Tutorial I: Regression

Acknowledgment

1. This tutorial was prepared by Milad Sangsefidi (milads@uark.edu) and Daniel Alejandro Peraza (daperaza@uark.edu) at the University of Arkansas.
2. The algorithms used in this tutorial are adopted from *J. K. Hoskins, H. Hu, and M. Zou, "Exploring Machine Learning and Machine Vision in Femtosecond Laser Machining," ASME Open Journal of Engineering, 2023, 2: 024501. [link]*. The data used in this tutorial is from *C. Dunlap, H. Pandey, E. Weems, and H. Hu, "Nonintrusive Heat Flux Quantification Using Acoustic Emissions During Pool Boiling," Appl. Therm. Eng., 228, (2023) 120558. [link]*
3. This tutorial was implemented in the MEEG tech elective "Machine Learning for Mechanical Engineers." [link]

Abstract

Reliable heat flux prediction in pool boiling experiments is a sensitive task, since in most cases it is wrapped by non-ideal realities of heat losses-particularly in systems with thin-film heaters like ITO heaters, which have poor thermal conductivities. Different finite-element simulations are more often used to account for these types of heat losses in the experiments to validate assumptions made at the experiment. However, machine learning offers a whole new perspective toward modeling such complex thermal phenomena by 'learning' from non-ideal situations using neural networks. In this Project, we would be showing an application in which the Multilayer Perceptron model is used to predict the heat flux using temperature data measured during a transient pool boiling test. Furthermore, another model using Gaussian Process Regression is also developed for comparison. Both models are trained and tested on the dataset of temperature and heat flux values. Training performance is determined by training curves, and k-fold cross-validation is used to prevent overfitting and to improve model generalization. This work compares the model performances of MLP and GPR regarding their prediction accuracy and computational efficiency that could be the first step toward establishing the relevance of machine learning for improving heat flux predictions in boiling experiments.

Introduction

Machine learning [1] has now emerged as a strong tool for uncovering patterns and making predictions from big data in various scientific and engineering applications. Among the various machine learning models, there exist two popular methods: Multilayer Perceptron and Gaussian Process Regression, which fall under the category of predictive tasks. MLP [2] is a form of artificial neural network that, due to its multi-layered structure of interconnected neurons, is particularly suited for nonlinear relationships and complex data sets. On the other hand, GPR [3] is a probabilistic model; therefore, by accounting for uncertainty in predictions, it offers flexibility towards regression tasks. Both models have been successful on a wide range of problems across the domains of thermodynamics, material science, and fluid dynamics, where prediction accuracy is paramount as well as computational efficiency.

In the present work, performances concerning the prediction accuracy and computational efficiency of MLP and GPR models are compared. The presented project compares the two methods to shed some light on the potential of machine learning in improving the heat flux prediction capability in boiling experiments. With such enhancements, thermal management systems will work considerably better and, thus, contribute to further insights on the real understanding of the heat transfer processes. This understanding is intended to offer further avenues for studies about applications of machine learning in thermofluidic systems and promote efficient designs and processes effectively.

Problem statement

Traditional heat flux calculation for pool boiling experiments is simply taking the supplied power into the system and dividing it by the surface area of the heater. This value is usually imprecise as there are several experimental difficulties, such as heat losses and other non-ideal factors that can occur with low thermal conductivity thin-film heaters like ITO heaters. This problem can be treated in a rather up-to-date way with the help of machine learning by using MLP and GPR models that capture the heat loss and experimental irregularities. The goal is to make a model which will predict the heat flux based on temperatures measured during a transient pool boiling experiment.

Overview:

1. Develop and train two models: MLP and GPR on predicting the heat flux from the temperature measurements.
2. This section presents the training performance in terms of plots about training/validation accuracy and loss across epochs along with time metrics such as time per epoch and total time to reach the optimal model.
3. It should perform k-fold cross-validation, aiming at a minimum size of 10 folds for better generalization.

