (dc_mean) using point-biserial correlation. The analysis pipeline consisted of:

1. **Data Integration**: Merged annotation metrics with clinical metadata from JSON files
2. **Binary Encoding**: Mapped diagnosis labels to numerical values (0=benign, 1=malignant)
3. **Statistical Analysis**: Calculated point-biserial correlation coefficient
4. **Visualization**: Generated comparative box plots of dc_mean distributions

The analysis revealed no significant relationship between lesion malignancy and annotation consensus ($r = 0.005$, $p = 0.828$). This indicates that:

- Annotator disagreement occurs independently of lesion malignancy status
- Other factors (e.g., border ambiguity, texture complexity) likely drive variability
- Model performance differences cannot be attributed to diagnosis distribution

Data Collection Note: Lesion images and metadata were obtained through the ISIC Archive API Image Downloader (provided by Kumar Abhishek), ensuring standardized retrieval of dermatological data.

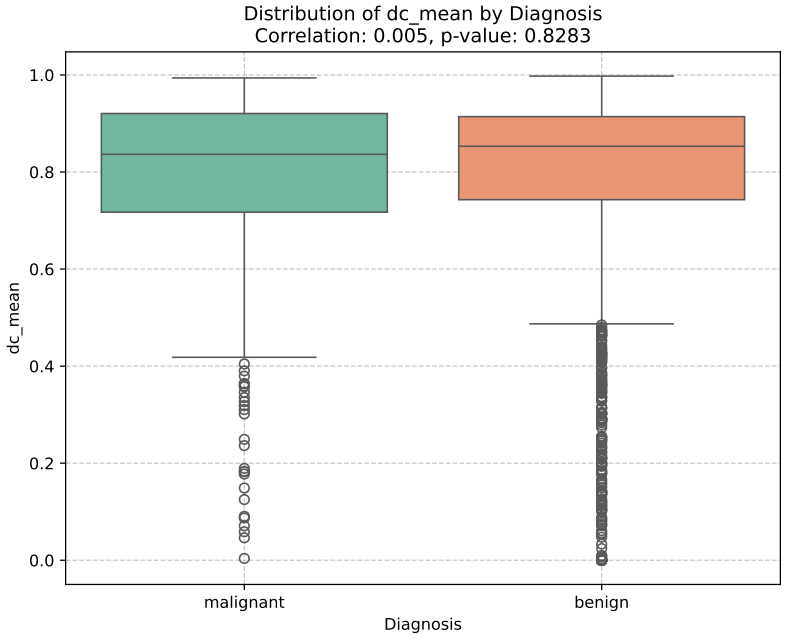

Fig. 5. Distribution of inter-annotator agreement (dc_mean) by lesion diagnosis. The weak correlation (r = 0.005, p = 0.828) suggests diagnosis type does not significantly influence annotation consistency.

Table 4. Point-biserial correlation results

Metric	**Correlation**	**p-value**
Diagnosis vs dc_mean	0.005	0.828

5 Accomplishments

Through this project, we became comfortable with key technical concepts such as regression, segmentation, and inter-annotator disagreement in medical imaging. We gained hands-on experience using Python libraries commonly used in machine learning workflows, including PyTorch, torchvision, and scikit-learn, and applied them effectively to solve a real-world problem.

One of the more challenging aspects was implementing Grad-CAM for model interpretability. Since our task involved predicting annotation variability rather than performing direct segmentation, generating meaningful heatmaps proved difficult. Despite this, we were able to experiment with various model architectures and gain a deeper understanding of the limitations and considerations when applying interpretability tools in regression contexts.

6 Conclusion and Discussions

We used deep regression models—ResNet18, EfficientNet-B0, and ViT-B-32 on 2,261 dermoscopic images to predict inter-annotator variability in skin lesion segmentation. Our goal was a system to estimate annotation discord from images, aiding real-time quality control in crowd-sourced medical imaging.

Results show transfer learning captures variability features, with models achieving similar MSE and MAE. ResNet18 had the lowest MSE (0.0340 ± 0.0603) and MAE (0.1380 ± 0.1223), a slight edge, but Kruskal-Wallis tests found no significant differences (p-values greater than 0.05). This may reflect few runs (10) or data variation, suggesting more tests are needed.

Diagnosis correlation analysis showed little link between malignancy and annotation agreement ($r = 0.005$, $p = 0.828$), indicating variability stems from factors like border fuzziness or texture, not diagnosis. This highlights the potential of image-based prediction to detect subtle influences.

Some of the challenges included Grad-CAM's poor interpretability for regression, suggesting a need for tailored methods. Limited dataset size and variety hindered generalization, and small model differences imply data quality matters more than architecture.

Our work forms the foundation for automated quality control, allowing expert verification and training. Real-time discord prediction would make pipeline optimization possible, enhancing medical AI performance. Low MSE/MAE correlations, however, suggest optimization with larger datasets or multi-task learning is required to achieve accuracy.

This work validates deep regression's potential for variability prediction, presenting a novel medical imaging solution. As a proof-of-concept, its generalizability and real-world applicability are left open, and further research outlines the way forward for broader dermatological use.

7 Future Work

While our study successfully demonstrates the feasibility of predicting annotation variability using deep regression models, several opportunities exist to enhance and extend this work:

- **Improved Interpretability**: The application of Grad-CAM revealed some challenges in generating meaningful heatmaps for regression tasks [9]. Future work could explore alternative interpretability methods, such as SHAP or even attention-based visualizations tailored for regression which better identify image features driving annotation disagreement predictions [10].
- **Expanded Dataset**: Getting a larger dataset would strengthen our findings and test model robustness across diverse skin lesions. It could address potential overfitting to the current 2,261 images [11].
- **Multi-Task Learning**: Now that our models simply predict the mean of the Dice score (dc mean), the extension of the framework to multi-task

learning—predicting dc mean, the number of annotators, and other variability measures (e.g., Fleiss' kappa)—could enable better interpretation of inter-annotator variability and higher utility of the models in quality control pipelines [12].
- **Real-Time Integration**: Integrating the prediction system with an active annotation process could enable real-time feedback to annotators. This would include optimizing the speed of model inference and building a user interface to highlight high-disagreement cases for review, potentially employing lightweight models like MobileNet [13].

Acknowledgements

We would like to thank Kumar Abhishek (Teaching Assistant at Simon Fraser University) for providing the data sets and offering valuable guidance throughout this project.

We are also grateful to Professor Ghassan Hamarneh (Simon Fraser University) for sharing supplemental material in class, as well as for his constructive feedback and insightful questions that helped keep our work on track.

References

1. R. Jensen et al. Variability in manual segmentation of skin lesions. *Journal of Medical Imaging*, 2(3):036001, 2015.
2. L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.
3. L. Maier-Hein et al. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature Communications*, 9(1):5217, 2018.
4. J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.
5. A. Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*, 2021.
6. O. Ronneberger et al. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241. Springer, 2015.
7. A. Jungo and M. Reyes. Assessing reliability and challenges of uncertainty estimation in medical image segmentation. *Medical Image Analysis*, 61:101664, 2020.
8. A. Sharma et al. Transfer learning for skin lesion classification using convolutional neural networks. *IEEE Journal of Biomedical and Health Informatics*, 23(6):2339–2347, 2019.
9. R. R. Selvaraju et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, 2017.
10. S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4765–4774, 2017.

11. International Skin Imaging Collaboration. Isic 2018: Skin lesion analysis towards melanoma detection, 2018.
12. R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.
13. A. G. Howard et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint*, 2017.