A Machine Learning Approach to Characterisation of Bacterial Flocculation in Wastewater Treatment

Diego Fernández, UG[1]; Prof. Cait MacPhee[2]; Gavin Melaugh, PhD[2]; Ryan Morris, PhD[2]
[1]University of Edinburgh

THE UNIVERSITY of EDINBURGH

School of Physics and Astronomy

Abstract

Over the course of the last decade, machine learning (ML) has proven an exceptional method in labelling and classifying a broad spectrum of images. We apply a novel ML approach to the characterisation of images of bacterial aggregation in wastewater treatment (WWT) via the development of a convolutional neural network (CNN). WWT is one of the most important biotechnological processes in the world. This phase of WWT, however, uses outdated biological and microscopic quality assessment methods, which have been challenged by the complex nature of the physics and biology of the system. In this project, we develop an effective prototype model that by far exceeds our pre-set specifications of 60% accuracy for proof of concept by consistently delivering >70% test accuracy. The success of this experiment will further our understanding of the science behind bacterial aggregation and may prove revolutionary in modernising global WWT process control.

Motivation and Objective

- Wastewater treatment (WWT) is crucial in maintaining the health of our environment and ourselves. The process is depicted in Fig. 1.
- Improving the characterisation of the flocculation process – the process of aggregation of bacteria and waste – in the secondary phase of WWT could increase the efficiency of the system.
- Automation of this kind reduces human error as well as saving time and energy.

OBJECTIVE: Provide proof of concept for a CNN effectively classifying flocculation images in WWT.

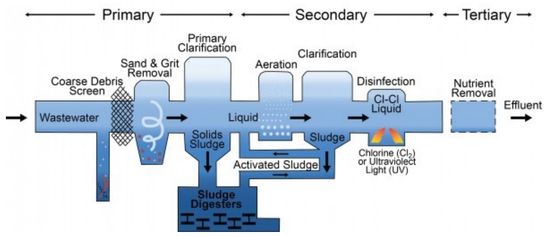


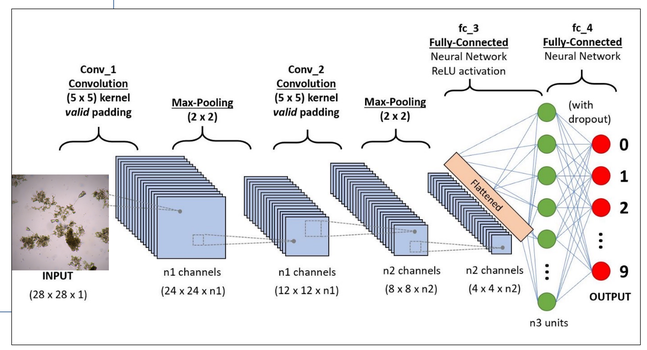
Figure 1. The wastewater treatment process [1].

How a Convolutional Neural Network (CNN) Works

- A CNN is a model that is trained to determine the most important features that classify/define an image.
- Training images are passed through a series of layers that determine feature weights (see Fig. 2).
- Accuracy is measured by predicting on a set of labelled images; this is referred to as *supervised learning*.
- Our images are labelled 'low', 'good' and 'high', representing the quality of the flocculation in an image.

Note: A lack of available 'low' images in the data set meant we had to show proof of concept using only 'high' and 'good' flocculation classifications.

Figure 2. The architecture of a CNN [2].



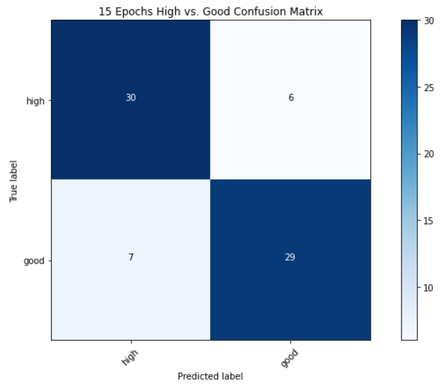


Table 1. The confusion matrix from the highest accuracy CNN developed.

