グラム染色画像からの酵母様真菌の分類と検出

田中 空翔[1,a)]　平田 耕一[2,b)]

概要：感染症の初期診断に利用されるグラム染色は，検体材料に存在する菌を染色液に染め上げ，その染色性と形状から顕微鏡検査により菌種を推定する手法である．本論文では，酵母様真菌とクロストリジウム・パーフリンゲンスとコリネバクテリウムという 2 種類のグラム陽性桿菌に着目する．そして，酵母様真菌と各菌のアノテーションを基に，画像分類器である VGG，MobileNet，DenseNet，Vision Transformer を利用してそれらを分類すると共に，物体検出器である YOLOv8，YOLO11，YOLO11 の改良版である SOD-YOLOv11，RT-DETR を利用しそれらを検出する．

キーワード：2170301 医療・福祉支援，2150203 画像分類，2150205 画像認識・理解，2130102 機械学習

Classifying and Detecting Yeast-Like Fungi from Gram Stained Smears Images

Abstract: Gram staining, which is used in the initial diagnosis of infectious diseases, is a technique to stain specimen materials with a staining solution and to estimate the species of bacteria by microscopic examination based on their staining properties and shapes. In this paper, we focus on yeast-like fungi that cause dermatophyte and two types of Gram positive bacilli that are Clostridium perfringens and Corynebacterium. Based on the annotations of each bacteria and yeast-like fungi in the Gram-stained images, we classify them by using image classifiers of VGG, MobileNet, DenseNet and Vison Transformer, and detect them by using object detectors of YOLOv8, YOLO11, SOD-YOLO11 as the improvement of YOLO11 and RT-DETR.

Keywords: 2170301 Medical・Welfare Support, 2150203 Image Classification 2150205 Image Recognition・Understanding, 2130102 Machine Learning

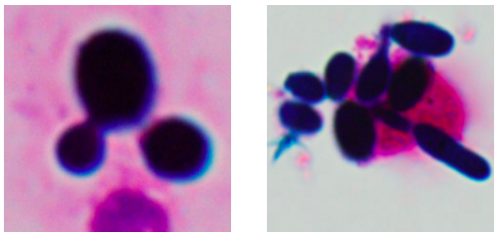
1. はじめに

感染症とは，病原体が体内に侵入し，さまざまな症状を引き起こす状態を指す．病原体には，菌や寄生虫などが含まれ，一般に感染症はこれらの病原体によって発生する．感染症診断において，適切な治療法の選択や感染症の迅速な特定を行うために，グラム染色を用いる手法がある．そのため，グラム染色は感染症の診断と治療において重要な意義を持つ．

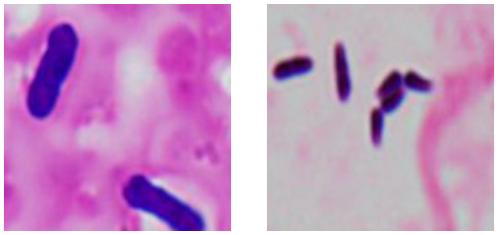
グラム染色 [1] は，デンマークの細菌学者 Hans Gram によって考案された，菌を染色する手法の一つである．グラム染色では，細胞壁の構造の違いによって，染まる色が異なりグラム陽性菌とグラム陰性菌の 2 種類に分類される．具体的には，グラム陽性菌は紫色・藍色に染まり，ペプチドグリカン層が厚い構造を有している．一方で，グラム陰性菌は赤色・桃色に染まり，ペプチドグリカン層が薄い構造を持っている．この違いにより，菌の性質や構造を視覚的に判断することが可能となる．また，グラム染色ではグラム陽性菌とグラム陰性菌から，さらに桿菌と球菌の 2 種類に分類される．桿菌は細長い棒状の形状を持つ菌であり，球菌は丸い形状を有している．これにより，グラム陽性球菌，グラム陽性桿菌，グラム陰性球菌，グラム陰性桿菌の 4 種類に分類される．

さて，グラム染色画像には本論文で対象とする真菌も出現する．真菌は，酵母様真菌と糸状真菌の 2 つに分類することができ，皮膚糸状菌感染症から重度の免疫不全患者に

[1] 九州工業大学大学院情報工学府
Graduate School of Computer Science and Systems Engineering, Kyushu Institute of Technology
[2] 九州工業大学情報工学研究院
Department of Artificial Intelligence, Kyushu Institute of Technology
[a)] tanaka.tsubasa391@mail.kyutech.jp
[b)] hirata@ai.kyutech.ac.jp



(a) YF



(b) CP (c) Cr

図 1: YF，CP，Cr のグラム染色画像
Fig. 1 Gram Stained Smears Images of YF, CP, and Cr.

