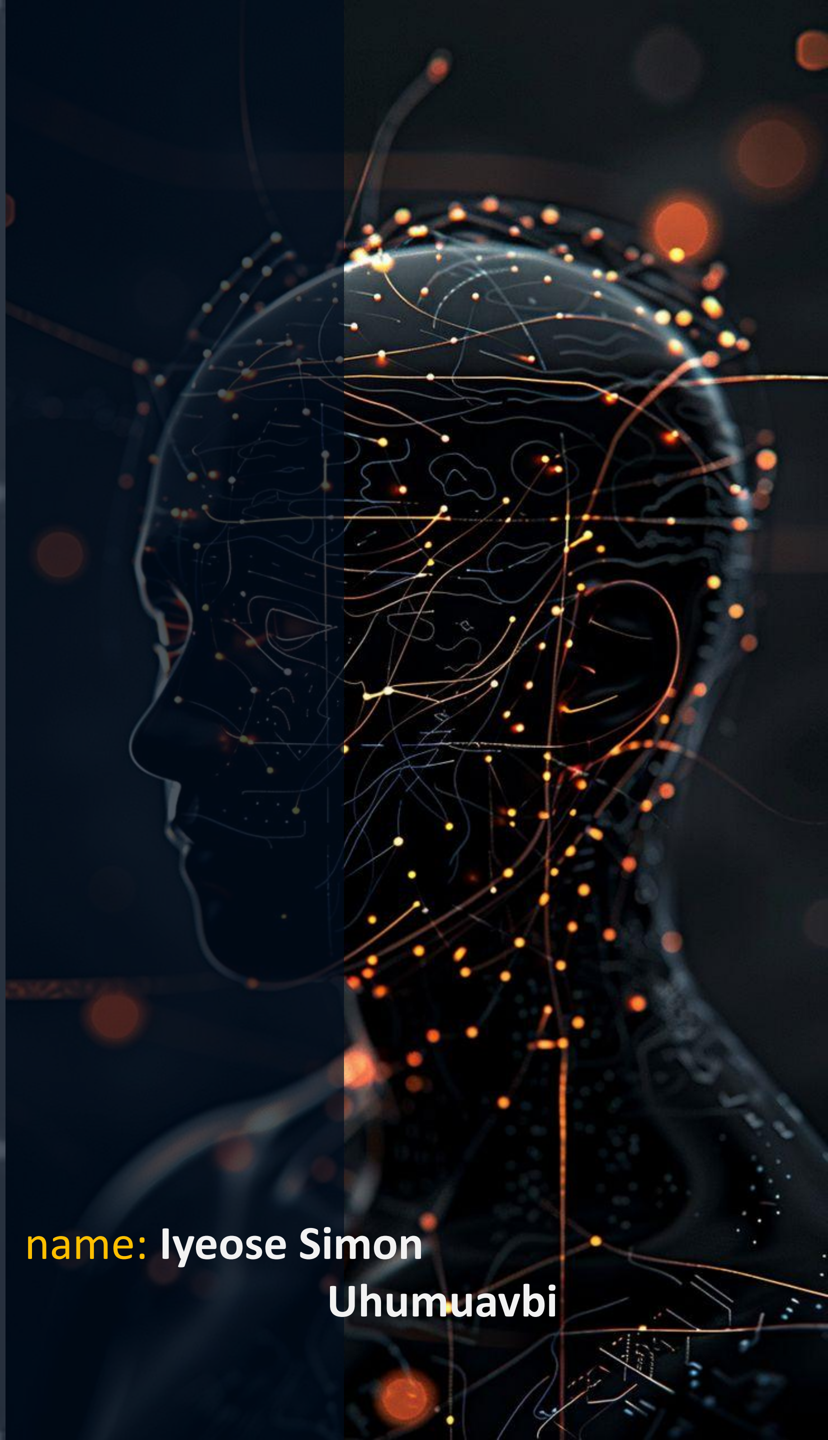
Data-Driven Prediction of Human Thermal Comfort

Using Machine Learning Approaches



name: Iyeose Simon Uhumuavbi

The Problem with "One-Size-Fits-All" Comfort

Thermal comfort is critical for well-being and productivity, as we spend 80-90% of our time indoors.

