OPSD PowerDesk: Electric Load Forecasting System

Advanced Time Series Analysis for European Electric Load Data

Vivardhan Reddy (Roll No. SE23UCSE006.)
Amogh Raj(Roll No. SE23UCSE022)
Koushik K 3 (Roll No. SE23UCSE096)

Course: **ATA**

November 26, 2025

Abstract

Electricity demand forecasting is critical for power system planning, stable grid operation, and cost-efficient scheduling. In this project, we build a complete machine learning pipeline for day-ahead electric load forecasting using the Open Power System Data (OPSD) time series dataset for three European countries: Austria (AT), Belgium (BE), and Bulgaria (BG). The pipeline combines classical statistical modelling (SARIMA) with deep learning architectures (LSTM, GRU, Vanilla RNN), and evaluates them on a common backtesting framework.

We work with 50,398 hourly observations from 2014–2020, perform an 80/10/10 train/validation/test split, and compare models using Mean Absolute Scaled Error (MASE) and other forecast metrics. The best overall performance is achieved by a GRU model, with an average MASE of 0.73 across the three countries and a best score of 0.41 for Austria. On top of forecasting, we implement anomaly detection using rolling z-scores and CUSUM, followed by a machine learning based anomaly classifier that achieves a PR-AUC of 1.0. Finally, we integrate the results into an interactive Streamlit dashboard that allows real-time visualization, model comparison, and live monitoring with periodic model refitting.

1 Introduction

1.1 Motivation

Electric power systems must continuously balance supply and demand. Overestimating load leads to unnecessary generation and cost, while underestimating can cause instability, blackouts, or expensive balancing actions. Short-term load forecasting (STLF), typically at hourly or day-ahead horizons, is therefore a core problem for utilities, transmission system operators, and energy markets.

Traditional forecasting approaches such as ARIMA/SARIMA exploit temporal structure, but may struggle to capture complex non-linearities, holiday effects, or changing patterns over time. Recent advances in deep learning, especially recurrent neural networks (RNNs) such as LSTM and GRU, offer more flexible function approximators that can learn rich temporal dependencies from large datasets.

1.2 Project Objectives

The main objectives of this project are:

- To build a complete, reproducible pipeline for hour-ahead / day-ahead electric load forecasting using OPSD data for AT, BE, and BG.

- To compare classical SARIMA models with deep learning models (LSTM, GRU, Vanilla RNN) on a common train/validation/test protocol.
- To design and evaluate an anomaly detection module that flags unusual load patterns using both statistical and machine learning techniques.
- To implement a live monitoring and online adaptation framework with periodic refitting of models.
- To expose all components via an interactive dashboard built with Streamlit for easy exploration by non-technical users.

2 Dataset and Problem Setup

2.1 OPSD Dataset

We use the Open Power System Data (OPSD) time series dataset, which provides hourly electricity load for multiple European countries. For this project we focus on the following three:

- **AT** – Austria
- **BE** – Belgium
- **BG** – Bulgaria

The raw dataset contains approximately 50,401 hourly timestamps. After cleaning and filtering we use 50,398 valid hourly observations per country, covering the period from 2014-12-31 to 2020-09-30. The main target variable is the electric load (in MW) at each hour.

2.2 Train/Validation/Test Split

We adopt a chronological split to respect the time ordering:

- **Training set**: 80% (40,319 hours)
- **Validation set**: 10% (5,039 hours)
- **Test set**: 10% (5,040 hours)

All models are trained on the training set, tuned on the validation set, and evaluated on the held-out test set using rolling backtesting.

2.3 Forecasting Task

The forecasting task can be summarized as:

> Given a window of past hourly loads and calendar information, predict the next 24 hours of electricity demand for each country.

For the deep learning models, we use a fixed lookback window (e.g. 168 hours) and predict the next step or multiple steps. For SARIMA, we perform multi-step forecasting directly from the fitted time series model.

3 Methodology

3.1 Project Pipeline Overview

The project is organized into multiple phases:

1. **Data Cleaning and Preprocessing**: validation of raw OPSD data, handling missing values, and creating the final single-index time series.
2. **Exploratory Data Analysis (EDA)**: visualization of trends, seasonality, and stationarity checks using STL, ACF/PACF, and statistical tests.
3. **Model Building**: SARIMA grid search and neural network model design (LSTM, GRU, Vanilla RNN).
4. **Forecasting and Backtesting**: rolling-origin evaluation on the test set with multiple metrics.
5. **Anomaly Detection**: statistical detection (z-score, CUSUM) and ML-based anomaly classification.
6. **Live Monitoring and Online Adaptation**: 3,500-hour simulation with periodic model refits.
7. **Dashboard**: a Streamlit application to visualize all phases and compare models.

