Smart Home Energy Predictions:

Using IoT and Machine Learning

Aaron Ramirez and Outhai Xayavongsa

Master of Science in Applied Artificial Intelligence, University of San Diego

AAI-530: Data Analytics and the Internet of Things

Professor Anna Marbut

February 24, 2025

Author Note

This work is a collaborative endeavor by Aaron Ramirez and Outhai Xayavongsa. The project demonstrates methods of deploying data-driven insights to facilitate more efficient and sustainable management of residential energy systems. The project code can be accessed at https://github.com/oxayavongsa/aai-530-iot-smart-house.

Abstract

This Internet of Things (IoT) project presents a comprehensive approach to multi-sensor indoor environmental monitoring using two distinct sensors, EcoLab Ground and WeatherLink Indoor, combined into a unified dataset of over 100,000 records. Rigorous preprocessing addressed anomalies such as spurious temperature readings and skewed relative humidity values, while feature engineering incorporated rolling averages and barometric deltas. A long short-term memory (LSTM) model demonstrated robust single-step temperature forecasting, achieving an RMSE near 0.81°C and a MAPE of about 3.4%. Meanwhile, a convolutional neural network (CNN) classified indoor conditions with 94% overall accuracy but faced challenges predicting the rare “unstable” class. Finally, a Seasonal ARIMAX (SARIMAX) model, leveraging humidity, dew point, barometer, and heat index as exogenous variables, excelled at short-term forecasting yet delivered smoother, less precise predictions at longer horizons.

Keywords: Smart Home, Energy Management, IoT, EDA, LSTM, CNN, ARIMA, SARIMAX, Time-Series Forecasting, Feature Engineering

Smart Home Energy Predictions

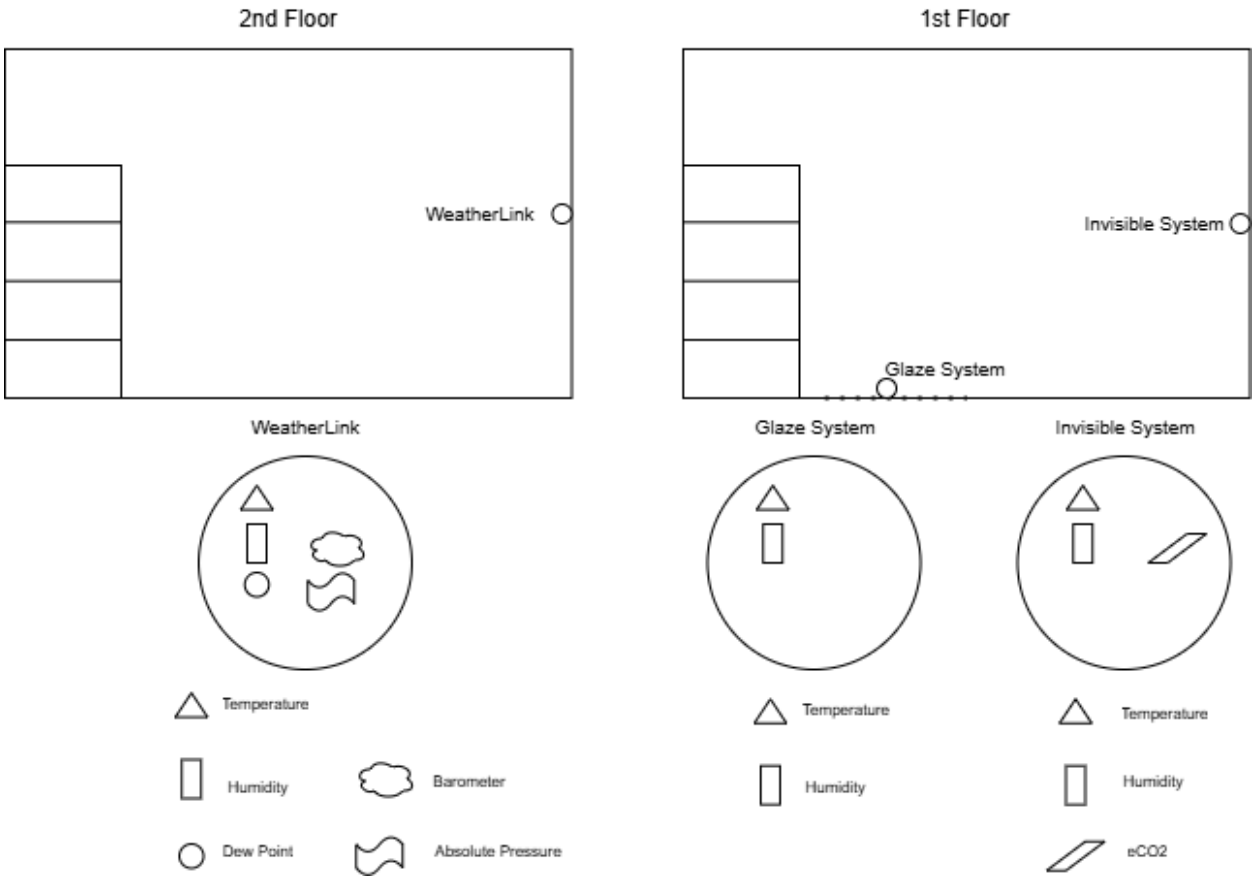
This project develops a targeted data pipeline to forecast and classify indoor conditions using two reliable sensors from the Suffolk Sustainability Institute's Smart House Data Pack (2022–2023). After discarding inconsistent data sources and extreme readings, we consolidated timestamps and performed exploratory data analysis (EDA), revealing correlations among temperature, dew point, and humidity, as well as rolling-average patterns indicative of seasonal fluctuations. Building on these insights, we pursued two modeling objectives: using ARIMA, SARIMAX, and LSTM for temperature forecasting and employing a convolutional neural network (CNN) to classify states (comfortable, humid, dry, unstable). We discuss limitations such as sparse "unstable" samples and sporadic sensor gaps while proposing future expansions, like occupant-aware modeling or extra sensor inputs to bolster real-time, energy-efficient control in modern smart homes.

