Data-Driven Prediction of Temperature in CFD Simulations Using Ensemble Models

Vishal Rasaniya
School of Engineering and Applied Sciences
Mechanical Engineering, Ahmedabad University
Ahmedabad, India 380009
Email: Vishal.r@ahduni.edu.in

***Abstract*—The study explores the application of a Random Forest Regressor (RFR) for predicting temperature evolution in a 2D rectangular enclosure subjected to constant heat flux on both sides. The dataset comprises 547 time-step CSV files generated from CFD simulations in ANSYS Fluent, containing nodenumber-wise temperature, velocity, and pressure values. To capture temporal dependencies, lag features (past temperature, velocity, and pressure values) are introduced. The model undergoes hyperparameter optimization using RandomizedSearchCV to enhance predictive accuracy. Performance evaluation using MAE, RMSE, and R^2 score demonstrates the model's reliability. Additionally, feature importance analysis, actual vs. predicted plots, residual error distributions, and contour plots provide insights into the model's effectiveness. The study concludes that Random Forest is a robust tool for CFD-based temperature prediction.**

1. Introduction

Predicting temperature distribution in enclosed domains is crucial for thermal system design, energy efficiency analysis, and heat transfer optimization. Computational Fluid Dynamics (CFD) simulations provide detailed insights into heat transfer, but they are often computationally expensive. Machine learning (ML) offers an alternative approach to learn temperature evolution patterns from CFD-generated data and make rapid predictions.

This study applies a Random Forest Regressor (RFR) to predict temperature variations over time in a 2D rectangular enclosure heated from both sides. The dataset consists of 547 sequential time-step files containing nodenumber-wise temperature, velocity, and pressure values. By introducing lag features, the model learns temporal dependencies in heat transfer.

To optimize predictive performance, the study employs RandomizedSearchCV for hyperparameter tuning and evaluates the model using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 score. Furthermore, feature importance analysis, residual error plots, and contour plots are used for model interpretation.

The findings demonstrate that Random Forest effectively captures heat transfer dynamics and provides reliable temperature predictions. Future work aims to enhance accuracy through hybrid models (RF-LSTM) and Physics-Informed Neural Networks (PINNs) to incorporate governing heat equations into the learning process.

2. Literature Survey

Several studies have explored machine learning (ML) techniques to develop reduced-order models (ROMs) for computational fluid dynamics (CFD) applications. These approaches aim to improve computational efficiency while maintaining high accuracy in predicting temperature, pressure, and velocity fields.

Mun Yoo (2024) - "Operating Key Factor Analysis of a Rotary Kiln Using a Predictive Model and SHAP" Summary: Mun Yoo developed a predictive model for temperature estimation in rotary kilns using XGBoost, LightGBM, CatBoost, and Gated Recurrent Units (GRU). They used Shapley Additive Explanations (SHAP) to interpret model decisions and optimize industrial processes. Key Contribution: Demonstrated the effectiveness of ensemble learning for improving predictive accuracy and interpretability. Limitation: High computational cost and need for optimization in industrial deployment.

Xiao et al. (2024) - "Multi-Model Fusion for Temperature Prediction in Cement Kilns" Summary: Proposed a hybrid deep learning model using Residual Networks and GRU to enhance temperature forecasting in dynamic environments. Key Contribution: Leveraged deep learning for accurate spatiotemporal predictions of temperature fluctuations. Limitation: Deep learning models require high computational resources, limiting real-time deployment.

Wang et al. (2024) - "Machine Learning for Real-Time Temperature Field Optimization in Industrial Kilns" Summary: Introduced Random Forest (RF) and Gradient Boosting to integrate ML with CFD simulations for real-time process optimization. Key Contribution: Demonstrated that tree-based models provide fast and reliable predictions when trained on high-fidelity CFD data. Limitation: Requires extensive CFD datasets for training, limiting adaptability to new operating conditions.

Schmelzer et al. (2020) - "Sparse Symbolic Regression for Turbulence Modeling" Summary: Developed SpaRTA (Sparse Regression of Turbulent Stress Anisotropy), a physics-informed ML approach for turbulence modeling.

Key Contribution: Generated interpretable symbolic expressions for ROM turbulence predictions. Limitation: Limited scalability for highly nonlinear turbulent flows, reducing accuracy in complex CFD cases.

Halder et al. (2024) - "Ensemble Learning for Reduced-Order Modeling in CFD" Summary: Utilized Bagging with LSTM and Autoencoders to enhance ROM stability and reduce error propagation. Key Contribution: Improved generalization and robustness in time-dependent CFD simulations. Limitation: High training costs due to multiple weak learners, making real-time deployment challenging.

3. Dataset Discussion

The dataset used in this study originates from Computational Fluid Dynamics (CFD) simulations, specifically focusing on temperature prediction over time. It consists of 547 time steps and contains multiple physical parameters that influence heat transfer. The dataset captures both spatial (nodal) and temporal (time-series) dependencies, making it suitable for machine learning-based reduced-order modeling (ROM).

3.1. Dataset Structure

- **Source:** CFD-generated simulation data.
- **Number of Files:** 500+ CSV files, each representing one second of simulation.
- **Time Range:** 3.38s to 549.38s.
- **File Format:** Each CSV file contains nodal data with multiple physical variables.

