Capstone Faculty-Sponsored Project - Bacteria classification problem

Data Science Institute
Columbia University

A. Rios, J. Dong, P. Singh, A. Punia, C. Provinciali, X. Cao

Faculty Mentor: Sining Chen
Faculty Sponsor: Tal Danino

Data Science Capstone Project
with Biomedical Engineering, SEAS

Problem Statement

Bacteria-related illnesses are responsible for approximately 5 million deaths per year worldwide[1]. Current microbial identification approaches are extraordinarily time consuming, laborious, and expensive which requires the development of a rapid, automated process to identify and characterize bacterial species from environmental and clinical samples.

Project Goal and Methods

The aim of this project is to build a computer vision model to identify and classify bacteria species from environmental samples that are grown on solid media in petri dishes containing colonies of mixed species (Figure 1).

We trained our model using images of petri dishes, each containing one out of 8 bacteria species in our sample. To annotate pixels corresponding to colonies ('*positive*' pixels), we used pixel value thresholding after detecting and removing the petri dish borders. We then removed noisy small dark regions using connected component labeling. From the processed images, we extracted '*positive*' and '*negative*' patches of different sizes for training.

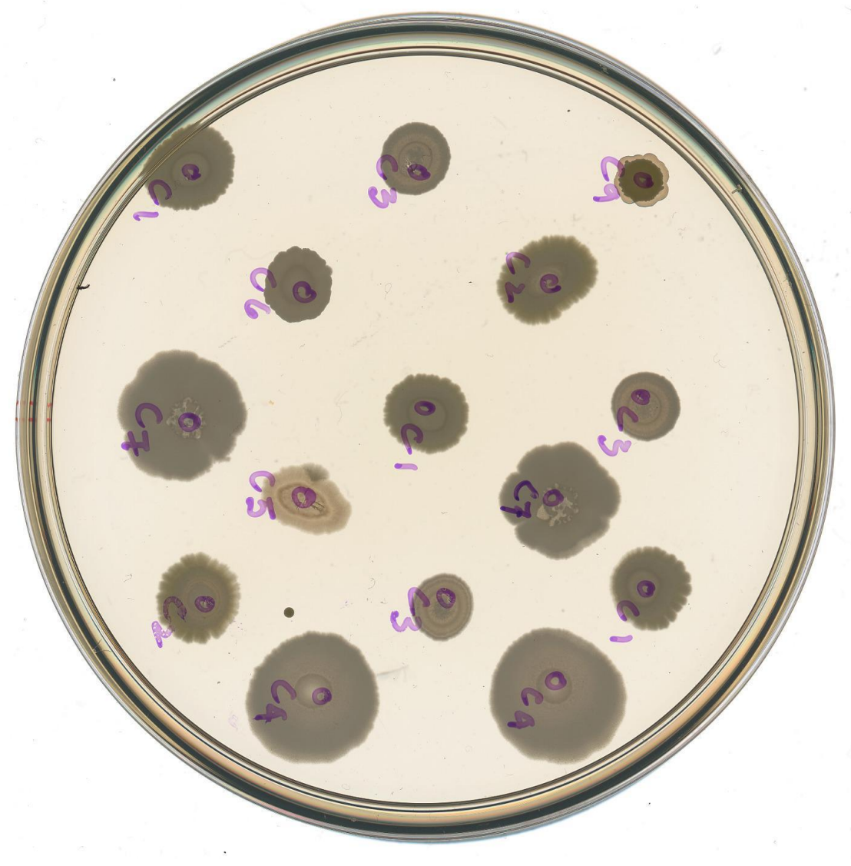
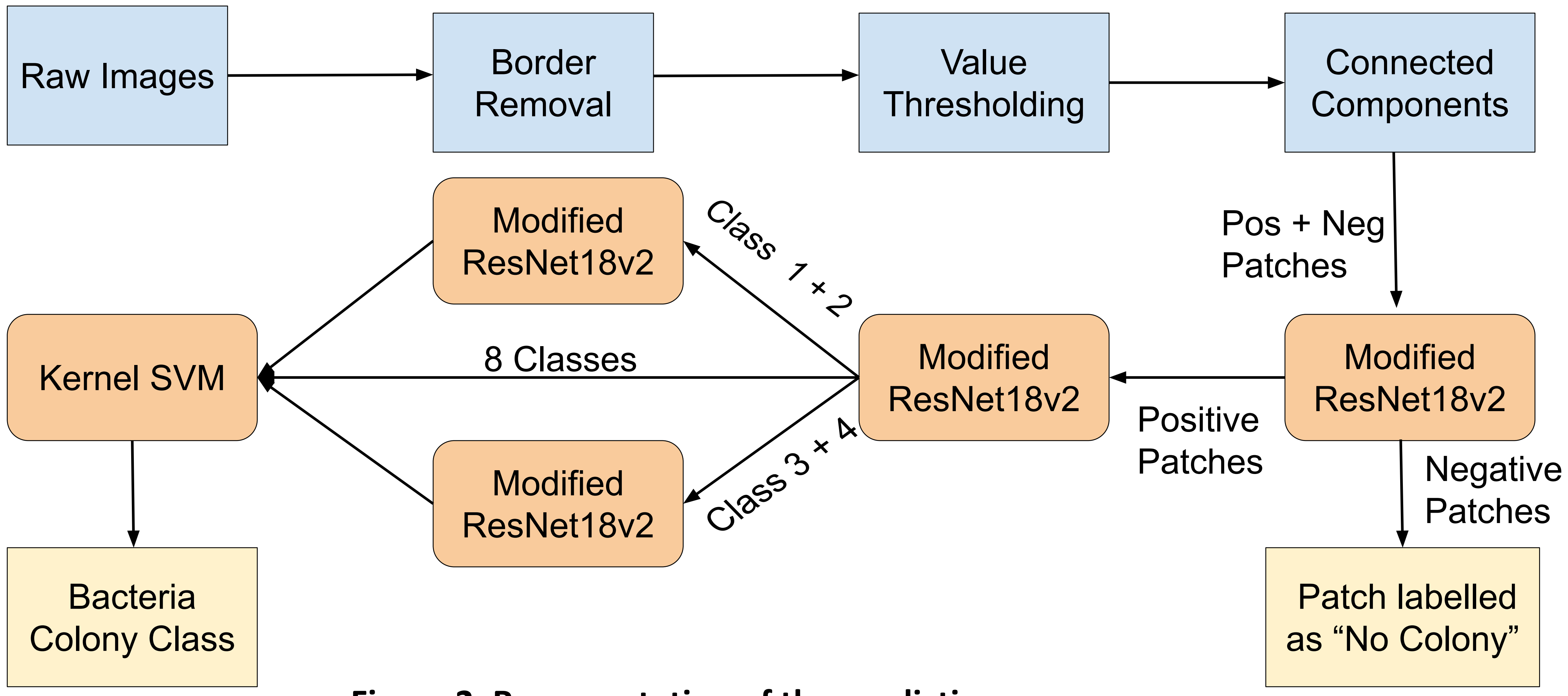