Procedure:

1. **Data Processing:** It is usually the preliminary step including cleaning of dataset, feature selection which shall provide only the relevant features to the model. Any preprocessing-normalization or encoding is dealt with here.
2. **Data Split:** The dataset is divided into three parts: training, validation, and test sets. The model will learn from the training data, its performance will be checked on the validation set, and the test set is reserved for the final evaluation after the model is built.
3. **Define Model:** The model is defined through MLP and GPR models for the comprehension and prediction between temperature and heat flux
4. **Training & Validation:** Training will take place by iterative weight adjustment to make better predictions on the training data.
5. **Prediction:** As a last step, test data is used to make the predictions with the model. Here, the generalization capability of the model is measured.

6. **Performance Review:** Many of these results are predictions behind which several performance metrics may be calculated for accuracy, loss, or anything else relevant to see how well the model has done.

MLP method

A multilayer perceptron is a neural network containing at least one hidden layer of neurons apart from the input and output layers. The MLP defined here takes in an input and sends it through three hidden layers and then an output layer. The first hidden layer has 64 neurons, the second layer 32 neurons, and the third layer 16 neurons. Hiding layers use the Leaky ReLU activation function with an alpha of 0.05 for a small negative slope, whereas the output layer consists of one neuron, making this model appropriate for regression tasks.

It uses the Adam optimizer during compilation and takes a Mean Squared Error loss, suited for regression. Moving ahead, Adam updater will update the weights of the model in a way to try and minimize the MSE through training. Moreover, the model uses early stopping in case its validation loss is monitored to save overfitting. It stops training the model after a fixed number of epochs if its performance does not improve.

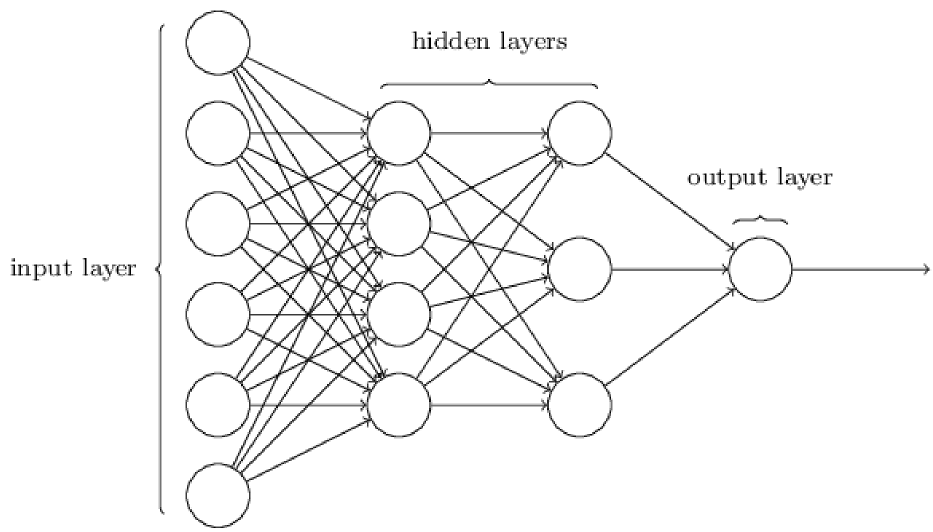

Fig 1. MLP structure

GRP method

GPR stands for a non-parametric probabilistic model used for regression. The following example is making the implementation of GPR with the RBF, Radial Basis Function, kernel, implying setting its smoothness with a fixed length scale of 1.0. The model includes the alpha parameter of value 0.001, defining the noise level in the observations. By default, the model should be normalized, and the random state will ensure reproducibility.

The GPR model is trained using the provided training data and reshaped to be two-dimensional as needed by the model. Then, after training, the GPR model will run predictions on the validation data, while the performances are checked with the MSE and RMSE, which provide the error between the predicted and actual values in the validation set.

K-Fold Cross Validation

The MLP and GPR models go through k-fold cross validation to understand the model's performance across data splits. The validation also helps to determine the best performing configuration. It returns a summary of validation losses and root mean square errors, RMSE, for each fold, best model in that fold, and average validation loss and RMSE across all folds. A fold is an equal-sized subset of the original data that is used independently to validate a model's performance.

How to run the code

1. Open the file in **Google Colab**.
2. Use the "**Run all**" option to execute the code.
3. In the "**Loading and plotting dataset**" section, a "**Choose file**" prompt will appear. Select the "**boiling-32_temp_heat_flux.txt**" file.
4. Navigate to the end of the code and wait for the results to be generated.