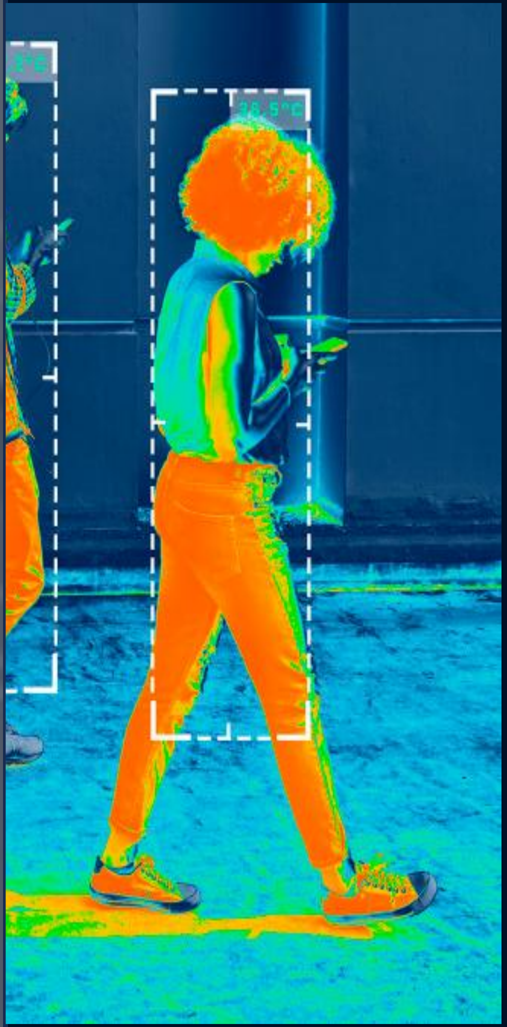
Current HVAC systems are inefficient, accounting for ~40% of building energy use, often failing to satisfy occupants due to a "one-size-fits-all" approach.

PMV Model (Static):

Based on lab conditions, often inaccurate in real-world buildings.

Adaptive Model (Dynamic):

Better, but too general and struggles with individual differences



Research Aims & Questions

Develop & evaluate ML models for predicting thermal comfort.
Move beyond one-size-fits-all models

Question 1

Can ML models predict thermal sensation more accurately than a baseline linear model?

Question 2

Which ML algorithm provides the best balance of accuracy and reliability after optimization?

Question 3

What are the most significant variables that influence thermal comfort, and what can they teach us about building design?

Methodology Overview

Dataset:

Utilized the **ASHRAE Global Thermal Comfort Database II** (>100,000 entries).

Data Preprocessing

- Handled missing values using imputation (Iterative Imputer).
- Corrected impossible data points (e.g., negative age).
- Selected key features to avoid multicollinearity.

Class Imbalance

Addressed severe imbalance in the target variable (Thermal Sensation) using **SMOTE** on the training data.

Model Selection

Benchmarked five diverse ML algorithms.

Optimization

Performed hyperparameter tuning on top-performing models

Evaluation & Interpretation:

Assessed models using accuracy and F1-scores and analyzed feature importance.

The Dataset & Exploratory Data Analysis (EDA)

