COMP9444 Project Summary

Skin Lesion Classification Using Deep Learning

Hao Yan: z5497450   Xiaxi Li: z5473897   Zhen Zhou: z5513332

Wenhao Chen: z5442760   Haoran Sun: z5461804

I. Introduction

In dermatology, skin lesion classification is the key to differentiating melanoma from nevi and other benign lesions using deep learning techniques. This project will therefore zero in on how deep learning can be used in automating this classification procedure to better facilitate early detection and improved patient outcomes. Unlike variability and complexity of skin conditions in standardized datasets, this makes it quite an alluring area of research.

II. Literature Review

Kassem et al. (2021) review some skin lesion classification techniques at the root of machine learning and deep learning. Reflections of high accuracy and strong robustness could be realized through CNNs and ResNets. They pointed to challenges regarding data quality, the quality of the labels, model interpretability, and most of the models not handling variability in skin lesion images and requiring extensive labeled datasets for specific tasks generally hard to get.

Lopez et al. applied deep learning in the classifying of dermoscopic images using fine-tuned Inception-v3 on the ISIC 2017 dataset in 2017. Their model revealed improved performance with an AUC of 0.943 for macro-level classification, thereby greatly differentiating melanoma from other benign lesions. They, however, have been noticed leading to limitations connected with the imbalanced distribution; data biasing results towards classes with a higher frequency, together with their reliance on high-quality and large-scale datasets as limiting factors.

Benyahia et al. (2022) investigated several feature extraction techniques via deep learning and evaluated the performance using ResNet50 and VGG16 models with the ISIC 2019 dataset. ResNet50 had accuracy of 91.23%. It was observed that ResNet50 could extract features very well. VGG16 showed accuracy of 83.89% and it was found to be stable across various data scales. However, ResNet50 expended huge computational time and resources during training. Therefore, it is suggested that optimization of computational efficiency by model pruning and quantization methods.

III. Methods

1. CNN: They have had immense power in image classification due to the fact that they can automatically learn many spatial hierarchies of features. We chose to use CNNs because of their high accuracy and robustness in handling complex classification tasks; hence, we aim at exploiting their feature extraction capability to achieve precise skin lesion classification.

2. VGG19 is a simple and deep model that makes use of small 3x3 convolutional filters, which capture the fine details in images. In this work, we will be using fine-tuning of a pre-trained VGG19 for reducing training time while enhancing performance due to its detailed feature extraction to distinguish different kinds of skin lesions.

3. ResNet: As the residual connections bypass some layers, it is designed to train very deep networks without suffering from the vanishing gradient problem. In this work, we have used ResNet for the

extraction of complex patterns in skin lesion images; this deep architecture is aimed at enhancing classification accuracy and robustness.

4. DenseNet: Since each layer is connected to all the other layers in DenseNet, it maximizes information flow between layers and reuses features very effectively. We have taken DenseNet for its high performance and parameter efficiency to leverage it in enhancing model robustness and accuracy in the classification of skin lesions.

IV. Data Processing and Augmentation

1. Data Augmentation: We applied a unified data augmentation technique to improve the robustness and generalization capability of the models.

Rotation: Randomly rotating images within a span of 10 degrees.

Zoom: Applying a random zoom within a range of 0.1.

Width and Height Shifts: Shifting images horizontally and vertically by up to 10% of the total width and height.

Horizontal Flip: Randomly flipping images horizontally to increase dataset diversity.

2. Model Parameters: For consistency across models, we standardized the input shape, number of classes, batch size, and number of training epochs:

Input Shape: Resized to 75x100x3 (height, width, and color channels).

Number of Classes: 7; Batch Size: 16; Epochs: 20

3. Regularization Techniques: To reduce overfitting, we used dropout and batch normalization across all models. These techniques help stabilize training and make models more robust, addressing limitations identified in related work.

4. Training and Evaluation: The augmented dataset was used to train the models, which were then linked to the validation and test sets in order to test their performance. This work includes some important evaluation metric measures, such as accuracy, confusion matrix, and F1 score; thus, it is able to show a full overview for how effective each model will be in classifying skin lesions.

V. Experimental Setup

The dataset used for this project is the ISIC 2018 dataset, which can be found publicly and accessed through ISIC Archive. It contains 10,015 train, 193 validation, and 1,512 test images of dermoscopic skin lesions images classified under seven classes: Melanoma, Nevus, Basal Cell Carcinoma, Actinic Keratoses and Intraepithelial Carcinoma, Benign Keratoses-like Lesions, Dermatofibroma, and Vascular Lesions. The dataset includes a variety of lesion sizes, shapes, and colors, making the classification task challenging. One of the main observations during data exploration was class imbalance—the Nevus group is overrepresented, while Dermatofibroma is underrepresented—so potential techniques are needed in data augmentation to rectify this imbalance and beef up model robustness.

VI. Results

1. Key experimental results

CNN: Validation Accuracy: 84%; Test Accuracy: 68%; F1: 66.26%

Observations: What was seen is that the model of CNN is overfitting, with a large drop in performance from validation to test set. It simply means that it does well on both training and validation data but poorly on any unseen data.

VGG19: Validation accuracy: 91.72%; Test accuracy: 74.14%; F1: 74.63%.

Observations: VGG19 showed stability and good characteristics of feature extraction but was, compared to the deeper ResNet-50, slightly worse in terms of test performance. It has performed consistently across validation and test sets.