Result and discussion

Plots show performance of the MLP and GPR models in performing the task of predicting heat flux from temperature. Both the models have a strong correlation between predicted and actual heat flux, as depicted by the clustering around the ideal line in the scatter plots. However, there were several subtle differences in their predictive abilities.

In the case of the MLP model, the accuracy is slightly higher than that of GPR. This observation could be an indication that MLP is more effective in capturing the underlying relationship between temperature and heat flux for this data set. These findings are further supported by residual plots where residuals of the MLP model tend to be closer to zero, which indicates a smaller deviation from actual values.

The number of layers in a neural network affects its ability to learn complex patterns. In this model, three hidden layers with 64, 32, and 16 neurons are used. More layers can capture more intricate relationships but also increase the risk of overfitting, especially with limited data. To prevent this, L2 regularization is applied. While deeper networks can improve performance, they also require more computation and may face training challenges like vanishing gradients, making it important to balance depth with regularization and data size.

While both models give good performances, the decision between MLP and GPR could fall under specific needs and considerations. For example, GPR would be advantageous in such cases that require interpretability because it has a probabilistic framework for quantifying uncertainty. On the other hand, since an MLP provides the highest accuracy and allows generalization, it could be preferable when one is interested in precise predictions only.

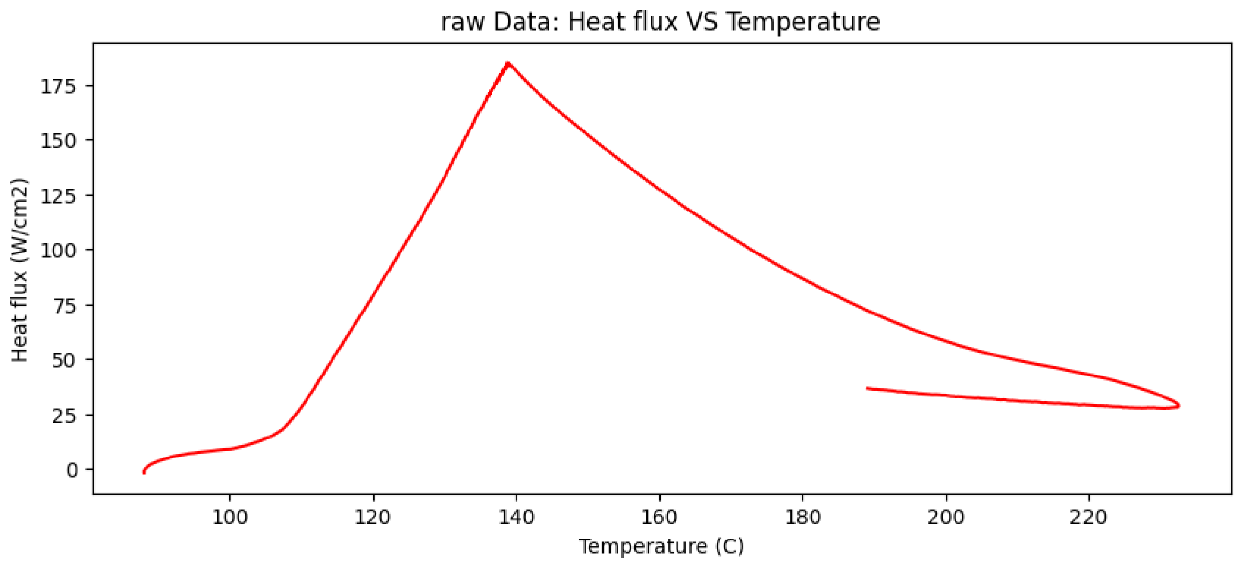

Fig 2. Original data plot

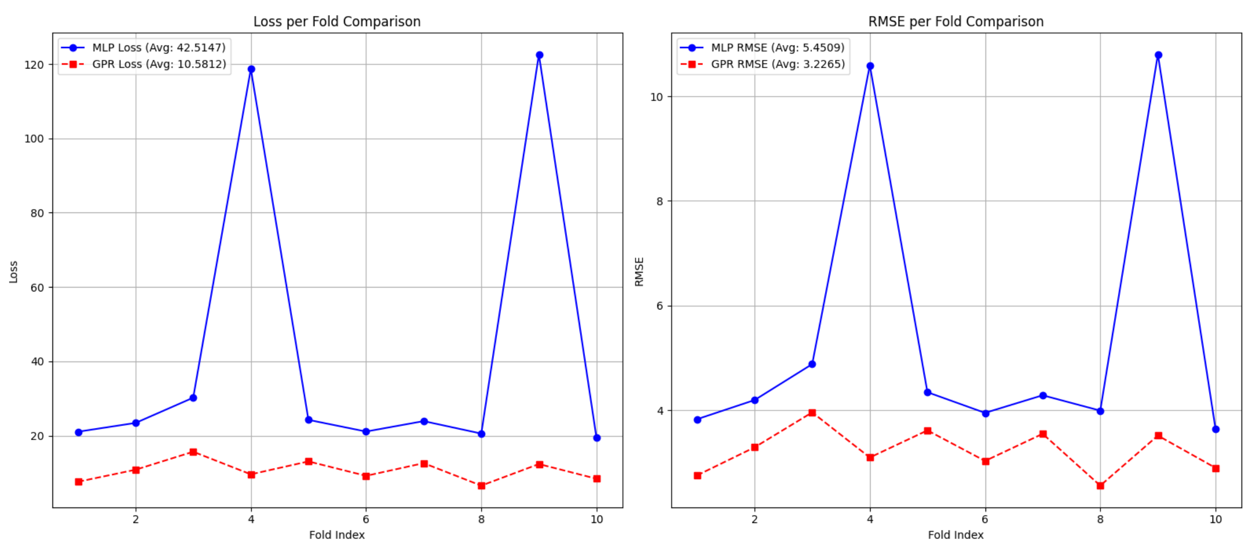

Fig 3. Loss and RMSE per fold (10 folds)

