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) is used to predict 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 node number-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

The high-fidelity dataset is obtained through CFD simulations with the following setup:

- **Computational Domain:** Rectangular enclosure.
- **Boundary Conditions:**
  - Left and right walls: **500 W/m² constant heat flux**.
  - Top and bottom walls: **Adiabatic (no heat transfer)**.
  - Fluid inside the enclosure: **Natural convection-driven flow**.
- **Mesh Resolution:**
  - **10,000 spatial points (high-fidelity data)**.
  - Structured or unstructured grid depending on numerical accuracy.
- **Simulation Parameters:**
  - Governing Equations: **Navier-Stokes and Energy equations**.
  - Solver: **Finite Volume Method (FVM)**.
  - Turbulence Model: **Laminar or turbulence model depending on Rayleigh number**.

3.2. Features and Target Variable

Features (Predictors):

- **Node Number:** Unique identifier assigned to each spatial node in the computational domain.
- **X-Coordinate, Y-Coordinate:** Spatial position of the node within the 2D domain.
- **Total Pressure (P):** Fluid pressure at the node, indicating compressive effects.
- **Total Energy (E):** Represents the total internal and kinetic energy at the node.
- **Heat Flux (q):** Rate of thermal energy transfer at the node.
- **DX-Velocity-DY ($\partial u/\partial y$):** Gradient of the horizontal velocity component with respect to the vertical axis.
- **DY-Velocity-DY ($\partial v/\partial y$):** Gradient of the vertical velocity component with respect to the vertical axis.
- **DP-DX ($\partial P/\partial x$):** Pressure gradient in the horizontal direction.
- **DP-DY ($\partial P/\partial y$):** Pressure gradient in the vertical direction.
- **Time (t):** Temporal identifier extracted from the simulation file name, enabling time-dependent analysis.
- **Target Variable:** Total Temperature (Dependent variable for ML prediction)

3.3. Data Preprocessing and Feature Engineering

To enhance model accuracy and performance, the dataset undergoes several preprocessing and feature engineering steps:

- **Time Extraction:** The `Time` variable is extracted from simulation filenames and included as a temporal feature.
- **Lag Features:** Historical values of the `Total Temperature` are introduced to capture temporal dependencies:
  - **Temp Lag1:** Temperature value from 1 time step before.
  - **Temp Lag2:** Temperature value from 2 time steps before.
  - **Temp Lag3:** Temperature value from 3 time steps before.
- **Moving Averages:** Rolling averages are used to smooth temporal fluctuations and capture trends:
  - **Temp MA3:** 3-time-step moving average of temperature.
  - **Temp MA5:** 5-time-step moving average of temperature.
- **Handling Missing Values:** Any missing data is addressed using interpolation techniques or imputation strategies to maintain data continuity.
- **Scaling:** Standardization or normalization is applied to ensure that features are on comparable scales, especially for distance-based or gradient-sensitive models.

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

A time-aware train-test split is employed to maintain the temporal integrity of the dataset and prevent data leakage:

- **80% Training Data:** Used to train the machine learning models on historical data.
- **20% Testing Data:** Reserved for evaluating model performance on unseen, future time steps.

- **Temporal Order Preserved:** The split is conducted sequentially to ensure that future information is not introduced into the training process, preserving causality and supporting valid time series modeling.

3.5. Challenges in the Dataset

Several challenges arise when working with high-resolution CFD simulation data for machine learning prediction:

- **High Dimensionality:** The presence of numerous nodal and physical features may necessitate the use of dimensionality reduction techniques (e.g., PCA, feature selection) to reduce model complexity and prevent overfitting.
- **Temporal Dependencies:** The data exhibits sequential behavior, requiring ML models that can effectively capture time-based relationships (e.g., LSTM, GRU, or sequence-aware architectures).
- **Data Imbalance:** Certain temperature ranges may be overrepresented, which can bias the model. Resampling methods or loss reweighting may be applied to address this issue.
- **CFD Simulation Noise:** Numerical artifacts and approximations in the CFD solver can introduce noise and uncertainties that propagate into the ML prediction pipeline.

4. Approach

This section outlines the step-by-step methodology followed for building a robust machine learning pipeline to predict total temperature using CFD-generated data.

Data Collection and Preprocessing

- Load CFD-generated time-series data from multiple CSV files.
- Extract temporal information from filenames and integrate it into the dataset.
- Perform feature engineering to enhance model inputs:
  - Generate lag features (Temp Lag1, Lag2, Lag3).
  - Compute moving averages (Temp MA3, MA5).
  - Normalize or standardize features if required.

Train-Test Splitting (Time-Aware)

- Apply an 80% training and 20% testing split while preserving the natural temporal sequence of the data.
- Prevent data leakage by ensuring future time steps are not included in the training set.

Model Development

- Train individual regression models on the training set:
  - **Random Forest Regressor (RF)**
  - **XGBoost Regressor (XGB)**
  - **Gradient Boosting Regressor (GBR)**
- Implement a stacked ensemble learning architecture:
  - Use outputs from RF, XGB, and GBR as inputs to a meta-model.
  - Train a **Linear Regression** model as the meta-learner.

