

****

**🙟🕮🙝**

Ministry of Education and Training

**Saigon International University**

**SPECIALTY: ENGINEERING AND COMPUTER SCIENCE**

Ảnh có chứa biểu tượng, phù hiệu, Đồ họa, Nhãn hiệu

Mô tả được tạo tự động

**FINAL ESSAY**

**TOPIC: APPLICATION OF MACHINE LEARNING ALGORITHMS IN MULTICLASS GARBAGE CLASSIFICATION PROBLEM**

**Instructor:**

**Member:**

Tran Quang Huy

Ho Truong Hai Trieu

Nguyen Truong Duy

Do Quoc Dat

Table of Contents

[**1. Introduction:** 1](#_Toc151287106)

[**1.1. Current Situation:** 1](#_Toc151287107)

[**1.2** **Objectives:** 2](#_Toc151287108)

[**2.** **Methodology:** 2](#_Toc151287109)

[**2.1** **Hypothesis:** 2](#_Toc151287110)

[**2.2** **Data Collection:** 2](#_Toc151287111)

[**2.3** **Data Analysis:** 3](#_Toc151287112)

[**2.4** **Data Preprocessing:** 3](#_Toc151287113)

[**2.5** **Augumentation:** 3](#_Toc151287114)

[**2.6** **Machine Learning algorithm:** 4](#_Toc151287115)

[**2.7** **Model Evaluation:** 4](#_Toc151287116)

[**3.** **Experiment:** 4](#_Toc151287117)

[**3.1** **Data processing:** 4](#_Toc151287118)

[**3.2** **Model Training:** 6](#_Toc151287119)

[**3.3 Evaluation:** 7](#_Toc151287120)

[**4.** **Overview:** 8](#_Toc151287121)

[**4.1** **Overview:** 8](#_Toc151287122)

[**4.2. Limitations** 8](#_Toc151287123)

[**5.** **Conclusion:** 9](#_Toc151287124)

[**6.** **References:** 9](#_Toc151287125)