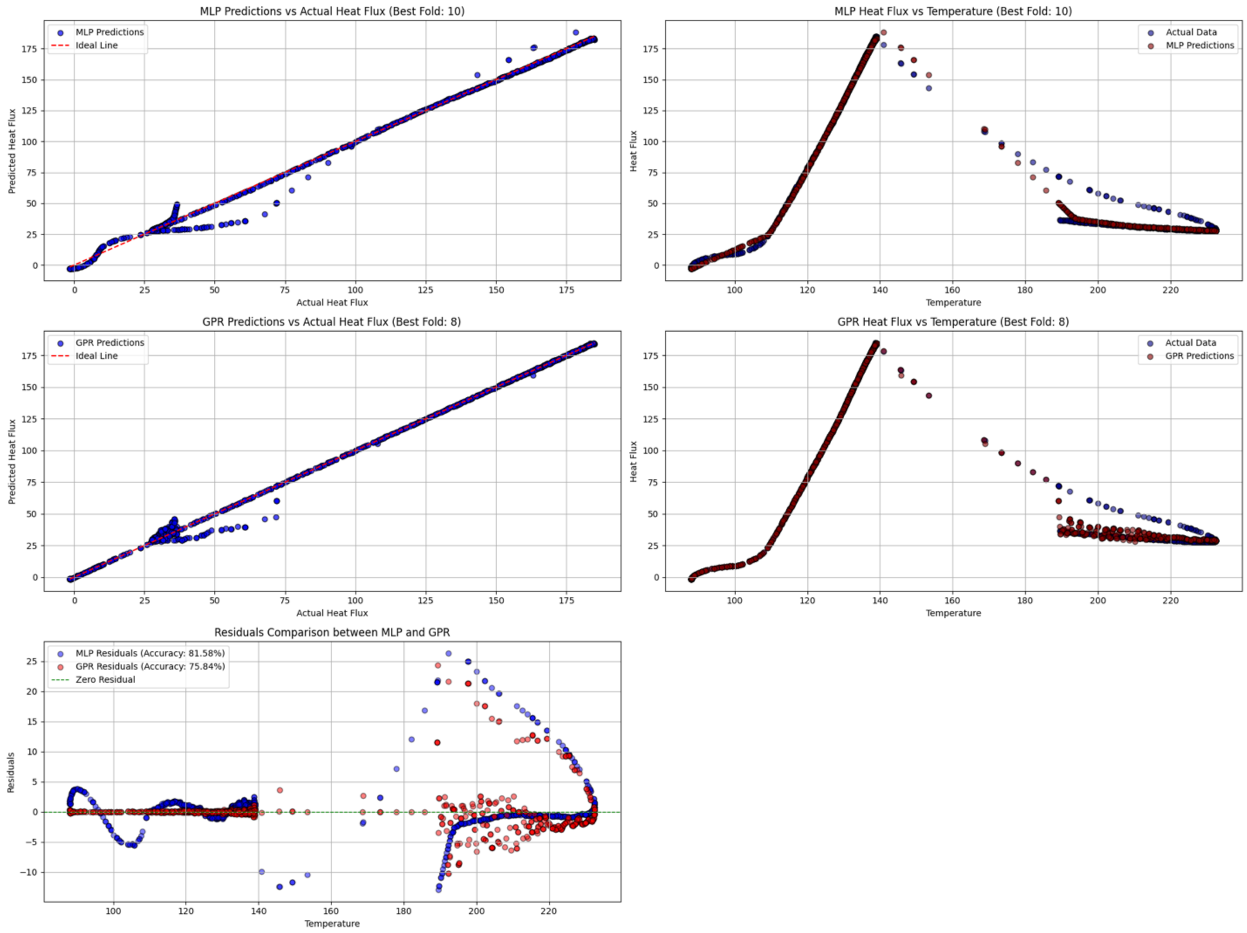

Fig 4. Results (10 folds): MLP vs Actual heat flux. MLP heat flux vs Temperature. GPR vs Actual heat flux. GPR