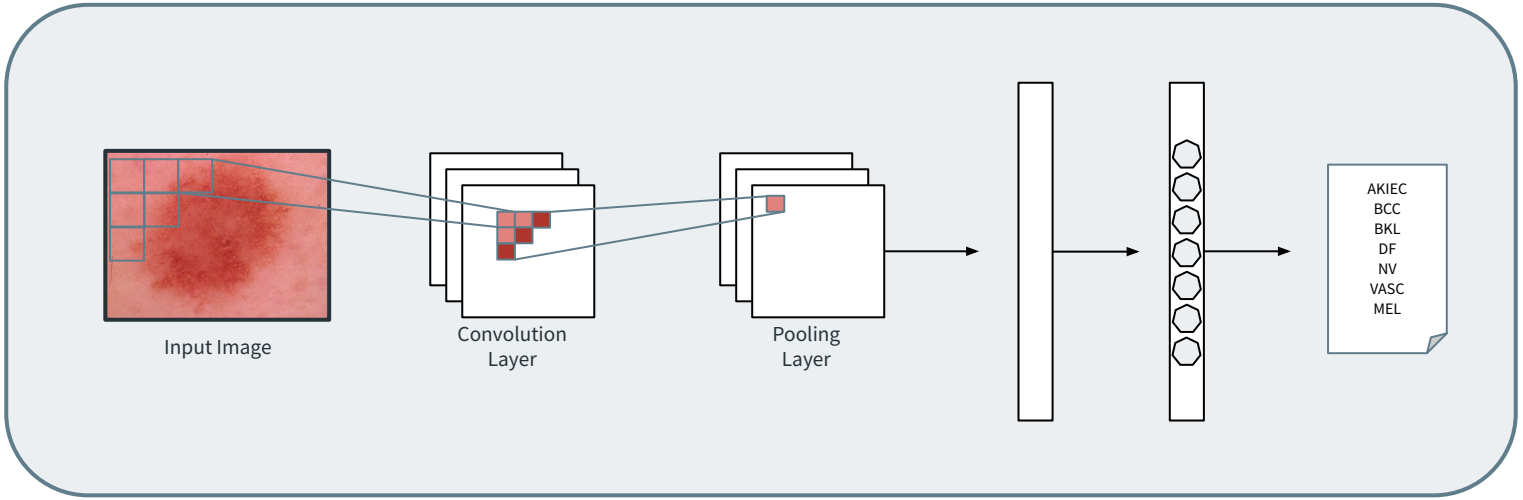
Skin Cancer Classifier - MNIST: HAM10000



Data Science 207

Spring 2023, Section 4, Team 4

207 Final Project Team

Fall 2023, Section 4, Team 4

Members

- Nathan Arias (nathanarias@berkeley.edu)
- Douglas Houghton (dchoughton@berkeley.edu)
- I-Hsiu Kao (ihsiukao@berkeley.edu)

Contributions

- Due to a shared interest in the entire process across our team, we each conducted an EDA and built a shared data set for model development. We then independently built models and iteratively tuned them, with frequent performance and insight check-ins throughout the process.

https://github.com/dchoughton/207_final_project

Dataset: MNIST HAM10000

Modified National Institute of Standards and Technology (MNIST) Human Against Machine (HAM) 10,000

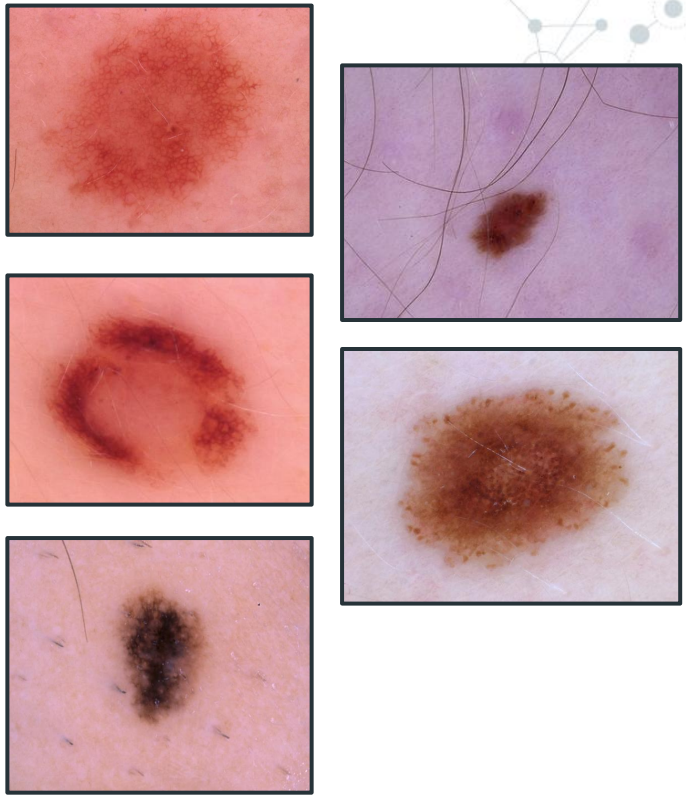
Training of neural networks for automated diagnosis of pigmented skin lesions is hampered by the small size and lack of diversity of available dataset of dermatoscopic images. We tackle this problem by releasing the HAM10000 ("Human Against Machine with 10000 training images") dataset. We collected dermatoscopic images from different populations, acquired and stored by different modalities. The final dataset consists of 10015 dermatoscopic images which can serve as a training set for academic machine learning purposes. Cases include a representative collection of all important diagnostic categories in the realm of pigmented lesions: Actinic keratoses and intraepithelial carcinoma / Bowen's disease (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (solar lentigines / seborrheic keratoses and lichen-planus like keratoses, bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv) and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage, vasc).

More than 50% of lesions are confirmed through histopathology (histo), the ground truth for the rest of the cases is either follow-up examination (follow_up), expert consensus (consensus), or confirmation by in-vivo confocal microscopy (confocal). The dataset includes lesions with multiple images, which can be tracked by the lesion_id-column within the HAM10000_metadata file.