おける深在性の感染症まで，幅広い疾患を引き起こす [4]．代表的な酵母様真菌には，カンジダ属がある．

酵母様真菌のグラム染色画像は，グラム陽性菌のように染色されており，形状は球状または桿状をしている．このような酵母様真菌の色と形状はグラム陽性桿菌であるクロストロジウム・パーフリンゲンスやコリネバクテリウムと似ている．**図 1** は，酵母様真菌（YF），クロストロジウム・パーフリンゲンス（CP），コリネバクテリウム（Cr）のグラム染色画像である．そこで本論文では，画像分類器と物体検出器を用いてグラム染色画像から酵母様真菌とグラム陽性桿菌であるクロストロジウム・パーフリンゲンス，コリネバクテリウムの分類と検出を行う．

グラム染色画像からの菌分類として，Sato ら [21] は，VGG16 [22] を使用して，グラム陰性桿菌である緑膿菌とアシネトバクター・バウマニを分類した．また，Yoshihara と Hirata [27] は，VGG16 とその改良を使用して，グラム陰性桿菌であるキャンピロバクターと白血球の貪食活性を分類した．さらに，Kawano ら [13] は，VGG16 と VGG19 [22]，MobileNet [6]，DenseNet [7] を使用して，4 種類のグラム陽性球菌，1 種類のグラム陰性球菌，2 種類のグラム陽性桿菌，6 種類のグラム陰性桿菌の計 13 種類の菌を分類した．これらの先行研究では，まず元の画像内の細菌の領域を抽出して収集し，次に画像分類器を適用して領域を分類しており，これは元の画像を分類するだけの本論文の分類とは異なっている．

グラム染色画像からの菌検出として，Yoshihara と Hirata [28] は，Faster R-CNN [20]，RetinaNet [15]，YOLOv5 [8] を使用して，白血球貪食とキャンピロバクターを検出した．また，Sugimoto と Hirata [23] は，Faster R-CNN，RetinaNet，YOLOv5 を使用して，3 種類のグラム陽性球菌とグラム陰性球菌であるブランハメラ・カタラーリスを検出した．これらの先行研究によって，グラム染色画像からの菌検出には，Faster R-CNN や RetinaNet よりも YOLOv5 の方が優れていると結論付蹴られた．さらに，Kashino ら [12] は，YOLOv5 と YOLOv7 [26] を使用して，13 種類の菌 [13] を検出した．Tanaka と Hirata [24] は，SSD [16]，M2Det [29]，RT-DETR [17]，YOLOv8 [9] を使用して，3 種類のグラム陽性球菌，1 種類のグラム陽性桿菌，3 種類のグラム陰性桿菌の計 7 種類の菌を検出した．Kashino ら [11] は，shifting convolution layers を備えた YOLOv5，YOLOv7，YOLOv8 を使用して，13 種類の菌 [13] と結核菌を検出した．

本論文では，YF を含み CP と Cr を含まないグラム染色画像を YF 画像，CP と Cr を含み YF を含まないグラム染色画像を非 YF 画像として，画像分類器である VGG，MobileNet，DenseNet，Vision Transformer（ViT）[3] を使用して，YF 画像または非 YF 画像に分類する．次に，物体検出器である YOLOv8，YOLO11，SOD-YOLO11，RT-DETR を使用して，グラム染色画像から YF，CP，Cr を検出する．

2. 画像分類器

2.1 VGGNet

VGGNet [22] は，3×3 の畳み込みフィルターを備えた基本的な畳み込みニューラルネットワーク（CNN）である．VGGNet の複数の CNN の後には，3 つの全結合層で構成されている．VGG16 は，13 層の CNN と 3 層の全結合層で合計 16 層を持つ VGGNet であり，VGG19 は，16 層の CNN と 3 層の全結合層で合計 19 層を持つ VGGNet である．本論文では，VGG16 と VGG19 を使用する．

