COM6003

SKIN CANCER DETECTION

Team ID: GROUP 4

Team Member：

GE Yuxin, P233336

KWOK Tsz Yi, P233340

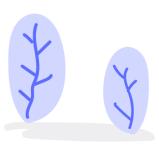LEI Linlin, P233341

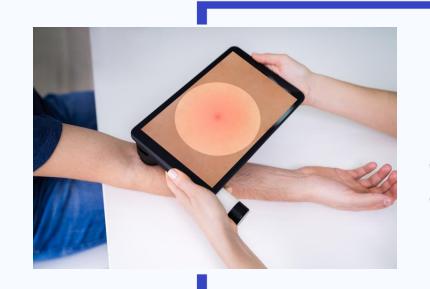WANG Yanyi, P233361

CONTENTS

Machine learning pipeline

01
Background & Objectives

• **Background**

Challenges in Hong Kong's Public Healthcare:

• **Extended Wait Times:** Patients in Hong Kong face long delays, sometimes over a year, for initial dermatology consultations in public hospitals.

• **Impact on Skin Cancer:** This delay is critical as non-melanoma skin cancer is one of the top 9 most common cancers in Hong Kong, according to the 2020 Hong Kong Cancer Registry.

Technological Advancements in Diagnosis:

• **Impact on Skin Cancer:** This delay is critical as non-melanoma skin cancer is one of the top 9 most common cancers in Hong Kong, according to the 2020 Hong Kong Cancer Registry.

Is there a possibility of developing a computer vision based pre-screening tools for patients to have a preliminary screening on skin cancer?
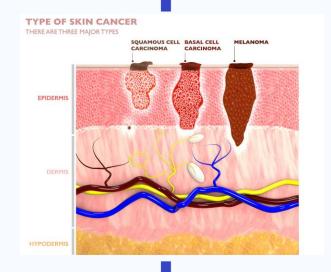
Objectives

Our project aims at **investigate the potential of developing a computer vision system with combination of patient information** to **enable earlier detection of skin cancer.**

This initiative aims to alleviate the healthcare system's load and improve patient outcomes by:

1. **Enabling proactive monitoring through photographic assessments**

2. **Significantly reducing diagnosis timelines**

3. **Facilitating quicker medical interventions, potentially lowering mortality and morbidity from late-stage diagnose**

02

**Data collection &
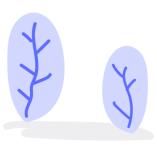Evaluation**

Data collection

Data from:

https://www.kaggle.com/datasets/mahdavi1202/skin-cancer

Published: 7 Jul 2020

Includes:

- metadata.csv (2298rows & 26 columns)

- 2298 images.

| No.	Attribute Name	Description
1	patient_id	Identifier of the patient under study.
2	lesion_id	Identifier of the lesion or wound under study in the patient.
3	smoke	Whether the patient has a history of smoking or not.
4	drink	Whether the patient has a history of alcohol consumption or not.
5	background_father	History of any diseases or health conditions related to the patient's father.
6	background_mother	History of any diseases or health conditions related to the patient's mother.
7	age	Age of the patient at the time of examination.
8	pesticide	Whether the patient has been exposed to pesticides or other chemicals.
9	gender	Gender of the patient.
10	skin_cancer_history	History of skin cancer in the patient's family.
11	cancer_history	History of cancer in the patient's family.
12	has_piped_water	Indicates whether the patient's residence has access to piped water.
13	has_sewage_system	Indicates whether the patient's residence has a proper sewage system.
14	fitspatrick	Skin tolerance to sunlight.
15	region	The area of the body where the lesion or wound has been examined.
16	diameter_1	Primary diameter of the lesion or wound.
17	diameter_2	Secondary diameter of the lesion or wound.
18	diagnostic	The type of lesion or wound diagnosed.
19	itch	Whether the lesion or wound has itched or not.
20	grew	Whether the size of the lesion or wound has grown or not.
21	hurt	Whether the lesion or wound has hurt or not.
22	changed	Whether the appearance of the lesion or wound has changed or not.
23	bleed	Whether the lesion or wound has bled or not.
24	elevation	Description of the lesion or wound relative to the skin surface.
25	img_id	Identifier of the image related to the lesion or wound.
26	biopsed	Whether the lesion or wound has been biopsied or not.

