Background-Aware Instance Segmentation for Early Detection of *E. coli* and *Salmonella* in Time-Stamped Microscopy Images

Bibek Koirala
Department of Computer Science, School of Computing
Southern Illinois University Carbondale (SIUC)
Carbondale, IL, USA
bibek.koirala@siu.edu

Anas M. Alsobeh
Department of Information Technology, School of Computing
Southern Illinois University Carbondale (SIUC)
Carbondale, IL, USA
anas.alsobeh@siu.edu

Namariq Dhahir
School of Agricultural Sciences
Southern Illinois University Carbondale (SIUC)
Carbondale, IL, USA
namariq@siu.edu

Amer AbuGhazaleh
School of Agricultural Sciences
Southern Illinois University Carbondale (SIUC)
Carbondale, IL, USA
aabugha@siu.edu

***Abstract*—Rapid detection of bacterial contamination in food supply chains is essential to prevent outbreaks and protect public health. Traditional microbiological methods, such as culture and biochemical testing, are slow and can delay necessary interventions. This study presents deep learning–based approaches for the early detection and classification of *Escherichia coli* and *Salmonella Typhimurium* in microscopy images captured during incubation periods ranging from 1.5 hours to 4 hours, with a step size of 30 min, under two different conditions: a plain background and an onion mixture background. We annotated a dataset of 2,200 high-resolution (2160×1620) images, collected under 60× magnification from the time frame 1.5 h to 4 h, using a combination of manual and semi-automated techniques with Mask R-CNN. For manual annotation, dataset was annotated using tkinter framework in python. For model training, only pure culture data were used. The initial pipeline employed Cellpose for segmentation and a ViT for classification across bacterial growth stages. However, Cellpose was unable to segment all instances in an image due to its requirement for a diameter parameter, which, under variable colony sizes, resulted in partial segmentation. To overcome these challenges, we adopted an end-to-end Mask R-CNN model for instance segmentation. Mask R-CNN achieved consistently strong mean Intersection over Union (mIoU) scores between 0.91 and 0.98 across growth stages. Over the full incubation period, the mean Average Precision (AP) and Average Recall (AR) at IoU threshold 0.5 were 0.93 and 0.96 for *E. coli*, and 0.945 and 0.975 for *Salmonella*, respectively, indicating robust detection performance across bacterial types and bacterial growth period. Using fine-tuned Mask R-CNN trained on background aware time-stamped microscopy datasets, our method achieves an mAP@0.5 of 0.95 for colonies cultured at 2 hours timepoint. This work focuses on proactive food safety monitoring in agricultural pipelines.**

***Index Terms*—Mask R-CNN, Computer Vision, Vision Transformer(ViT), Cellpose, bacterial detection, deep learning**

I. Introduction

Foodborne bacterial contamination poses significant public health risks, leading to illness outbreaks and economic losses across global food supply chains. Food manufacturers aim to identify harmful pathogens before products reach consumers, ideally within hours of processing [1]. Unfortunately, traditional culture-based methods require several days, often after products have already entered the supply chain or been consumed. Beyond the health risks, faster detection can also help companies avoid costly recalls and liabilities [2].

Traditional detection of bacteria includes multiple steps from preenrichment to biochemical testing, which often takes 5-7 days, liquid growth medium requires at least 18 hours prior to final read-out [3]. These timelines motivate screening approaches that flag likely-positive samples early while preserving confirmatory workflows. Significant research has been conducted on the early detection of bacteria. For instance, Kang et al. [4] proposed an approach that integrates hyperspectral microscope imaging (HMI) with a U-Net-based segmentation model and a convolutional neural network (CNN) for classification of foodborne bacteria, achieving over 90% accuracy. More recently, transformer-based architectures [5] such as the ViT [6], introduced by Vaswani et al., have been widely adopted for both classification and segmentation tasks, demonstrating high accuracy. In the work by Borhani et al. [7], a hybrid model combining CNN and ViT was applied to plant disease classification and achieved an impressive accuracy of 97.99%. Similarly, Santiago et al. [8] employed a ViT-based model, TransCrowd [9], for microorganism enumeration, and conducted comprehensive comparisons with various architectures including CNN [10], ResNet [11].

In another recent study, Mask R-CNN [12] was adapted

for instance segmentation of bacteria from soil samples collected across geographically diverse ecosystems—Sweden, Greenland, and Kenya [13]. The authors trained the model to detect and classify multiple microbial structures using a stepwise training approach involving transfer learning, manual annotations, and iterative self-labeling. They demonstrated that despite challenges such as image quality variation and morphological similarities among microbes, the model achieved high detection performance, with average precision and recall values exceeding 90% across all test sets.

In designing our methodology, we initially experimented with Cellpose for colony segmentation, as it was trained on diverse biological images including bacterial cytoplasmic and phase-contrast microscopy from the LiceCell dataset. We paired this with ViT for classification due to its ability to capture global contextual information across colonies. However, due to limitations in Cellpose's segmentation performance across varying colony sizes, we adopted Mask R-CNN for robust instance segmentation and simultaneous multi-class classification.

