OPERATION SENTINEL

PROBLEM STATEMENT

AIM

Develop a **AI-based skin lesion detection system** that can analyze patient data & images to flag **malignant cases** early and escalate them to dermatologists.

WHAT WE NEED TO TACKLE

01 Rapid diagnosis from both metadata and image data.

02 Accuracy in detection to reduce false negatives.

03 Scalability to work in low-resource and remote settings.

04 Seamless integration for dermatologist review.

INFORMATION PROVIDED

METADATA

01 Patient Data:
- Patient ID
- Age
- and Sex.

02 Lesion Info:
- Lesion ID
- site of occurrence
- max diameter of the lesion
- confidence score of other Neural Networks
- lighting modality, etc.

03 Diagnosis Analysis: Data about multiple levels of classification

04 Other Technical Details:
- Light Modality, Chroma,
- Eccentricity, Color variation, etc.

LESION IMAGES

01 Well-lit images for the lesion to be inspected
- These can be identified by the ISIC Identifiers provided for each row

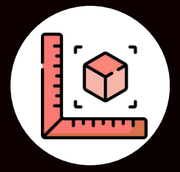
02 Image Size:
- 139x139 RGB images are provided. These prove to be too large and memory-intensive for efficient training.
- We reshape these images to 64x64 and use data augmentation in order to deal with class imbalance

EDA

KEY INSIGHTS

EXTREME IMBALANCE

Malignant cases are rare (~0.1%), requiring careful model training.



LESION CHARACTERIESTICS

Size, color, and geometry are strong indicators.

DATA LEAKAGE

Diagnostic columns must be removed from training to prevent bias.

OUTLIERS

Several numeric columns contain extreme outliers (very large area/perimeter values, odd coordinate ranges)

DATASET SNAPSHOT AND CLASS IMBALANCE

The dataset contains 400,666 benign cases versus only 393 malignant cases, representing an extreme class imbalance of approximately 0.1% malignant.

This imbalance is critical for model training, as a naive model could achieve high accuracy by simply predicting "benign" for all cases.

CRITICAL: DATA LEAKAGE AND QUALITY ISSUES

Leakage Risk

Features like iddx_* and diagnosis-like columns directly encode labels, leading to artificially high performance in training but failure in real-world screening.

Action Required

These columns must be identified and removed from model training. Treat them as evaluation metadata only.

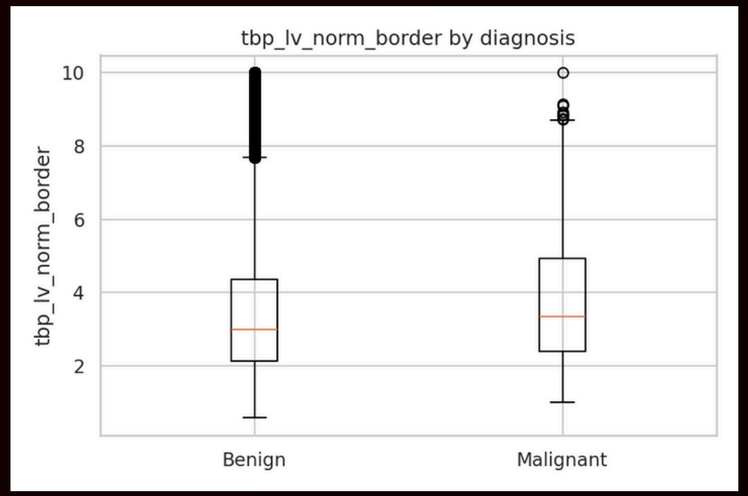
Outlier Management

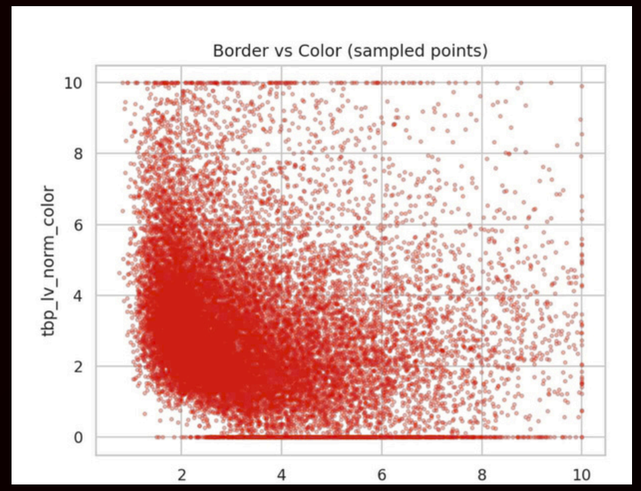
Extreme outliers in numeric columns (e.g., area/perimeter) need to be cleaned or winsorized to prevent model instability.