MNIST HAM10000: Core Imagery

- 10,015 total images
  - All standardized to 600x450

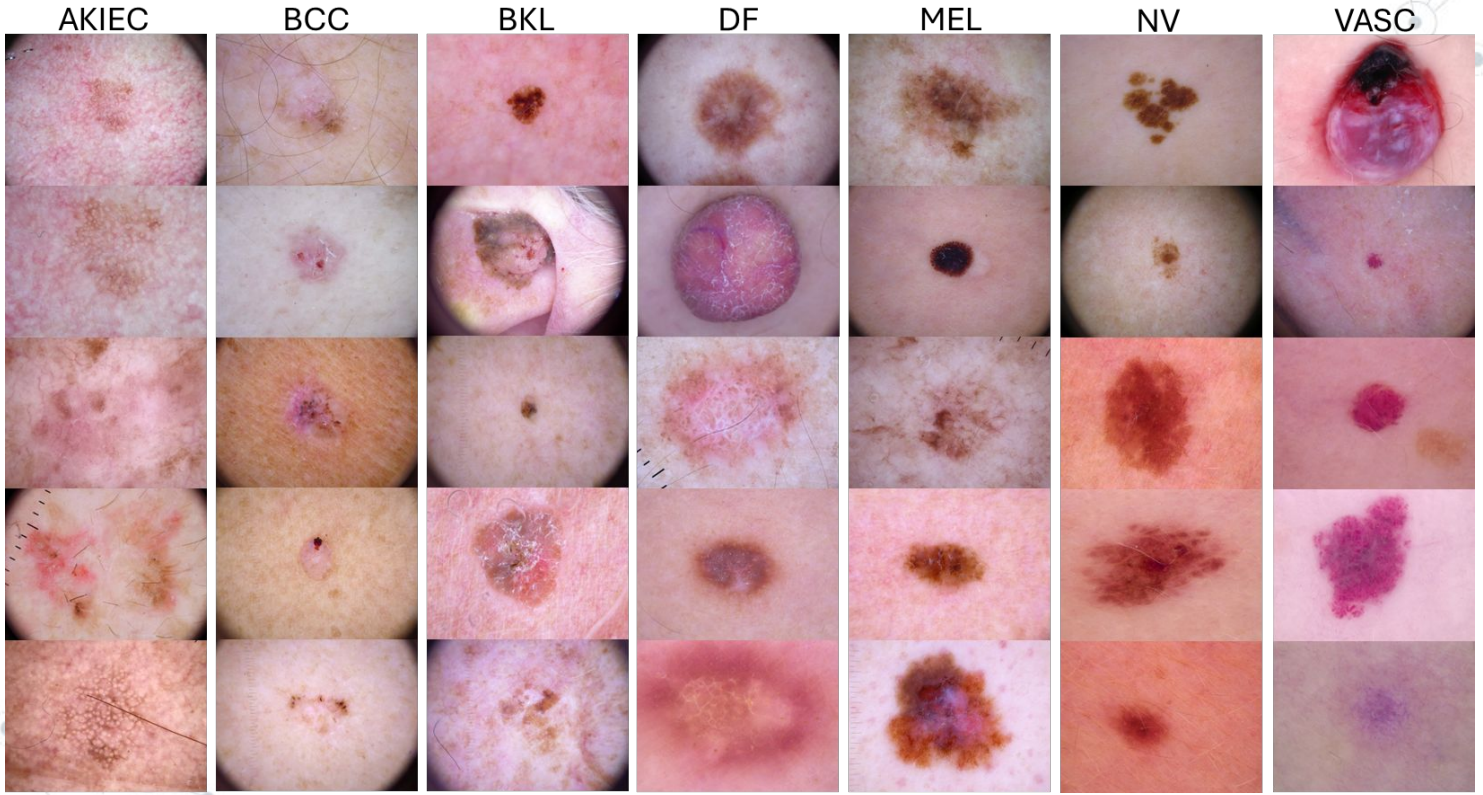
- 7 Different Lesion Types
  - AKIEC**:** Actinic Keratoses and Intraepithelial Carcinoma / Bowens disease
  - BCC**:** Basal Cell Carcinoma
  - BKL**:** Benign Keratosis-like Lesions
  - DF**:** Dermatofibroma
  - NV**:** Melanocytic Nevi
  - VASC**:** Vascular Lesions
  - MEL**:** Melanoma



The next slide has an array of **skin lesions** that may be considered graphic for some viewers.

(Trigger Warning!)

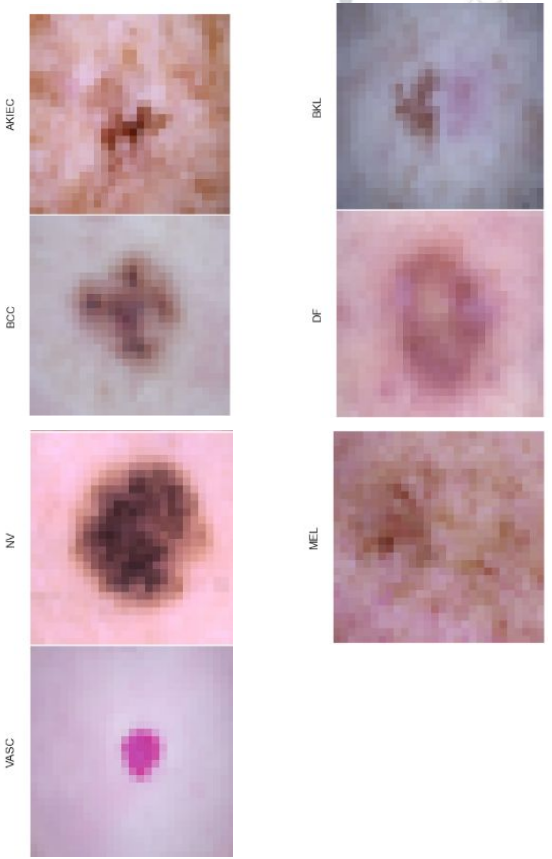
MNIST HAM10000: Imagery Variations



MNIST HAM10000: Additional Metadata

- Seven included metadata fields
  - lesion_id: unique identifier
  - image_id: name of corresponding image file
  - dx: diagnosis using six categories identified below
  - dx_type: method of diagnosis confirmation
  - **age:** age of image source patient
  - **sex:** sex of image source patient
  - **localization:** location of the lesion on patient

- Scaled 28x28 pixel RGB values (via Excel)