Data Evaluation
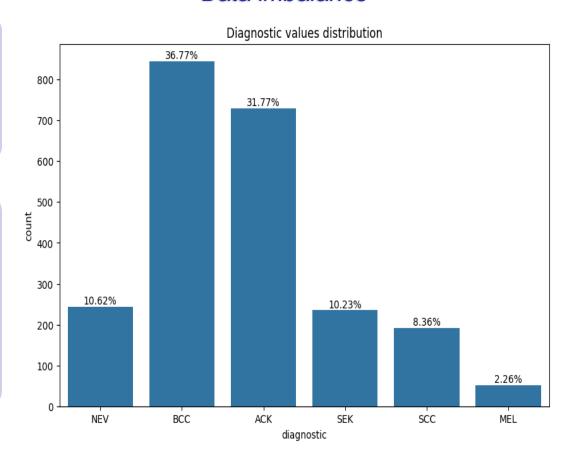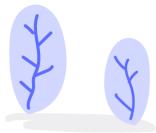
There were a total of 1,373 patients with 1,641 skin lesions in the dataset. Each image/sample has a reference to the patient and skin lesion in the metadata. There are situations where a patient has multiple skin lesions, multiple skin lesions, and corresponds to multiple pictures.

Skin lesions include: basal cell carcinoma (BCC), squamous cell carcinoma (SCC), actinic keratosis (ACK), seborrheic keratosis (SEK), melanoma (MEL) and nevus (NEV) 6 There are three kinds of skin lesions, there is the problem of sample data imbalance, three skin cancers (BCC, MEL and SCC) and three skin diseases (ACK, NEV and SEK). Of these, all BCCs, SCCs, and MELs were biopsy proven, with a total of approximately 58% of samples being biopsy proven.

Data imbalance

Diagnostic values distribution

03

**Data processing &
Feature engineering**

Data preprocessing

CSV data:

1. Use -1 constant value to fill in missing values
2. Combine upsampling and undersampling to deal with imbalance

```python
pipe = make_pipeline(
    SMOTE(sampling_strategy=over_sample_strategy),
    NearMiss(sampling_strategy=under_sample_strategy)
)
```

3. Delete meaningless columns ('patient_id', 'lesion_id', 'img_id') and columns that may cause overfitting ('biopsed')

Missing values percentage in columns with missing values

A total of 13 columns have missing values, and the proportion of missing values is about 35%.
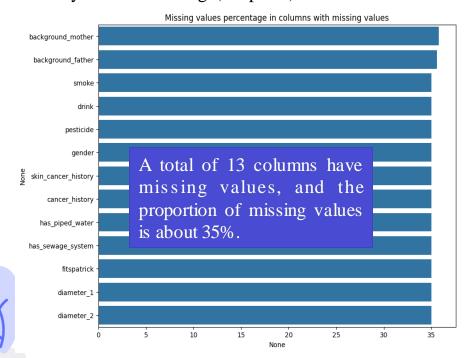
IMAGE DATA:

1.Reading->Resizing -> Normalization

```python
IMG_SIZE = (256, 256)
BATCH_SIZE = 16
# Reading -> Resizing -> Normalization
    def img_preprocessing(img, label):
        """ Image preprocessing function """
        img = tf.io.read_file(img)  # Read the image file
        img = tf.image.decode_png(img, channels=3)  #
Decode the JPEG image
        img = tf.image.resize(img, img_size)  # Resize the
image
        img = tf.cast(img, tf.float32) / 255.0  #
Normalize pixel values to [0, 1] range
        return img, label
```
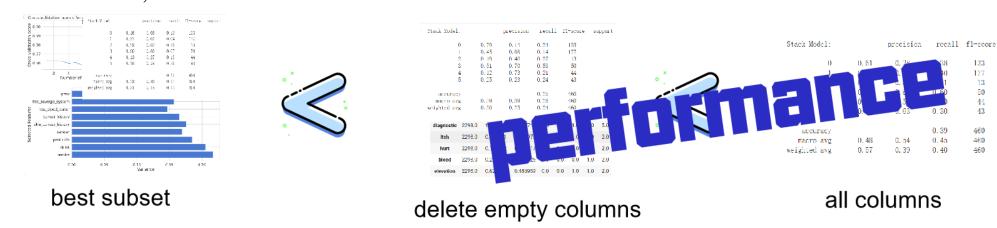
2.Use undersampling to deal with imbalance problems

```python
class_counts = Counter(y)
min_class_count = min(class_counts.values())
under_sample_strategy = {label: min_class_count for
label in class_counts.keys()}
```
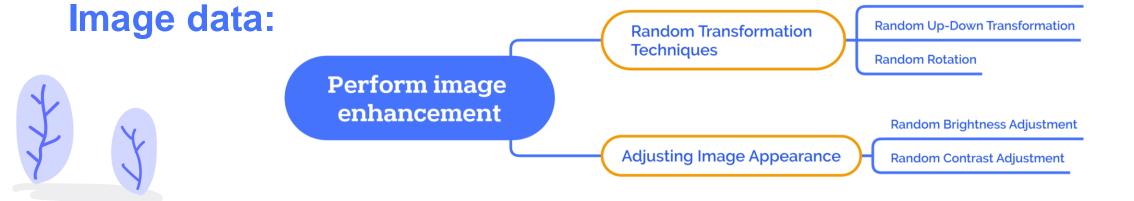
- **Feature engineering**