Figure 1. Petri dish with mixed bacteria species.



Raw Images
Border Removal
Value Thresholding
Connected Components
Pos + Neg Patches
Modified ResNet18v2
Negative Patches
Patch labelled as "No Colony"
Positive Patches
Modified ResNet18v2
Class 1 + 2
Modified ResNet18v2
Class 3 + 4
Modified ResNet18v2
8 Classes
Kernel SVM
Bacteria Colony Class

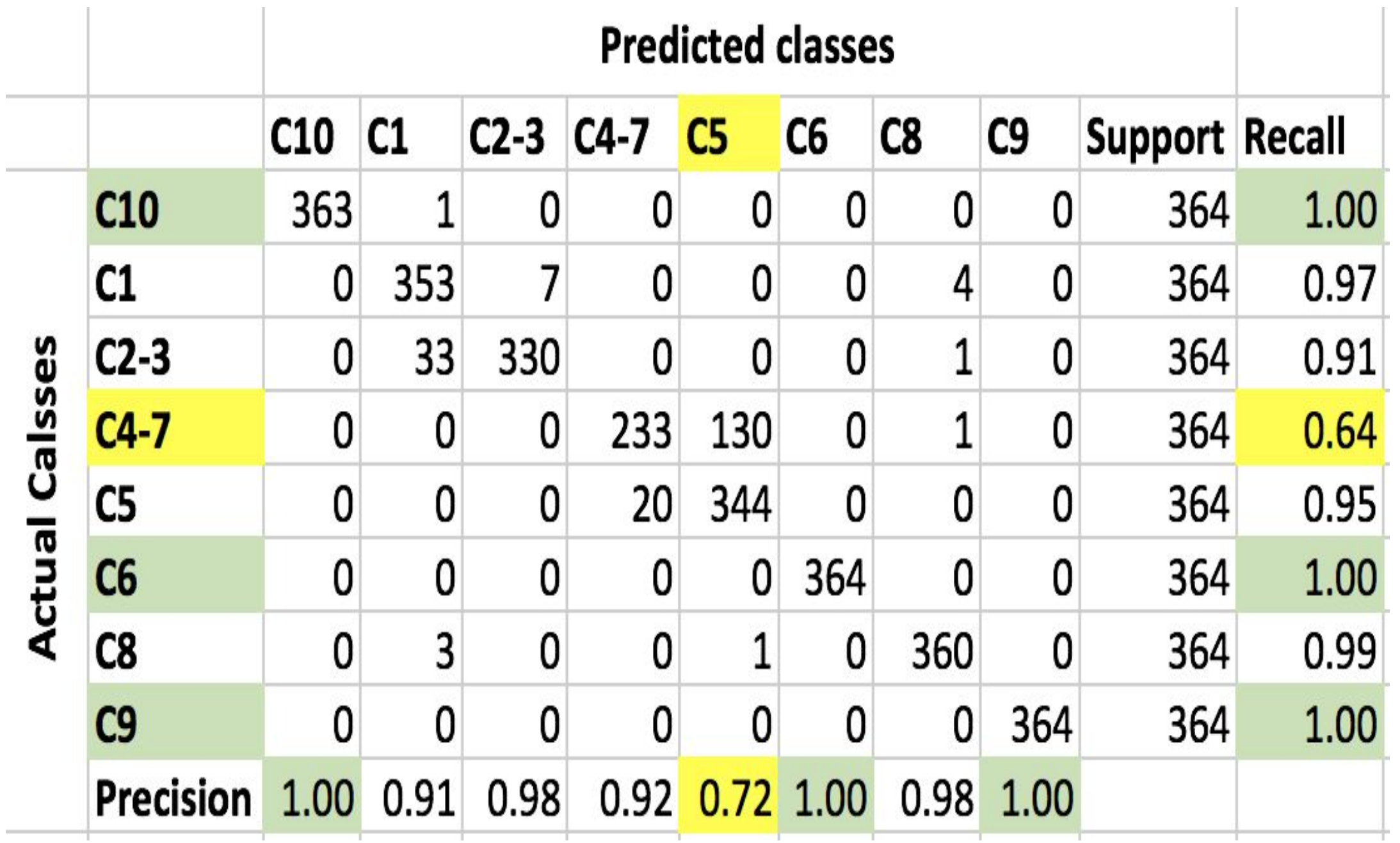
Figure 2. Representation of the prediction process.

We used the resulting patches to train 4 models using a modified ResNet18v2[2]. The first model classifies '*positive*' vs '*negative*' (i.e. no colonies) patches. The second model classifies each 'positive' patch as one of the 8 bacterial species. Due to some class pairs being harder to discriminate, we trained 2 other models that are specialized at classifying the more 'problematic' classes. Finally, we used the probability outputs from the 4 models as input to a kernel SVM for final prediction (Figure 2).

Results

Figure 3 shows the results of the patch classification: the model performs well on most classes, expect for C5 and C4-7, which may be due to their similar visual features. The overall prediction accuracy on balanced test datasets reaches 0.93.

Following the patch prediction, we use the heat map to integrate all patch classifications' results. By setting a threshold on the minimum positive predictions, we are able to filter out uncertain outputs. As shown in Figure 4. The model detects C10 successfully from other colonies.



		Predicted classes									
		C10	C1	C2-3	C4-7	C5	C6	C8	C9	Support	Recall
Actual Calsses	C10	363	1	0	0	0	0	0	0	364	1.00
	C1	0	353	7	0	0	0	4	0	364	0.97
	C2-3	0	33	330	0	0	0	1	0	364	0.91
	C4-7	0	0	0	233	130	0	1	0	364	0.64
	C5	0	0	0	20	344	0	0	0	364	0.95
	C6	0	0	0	0	0	364	0	0	364	1.00
	C8	0	3	0	0	1	0	360	0	364	0.99
	C9	0	0	0	0	0	0	0	364	364	1.00
	Precision	1.00	0.91	0.98	0.92	0.72	1.00	0.98	1.00		