IoT System

Figure 1 depicts the house layout, with the WeatherLink system upstairs, the Glaze System by the front door, and the Invisible System in the kitchen on the ground floor. Each system communicates over Wi-Fi every 7.5 to 15 minutes via HTTPS. Minimal edge computing discards incomplete readings and calculates heat index values. All data are stored on AWS, using both an S3 bucket and an SQL database, which automatically scales based on sensor output, supporting future machine-learning applications.

Figure 1

IoT Sensor Diagram and Data Architecture



Temperature	Measured using a digital sensor that must avoid direct sunlight to prevent skewed readings
Humidity	A resistive sensor measuring changes in air moisture; can be affected by condensation or airflow.
Dew Point	Derived from a chilled mirror hygrometer.
***Heat Index**	Computed from temperature and humidity, not directly sensed.
Barometer	A MEMS-based sensor for atmospheric pressure; minor altitude effects indoors.
Absolute Pressure	Similar to a Barometer but includes altitude correction.
eCO_2	A non-dispersive infrared sensor detecting CO_2 absorption in IR light.

Cleaning and Preprocessing

Initially, the project utilized three sensors: the front door sensor (Glaze System), the ground floor sensor (Invisible System), and the weather station sensor (WeatherLink). The front door sensor recorded timestamp, serial, name, mode, alarm, relative humidity, and temperature, but only timestamp, humidity, and temperature were relevant. Humidity readings were mostly zero, with occasional spikes, indicating malfunction. Timestamps were sporadic, with only 1-6 readings per day, making trend analysis unfeasible, leading to its removal.

The ground sensor provided more reliable data, recording sensor type, DateTime, Value, Max, Min, and Unit. Max and Min were dropped due to a lack of data. Readings were recorded every 7.5 minutes, sufficient for trend analysis. Data included eCO_2, relative humidity, and temperature in PPM, which were split into separate rows.

