Non-intrusive Testing of Liquid Culture Medium using Online NIR Spectroscopy and Machine Learning for Qualitative Analysis

Connor Reintjes, Paola González Pérez, Benjamin Samuel, Shiza Hassan

Supervisor: Dr. Amin Reza Rajabzadeh

McMaster University, W. Booth School of Engineering Practice and Technology

BIOTECH 4TR3 - Capstone

Abstract

Insufficient quality assurance is a major expense associated with laboratories that can 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 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 manufactured 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. In phase 2, further data collection was conducted using a quartz cuvette and Teflon background to reduce noise and spectra interference. A data augmentation pipeline was constructed to overcome data size limitations. The processed data was then fed into a 1D-CNN model to obtain preliminary results on its performance. Implementation of one-class classification resulted in the overfitting of the model. The pipeline and 1D-CNN model will continue to be developed in order to improve their performance as more diverse data is collected.

Contents

1. Introduction

Quality assurance issues are responsible for repeated testing, misdiagnosis, improper treatment and unpredictable costs.[1] The contamination of samples is the most common problem experienced in microbial laboratories. The real-time monitoring of samples can represent a significant cost-saving solution to improve quality assurance practices in laboratory and industrial settings.[2] To improve these issues, Near Infrared (NIR) spectroscopy has been combined with deep learning to provide real-time analysis of sample quality. This online analysis increases the testing and result obtention speed while remaining non-invasive to reduce further contamination of the samples.[3] This technology also symbolizes a financial advantage by reducing the loss of biological samples using low computational requirements that can measure and classify the analyzed data as required. A deep learning model will be used to optimize this analysis as it offers advantages such as automatically integrating feature extraction, and incorporating hundreds of layers with many parameters compared to artificial neural networks.[4]

The NIR spectral region has a range from 700 nm to 2500 nm. It is primarily characterized by absorption bands corresponding to the overtone and combination bands of C−H, N−H, O−H, and S−H stretching Figure 1. Hence, the spectra of most compounds will show unique absorption peaks in the NIR spectral region. These peaks can be used for the quantitative and qualitative analysis of the compounds in gases, solids or liquids.

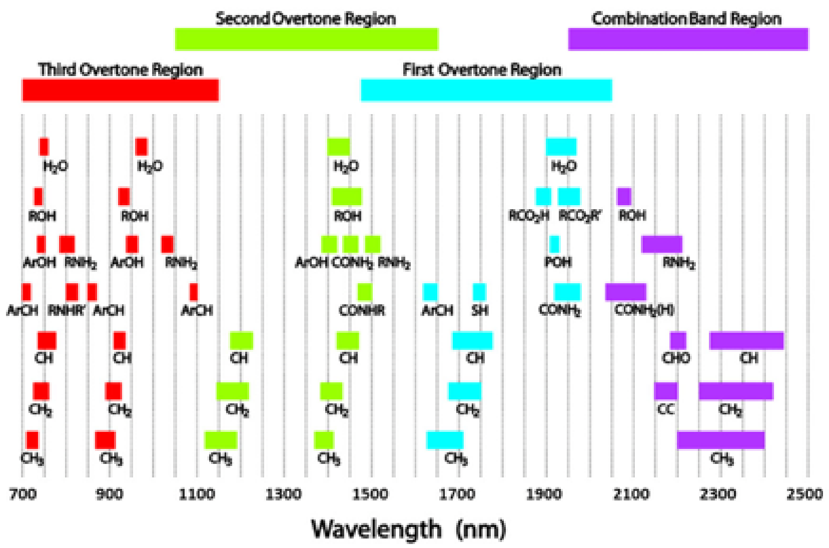


*Figure 1: Principal analytical bands and relative peak locations for absorptions in the near-infrared region.*NEEDS CITATION *Unique absorptions found in the majority of chemical and biological products can be utilized in both qualitative and quantitative examinations.*

Ethanol and water have characteristic NIR spectrum patterns that allow for their identification and quantification as seen in Figure 2. Ethanol has a unique narrow peak at 1200 nm representing the second overtone of the C−H stretching vibration.[5] Water, on the other hand, shows a flat behaviour at this wavelength due to the lack of C−H bonds.

