Melanoma Detection using Knowledge Distillation for Mobile Phones

Heejeong Yoon
School of Data Science
University of Virginia
Charlottesville, Virginia
rpk3ve@virginia.edu

Ryan Healy
School of Data Science
University of Virginia
Charlottesville, Virginia
rah5ff@virginia.edu

***bstract*—We explore using knowledge distillation to classify melanoma specifically for use in a model deployed on a mobile device.**

***Index Terms*—Melanoma Image Classification, Mobile deployment, Kaggle, H M10000, Edge deployment**

I. Introduction

Motiv tion

Convenient smartphone applications for detecting melanoma are becoming more popular. Despite improvements, the cost of misdiagnosing melanoma remains high; especially for elevated risk individuals with a personal or family history of skin cancer and those taking long-term immunosuppressive drugs.

We explore current benchmarks, architectures, and trade-offs in effective knowledge distillation. We propose a model that can work locally on most smart phones without the need for an Internet connection and focus on building user trust while promoting explainable I.

Method Overview

Knowledge distillation (KD) uses the teacher-student model paradigm, which involves training two different models.

To train the larger teacher model, we will use full fine-tuning on a frozen ResNet50 to reasonably approximate some benchmark learning. Given the high cost of missing a potential diagnosis of melanoma, we optimize for high model sensitivity and explore the trade-offs at a high level of model certainty.

To train a small and efficient student model, we will use offline knowledge distillation, where the larger and more powerful frozen teacher model "guides" a smaller phone-sized student model in learning; using both the teacher's outputs and hard ground truth labels as learning targets.

We explore the (KD) tuning temperature T, the mix weighting parameter , as well as different losses, and the effect of floating point quantization of the model weights on the performance of the unseen holdout set.

The approximate model sizes that we will target during model inference are 25 MB for mobile deployments and 2MB for edge deployments.

D t set

The data set on which we will train and evaluate is H M10000 - (Humans gainst Machines 10,000): https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000

This widely used benchmark data set contains (n = 10,015) images of expert-verified "cancerous" and "noncancerous" lesions in a wide variety of modalities.

The data card for H M10000 suggests using the "competition" website for access to a fair test set. We were unable to access this test set, and so we will instead use a standard 80-20% train and holdout split to fairly approximate general model performance.

We focus on cancerous melanoma (mel) and ignore similarly cancerous basal cell carcinoma (bcc). Our target class of melanoma is 11% of the total data.

Liter ture Review

Previous Methods

Several recent studies have used KD for Dermoscopy images.

Islam et al. used a large teacher (ResNet152V2, ConvNeXt-Base, ViT-Base, 236M parameters) to train a small student. Global max pooling outperformed average pooling. Simple knowledge distillation settings (T 1, 0.5) worked well on H M10000. [1]

Kabir et al. used smaller teachers (ResNet50, DenseNet161) and students (¡1M parameters), showing small models can perform well. [2]

Nazari et al. proposed compact attention-augmented EfficientNet-B3 models, achieving high accuracy with 98% fewer parameters and 20 faster inference. [3]

Gaps in urrent Research

We aim to address some of the following gaps in research

- Most KD studies report accuracy, F score, and model parameter size, but studies on sensitivity and model calibration are often missing. For both patients and providers, detection at a clear operating point matters most (e.g., precision / specificity at high sensitivity).

- Student models can inherit teacher bias, and discussion of that bias can be limited. To address this, we will focus on actionable uncertainty estimates and clear model calibrations.
- Performance of melanoma classification in edge devices is often lacking where power and memory constraints are even more essential. We expect to report models that can fit on edge device sizes and comment on their trade-offs

EXPERIMENT PL N

Teacher and Student Models

Our baseline teacher-student modeling begins with

Teacher: ResNet50 (ImageNet weights, global average pooling, frozen, fine-tuned).
Student: MobileNetV3-Small, lightweight head.
we will use the standard generalizable D M Optimizer Kaiming Weight Initialization

To validate successful teacher training, we will also benchmark that model against some more naive Classification models such as Logistic Regression, Gradient Boosted Machines, and Random Forests to motivate the complexity of the teacher model before Knowledge Distillation.