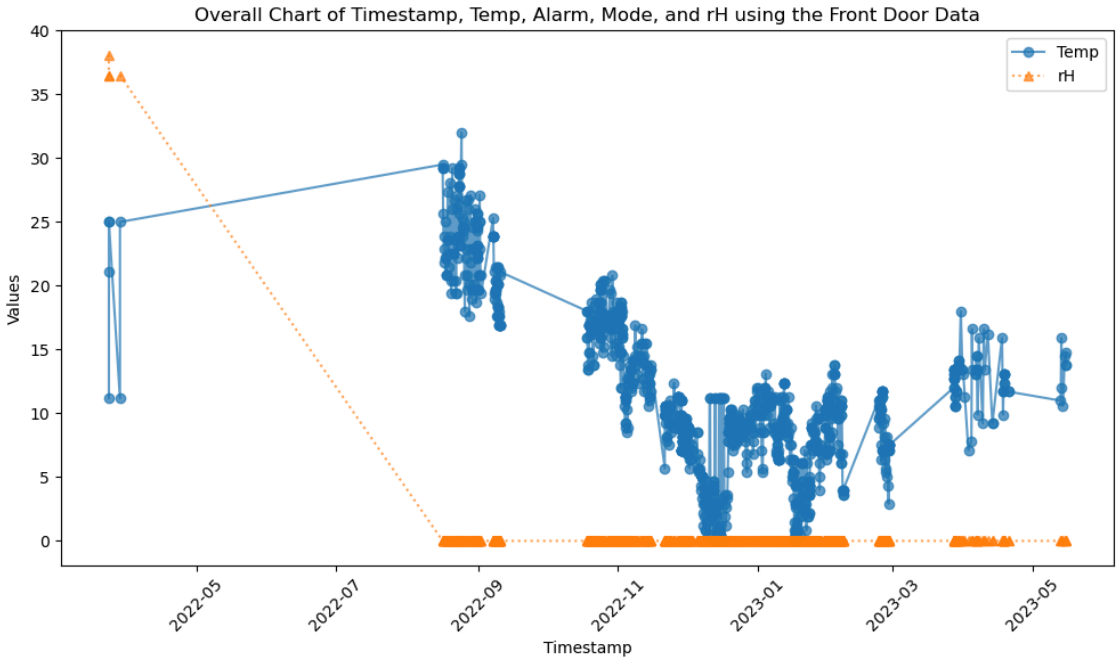
The WeatherLink sensor recorded data every 15 minutes, capturing inside temperature, humidity, dew point, heat index, barometer, and absolute pressure. Highs and lows were removed due to low variability. This dataset was merged with the ground sensor data to create the final dataset. Unrealistic values, such as extreme temperatures, were removed, and missing values were handled through backward and forward fill. Similar columns (e.g., relative humidity, temperature, and DateTime) were consolidated. Feature engineering included computing dew point and heat index, with temperature converted to Fahrenheit where necessary. When the temperature exceeded 80.6°F (~27°C), we used the default temperature.

Exploratory Data Analysis (EDA)

Our first analysis was on the Front Door sensor (GlazeAlarm 2023), we discovered that its relative humidity readings were mostly zeros and contained severe temperature anomalies, undermining data reliability (Figure 2). Consequently, we excluded this sensor from the final modeling and focused on the remaining two sensors.

Figure 2

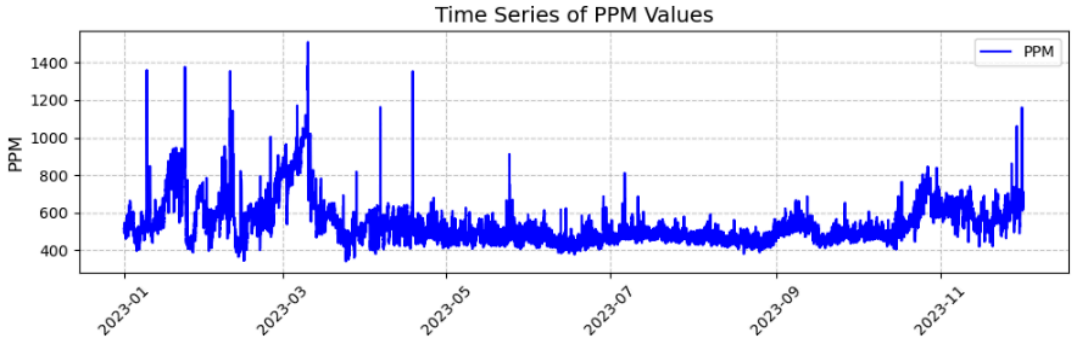
Overall Chart of the Front Door Sensor (Glaze Alarm 2023)



Before cleaning, the EcoLab Ground plots (PPM, %RH, and °C) showed abrupt spikes, particularly in CO_2, and occasional irregularities in humidity and temperature values (Figure 3). Applying thresholds at 3000 ppm, and then linearly interpolating the resulting gaps, revealed a smoother and more coherent time series.