2.2 MobileNet

MobileNet [6] は，最初の層は通常の CNN であり，その後の層は Depthwise Separable Convolution で構成されている．通常の CNN では，入力をフィルタリングして出力する処理を 1 ステップで処理するが，Depthwise Separable Convolution では，空間方向に畳み込みした後チャンネル方向に畳み込みする．MobileNet では，すべての層の後にはバッチ正規化，ReLU が続く．本論文では，MobileNetV2（MNV2），MobileNetV3-Small（MNV3-S），MobileNetV3-Large（MNV3-L）を使用する．

2.3 DenseNet

DenseNet [7] は，すべての層を互いに直接接続するような構造となっている．これにより，パラメータ数が削減されるとともに，勾配消失問題の緩和にも寄与する．本論文では，層の数が 121 層，161 層，169 層，201 層ある DenseNet-121（DN-121），DenseNet-161（DN-161），DenseNet-169

（DN-169），DenseNet-201（DN-201）を使用する．

2.4 Vision Transformer

Vision Transformer（ViT） [3] は，従来の CNN とは異なり，自然言語処理で使用される Transformer [25] を画像処理に応用したものである．画像を小さなパッチに分割し，それをトークンとして Transformer エンコーダーで処理する．Transformer エンコーダーは，Self-Attention を活用して，画像全体の文脈情報を効率的に学習する．ViT には，Base，Large，Huge のモデルがあり，それぞれ層の数とハイパーパラメータが異なっている．また，パッチに分割する際のパッチサイズが 16 または 32 のモデルがある．本論文では，ViT-B/16，ViT-L/16，ViT-B/32，ViT-L/32 を使用する．

2.5 画像分類の評価方法

画像分類での評価指標として，適合率，再現率，F 値を使用する．

$$\text{適合率} = \frac{TP}{TP + FP}$$

$$\text{再現率} = \frac{TP}{TP + FN}$$

$$\text{F 値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}}$$

TP（真陽性）は，モデルが真と予測した画像のうち，実際に真である画像の数，FP（偽陽性）は，モデルが真と予測した画像のうち，実際には偽である画像の数，FN（偽陰性）は，モデルが偽と予測した画像のうち，実際には真である画像の数である．

3. 物体検出器

3.1 YOLOv8

Redmon ら [19] によって開発された YOLO（You Only Look Once）は，物体の位置とそのクラスを同時に予測（one-stage）することによって，従来の物体検出手法より高速に処理することができる．

YOLOv8 [9] は，Ultralytics 社によって提案された手法で，Backbone，Neck，Output によって構成されている．**図 2** は，YOLOv8 のネットワークの構造図である．YOLOv8 では，Backbone と Neck に C2f を使用している．C2f は，まず入力を 1 層の CNN で処理した結果を 2 つに分割する．その後，片方を BottleNeck で複数回処理し，それぞれの途中結果と最初に 2 つに分割した結果を結合する．最後に，1 層の 1×1 の CNN で処理する．BottleNeck は，CNN を 2 層連結させた構造となっている．

YOLOv8 は，深さ乗数と幅乗数の違いにより，n，s，m，l，x の 5 つの事前モデルが準備されている．**表 1** は，YOLOv8 の事前モデルの深さ乗数と幅乗数である． 本論文では，YOLOv8 の 5 つの事前モデルをすべて利用する．

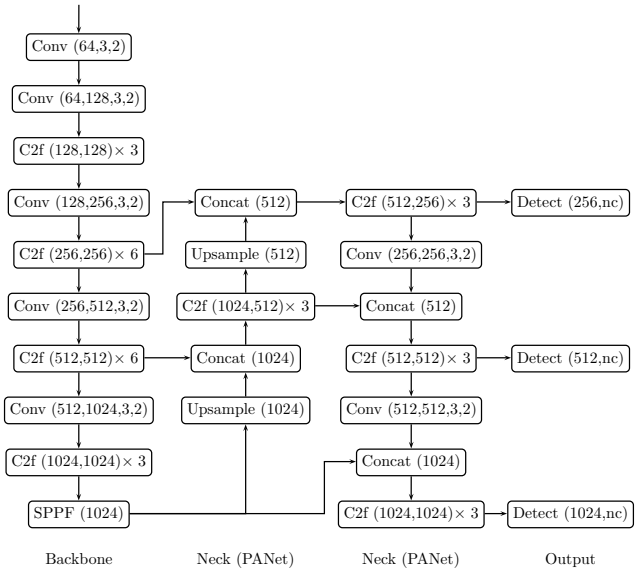


図 2: YOLOv8 の構造
Fig. 2 Architecture of YOLOv8.

