Skin Lesion Segmentation and Classification

Nagarjuna Devarayi
Data Science
Anderson College of Business and Computing.
Regis University, Denver, CO, USA
ndevarayi@regis.edu

Abstract

Melanoma has great mortality rates in case of late diagnosis, but is much susceptible to therapy when infested at a very early age. Tens of thousands of cutaneous lesions are screened by dermatologists each year, but the equal opportunity to receive an assessment of specialists is unevenly distributed between geographic areas and insurance coverage. Although automated diagnostic systems have potential, they all tend to fail at two critical points: good definition of lesions with normal dermal tissue and proper control of extreme inaccuracy in the percentage composition between malignant melanomas and benign nevi, which are scaled at disturbing percentages. This project suggests an extensive analytical chain, which starts with lesion segmentation with the help of adversarial neural networks and moves to lesion classification with EfficientNet architectures. To address the issue of class imbalance, several mechanisms of methodology will be strictly compared, namely, loss function reweighting, SMOTE-based oversampling, and conditional Generative Adversarial Networks of synthetic data augmentation. Segmentation module will demarcate the lesions in dermoscopic imagery before classification, then segmented and raw images will be regarded by EfficientNet classifiers to determine the quantitative advantage of the accurate segmentation.

I. Introduction/Background

Dermatologists use dermoscopy, amplified imaging technology, which is supplemented with special lighting, to examine lesions of uncertain etiology. They evaluate them based on morphological characteristics, including asymmetry, irregular shapes, heterogeneity of color and size change. Skills in these assessments take a lot of training. The limited number of dermatologic specialists in a rural and underserved environment also increases the delay in diagnosis, therefore, automated screening systems as an initial step in triaging case on the basis of lesions of most at-risk of expert referral can reduce the disparities. The technical issues of this sphere are great. Confounding factors that are commonly used in dermoscopic imaging are cutaneous hair, complicated skin texture and poor access to light and demarcation between neoplastic and normal epidermis. In addition, the underlying medical data are characterized by the significant imbalance of the classes, melanomas make only about 120% of the tested lesions, but the correct determination of such pathologies is of the highest priority. The current study deals with the segmentation problem to find the localization of lesions, and the classification problem to find the lesion taxonomy. Conditional generative adversarial networks (cGANs) will be used to learn lesion contours to be used in segmentation. The experiments on classification will entail EfficientNet architectures that have been demonstrated to have better trade-offs between predictive accuracy and computing requirements compared to traditional ResNet vestigues. It will compare three class-balancing methods which include:

(i) class-weighting
(ii) The synthetic minority over-sampling technique (SMOTE) augmentation
(iii) GAN-based data augmentation

This work will require the application of adversarial segmentation models, the training of multiple instances of the classifier, the use of complementary methods to address a class imbalance and the evaluation of a work on clinical metrics that do not prioritize overall performance but recall. The methodology plan will involve the application of GAN- learned segmentation models, the training of EfficientNet-based classifiers, and the actual evaluation of three balancing methods on the segmented and the unsegmented

image sets. The metric of performance will be determined through a set of measures, such as a per-class accuracy and recall, and an error analysis will concentrate on the mistaken examples.

II. Problem Statement

Can one effectively construct a robust skin-lesion classifier that can withstand significant class imbalance and at the same time can admit variation in quality of image? This work organically repeats the parts of the medical imaging pipeline: data gathering consists of loading dermoscopic image datasets with associated metadata (age, sex, location of lesion, diagnosis), data preparation includes segmenting data by patient identifier to avoid data leakage, dealing with data class imbalance, and performing relevant study of augmentation methods, data exploration consists of visualization of class distributions, sensitivity evaluations of segmentation performance, and interspecies misdiagnosis and failure, model construction involves training standalone classifiers with various balancing schemes, model optimization involves comparing performance between raw and segmented images.

What is the effect of different data-balancing mechanisms, such as re-weighting of the class, SMOTE, and GAN-based augmentation, and pre-processing the segmentation on the classifier performance on minority classes in dermoscopic image, namely melanoma?

III. Related Work

Tschandl et al. (2018) released HAM10000, the benchmark dataset containing 10,015 dermoscopic images across seven diagnostic categories [1]. Over 50% got confirmed by histopathology. The dataset is severely imbalanced - melanocytic nevi make up 67% while melanoma is just 11%. Esteva et al.(2017) trained a deep CNN on 129,000 clinical images and matched dermatologist performance on binary classification tasks [2]. Codella et al. (2019) won the ISIC challenge using ensemble methods combining multiple CNN architectures [3]. Pacheco et al. (2020) compared different CNN architectures on PAD-UFES-20, another dermoscopy dataset [4]. Barata et al. (2019) reviewed explainability methods for skin lesion classifiers [5]. They argued that clinical deployment requires understanding what features models learn.