Figure 3

Ecolab Ground Times Series PPM Graph (Invisible Systems 2023)



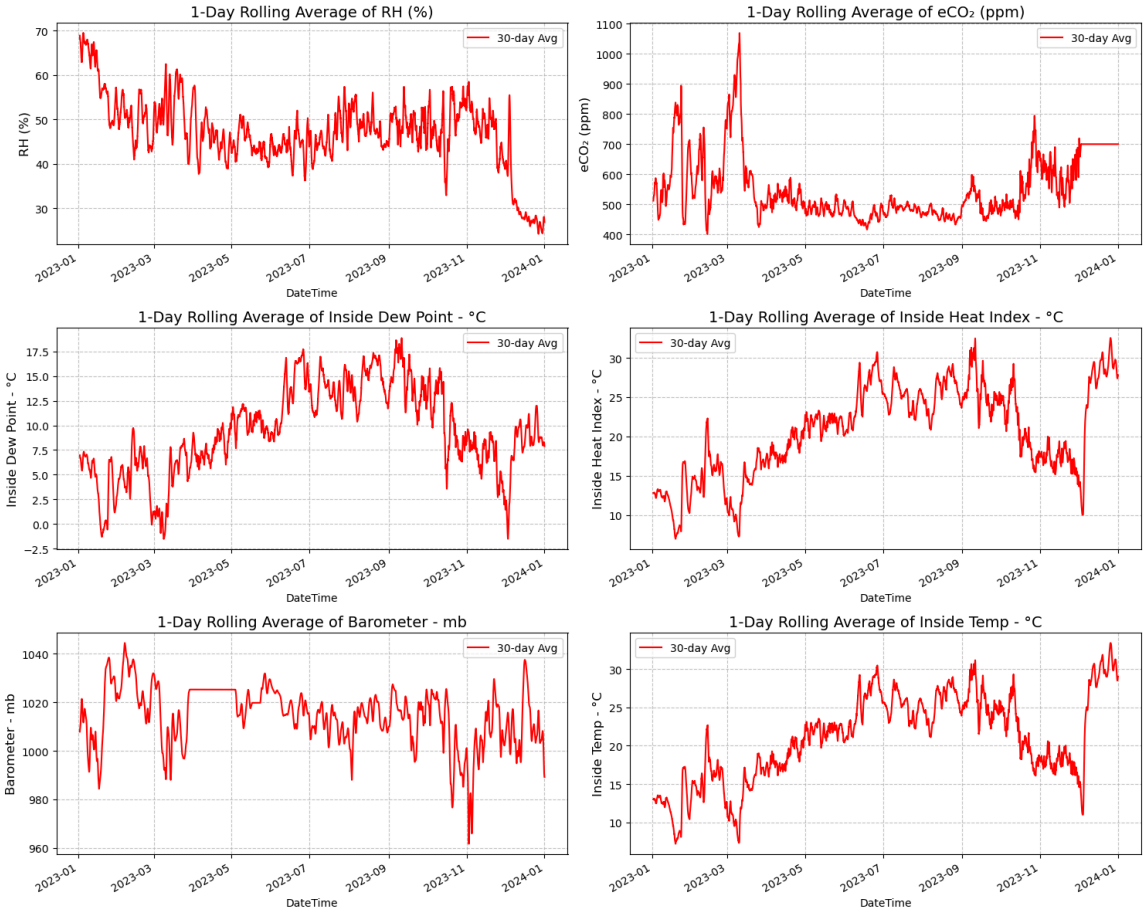
Our final analysis of the WeatherLink Indoor dataset reveals a strong correlation (~0.99) between the Inside Temperature (°C) and the Inside Heat Index (°C), indicating that the heat

index can be inferred from temperature and humidity. Similarly, Barometer (mb) and Absolute Pressure (mb) track so closely that only one may be needed for predictive modeling. Inside Humidity (%) and Inside Dew Point (°C) also show a moderate-to-strong positive correlation due to their shared dependence on moisture content.

The 1-day rolling averages highlight seasonal trends in temperature, heat index, and dew point while revealing an unusual dip in barometric pressure around March (Figure 4). Additionally, eCO_2 (ppm) rises sharply toward December, possibly due to sensor drift or reduced ventilation. These rolling averages provide a clearer view of environmental patterns, helping to differentiate large-scale trends from short-term fluctuations.