表 1: YOLOv8 の事前モデルの深さ乗数と幅乗数
Table 1 Depth Multiple and Width Multiple for Each Prepared Model in YOLOv8.

モデル	深さ乗数	幅乗数
n	0.33	0.25
s	0.33	0.50
m	0.67	0.75
l	1.00	1.00
x	1.00	1.25

表 2: YOLO11 の事前モデルの深さ乗数と幅乗数
Table 2 Depth Multiple and Width Multiple for Each Prepared Model in YOLO11.

モデル	深さ乗数	幅乗数
n	0.50	0.25
s	0.50	0.50
m	0.50	0.75
l	1.00	1.00
x	1.00	1.50

3.2 YOLO11

YOLO11 [10] は，YOLOv8 と同様に Ultralytics 社によって提案された手法である．**図 3** は，YOLO11 のネットワークの構造図である．YOLOv8 で使用されていた C2f を C3k2 に置き換えた構造になっている．C3k2 は，まず入力を 1 層の CNN で処理する．その後，C3k で複数回処理し，その結果と最初に CNN で処理した結果を結合する．最後に，1 層の 1×1 の CNN で処理する．C3k は，まず入力を 1 層の CNN で処理する．その後，BottleNeck で複数回処理し，その結果と最初に CNN で処理した結果を結合する．最後に，1 層の 1×1 の CNN で処理する．

YOLOv8 と同様に，YOLO11 にも深さ乗数と幅乗数の違いにより，n，s，m，l，x の 5 つのモデルが準備されている．**表 2** は，YOLO11 の事前モデルの深さ乗数と幅乗数である．本論文では，YOLO11 の 5 つの事前モデルをすべて利用する．

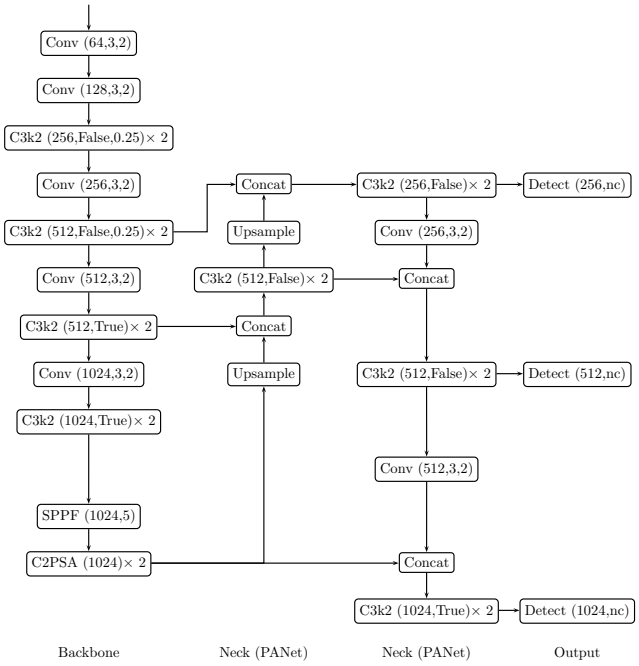


図 3: YOLO11 の構造

Fig. 3 Architecture of YOLO11.

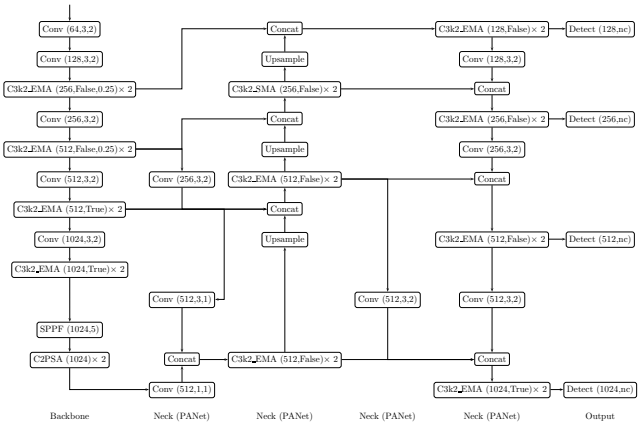


図 4: SOD-YOLO11 の構造

Fig. 4 Architecture of SOD-YOLO11.