KEY UNIVARIATE AND BIVARIATE FINDINGS

LESION SIZE (DIAMETER)

Malignant lesions are larger median (5.14 mm) compared to benign (3.37 mm), with a longer upper tail.

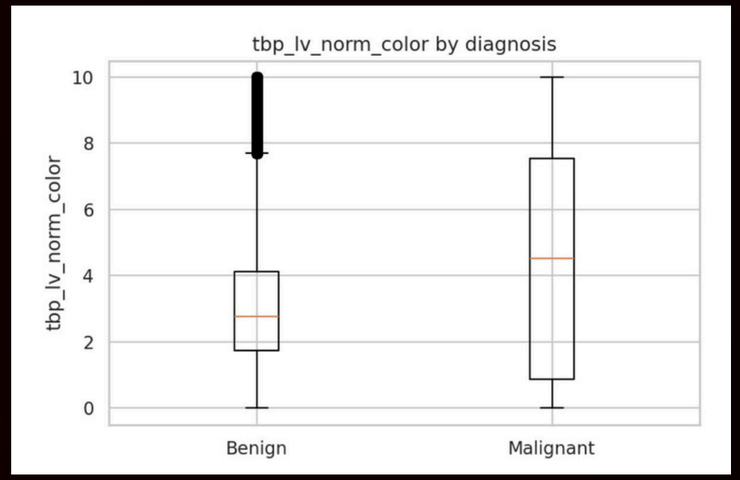


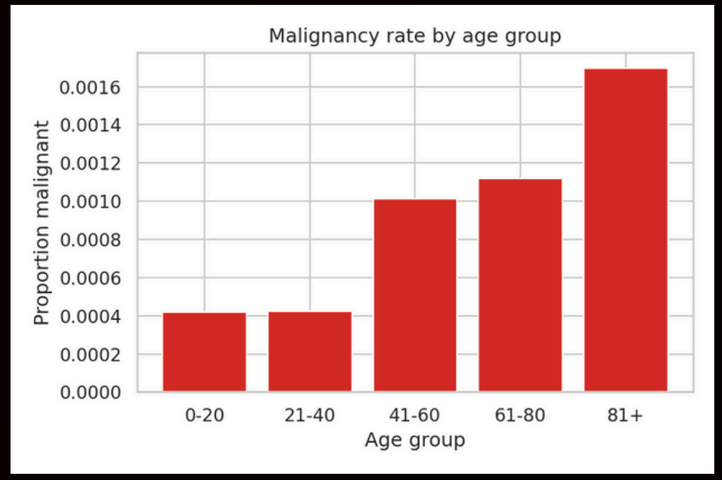


Lesion diameter (mm): malignant lesions are larger on median.

COLOR & BORDER IRREGULARITY

Malignant lesions show higher median color heterogeneity (4.53 vs 2.76) and border irregularity (3.35 vs 2.99), aligning with clinical ABCDE rules.