The intensity of NIR absorption bands is $10 - 100$ times lower than that of the equivalent basic mid-IR absorption bands. This facilitates the direct analysis of strong absorbances and high light-scattering efficiency. Band overlap and penetration depth reduce in the near-infrared (NIR) spectral region, whereas the efficiency of light scattering and absorptivity improves with wavelength. As NIR spectroscopy depends on light absorption, spectral data can be acquired either in transmittance or reflectance mode. Diffuse reflectance ($\log \frac{1}{R}$) measurements are favored for opaque or light-scattering matrices whereas translucent samples used transmittance ($\log \frac{1}{R}$).

NIR spectroscopy's quickness, adaptability to a variety of materials, and capacity to examine liquid samples make it a promising tool for bio applications.[6] Recent advancements are making NIR spectroscopy more accessible with portable handheld instruments that allow for real-time, on-site examination. Cell culture media (CCM) plays a vital role, directly impacting process yield and product quality.[7] CCM's primary objective is to create and sustain an optimal physiological environment for large-scale cell culturing, ensuring cell health and expression of desired Critical Quality Attributes (CQAs). However, it's crucial to recognize that the chemical and physical properties of CCM are sensitive to microbial growth, chemical reactions, and environmental factors like temperature and light. Without sterilization, CCM can rapidly change due to microbial contamination, leading to increased turbidity and light scatter, consequently elevating baselines and noise in spectra, potentially compromising measurements. While chemometrics can mitigate some measurement errors, identifying errors induced by samples and measurements in CCM analysis can be challenging.

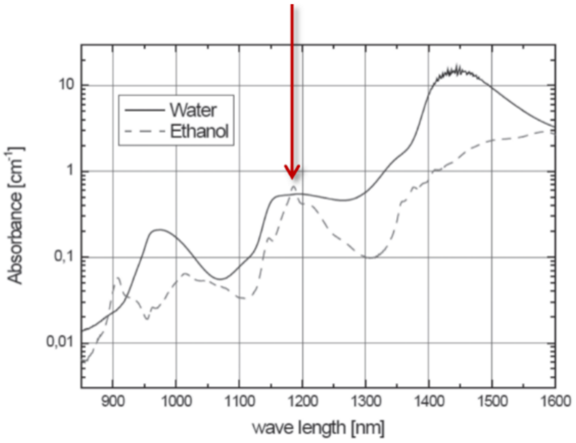


Figure 2: NIR spectrum of ethanol and water[5]

NIR spectra are known to be high-dimensional data due to the baseline drift and other noise signals present. This results in the need for preprocessing techniques for noise filtering and dimensionality reduction before analyzing the true compound signals.[8] Hence, NIR spectra usually involve the use of chemometric algorithms like partial least squares regression (PLSR) and support vector regression (SVR) to clean the data from baseline drift as well as to understand the chemical changes in the samples.[9,10] These algorithms require trained individuals to obtain the necessary parameters used to analyze the spectra.[11] Similarly, chemometric models are often limited by their inability to generalize for variance in spectra from different instruments, or changing storage and growing conditions. Deep Learning[12] Machine Learning (ML) can be used to automate extracting the main compound features from high-dimensional spectral data and eliminate the need for feature-selection techniques. While different classes of ML algorithms have successfully achieved the classification of NIR spectra,[8] traditional ML methods such as partial least squares (PLS), K-nearest neighbor (KNN), and principal component analysis (PCA) require a higher level of expertise to design suitable features of the models architecture.[13] Deep learning, on the other hand, does not require a high level of expertise as it utilizes the raw features of data to analyze and classify it as needed. This is achieved through multiple hidden layers that are trained end to end. Some of these layers can be specialized convolution layers that allow for the learning and identification of local feature patterns. This aspect enables deep learning architectures to employ less preprocessing for high-dimensional data.[14]

Convolutional neural networks (CNNs) are a class of deep neural networks commonly used for data analysis due to their successful results in image processing and classification,[15] speech recognition,[3] and other computer vision tasks. CNNs are a feed-forward multi-layer architecture in which a kernel or filter takes specific features from local regions from the upper layer. This architecture allows for an autonomous extraction of important features from complex data for analysis and learning.[9] A common drawback of the use of CNNs for spectral analysis is the requirement of large data sets during the training process of the model.[16] To avoid overfitting due to a limited number of data samples, one-dimensional CNN (1D-CNN) can be used. These models are similar to traditional CNNs except for the data size required and low computational requirements. 1D-CNNs have proven to have good information extraction and high classification accuracy with simple preprocessing techniques. It has been applied by Shang et al.[9] for the analysis and classification of NIR spectra from breast cancer tissue to aid in cancer diagnosis, demonstrating a 94.67% classification accuracy. Chai et al.[17] also developed an improved 1D-CNN structure to discriminate Anoectochilus roxburghii from its counterfeits using NIR spectra.