Although these studies have demonstrated the versatility of deep learning in bacterial detection and classification, models tailored to specific experimental conditions and capable of maintaining high accuracy across varying image qualities, bacterial morphologies, and environmental contexts remain necessary. In this work, we present a deep learning-based instance segmentation approach optimized for the early detection and classification of *E. coli* and *Salmonella* under controlled experimental setups. Our method integrates post-processing techniques to enhance segmentation accuracy, achieving improvements in precision, recall, and intersection-over-union (IoU) across multiple growth time frames and experimental groups. By providing high-resolution morphological segmentation coupled with quantitative performance evaluation, our study contributes to the development of robust, automated tools for rapid bacterial analysis, with potential applications in food safety monitoring and clinical diagnostics.

We present two key contributions: (i) the first detection of *E. coli* and *Salmonella* in background-aware, time-stamped microscopy dataset across plain and food-matrix environments, and (ii) a novel instance-segmentation pipeline that achieves 95% mAP@0.5 for early bacterial detection, enabling rapid food safety screening.

II. Materials and Methods

A. Dataset Collection

The dataset consisted of microscopic images of *Escherichia coli (E. coli)* and *Salmonella Typhimurium (Salmonella)* in three configurations: pure *E. coli* culture, pure *Salmonella* culture, and mixed cultures containing both species. We collected images at time intervals from 1.5 h to 4 h, with a step size of 30 min, under two experimental conditions: a plain background and an onion mixture background. At each time point, for both *E. coli* and *Salmonella*, approximately 100 images were collected per background condition, yielding a total of 2,200 images. Only pure culture images with both plain and onion mixture backgrounds were used for model training, except for *Salmonella* at 1.5 h and 2 h with the onion mixture background, due to the small size of the bacterial colonies, which was comparable to onion particles and made annotation difficult.

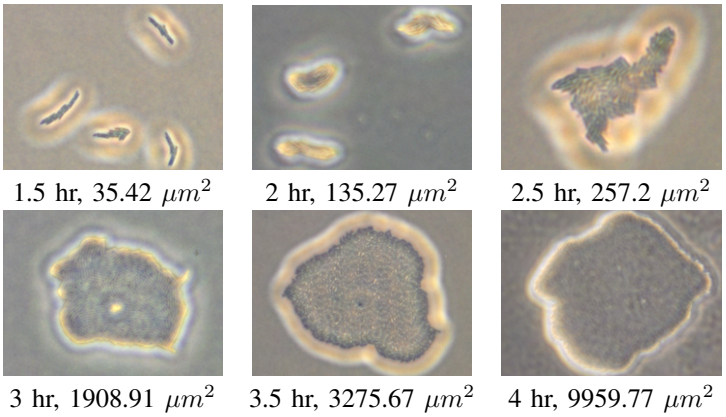


Fig. 1: Samples of *E. coli* colonies at different time frames with average area in μm^2.

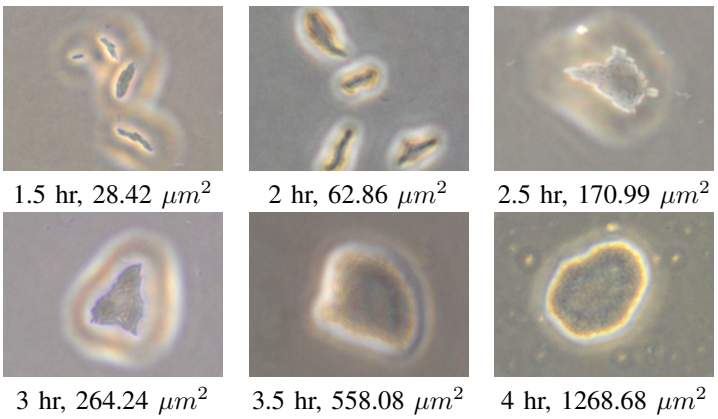


Fig. 2: Samples of *Salmonella* colonies at different time frames with average area in μm^2.

We captured all images in TIFF format at 2160×1620 pixel resolution, providing a corresponding to a physical field of view of 270×202.5 μm^2. This configuration yields a spatial resolution of 0.125 μm per pixel. We randomly partitioned the dataset into training (80%), validation (10%), and testing (10%) sets, reserving mixed culture images exclusively for generalization testing.

Figures 1 and 2 illustrate the growth progression of *E. coli* and *Salmonella* colonies, respectively, monitored from 1.5 h to 4 h at 30-minute intervals. A clear increase in colony size is observed with increasing incubation time.

B. Annotation Approach

Two annotation procedures were used: manual and semi-automatic. For the manual annotations, we developed a **Tkinter-based GUI annotation tool** [14], which can be used offline, supports freehand drawing, pixel-level segmen-

tation, and incorporates a human-in-the-loop workflow with correction capabilities, thereby making it easier to accurately annotate the dataset. The tool captured each instance and saved the annotations as `.png` images, with instance values set to 255 and the background set to 0. These instance masks were later merged during training.