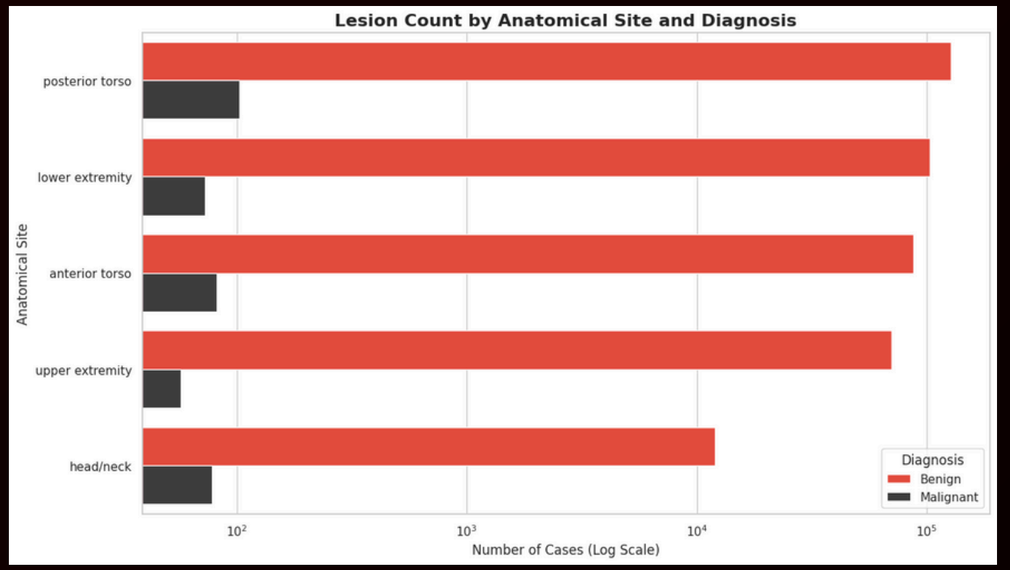


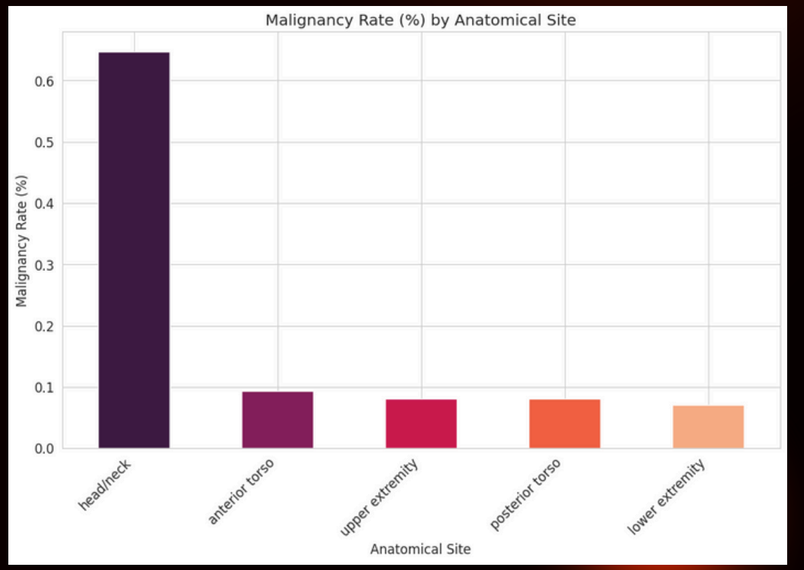
Color heterogeneity and border irregularity — both elevated in malignancy.

AGE AND ANATOMICAL SITE IMPACT

ANATOMICAL SITE

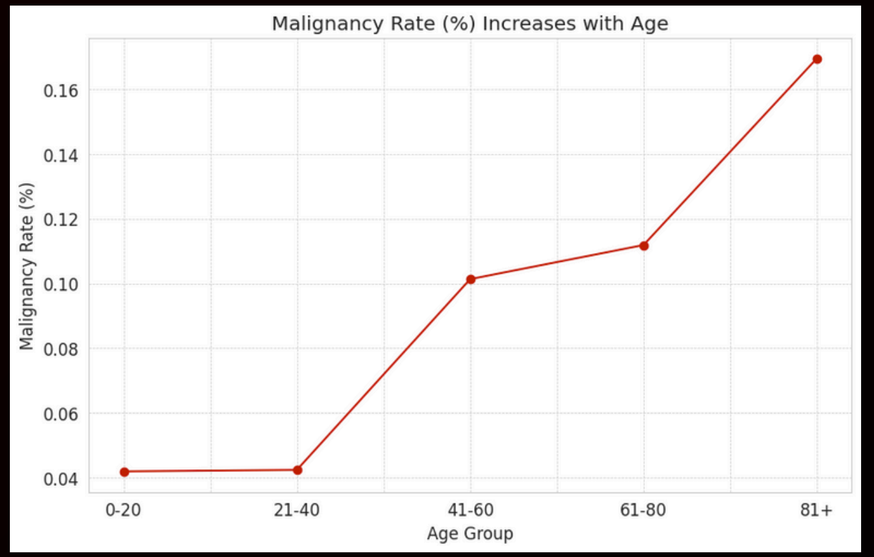
Head/neck lesions show the highest site-specific malignancy rate (~0.65%), suggesting sun exposure and cumulative damage as drivers.





AGE DISTRIBUTION

Malignancy proportion increases with age, especially after 50. Median age for both benign and malignant cases is ~60 years



MISSION & APPROACH

INITIAL APPROACH

- **Reasoning:**

  We initially assumed that most predictive power would come from the image itself, as cancer diagnosis is largely visual

- **Models Tried:**

  CNN (baseline) - basic convolution layers to test feasibility.
  ResNet - deeper network with skip connections to prevent vanishing gradients.
  EfficientNet - scaled network balancing depth, width, and resolution.

- **Setup:**

  Only image data, no tabular features.
  Classification into cancerous vs non-cancerous

- **Goal:**

  Test if deep CNNs could learn discriminative visual features without additional data.

