Machine Learning Model for Metabolite Profiling and Quantification in Complex Mixtures from NMR Data

Rahul Madhav M
School of Biological Sciences
National Institute of Science Education and Research,Bhubaneswar, HBNI
rahulmadhav.m@niser.ac.in

Rabmit Das
School of Biological Sciences
National Institute of Science Education and Research,Bhubaneswar, HBNI
rabmit.das@niser.ac.in

Abstract

Urine metabolomics, analyzing small molecules in urine, holds promise for disease biomarker discovery. This study explores a machine learning-based approach to streamline data processing and biomarker identification from human urine NMR spectra. We implemented linear regression, ensemble learning, and deep learning techniques using simulated spectra generated from the Human Metabolome Database (HMDB). Our findings demonstrate the potential of this approach, with a Mean Squared Error (MSE) of 5.06×10^{-16} achieved using linear regression. Further validation with real NMR data is crucial to translate this approach into improved disease diagnosis and personalized medicine.

1 Introduction

Metabolomics of bio-fluids offer deep insights into the physical, chemical and biological processes that interact with the corresponding fluid. Urine as a bio-fluid has been abundantly used for metabolomic study in diverse aspects. Urine's sterile nature, abundant availability in large volumes, minimal interference from proteins or lipids, and chemical complexity make it highly favored for this purpose [1]. The multifaceted chemical composition of urine has rendered it a notably intricate substrate to decipher comprehensively. Serving as a biological waste material, urine typically encompasses a diverse array of metabolic breakdown products originating from an assortment of sources including foods, beverages, medications, environmental contaminants, endogenous waste metabolites, and bacterial by-products. A detailed metabolomic study can also reveal the anomalies in physiological processes of a human from whom the sample has been collected. By analyzing the unique fingerprint of small molecules in urine, we can uncover biomarkers for various diseases, including cancer.This paves the way for:

- **Early detection**: Catching diseases at their earliest stages for better treatment outcomes.
- **Personalized medicine**: Tailoring treatment strategies based on an individual's unique biochemical profile.
- **Improved diagnosis**: Developing more accurate and non-invasive diagnostic tools.

On account of chemical complexity of urine, Nuclear Magnetic Resonance (NMR) Spectroscopy becomes an ideal tool to probe the metabolomic profile. It offers high-resolution, non-destructive analysis and superior metabolite identification. NMR can even distinguish between isomers of

Preprint. Under review.

a the same molecule making it a powerful tool in bio-medical marking and treatment [2]. NMR not-only gives an insight to the different chemical species present in the sample, but also gives quantitative results for the chemical abundances, which can be obtained by comparing the amplitudes of the corresponding chemical species. Thus we have an ideal method that can give quantitative and qualitative analysis in the same go. It is used widely for biomarker characterisation of various diseases, and thus needs a speedy and accurate prediction method.

The bottleneck to this issue arises mainly from the hindrance provided by classical methods. The usual NMR has a huge volume of data, as the fluid under consideration, has high level of chemical complexity. For proper quantification and qualitative analysis, we need to go through a series of pre-processing steps which get computationally heavy for our case. For the high volume of data, it is difficult for to provide speedy reports for diagnosis, and thus diagnosis gets delayed, causing a serious hindrance to medical fields.
We therefore propose a Machine Learning based prediction system, that predicts both the chemical species and what chemical species it corresponds to, given an NMR spectra for human urine. We specify human urine, because there are certain constraints that human physiology imposes on bio-fluids derived from the corresponding source.

2 Methodologies

2.1 Data Generation and Acquisition

We went through different sources and selected 31 metabolites to study. Then, we extracted raw NMR data for each metabolite from the Human Metabolome Database(HMDB) and the concentration range of each metabolite present in human urine from Bouatra et al. (2013). We generate the NMR spectrum for each metabolite using the Lorentzian function and this raw data(Chemical shift and intensity). We combined these spectra in proportion to their concentration randomly chosen from the available range. After combining these spectra, we added Gaussian noise to replicate real data. This dataset was then split into training, validation and test datasets for training and evaluation of models.