3.2 Data Cleaning and Preprocessing

Data cleaning scripts validate date indices, remove invalid timestamps, and ensure a consistent hourly frequency. A single time index is kept for all three countries, and missing values are handled using interpolation or forward-filling, depending on context.

For the neural network models, we normalize the load series using training-set statistics and construct sliding windows:

- Input: past L hours of load (e.g. $L = 168$).
- Target: next hour (or next 24 hours) of load.

3.3 SARIMA Modelling

Seasonal ARIMA (SARIMA) models are built per country using a grid search over $p, d, q \in \{0, 1, 2\}$ and seasonal parameters $P, D, Q \in \{0, 1\}$ with period $s = 24$ hours. Model selection is performed using information criteria (AIC/BIC) and validation MSE.

The final selected orders (by BIC) are:

- AT: SARIMA(2,0,2)(1,1,1,24)
- BE: SARIMA(2,0,2)(1,1,1,24)
- BG: SARIMA(2,1,2)(1,1,1,24)

3.4 Neural Network Models

We implement three recurrent neural network architectures using PyTorch:

- **LSTM**: 2 layers, 128 units, lookback window of 168 hours, trained with early stopping on validation loss.

- **GRU**: 2 layers, 64 units, batch size 16, with approximately 159k trainable parameters.
- **Vanilla RNN**: 3 layers, 128 units, batch size 32.

All models are trained with GPU acceleration (NVIDIA RTX 4050) and use standard optimizers (e.g. Adam) with learning rate scheduling and early stopping.

3.5 Anomaly Detection

Anomalies are defined as unusual deviations between actual and predicted load. We apply a two-stage approach:

1. **Statistical Detection**:
   - Rolling z-score with a 336-hour window and threshold $|z| \geq 3$.
   - CUSUM (Cumulative Sum) to detect sustained drifts (parameters $k = 0.5$, $h = 5.0$).

   On the last 1000 hours, this yields 21 z-score anomalies across all countries (about 0.72% of points).
2. **ML-based Classifier**: We construct a small labelled dataset (153 samples) with engineered features: lag values, rolling statistics, calendar features, and CUSUM-based measures. A Logistic Regression and a LightGBM classifier are trained to predict anomaly vs normal, achieving PR-AUC of 1.0 and F1-score of 1.0 at high precision.

3.6 Live Monitoring and Online Adaptation

To simulate deployment, we run a 3,500-hour (146 days) live monitoring experiment. Every 336 hours (two weeks) we:

- Generate forecasts using the current models.
- Update performance metrics on the new data.
- Optionally refit SARIMA models using an expanding training window with at least 1,440 hours of history.

This leads to 10 refits per country and demonstrates how forecasting quality evolves over time under live conditions.

3.7 Interactive Dashboard

All results are integrated into a Streamlit dashboard with:

- Comparison plots of actual vs. predicted load for each model and country.
- Error metric summaries over the test period.
- Anomaly detection visualizations.
- Live monitoring views showing how metrics change across refits.

The dashboard is launched with:

```
python -m streamlit run dashboard.py
```

and is accessible locally at `http://localhost:8501`.

4 Results and Discussion

4.1 Forecasting Performance

Table 1 summarizes the MASE scores on the test set for each model and country.

Table 1: MASE scores for each model and country. Lower is better.

Model	AT (MASE)	BE (MASE)	BG (MASE)	Avg. MASE
SARIMA	0.96	0.96	**0.85**	0.92
LSTM	0.67	0.63	1.25	0.85
Vanilla RNN	0.46	**0.69**	1.11	0.75
GRU	**0.41**	0.95	0.82	**0.73**