DATASET CHALLENGES

1. Ambiguous Images

Many samples were so unclear that even humans couldn't confidently label them.
Action: Removed these to avoid confusing the model.

2. False Visual Cues

Some images looked cancerous but had non-cancerous labels (label noise).
Action: Kept these to mimic real-world diagnostic uncertainty.

3. Class Imbalance

Large difference between cancerous and non-cancerous sample counts.
Action: Used WeightedRandomSampler to ensure balanced training batches.

TRANSITION TO HYBRID MODEL

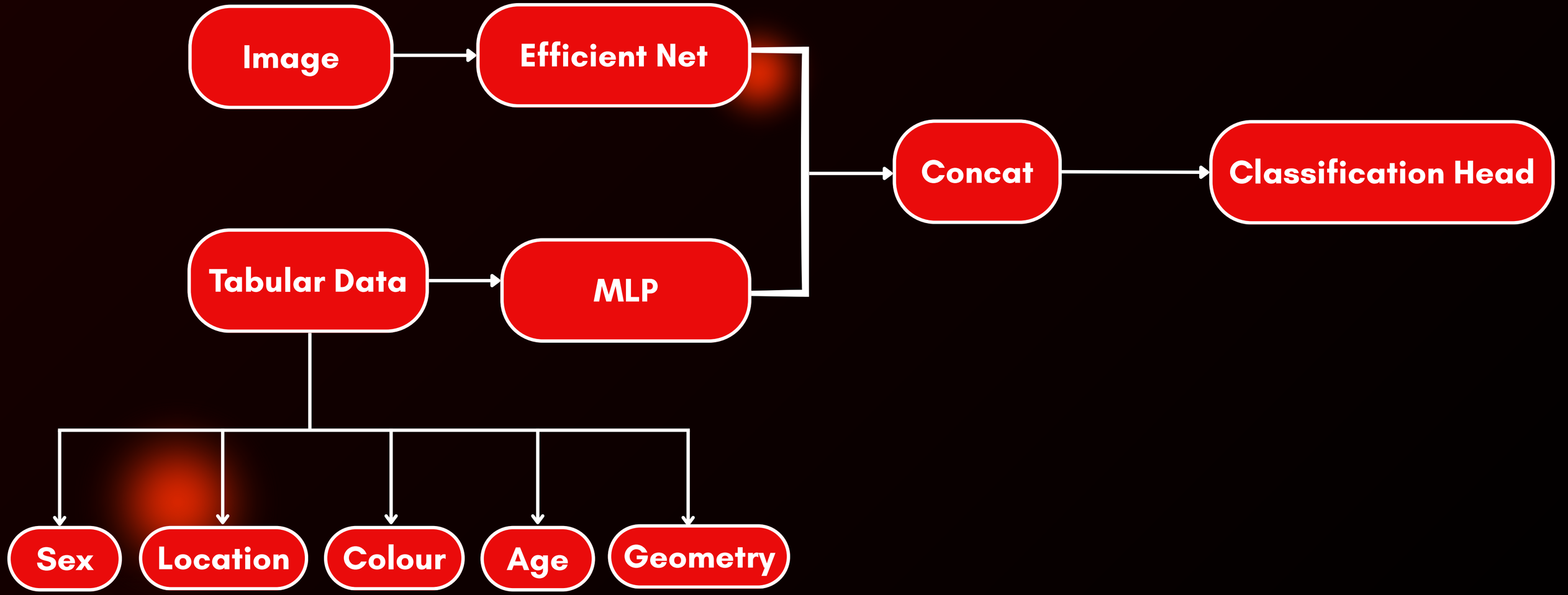
Image-only models plateaued in accuracy.
Tabular features could provide contextual signals not visible in images (e.g., patient age, lesion location, colour).
We tried out a few hybrid approaches involving a pre-trained Vision Transformer

Proposed Solution: Combine image features with tabular embeddings.

ARCHITECTURE

- EfficientNet backbone for high-quality image features.
- Tabular data → Embedding layer to turn numeric/categorical features into a learnable vector.
- Concatenate both vectors.
- Pass through fully connected layers → final classification.