**1. Introduction:**

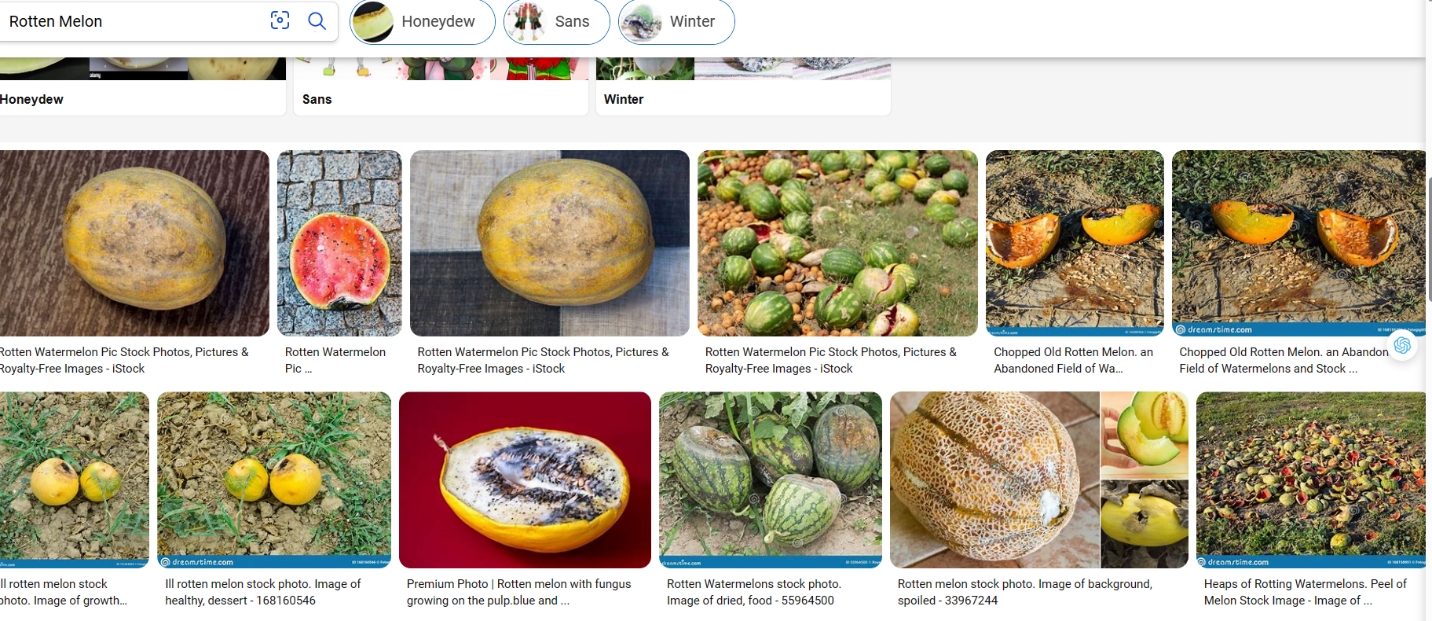
**1.1. Current Situation:**

Currently, the environmental protection issue through solid waste classification at the source is regulated by the Environmental Protection Law 2020 [(1)](#Figure1). This law addresses the classification, storage, transfer, collection and transportation costs, pollution treatment, etc. However, the problem of solid waste classification remains unresolved due to various issues. For instance, although citizens may separate their waste, sanitation workers often collect it together, or the cramped conditions of townhouses lead residents to store bottles and cans indoors, causing unpleasant odors [(2)](#Figure2).

* Additionally, there is an existing issue related to the awareness of the citizens. Most source separation programs are encouraging and experimental, lacking strict and compulsory measures, as seen in China and other advanced countries.
* Applications of automated waste classification primarily operate after waste has been collected by workers and brought to the processing facility. There is currently no application for automated waste classification at the source, such as in parks, beaches, or indoors.
  1. **Objectives:**
* To develop and create a machine learning model capable of waste classification integrated into robots and cameras, supporting improved waste sorting, enhancing environmental protection, reducing dependence on human cognitive factors, and minimizing the time spent on household chores.
* This article is divided it into four main parts as follows: Part 1 (introduction) outlines the current issues of waste classification, the purpose of the project; Part 2 (methodology) presents a general overview of the methods we used in this project, including hypotheses, data collection, data analysis, data preprocessing, data augmentation, machine learning algorithms, and metrics for model evaluation; Part 3 (experiment) discloses the processed data (including data augmentation and data preprocessing), model parameters, and metrics derived from the results of the machine learning model; Part 4 reviews the results, identifying limitations, and discussing the prospects of the project in the future.

1. **Methodology:**
   1. **Hypothesis:**

In this project, we assume that we have collected images of waste objects after extraction from the waste detection problem. Subsequently, we will use these images to classify waste using machine learning models for a 3-class classification problem, including Recyclable Waste, Non-Recyclable Waste, and Compostable Waste.

* 1. **Data Collection:**
* In this study, we primarily collected data from Google Images [(Figure 1)](#Figure1) , Kaggle, and a small portion from real-world environments such as homes, schools, and parks. For each type, we gathered three categories, including compostable waste, non-recyclable waste, and recyclable waste, with each category containing a quantity of images ranging from 400 to 600.
* Additionally, for the Recyclable and Non-Recyclable waste types, we obtained data from Kaggle and GitHub. Subsequently, we manually cropped the images and stored them in the dataset.

[Figure 1](#Figure1)

* 1. **Data Analysis:**

In the data analysis section, we utilized basic statistical measures such as standard deviation, mean, maximum value, and minimum value to statistically describe the length and width dimensions of the images in the dataset.

* 1. **Data Preprocessing:**

Data preprocessing is a crucial step in training any machine learning model. In this project, we performed two main steps for image preprocessing: resizing the images and scaling the entire images to the range [0,1].

