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

Rabmit Das, Rahul Madhav M, Aniket Nath, Subhankar Mishra, Arindam Ghosh
School of Computer Sciences, NISER Bhubaneswar

## 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 x $10^{-16}$ achieved using linear regression. Further validation with real NMR data is crucial to translate this approach into improved disease diagnosis and personalized medicine.

## Introduction

Urine metabolomics offers a window into our body's inner workings. By analyzing the unique fingerprint of small molecules in urine, we can uncover biomarkers for various diseases, including cancer. This paves the way for:

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

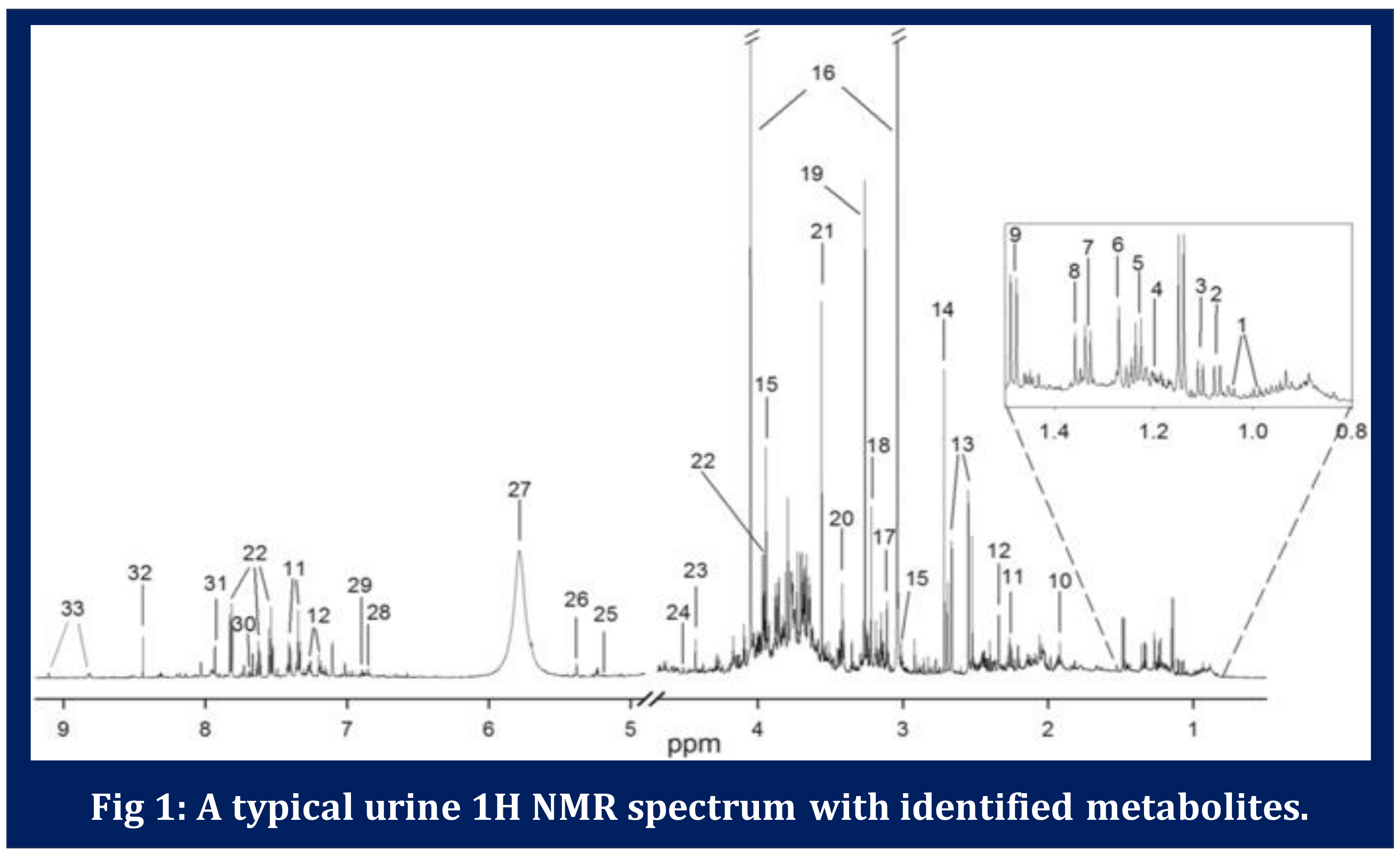
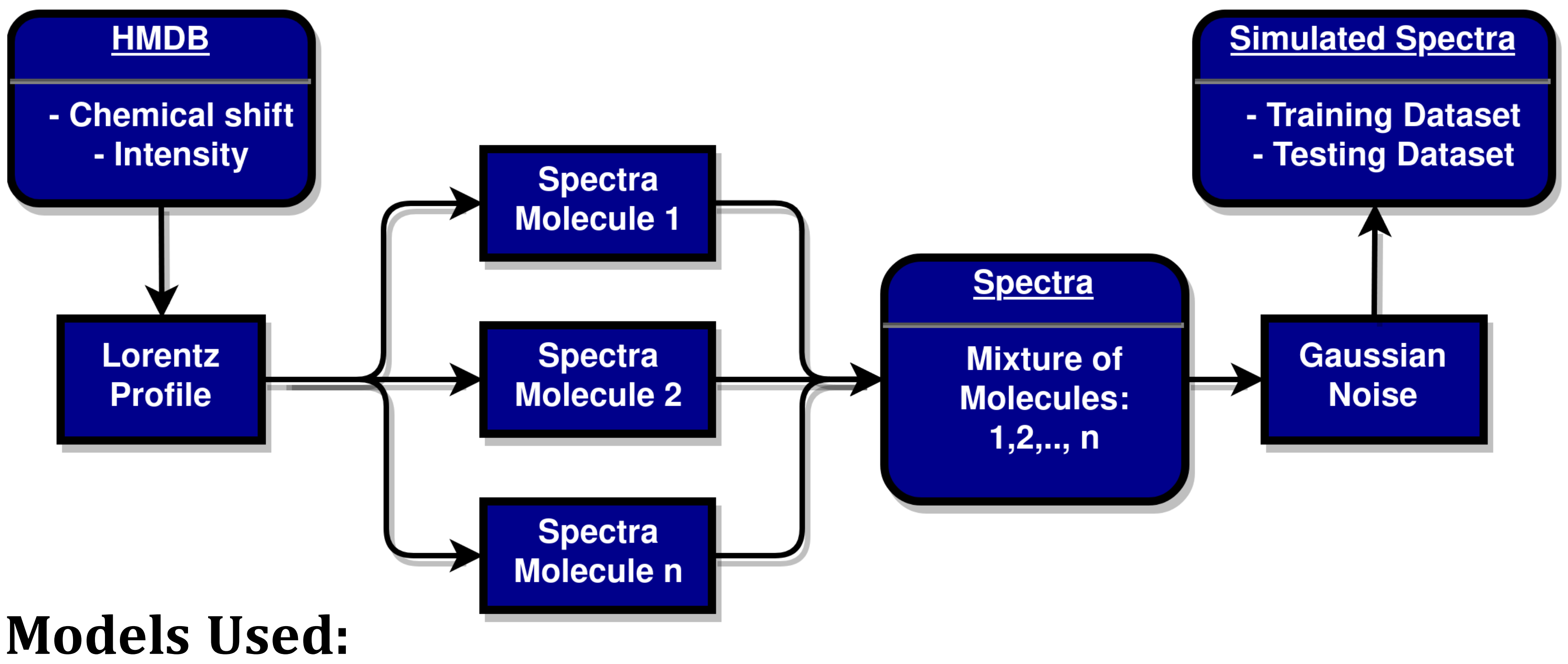