heat flux vs Temperature. MRP and GPR residuals comparison

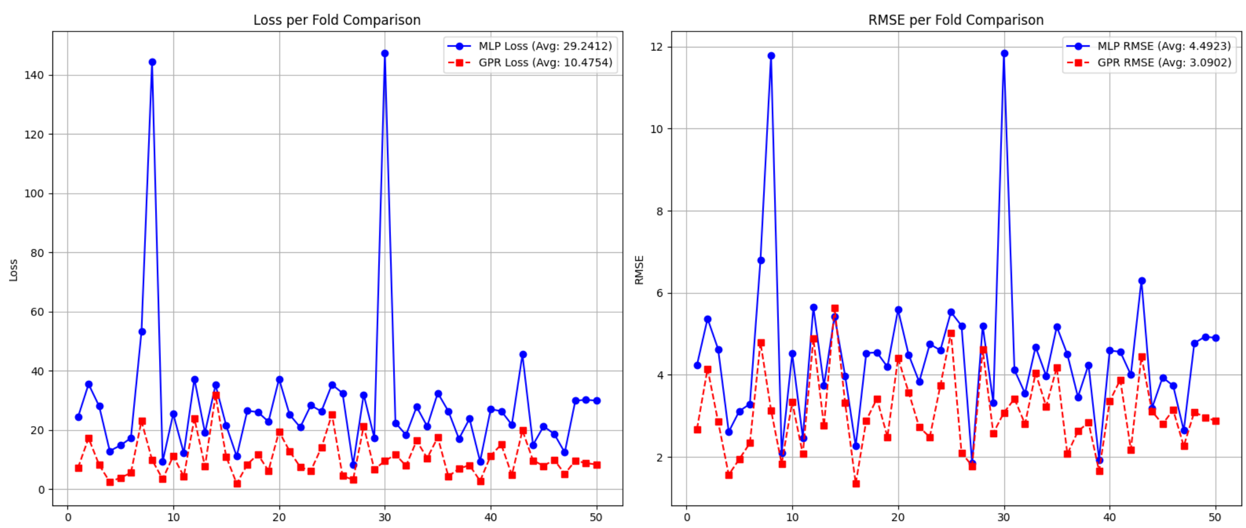

Fig 5. Loss and RMSE per fold (50 folds)