For training Cellpose, we used the **Cellpose GUI** to annotate the dataset. Instance (colony) masks were labeled with integer values $\{0, 1, 2, 3, \dots\}$, where 0 represented the background and each non-zero integer corresponded to an individual bacterial colony (for both *E. coli* and *Salmonella*). After training the Mask R-CNN to a stabilized IoU of over 0.9, we used the model to generate annotations with manual verification. Representative instance annotations are shown in Figure 3. The manually annotated dataset in `.png` format was subsequently converted into COCO-style annotations for Mask R-CNN training. In this format, three classes were defined: *E. coli*, *Salmonella*, and background. The segmentation masks were encoded using Run-Length Encoding (RLE) to ensure compatibility with the COCO format, enabling efficient training and evaluation.

C. Microscopy Setup

All images were acquired using an IX73 inverted microscope (Evident Scientific, Inc., USA) coupled with a digital color camera (Olympus LC35, USA).

D. Model Architectures

1) Mask R-CNN: Mask R-CNN was trained using the **ResNet50** backbone with pretrained weights set to `DEFAULT`, which is equivalent to using the `COCO_V1` pretrained weights trained on the COCO dataset. The classification and mask prediction heads were replaced to match the bacterial classes and the background. The model was trained with a learning rate of 1×10^{-4}, a scheduler step of **5**, and a batch size of **4**. Run-length encoding (RLE) segmentation output was adopted. It was trained for 20 epochs, which produced consistent instance segmentation.

Figure 4 illustrates our adapted Mask R-CNN architecture for bacterial colony detection in microscopy images. Given an input image $I \in \mathbb{R}^{H \times W \times 3}$ of dimensions 2160×1620 pixels, the ResNet-50 backbone with Feature Pyramid Network (FPN) extracts multi-scale feature maps $\{C_2, C_3, C_4, C_5\}$ through successive convolutional layers, where each C_i has spatial resolution $\frac{H}{2^i} \times \frac{W}{2^i}$. The FPN constructs pyramid features $\{P_2, P_3, P_4, P_5, P_6\}$ to capture bacterial colonies across our observed size range $A \in [28, 9000]$ μm^2. The Region Proposal Network generates N candidate regions $\{r_1, r_2, ..., r_N\}$, which undergo RoI Align to extract fixed-size feature representations. Our modified architecture outputs three predictions per region: (1) classification scores $p_i = \text{softmax}(W_c^T f_i + b_c) \in \mathbb{R}^{K+1}$ for $K = 2$ bacterial classes plus background, (2) bounding box refinements $\Delta = (t_x, t_y, t_w, t_h) \in \mathbb{R}^4$, and (3) binary masks $M_i \in \{0, 1\}^{m \times m}$ where $m = 28$ for pixel-level segmentation.

E. Post-Processing

During inference, the model occasionally produced multiple candidate regions for the same object. To address this, we implemented a non-maximum suppression post-processing step. Multiple predictions with IoU > 0.9 for the same colony were filtered, retaining only the highest confidence prediction. Bacterial sizes were categorized into three classes based on mask area in pixels:

$$\text{Small: } A \leq 64^2 \text{ px}^2 \quad (\approx 8^2 \ \mu\text{m}^2) \tag{1}$$

$$\text{Medium: } 64^2 < A \leq 256^2 \text{ px}^2 \quad (\approx 8^2 - 32^2 \ \mu\text{m}^2) \tag{2}$$

$$\text{Large: } A > 256^2 \text{ px}^2 \quad (>\approx 32^2 \ \mu\text{m}^2) \tag{3}$$

Figure 5 presents a comparison of the total number of annotations for small, medium, and large instances between the default COCO annotations and those categorized using our defined size ranges. Figure 5 demonstrates that our size categorization better represents the actual distribution compared to COCO's default thresholds, which inappropriately shift small bacterial colonies into the medium category, potentially biasing evaluation metrics for early detection scenarios. When using the default COCO definitions, the majority of small instances are shifted to the medium category, introducing bias during the evaluation of small instances. In contrast, the size distribution in Figure 5 shows that the small category is more accurately represented using our defined ranges, leading to a more balanced and reliable evaluation.

F. Evaluation Metrics

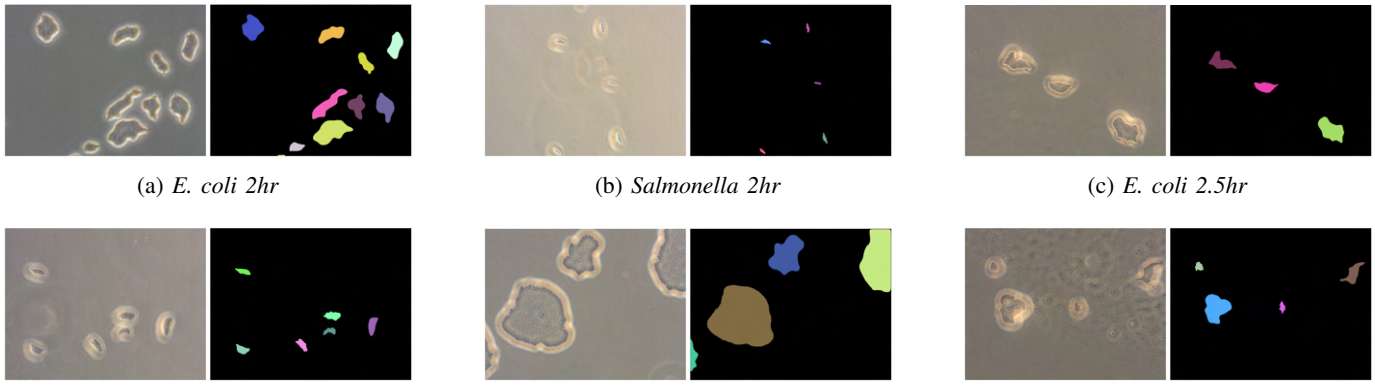
We evaluated model performance using standard instance segmentation metrics adapted for bacterial detection. For each prediction, we computed:

