第3章 YOLOv8で読み込める形にして類似画像がないか調べる

2万枚の画像を整理…データセット作成&評価

佐藤 聖

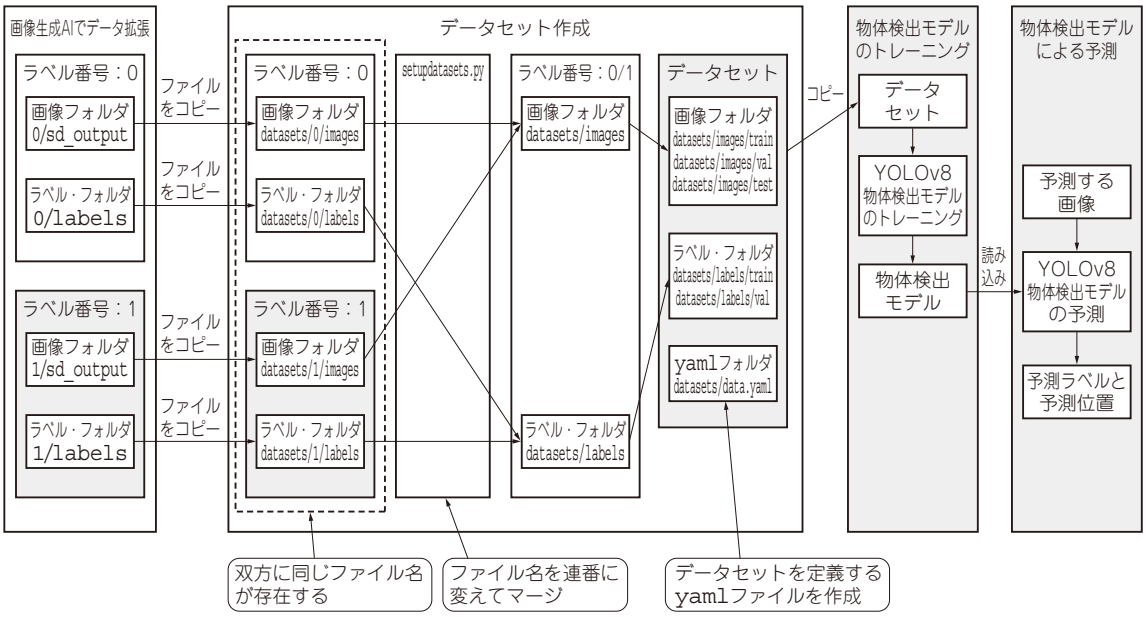

図1 データセット作成から物体検出モデルによる予測までの流れ

● データセットの整理が必要な理由

前章で2種類のボルトの画像をOKとNGのクラスに分類し，ラベルには0と1を設定しました．各クラスで1万枚の画像を用意しましたので合計で2万枚の画像があります．本稿ではこれらをデータセットにまとめる作業を行います．**図1**にデータセット作成から物体検出モデルでの検出までの流れを示します．

一般的には，後半の物体検出モデルのトレーニングや予測が難しいと思われるかもしれません．しかし，アルゴリズムやプログラムは，既に広く公開された実績のあるYOLOv8を使用しますので，失敗することは稀だと思います．

失敗する可能性のある部分は，意外にもデータセットを作成する段階です．AIモデルのトレーニングや予測の段階で気づくことになります．データセットの作成は，画像を集めて，フォルダにまとめるだけと思われがちですが，その分け方に偏りがあるとバイアスとなります．バイアスも含めて，AIモデルは学習してしまうので，偏った予測対象の画像が変わると精度に大きなばらつきが出るなどは典型例です．

● データセット作りの流れ

データセット作成のプロセスを段階に分けて説明します．ここでの目的は，最終的に物体検出モデルのトレーニングに使用する各クラスの画像ファイルとラベル・ファイルを構造化したフォルダにまとめることです．そのときに，同一フォルダの中で，同一ファイル名があると上書きされてしまいますので，ファイルをフォルダに移動する前に，ファイル名に連番を付けてユニークなファイル名に変えます(**図2**)．

● データセット用の実行環境へ移行

仮想環境env-aiのuserフォルダ内にdatasetsフォルダを作成し，その中に前章で使用した仮想環境