3.3 SOD-YOLO11

SOD-YOLOv8 [14] は，YOLOv8 を小物体検出用に改良したものである．**図 4** は，SOD-YOLO11 のネットワークの構造図である．具体的には，通常の YOLO では，80×80, 40×40，20×20 の 3 つのサイズで最終的に Head で処理されるが，SOD-YOLO では，小物体を検出しやすくするために 160×160，80×80，40×40，20×20 の 4 つのサイズで処理される．また，それに伴って Neck にいくつかの層が追加されている．さらに，Neck では C2f を C2f-Att に置き換えている．C2f-Att は C2f の途中に Attention を追加した構造となっている．

本論文では，YOLOv8 でなく YOLO11 に対して同様の改良を施した検出器 SOD-YOLO11（SOD11）を利用する．また，Attention には EMA [18] を使用する．

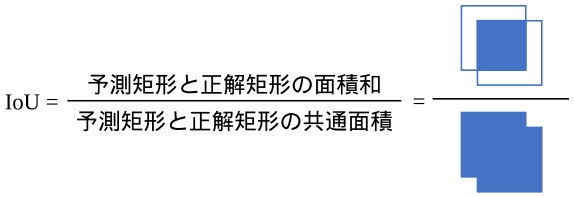


図 5: IoU

Fig. 5 IoU.

3.4 RT-DETR

DETR（Object Detection with Transformers） [2] は，自然言語処理で使用される Transformer [25] を使用した物体検出器である．DETR は，Backbone で特徴抽出した後，Transformer で処理される．その後，矩形の位置と大きさを予測する層とクラス分類する層を連結した構造となっている．そして，スコア最小の 2 部グラフマッチングを求めるアルゴリズムであるハンガリアン法により，最終的に予測した矩形と正解矩形とを対応付ける．

RT-DETR（Real-time Object Detection）[17] は，Baidu らによって開発された，高い検出性能を維持したままリアルタイムに物体を検出できる検出器である．RT-DETR は，Backbone，Efficient Hybrid Encoder，Transformer のデコーダによって構成されている．Backbone は，複数回の畳み込み層を行いマルチスケール特徴抽出をする．Efficient Hybrid Encoder では，Backbone によって処理されたマルチスケール特徴をスケール内での相互作用（AIFI）とスケール間での結合（CCFM）によって，画像特徴に変換する．その後，IoU-aware Query Selection によって Efficient Hybrid Encoder での出力を指定された数だけ選択する．最後に，auxiliary prediction heads を備えた Transformer のデコーダで処理され，矩形の位置と大きさとクラスを予測する．

RT-DETR には，l，x とバックボーンに ResNet50, ResNet101 [5] を適用した 4 種類のモデルがある．本論文では，RT-DETRl（DTl），RT-DETRx（DTx），RT-DETR-R50（DT-R50），RT-DETR-R101（DT-R101）を使用する．

3.5 物体検出の評価方法

IoU（Intersection over Union）は，予測矩形と正解矩形の面積和に対する予測矩形と正解矩形の共通部分の面積比である．**図 5** は，IoU の計算方法を表している．そして，閾値 δ（$0 \leq \delta \leq 1$）に対して，IoU が δ 以上となる予測矩形を正解とする．適合率は，ネットワークが予測した矩形の中で実際に正解の矩形である割合である．また，再現率は，正解矩形の中でネットワークが正しく正解と予測した矩形の割合である．さらに，複数画像の検出結果を評価するためには，平均適合率（Average Precision，AP）を利用する．これは，同一再現率の画像の適合率の平均をすべ

表 3: 計算機環境
Table 3 Computer Environment.

OS	Windows 10 Home
CPU	11th Gen Intel(R) Core(TM) i9-11900 @ 2.50GHz
RAM	16.0 GB
GPU	NVIDIA GeForce RTX 3060

表 4: 分類における YF，CP，Cr の画像数
Table 4 Number of Images in the Classification.

