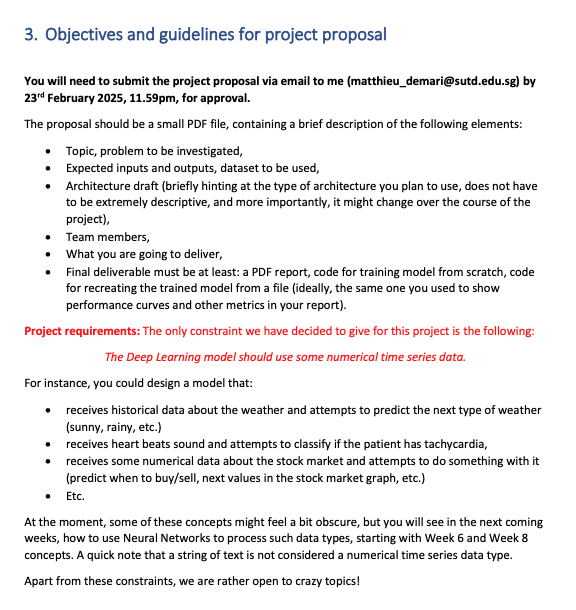


Project Proposal

**50.039 Deep Learning**

**Group 09**

|  |  |
| --- | --- |
| **Student ID** | **Full name** |
| 1006232 | Benetta Cheng Jia Wen |
| 1005935 | Ong Shao Ren |
| 1005242 | Tran Cong Nam Anh |



**Project Proposal**

## Problem statement:

Weather forecasting has long been an essential element of urban planning, agriculture, public health, disaster management, and public transportation management. It is even more crucial for Delhi, one of the most populous cities in the world, as accurate weather predictions enable individuals, organizations, and city authorities to optimize resource allocation, develop comprehensive plans for extreme weather conditions, and mitigate all sorts of risks.

However, achieving a high level of accuracy in weather forecasting remains a significant challenge due to the inherently dynamic nature of weather conditions and the complex interactions between various factors, including both natural phenomena and human activities.

## Proposed solution:

To tackle this problem, we will train deep neural networks using the daily climate time series dataset from January 1, 2013, to April 24, 2017, with four features: average temperature, humidity, wind speed, and average air pressure. After training, we will predict weather conditions for these four variables over the following months and compare the predictions with real recorded data to evaluate the accuracy and performance of our deep learning model.

We hope that our trained model will achieve accuracy within an acceptable range and become a reliable weather prediction tool for households, organizations, and authorities in Delhi.

## Understanding Dataset:

### 3.1 Expected Inputs:

The dataset consists of time series data of daily weather records spanning from January 1, 2013, to April 24, 2017, for Delhi, India. The dataset also includes four key variables as below:

* Average Temperature (meantemp): represents the mean daily temperature;
* Humidity: indicates the daily average relative humidity;
* Wind Speed: measures the average wind speed;
* Average Air Pressure (mean pressure): measure the average air pressure throughout the day at Delhi.

Additionally, a Time Series component is included as a separate column, associating each observation with a specific date to capture temporal patterns, seasonality, and trends.

### 3.2 Expected Outputs:

* Predicted Weather Variables:
  + Forecasted Average Temperature
  + Forecasted Humidity
  + Forecasted Wind Speed
  + Forecasted Mean Air Pressure
* Prediction Horizon: The outputs will be generated over a specified future period. For instance, daily forecasts for the upcoming month. To facilitate evaluation against actual recorded data.
* Evaluation Metrics: The performance of the model will be assessed using regression-based metrics. For instance, Mean Absolute Error and Root Mean Squared Error to ensure that predictions are within an acceptable accuracy range.

### 3.3 Source of Dataset:

The project will leverage a Kaggle dataset that contains the daily climate for Delhi, consisting of two CSV files: DailyDelhiClimateTrain.csv and DailyDelhiClimateTest.csv <https://www.kaggle.com/datasets/sumanthvrao/daily-climate-time-series-data>

## **Architecture**:

### 4.1 Data Preprocessing

* Handling Missing Values: Impute missing values using statistical methods (e.g., mean or interpolation).
* Feature Engineering:
  + Temporal Features: Day, Month, Year as cyclical features using sine and cosine transformations.
  + Lagged Features: Previous days’ climate values as input features.
  + Moving Averages: Rolling means for smoothing fluctuations.
  + Normalization: Scale the data using Min-Max normalization or Standardization.

### 4.2 Model Architecture

* Base Model: Recurrent Neural Network (RNN) variants for time series forecasting.

Proposed Models:

**LSTM (Long Short-Term Memory)**

* Input layer will take in sequences of historical climate data and capture the temporal patterns. We think LSTM is promising since LSTM units are good for learning long-term dependencies which is important to predict climate patterns.
* 2 stacked LSTM layers to capture long-term dependencies.
* Can implement dropout layers to prevent overfitting by randomly ignore fraction of units during training to encourage better generalization.
* Dense Output Layer: Output from last LSTM layer will be passed through this fully connected layer for multi-variable output.
* Activation functions we have in mind currently are tanh and sigmoid since they are typically used in LSTM layers. Dense output will use linear activation function.

