Machine Learning on ISIC-2024-challenge dataset in Chaos and Clues Algorithm

Work Progress

20241028	**Receive Isic-2024-challenge dataset from Sakib**
20241102	**Start AI analysis by using Non-Deep-Learning Machine Learning learned from Year 2 study**
20241111	**Present Result to Sakib and Nurjahan, start writing the report**
20241118	**Present Report to Sakib and Nurjahan**
20241202	**Producing third result using XGB, LGBM and CatBoost Classifier**

Consideration of AI approach

As the dataset consists of diagnostically labelled images with many additional metadata which are derived from diagnostically labelled images, the metadata are to be studied in Non-Deep-Learning Machine Learning.

Data preparation

In ISIC-2024-challenge dataset, there are 400,666 samples of 0: benign cases, 393 samples of 1 : malignant cases. After discovering the severe bias of the data, Dr Adrian K. Davison was consulted. The 393 samples are suggested to be randomly extracted from the 400,666 samples of 0: benign cases, and are added to the 393 samples of 1 :malignant cases for Non-Deep-Learning Machine Learning.

The following features (columns) are deleted from the dataset before the Non-Deep-Learning Machine Learning algorithms are executed, as these are just sample numbers, or less relevant or not relevant features:

isic_id

patient_id

image_type

tbp_tile_type

tbp_lv_location

attribution

copyright_license

lesion_id

iddx_full

iddx_1

iddx_2

iddx_3

iddx_4

iddx_5

mel_mitotic_index

mel_thick_mm

Sex feature is set:

male as 1

female as 0

anatom_site_generalis set

head/neck as 1

anterior torso as 2

posterior torso as 3

upper extremity as 4

lower extremity as 5

tbp_lv_location_simple is set

Head & Neck as 1

Torso Front as2

Torso Back as	3
Left Arm as	4
Right Arm as	5
Left Leg as	6
Right Leg as	7
Unknown as	8

The revised dataset is named “train-metadata V2.csv”.

Methods of Machine Learning used:

1 K-nearest neighbours (K-NN) classifier

2. Naïve Bayes (NB) classifier

3 Decision Tree (DT) Classifier

4. Bagging Classifier

5 Random Forest (RF) Classifier

The Python Code is named “Synosis Project 1 V2.ipynb”.

First Result

Methodology	Accuracy
1 K-nearest neighbours (K-NN) classifier	64.4 % (K = 6)
2 Naive Bayes (NB) classifier	81.9 %
3 Decision Tree (DT) Classifier	86.2 %
4 Bagging Classifier	90.4 %
5 Random Forest (RF) Classifier	90.4 %

Features with Feature Importance more than 0.03 in the First, Second and Third Decision Tree in Bagging Classifier and in the First, Second and Third Decision Tree in Random Forest Classifier :

tbp_lv_H

clin_size_long_diam_mm

tbp_lv_norm_color

tbp_lv_perimeterMM

tbp_lv_y

tbp_lv_minorAxisMM

tbp_lv_dnn_lesion_confidence

tbp_lv_radial_color_std_max

tbp_lv_deltaLBnorm

tbp_lv_location_simple

tbp_lv_deltaA

tbp_lv_deltaB

tbp_lv_B

anatom_site_general

tbp_lv_deltaL

tbp_lv_Hext

tbp_lv_nevi_confidence

tbp_lv_deltaLB

tbp_lv_symm_2axis

tbp_lv_z

tbp_lv_stdL

tbp_lv_areaMM2

tbp_lv_color_std_mean

tbp_lv_Bext

tbp_lv_stdLExt

tbp_lv_L

Final Result

Different maximum depth of the Decision Tree Classifier, Bagging Classifier and Random Forest (RF) Classifier and different datasets including less features are tried to see if such combination can maintain good or better results.