	対象	画像数
YF 画像	YF	208
非 YF 画像	CP	12
	Cr	55

ての再現率に対して平均した値である．複数クラスが存在する場合は，クラスごとに AP を計算をする．IoU の閾値 $\delta = 0.5$ のときの AP を AP50 という．また，mAP（mean AP）とは，δ を 0.5 から 0.95 まで 0.05 刻みで算出した AP の平均である．本論文では，AP50 と mAP を検出の評価指標として使用する．

4. 酵母様真菌の分類と検出

4.1 計算機環境

表 3 は，酵母様真菌の分類と検出における計算機環境である．この環境において，分類と検出の両方で 5 分割交差検証を適用する．5 分割交差検証では，まず，すべてのデータを 5 等分し，そのうち 3 つを訓練データ，1 つを検証データ，1 つをテストデータとする．そして，訓練データによって訓練し，モデルのパラメータを調整する．また，訓練中に検証データを使って過学習が起きていないことなどを検証する．訓練が終了した後，テストデータを使用してモデルを評価する．これを 1 セットとし，訓練データ，検証データ，テストデータの分け方を分割したデータがすべてテストデータとなるように変更しながら合計 5 セット行い，5 回の平均を評価指標として使用する．

4.2 酵母様真菌の分類

表 4 は，分類における YF，CP，Cr の画像数である．また，学習回数を 100 エポック，バッチサイズを 8 に設定している．**表 5** は，それぞれの画像分類器による適合率，再現率，F 値である．ここで，太字はそれぞれの指標での最大値を表している．また，時間は画像 1 枚あたりの推論時間の平均である．

表 5 より，F 値に着目すると，YF 画像はすべての分類器で 0.9 以上となっており，どの分類器も十分大きな F 値であることが分かる．特に，VGG と ViT は他の分類器と比較して YF 画像と非 YF 画像の F 値が大きく，その中でも ViT-B/16 が最も F 値が大きい分類器であるため，酵母様真菌の分類において，これらの分類器は有用である．

表 5: 酵母様真菌の分類
Table 5 Classification Results of Yeast-Like Fungi.

	時間	YF 画像			非 YF 画像		
画像分類器	(ms)	適合率	再現率	F 値	適合率	再現率	F 値
VGG16	7.77	0.991	0.990	0.990	0.973	0.971	0.970
VGG19	7.43	**0.995**	0.990	0.993	0.973	**0.986**	0.978
MNV2	9.95	0.910	**1.000**	0.952	**1.000**	0.685	0.806
MNV3-S	9.55	0.917	0.966	0.940	0.888	0.718	0.777
MNV3-L	9.80	0.913	0.990	0.949	0.968	0.692	0.772
DN-121	12.28	0.991	0.971	0.980	0.927	0.969	0.943
DN-161	11.61	0.995	0.957	0.975	0.886	**0.986**	0.932
DN-169	12.01	0.971	0.971	0.971	0.915	0.911	0.911
DN-201	12.45	0.990	0.933	0.960	0.834	0.970	0.894
ViT-B/16	8.96	**0.995**	0.995	**0.995**	0.986	**0.986**	**0.985**
ViT-L/16	8.64	**0.995**	0.990	0.993	0.971	**0.986**	0.978
ViT-B/32	7.97	**0.995**	0.990	0.993	0.971	**0.986**	0.978
ViT-L/32	8.50	0.995	0.990	0.993	0.972	**0.986**	0.979

表 6: 検出における YF，CP，Cr の画像数と矩形数
Table 6 Number of Images and Regions in the Detection.

対象	画像数	矩形数
YF	208	2,380
CP	12	144
Cr	55	3,041

4.3 酵母様真菌の検出

表 6 は，検出における YF，CP，Cr の画像数と矩形数である．また，学習回数をすべての検出器で 500 エポック，バッチサイズを YOLOv8 と YOLO11 は 8, SOD-YOLO11 は 4, RT-DETR は 2 に設定している．**表 7** は，それぞれの物体検出器による AP50 と mAP である．ここで，太字はそれぞれの指標での最大値を表している．また，時間は画像 1 枚あたりの推論時間の平均である．

表 7 より，AP50 に着目すると YF は YOLOv8s，CP は YOLOv8x，Cr は RT-DETR-R101 が最大となる．また，mAP に着目すると YF は SOD-YOLO11m，CP は YOLO11x，Cr は RT-DETR-R101 が最大となる．RE-DETR の mAP は他の検出器の mAP よりも大きいが，これは Cr の大きさが YF や CP と比較して大きいことが原因であると考えられる．したがって，小さい菌の検出には RT-DETR の方が，酵母様真菌や大きい菌の検出には YOLO の方が適していると考えることができる．

5. まとめと今後の課題

本論文では，酵母様真菌とグラム陽性桿菌を分類し，検出した．分類では ViT と VGG，検出では酵母用真菌は SOD-YOLO11m，クロストリジウム・パーフリンゲンスは YOLO11x，コリネバクテリウムは RT-DETR-R101 が有

表 7: 酵母様真菌の検出
Table 7 Detection Results of Yeast-Like Fungi.