Text data:

1.Feature extraction, use all columns has the best effect.

best subset   <   delete empty columns   all columns

performance

2.Ordinal Encoding all columns and normalize using MinMaxScaler.

Image data:

Perform image enhancement

Random Transformation Techniques
- Random Left-Right Transformation
- Random Up-Down Transformation
- Random Rotation

Adjusting Image Appearance
- Random Brightness Adjustment
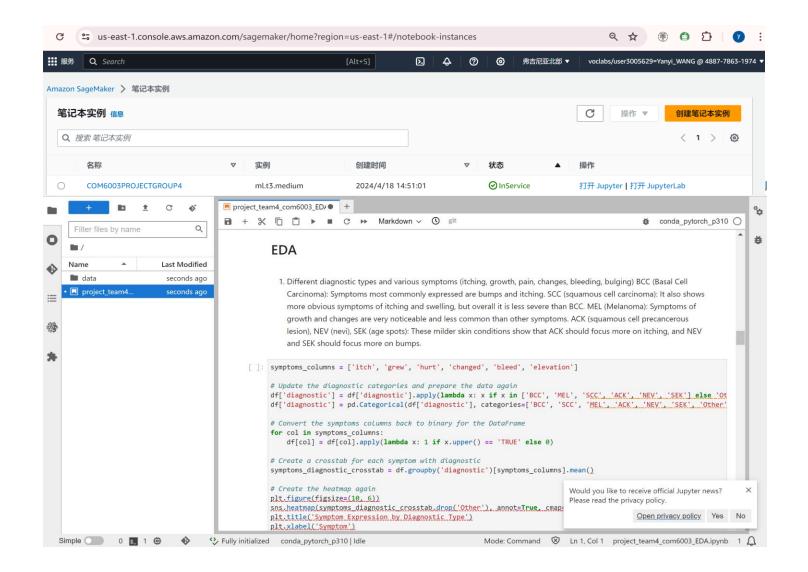- Random Contrast Adjustment

04

Exploratory data analysis

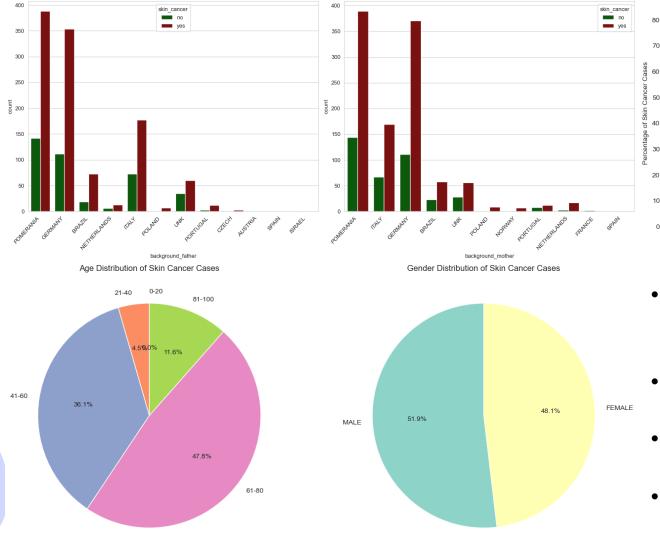# • EDA

EDA is completed on the AWS platform
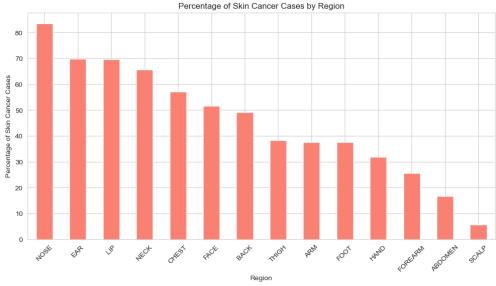
EDA

Father Background vs Skin Cancer

Mother Background vs Skin Cancer

Percentage of Skin Cancer Cases by Region

Age Distribution of Skin Cancer Cases

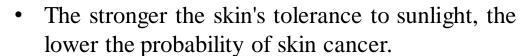Gender Distribution of Skin Cancer Cases

- People whose parents' background is GERMANY or POMERANIA have a higher risk of skin cancer.
- The nose is where the most skin cancers are diagnosed.
- Middle-aged and elderly people (41-80 years old) have a higher risk of skin cancer.
- Men have a higher risk of skin cancer than women.