ARCHITECTURE OF THE MODEL



MODEL OVERVIEW

Image Input (B, 3, H, W)

Efficient Net Backbone

[B, in_ch, 1, 1] → Flatten → [B,in_ch]

Meta features

Input: [B, n_meta_features]

Linear (n_meta_features → 512)

BatchNorm1d + Swish

Dropout (p=0.3)

Linear (n_meta_dim[0] → 128)

BatchNorm1d + Swish

Concatenated Feature Vector [B, in_ch + meta_output_dim]

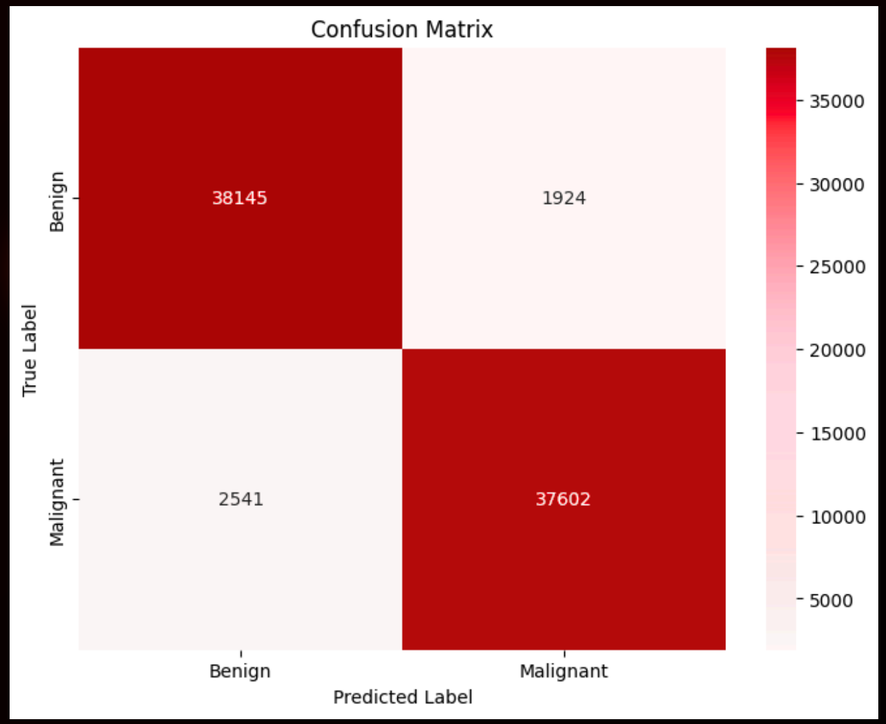
Dropout (p=0.5) + Linear FC (same weights each time)

Average logits over 5 dropout passes

Output: [B, 2]

METRICS

F1 SCORE	0.9440
RECALL	0.9367
ACCURACY	0.9443
POSITIVE PREDICTIVE VALUE	0.9513
NEGATIVE PREDICTIVE VALUE	0.9375



MEET OUR TEAM

AKSHAT SHUKLA

TADI JOSHUA RAJ

MURARI MARUPU

ARNAV GAWADE

KUMAR SAKCHHAM

LIKHITHA TUGITI

SIDDHANT KODOLKAR

CHIRAG KHANDELWAL

DEBOSMITA ROY

REFERENCES

1) HTTPS://ARXIV.ORG/PDF/2106.00131

CLUSTERING-FRIENDLY REPRESENTATION LEARNING VIA INSTANCE DISCRIMINATION AND FEATURE DECORRELATION

Yaling Tao, Kentaro Takagi & Kouta Nakata
Corporate R&D Center, Toshiba
1, Komukai Toshiba-cho, Saiwai-ku, Kawasaki, Kanagawa, Japan
{yaling1.tao,kentaro1.takagi,kouta.nakata}@toshiba.co.jp

2) HTTPS://PUBMED.NCBI.NLM.NIH.GOV/38600009/

Data cleaning for clinician researchers: Application and explanation of a data-quality framework

Julia K Pilowsky [1], Rosalind Elliott [2], Michael A Roche [3]

3) HTTPS://PMC.NCBI.NLM.NIH.GOV/ARTICLES/PMC11607306/#SEC2

Classification of melanoma skin Cancer based on Image Data Set using different neural networks

Rukhsar Sabir [1,✉], Tahir Mehmood [1]

4) HTTPS://JMLR.ORG/PAPERS/V24/21-1518.HTML

HiClass: a Python Library for Local Hierarchical Classification Compatible with Scikit-learn

Fábio M. Miranda, Niklas Köhnecke, Bernhard Y. Renard; 24(29):1–17, 2023.

5) HTTPS://ARXIV.ORG/PDF/2106.00

Identifying Melanoma Images using EfficientNet Ensemble

October 2020
DOI: 10.48550/arXiv.2010.05351

Qishen Ha · Bo Liu · Fuxu Liu

THANK YOU