Personal Data

Clothing (clo),
Metabolic Rate (met),
Age

Contextual Data

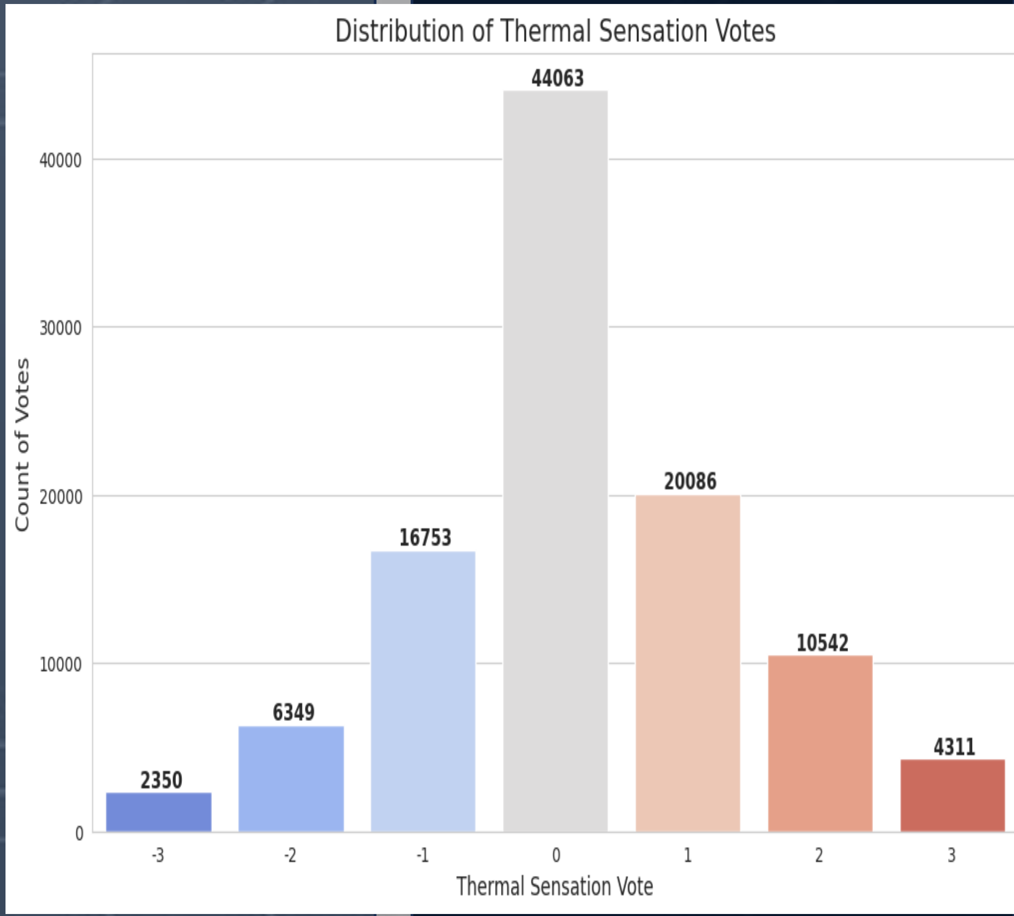
Season, Building Type,
Köppen Climate
Classification

Thermal Sensation Vote (-3 to +3)

Environmental Data

Operative Temp,
Humidity,
Air Velocity

Key EDA Finding: Severe Class Imbalance



The "Neutral" (0) class heavily dominates the dataset.
This required the use of SMOTE to prevent the model from simply guessing the most common class.

Initial Model Benchmarking

Model	Accuracy	Key Takeaway
Random Forest	**53.0%**	**Clear top performer out-of-the-box.**
Deep Neural Network	47.0%	Good, but inferior to ensembles.
XGBoost	45.0%	Good, but inferior to ensembles.
Support Vector Machine	35.0%	Poor performance, computationally expensive.
Logistic Regression (Baseline)	**23.0%**	**Confirms that a simple linear model is inadequate.**

Non-linear, tree-based ensemble models (Random Forest, XGBoost) are far superior for this task.

Model Optimization & Champion Selection

51%

Random Forest
Performance remained stable at ~51%.