In ablation, we will explore different losses, student pooling layers, KD hyperparameters, and model hyperparameters. We elaborate on each in the following.

Different Losses and Pool Methods

For the Student Losses (L)

unweighted BCE - no opinion on how rare the target happens to be
weighted BCE - we give more weight to the relatively rare melanoma cases - exploring melanoma weights of [0.6, 0.7, 0.8]
focal loss - using γ - [1, 2, 4]

Variation in Pooling Layers to evaluate which best captures patterns for the smaller student model.

Global Max Pooling - to learn the most discriminative regions of the image
verage Pooling - which can focus on capturing overall patterns

KD Hyperparameters

How "strict" the teacher is with its probability predictions is governed by temperature T. higher T means that the teacher's probability distribution is more uniform and a lower T is more opinionated. [4]

The respective weights of the teacher/ground truth mixture of losses are governed by . [4]

We test how the "opinion" of the teacher model and the "mix" with ground-truth labels affect the quality of the student model. These are respectively governed by:

Temperature (T) - [1,2]
Mixing Level ( ) - [0.1, 0.5, 0.7, 0.9]

Standard Training Hyperparameters

We explore the following standard hyperparameters.

Learning Rate - [0.00001, 0.0001, 0.001, 0.01]
Batch Size - [16, 32, 64, 128, 256]
Dropout Rate pplied to the fully connected layers to reduce overfitting - [0.1, 0.2, 0.3, 0.4, 0.5]
Quantization - we explore floating point bit sizes of [4,8,16,32,64] to evaluate smaller models and their respective performance trade offs

The search space for these parameters is a starting point and will be adjusted based on model performance. Bit quantization currently depends on training with an NVIDI GPU, which is subject to availability, as GPU budgets can be variable.

EV LU TION PL N

Data and Splits

We will use H M10000 with binary labels showing melanoma and non-melanoma images. We split the data using the train/test/holdout method, where the original data are first divided into training and holdout sets, and then the training set is further divided into train/test splits. Each split is done using the standard 80/20 split method using reproducible seeds.

Given the smaller sizes of the data and models and the fact that classification generalization is so important in detecting melanoma, we will fit the training using K-fold target stratified cross-validation.

We might reduce K if the training time becomes prohibitive.

Training will be done using NVIDI GPUs on the UV Rivanna Cluster. If allocation is scarce, the training can still be done locally using a Mac M3 MPS chip, what can be done with Bit quantization will be limited.

Main Metrics

The main metrics for which we will optimize to demonstrate model quality include; Receiver Operating Curve - rea under the curve (ROC- UC), and Precision, rea Under the Curve (PR- UC). We choose the threshold for 95% sensitivity and then report the specificity, PPV, NPV at that threshold.

REPORTING

For each ablation, we report the following using the holdout scores for model comparison.

rea under the Curve (UC) scores
performance at 95% Sensitivity and at that threshold report Specificity, Positive Predicted Value (PPV), Negative Predicted Value (NPV)
Model Parameter Count
Floating-Point Operations Per Second (FLOPs)
F score
Model Latency
Model size (MB)

REFERENCES

[1] N. Islam, K. M. Hasib, F. . Joti, . Karim, and S. zam, "Leveraging Knowledge Distillation for Lightweight Skin Cancer Classification: Balancing ccuracy and Computational Efficiency," arXiv:2406.17051, 2024. vailable: https://arxiv.org/abs/2406.17051.

[2] M. R. Kabir, R. H. Borshon, M. K. Wasi, R. M. Sultan, . Hossain, and R. Khan, "Skin cancer detection using lightweight model souping and ensembling knowledge distillation for memory-constrained devices," Intelligence-Based Medicine, vol. 10, rt. no. 100176, 2024, doi:10.1016/j.ibmed.2024.100176.

[3] S. Nazari and R. Garcia, "Going smaller: ttention-based models for automated melanoma diagnosis," Computers in Biology and Medicine, vol. 185, rt. no. 109492, 2025. vailable: https://www.sciencedirect.com/science/article/pii/S0010482524015774.

[4] PyTorch. "Knowledge Distillation Tutorial". *PyTorch Tutorials*, https://docs.pytorch.org/tutorials/beginner/knowledge_distillation_tutorial.html. ccessed [October 24, 2025].