**Fig 1: A typical urine 1H NMR spectrum with identified metabolites.**

NMR, a powerful tool for metabolomics, offers high-resolution, non-destructive analysis and superior metabolite identification. However, the massive data volume hinders traditional analysis. We propose machine learning to streamline processing, identify key disease markers, and develop predictive models for diagnosis, unlocking the full potential of NMR in metabolomics.

## Methodologies

**Synthetic Data Generation Pipeline:**


HMDB
- Chemical shift
- Intensity

Lorentz Profile

Spectra Molecule 1

Spectra Molecule 2

Spectra Molecule n

Spectra
Mixture of Molecules: 1,2,.., n

Gaussian Noise

Simulated Spectra
- Training Dataset
- Testing Dataset

**Models Used:**

1. Baseline Model: Linear regression to establish reference.
2. Dimensionality Reduction: Principal Component Analysis (PCA) to explore data structure. Visualization: t-distributed Stochastic Neighbour Embedding (t-SNE) for data visualization.
3. Deep Learning Models: Investigated CNN and RNN architectures for complex pattern recognition.

## Results

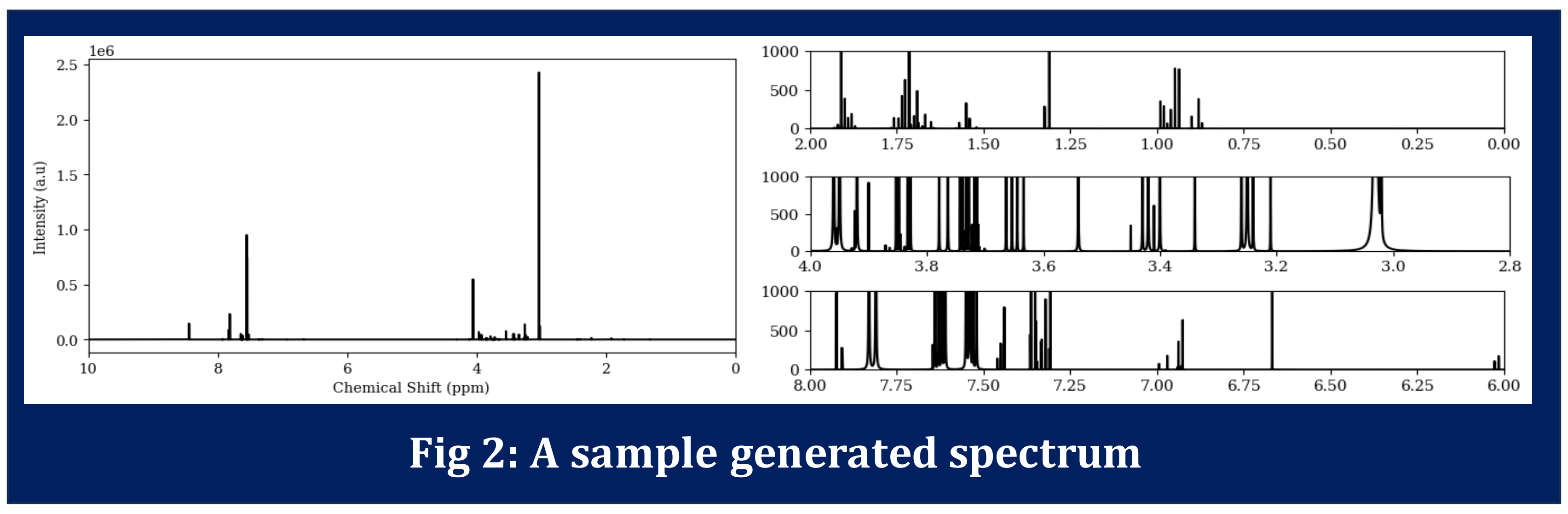


**Fig 2: A sample generated spectrum**

- ❑ Linear Regression: Excellent performance, MSE < $10^{-15}$ (Validation:4.6 x $10^{-16}$; Test: 5.0 x $10^{-16}$)
- ❑ Dimensionality Reduction: PCA (2 components) slightly increases MSE (Validation: 29,355; Test: 29,285), t-SNE more significant (Validation: 6,478,167; Test: 6,471,003)