Previous studies performed on 1D-CNN for the analysis of NIR spectral data demonstrate the viability of its application for the online analysis of culture media. The present study aims to develop a method based on NIR spectra acquired on a portable device, as a non-destructive, online method to assess quality attributes of culture media.

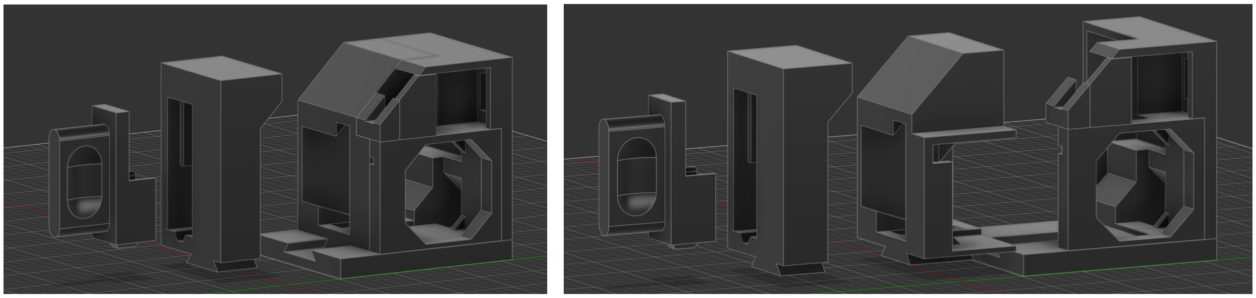
Due to the expected nature of the data collection, it was necessary to employ a time series analysis when designing the CNN structure. Within this experiment *S. cerevisiae* was used as the model, testing it under two conditions, normal and contaminated with Lactic Acid Bacteria. *Lactobacillus rhamnosus* was used as the contaminant of choice due to its improved growth in the presence of *S. cerevisiae.* Also, using this species ensures the proper growth of *S. cerevisiae* as it remains unaffected in its presence.[18] The NIR spectra of ethanol were analyzed in each case and passed to the deep learning model which classified the normal baseline against a non-normal sample.

2. Methodology

2.1. NIR Housing Design

Due to the design of the DLP NIRscan Nano (Texas Instruments) and the location of the sensor window on the unit, it required a housing to be designed. This housing was designed to both hold the NIR as well as the sample to ensure consistent scan conditions. The housing was created using a 3D printer (P1S, Bambu Labs) using Matte PLA filament (Bambu Labs) for the test models, and was reprinted in ABS filament (Bambu Labs) for the final housing.

The housing underwent multiple iterations with all design and modeling performed within AutoDesk Fusion360 parametric CAD software.[19] The final iteration of the housing used can be seen below in Figure 3. All of the components slide along a single dovetail mount along the bottom plate to allow for quick assembly while keeping all components secure. Due to slight problems with temperature regulation, a NF-A4x10 5V PWM fan from Noctua was added in addition to a fan speed controller (NA-FC1), to prevent the NIR from overheating during scans. A cuvette holder was also designed to allow for the quick change of samples, while also ensuring that the cuvette remained fixed in place for scans.



(a) NIR Housing

(b) NIR Housing Exploded

Figure 3: NIR Housing Design

Variable	Setpoint
Start wavelength (nm):	900
End wavelength (nm):	4.68
Pattern Pixel Width (nm):	1500
Exposure (ms):	1.27
Digital Resolution:	264
Scan Repeats:	9
PGA Gain:	64

Table 1: Scan Setting for DLP NIRScan Nano

The 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,[20] and is one factor attributed to improving the signal-noise-ratio (SNR) within the scans.

