OCR

OCRとは、画像内の文字を検出し、テキストデータとして抽出する技術。

本研究では電子カルテをOCRによってテキストデータ化して、そのデータの二値分類を行いたい。

次ページのコードで実際にOCRを行ってみてください

```
import cv2
import pytesseract
from PIL import Image
import matplotlib.pyplot as plt

# サンプル画像の読み込み
image_path = "sample_image.png"
image = cv2.imread(image_path)

# OpenCVでグレースケール化
gray = cv2.cvtColor(image, cv2.COLOR_BGR2GRAY)

# ノイズ除去
gray = cv2.medianBlur(gray, 3)

# 画像表示
plt.imshow(gray, cmap="gray")
plt.title("Preprocessed Image")
plt.axis("off")
plt.show()

# OCR処理
text = pytesseract.image_to_string(gray, lang="eng")
print("抽出テキスト:")
print(text)
```