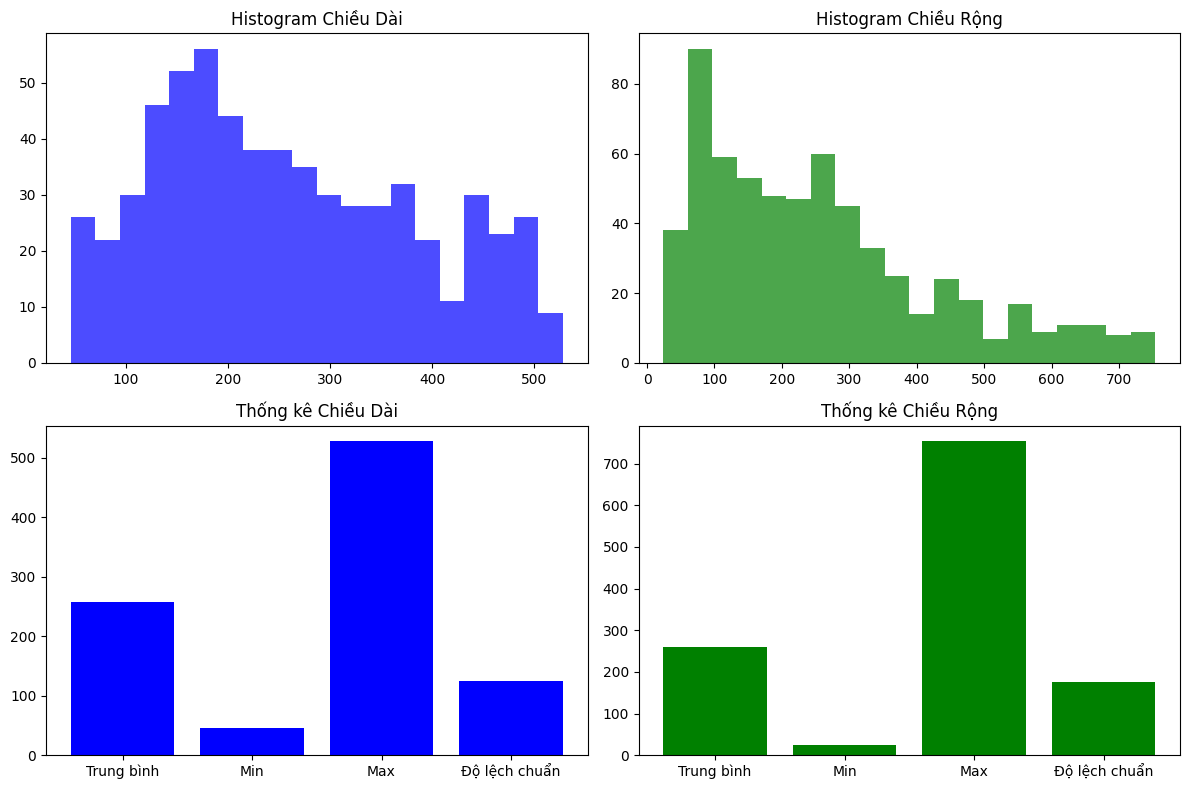
* 1. **Augmentation:**

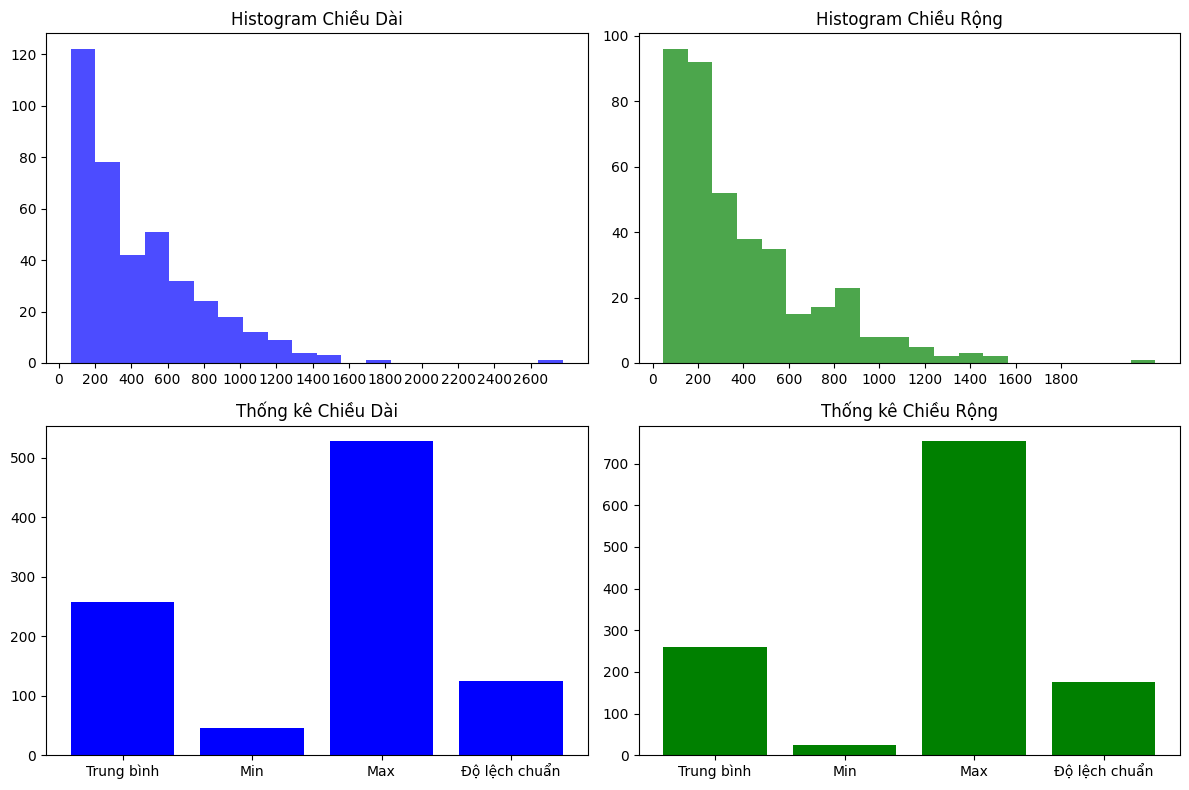
Due to the need for a large amount of data to generalize various classification scenarios in real-life waste sorting problems, applying data augmentation techniques such as image rotation, translation, and adjusting brightness is essential. Data augmentation serves purposes such as increasing data diversity, preventing overfitting, and enhancing model performance. In this project, for the sake of convenience and ease of use, we used the augmentation techniques available in the Albumentations library.