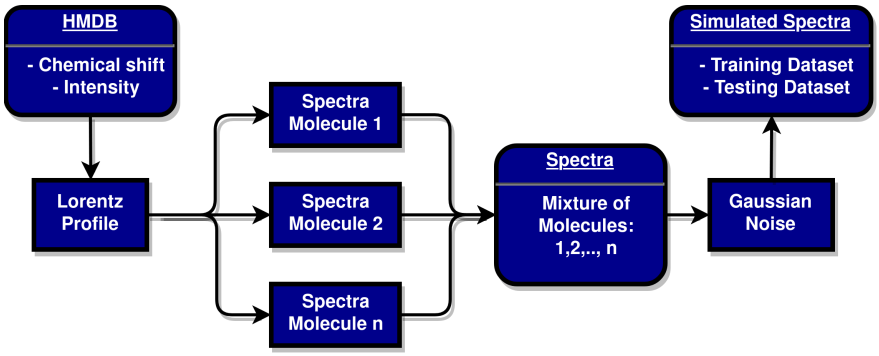

Figure 1: Data generation pipe-line

2.2 Machine Learning Models

2.2.1 Baseline Model: Linear Regression

As our foundation, we will employ Linear Regression, a well-understood statistical method adept at modeling linear relationships between independent and dependent variables.

In the context of metabolite prediction, Linear Regression will be tasked with learning the connection between the raw features extracted from the NMR spectra (independent variables) and the corresponding concentrations of specific metabolites (dependent variables).

2.2.2 Dimensionality Reduction and Linear Regression

- **Principal Component Analysis (PCA):** This technique will be used to reduce the dimensionality of the NMR spectral data while preserving the most significant variations. This can be beneficial for improving computational efficiency and potentially reducing noise in the data.
- **t-Distributed Stochastic Neighbor Embedding (t-SNE):** We will also explore t-SNE as an alternative dimensionality reduction method. Unlike PCA, which focuses on preserving global variance, t-SNE aims to maintain the local structure of the data points. This might be useful for visualizing relationships between samples in lower dimensions.

Following PCA and t-SNE, we will employ Linear Regression to predict metabolite concentrations from the reduced-dimensionality data. By comparing the prediction accuracy of Linear Regression with and without PCA or t-SNE, we can assess the effectiveness of dimensionality reduction in this context.

2.2.3 Future Exploration: Complex Models

While Linear Regression provides a strong starting point, we anticipate the need for more intricate models to potentially achieve even higher accuracy, especially when encountering unseen data or the inherent complexities of real-world urine samples. Here, we outline the complex models we plan to investigate in future work:

1. **Convolutional Neural Networks (CNNs):** These models excel at extracting features from intricate data. In our case, CNNs will be used to analyze the NMR spectra as input.
2. **Recurrent Neural Networks (RNNs):** RNNs are powerful for modeling complex relationships in data. We will explore their potential for analyzing NMR spectra.
3. **Convolutional Recurrent Neural Networks (CRNNs):** Combining the strengths of CNNs and RNNs, CRNNs offer a potentially powerful approach for analyzing complex data like NMR spectra.

3 Results

We successfully generated a large dataset (2,000,000 data points) of simulated urine NMR spectra containing 31 metabolites. This synthetic data serves both as a foundation for evaluating machine learning models as well as their training in metabolite prediction from urine NMR spectroscopy.

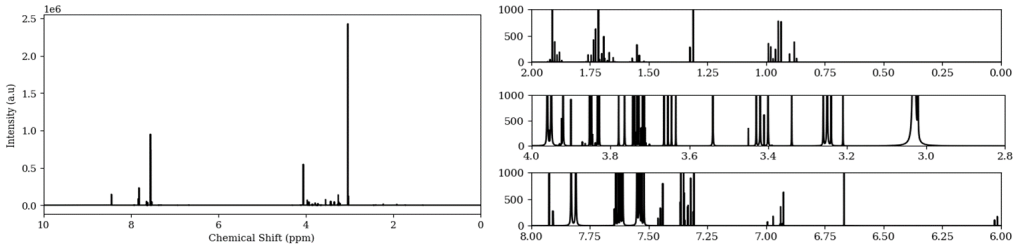

Figure 2: A sample genrated spectrum; On the left side we have the entire spectrum, On the right hand side, we have inspected the spectra at a zoomed in scale in three different ranges of Chemical Shifts.

Linear Regression, our baseline model, achieved exceptional performance with a Mean Squared Error (MSE) of less than 10^{-15} (Validation: 4.6×10^{-16} Test: 5.0×10^{-16}). This suggests strong potential for accurate metabolite prediction using raw data.

Dimensionality reduction techniques were further investigated, which involved PCA and t-SNE as discussed.

Principal Component Analysis (PCA) with two components resulted in a slight increase in MSE (Validation: 29,355; Test: 29,285). t-SNE transformation led to a more significant rise in MSE (Validation: 6,478,167; Test: 6,471,003).