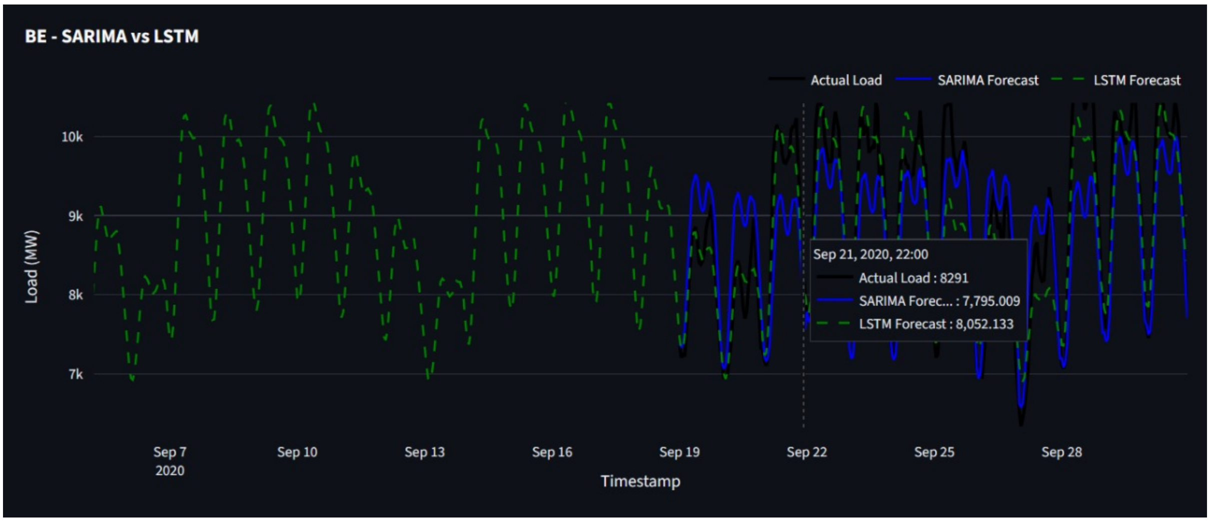

Figure 1: BE: SARIMA vs LSTM forecasting comparison showing actual load vs predicted values over the September 2020 period.

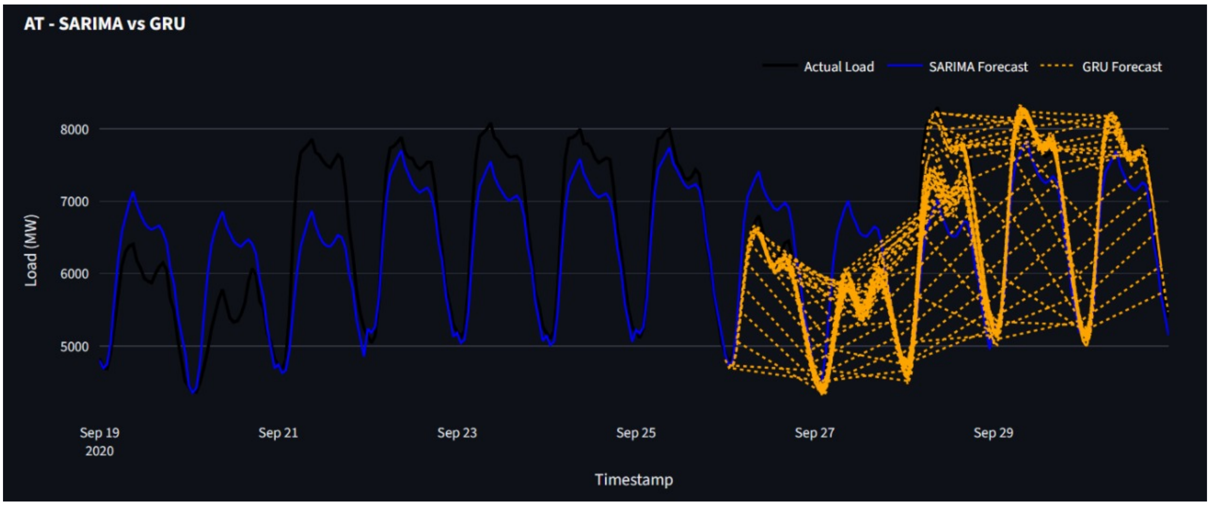

Figure 2: AT: SARIMA vs GRU forecasting comparison showing the performance difference on Austrian load data.

Key observations:

- The **GRU** model achieves the best average MASE (0.73) across AT, BE, and BG, and the best single score (0.41) for Austria.
- **Vanilla RNN** performs competitively, especially for Belgium (MASE 0.69).
- **LSTM** is slightly worse on average but still clearly improves over the SARIMA baseline.
- **SARIMA** remains a strong classical baseline, particularly for Bulgaria (MASE 0.85), confirming that traditional models can still be effective on relatively regular series.