2.2. Ethanol Standard Curves

2.3. Yeast Culture

2.4. Lactobacillus Viability Test

2.4.1. Culture Preparation

2.4.2. Plating

2.5. Spectral Measurement

2.5.1. Spectral Processing

Hadamard Transform is one of the available protocols built into the DLP NIRscan Nano. Unlike the column protocol which only reads one wavelength at a time, the Hadamard protocol takes a multiplex scan of multiple wavelengths simultaneously. An algorithm is then used to decode the individual wavelengths from the multiplex scan. The benefit of using Hadamard over Column is that Hadamard has a much better signal-noise ratio which will minimize the required preprocessing later.

2.6. Data Pipeline

2.7. 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.[21] The Tensorflow backend allows for accelerated training when used with GPU.[22] 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

2.7.1. Model Architecture

The architecture of the model employs five different layers repeated at different points, as shown in figure x. 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.[23] These features set the parameters that will be learned throughout training. To avoid overfitting and reduce the dimensionality of these parameters, MaxPooling1D 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.[24] 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.[25]

Figure x. 1D-CNN architecture diagram

2.7.2. Optimizer and Loss Function

Optimizers are functions responsible for adjusting the parameters as training progresses.[26] 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.[21] 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.

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

The binary cross-entropy loss function was used in the model due to the binary classification nature of the project. The loss function is defined in Equation 1, 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 models prediction confidence.

References

1 Randox Laboratories. *The Cost of Poor Quality in the Clinical Laboratory.Pdf.* URL: https://www.labmedica.com/whitepapers/The%20Cost%20of%20Poor%20Quality%20in%20the%20%20%20Clinical%20Laboratory.pdf.

2 Endeshaw Abatenh, Birhanu Gizaw, and Zerihun Tsegaye. “Contamination in a Microbiological Laboratory”. *International Journal of Research Studies in Biosciences* 6.4 (2018). DOI: 10.20431/2349-0365.0604002.

3 A. Alsobhani, H. M. A. ALabboodi, and H. Mahdi. “Speech Recognition Using Convolution Deep Neural Networks”. *Journal of Physics: Conference Series* 1973.1 (2021), p. 012166. DOI: 10.1088/1742-6596/1973/1/012166.

4 P. Mishra, D. Passos, F. Marini, J. Xu, J. M. Amigo, A. A. Gowen, J. J. Jansen, A. Biancolillo, J. M. Roger, D. N. Rutledge, and A. Nordon. “Deep Learning for Near-Infrared Spectral Data Modelling: Hypes and Benefits”. *TrAC Trends in Analytical Chemistry* 157 (1, 2022), p. 116804. DOI: 10.1016/j.trac.2022.116804.

5 *Near Infrared Spectroscopy (NIR) | Anton Paar Wiki.* Anton Paar. URL: https://wiki.anton-paar.com/en/near-infrared-spectroscopy-nir/.

6 K. B. Be, J. Grabska, and C. W. Huck. “Near-Infrared Spectroscopy in Bio-Applications”. *Molecules* 25.12 (12 2020), p. 2948. DOI: 10.3390/molecules25122948.

7 A. G. Ryder. “Cell Culture Media Analysis Using Rapid Spectroscopic Methods”. *Current Opinion in Chemical Engineering.* Biotechnology and Bioprocess Engineering 22 (1, 2018), pp. 11–17. DOI: 10.1016/j.coche.2018.08.008.

8 B. A. Krohling and R. A. Krohling. *1D Convolutional Neural Networks and Machine Learning Algorithms for Spectral Data Classification with a Case Study for Covid-19.* 24, 2023. DOI: 10.48550/arXiv.2301.10746. arXiv: 2301.10746 cs. URL: http://arxiv.org/abs/2301.10746. Pre-published.

9 H. Shang, L. Shang, J. Wu, Z. Xu, S. Zhou, Z. Wang, H. Wang, and J. Yin. “NIR Spectroscopy Combined with 1D-convolutional Neural Network for Breast Cancerization Analysis and Diagnosis”. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 287 (15, 2023), p. 121990. DOI: 10.1016/j.saa.2022.121990.