	時間	YF		CP		Cr	
物体検出器	(ms)	AP50	mAP	AP50	mAP	AP50	mAP
YOLOv8n	14.66	0.828	0.389	0.899	0.632	0.625	0.300
YOLOv8s	15.16	**0.854**	0.412	0.944	0.642	0.689	0.365
YOLOv8m	24.04	0.833	0.397	0.952	0.665	0.728	0.393
YOLOv8l	33.30	0.844	0.416	0.950	0.685	0.731	0.401
YOLOv8x	50.80	0.840	0.407	**0.963**	0.688	0.745	0.411
YOLO11n	16.26	0.834	0.396	0.931	0.636	0.620	0.301
YOLO11s	15.90	0.847	0.406	0.935	0.646	0.701	0.367
YOLO11m	26.18	0.832	0.407	0.942	0.683	0.747	0.416
YOLO11l	28.92	0.836	0.406	0.945	0.687	0.741	0.400
YOLO11x	38.14	0.845	0.415	0.954	**0.692**	0.750	0.421
SOD11n	25.50	0.841	0.405	0.928	0.641	0.799	0.450
SOD11s	29.74	0.848	0.411	0.932	0.654	0.844	0.493
SOD11m	47.16	0.843	**0.416**	0.930	0.663	0.863	0.514
SOD11l	56.24	0.842	0.411	0.941	0.690	0.864	0.515
SOD11x	87.44	0.838	0.410	0.931	0.658	0.865	0.519
DTl	26.28	0.713	0.375	0.813	0.478	0.901	0.663
DTx	41.40	0.585	0.276	0.723	0.394	0.787	0.530
DT-R50	30.66	0.732	0.388	0.825	0.484	0.911	0.667
DT-R101	35.76	0.737	0.385	0.832	0.489	**0.921**	**0.675**

用であった．一方で，分類の F 値は大きかったが，検出における酵母用真菌の mAP は最大で 0.416 であり，数値としては大きくはない．そこで今後は，YOLO の層の追加や改良を施し，酵母用真菌の mAP を上げることが重要な課題である．

参考文献

[1] J. W. Bartholomew, T. Mittwer: *The Gram stain*, Bacteriol. Rev. **16**, 1–29 (1952). https://doi.org/10.1128/br.16.1.1- 29.1952.

[2] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov and S. Zagoruyko: *End-to-end object detection with transformers*, Proc. ECCV'20, LNCS **12346**, 213–229 (2020). https://doi.org/10.1007/978-3-030-58452-8_13.

[3] A. Dosoviskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby: *An image is worth 16 × 16 words: Transformers for image recognition at scale*, Proc. ICLR'21 (2021). https://openreview.net/forum?id=YicbFdNTTy.

[4] S. Gillespie, K. Bamford: *Medical microbiology and infection at a glance* (3rd edition), Blackwell Publishing (2007).

[5] K. He, X. Zhang, S. Ren, J. Sun: *Deep Residual Learning for Image Recognition*, Proc. CVPR'16, 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90.

[6] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M Andreetto, H. Adams: *MobileNets: Efficient convolution neural networks for mobile vision applications*, arXiv:1704.04861 (2017).

[7] G. Huang, Z. Liu, L. van der Maaten, K. Q. Weinberger: *Densely connected convolution networks*, Proc. CVPR'17, 2261–2269 (2017). https://doi.org/10.1109/CVPR.2017.243.

[8] G. Jocher: *YOLOv5*, https://github.com/ultralytics/yolov5 (2020).

[9] G. Jocher, A Chaurasia, J. Qui: *YOLOv8 by Ultralytics*, https://github.com/ultralytics/ultralytics (2023).

[10] G. Jocher: *YOLO11 by Ultralytics*, https://github.com/ultralytics/ultralytics (2024).