IV. Related Works

Data Sources: HAM10000 Dataset - Dermoscopic images across seven lesion types from multiple institutions (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T) Kaggle Skin Lesion Dataset - Dermoscopic images with expert pathologist annotations for segmentation validation (https://www.kaggle.com/datasets/ahmedxc4/skin-ds)

V. Expected Outcomes

Segmentation networks and EfficientNets with various configurations are to be stored in the form of trained models. Segmentation will come in handy where the boundaries of the lesions are not distinctly defined.He/she is supposed to make a parity compare to choose the most effective approach to minority classes. Hypothesising that SMOTE will have a weak effect, class re-weighting stronger, and GAN based augmentation strongest in case the generated samples are really real. Per-class analysis will also show those lesions that are most confounded. The expected trend implies that basal cell carcinoma and benign keratosis rival each other in terms of classification because they are similar in terms of being scaly lesions. Unlike the melanocytic nevi, melanoma is highly differentiated, and therefore, sensitivity is the most important factor in detection. Dermatofibroma has several other features like central dimple when compressed and these may not be seen in still images which can affect the accuracy of classification. Analysis of the failure will be based on the analysis of the errors. The coloured lesions can be mistakenly identified due to the supremacy of colour. Images with more than one lesion or uncertain edges are bound to induce segmentation failures. Lesions detected are less than 50 pixels in diameter which can not provide enough detail to be classified accurately.

VI. METHODOLOGY AND APPROACH

A. Research Design and Data Sources My general research design will be an experimental, quantitative design, which will determine the effectiveness of different preprocessing, segmentation, and class-imbalance mitigation mechanisms in classifying dermoscopic images. I used the dataset of the ISIC 2019 Challenge that contains eight clinical classes (AK, BCC, BKL, DF, MEL, NV, VASC). I further used the ISIC 2018 Task 1 dataset to train an exclusive segmentation model. The major drawback of the primary data is that the ratio between the classes is extremely skewed (e.g., the majority group, NV, is 58 times larger than the minority group,DF). In order to counter this bias, my approach relied more on methods of advanced data augmentation and balancing instead of depending on traditional, unadapted learning pipelines.

B. Technical Infrastructure The whole project was written in Python as the main programming language. The essence of machine learning and deep learning pipelines were developed on TensorFlow and Keras and optimized on the basis of GPU acceleration. I used OpenCV and PIL to process the images on an advanced level, and data manipulation and statistical analysis were processed through Pandas and NumPy. To the specific imbalance methods, I used imbalanced-learn (imblarn) library, and to visualize the performance and analyze the feature-space, I used Matplotlib, Seaborn, and scikit-learn (PCA, t-SNE).

C. Analytical Techniques and Machine Learning Algorithms To construct a robust pathway from raw data to accurate classification, I engineered a multi-stage analytical pipeline: Image Preprocessing: In order to resolve lighting artifacts and other visual artifacts, I used a pipeline based on bilinear interpolation to resize images to 256 x 256 and CLAHE (Contrast Limited Adaptive Histogram Equalization) in the LAB color space to normalize luminance and Morphological Blackhat filtering with Telea inpainting to detect and removes hair artifacts algorithmically. Lesion Segmentation: I trained a Bi-directional ConvLSTM U-Net with Densely connected con- volutions on Binary Cross-Entropy and Dice Loss function. The lesion ROI (Region of Interest) of this model was isolated by zeroing out the background pixels and hypothetically, this allowed the downstream classifier to concentrate on the pathological features only. Classification Backbone: I have chosen the EfficientNet that is initialized with weights of the ImageNet as the main feature extractor. I used the strategy of two-phase transfer learning, i.e., first, freeze the base and then train it to a personal classification head (with a learning rate of 1e-2), and then thaw the entire network (with a learning rate of 1e-4).

Imbalance Mitigation Strategies: I tried several methods to fix the imbalance of classes (58:1):

1) **Algorithmic Class Weighting:** Proportional scaling of categorical cross-entropy loss by inverse class frequencies.
2) **Generative Adversarial Networks (CGAN):** I trained a Conditional GAN that can synthesize realistic, 64x64 pixel-level dermoscopic images to numerically balance the minority classes to the majority number (creating more than 25,000 synthetic images).
3) **Feature-Space SMOTE:** I obtained frozen EfficientNet feature vectors with 1,408 dimensions and directly used Synthetic Minority Oversampling Technique (SMOTE) on these feature vectors, and then trained a custom Multi-Layer Perceptron (MLP) head on those feature-balanced vectors.

D. Validation Strategies and Evaluation Metrics To provide the rigor and reproducibility of my results, I stratified the dataset into a strict split between Train and Validation (70% and 15% respectively) and Test, in order to make sure the same classes balanced within each subset. I also applied Early Stopping (to avoid overfitting) and dynamic learning rate schedule (ReduceLROnPlateau) during training. Since general accuracy is very deceiving when data is imbalanced, I considered model rigor with Macro F1-Score, Per-Class Recall (which is very important in minority malignant types of data such as Melanoma and Squamous Cell Carcinoma), and one-vs-Rest ROC-AUC curves. Lastly, in order to confirm internal consistency of the model, I used the following techniques to support visualization of the model: Grad-CAM to visualize spatial attention heatmaps and t-SNE/PCA to demonstrate feature-space separability.