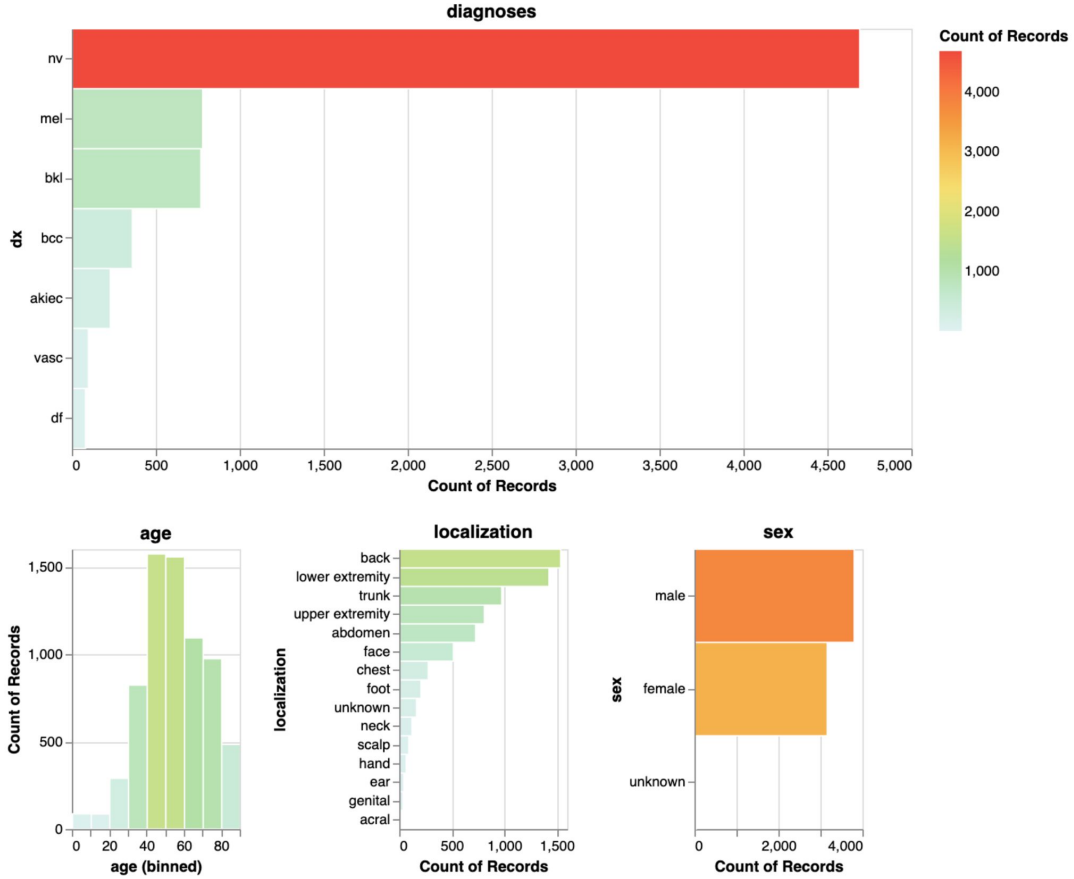
Problem Statement & Hypothesis

- **Problem Statement**
  - Machine learning models/ neural networks for automated diagnosis of pigmented skin lesions is hampered by the small size and lack of diversity of available dataset of dermatoscopic images.

- **Hypothesis**
  - Machine learning models can be applied to a variety of dermatoscopic pigmented skin lesion images with sufficient accuracy to assist physicians in the diagnosis and confirmation of several important diagnostic categories.

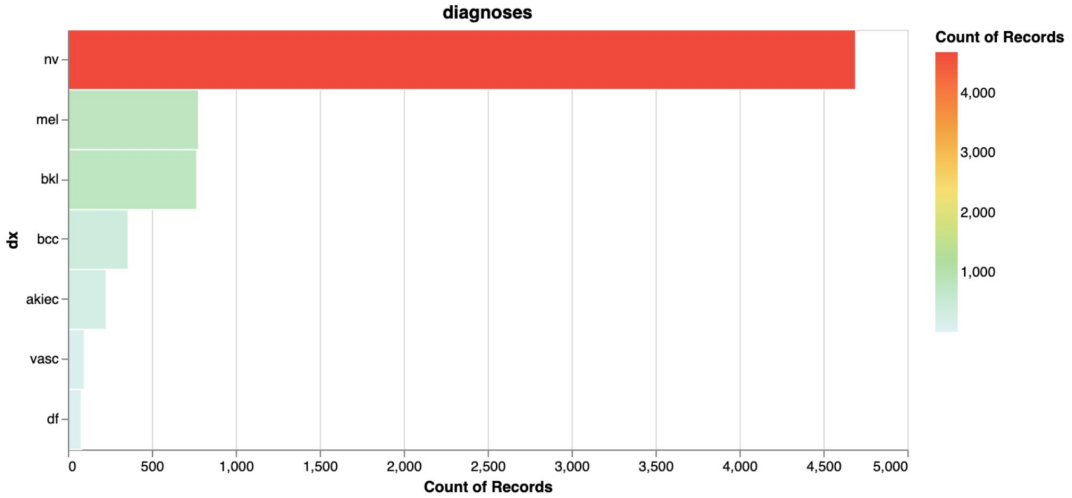
Training Set EDA

`akiec` - Actinic keratoses and intraepithelial carcinoma
`bcc` - basal cell carcinoma
`mel` - melanoma
`bkl` - benign keratosis-like lesions **(benign entitiy)**
`df` - dermatofibroma ***(benign cutaneous entity)***
`nv` - melanocytic nevi ***(mole)***
`vasc` - vascular lesions ***(birthmark)***



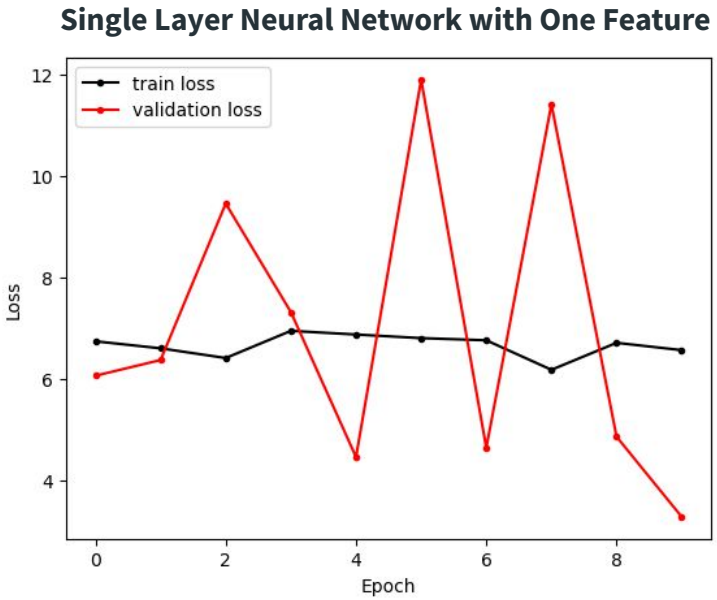
Baseline: Majority Model

- **Description:** Initial baseline model that simply assumes diagnosis of NV (66.86% of labels).
- **Accuracy:** 66.86%
- **Benefits:** Provides simple naive baseline for comparison.
- **Limitations:** Has no logic and simply assumes the most prominent class for all samples.



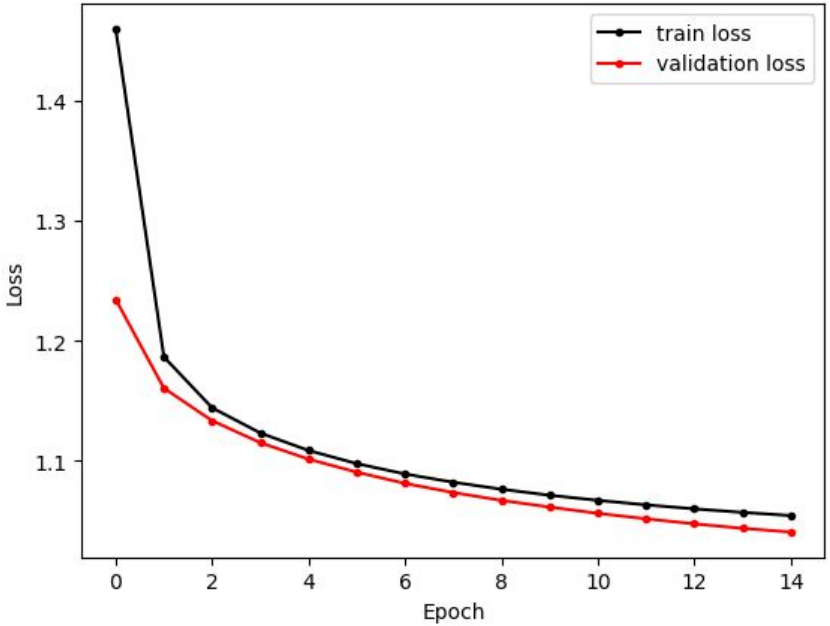
Iteration 1: Metadata Models

Features/Results	Loss	Val Loss	Accuracy	Val Accuracy
Age (single layer NN)	6.5578	3.2716	0.4638	0.6645
Age Bucket (single layer NN)	1.0632	1.0436	0.6696	0.6645
Age Bucket + Sex (single layer NN)	1.0635	1.0476	0.6696	0.6645
Age Bucket + Sex + Localization (single layer NN)	1.0544	1.0407	0.6696	0.6645
Age Bucket + Sex + Localization (Multi-layer NN)	0.9430	0.9356	0.6952	0.6936