- # EDA

Percentage Distribution of Diagnostics by Variable

Variables	BCC	MEL	SCC	Other
pesticide-False	55.0	3.1	12.6	29.3
pesticide-True	59.0	4.0	13.2	23.8
has_piped_water-False	59.0	4.9	10.2	25.9
has_piped_water-True	55.0	2.6	14.5	27.8
has_sewage_system-False	60.0	4.5	10.3	25.2
has_sewage_system-True	53.9	2.7	14.8	28.6
fitspatrick-True	66.0	2.0	10.5	21.6
fitspatrick-2.0	57.3	4.0	14.4	24.3
fitspatrick-3.0	55.4	2.6	11.5	30.6
fitspatrick-4.0	35.5	6.5	4.8	53.2
fitspatrick-5.0	30.0	0.0	20.0	50.0
fitspatrick-6.0	0.0	0.0	0.0	100.0
skin_cancer_history-False	54.0	4.6	12.7	28.8
skin_cancer_history-True	59.6	2.2	13.1	25.1
cancer_history-False	57.9	2.9	12.3	26.9
cancer_history-True	55.3	4.0	13.4	27.3

Diagnostic Categories

- The stronger the skin's tolerance to sunlight, the lower the probability of skin cancer.
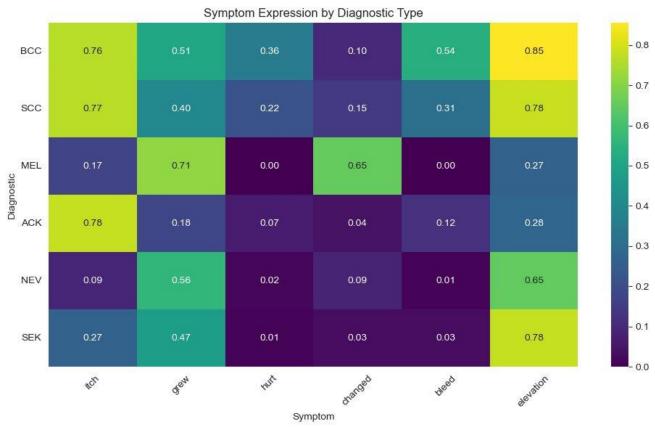- People who have been exposed to pesticides or other chemicals may have a higher probability of developing skin cancer.
- Patients who live in a location or area without access to running water or a proper sewage system may have a higher probability of developing skin cancer
- Patients with a history of skin cancer in their families may be more susceptible to skin cancer.
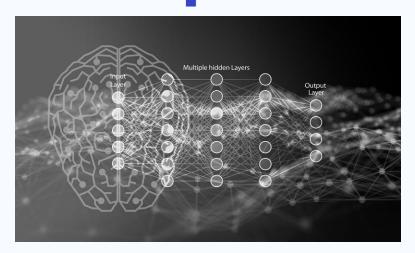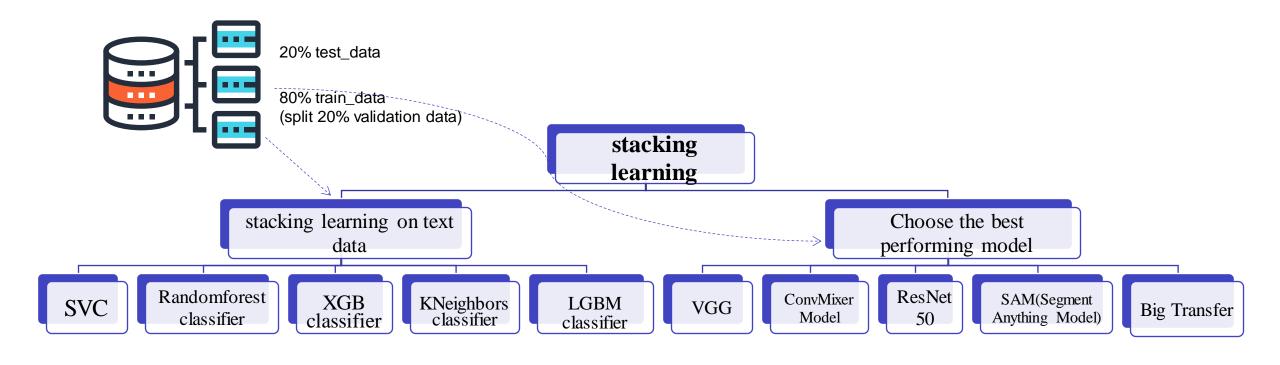
EDA