After several trials, the maximum depth of the Decision Tree (DT) Classifier, Bagging Classifier and Random Forest (RF) Classifier are set as 9 and in the first result, features with Feature Importance more than 0.03 in the First, Second and Third Decision Tree in Bagging Classiffier and in the First, Second and third Decision Tree in Random Forest(RF) Classifier

tbp_lv_H
clin_size_long_diam_mm
tbp_lv_norm_color
tbp_lv_perimeterMM
tbp_lv_y (not used and deleted from the dataset as it is the Y-coordinate of the lesion on 3D TBP)
tbp_lv_minorAxisMM
tbp_lv_dnn_lesion_confidence
tbp_lv_radial_color_std_max
tbp_lv_deltaLBnorm
tbp_lv_location_simple (not used and deleted from the dataset as it is the classification of anatomical location, simple)
tbp_lv_deltaA
tbp_lv_deltaB
tbp_lv_B
anatom_site_general (not used and deleted from the dataset as it is the location of the lesion on the patient's body)
tbp_lv_deltaL
tbp_lv_Hext
tbp_lv_nevi_confidence
tbp_lv_deltaLB
tbp_lv_symm_2axis
tbp_lv_z (not used and deleted from the dataset as it is the Z-coordinate of the lesion on 3D TBP.)
tbp_lv_stdL
tbp_lv_areaMM2
tbp_lv_color_std_mean
tbp_lv_Bext
tbp_lv_stdLExt
tbp_lv_L

are used to maintain a good or better result. The revised dataset is called "train-metadata V4.csv".

The Python Code is named "Synosis Project 1 V5.ipynb".

Methodology	Accuracy
1 K-nearest neighbours (K-NN) classifier	86.7 % (K = 6)
2 Naive Bayes (NB) classifier	80.1 %
3 Decision Tree (DT) Classifier	83.7 %
4 Bagging Classifier	89.3 %
5 Random Forest (RF) Classifier	90.8 %

Conclusion

Random Forest (RF) Classifier gives Accuracy of 90.8%. Many Features used are similar to the decision criteria as stated in Chaos and Clues algorithm and Decision algorithm for non-pigmented skin malignancy. For example:

tbp_lv_color_std_mean: Colour irregularity, calculated as the variance of colours within the lesion's boundary.

Tbp_lv_norm_color: Colour variation (0-10 scale); the normalized average of colour asymmetry and colour irregularity.

As the sample size is too small (just 786), the result would not be well representative. The dataset finally used can later be executed in Deep Learning Algorithm to find a perhaps better result.

Important future suggestion

1 As metadata seems to help in good results on AI analysis, it is worth to invite several Dermatology doctors to examine all the photos of the curated balanced dataset in "Cassidy, B., Kendrick, C., Brodzicki, A., Jaworek-Korjakowska, J. and Yap, M.H., 2022. Analysis of the ISIC image datasets: usage, benchmarks and recommendations. Medical image analysis, 75, p.102305." to find out features such as Chaos, Clues, Black Dots, Clods, Ulceration, White Clues, Monomorphous or Polymorphous vessels… etc. (i.e.all the decision criteria as stated in Chaos and Clues algorithm and Decision algorithm for non-pigmented skin malignancy) and also to predict more precisely what type of cancer and non-cancer it is.

By examining each of the photos of the curated balanced dataset to find out features in Chaos and Clues algorithm and Decision algorithm for non-pigmented skin malignancy, it should improve the AI analysis completely. It may also have some insights on the change of Chaos and Clues algorithm and Decision algorithm for non-pigmented skin malignancy for doctors to do clinical examination better.

By predicting more precisely what type of cancer and non-cancer it is, it should also improve the AI analysis completely.

2 Some features used in the final result have a scale representation. The scale presentation may help in the AI analysis of the curated balanced dataset.

3 As some features in ISIC-2024_challenge dataset have in more or less similar type, some features of similar types should be eliminated to reduce correlation error.