- Intersection over Union (IoU): Overlap between predicted and ground truth masks, measuring segmentation accuracy
- Mean Average Precision (mAP@0.5): Average precision across all classes at IoU threshold 0.5
- mAP@[0.5:0.95:0.05]: Average precision computed across IoU thresholds from 0.5 to 0.95 in 0.05 increments, providing comprehensive overlap assessment
- Average Recall (AR): Mean recall across all classes and IoU thresholds
- Precision, Recall, and F1-score: Computed per class and time point to assess detection performance across bacterial growth stages

All test set evaluations used IoU threshold 0.5 unless otherwise specified. We computed size-stratified metrics (small, medium, large) using our defined area thresholds to evaluate performance across different colony growth stages. For mixed culture testing, we relied on expert validation due to the absence of ground truth annotations.

G. Computational Resources

All experiments were conducted on a workstation equipped with an NVIDIA RTX A6000 GPU (12 GB VRAM). Fine-tuning the Mask R-CNN model for 20 epochs required approximately 8 hours of training time on this hardware.



(a) *E. coli 2hr* (b) *Salmonella 2hr* (c) *E. coli 2.5hr*

(d) *Salmonella 2.5hr* (e) *E. coli 3.5hr* (f) *Salmonella 3.5hr*

Fig. 3: Annotation showing instance-level bacterial masks generated during the annotation process at various incubation periods for size comparison.

III. Results

A. Segmentation and Classification Performance

1) Cellpose Segmentation: Since Cellpose was unable to segment all colonies of varying size comprehensive AP and AR values could not be computed for this method. Therefore, quantitative comparisons of AP and AR are only presented for the Mask R-CNN model. Some example segmentation results from Cellpose are shown in Figure 6.

Figure 6 demonstrates this fundamental limitation: at 1.5hr with diameter = 100px, smaller colonies were segmented but larger ones missed; conversely, diameter = 200px captured larger colonies but failed on smaller ones. At 3 h, using a diameter of 100 px failed to segment any colonies, with larger colonies consistently missed at this setting. This partial segmentation prevented computation of meaningful average precision and recall metrics, as the denominator (total ground truth colonies) remained unknown. The diameter-dependent performance made Cellpose unsuitable for time-series bacterial analysis where colony sizes vary dramatically across growth stages.

2) Mask R-CNN on Pure Culture: A Mask R-CNN model with a ResNet-50 backbone and Feature Pyramid Network (FPN) was fine-tuned on the annotated dataset for instance segmentation of *E. coli* and *Salmonella*. The architecture was adapted by replacing the default box and mask predictors to match bacteria classes. The training was performed for 20 epochs with a learning rate of 1×10^{-4}, a learning rate decay step of 5 epochs (factor 0.1), and a batch size of 4.

The loss and mIoU curves during training are shown in Figure 7. IoU starts low - Poor segmentation with rough, inaccurate masks, Rapid improvement to 0.9 by epoch 3-4, Colonies are detected but masks are imprecise, often over- or under-segmenting, and Many false positives as the model learns to distinguish colonies from background. The loss steadily decreases to 0.10, while IoU converges around 0.93 by the end of training, indicating stable convergence and strong segmentation performance.

Figure 8 shows the mean AP (mAP) at IoU thresholds ranging from 0.50 to 0.95. At IoU = 0.50, the model reached an mAP of ≈0.80 by the third epoch and stabilized around 0.85 after sixth epoch. As the IoU threshold increased, a gradual decline in mAP was observed: mAP@0.85 stabilized around 0.79, and at IoU = 0.90 it remained ≈0.71. The strictest threshold, IoU = 0.95, yielded a substantially lower mAP (≈0.35), reflecting the increased difficulty of achieving precise mask alignment at high overlap requirements.

The breakdown of AP and AR by object size is presented in Figure 9. Large objects achieve near-perfect AP and AR (≈0.99), medium objects maintain AP around 0.92, while small objects are more challenging, with AP peaking at ≈0.75.

The per-timepoint performance metrics for each bacterial species are summarized in Table I. These values were obtained by averaging the plain background and onion mixture background for each time frame, except for *Salmonella* at 1.5 hr and 2 hr, where only the plain background was used since the onion mixture background was not included. Since only 10% of the total data was used for testing, the results show consistently high precision, recall, and mIoU across all growth stages; however, smaller colonies under the onion mixture remain difficult to segment and classify accurately. To improve performance on the test dataset, a post-processing step was applied in which multiple segmentations of the same object with IoU greater than 0.9 were removed, retaining only the highest-confidence prediction.