53.1%

XGBoost
Performance **dramatically improved from 45.0% to 53.1%.**

46%

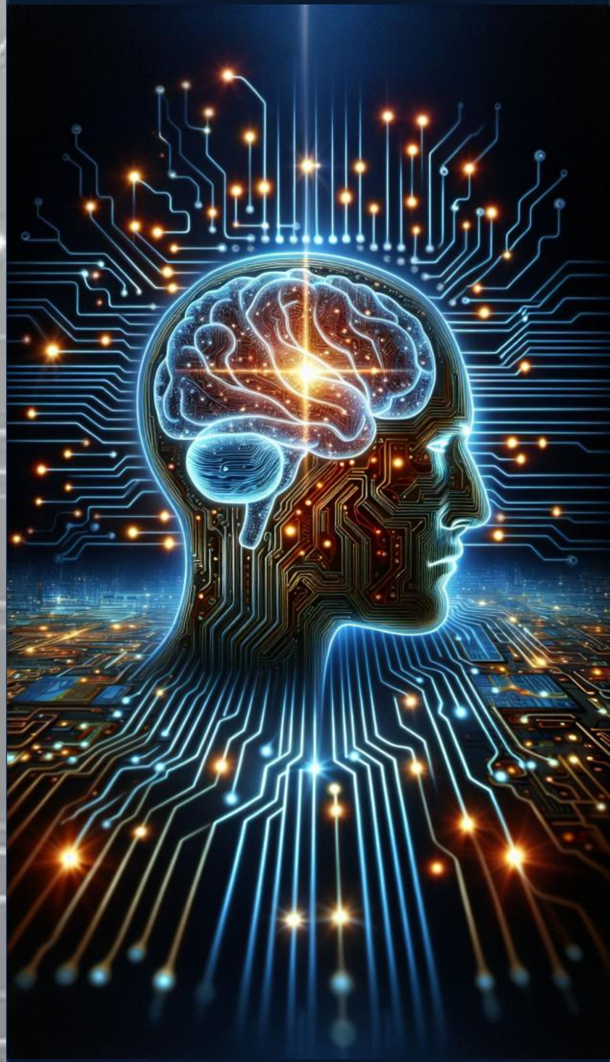
DNN
Tuning failed, creating a flawed model that ignored an entire class.(+3)

The Champion Model: The **Tuned XGBoost** model was selected as the definitive champion due to its superior accuracy (53.1%) and demonstrated capacity for optimization.

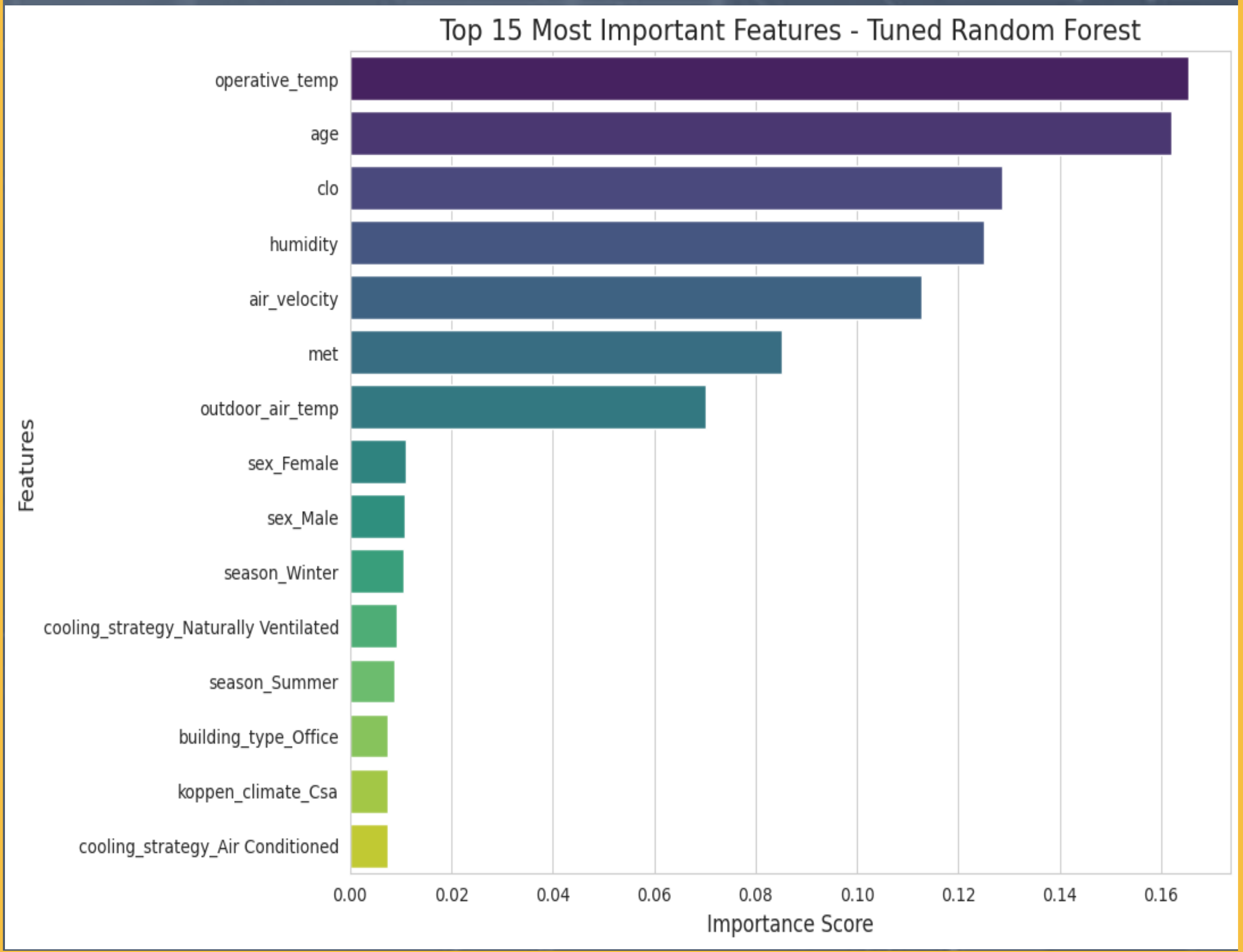
The Core Finding: Two Models, Two Different Strategies