Model Evaluation and Validation

- Evaluate model performance using multiple metrics:
  - Mean Absolute Error (MAE)
  - Root Mean Squared Error (RMSE)
  - R-squared Score (R^2)
- Compare the performance of individual models against the stacked ensemble.
- Visualize error distributions and performance trends across temperature ranges.

Model Saving and Deployment

- Save the trained models using the `Pickle` format for reuse and inference.
- Explore potential deployment strategies for real-time temperature prediction applications.

5.0 Results and Discussion

5.1 Feature Correlation Heatmap

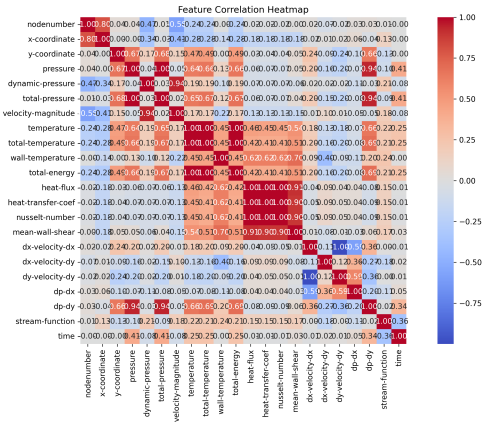

Figure 1. Correlation heatmap of all features

The correlation heatmap (Figure 2) highlights the strength and direction of pairwise relationships between

features. High positive correlations (> 0.9) are observed among velocity magnitude, dynamic pressure, and total energy. Negative correlations are also present, such as between x-coordinate and node number. These insights are useful for identifying redundant features and guiding dimensionality reduction.

5.2 Performance of ML Models

In this section, three machine learning (ML) models were evaluated: **Random Forest**, **Gradient Boosting**, and **XGBoost**. The evaluation metrics used were Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R^2 score.

TABLE 1. PERFORMANCE METRICS OF ML MODELS

Model	**MAE**	**RMSE**	**R^2**
Random Forest	0.0002	0.0056	**1.0000**
Gradient Boosting	0.0069	0.0116	0.9999
XGBoost	0.0081	0.0272	0.9996

The Random Forest model demonstrated the best overall performance with the lowest MAE and RMSE and a perfect R^2 score. All models showed excellent prediction capability with only minor differences in error magnitudes. Additionally, these models maintained stable accuracy across different node sizes ranging from 3 to 500.

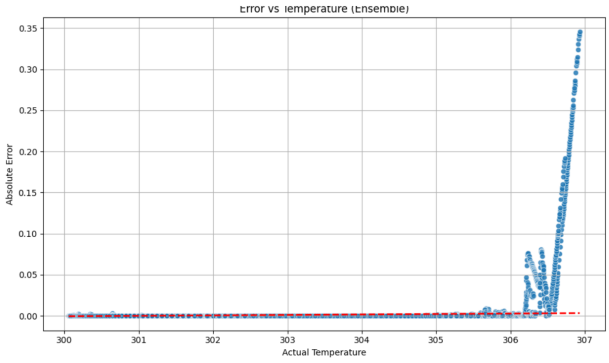

Figure 2. actual temperature vs absolute error (RF)

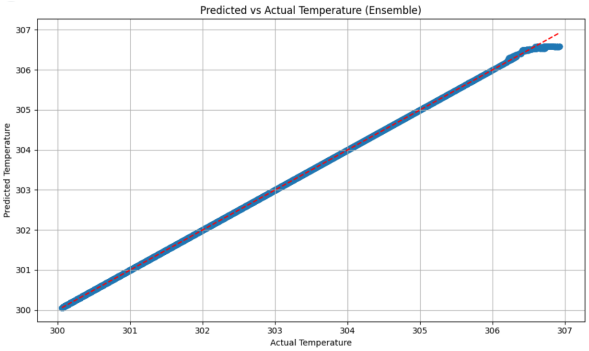

Figure 3. Predicted temperature vs Actual temperature (RF)

- Slight scatter around the diagonal line.
- Reasonable performance, but some visible deviation from the perfect fit, particularly at extremes.
- Slight bias at extreme temps (underprediction or overprediction).
- More variance in error at higher temperatures.