Figure 4

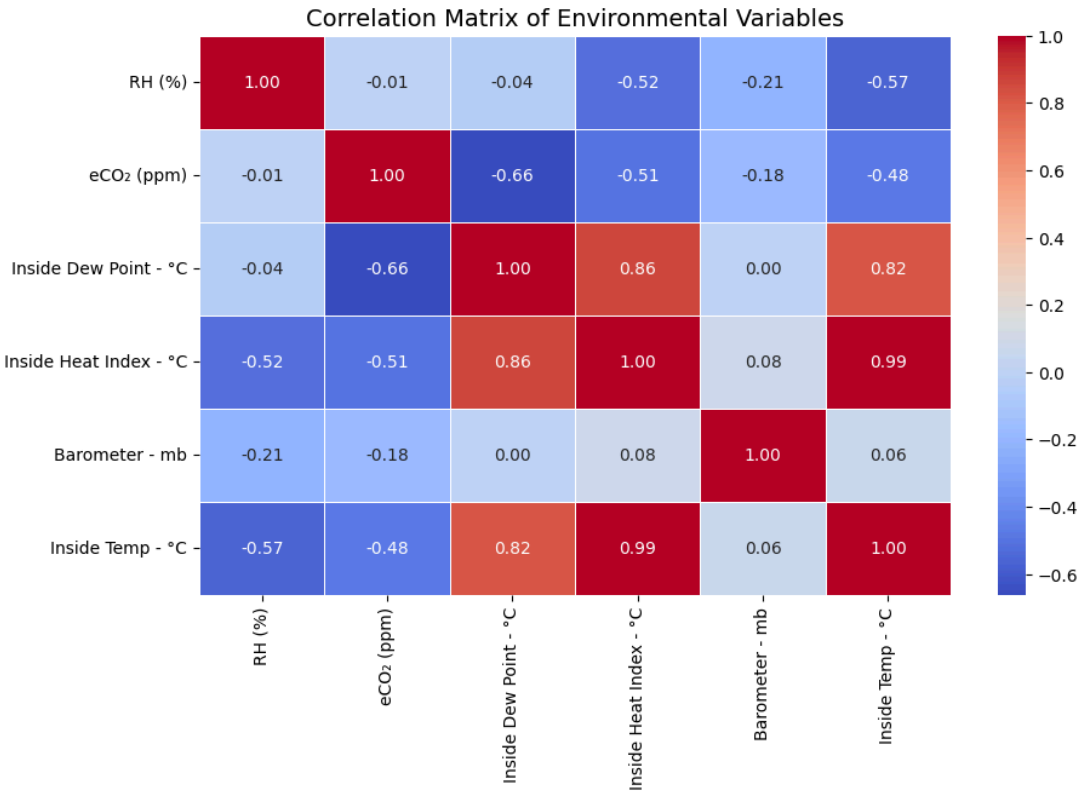
1-Day Rolling Average for Multiple Variables



Using a correlation matrix (Figure 5) to examine our final dataset, we found a *dew point* correlating with *inside temperature* at 0.82 and *heat index* at 0.86, while *inside temperature* and *heat index* exhibited an almost perfect correlation (0.99). A complementary pairwise plot confirmed these relationships. For modeling, we focused on *relative humidity*, *dew point*, *heat index*, and *barometer* due to their moderate-to-strong intercorrelations, and we retained *eCO₂* (ppm) because elevated CO_2 can influence temperature and overall indoor comfort.

Figure 5

Correlation Matrix of Final Variables



Model's Scope and Hyperparameters

For the time series and classification models, Humidity, Dew Point, Barometer, and Heat Index were used to predict temperature, while eCO_2 was excluded due to a lack of correlation. The modeling window spanned June 1, 2023 – November 30, 2023, as data from May and December contained faults in certain parameters. For classification, the dataset was expanded to capture daily and monthly cycles, defining conditions such as comfortable, humid, dry, and unstable. Time itself was not a direct factor in the classification problem.