* 1. **Machine Learning algorithm:**
* In this project, we employed supervised machine learning algorithms, including K-Nearest Neighbor, Decision Tree, and Ensemble Learning method with the Random Forest algorithm to solve the multi-class waste classification problem.
* **The KNN model**, relying on the similarity between data points, can provide accurate predictions in cases where the nearest neighbors have similar characteristics.
* **Decision Tree**, with classification based on decisions at nodes, provides insights into the model's decision-making but carries the risk of overfitting if not controlled.
* **Random Forest**, combining the strength of multiple decision trees, often exhibits good generalization and reduces the risk of overfitting.
  1. **Model Evaluation:**

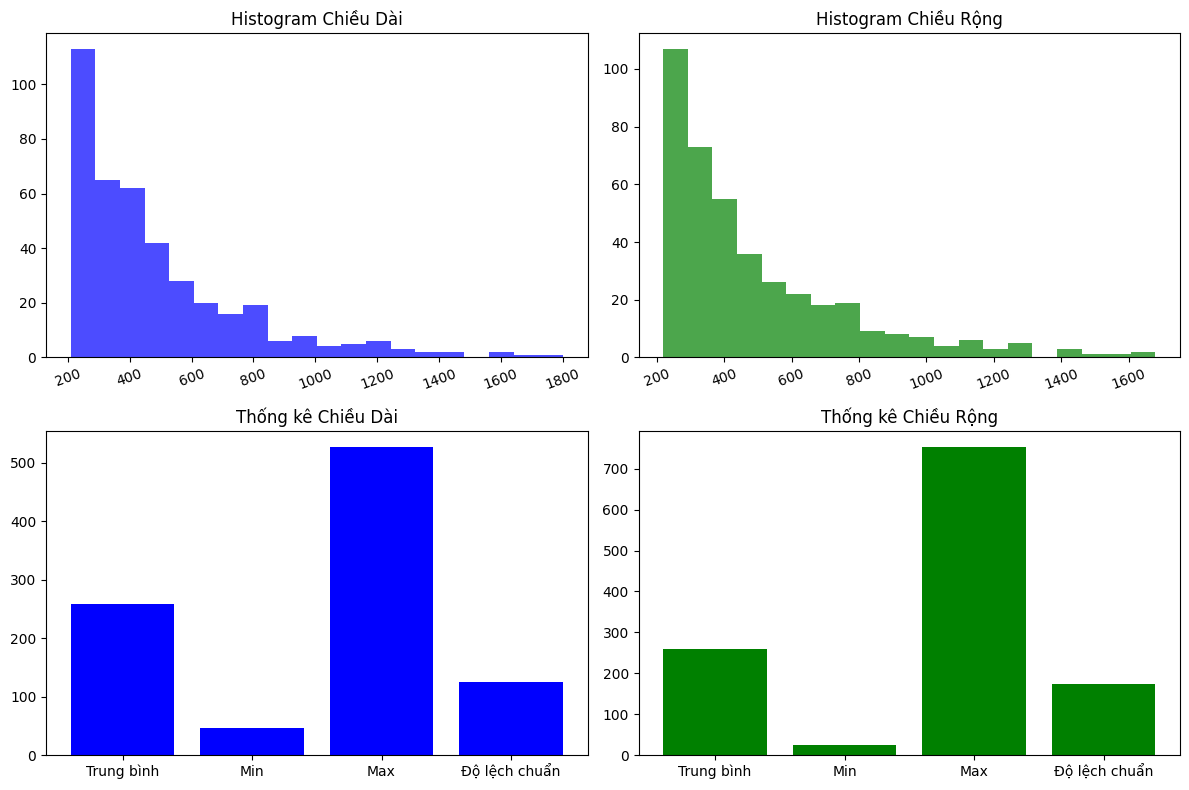
To diversify the evaluation criteria for the model's feasibility in applying machine learning models in real-world scenarios, we proposed four metrics for the problem, including Precision, Recall, F1-Score, and Accuracy Score.