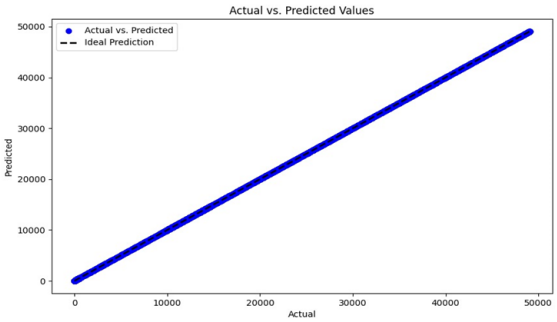

Figure 3: Linear Regression results as obtained after training of the base model.

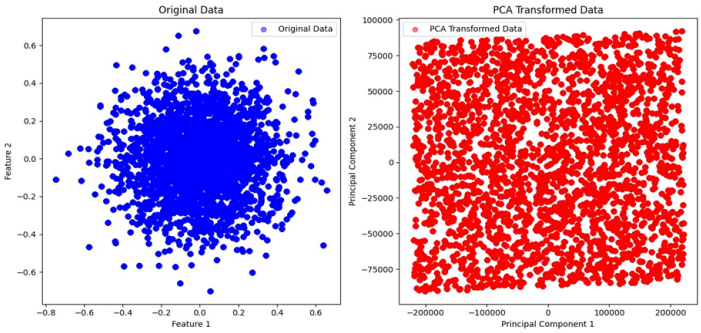

Figure 4: Original data distribution and data distribution after doing PCA with two components

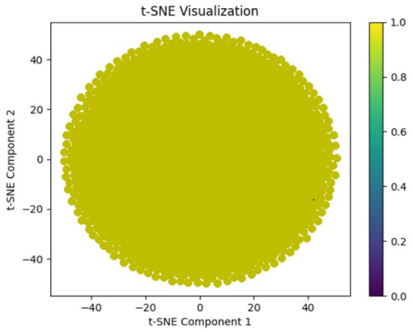

Figure 5: t-SNE data visualization

4 Inferences

Linear Regression's exceptional performance (MSE < 10^{-15}) suggests strong potential for accurate metabolite prediction directly from raw NMR data. This highlights its effectiveness as a baseline model in this context.

The impact of dimensionality reduction techniques varied. PCA with two components resulted in a slight increase in MSE, while t-SNE caused a more significant rise. This implies that minimal data manipulation might be optimal for Linear Regression in this case. However, further investigation into dimensionality reduction techniques for more complex models is warranted.

5 Future Works

Building on these findings, we will explore the capabilities of more complex machine learning models, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Convolutional Recurrent Neural Networks (CRNNs). These models hold promise for potentially surpassing the current accuracy of Linear Regression, especially when encountering a wider range of scenarios:

1. Unseen metabolite concentrations: We will test the models' ability to predict metabolite concentrations not present in the training data.
2. Novel molecules: Previously unencountered molecules will be included in the test data to evaluate the models' prediction accuracy for known metabolites.
3. Zero-concentration metabolites: The models' performance will be assessed in scenarios where specific metabolites are entirely absent from the sample.
4. Real urine NMR data: The models' generalizability will be evaluated on real-world urine samples, which can be more complex and variable than simulated data.

The performance of Linear Regression will be compared in each of these scenarios. By investigating these scenarios with all models, we aim to develop robust and generalizable machine learning methods for metabolite prediction from urine NMR spectroscopy. This could have significant implications for disease diagnosis, treatment monitoring, and personalized medicine

References

[1] S. Bouatra, C. Stewart, D. Vigneron, A. Tasse, M. Amit, J. V. Li, ... & D. S. Wishart, "The human urine metabolome," PLoS ONE, vol. 8, no. 9, e73076, Sep. 2013.

[2] C. Corsaro, S. Vasi, F. Neri, A. M. Mezzasalma, G. Neri, & E. Fazio, "NMR in metabolomics: From conventional statistics to machine learning and neural network approaches," Applied Sciences, vol. 12, no. 6, Art. no. 6, Jan. 2022.

[3] W. Wang, L.-H. Ma, M. Maletic-Savatic, & Z. Liu, "NMRQNet: A deep learning approach for automatic identification and quantification of metabolites using Nuclear Magnetic Resonance (NMR) in human plasma samples," bioRxiv, p. 2023.03.01.530642, Mar. 02, 2023.