[11] U. Kashino, K. Taira, K. Hirata: *Detecting bacteria from Gram stained smears images by the family of YOLOs*, Proc. DMIP'24, 6–10 (2025). https://doi.org/10.1145/3705927.3705929.

[12] U. Kashino, S. Terada, K. Hirata: *Detecting infectious disease-causing bacteria from Gram stained smear images*, Proc. ESKM'23, 13-18 (2023). https://doi.org/10.1109/IIAI-AAI59060.2023.00013.

[13] I. Kawano, E. Kurumida, S. Terada, K. Hirata: *Classifying Gram positive cocci and Gram negative bacilli in Gram stained smears images*, Proc. ESKM'22, 55–60 (2022). https://doi.org/10.1109/IIAIAAI55812.2022.00021.

[14] B. Khalili1, A. W. Smyth: *SOD-YOLOv8: Enhancing YOLOv8 for Small Object Detection in Traffic Scenes*, https://doi.org/10.48550/arXiv.2408.04786.

[15] T. Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár: *Focal loss for dense object detection*, Proc. ICCV'17, 2980–2988 (2017). https://doi.org/10.1109/ICCV.2017.324.

[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu and A. C. Berg: *SSD: Single shot multibox detector*, Proc. ECCV'16, LNCS **9905**, 21–37 (2016). https://doi.org/10.1007/978-3-319-46448-0_2.

[17] W. Lv, Y. Zhao, S. Xu, J. Wei, G. Wang, C. Cui, Y. Du, Q. Dang, Y. Liu: *DETRs beat YOLOs on real-time object detection*, arXiv:2304.08069 (2023).

[18] D. Ouyang, S. He, G. Zhang, M. Luo, H. Guo, J. Zhan, Z. Huang: *Efficient Multi-Scale Attention Module with Cross-Spatial Learning*, Proc. ICASSP'23, (2023). https://doi.org/10.1109/ICASSP49357.2023.10096516.

[19] J. Redmon, S. Divvala, R. Girshick, A. Farhadi: *You only look once: Unified, real-time object detection*, Proc. CVPR'16, 778-788, (2016).

[20] S. Ren, K. He, R. Girshick, J. Sun: Faster R-CNN: *Towards real-time object detection with region proposal networks*, IEEE Trans. Pattern Anal. Mach. Intell. **39**, 1137–1149 (2017). https://doi.org/10.1109/TPAMI.2016.2577031.

[21] B. D. Satoto, I. Utoyo, R. Rulaningtyas, E. B. Khoendori: *An improvement of Gram-negative bacteria identification using convolution neural network with fine tuning* Telekomnika **18**, 1397–1405 (2020). https://doi.org/10.12928/TELEKOMNIKA.vl8i3.14890.

[22] K. Simonyan, A. Zisserman: *Very deep convolution networks for large-scale image recognition*, Proc. ICLR'15 (2015).

[23] H. Sugimoto, K. Hirata: *Object detection as Gram positive cocci in Gram stained smear images*, Proc. ESKM'22, 134–137 (2022). https://doi.org/10.1109/IIAIAAI55812.2022.00035.

[24] T. Tanaka, K. Hirata: *Comparison with detection of bacteria from Gram stained smears images by various object detectors*, Proc. ESKM'24, 58–61 (2024). https://doi.org/10.1109/IIAI-AAI63651.2024.00020.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin: *Attention Is All You Need*, Proc. NIPS'17, 6000–6010 (2017).

[26] C. Y. Wang, A. Bochkovskiy, H. Y. M. Liao: *YOLOv7:*

Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, arXiv.2207.02696 (2022).

[27] K. Yoshihara, K. Hirata: *Detecting Campylobacter bacteria and phagocytotic activity of leukocytes from Gram stained smears images*, Proc. ESKM'21, 10–15 (2021). https://doi.org/10.1109/IIAI-AAI53430.2021.00002.

[28] K. Yoshihara, K. Hirata: *Object detection as Campylobacter bacteria and phagocytotic activity of leukocytes from Gram stained smears images*, Proc. ICPRAM'22, 534–541 (2022). https://doi.org/10.5220/0000155500003122.

[29] Q. Zhao, T. Sheng, Y. Wang, Z. Tang, Y. Chen, L. Cai, H. Ling: *M2Det: A single-shot object detector based on multi-level feature pyramid network*, Proc. AAAI'19, 9259–9266 (2019). https://doi.org/10.1609/aaai.v33i01.33019259.