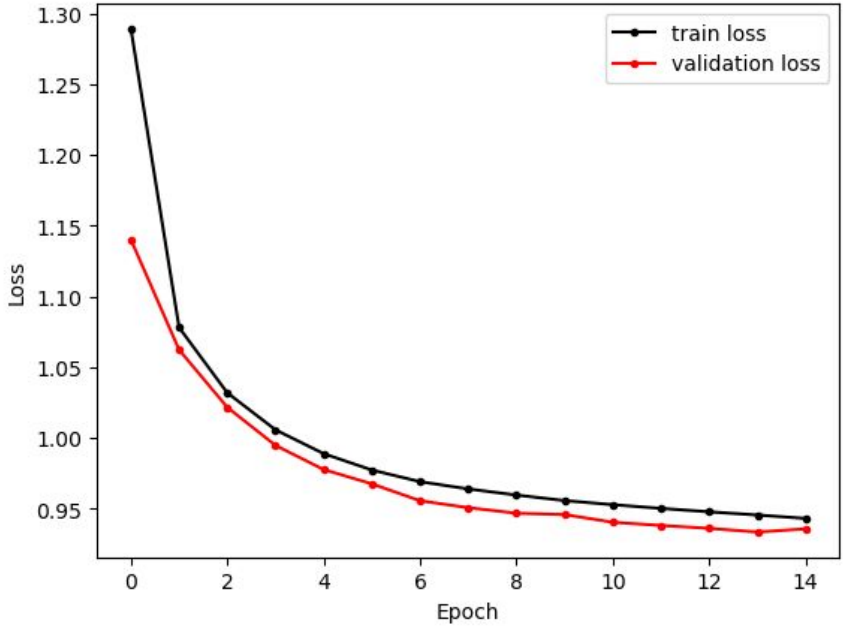


Iteration 1: Metadata Models

Single Layer Neural Network with Three Features



Multi-layer Neural Network with Three Features



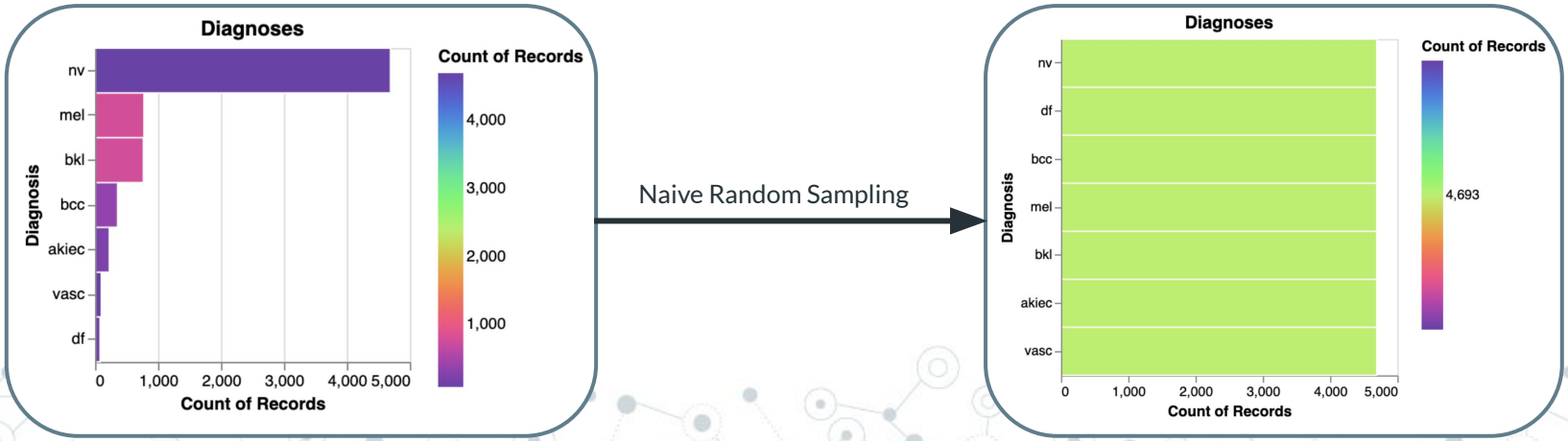
CNN Models

Balancing Minority Classes

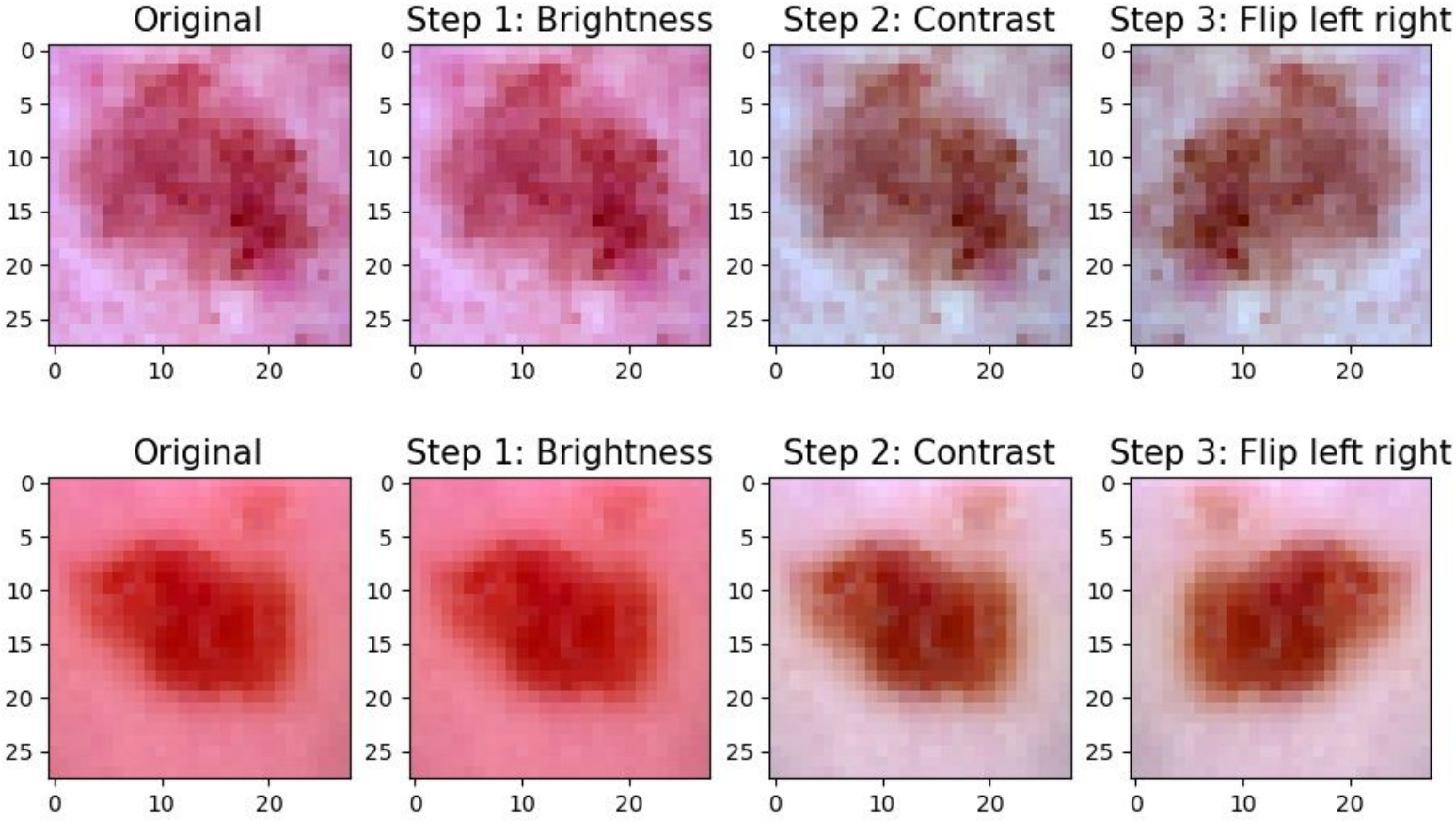
Original MNIST dataset heavily leans toward melanocytic nevi diagnosis. Using **Sklearn RandomOversampler** we balanced class labels.