**GRU (Gated Recurrent Unit):**

* Similar to LSTM but with GRU layers for faster training and reduced complexity.
* First GRU layer takes sequence of climate data and returns sequences.
* Second GRU layer refines predictions.
* Dropout layer can also be implemented for same reason.

**Transformer Model:**

* Self-attention mechanism for capturing global dependencies in the time series data.
* Will have the self-attention layer, positional encoding, multi-head attention and feedforward network + output layer

**Hybrid CNN-LSTM Model:**

* + CNN Layers: For feature extraction and pattern recognition in time series.
  + LSTM Layers: To capture sequential dependencies.

## Team Member:

|  |  |
| --- | --- |
| **Student ID** | **Full name** |
| 1006232 | Benetta Cheng Jia Wen |
| 1005935 | Ong Shao Ren |
| 1005242 | Tran Cong Nam Anh |

## Deliverables:

* ​​A GitHub repository containing our data preprocessing, data cleaning, and model training code, along with relevant notebooks/files.
* A comprehensive PDF report summarizing the technical details, algorithms, model architecture, and performance results.
* Instructions on how to recreate our trained PyTorch model using the saved PyTorch weights and recreate the performance metrics claimed in the PDF report.

Brainstorm - delete later

Louis proposed dataset:

<https://www.kaggle.com/datasets/sumanthvrao/daily-climate-time-series-data>

https://www.kaggle.com/datasets/htagholdings/property-sales

https://www.kaggle.com/datasets/marquis03/heartbeat-classification

PC components classification:

<https://www.kaggle.com/datasets/asaniczka/pc-parts-images-dataset-classification>

ECG Heartbeat Categorization Dataset

<https://www.kaggle.com/datasets/shayanfazeli/heartbeat>

Weather Classification

<https://www.kaggle.com/datasets?search=class>

Drug Classification

<https://www.kaggle.com/datasets/prathamtripathi/drug-classification>

Customer Segmentation

<https://www.kaggle.com/datasets/kaushiksuresh147/customer-segmentation>

HeartBeat

<https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset>

**Financial/Stock Market Time-Series**

Stock Market Data - Nasdaq, S&P 500, Dow Jones

<https://www.kaggle.com/datasets/szrlee/stock-time-series-20050101-to-20171231>

• Contains stock prices from 2005-2017.

• Can be used for stock price prediction, market trend analysis, or anomaly detection.

Bitcoin Historical Data

https://www.kaggle.com/datasets/sudalairajkumar/cryptocurrencypricehistory

• Price and volume data for various cryptocurrencies.

• Can be used for crypto trading strategy models or volatility prediction.

**Health/Medical Time-Series**

ICU Patient Vital Signs Monitoring

https://www.kaggle.com/datasets/marcopeix/icu-patient-vital-signs

• Contains real ICU patient data, including heart rate, respiration rate, and blood pressure.

• Can be used for predicting patient deterioration or ICU outcome forecasting.

**Weather & Environmental Data**

Historical Weather Data for 1000+ Cities

https://www.kaggle.com/datasets/selfishgene/historical-hourly-weather-data

• Hourly temperature, humidity, wind speed, etc.

• Can be used for weather prediction models or climate anomaly detection.

Air Quality Data from the U.S.

https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india

• Pollution levels over time in different cities.

• Can be used for forecasting air pollution trends.

**Industrial & Anomaly Detection**

NASA Bearing Failure Prediction Data

https://www.kaggle.com/datasets/wkirgsn/bearing-dataset

• Sensor data from rotating machinery.

• Can be used for predictive maintenance and anomaly detection.

Electric Power Consumption Dataset

https://www.kaggle.com/datasets/uciml/electric-power-consumption-data-set

• Measures household power consumption in different categories.

• Can be used for energy consumption forecasting.

**Ben**

1. Predict S&P value and/or stock prices.: https://www.kaggle.com/datasets/andrewmvd/sp-500-stocks

**Technical documentation**

First architecture: Ben

NN with history observations

normal RNN with hidden vector

transformer

data visualization for your model performance

save weight of your best model performance

Second architecture: Louis

Plain LSTM

non-autoregressive LSTM

auto-regressive LSTM

data visualization for your model performance

save weight of your best model performance

Third architecture: Shaoren

plain GRU

Non-autoregressive GRU

auto-regressive GRU

data visualization for your model performance

save weight of your best model performance

Forth architecture: who finish the work first can take this

Compare to legacy or pretrained-large RNN models

Performance comparison cross models: Shaoren

Report: each person write a report for your respective work

Optional - Application to demo difference model - Louis

Create github repo: Louis