Following the quantitative metrics in Table I, Figure 10 demonstrates our Mask R-CNN model's superior performance compared to the Cellpose limitations shown in Figure 6. Scale-Invariant Detection Success: Unlike Cellpose's diameter-dependent failures, our Mask R-CNN approach successfully detects colonies across the entire size spectrum within single images. At 1.5 hours, the model accurately identifies both small emerging colonies (28 μm²) and larger established ones without parameter adjustment. This scale invariance proves critical for time-series analysis where colony sizes vary dramatically. The segmentation quality remains

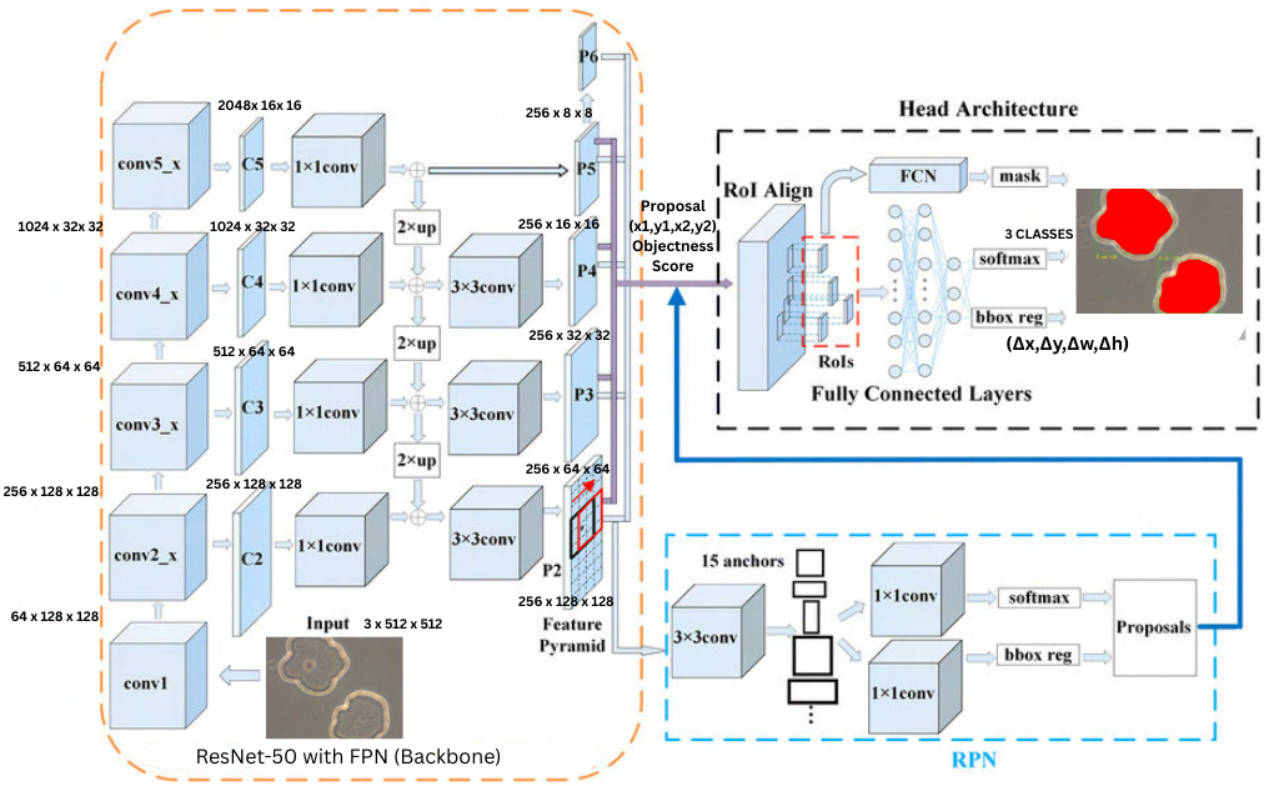


Fig. 4: Architecture of the Mask R-CNN model with ResNet50 backbone and Feature Pyramid Network [15], adapted for bacterial instance segmentation.

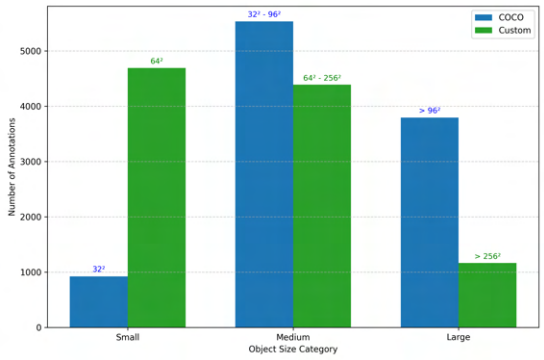


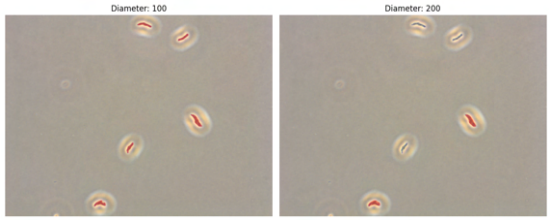
Fig. 5: Number of annotations distribution across small, medium, and large categories using COCO and our defined size thresholds

TABLE I: Per-timepoint performance metrics over time for *E. coli* and *Salmonella*.