Symptom Expression by Diagnostic Type

- BCC (Basal Cell Carcinoma): Symptoms most commonly expressed are bumps and itching.
- SCC (squamous cell carcinoma): It also shows more obvious symptoms of itching and swelling, but overall it is less severe than BCC.
- MEL (Melanoma): Symptoms of growth and changes are very noticeable and less common than other symptoms.
- ACK (squamous cell precancerous lesion), NEV (nevi), SEK (age spots): These milder skin conditions show that ACK should focus more on itching, and NEV and SEK should focus more on bumps.
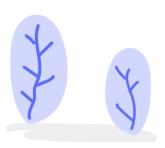
05

Model Training, Evaluation and Tunning

Model Training Architecture

20% test_data

80% train_data
(split 20% validation data)

stacking learning

stacking learning on text data

Choose the best performing model

SVC

Randomforest classifier

XGB classifier

KNeighbors classifier

LGBM classifier

VGG

ConvMixer Model

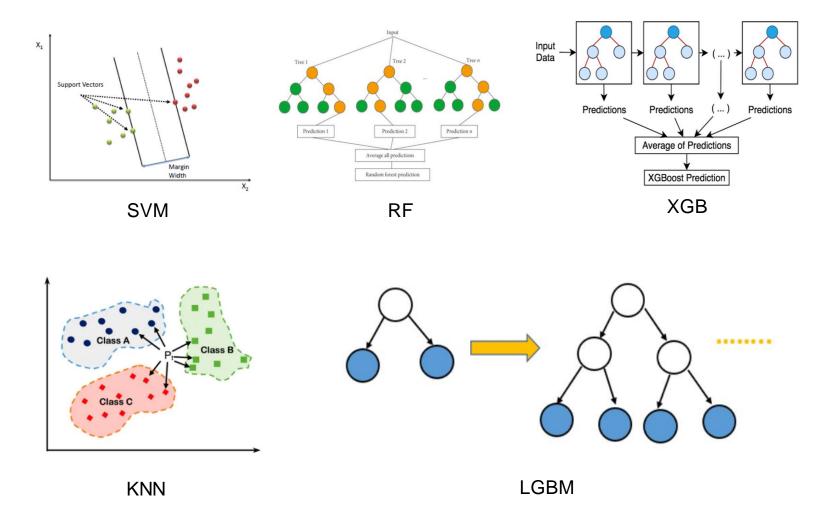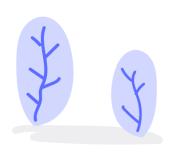ResNet 50

SAM(Segment Anything Model)

Big Transfer

Model Training Architecture – Text Analysis

- Dimension of data:

Only 1D, structured data type

- Explain:

Traditional machine learning excels in structured 1D text data, offering simplicity and interpretability.
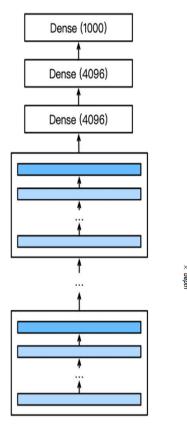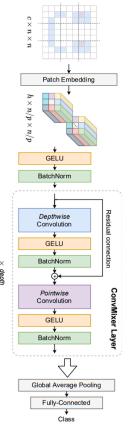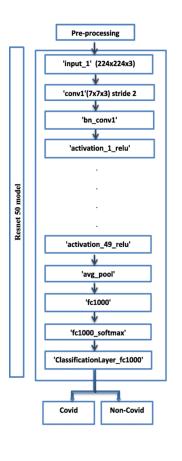

SVM

RF

XGB

KNN

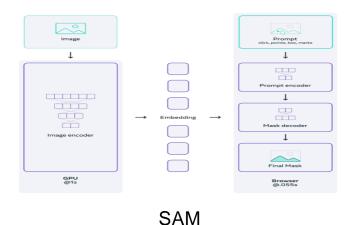
LGBM

Model Training Architecture – Image analysis

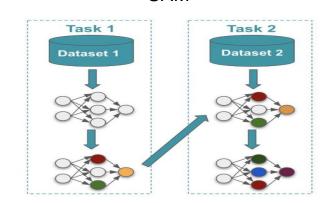- Dimension of data:

2D, unstructured data

- Explain:

Deep learning models capture complex patterns in 2D unstructured image data, excelling at hierarchical feature extraction.
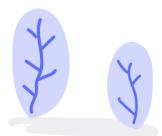