ResNet-50: Validation Accuracy: 86.53%; Test Accuracy: 79.10%; F1:78.27%.

Observations: Among the models tested, ResNet-50 returned the highest test accuracy and F1 score with an enhanced ability for feature extraction, due to its deeper architecture. However, this was prone to overfitting, manifesting itself in an abnormally high validation accuracy compared to test accuracy.

DenseNet: Validation accuracy: 81.35%; Test accuracy: 75.40%; F1: 74%.

Observations: DenseNet was very powerful, due to the efficient reuse of features; its performance is very close to that of VGG19. There is, however, still room for optimization in order to increase the accuracy on the test set.

2. Analysis based on the F1 score

CNN: The worst performance so far, the F1 score obtained by this algorithm was 66.26%, hence indicating that generalization to new data was difficult due to overfitting.

VGG19: This model gave an F1 score of 74.63%. The performance given by the VGG19 was really stable, and thus it is possible to apply it for certain tasks for which one needs stable feature extraction.

ResNet-50: The best among these had an F1 score of 78.27%. This is because ResNet-50 could extract deep features while generalizing well.

DenseNet: DenseNet scored an F1 score of 74%, indicating that it reuses features efficiently, though slightly behind ResNet-50 and therefore leaving scope for further tuning.

2. The main findings

Experiments showed that ResNet-50 performed skin lesion classification the best, gaining the highest F1 score and having very robust performance. VGG19 with its stable feature extraction, DenseNet with efficient reusage of features, have also given good results. The foundational CNN model suffered from overfitting and didn't generalize well.

3. Real-World Applications Analysis

These results suggest that ResNet-50, as the sharpest of the experimented models, becomes a very strong candidate for deployment in real-world clinical settings. Further enhancement of the robustness and reliability of such models needs improvements at the fronts of data augmentation, regularization, and class imbalance handling. Techniques related to synthetic data generation, different transfer learning criteria, and others would further boost performance toward applicability in practice.

4. Comparison with Other Proposed Solutions

Our solution is competitive w.r.t. performance compared to the existing literature. For instance, Kassem et al. (2021) raised issues with regard to data quality and model interpretability, which our approach partially resolves through better data augmentation and robust model selection. Lopez et al. achieved an AUC of 0.943 in 2017 using Inception-v3, which is way better, but our models like VGG19 really balance feature extraction and efficiency in training. Benyahia et al. (2022) have reported high accuracy values for ResNet50 and VGG16; however, our ResNet-50 shows similar performance but is much less resource-demanding.

5. Stand in Terms of the Standard Evaluation

In terms of standard evaluation metrics, our ResNet-50 model goes out of its way to yield test accuracy at 0.7910% and an F1 score of 0.7827%. Our ResNet-50 is powerful and serves well in clinical applicability; it compares favorably well against the state-of-the-art models for a strong balance of depth with computational efficacy—making this model a valid candidate for clinical application.

VII. Conclusions

1. Contributions

Model Evaluation: On ISIC 2018, we have systematically evaluated four deep models at the forefront of deep learning and drawn comprehensive comparisons concerning their performance.

Data Augmentation: Considerable data augmentation techniques were designed and implemented to improve the generalization of the model and to solve class imbalance problems.

Optimization: It has shown that ResNet-50 is efficient in deep feature extraction, holding the highest test accuracy and F1 score while remaining computationally efficient.

2. Key Strengths

Strong Performance: ResNet-50 has the best trade-off of depth against computational efficiency, making it very applicable in real-world scenarios.

Comprehensive Evaluation: The inclusion of multiple models and extensive data augmentation techniques ensures a thorough understanding of the strengths and weaknesses of each approach.

3. Limitations

Overfitting: The CNN model was worse at overfitting, which could suggest the requirement of further methods of regularization or more diversified training data.

Computational Constraints: It was endowed with limited computational resources and thus could not examine more complex models with larger batch sizes. This limitation affected the choice of parameter settings and the depth of models we could train.

Class Imbalance: Although data augmentation helped, this intrinsic imbalance in the dataset created many problems. Some classes had significantly fewer samples, and sometimes, predicting models are biased towards those classes, leading to reduced performance for underrepresented classes.

4. Future Improvements

Hybrid models: Hybrid models can also be explored, which have the benefits coming out of different architectures, like ResNet and DenseNet, for their performance advantage.

Better Computational Resources: The more powerful computational resources would allow one to investigate more complex models and, perhaps more importantly, larger parameter settings.

Reference：

[1] M. A. Kassem, K. M. Hosny, R. Damaševičius, and M. M. Eltoukhy, “Machine Learning and Deep Learning Methods for Skin Lesion Classification and Diagnosis: A Systematic Review,” Diagnostics, vol. 11, no. 8, p. 1390, Jul. 2021, doi: https://doi.org/10.3390/diagnostics11081390.

[2] A. Romero-Lopez, X. Giro-i-Nieto, J. Burdick, and O. Marques, “Skin Lesion Classification from Dermoscopic Images Using Deep Learning Techniques,” *Biomedical Engineering*, 2017, doi: https://doi.org/10.2316/p.2017.852-053.

[3] S. Benyahia, B. Meftah, and O. Lézoray, “multi-features extraction based on deep learning for skin lesion classification,” *Tissue and Cell*, vol. 74, p. 101701, Feb. 2022, doi: https://doi.org/10.1016/j.tice.2021.101701.