4TR3 - Capstone Midterm Report
Non-intrusive Testing of Liquid Culture Medium using Online NIR Spectroscopy and Machine Learning for Qualitative Analysis

Benjamin Samuel[a,3], Connor Reintjes[a,1,2], Paola González Pérez[a,2], Shiza Hassan[a,3], Dr. Amin Reza Rajabzadeh[a,4]

[a] *McMaster University, W. Booth School of Engineering Practice and Technology, Hamilton, ON., Canada*

Abstract

Insufficient quality assurance is a major expense associated with laboractories that an result in contamination, poor sample integrity, and lead to lost time due to repetitive sample testing. To ameliorate these issues, NIR spectroscopy has been combined with machine learning in this approach to qualitatively analyze the composition of liquid cultures in a non-intrusive and online manner. This will allow laboratories to save time by identifying contamination as soon as it happens and proceeding accordingly, as opposed to finding out after a full protocol has been performed. The method to achieve this involved creating a handheld casing to house the NIR which would take spectra of the sample and pass it to a machine learning model that would then identify whether the sample is in a normal or contaminated state. In phase 1 of the experiment, the NIR housing was built and initial testing was conducted for both contaminated and non-contaminated states, with the contaminant being Lactobacillus rhamnosus. The results achieved indicate that the NIR was able to differentiate between contaminated and non−contaminated samples. To analyze these readings a 1D−CNN model will be trained with the data collected thus far in the second phase of the experiment. Moving forward, further testing will be performed using quartz cuvettes to achieve more accurate and precise results as well as plating of bacteria to generate a standard curve to achieve quantitative results.

Keywords: Bioprocess Monitoring, Machine Learning, Near−Infared Spectroscopy (NIR), 1D Convolutional Neural Network (1D−CNN), Qualitative Analysis

[1] Conceptualization
[2] Validation & Software
[3] Investigation & Methodology
[4] Principle Investigator & Capstone Supervisor

1. Materials and Methods

1.1. NIR Housing Design

The custom housing for the NIR was 3D printed using a Bambu Lab P1S. The majority of the housing was constructed using Acrylonitrile butadiene (ABS) filament however, the cuvette holder was printed using polylactic acid (PLA) filament. The sides and bottom of the cuvette holder were printed in black PLA, while the background was printed in white PLA to increase reflectance.

This white PLA background was not as reflective as originally theorized, so the use of teflon (PTFE) was considered. A high-reflectance PTFE sheet with 3M Adhesive backing was purchased from ThorLabs. This sheet has a reflectance ($> 90\%$) in the UV spectrum, and is one factor attributed to improving the SNR within the scans.

1.2. Data Augmentation

Due to the limited number of samples that were able to be collected, the data was augmented for the sake of training. Different data augmentation methods were explored to best fit our use case. Factors such as simplicity, and reducing the risks of overfitting were considered. The use of subsampling was determined to be the optimal method and was applied to future data collection.

1.2.1. Pipeline Subsampling

The data augmentation pipeline that was created is based on a type of subsampling called cyclic subsampling. Cyclic subsampling divides the dataset into non-overlapping subsets using an offset starting index and a set interval. This method has a few key features that make it optimal for this dataset when compared to other methods like temporal subsampling with sliding windows or random sampling. Since the interval remains the same across all subsamples, the temporal order of the dataset can be maintained when training a neural network. The non-overlapping sub-samples have a lower redundancy within the datasets compared to other methods, which will help prevent overfitting. The improved diversity would also allow the model to account for variations in starting time or time accumulation from scan duration.

1.2.2. Subsampling Considerations

Using cyclic subsampling had some major considerations that changed future data collection. The subsample interval would have to be large enough to prevent redundancy, while also remaining small enough to not create a large margin of error within the model. The frequency of scans would also have to be high enough to prevent the offset time points from approaching the cycle interval. Due to these considerations, our scanning frequency was changed to 60 second intervals. Our default augmentation parameters were set at a 60 second offset with a 15 minute (900 second) cycle interval with 3 subsamples per dataset.

1.3. One-Dimensional Convolutional Neural Network

The classification of near-infrared (NIR) spectra to determine the contamination state of liquid media was performed with a one-dimensional convolutional neural network (1D-CNN). This neural network model uses the Keras library with Tensorflow as the backend. The use of Keras facilitates the modelling of the CNN due to its intuitive and user-friendly interface (Keras Team, n.d.). The Tensorflow backend allows for accelerated training when used with GPU (AceCloud Team, 2024). It also enables a detailed visualization through Tensorboard. Their combination results in a flexible model with easily customizable variables, like the layers and optimizers, along with shorter training times

1.3.1. Model Architecture

The model consists of two Conv1D layers, with 64 and 32 filters respectively. These filters allow for the extraction of significant features in the input data (Nisha & Meeral, 2020). These features set the parameters that will be learned throughout training. To avoid overfitting and reduce the dimensionality of these parameters, MazPooling1D layers and Dropout layers are implemented after the Conv1D layers. The activation function, or function that transforms the input of a node into an output of that node, is a key component of a neural network. The model uses the Rectified Linear Unit (ReLU) function as the activation function, which only outputs non-negative outputs or zeros (Brownlee, 2019). This is commonly used in CNNs due to its ease of training and superior performance. A final dense layer is implemented with a sigmoidal activation function. This function outputs a probability score between 0 and 1 to determine the contamination state of the sample: 0 for non-contaminated and 1 for contaminated (Saeed, 2021).