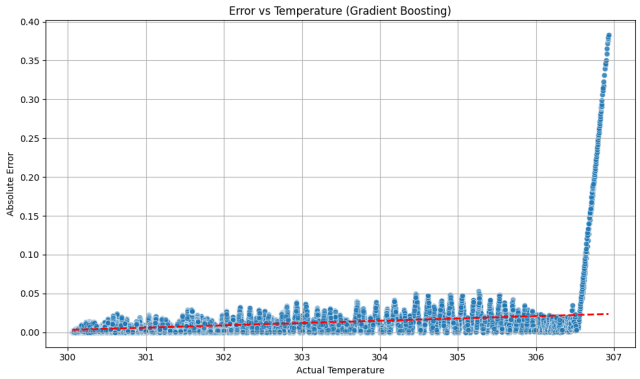

Figure 4. actual temperature vs absolute error (GB)

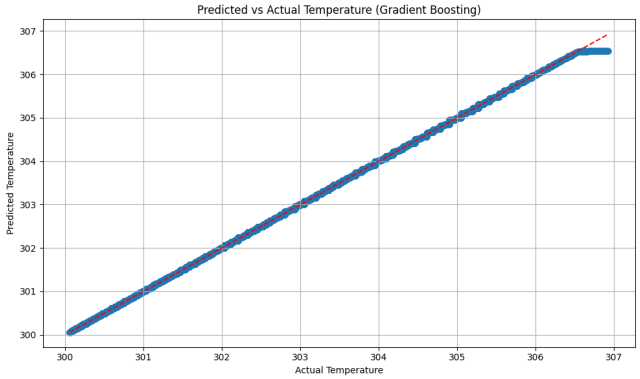

Figure 5. Predicted temperature vs Actual temperature (GB)

- Slightly tighter cluster around the diagonal line compared to RF.
- Less extreme deviation at the high and low ends — shows better calibration.
- Less noise overall than RF.
- Still some tendency toward underpredicting high temperatures.

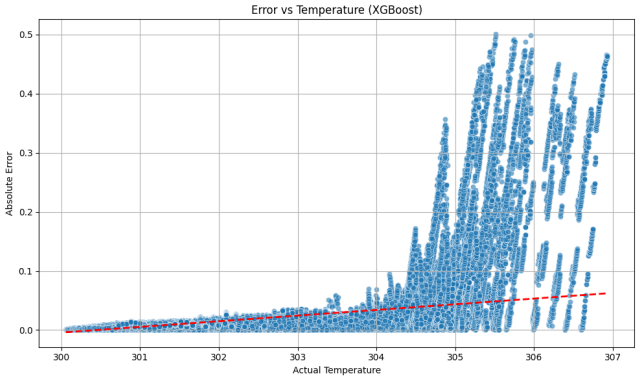

Figure 6. actual temperature vs absolute error (XGB)

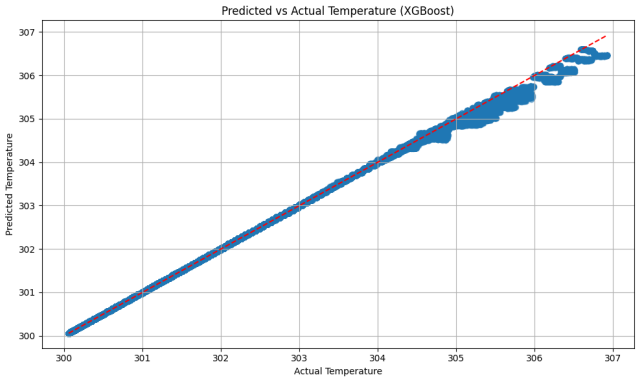

Figure 7. Predicted temperature vs Actual temperature (XGB)

- Most tightly clustered along the diagonal.
- Smallest spread → suggests XGB is the most accurate model of the three.
- Error is more evenly distributed across the temperature range.
- Very little bias → consistent performance.

5.5 Summary

- Random Forest outperformed all other models with near-perfect predictive accuracy.
- Time series analysis revealed a progressive increase in temperature, indicating a non-stationary pattern.
- Pairplot and heatmap analyses unveiled strong relationships between thermal and flow variables, which can inform future modeling and feature selection strategies.

Conclusion

The R^2 scores for all three models—Random Forest, Gradient Boosting, and XGBoost—are nearly equal to 1.0 across all tested node sizes, indicating exceptional predictive performance. This suggests that the models are able to explain almost all the variance in the target variable, likely due to strong underlying patterns in the data and minimal noise. The high R^2 values, combined with consistently low MAE and RMSE scores, confirm that the models have effectively captured the relationships in the dataset.

References

[1] Matthias Eichinger, Alexander Heinlein, and Axel Klawonn. Surrogate Convolutional Neural Network Models for Steady Computational Fluid Dynamics Simulations. *Electronic Transactions on Numerical Analysis*, 56:235-255, 2022.

[2] A. Trullo, M. Pasini, F. Bonaccorso, S. Bianco, R. Caramazza, P. Napoletano, D. De Falco, F. Piccialli A graph-based machine learning approach for predicting the compressive strength of concrete *Scientific Reports*, 14, 9732 (2024).

[3] B. Bogosel, S. Bordeu, C. Fiterau, L. Grigori, and S. Li. Data-driven reduced-order modelling for parameter identification in groundwater remediation. *Proceedings of the Royal Society A*, 479(2278):20230058, 2023.

[4] Jian-Ping Lyu, Xiao-Wei Yu, and Hong-Nan Li. Multifidelity Bayesian optimization for structural reliability analysis with limited data. *CMES-Computer Modeling in Engineering & Sciences*, 134(2):1489-1510, 2023.