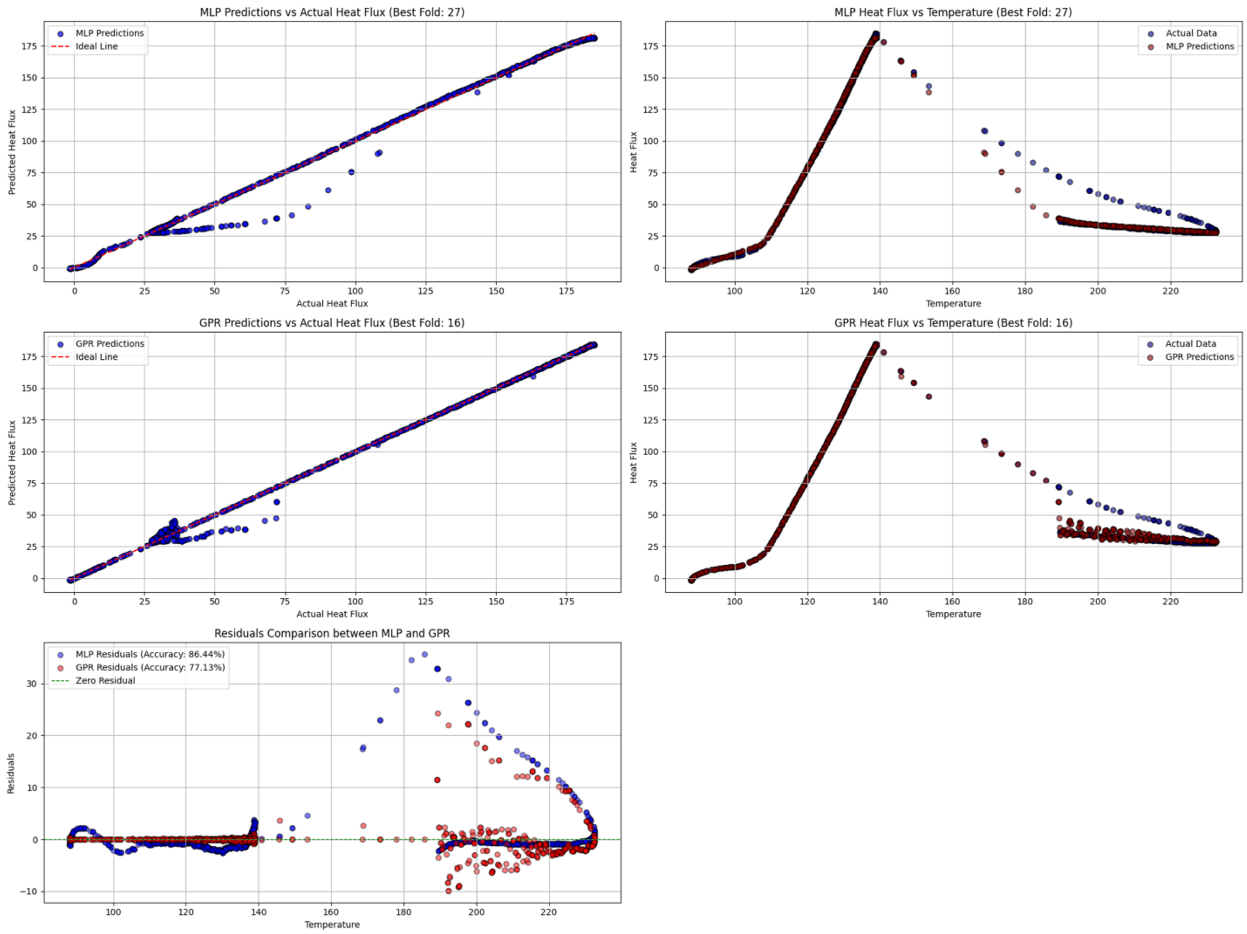

Fig 6. Results (50 folds): MLP vs Actual heat flux. MLP heat flux vs Temperature. GPR vs Actual heat flux. GPR heat flux vs Temperature. MRP and GPR residuals comparison

References

[1] Alpaydin, Ethem. Machine learning. MIT press, 2021.
[2] Delashmit, Walter H., and Michael T. Manry. "Recent developments in multilayer perceptron neural networks." In Proceedings of the seventh annual memphis area engineering and science conference, MAESC, vol. 7, p. 33. 2005.
[3] Deringer, Volker L., Albert P. Bartók, Noam Bernstein, David M. Wilkins, Michele Ceriotti, and Gábor Csányi. "Gaussian process regression for materials and molecules." Chemical Reviews 121, no. 16 (2021): 10073-10141.

Acknowledgment

ChatGPT was utilized to explore new approaches for writing code and to significantly enhance the overall writing style.