The most fascinating result came from comparing the **feature importances** of the two best models: Random Forest and XGBoost.

Despite achieving nearly identical accuracy (~53%), they learned completely different ways to predict comfort.



The "Physicist" Model: Random Forest



Top Predictors:

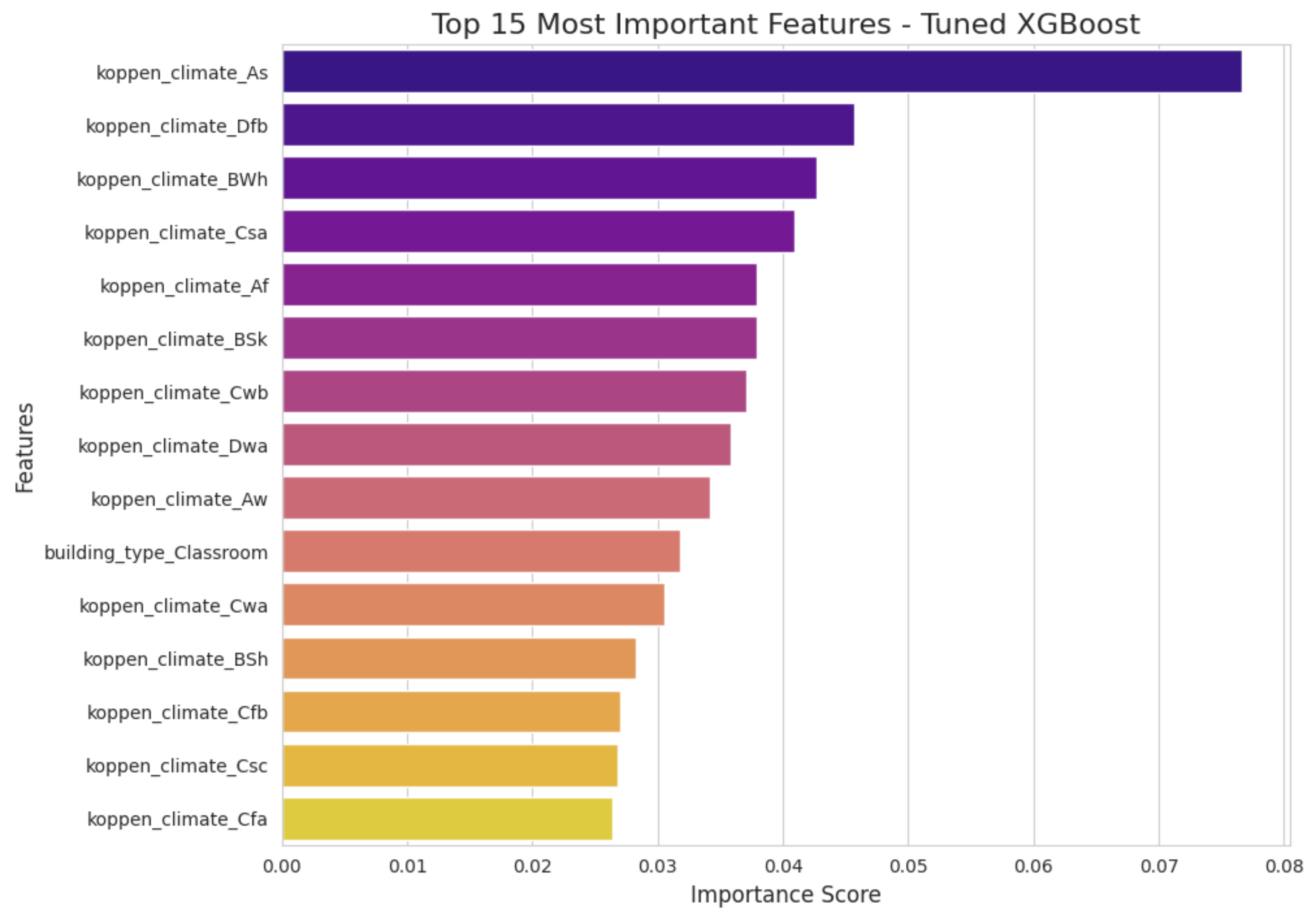
1. Operative
2. Temperature
3. Age
4. Clothing (clo)