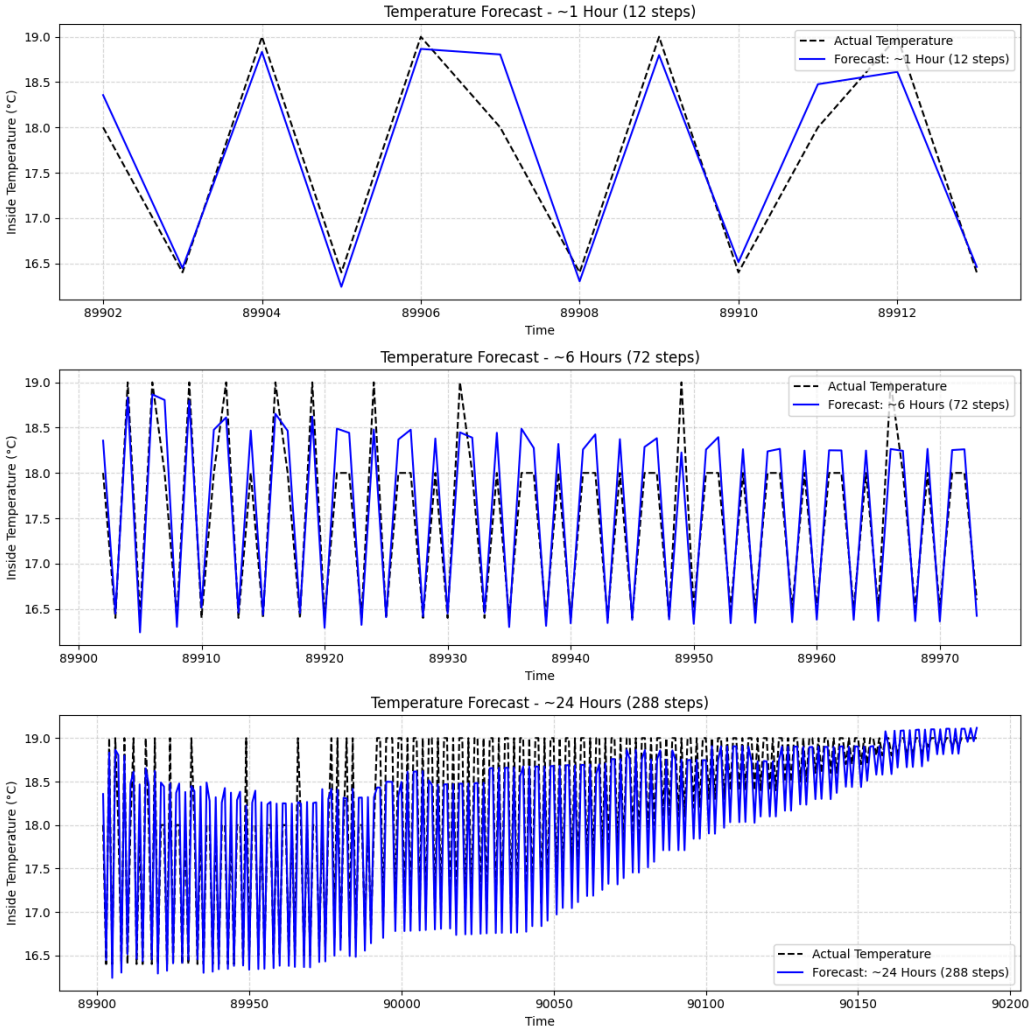
SARIMAX (Seasonal ARIMA with Exogenous Regressor)

SARIMAX was chosen for its ability to model short-term fluctuations, seasonal patterns, and dependencies between past and future values, making it ideal for weather prediction. Parameter selection (p=7, d=1, q=5) was guided by autocorrelation and partial autocorrelation analysis, while humidity, barometer, dew point, and heat index were incorporated as exogenous

variables, improving model efficiency and reducing the Akaike Information Criterion (AIC) from 120,000 to -3,000. The model performed well for short-term forecasts, with performance metrics of RMSE (1.04), MAE (0.82), and R^2 (0.47), and 1-hour and 6-hour predictions closely tracking actual trends. However, accuracy declined at 24-hour forecasts, showing significant deviations, revealing SARIMAX's limitations in extended forecasts due to error accumulation. Figure 6 illustrates that while SARIMAX captures short-term temperature fluctuations effectively, it struggles with long-range predictions, making it most suitable for near-term forecasting.

Figure 6

SARIMAX Prediction vs Actual on a single point in time



Long Short-Term Memory (LSTM)

LSTM was selected for its capacity to learn temporal dependencies in sequential sensor data and proved effective for short-horizon indoor temperature predictions. The same timeframe and data split as other models were used, with the relative heat index, dew point, humidity, barometric pressure, and temperature scaled via MinMax normalization. Input sequences were created by extracting consecutive observations to predict the next temperature reading. The LSTM architecture featured two stacked layers (256 and 128 units) with 50% dropout and L2 regularization, followed by two dense layers. Evaluation yielded RMSE = 0.81°C, MAE = 0.63°C, MAPE = 3.68%, and R^2 = 0.68, indicating strong predictive accuracy. Notably, training loss remained stable (Figure 7), and predictions closely matched actual values (Figure 8), confirming the model's reliability.