File Name Example	Time Extracted
both_side_1KWcase–3.38.csv	3.38 sec
both_side_1KWcase–4.38.csv	4.38 sec
both_side_1KWcase–549.38.csv	549.38 sec

TABLE 1. EXAMPLE OF FILE NAMING AND TIME EXTRACTION

3.2. Features and Target Variable

Features (Predictors):

- NodeNumber (Spatial node identifier)
- X-Coordinate, Y-Coordinate (Spatial position)
- Total-Pressure (Fluid pressure at the node)
- Total-Energy (Energy contained in the system)
- Heat-Flux (Heat transfer rate)
- DX-Velocity-DY, DY-Velocity-DY (Velocity gradients)
- DP-DX, DP-DY (Pressure gradients)
- Time (Extracted from filenames)

Target Variable: *Total Temperature* (Dependent variable for ML prediction)

3.3. Data Preprocessing and Feature Engineering

To improve model accuracy, the dataset undergoes various preprocessing and feature engineering steps:

- **Time Extraction:** The time variable is derived from filenames and added as a feature.
- **Lag Features:** Previous time step values of Total Temperature are used:
  - Temp_Lag1 (1 step before)
  - Temp_Lag2 (2 steps before)
  - Temp_Lag3 (3 steps before)
- **Moving Averages:** Capturing temperature trends over different time intervals:
  - Temp_MA3 (3-time-step rolling average)
  - Temp_MA5 (5-time-step rolling average)
- **Handling Missing Values:** Missing values are handled using interpolation or imputation methods.
- **Scaling:** Standardization or normalization is applied if required.

3.4. Train-Test Splitting (Time-Aware Approach)

A time-aware train-test split is applied to preserve the sequence:

- **80% Training Data:** Used for training ML models.
- **20% Testing Data:** Used for model evaluation.
- Ensures no future data is leaked into the training phase.

3.5. Challenges in the Dataset

- **High Dimensionality:** Large nodal and physical features may require dimensionality reduction techniques.
- **Temporal Dependencies:** ML models need to capture sequential relationships effectively.
- **Data Imbalance:** Some temperature ranges may be overrepresented.
- **CFD Simulation Noise:** Numerical errors from simulations may introduce uncertainties in ML predictions.

3.6. Justification for Dataset Selection

- **Real-World CFD Data:** Derived from high-fidelity CFD simulations, ensuring accuracy.
- **Rich Temporal and Spatial Information:** Contains both time-dependent and spatial features.
- **Scalability:** Can be extended to incorporate physics-based constraints such as PINNs.

4. Approach

This study employs a machine learning-based reduced-order modeling (ROM) approach to predict temperature evolution in CFD simulations. The methodology consists of several key steps, as outlined below:

4.1. Data Collection and Preprocessing

- Load CFD-generated time-series data from multiple CSV files.
- Extract time information from filenames and structure the dataset accordingly.
- Perform feature engineering, including lag features, moving averages, and normalization.

4.2. Train-Test Splitting (Time-Aware)

- Ensure an 80% training and 20% testing split while maintaining temporal order.
- Avoid data leakage by ensuring future time steps are not included in training data.

4.3. Model Development

- Train individual machine learning models:
  - Random Forest Regressor (RF)
  - XGBoost Regressor (XGB)
  - Gradient Boosting Regressor (GBR)
- Implement a stacked ensemble learning approach:
  - Use predictions from RF, XGB, and GBR as inputs to a meta-model.
  - Train a Linear Regression model as the meta-learner.

4.4. Model Evaluation and Validation

- Evaluate models using:
  - Mean Absolute Error (MAE)
  - Root Mean Squared Error (RMSE)
  - R-squared Score (R^2)
- Compare individual models with the ensemble model.
- Visualize error distributions and trends over temperature values.

4.5. Model Saving and Deployment

- Save trained models using the Pickle format for future inference.
- Explore potential deployment strategies for real-time prediction applications.

5. Future Work

While the proposed approach provides a reliable framework for CFD temperature prediction, further improvements can be made. Future work will focus on the following aspects:

5.1. Integration of Physics-Informed Neural Networks (PINNs)

- Incorporate physical constraints into the learning process to enhance generalization.
- Improve interpretability by aligning ML predictions with CFD governing equations.

5.2. Hyperparameter Optimization

- Utilize Bayesian Optimization or Grid Search to fine-tune model parameters.
- Optimize the number of estimators, learning rate, and depth for ensemble models.

5.3. Feature Selection and Dimensionality Reduction

- Implement Principal Component Analysis (PCA) or Autoencoders to reduce computational complexity.
- Identify the most influential features contributing to temperature variations.

5.4. Comparative Analysis with Deep Learning Approaches

- Investigate Convolutional Neural Networks (CNNs) or Long Short-Term Memory (LSTM) networks.
- Compare deep learning models with ensemble learning for accuracy and computational efficiency.

The implementation of these future improvements will further refine the predictive capability and robustness of the proposed ROM framework for CFD applications.

References

[1] Mun, S., & Yoo, J. (2024). Operating Key Factor Analysis of a Rotary Kiln Using a Predictive Model and Shapley Additive Explanations. *Electronics, 13*(22), 4413.

[2] Xu, X., et al. (2023). Cement rotary kiln temperature prediction based on time-delay calculation and residual network and bidirectional novel gated recurrent unit multi-model fusion. *Measurement, 218*, 113123.

[3] Wang, Y., et al. (2024). A soft measurement model construction method based on machine learning and CFD. *METALLURGIA ITALIANA, 11-12*.

[4] Halder, R., et al. (2024). Reduced-order modeling of unsteady fluid flow using neural network ensembles. *Physics of Fluids, 36*(7).

[5] Schmelzer, M., Dwight, R. P., & Cinnella, P. (2020). Discovery of algebraic Reynolds-stress models using sparse symbolic regression. *Flow, Turbulence and Combustion, 104*, 579-603.

[6] Ling, J., Kurzawski, A., & Templeton, J. (2016). Reynolds averaged turbulence modelling using deep neural networks with embedded invariance. *Journal of Fluid Mechanics, 807*, 155-166.