VGG
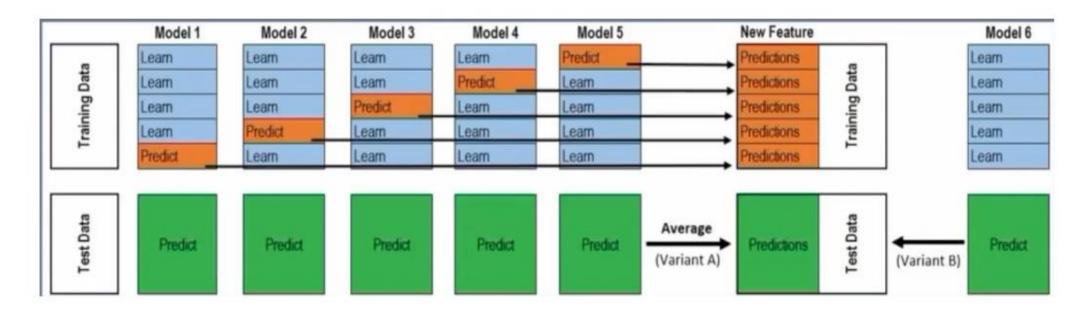
Convmixer

ResNet50

SAM

Big transfer
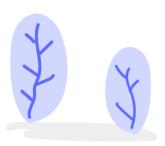
Model Training Architecture – Stacking of Image and Text model

- Stacking
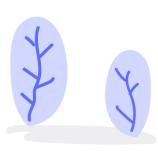
Preliminary Model Evaluation

- Text data

Model	Accuracy	Recall	Precision
SVM	0.40	0.38	0.54
RondamForest	0.43	0.43	0.59
XGBoost	0.42	0.42	0.62
KNN	0.40	0.39	0.52
LGBM	0.43	0.43	0.63
Stacking	0.41	0.40	0.59

- Image data
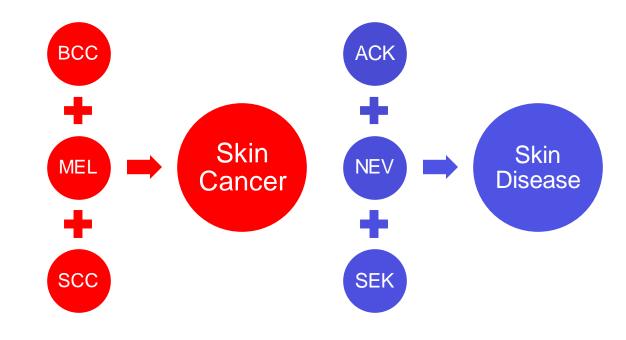
Model	Accuracy	Recall	Precision
Resnet 50	0.01	0.01	0.01
VGG	0.01	0.01	0.01
Big Transfer	0.11	0.10	0.10
ConvMixer	0.09	0.30	0.30
SAM	0.15	0.38	0.38
Stacking	0.31	0.19	0.19
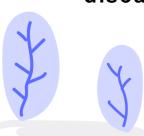
Model Tuning

- Based on the evaluation results, **the maximum accuracy of our preliminary model has only 61.6%,** however the acceptance performance for healthcare machine learning model **should has at least 80%. Our pre-screening model for skin cancer diagnosis is inadequate for public use**.

- To improve model accuracy, **we simplified the output classification from six categories to two: skin cancers (BCC, MEL, SCC) and skin diseases (ACK, NEV, SEK).**

BCC + MEL + SCC → Skin Cancer

ACK + NEV + SEK → Skin Disease
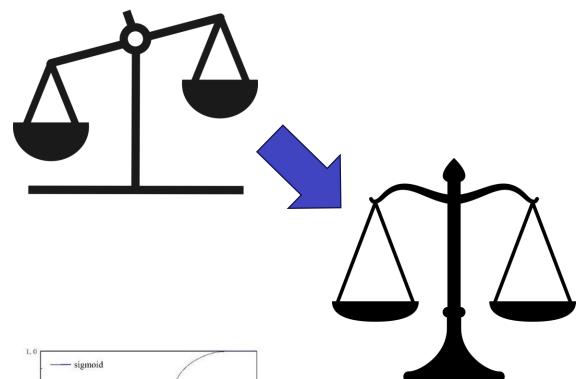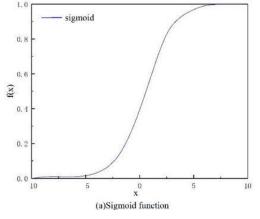
Model Tuning