1. **Experiment:**
   1. **Data processing:**

During data collection, we observed that the majority of image dimensions fell within the ranges of [120, 400] and [80, 280] for compost data, [100, 300], and [100, 300] for Recycle data, [200, 400] and [200, 400] for Non-Recycle data (Figure [2](#Figure2) , [3](#Figure3), [4](#Figure4)). Therefore, we decided to resize the images to 256x256 dimensions and normalized the images by dividing each pixel value by 255 to scale the images to the range [0,1]. This aimed to accelerate computational processing speed and enhance the accuracy of machine [](#Figure2)learning models.

[](#Figure3)[Figure 2: Compost Parameters](#Figure2)

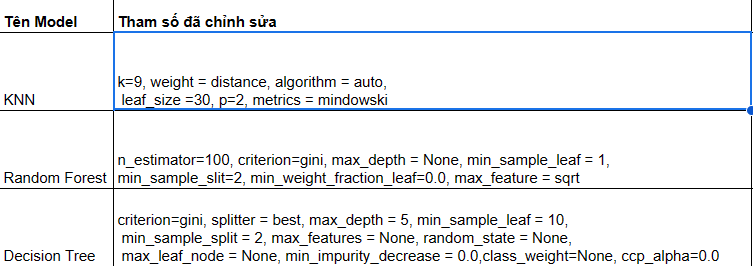
[Figure 3: Recycle Parameters](#Figure3)

[*Figure 4: Non-Recycle Parameters*](#Figure4)

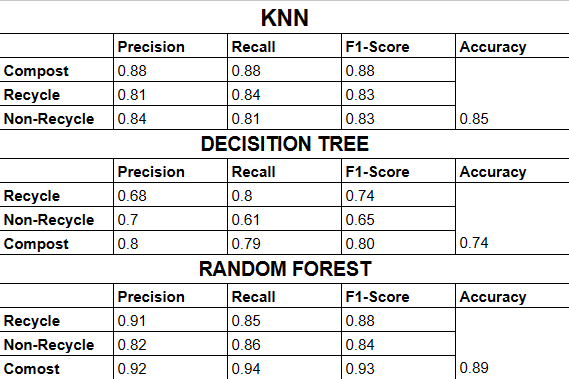
* When using deep learning networks such as Neural Networks, after training with approximately 1600 images, we noticed that the model did not achieve exceptionally good quality. Therefore, we applied data augmentation techniques, including image rotation at random angles between -15 to 15 degrees, to enhance the dataset, to enhance the dataset.
  1. **Model Training:**

In this project, we utilized three pre-built machine learning models from the Scikit-learn library, including K-Nearest-Neighbor, Decision Tree, and Random Forest:

* For KNN: The parameter n of neighbors was set to 9, weight was set to distance, and the search algorithm was set to automatically configure by scikit-learn. The leaf size was set to 30, and the algorithm used for distance measurement was Euclidean distance (L2) (Figure 5).
* For Random Forest: The number of trees was set to 100, and the remaining parameters were set to their default values (Figure 6) according to the documentation of the Scikit-learn library.
* For Decision Tree: The impurity measure was set to Gini, the maximum depth of the tree was set to 5, the minimum number of samples at each leaf was set to 10, and the other parameters were set to their default values (Figure 6) according to the documentation of the Scikit-learn library.