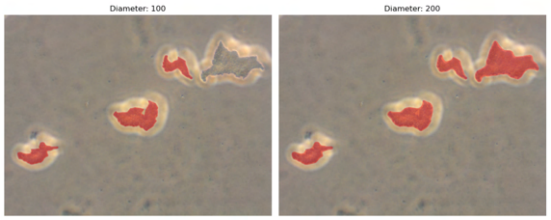
The column AvgSize is average size of the colonies taken in the time point as given

Category	Time	AvgSize (μm²)	Precision	Recall	F1	Sup	mIoU
E. coli	1.5 hr	34.46	0.92	0.94	0.93	534	0.91
	2 hr	123.65	0.95	0.92	0.93	467	0.93
	2.5 hr	421.79	0.93	0.96	0.95	223	0.96
	3 hr	1517.23	0.85	0.99	0.92	88	0.98
	3.5 hr	3398.80	0.92	0.98	0.95	109	0.98
	4 hr	4345.38	0.95	0.99	0.97	120	0.96
Salmonella	1.5 hr	49.87	0.88	0.98	0.93	320	0.93
	2 hr	47.96	0.95	0.93	0.94	182	0.93
	2.5 hr	125.25	0.97	0.98	0.97	224	0.95
	3 hr	105.40	0.92	0.99	0.95	203	0.95
	3.5 hr	566.88	0.97	0.98	0.98	123	0.97
	4 hr	929.74	0.98	0.99	0.99	122	0.96

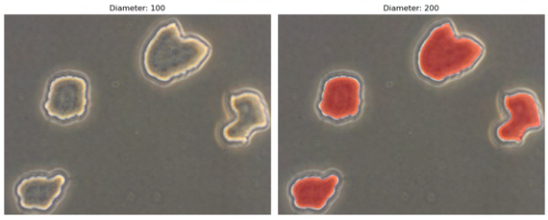
robust across growth stages. At 2 hours, when Cellpose would require different diameter settings for optimal performance, our model maintains consistent detection accuracy (95% precision for *E. coli*). The progression from 2.5 to 4 hours shows increasingly mature colonies with complex morphologies, all successfully segmented and classified with high confidence scores. The model reliably distinguishes between *E. coli* and *Salmonella* morphologies throughout the growth timeline. Our model successfully delineates these complex boundaries, as evidenced by the precise red mask overlays that closely follow colony contours. The bounding boxes (green) appropriately encompass the full colony extent, while confidence scores remain high despite morphological complexity. The precise segmentation boundaries demonstrate the model's ability to distinguish bacterial colonies from background artifacts. Each prediction pair (original left, segmentation right) illustrates the practical utility of our approach: clear visual confirmation of



(a) E. coli - 1.5 hr



(b) E. coli - 2.5 hr



(c) E. coli - 3 hr

Fig. 6: Segmentation results from Cellpose. (a) At 1.5 hr, Cellpose segmented smaller colonies with diameter 100 px, but diameter 200 px missed smaller colonies. (b) At 2.5 hr, larger colonies were missed with diameter 100 px. (c) At 3 hr, all the larger colonies were missed with diameter 100 px.

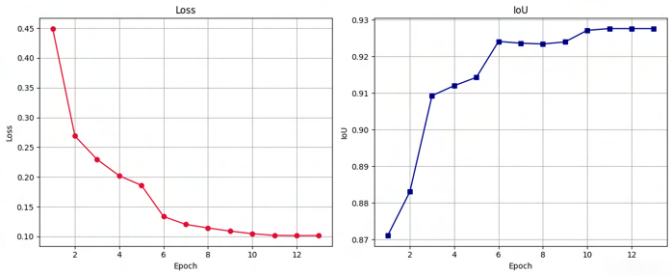


Fig. 7: Training loss and IoU trends for Mask R-CNN.

detected colonies with species identification and confidence quantification–essential features for food safety applications where rapid, reliable bacterial identification is critical.

3) Mask R-CNN on Mixed Culture: Accurate identification of bacteria in colonies under 4 hours requires expert knowledge. For testing purposes, images of mixed colonies were collected and visually inspected by experts. These images were also processed using Mask R-CNN. Although the model was not explicitly trained on mixed colonies, its predictions closely matched the expert identification, demonstrating robust generalization. Some sample predictions for different time points are shown in Figure 11.

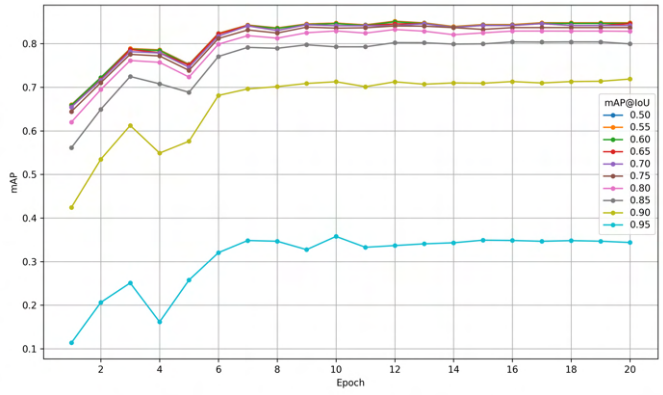


Fig. 8: mAP performance of Mask R-CNN across IoU thresholds (0.50 to 0.95) over 20 training epochs.

