Computational Biology & Bioinformatics Lab.

Project Report

Cytotoxicity of Nanoparticles

Applicant: **Tan-Thinh Duong**

Table of contents

Introduction

- **Cytotoxicity of Nanoparticles** refers to the capacity of materials between 1 and 100 nanometers in size to cause cellular damage or death due to their interaction with biological entities at a molecular level.
- **Tasks:**
    - **Regression:** predicting continuous values representing cell viability.
    - **Classification:** categorizing cell viability as either high (viability ≥ 85%) or low.
- **Outcomes:**
    1. Optimized data pipelines and hyperparameters for TabPFN-v2.
    2. Analysis on the optimization process.

Data - Properties

Table 1. Visualization of Target variable of Regression Task

	Training set		Test set	
	Regression	Classification	Regression	Classification
Features	20			
Examples	1775		762	
Missing-value examples	1102		475	
High-viability examples	-	910	-	
Low-viability examples	-	865	-	

Data - Properties

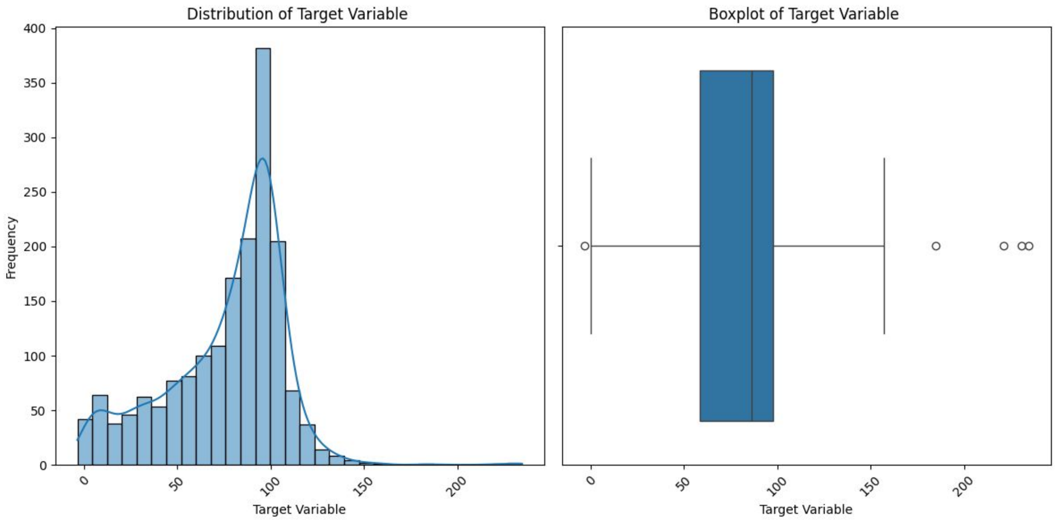

Fig 1. Visualization of Target variable of Regression Task

Data - Processing

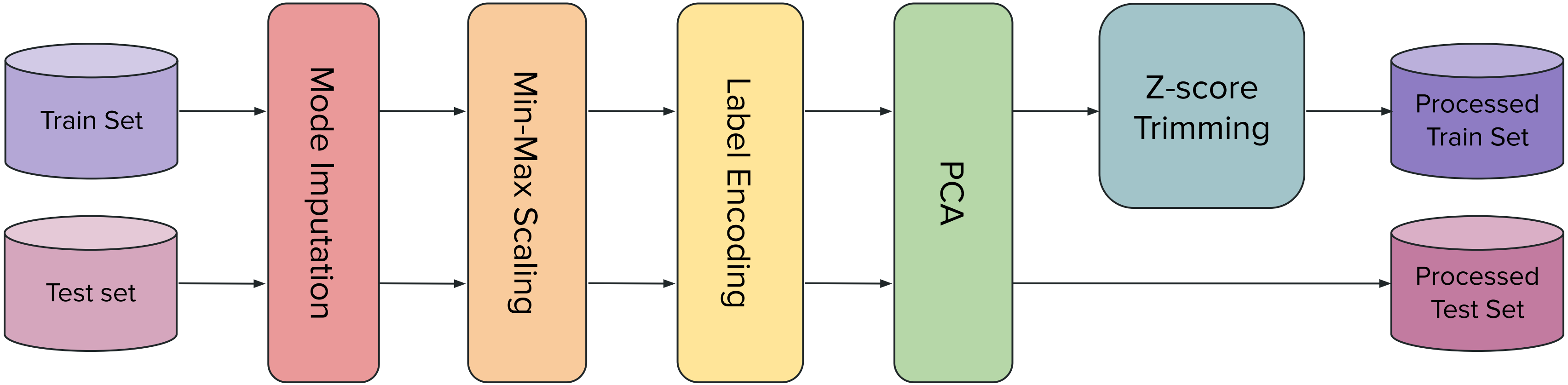

Fig 2. Optimized Data Pipeline for Regression Task

Data - Processing

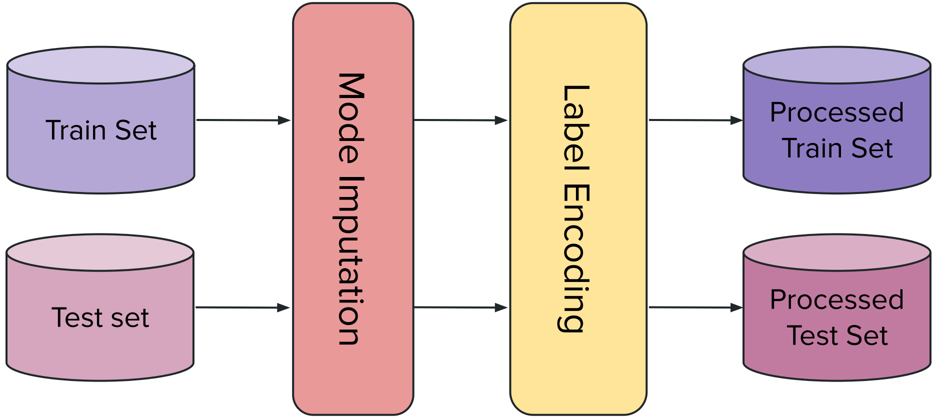

Fig 3. Optimized Data Pipeline for Classification Task

Architecture

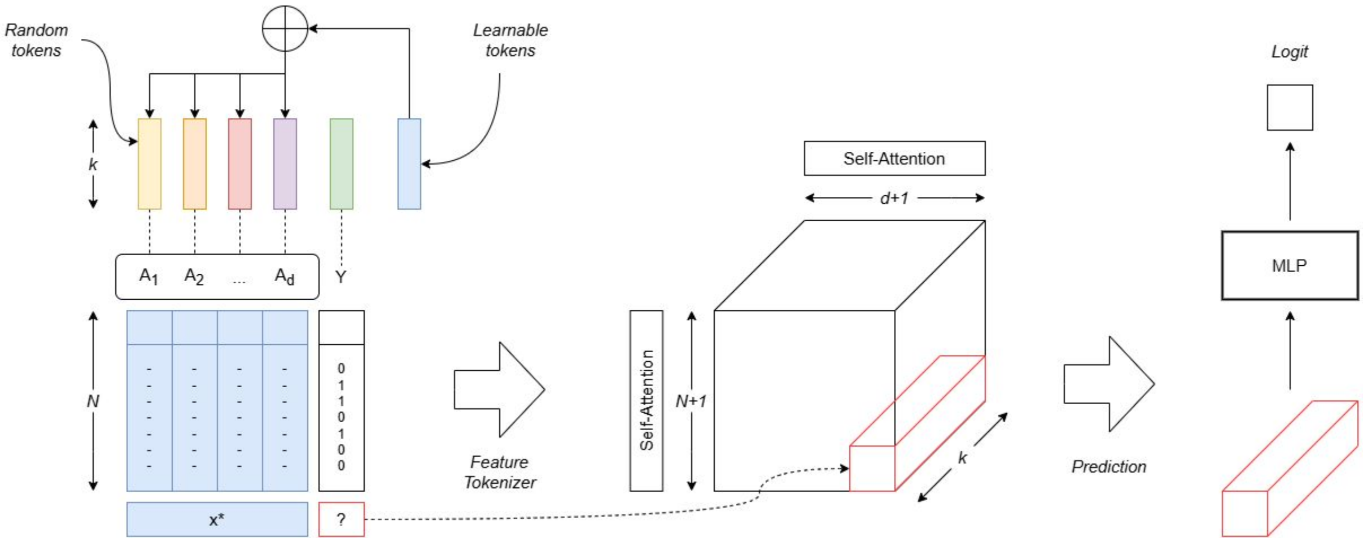

Fig 4. Architecture of TabPFN-v2 Regressor

Architecture

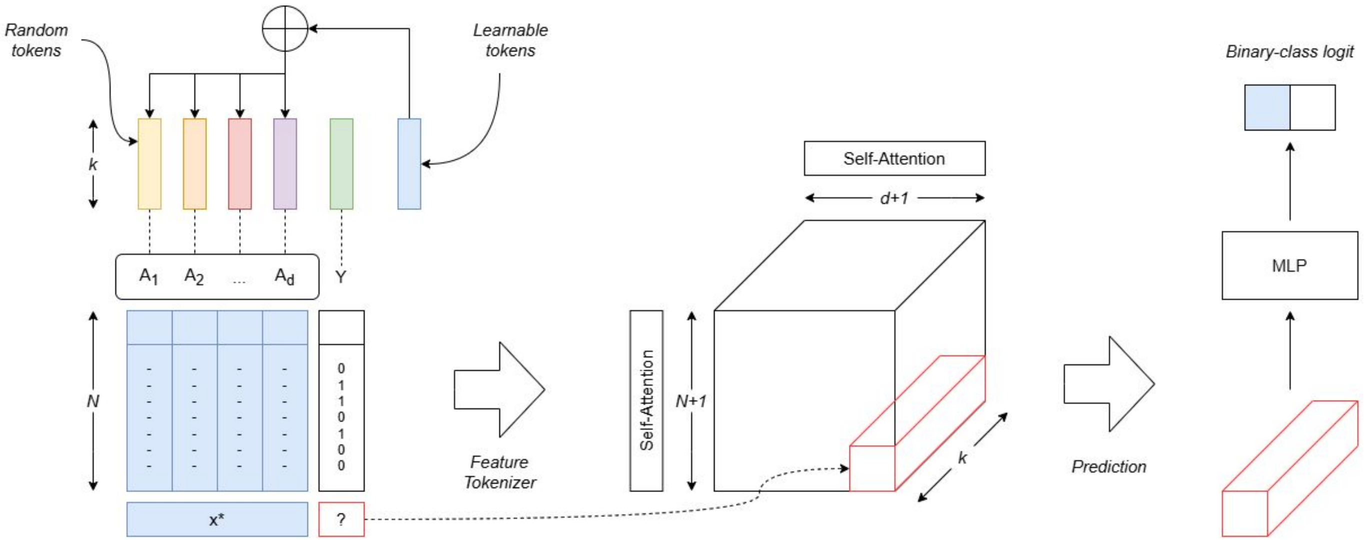

Fig 5. Architecture of TabPFN-v2 Classifier

Implementation Details

- All experiments were conducted on a **P100 16GB GPU** hosted on Kaggle and a local **NVIDIA GTX1650Ti 4GB GPU**
- The random seed consistently set to **42** to ensure reproducibility.
- The models included in the comparative analysis can be broadly categorized into two types: **data-driven** and **knowledge-driven**.
- Baseline models and ablation studies used **default hyperparameters** and a **simple data pipeline** with only a label encoding step, while the optimized ones utilized the **optimized hyperparameters and data pipelines**.

Table 2. Optimized Hyperparameters for TabPFN-v2

Hyperparameter	Regressor	Classifier
Number of estimators	10	10
Softmax temperature	0.5	0.5
Average before softmax	True	True

Table 3. Training hyperparameters for LLMs

Hyperparameter	GPT2	BioGPT
Training batch size	72	45
Number of epochs	40	
Warm-up ratio (%)	10	
Learning rate	1e-4	
Weight decay	1	
BF-16	True	