Figure 3. Confusion matrix of the patch classifier

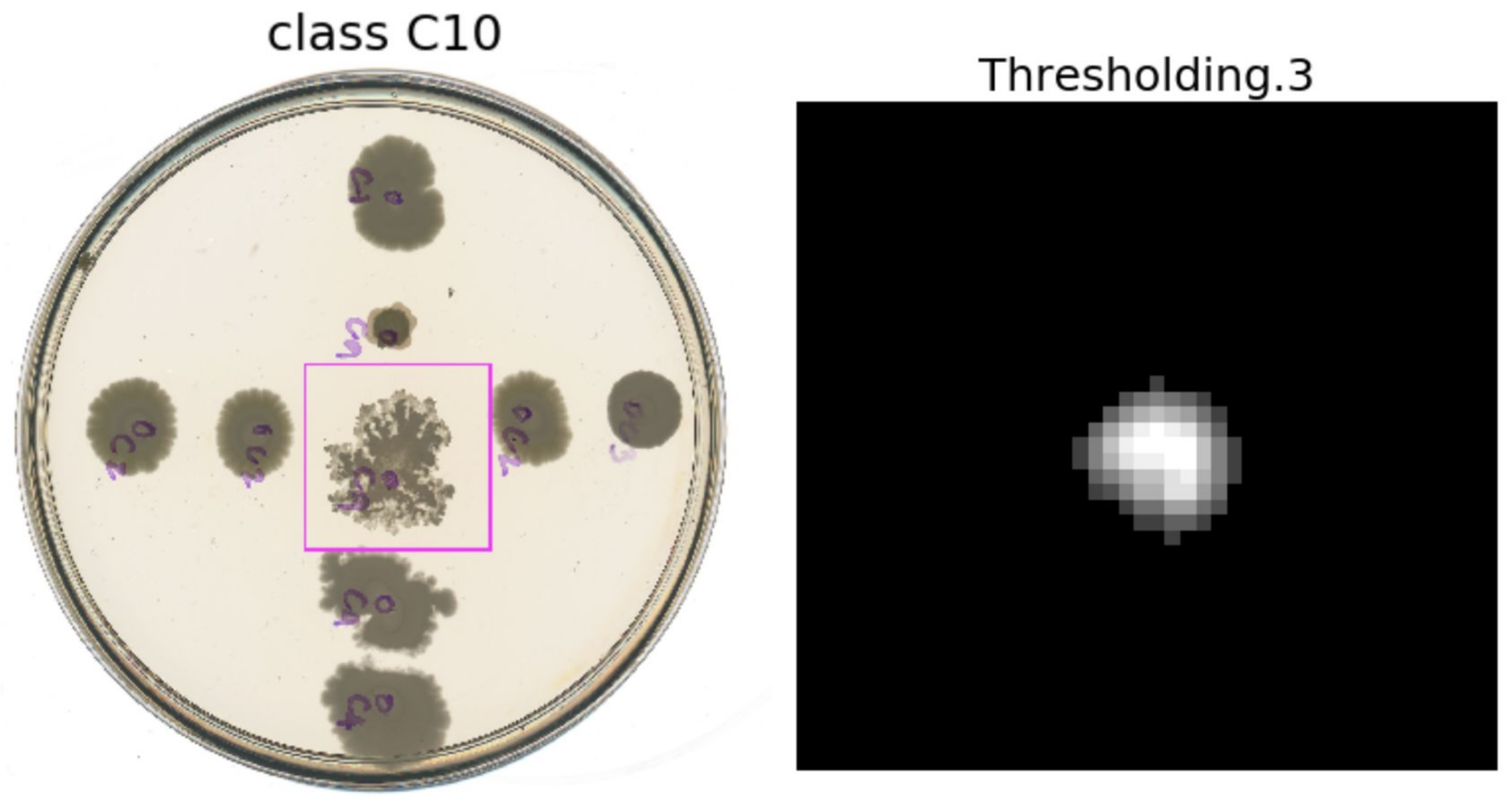


Figure 4. Successfully detected C10 from other bacterial colonies.

Conclusions and Recommendations

It is possible to classify bacterial colonies with moderate degrees of accuracy even with small sample sizes and low resolution images. Our model exploits static features from raw image pixels. However, some colonies are not easily distinguished from others when using this approach.

Recommendations:

1 Additional samples of bacteria colonies can help improve the accuracy of the model and prevent overfitting on training data.

2 In the future, dynamic time features like growing rate the growing rate of a colony can be explored.

3 The classifiers are trained on independent colony patches. In future, we can include streak data which shows the interaction features of colonies.

Acknowledgments

This research project is supported by SYNTHETIC BIOLOGICAL SYSTEMS LABORATORY at Columbia University. We sincerely appreciate the whole team's support.

References

[1] Capstone Faculty-Sponsored Project (FS#7) description, 2019.

[2] K. He, X. Zhang, S. Ren, J. Sun. Identity Mappings in Deep Residual Networks, 2016.