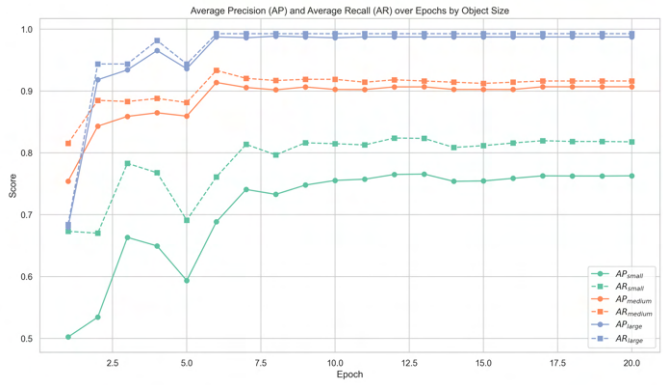


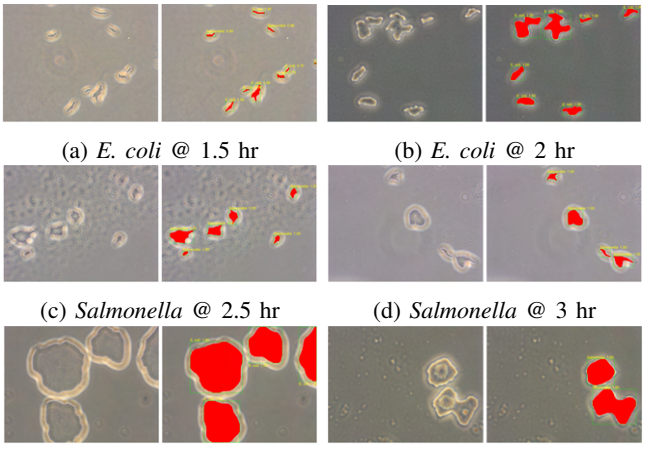
Fig. 9: Average Precision (solid lines) and Average Recall (dashed lines) for small, medium, and large bacterial instances over training epochs.

IV. Discussion

This study evaluated the performance of Mask R-CNN as an instance segmentation model for the early detection and classification of *E. coli* and *Salmonella* in microscopy images. Although a Cellpose–ViT pipeline was originally planned for comparison, it could not be implemented within the scope of this work and is left for future investigation.

Mask R-CNN performed very well in segmentation. It achieved high mIoU values (0.91–0.98) and strong average precision and recall across different growth stages. Its robustness extended to mixed culture scenarios, where it was not explicitly trained yet still closely matched expert identification, indicating strong generalization capability and suitability for real-world applications where mixed bacterial populations are common.

In mixed culture scenarios, *E. coli* colonies tend to grow faster than *Salmonella*, and when Cellpose is applied with a fixed diameter parameter, it often fails to capture both small and large colonies simultaneously. As a result, either the smaller or larger colonies are missed depending on the chosen diameter, making instance-level classification challenging. In contrast applying Mask R-CNN to mixed culture images

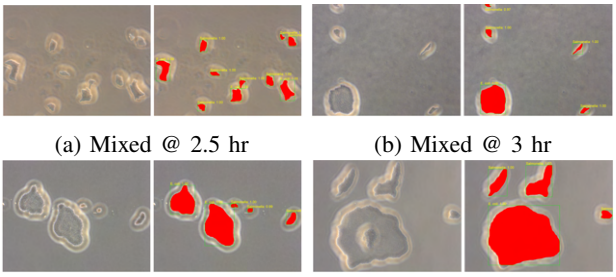


(a) *E. coli* @ 1.5 hr (b) *E. coli* @ 2 hr

(c) *Salmonella* @ 2.5 hr (d) *Salmonella* @ 3 hr

(e) *E. coli* @ 3.5 hr (f) *Salmonella* @ 4 hr

Fig. 10: Instance segmentation results from Mask R-CNN for *E. coli* and *Salmonella* across various time points. For each image pair, the original image is shown on the left, and the predicted segmentation and classification result is shown on the right. Red overlays indicate predicted masks, green bounding boxes highlight detected objects, and yellow labels show the predicted class with confidence scores.



(a) Mixed @ 2.5 hr (b) Mixed @ 3 hr

(c) Mixed @ 3.5 hr (d) Mixed @ 4 hr

Fig. 11: Pairwise visualization of Mask R-CNN predictions on mixed bacterial colonies at different time points. For each pair, the original image is shown on the left, and the corresponding segmentation and classification result is shown on the right. Predicted masks are overlaid in red, bounding boxes are shown in green, and labels in yellow indicate the predicted class and confidence score.

confirmed its capacity to detect and classify bacteria without retraining on mixed samples. The accurate segmentation overlays provided visual clarity and enabled expert validation. This finding is significant for real-time quality control in food supply chains, where samples may contain multiple bacterial species.