10 I. A. Naguib, E. A. Abdelaleem, M. E. Draz, and H. E. Zaazaa. “Linear Support Vector Regression and Partial Least Squares Chemometric Models for Determination of Hydrochlorothiazide and Benazepril Hydrochloride in Presence of Related Impurities: A Comparative Study”. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 130 (15, 2014), pp. 350–356. DOI: 10.1016/j.saa.2014.04.024.

11 W. Wang, H. Jiang, G. Liu, Q. Chen, C. Mei, K. Li, and Y. Huang. “Quantitative Analysis of Yeast Growth Process Based on FT-NIR Spectroscopy Integrated with Gaussian Mixture Regression”. *RSC Advances* 7.40 (5, 2017), pp. 24988–24994. DOI: 10.1039/C7RA02774E.

12 J. Walsh, A. Neupane, A. Koirala, M. Li, and N. Anderson. “Review: The Evolution of Chemometrics Coupled with near Infrared Spectroscopy for Fruit Quality Evaluation. II. The Rise of Convolutional Neural Networks”. *Journal of Near Infrared Spectroscopy* 31.3 (1, 2023), pp. 109–125. DOI: 10.1177/09670335231173140.

13 W. Zhang, L. C. Kasun, Q. J. Wang, Y. Zheng, and Z. Lin. “A Review of Machine Learning for Near-Infrared Spectroscopy”. *Sensors* 22.24 (24 2022), p. 9764. DOI: 10.3390/s22249764.

14 W. Yue, J. Yiming, and L. Julong. “A Fast Deep Learning Method for Network Intrusion Detection Without Manual Feature Extraction”. *Journal of Physics: Conference Series* 1738.1 (2021), p. 012127. DOI: 10.1088/1742-6596/1738/1/012127.

15 N. Sharma, V. Jain, and A. Mishra. “An Analysis Of Convolutional Neural Networks For Image Classification”. *Procedia Computer Science*. International Conference on Computational Intelligence and Data Science 132 (1, 2018), pp. 377–384. DOI: 10.1016/j.procs.2018.05.198.

16 G. Zhang, X. Tuo, S. Zhai, X. Zhu, L. Luo, and X. Zeng. “Near-Infrared Spectral Characteristic Extraction and Qualitative Analysis Method for Complex Multi-Component Mixtures Based on TRPCA-SVM”. *Sensors* 22.4 (20, 2022), p. 1654. DOI: 10.3390/s22041654.

17 Q. Chai, J. Zeng, D. Lin, X. Li, J. Huang, and W. Wang. “Improved 1D Convolutional Neural Network Adapted to Near-Infrared Spectroscopy for Rapid Discrimination of *Anoectochilus Roxburghii* and Its Counterfeits”. *Journal of Pharmaceutical and Biomedical Analysis* 199 (30, 2021), p. 114035. DOI: 10.1016/j.jpba.2021.114035.

18 S. Nenciarini, S. Renzi, M. di Paola, N. Meriggi, and D. Cavalieri. “The YeastHuman Coevolution: Fungal Transition from Passengers, Colonizers, and Invaders”. *WIREs Mechanisms of Disease* 16.3 (2024), e1639. DOI: 10.1002/wsbm.1639.

19 *Fusion 360*. Version 2.0.20508. Autodesk.

20 ThorLabs. *High-Reflectance PTFE Sheets*. URL: https://www.thorlabs.com.

21 K. Team. *Keras: Deep Learning for Humans*. Keras.io.

22 A. Team. *TensorFlow GPU: Basic Operations & Multi-GPU Setup [2024 Guide]*. AceCloud, 2024.

23 S. S. Nisha and M. N. Meeral. “Applications of Deep Learning in Biomedical Engineering”. *Elsevier eBooks* (2020), pp. 245–270. DOI: 10.1016/b978-0-12-823014-5.00008-9.

24 J. Brownlee. *A Gentle Introduction to the Rectified Linear Unit (ReLU) - Machinelearningmastery.Com*. MachineLearningMastery.com, 2019.

25 M. Saeed. *A Gentle Introduction to Sigmoid Function - Machinelearningmastery.Com*. MachineLearningMastery.com, 2021.

26 E. I. T. C. A. Academy. *What Is the Purpose of the Optimizer and Loss Function in Training a Convolutional Neural Network (CNN)? - EITCA Academy*. EITCA Academy, 2023.