[Figure 5: Parameters](#Figure5)

## **3.3 Evaluation:**

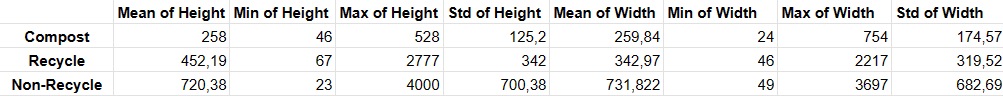
* [](#Figure6)The performance of the models on the three algorithms, KNN, Decision Tree, and Random Forest, was assessed using metrics such as Precision, Recall, F1-Score, and Accuracy, and was relatively good on the self-created dataset (Figure 6). Despite having only about 1600 images, the machine learning algorithms demonstrated effective waste classification capabilities, proving that traditional machine learning algorithms are applicable to real-world waste classification problems. Additionally, given the compact nature of these models, integration into robots is entirely feasible.

[Figure 6: Statistical Indicators of the Machine Learning Model](#Figure6)

1. **Overview:**
   1. **Overview:**

Applying machine learning models for waste classification with limited data is a challenge but also an opportunity to create effective solutions in waste management. To make this process feasible, we delved into the specific issues of waste classification, creating a context for collecting reasonable data. Concurrently, model selection, using augmentation and transfer learning techniques, along with controlling overfitting, helped expand the training dataset and improve generalization capabilities. Evaluating performance through metrics such as precision, recall, and F1-score is crucial to ensure the model operates effectively on limited data and among different types of waste. Despite challenges, with a flexible strategy and appropriate data processing and model training techniques, applying machine learning to the waste classification problem with limited data is entirely feasible and can yield positive results for environmental management.

## **4.2. Limitations**

[](#Figure7)However, there are several limitations to the current framework. Firstly, due to the low diversity of the data, attributed to the short data collection time and limited data collection experience, images were primarily sourced from Google Images. Consequently, many types of waste may not be represented in the dataset, leading to a lack of generalizability when applying the model in real-world scenarios. Secondly, as this is a sub-problem of Object Detection, it assumes that Object Detection performs well on the waste detection problem. Thirdly, because the images are cropped from a much larger image multiple times the size of the cropped image [(Figure 8)](#Figure8), the extracted image's size is very small, as evident in the large standard deviation of length and width in all three classes [(Figure 7)](#Figure7). This inconsistency is attributed to variations in data collection among team members.

[Figure 7: Statistical Indicators on the Dataset](#Figure7)

[](#Figure8) [](#Figure8)

[Figure 8: Illustration of a Small-sized Image Compared to a Large-sized Image](#Figure8)

1. **Conclusion:**

In summary, this project has provided a comprehensive insight into basic image processing techniques such as image resizing, normalization, and image data augmentation, as well as the use of machine learning models for multi-class classification problems. Additionally, a clear understanding of the context of waste classification issues was established to achieve optimal performance when applied in real-world scenarios. This project could mark a significant step in finding intelligent solutions for waste management and environmental protection.

1. **References:**

(1)[Quy định về phân loại rác tại nguồn - Tin liên quan - Cổng thông tin Bộ Y tế (moh.gov.vn)](https://moh.gov.vn/tin-lien-quan/-/asset_publisher/vjYyM7O9aWnX/content/quy-inh-ve-phan-loai-rac-tai-nguon?inheritRedirect=false)

(2) [Hơn 20 năm loay hoay phân loại rác - VnExpress](https://vnexpress.net/hon-20-nam-loay-hoay-phan-loai-rac-4673845.html)

(3) [Phân loại rác tại nguồn, làm thế nào cho đúng? | VTV24 - YouTube](https://www.youtube.com/watch?v=Mt8VwTLFUlg)

(4) [Tiêu điểm: Bất cập phân loại rác | VTV24 - YouTube](https://www.youtube.com/watch?v=7m4IMhiRfJc)

(5) [Albumentations Documentation](https://albumentations.ai/docs/)

(6) [scikit-learn Tutorials — scikit-learn 1.3.2 documentation](https://scikit-learn.org/stable/tutorial/index.html)