Fine-tuning the CNN

- Many hyperparameters can be optimised when developing a CNN. Here, however, we only focus on determining the ideal number of epochs - the amount of time in which we train the model.
- We tested accuracy for 5, 10, 15 and 20 epochs on three different train/validate/test sets of data.
- We determined that the ideal epoch range is between 10 and 15 epochs to maximise accuracy (see Chart 1).

Methodology: Running the CNN

1. Prepare the data by splitting it into train, validate and test sets
2. Pre-process the images so they can be interpreted by the CNN (see Fig. 3)
3. Create a sequential CNN manually or using existing models such as the VGG16 model
4. Use the train and validate data to train the model
5. Determine the accuracy of the model on the test data
6. Produce a confusion matrix depicting the predicted vs. actual image labels (see Table 1) – this helps visualise the accuracy

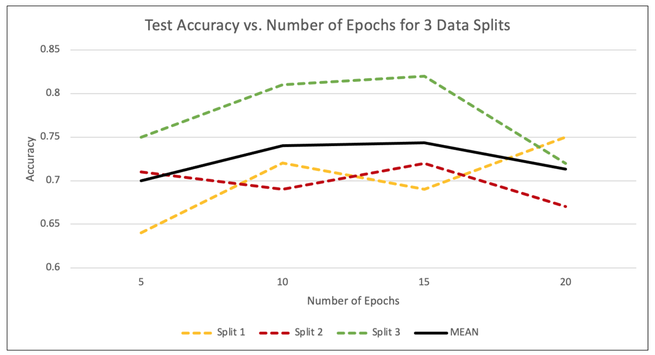


Chart 1. The model accuracy for three different data splits at 5, 10, 15 and 20 epochs. The mean is labelled in black showing an ideal number of epochs between 10 and 15.

Results

The best binary model classified with 83% accuracy with an average accuracy of 72.4%.

We determined that when using a pre-developed model known as the VGG16 model, 13 epochs is the ideal number of epochs (see Chart 2).

When using a 3-class model for 'low', 'good' and 'high' classifications, we achieved 50% accuracy, which also confirms proof of concept.

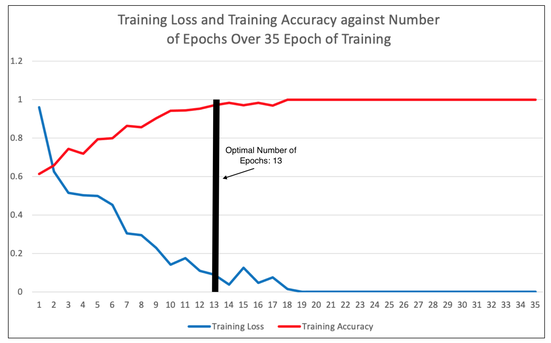


Chart 2. A plot of the training loss and accuracy against number of epochs for the VGG16 model over a period of 35 epochs. 13 epochs is deemed ideal because the loss stops decreasing and the accuracy is not so high that the model is overfit.

Conclusions and Next Steps

CONCLUSION: Proof of concept for the characterisation of flocculation in WWT via a CNN has been achieved via the successful development of prototype models.

THE FUTURE

- Larger training data sets should vastly improve the accuracy of the model.
- Future models should fine-tune different hyperparameters to maximise model accuracy.
- This CNN could be used to learn more about the biological and physical process of the system.

Contact Information

Diego Fernández
University of Edinburgh
Email: fernandezd1998@gmail.com
Phone: +44 7474047874

References

[1] "U.S. wastewater treatment factsheet," 2020. [Online]. Available: http://css.umich.edu/factsheets/us-wastewater-treatment-factsheet
[2] S. Saha, "A comprehensive guide to convolutional neural networks - the eli5 way," Dec 2018. [Online]. Available: https://towardsdatascience.com/ a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

Acknowledgements

The author would like to thank advisors Prof. Cait MacPhee, Dr. Gavin Melaugh, and Dr. Ryan Morris of the Soft Condensed Matter and Biological Physics Groups in the Institute of Condensed Matter and Complex Systems at the University of Edinburgh for their support in this Senior Honours Project. We thank Paul Banfield, technical manager at Veolia, for his provision of images from wastewater treatment sites across Scotland, without which the project could not have been completed. Finally, we would like to thank the National Biofilms Innovation Centre for their funding and the Biotechnology and Biological Sciences Research Council for their grant.