4.2 Statistical Metrics

Beyond MASE, the project evaluates:

- Mean Absolute Percentage Error (MAPE)
- Symmetric MAPE (sMAPE)
- Root Mean Squared Error (RMSE)
- Prediction Interval (PI) coverage for SARIMA

For example, SARIMA on BG achieves a MAPE of 2.95% and 86% PI coverage, indicating a good balance between accuracy and calibrated uncertainty estimates.

4.3 Anomaly Detection Results

On the last 1,000 hours of data:

- 21 anomalous points are detected using rolling z-score across all countries: 7 in AT (0.72%), 2 in BE (0.20%), and 12 in BG (1.23%).
- CUSUM identifies around 63–65% of points as part of a drift regime, highlighting slow shifts rather than spikes.
- The ML-based anomaly classifier, trained on 153 samples (3 positive, 150 negative), achieves a PR-AUC of 1.0 and F1-score of 1.0 at high precision, with the absolute current z-score being the most important feature.

4.4 Live Monitoring Performance

During the 3,500-hour live simulation:

- SARIMA models are refitted 10 times per country, with expanding windows.
- Typical MAPE values in live mode are around 10–16% depending on the country (e.g. AT: 15.73%, BE: 10.16%, BG: 14.41%).
- This experiment shows that even well-tuned models may degrade over time and benefit from periodic retraining as conditions change.

5 Conclusion

Overall, the project shows that deep learning models, especially GRUs, can significantly improve short-term load forecasting while traditional SARIMA models still provide strong and interpretable baselines. The anomaly detection and live monitoring modules add practical value for real-world deployment.

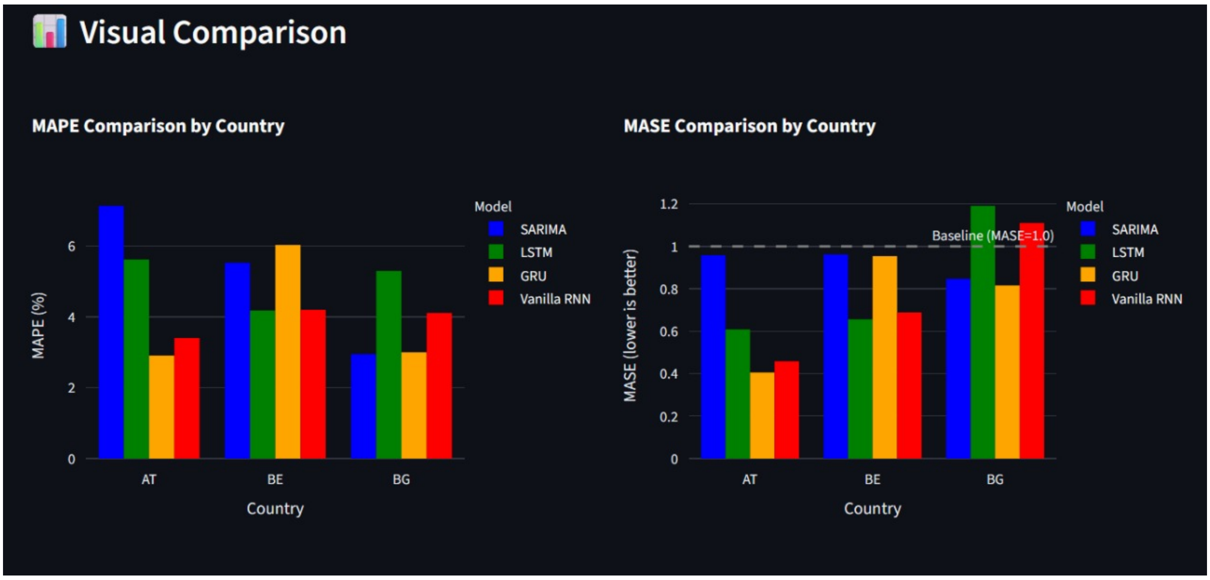

Figure 3: Visual comparison of models across countries using MAPE and MASE metrics. GRU and Vanilla RNN outperform SARIMA in most scenarios.

Code and Reproducibility: All scripts for data cleaning, model training, forecasting, anomaly detection, and live monitoring are organized under the `src/` directory, and all outputs are saved in structured `results/` and `outputs/` folders. The project can be reproduced by running the pipeline steps in order as documented in the README.

References

[1] Open Power System Data. Time series for Germany and other European countries, https://open-power-system-data.org/.

[2] PyTorch: An open source machine learning framework, https://pytorch.org/.

[3] Statsmodels: Statistical modelling in Python, https://www.statsmodels.org/.