Class Labels

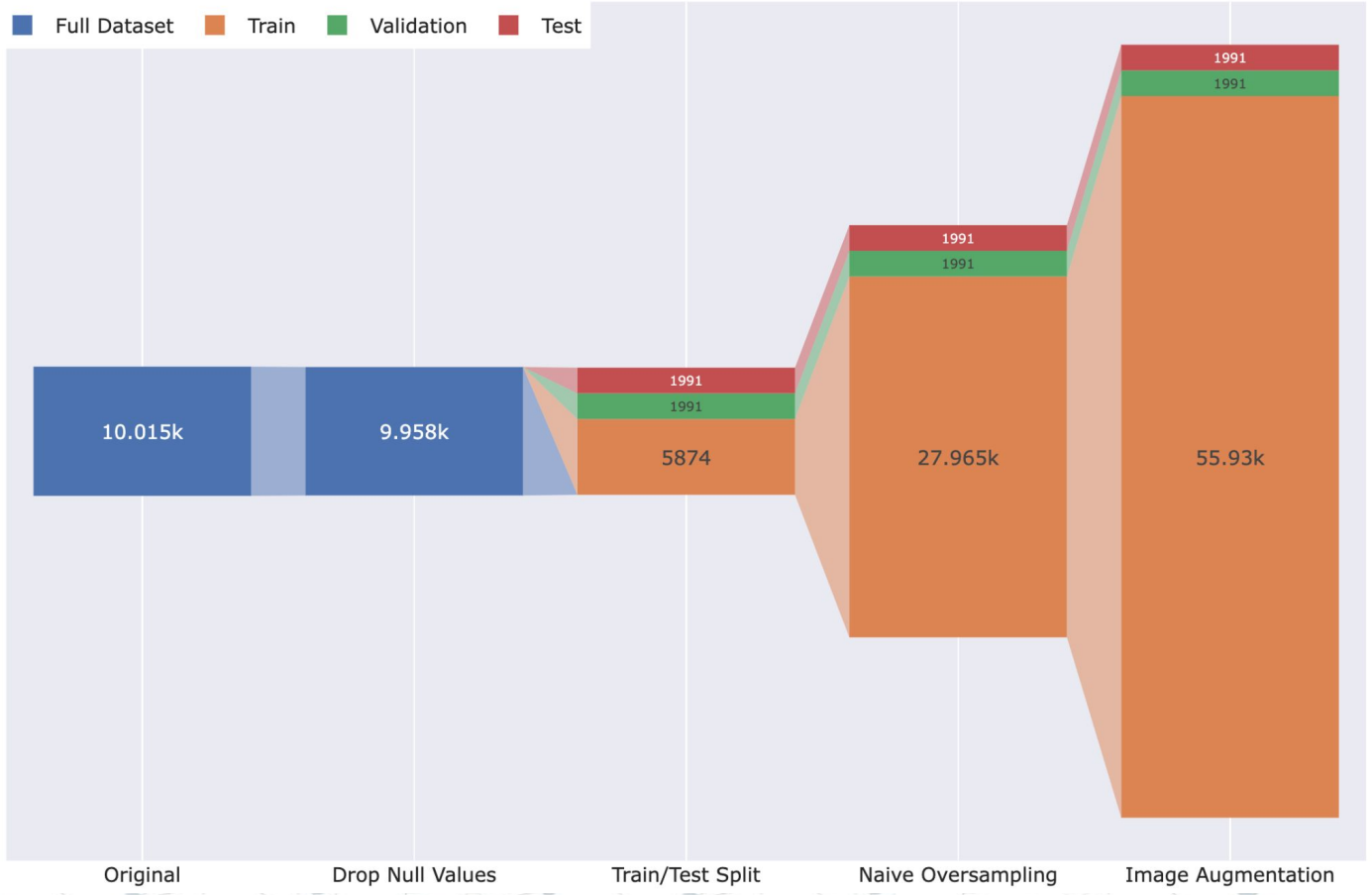
`akiec` - Actinic keratoses and intraepithelial carcinoma
`bcc` - basal cell carcinoma
`mel` - melanoma
`bkl` - benign keratosis-like lesions **(benign entitiy)**
`df` - dermatofibroma ***(benign cutaneous entity)***
`nv` - melanocytic nevi ***(mole)***
`vasc` - vascular lesions ***(birthmark)***