- This change addressed data imbalances, **creating a more even distribution between skin cancer and skin disease categories**, which eliminated the need for oversampling **and allowed us to use the entire dataset for training,** enhancing the model's learning potential.
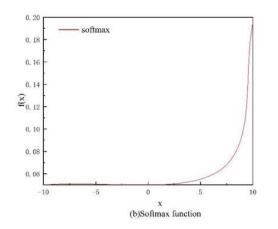
- In technical terms, while a SoftMax function is typically used for multi-class classification to calculate a dependent probability distribution across classes, **we found the sigmoid function more suitable for our revised two-category system as it treats each output independently.**

(a)Sigmoid function

$$\sigma(z) = \cdot \frac{1}{1 + \cdot e^{-z}}$$

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \cdot for \cdot i = 1, \dots, K \cdot and \cdot z = (z_1, \dots, z_K)$$

(b)Softmax function

Model performance comparison

The performance analysis of our skin cancer detection system indicates that simpler binary classifications substantially outperform more complex six-class categorizations.

Text Model Analysis:

- Light Gradient Boosting Machine (LGBM) excels with 0.93 accuracy in binary classifications compared to 0.43 in six classes.

- Random Forest (RF) and Extreme Gradient Boosting (XGB) both achieve 0.92 accuracy in binary classifications.

- Stacking models show a balanced performance with up to 0.91 accuracy for binary classifications.

Image Model Analysis:

- SAM model perform the best in 6 classification, having 0.38 of accuracy, 0.38 recall and 0.15 precision. For binary class classification, VGG model perform the best, having 0.72 of accuracy, 0.74 precision and 0.72 recall.

- The Stacking model, which integrates outputs from multiple architectures, also performs commendably with accuracies of 0.19, precision of 0.31 and recall of 0.19 for six classes; accuracies of 0.52, precision of 0.52 and recall of 0.52 for binary classes.

Model performance comparison

Text classification model strategy:

Approach: Utilized stacking model for meta-stacking.

Rationale: Minimal performance variation among models (top model only 10% more accurate than the least).

Benefit: Ensures consistent model performance by leveraging collective strengths.

Image classification model strategy:

Approach: Selected the best-performing model for meta-stacking.

Rationale: Significant performance disparity (top model 50% more accurate than the lowest).

Benefit: Maximizes accuracy by utilizing the strongest model, avoiding dilution from weaker models.

Meta Model strategy:

Six-Class Classification: Combined text stacking model with image SAM model.

Binary Classification: Paired text stacking model with image VGG model.

Goal: Tailored strategies optimize accuracy and reliability for both text and image data, addressing specific system challenges effectively.

Model performance comparison

Text Model:

Model	Precision		Recall		Accuracy	
No. of classes	Six	Two	Six	Two	Six	Two
SVM	0.54	0.87	0.38	0.87	0.40	0.87
RF	0.59	0.93	0.43	0.92	0.43	0.92
XGB	0.62	0.92	0.42	0.92	0.42	0.92
KNN	0.52	0.86	0.39	0.86	0.40	0.86
LGBM	0.63	0.93	0.43	0.93	0.43	0.93
Stacking	0.59	0.91	0.40	0.91	0.41	0.91

Image Model:

Model	Precision		Recall		Accuracy	
No. of classes	Six	Two	Six	Two	Six	Two
ConvMixer	0.09	0.22	0.30	0.47	0.30	0.47
Resnet 50	0.01	0.22	0.01	0.47	0.01	0.47
VGG	0.01	0.74	0.01	0.72	0.01	0.72
Big Transfer	0.11	0.74	0.10	0.58	0.10	0.58
SAM	0.15	0.55	0.38	0.54	0.38	0.54
Stacking	0.31	0.52	0.19	0.52	0.19	0.52

Stacked model performance for Image and Text:

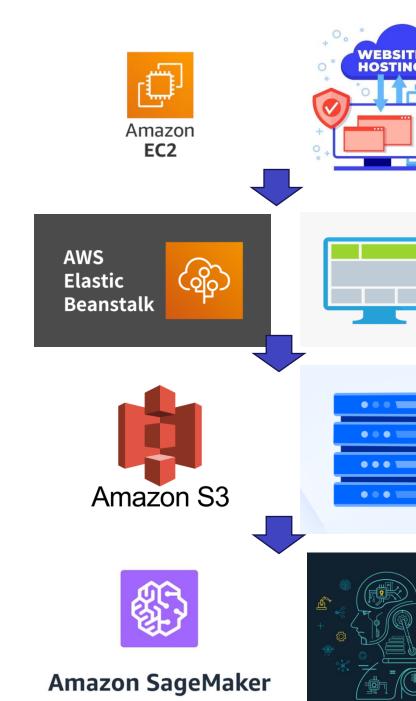| Model	Precision		Recall		Accuracy	
No. of classes	Six	Two	Six	Two	Six	Two
Stacking	0.37	0.83	0.39	0.82	0.39	0.82