Implementation Details

Table 4. Examples of Input Text for LLMs

Training	Inference
Predict the viability of a cell. The material of the nanoparticle is Pt. The nanoparticle is inorganic. The morphology of the nanoparticle is Sphere. The fabrication method is Chemical Reduction. The surface coating is PVP. The cell type is IMR90. The number of cells (cells/well) is 5000.0. The origin species of the cell is human. The source of the cell line is Human. The type of cell tissue is Lung. The morphology of the cell is Fibroblast. The cell is in Adult stage. The cell is cell line. The exposure time is 24 hours. The exposure concentration is 25.0 ug/ml. The type of cytotoxicity test is CellTiterGlo. The test mechanism is LuciferaseEnzyme. The size of the nanoparticle is 4.0 nm. The zeta potential indicating surface charge stability of the nanoparticle is -8.0 mV. The surface charge is Negative. Viability (%): 98.293	Predict the viability of a cell. The material of the nanoparticle is Ag. The nanoparticle is inorganic. The morphology of the nanoparticle is Sphere. The fabrication method is Commercial. The surface coating is Citrate. The cell type is CCL-110. The number of cells (cells/well) is 5000.0. The origin species of the cell is human. The source of the cell line is Human. The type of cell tissue is Skin. The morphology of the cell is Fibroblast. The cell is in Fetus stage. The cell is primary. The exposure time is 24 hours. The exposure concentration is 0.5 ug/ml. The type of cytotoxicity test is MTS. The test mechanism is TetrazoliumSalt. The size of the nanoparticle is 39.94 nm. The zeta potential indicating surface charge stability of the nanoparticle is -23.5 mV. The surface charge is Negative. Viability (%):

Main Results

Table 5. Main Results

Drive	Method	Data (%)	Cross-validation		Regression test set		Classification test set	
			Avg. R^2	Avg. MCC	R^2	Inference time (s)	MCC	Inference time (s)
Data	Logistic Regression	100	-	0.2750	-	-	0.16989	0.0041
	Linear Regression	100	0.1816	-	0.16989	0.0048	-	-
	Decision Tree	100	0.6677	0.6460	0.76673	**0.0008**	0.71634	**0.0012**
	Random Forests	100	0.8044	0.6966	0.84581	0.0086	0.76629	0.0062
	Gradient Boosting	100	0.6607	0.6656	0.65157	0.0026	0.69814	0.0017
	TabPFN-v2	90	0.8051	0.7245	0.85896	0.6432	0.76895	0.3127
	TabPFN-v2	100	0.8122	0.7237	0.86839	0.6862	0.76104	0.3397
	TabPFN-v2 (Optimized)	100	**0.8480**	**0.7300**	**0.88296**	0.8713	**0.79255**	0.8338
Knowledge	GPT2	100	-	-	0.78022	2.4049	-	-
	BioGPT	100	-	-	0.79314	4.1377	-	-

Ablation Studies - Missing Values

Table 6. Results of Missing Values Handling

Method	Cross-validation		Test set	
	Avg. R^2	Avg. MCC	R^2	MCC
None	0.8122	**0.7237**	0.86839	0.76104
Mode	**0.8200**	0.7173	**0.87879**	**0.78216**

Ablation Studies - Feature Scaling

Table 7. Results of Feature Scaling

Method	Cross-validation		Test set	
	Avg. R^2	Avg. MCC	R^2	MCC
None	0.8122	**0.7237**	0.86839	0.76104
Standard	0.8115	0.7216	0.86839	0.76625
Min-Max	**0.8131**	0.7186	**0.86908**	**0.76629**

Ablation Studies - Feature Encoding

Table 8. Results of Feature Encoding

Method	Cross-validation		Test set	
	Avg. R^2	Avg. MCC	R^2	MCC
Label	**0.8122**	**0.7237**	**0.86839**	0.76104
One-hot	0.8096	0.7007	0.86474	**0.77946**
Target	0.7869	0.6960	0.85966	0.76125