Data Augmentation

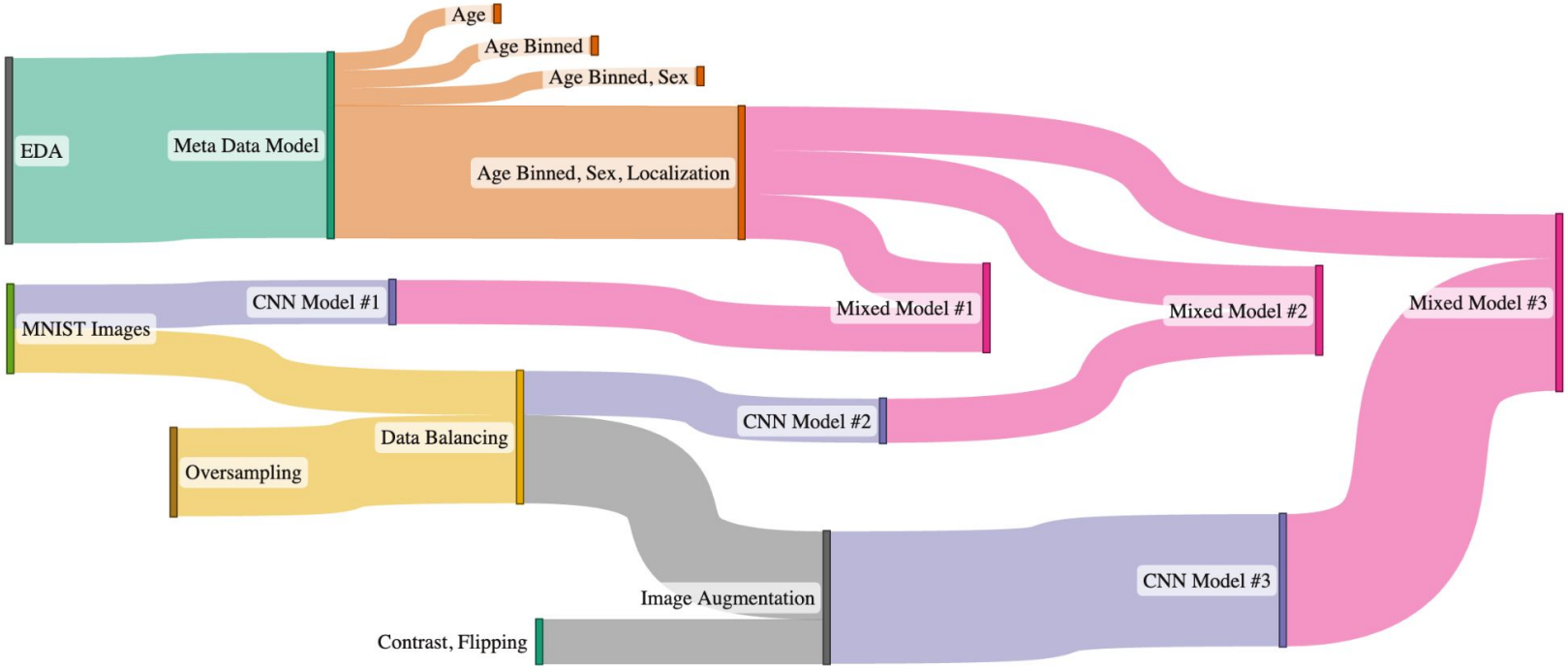


Data Accounting



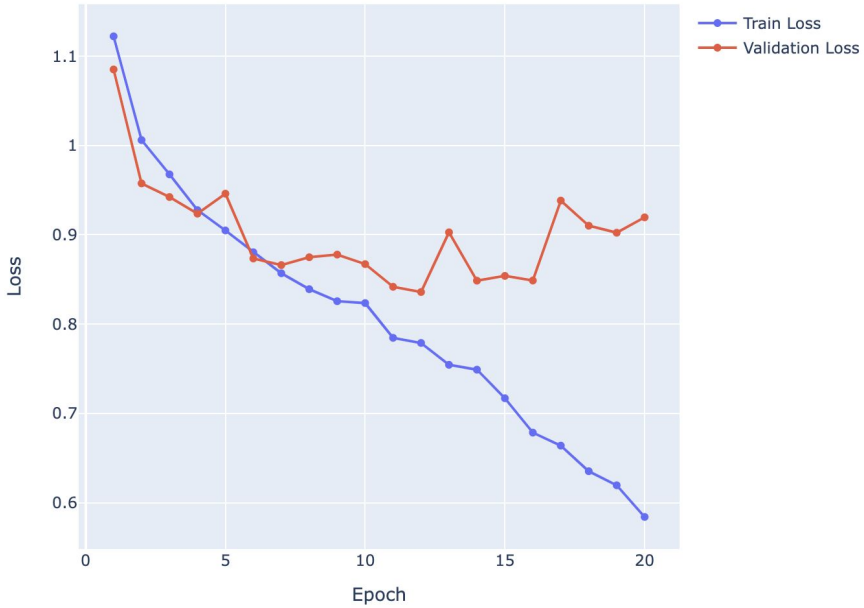
Strategy

Model Development Outline

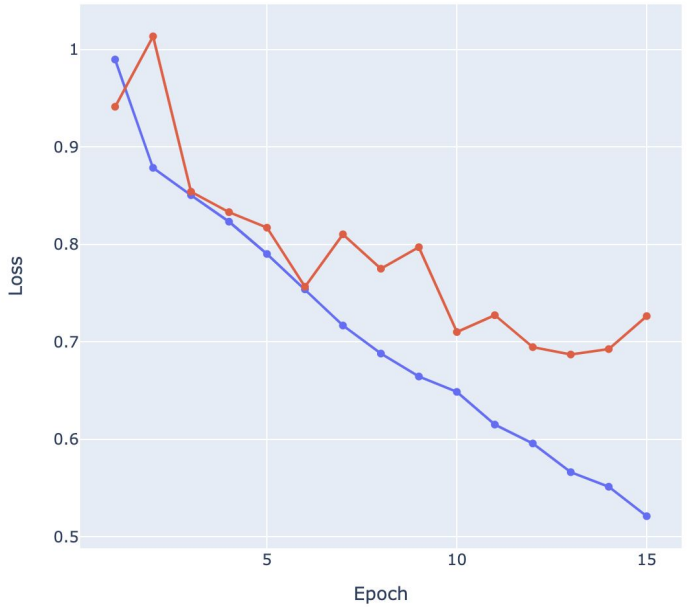


CNN Model - Grayscale vs. Color Images

Grayscale



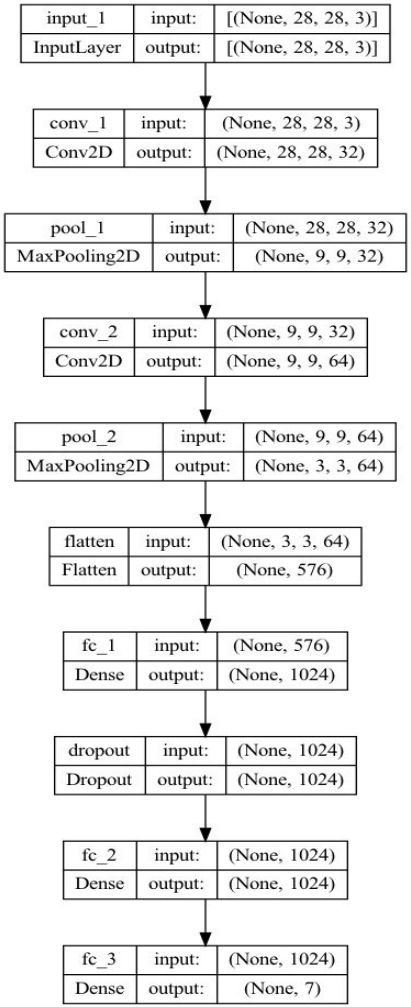
RGB



The incorporation of color into images provides more power to the CNN model

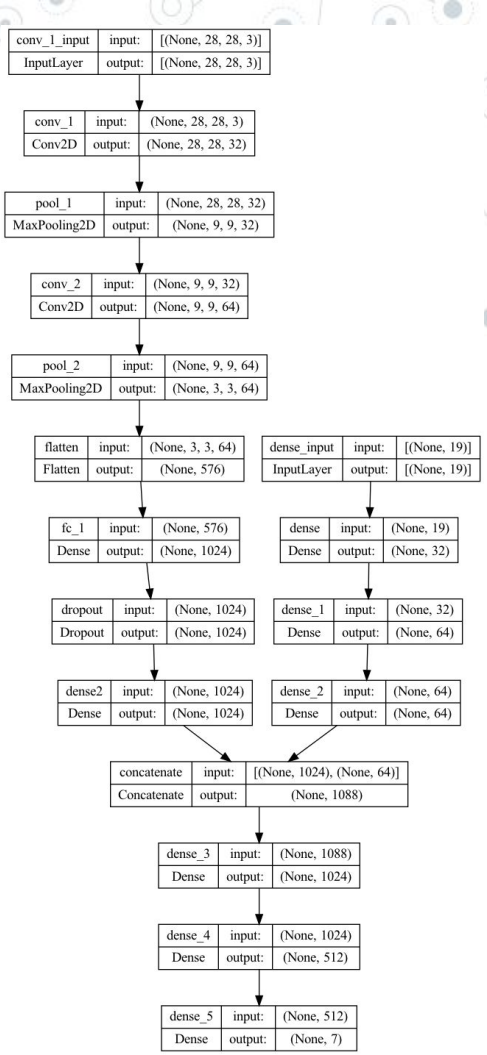
Results / Architecture

Features/Results	Loss	Val Loss	Accuracy	Val Accuracy
Grayscale	0.6106	0.8873	0.7869	0.7121
RGB Training Set	0.6751	0.7737	0.7508	0.7223
RGB with Oversampling	0.1014	1.2833	0.9652	0.7142
RGB with Image Augmentation	0.8094	0.8632	0.7034	0.7152
RGB with Oversampling and Image Augmentation	0.1462	1.2583	0.9498	0.7383



Mixed Model Results

Features/Results	Loss	Val Loss	Accuracy	Val Accuracy
RGB Training Set	0.5942	0.6009	0.7775	0.7835
RGB - Oversampling	0.1158	1.0701	0.9601	0.7544