VII. Data Description

A. Dataset Overview and Quality Assurance The basic dataset is composed of high-resolution dermoscopic images offered by the International Skin Imaging Collaboration (ISIC). All the images have a ground-truth CSV file that corresponds to each image, which is one of eight mutually-exclusive clinical diagnoses that are strictly exclusive (authenticated through one-hot encoding logic in my data loading script). I performed a wide Exploratory Data Analysis (EDA) so that it could be assured that the quality of data is rigorous. I checked that no labels were missing, and programmatically checked the image dimensions, which had a uniform mean aspect ratio, and extremely varied lighting conditions and artifact interaction (in a 200-image random sample, potential hair artifacts in 98%). I used these outliers directly in the automated artifact removal and CLAHE luminance standardization pipeline in my methodology.

B. Ethics, Governance, and Bias Mitigation ISIC archive contains medical data, which is heavily vetted, anonymized and publicly available; therefore, the issue of direct patient privacy, GDPR/HIPAA compliance and informed consent were addressed before the dataset creators published their findings. Nonetheless, one of the ethical issues in this project was a significant problem: algorithmic bias. Unchecked, the pathological excessive over-representation of benign Melanocytic Nevi (NV) would result in the model defaulting to the benign variable, and result in a failure to identify life-threatening melanomas (MEL) or basal cell carcinoma (BCC) in a clinical practice. One way that I have shown a level of ethical data science is by not using naive metrics of accuracy. To reduce the effect of algorithmic bias I was actively involved in designing and testing various balancing models (SMOTE, CGANs, Class Weights) and critically grading the models based on minority-class recall. This will make the resulting model clinically safe and not a shallow statistical model.

VIII. Conclusion

This practicum proposal will constitute a solution to a very serious issue in medical computer vision: the multi-class diagnosis of skin lesions, which will be accomplished in an end-to-end manner and without explicitly ignoring the presence of severe dataset imbalance and difference in quality of images.
In this project, I have sequentially planned and implemented a very advanced pipeline which includes image processing (CLAHE, morphological artifact removal), semantic segmentation (BCDU-Net), transfer learning (EfficientNet) and generative augmentation (GAN and SMOTE).
The originality of my work is the strict comparative analysis of strategies of imbalance mitigation. My empirical study of image-space augmentation (GAN and segmentation) versus feature-space balancing (SMOTE + MLP) allowed showing that the feature-space oversampling method has the greatest diagnostic capabilities with a test accuracy of 77.6 percent and a Macro F1-score of 0.566 percent (far outperforming the unbalanced and segmented ones). Moreover, the combination of Grad-CAM and t-SNE visual validations transforms the model into an interpretable device, which is an unquestionable requirement of the current medical AI.
Educationally, this project represents the final project on data science. It fills the distance between the deep learning theory and the dirty, wild, real world. Shifting the data ingestion and pipeline architecture to deep learning diagnostics and mitigation of ethical bias, it will show that I am ready to practice in the profession. It is a strong example of how I was able not only to create elaborate machine learning systems but also to critically analyze them to provide actionable, scientifically valid information.

IX. Timeline

Week 1: Download and process the data sets, study classes and image quality, study metadata.preparation preprocess pipelines, integrity.

Week 2: Baseline training, data augmentation and balancing (reweighting, SMOTE). unbalanced data equipments to determine the bottom of performance.

Week 3: Select and optimize segmentation network, check the quality of segmentation in the validation set, poisson distribution of the whole datasets.

Week 4: Fine train conditional GANs to accomplish lesion segmentation and data generation, analyze generated. encasements of reality and assortment.

Week 5: Categorize the trains which have various balancing options on both segmented and non-segmented images, validation set hyperparameter optimization.

Week 6: Final examination of the test set, construct confusion matrices and ROC curves, determine per-class. determine and detect optimal settings.

Week 7: Weekly review on the errors committed in the misclassified cases, findings of the report, cleaned code to be handed in, compose final medical implicating report.

REFERENCES

1) P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," Scientific Data, vol. 5, no. 1, pp. 1–9, 2018.
2) A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115–118, 2017.
3) N. C. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. W. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti et al., "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)," in arXiv preprint arXiv:1902.03368, 2019.
4) A. G. Pacheco, G. R. Lima, A. S. Salomao, B. Krohling, I. P. Biral, G. G. de Angelo, F. C. Alves Jr, J. G. Esgario, A. C. Simora, P. B. Castro et al., "Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones," Data in Brief, vol. 32, p. 106221, 2020.
5) C. Barata, M. E. Celebi, and J. S. Marques, "Explainable skin lesion diagnosis using taxonomies," Pattern Recognition, vol. 110, p. 107413, 2021.