Ablation Studies - Dimensionality Reduction

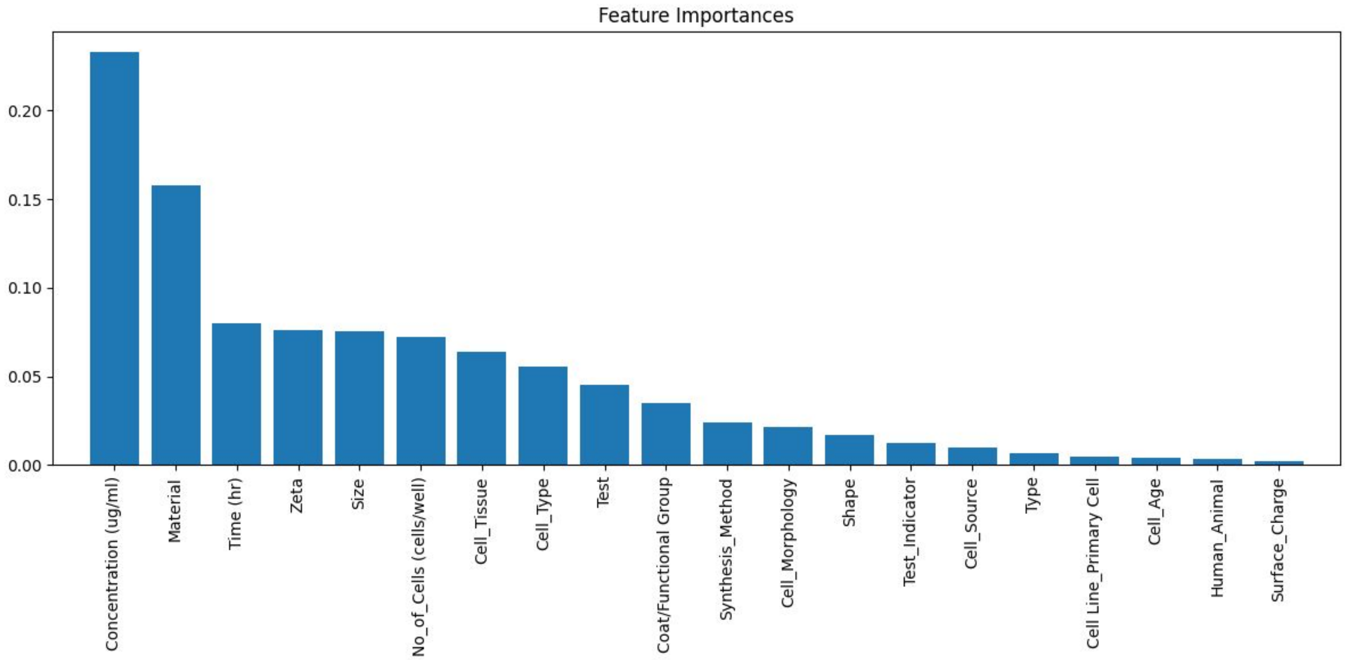

Fig 6. Feature importance in Regression Task

Ablation Studies - Dimensionality Reduction

Table 9. Results of Dimensionality Reduction

Reduced features				Cross-validation		Test set	
Surface_charge	**Human_Animal**	**Cell_Age**	**Cell Line_Primary Cell**	**Avg. R^2**	**Avg. MCC**	**R^2**	**MCC**
				0.8122	0.7237	0.86839	0.76104
✓				0.8130	-	0.86905	-
	✓			0.8107	-	0.87177	-
		✓		0.8072	-	0.86846	-
			✓	0.8099	-	0.87034	-
✓	✓			0.8163	-	0.86997	-
✓	✓	✓		0.8172	-	**0.87186**	-
✓	✓	✓	✓	**0.8177**	-	0.87095	-

Ablation Studies - Outliers

Table 10. Results of Outliers Handling in Regression Task

Detection	Handling		Cross-validation		Test set	
	Trim	Clip	Avg. R^2	Avg. MCC	R^2	MCC
None			0.8122	0.7237	0.86839	0.76104
IQR	✓		0.8408	-	0.86913	-
		✓	0.8288	-	0.86816	-
Z-score	✓		**0.8418**	-	0.86737	-
		✓	0.8291	-	**0.86928**	-

Conclusion

- **Findings:**
  - The experiments demonstrated the effectiveness of the TabPFN-v2 architecture, especially when combined with an optimized data pipeline.
  - The optimized model outperforming other evaluated methods in both regression (R-squared: 0.88296) and classification (MCC: 0.79255) tasks on the test set.
- **Limitations:** The inference time of TabPFN-v2 was substantially longer compared to traditional machine learning models and appeared to increase with the size of the training data.
- **Future Directions:**
  - Future research could focus on reducing the inference time of the TabPFN-v2 model.
  - Exploring the applicability of TabPFN-v2 to larger, high-dimensional datasets is another potential area for future investigation.

CBBL

Computational Biology & Bioinformatics Lab.

Q & A

* Code and materials can be found at: https://github.com/tanthinhdt/cbbl-2025