RGB - Image Aug.	0.6145	0.6824	0.7658	0.7675
RGB - Oversampling + Image Aug.	0.1804	1.1344	0.9359	0.7388



Hyperparameter Tuning

Kernel Size & Pool Size

Kernel size: 1, 2, 3, 4, 5

Pool size: 1, 2, 3, 4, 5

Learning Rate

0.001, 0.01, 0.1, 1

Adam, SGD, Adagrad

Optimizer

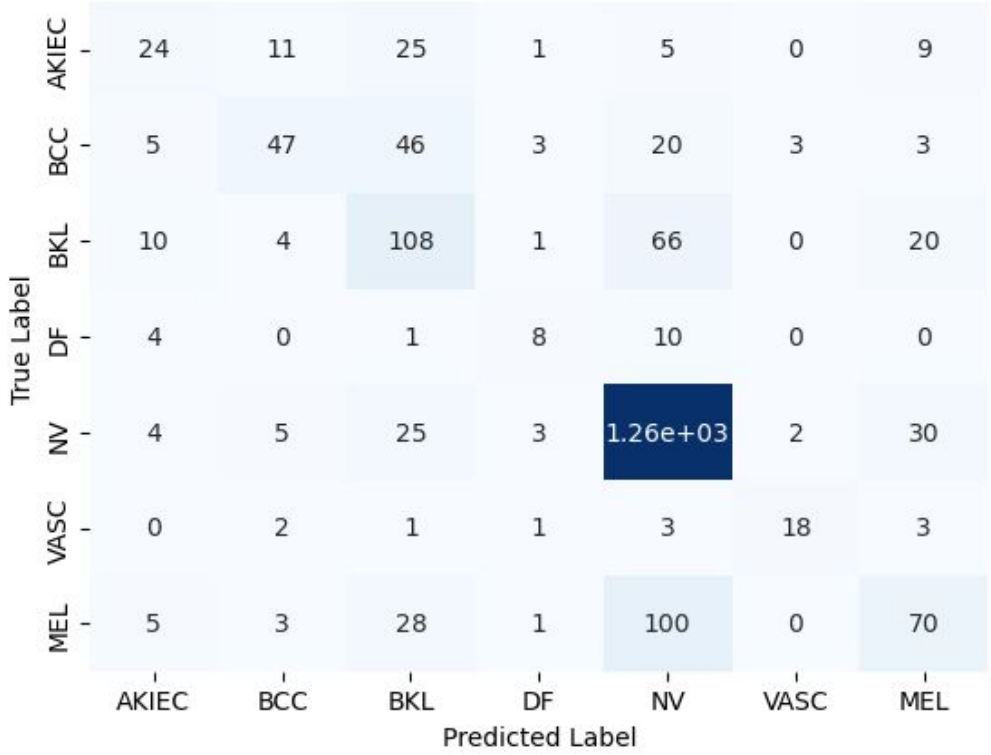
32, 64, 128, 256

Batch Size

Evaluating the Final Model

- **Final Key Hyperparameter Settings**
  - **Convolution layer kernel size: 3**
  - **Pooling layer pool size: 3**
  - **Learning rate: 0.001**
  - **Optimizer: Adam**
  - **Epoch: 10**

- **Result**
  - **Loss: 0.6118**
  - **Accuracy: 76.77%**

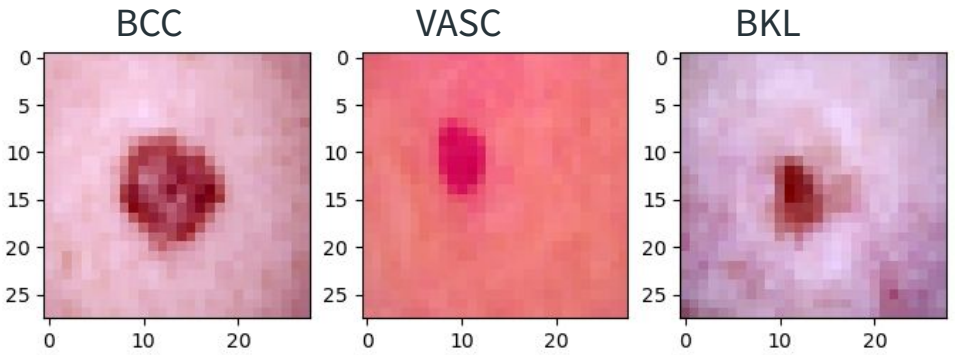


Understanding the Final Model

Label	Precision	Recall	F1-Score	Support
AKIEC	0.46	0.32	0.38	75
BCC	0.65	0.37	0.47	127
BKL	0.46	0.52	0.49	209
DF	0.44	0.35	0.39	23
NV	0.86	0.95	0.90	1324
VASC	0.78	0.64	0.71	28
MEL	0.52	0.34	0.41	207