V. Limitations and Future Work

Detection of microcolonies below 2 hours remains challenging. This limitation may be addressed by training with additional early-stage data, expanding the dataset size, and employing augmentation or advanced feature extraction strategies. Future work should therefore focus on:

- Expanding the dataset to include mixed cultures and diverse environmental backgrounds.
- Introducing advanced or multi-scale feature learning prior to classification to improve detection of small bacterial instances.
- Developing hybrid architectures capable of achieving both high classification accuracy and precise instance segmentation [16], [17], and use High Performance Computing workstation to run complex models.
- Designing real-time deployment pipelines for in-field bacterial detection in food safety monitoring applications.

VI. Conclusion

This study demonstrates that both *E. coli* and *Salmonella* can be reliably detected and classified in microscopy images between 2–4 hours of growth using deep learning approaches. Among the evaluated models, Mask R-CNN provided the highest reliability for precise bacterial instance segmentation in both pure and mixed cultures, while also offering faster inference. These findings highlight the potential of deep learning for improving food safety monitoring through earlier and more accurate bacterial detection.

Acknowledgment

This work is supported by the United States Department of Agriculture, National Institute of Food and Agriculture (USDA-NIFA), through the Capacity Building Grants for Non-Land-Grant Colleges of Agriculture Grant Number: 2024-70001-43667 (Proposal No. 2024-02854). The authors gratefully acknowledge the following contributions: Dr. Namriq for data collection; Bibek for preprocessing and annotation work that formed the foundation of this research, as well as for building and implementing the model, conducting experiments, and performing the results analysis.

References

[1] A. AlSobeh, A. AbuGhazaleh, N. Dhahir, and M. Rababa, "XAIPath: Temporal-Environmental Explainable AI Framework for Co-Contaminated Food Pathogen Detection in Microscopic Imaging," in *Proceedings of the 54th International Conference on Parallel Processing Companion (ICPP Companion '25)*. ACM, Sep. 2025. [Online]. Available: https://doi.org/10.1145/3750720.3758080

[2] L. Ma, J. Yi, N. Wisuthiphaet, M. Earles, and N. Nitin, "Accelerating the detection of bacteria in food using artificial intelligence and optical imaging," *Applied and Environmental Microbiology*, vol. 89, no. 1, pp. e01 828–22, 2023.

[3] H. Wang, H. Koydemir, Y. Qiu, B. Bai, Y. Zhang, Y. Jin, S. Tok, E. Yilmaz, E. Gumustekin, Y. Rivenson *et al.*, "Early-detection and classification of live bacteria using time-lapse coherent imaging and deep learning. light sci. appl. 9, 118 (2020)."

[4] R. Kang, B. Park, M. Eady, Q. Ouyang, and K. Chen, "Classification of foodborne bacteria using hyperspectral microscope imaging technology coupled with convolutional neural networks," *Applied microbiology and biotechnology*, vol. 104, no. 7, pp. 3157–3166, 2020.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[7] Y. Borhani, J. Khoramdel, and E. Najafi, "A deep learning based approach for automated plant disease classification using vision transformer," *Scientific Reports*, vol. 12, no. 1, p. 11554, 2022.

[8] J. U. Santiago, T. Ströhle, A. Rodríguez-Sánchez, and R. Breu, "Vision transformers for weakly-supervised microorganism enumeration," in *2024 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2024, pp. 126–133.

[9] D. Liang, X. Chen, W. Xu, Y. Zhou, and X. Bai, "Transcrowd: weakly-supervised crowd counting with transformers," *Science China Information Sciences*, vol. 65, no. 6, p. 160104, 2022.

[10] K. O'shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[13] H. Zou, A. Sopasakis, F. Maillard, E. Karlsson, J. Duljas, S. Silwer, P. Ohlsson, and E. C. Hammer, "Bacterial community characterization by deep learning aided image analysis in soil chips," *Ecological Informatics*, vol. 81, p. 102562, 2024.

[14] K. Maeda, X. Ma, H. Lee, and S. Bird, "The annotation graphs toolkit (version 1.0): Application developer's manual," *Linguistic Data Consortium, University of Pennsylvania*, 2002.

[15] H. Huang, S. Zhao, D. Zhang, and J. Chen, "Deep learning-based instance segmentation of cracks from shield tunnel lining images," *Structure and Infrastructure Engineering*, vol. 18, no. 2, pp. 183–196, 2022.

[16] E. M. Al-Shawakfa, A. M. Alsobeh, S. Omari, and A. Shatnawi, "Radar#: An ensemble approach for radicalization detection in arabic social media using hybrid deep learning and transformer models," *Information*, vol. 16, no. 7, p. 522, 2025.

[17] A. AlSobeh, A. Shatnawi, B. Al-Ahmad, A. Aljmal, and S. Khamaiseh, "Ai-powered aop: Enhancing runtime monitoring with large language models and statistical learning." *International Journal of Advanced Computer Science & Applications*, vol. 15, no. 11, 2024.