



An Automated Neural Centric Machine Learning Pipeline

Authors: Mangaliso Mngomezulu¹, Gciniwe Baloyi², Sibusisiwe Makhanya³, on behalf of the HE²AT Centre Consortium
IBM Research^{1,2,3}



Motivation

Health researchers with no deep learning (DL) background frequently work with complex clinical and biomedical datasets and face significant barriers when applying DL methods where classical ML and Statistical methods are not efficient. Working with basic DL requires decisions to be made on model architectures which is not always a trivial task. To speed up experimentation for non-experts, there is a need for an accessible, robust, and interpretable modelling framework that enables rapid experimentation.

System Overview

We present a fully integrated, neural-centric machine-learning pipeline designed specifically for health data analysis. The system operates directly on generic CSV datasets and supports a broad range of modelling tasks, including classification, regression, clustering, anomaly detection, and dimensionality reduction. All modelling is executed through neural networks and established machine-learning methods within a unified, automated workflow. Classical ML models are featured for comparison and justification of necessity for DL methods. The system is built mainly using PyTorch & Scikit learn.

Capabilities

- Direct ingestion of tabular health data (CSV)
- Support for multiple analytical tasks:
 - Classification and regression
 - Clustering and anomaly detection
 - Dimensionality reduction (e.g., PCA)
- Architecture-agnostic neural modelling
- End-to-end automation from preprocessing to evaluation
- Built-in model interpretability via SHAP

Hyper-parameter Optimization

A defining feature of the pipeline is its full automation of model selection and hyper-parameter tuning. Rather than relying on predefined architectures, the system systematically searches across: network depth and width, activation functions, learning rates and schedules, regularization strategies. This step ensures that users of the pipeline not familiar with ML save time that would have been used for deciding on hyperparameters, which is not a trivial task for non-experts.

User-Centric Preprocessing Interface

Feature engineering and preprocessing are performed through an interactive, click-driven interface within a Jupyter notebook environment. Users can: drop or select features, normalize variables, apply dimensionality reduction (e.g. using PCA). This design allows non-programmer domain experts to initiate modelling, iterate rapidly, and explore results without writing code since they already understand data.

Practical Applications

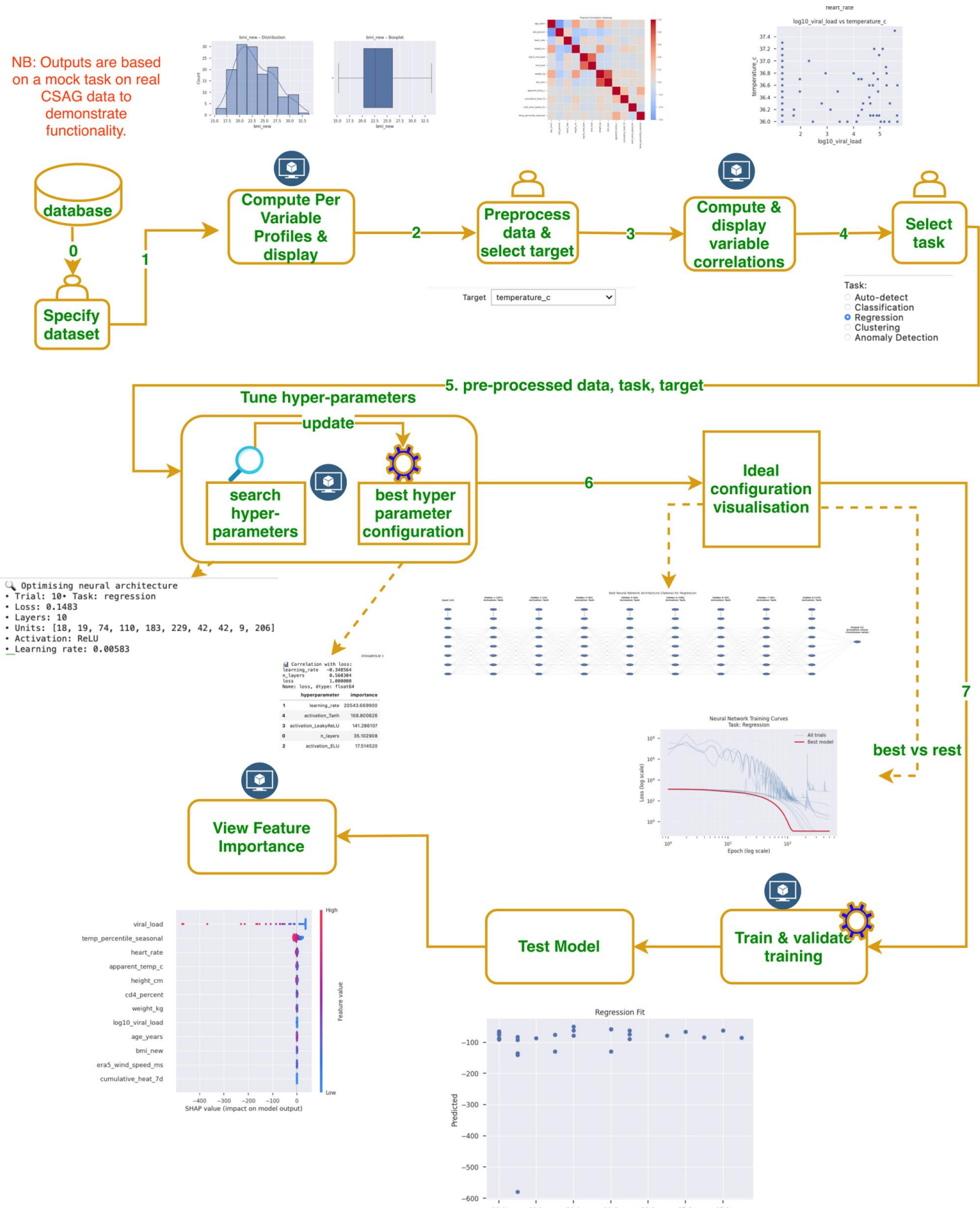
The pipeline is applicable to a wide range of health and biomedical use cases, including:

- Clinical outcome prediction
- Risk stratification
- Patient subgroup discovery
- Exploratory analysis of observational datasets

Conflict of Interests: The authors declare that there are no conflicts of interests associated with this work.

Funding: HE²AT Center is supported by the NIH Common Fund, managed by the Fogarty International Centre (FIC), National Institute of Environmental Health Sciences (NIEHS), and OD/Office of Strategic Coordination (OSC) of the National Institutes of Health (NIH award number: U54TW012083). Disclaimer: The poster content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Fig.1: Pipeline Workflow Overview



Impact on the HEAT projects

By lowering the technical barrier while maintaining methodological rigor, the pipeline supports faster, more reliable, and more transparent health data analysis. What would have taken a days for a non-expert to implement with trust of correctness now takes a few minutes. **This saves over 95% of the time** that would have been otherwise spent on simpler initial analysis. The pipeline has notes on its functionality, hence it also serves as a learning tutorial.

Conclusion

By combining automated neural modelling, interactive preprocessing, and built-in explainability, this pipeline bridges the gap between advanced machine learning and practical health research. It enables domain experts to perform rigorous analyses efficiently while preserving interpretability.

References:

1. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. and Desmaison, A., 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.
2. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12, pp.2825-2830

Corresponding Author: Mangaliso Mngomezulu. Email: Mngomezulum@ibm.com