Third Result

Methodology	Accuracy
1 XGB Classifier	85.8%
2 LGBM Classifier	86.3%
3 CatBoost Classifier	86.8%

Methodology	Accuracy
1 K-nearest neighbours (K-NN) classifier	86.7 % (K = 6)
2 Naive Bayes (NB) classifier	80.1 %
3 Decision Tree (DT) Classifier	83.7 %
4 Bagging Classifier	89.3 %
5 Random Forest (RF) Classifier	90.8 %

LGBM Classifier has 16 features with non-zero Feature Importance

Feature	Importance

tbp_lv_norm_color	0.258648
tbp_lv_H	0.228952
clin_size_long_diam_mm	0.178502
tbp_lv_minorAxisMM	0.067067
tbp_lv_dnn_lesion_confidence	0.058715
tbp_lv_areaMM2	0.05441
tbp_lv_nevi_confidence	0.039744
tbp_lv_stdLExt	0.026878
tbp_lv_deltaLBnorm	0.022708
tbp_lv_Bext	0.021274
tbp_lv_symm_2axis	0.014718
tbp_lv_L	0.009396
tbp_lv_deltaL	0.006898
tbp_lv_stdL	0.005675
tbp_lv_B	0.004694
tbp_lv_radial_color_std_max	0.001721
tbp_lv_Hext	0
tbp_lv_perimeterMM	0
tbp_lv_deltaB	0
tbp_lv_deltaA	0
tbp_lv_color_std_mean	0
tbp_lv_deltaLB	0

LGBM Classifier has only 9 features with non-zero Feature Importance

Feature	Importance
clin_size_long_diam_mm	3

tbp_lv_H	3
tbp_lv_L	3
tbp_lv_areaMM2	2
tbp_lv_deltaLBnorm	1
tbp_lv_stdLExt	1
tbp_lv_norm_color	1
tbp_lv_minorAxisMM	1
tbp_lv_dnn_lesion_confidence	1
tbp_lv_deltaB	0
tbp_lv_deltaL	0
tbp_lv_B	0
tbp_lv_deltaA	0
tbp_lv_color_std_mean	0
tbp_lv_nevi_confidence	0
tbp_lv_Hext	0
tbp_lv_perimeterMM	0
tbp_lv_radial_color_std_max	0
tbp_lv_stdL	0
tbp_lv_Bext	0
tbp_lv_symm_2axis	0
tbp_lv_deltaLB	0

CatBoost Classifier has only 6 features with non-zero Feature Importance

Feature	Importance
tbp_lv_areaMM2	37.77552
tbp_lv_H	25.75597

clin_size_long_diam_mm	18.19444
tbp_lv_perimeterMM	7.320137
tbp_lv_dnn_lesion_confidence	6.481997
tbp_lv_deltaB	4.471939
tbp_lv_L	0
tbp_lv_nevi_confidence	0
tbp_lv_symm_2axis	0
tbp_lv_stdLExt	0
tbp_lv_stdL	0
tbp_lv_radial_color_std_max	0
tbp_lv_Bext	0
tbp_lv_norm_color	0
tbp_lv_minorAxisMM	0
tbp_lv_Hext	0
tbp_lv_deltaLBnorm	0
tbp_lv_B	0
tbp_lv_deltaL	0
tbp_lv_deltaA	0
tbp_lv_color_std_mean	0
tbp_lv_deltaLB	0

Dataset Citation

International Skin Imaging Collaboration. SLICE-3D 2024 Challenge Dataset. International Skin Imaging Collaboration https://doi.org/10.34970/2024-slice-3d (2024). Creative Commons Attribution-Non Commercial 4.0 International License. The dataset was generated by the International Skin Imaging Collaboration (ISIC) and images are from the following sources: Hospital Clínic de Barcelona, Memorial Sloan Kettering Cancer Center, Hospital of Basel, FNQH Cairns, The University of Queensland, Melanoma Institute Australia, Monash University and Alfred Health, University of Athens Medical School, and Medical University of Vienna.