1.3.2. Optimizer and Loss Function

Optimizers are functions responsible for adjusting the parameters as training progresses (EITCA Academy, 2023). This is done to reduce the loss function, a mathematical quantification of the error margin between the prediction and the ground truth. The Adam algorithm was chosen as the optimizer. This follows a stochastic gradient descent based on adaptive estimation and regulated by the learning rate (Keras Team, n.d.). It allows for the model to reach an optimal point faster without taking large learning steps that could lead to skipping key information during training. The learning rate was set to 0.0005 to avoid reaching a convergence between the ground truth and prediction too early during training.

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i))\right]$$

The binary cross-entropy loss function was used in the model due to the binary classification nature of the project. This loss function is defined as where y is the label (1 for contaminated and 0 for non-contaminated) and $p(y)$ is the predicted probability of the point being contaminated for all N points.

Binary cross-entropy quantifies the differences between the logarithmic probability distributions and penalizes inaccurate predictions which helps assess the model's prediction confidence.

1.3.3. Training Parameters

Since the small size of the training dataset represents a major problem for the accuracy of the model, the number of epochs, or cycles that the entire dataset passes through the algorithm during training, was adjusted to offset this limitation. The model was trained with 15 epochs to reduce the risk of overfitting, and an early stopping callback was included to stop the training process once the model showed no further improvement after each epoch. To ensure an efficient use of the limited dataset for both training and validation, cross-validation was implemented. This allowed for both the training and validation batch size to be 18, the total dataset size.

2. Results

3. Discussion

4. Problems and Future Steps

5. Conclusion

Acknowledgments

Acknowledge any individuals or organizations that contributed to the research, funding bodies, and any supporting institutions.

References

Appendix A

Figure A.1: Ethanol Standard Curve from Quartz Cuvette.

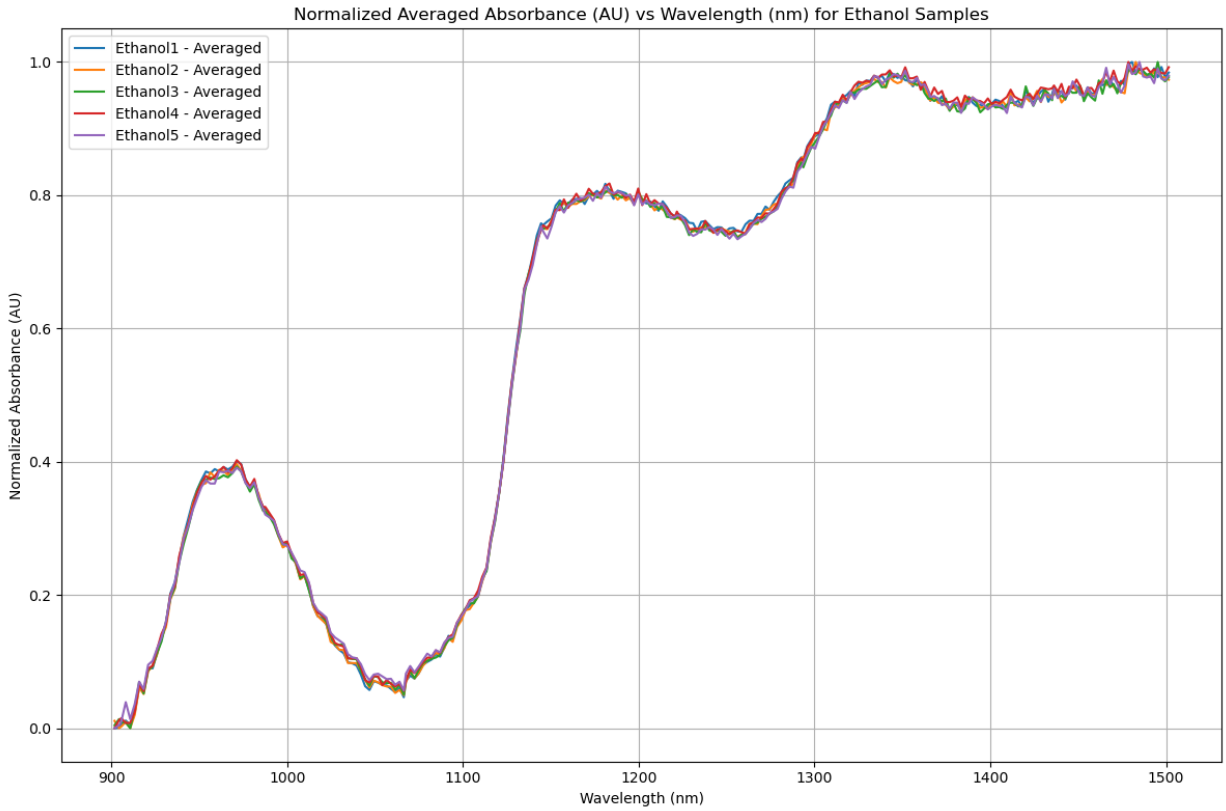

Figure A.2: Ethanol Standard Curve from Plastic Cuvette.