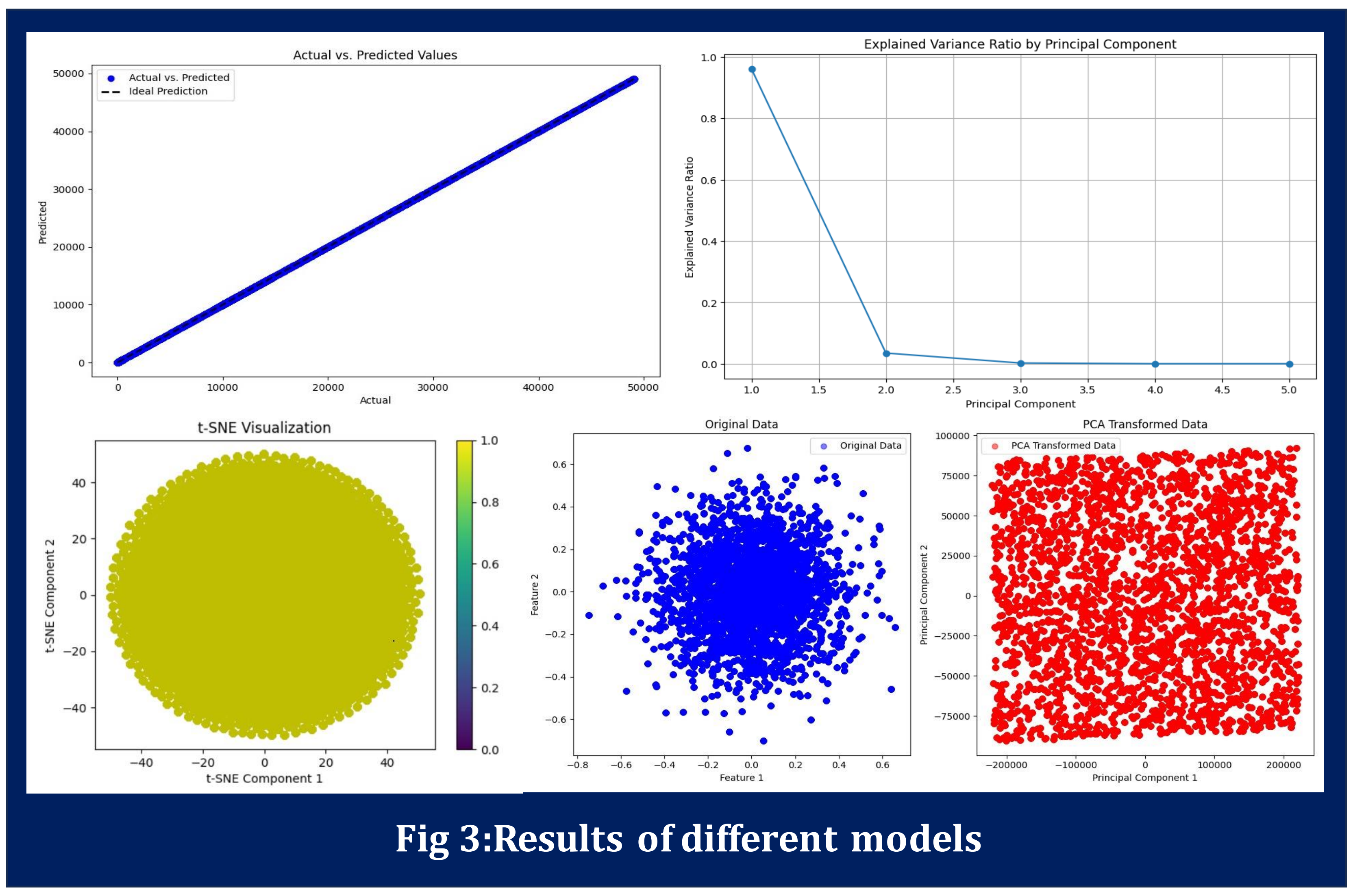


**Fig 3:Results of different models**

## Future Directions

Investigate more complex models (CNN, RNN, CRNN) & use unseen metabolite data for testing the model (adding other metabolites, unseen metabolite concentration., real urine NMR data).

## References

1. S. Bouatra *et al.*, "The Human Urine Metabolome," *PLOS ONE*, vol. 8, no. 9, p. e73076, Sep. 2013, doi: 10.1371/journal.pone.0073076.
2. C. Corsaro, S. Vasi, F. Neri, A. M. Mezzasalma, G. Neri, and 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, doi: 10.3390/app12062824.
3. W. Wang, L.-H. Ma, M. Maletic-Savatic, and 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. doi: 10.1101/2023.03.01.530642.

## Acknowledgment

1. Dr. Subhankar Mishra (Instructor), Aniket Nath (TA) for guidance and support.
2. Dr. Arindam Gosh & Pratiman De for dataset creation.
3. Fellow students for valuable feedback.
4. NISER for resources and DAE for scholarship.

CS460: Introduction to Machine Learning