Figure 7

Training & Validation Loss for LSTM

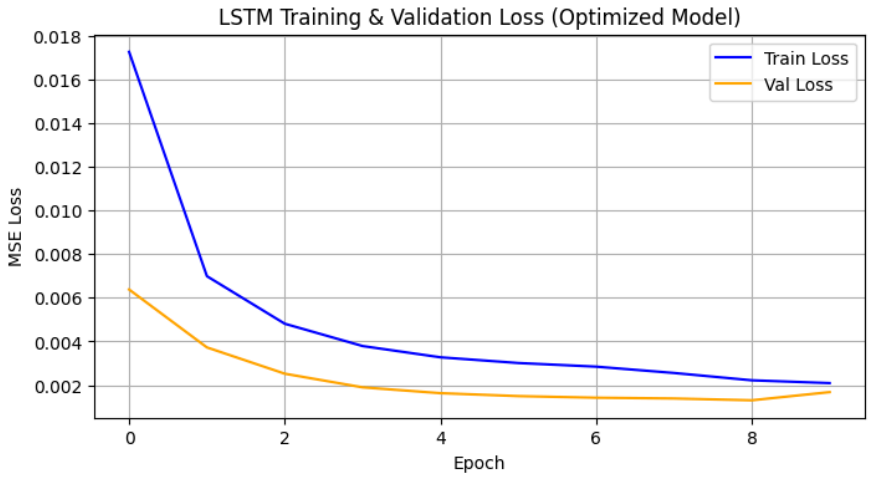
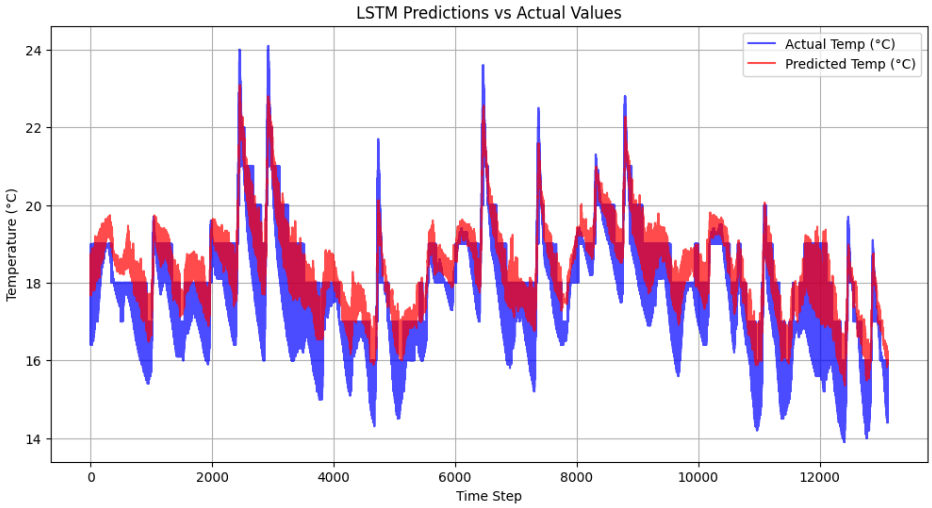


Figure 8

Predictions vs Actual Values for LSTM



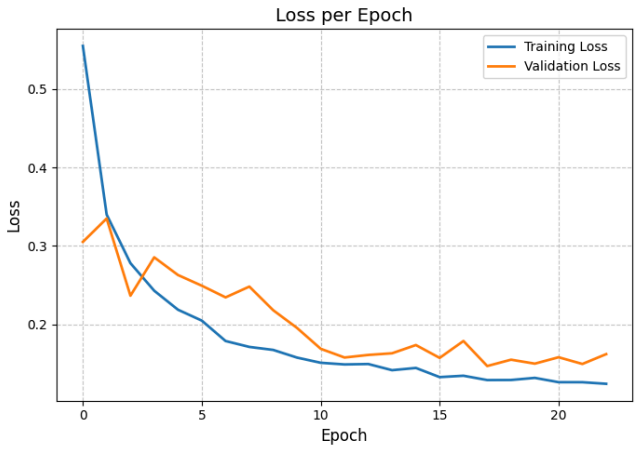
Convolution Neural Network (CNN)

The CNN model was trained on the same dataset but expanded to classify weather conditions and capture cyclic patterns across hours, weeks, and months. Data preprocessing involved removing non-essential columns (hour, week, month, eCO_2, sensor ID, and time), scaling for consistency, and adjusting class weights to balance underrepresented categories. The model architecture included a Conv1D layer with regularization, max pooling, batch normalization, and dropout to mitigate overfitting.