06

Conclusion

Deployment

To deploy our model with a user interface, we plan to utilize Amazon Web Services (AWS) for a robust and scalable solution:

1. **Backend Hosting:** The application's backend, responsible for data intake and processing, will be hosted on **Amazon EC2 instances**

2. **User Interface:** The patient questionnaire and image upload interface will be integrated into a web service managed on **AWS Elastic Beanstalk**

3. **Data Storage:** Patient data from the questionnaire and uploaded images will be securely stored on **Amazon S3**

4. **Model Development and Training**: We will use **Amazon SageMaker** for developing, training, and deploying our text and image classification models.

Deployment

5. **Model Deployment:** Once trained, the models will be deployed as **SageMaker endpoints**, which the application will use to send user-input data for real-time analysis. The text and image data are processed independently by their respective models to generate preliminary outputs.

6. **Output Combination: AWS Lambda functions** will implement the stacking technique, triggered after both models have processed the data. These functions will combine outputs from the text and image models and compute the final binary classification.

7. **User Interaction and Results:** Users interact with the system by completing a digital questionnaire and uploading necessary images via the web service. The system then returns the result—'0' for no cancer or '1' for cancer—through the web interface.

This AWS-based architecture ensures that our skin cancer detection system is not only efficient and reliable but also user-friendly and accessible to a wide audience.
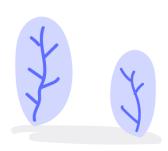
Conclusion

- Our project explores the feasibility of enabling the public to conduct preliminary skin cancer screenings independently, and preliminary results confirm its viability.

- The meta model of traditional machine learning model on patient textual information and deep learning model on patient skin image under binary classification output has a satisfactory performance, i.e. 82%.

- **In order to provide an accurate pre-screening assessment to general public, a combination of simple photographic assessments and questionnaires is required.**

- However, before developing this into a publicly available skin cancer pre-screening application, several critical factors must be addressed. These include assessing potential biases within our model and ensuring the credibility of the data used for model training.

References

1. H. A. Hong Kong Cancer Registry, "Leading Cancer Sites in Hong Kong in 2020," Hong Kong Cancer Registry, Hospital Authority, Hong Kong, 2020.
2. E. e. al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature (London), vol. vol. 542, no. no. 7639, p. pp. 115–118, 2017.
3. L. G. S. A. K. B. B. I. d. A. G. A. F. J. E. J. S. A. C. P. R. F. F. P. K. R. K. H. S. M. d. E. S. R. M. T. C. T. d. B. L. Pacheco AGC, "PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. Data Brief.," 2020 Aug.
4. A. C. Society, "What Is Basal and Squamous Cell Skin Cancer?," [Online]. Available: https://www.cancer.org/cancer/types/basal-and-squamous-cell-skin-cancer/about/what-is-basal-and-squamous-cell.html.. [Accessed 15 Apr 2024].
5. S. H. a. F. K. F. Thabtah, "Data imbalance in classification: Experimental evaluation," Information Sciences, vol. 513, pp. 429-441, 2020.
6. M. e. a. Mohammed, "A stacking ensemble deep learning approach to cancer type classification based on TCGA data," Scientific Reports, vol. 11, no. 1, p. 15626, 2021.
7. B. Ummepure, "Classification of Skin Cancer Using ResNet and VGG Deep Learning Network," in Proceedings of the International Journal of Computer Applications, vol. 184, no. 42, 2023.
8. D. H. a. Z. Jie, "Research on Skin Cancer Classification Method Based on Improved ResNet50," Computer Technology and Development, vol. 2023, no. 02.
9. A. T. a. J. Z. Kolter, "Patches Are All You Need?," arXiv:2201.09792. .
10. Y. L. a. X. Y. M. Hu, "SkinSAM: Empowering Skin Cancer Segmentation with Segment Anything Model," arXiv:2304.13973.
11. A. K. e. al., "Big Transfer (BiT): General Visual Representation Learning," ar5iv.org, 2023.
12. Q. M. T. F. T. W. W. C. W. M. Q. Y. a. T.-Y. L. G. Ke, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," Advances in neural information processing systems, no. 30, 2017.
13. H. a. S. Gohel, "Machine Learning in Healthcare," Current genomics, vol. 22, no. 4, p. 291–300, 2021.

THANKS

- GROUP 4