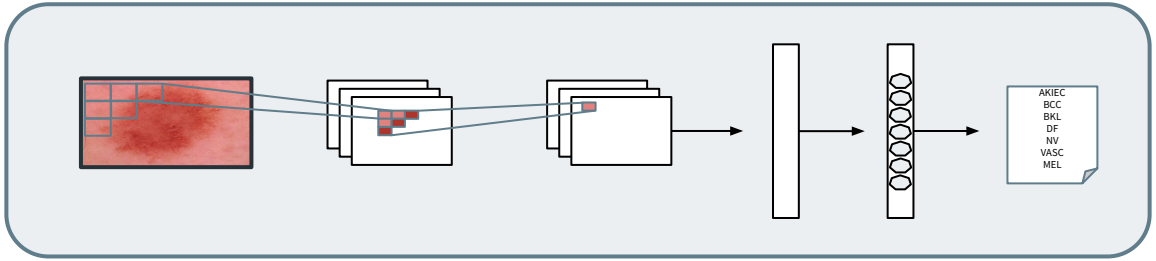
Metric	Precision	Recall	F1-Score	Support
Macro Avg	0.50	0.46	0.47	1503
Weighted Avg	0.73	0.75	0.73	1503

`akiec` - Actinic keratoses and intraepithelial carcinoma
`bcc` - basal cell carcinoma
`mel` - melanoma
`bkl` - benign keratosis-like lesions **(benign entitiy)**
`df` - dermatofibroma ***(benign cutaneous entity)***
`nv` - melanocytic nevi ***(mole)***
`vasc` - vascular lesions ***(birthmark)***



False Positive for NV

Future Work



- **Future Work for This Model**
  - **Convert to full image model.** With additional time and compute power, run the model using the more detailed full size imagery samples.
  - **Optimize sampling and augmentation.** Review methods and consider alternate strategies.
  - **Continued exploration of performance with additional model complexity.** Exploration in this project indicated additional convolutions may continue to improve the accuracy of the model.

- **Future Work for Problem Space**
  - **Expand sample imagery.** Expand underrepresented lesion types and skin tones.
  - **Expand metadata.** Additional non-image data points may improve mix model results.

NeurIPS Discussion: Societal Impact

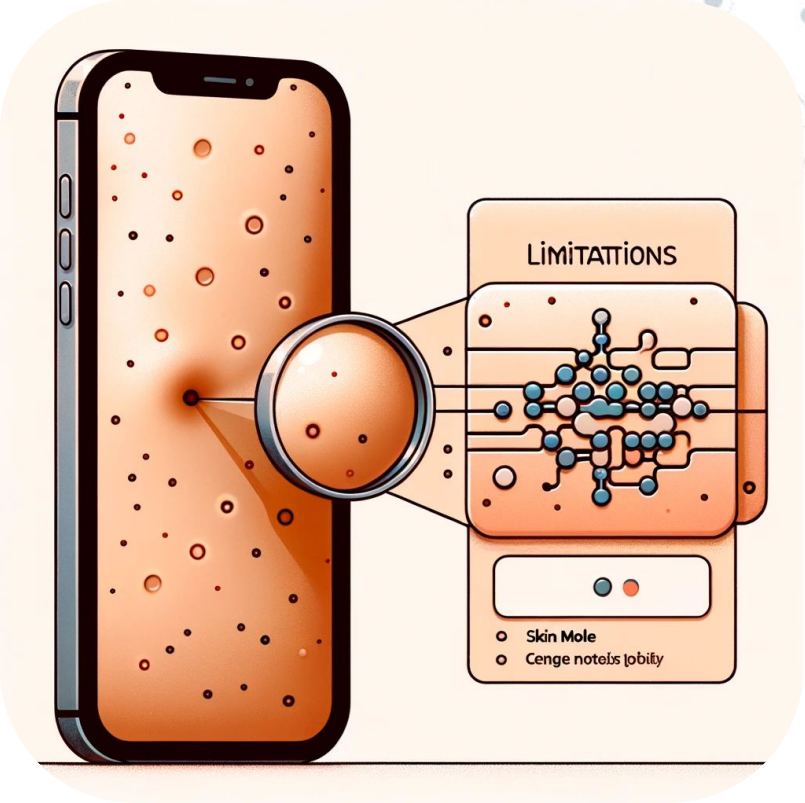
- **Positive Impacts**
  - **Equal access to expertise.** Providing access to model expertise to underserved populations.
  - **Increased diagnostic capacity.** Model can expedite physician assessment of condition.

- **Negative Impacts**
  - **Model Bias.** Limited set of skin tones, location of sampled individuals unknown.
  - **Reduced expert analysis.** Leaning on models to heavily could reduce expert analysis.

NeurIPS Discussion: Limitations

- **Dataset Limitations**
  - **Imagery distribution.** More limited samples of each lesion types.
  - **Universe size.** Limited scope of lesions types, only addresses seven specific variants.
  - **Naive Oversampling.** Consider using expert advice and more images of other labels.

- **Modelling Limitations**
  - **Compute.** Utilized lower resolution image input to accommodate compute



NeurIPS: General Compliance

- **YES.** Contributions are clearly stated.
- **YES.** Important assumptions and limitation are stated.
- **YES.** The ethics review guidelines were reviewed and conformed with.
- **YES.** Potential negative societal impact is discussed.
- **YES.** Potential limitations are discussed.
- **N/A.** Theoretical results are not a component of this project.
- **YES.** Code and instruction needed to reproduce experimental results are included in the identified Git repo.
  - **YES.** All training details are included (data splits, augmentation, hyperparameters, etc).
  - **YES.** Compute is discussed.
- **YES.** Data is accessible from identified Kaggle link.
  - **YES.** Data set creators are identified.
  - **YES.** Data license is identified.
  - **N/A.** No new assets nor supplemental material was incorporated.
  - **N/A.** Consent from individuals was not part of the data set requirements, this was provided.
  - **YES.** Personally identifiable information and privacy are discussed.
- **N/A.** No human research or crowdsourcing as directly a component of this project.

References

- **Citations**
  - **Dataset:** Tschandl, Philipp, 2018, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions", https://doi.org/10.7910/DVN/DBW86T, Harvard Dataverse, V4, UNF:6:KCZFcBLiFE5ObWcTc2ZBOA== [fileUNF]
  - **Kaggle:** Mader, K. S. (2018, September 20). Skin Cancer MNIST: HAM10000. Kaggle. https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000
  - **Dataverse:** Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, Allan Halpern: “Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)”, 2018; https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FDBW86T

- **NeurIPS Existing Assets Checklist**
  - **Creator(s):** Tschandl, P., Rosendahl, C. & Kittler, H. via MNIST and Harvard Dataverse
  - **License:** Creative Commons CC BY-NC-SA 4.0, https://creativecommons.org/licenses/by-nc-sa/4.0/
  - **New/ Additional Assets:** No additional assets were added to the existing dataset.
  - **Privacy:** Data identifies subjects by age and sex, no personally identifiable information is included.

Questions?