Hyperparameters, such as filter size, kernel size, dropout rate, and regularization strength were optimized using grid search, with early stopping and plateau reduction improving stability. The final model achieved 94% accuracy, learning at a stable rate without rapid loss spikes (Figure 9). However, classification struggled with unstable conditions due to low representation, as reflected in Table 1, where the unstable class had an F1-score of 0.32 compared to significantly higher scores for other categories.

Figure 9

CNN Loss and Accuracy



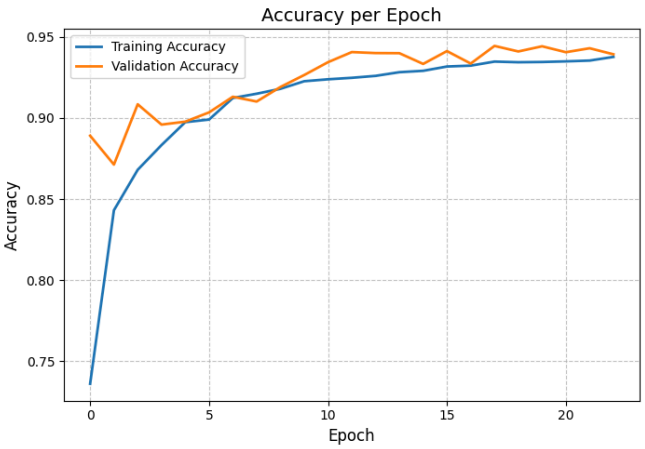


Table 1

CNN Classification Report

	precision	recall	f1-score	support
comfortable	1.00	0.95	0.97	9341
humid condition	0.97	0.99	0.98	3026
dry condition	0.99	0.99	0.99	663
unstable	0.19	0.90	0.32	112
accuracy			0.96	13142
macro avg	0.79	0.96	0.81	13142
weighted avg	0.98	0.96	0.97	13142

Dashboard and Visualization Design Choices

The Smart Home Energy Prediction Dashboard was designed to effectively communicate machine learning model performance in temperature and energy forecasting. The SARIMAX and LSTM models compare predicted vs. actual indoor temperatures, with clear trend lines marking fluctuations over time, allowing for easy evaluation of accuracy. The eCO_2 Levels Over Time and Inside Temperature Trends provide insight into environmental conditions, showing long-term patterns in air quality and temperature variability. The Multi-Class CNN Predictions visualization utilizes jittering to separate overlapping data points, ensuring clarity in classification results. Distinct shapes and colors differentiate predicted vs. true classes, making

misclassifications easily identifiable. The legend was included to improve interpretability, allowing users to quickly associate each symbol with its corresponding class.

Discussion, Limitations, and Comparative Analysis

This study showcases how different machine-learning models can collectively advance IoT-based energy management. LSTM outperformed both ARIMA and SARIMAX at short-horizon temperature forecasting, capturing nuanced temporal patterns with lower error and smoother convergence. ARIMA and SARIMAX, despite offering interpretability and decent performance over a few hours, struggled with longer forecasts. Meanwhile, the CNN classification effectively flagged comfortable, humid, dry, and unstable indoor conditions, but rare “unstable” data constrained precision. These class imbalances, along with sporadic sensor anomalies—such as near-zero humidity readings—illustrate the importance of further data cleaning and threshold tuning. Occupant behavior, building materials, and external climate influences remain outside the current pipeline, suggesting additional sensors or external weather data could refine both forecasting and classification. Nonetheless, combining each model’s strengths—LSTM for near-term predictions, CNN for rapid state detection, and SARIMAX for interpretability.

Conclusion

Overall, this project delivers a robust IoT energy management pipeline, integrating predictive and classification models for real-time insights into indoor conditions. By uniting LSTM’s sequence learning, CNN’s classification accuracy, and ARIMA/SARIMAX’s interpretability, we achieve comprehensive temperature forecasts and actionable condition alerts. These findings underline the tangible benefits of applying advanced machine learning in smart homes, paving the way for further improvements—such as broader sensor coverage, refined outlier handling, and deeper integration with automated HVAC controls—to maximize efficiency and occupant well-being.

Reference

Suffolk Sustainability Institute. (2023). *Smart House Data Pack.*

https://www.kaggle.com/datasets/ssiatuos/smart-house-data-pack