Strategy

Prioritizes the **immediate physical reality** of the occupant.

This model thinks like a traditional building scientist, learning an advanced, non-linear version of the PMV heat-balance logic. It focuses on measurable, physical variables.

The "Geographer" Model: XGBoost (The Champion)



Strategy

Prioritizes the **immediate physical reality** of the occupant.

This model learned that knowing *where* you are is a more powerful predictor than knowing the exact temperature. Climate acts as a powerful **proxy** for long-term adaptation, expectations, and building styles.

Top Predictors:

1. **Köppen Climate_As = Tropical Dry Savanna**
2. **Köppen Climate_Dfb = Warm Summer Humid Continental**
3. **Köppen Climate Classification_BW = Hot Desert'**

Synthesizing the Findings: A Hierarchy of Comfort Drivers

Tier 1

Immediate Physical State: Operative Temp, Age, Clothing

Overarching Context: Köppen Climate Classification

Tier 2

Humidity, Air Velocity, Metabolic Rate

Tier 3

Season, Cooling Strategy, Building Type

A successful comfort prediction requires understanding BOTH the occupant's immediate physical reality AND their broader environmental context.

Answering the Research Questions

RQ1

Can ML models be more accurate?

Yes. A 130% relative improvement over the baseline (23.0% -> 53.1%).

RQ2

Which algorithm is optimal?

Tuned XGBoost provides the highest accuracy. Random Forest offers a more intuitive, physics-based interpretation.

RQ3

What are the key variables for design?

The key drivers are a hybrid of physical factors (**operative temp, age, clothing**) and context (**Köppen climate**).

This empirically validates the need for **climate-responsive architecture** and **personalized, occupant-specific controls**.

Practical Implications & Future Impact

Advanced HVAC Control: Models can be integrated into Building Management Systems (BMS) for predictive, proactive control, saving energy and improving satisfaction.

Personalized Comfort Systems (PCS): The model can act as the "brain" for smart vents, desk fans, or heated chairs, adjusting an individual's microclimate automatically.

Informing Architectural Design: Provides strong data-driven evidence that designing buildings in harmony with their local climate is fundamental to achieving comfort and efficiency.

Limitations & Future Work

Limitations.

The 53% Accuracy Ceiling: Comfort is subjective ("a condition of mind"). The model cannot capture unmeasured factors like mood or stress.

Data Quality: Relies on self-reported and estimated values (e.g., clo, met).

Interpretability: Models are still partially "black boxes."

Future Work:

Incorporate real-time physiological data (e.g., skin temperature).

Develop hybrid models that combine physics-based equations with ML.

Test models in live building control systems.

This research successfully developed a machine learning framework that significantly outperforms traditional models in predicting thermal comfort.

The most accurate predictions emerge from combining immediate physical measurements with broad geographical context, validating the principles of both physics-based and adaptive comfort theories.