National College of Ireland

Configuration Manual

MSc Research Project
Data Analytics

Rishabh Mishra
Student ID: x23350717

School of Computing
National College of Ireland

Supervisor: Harshani Nagahamulla

Declaration

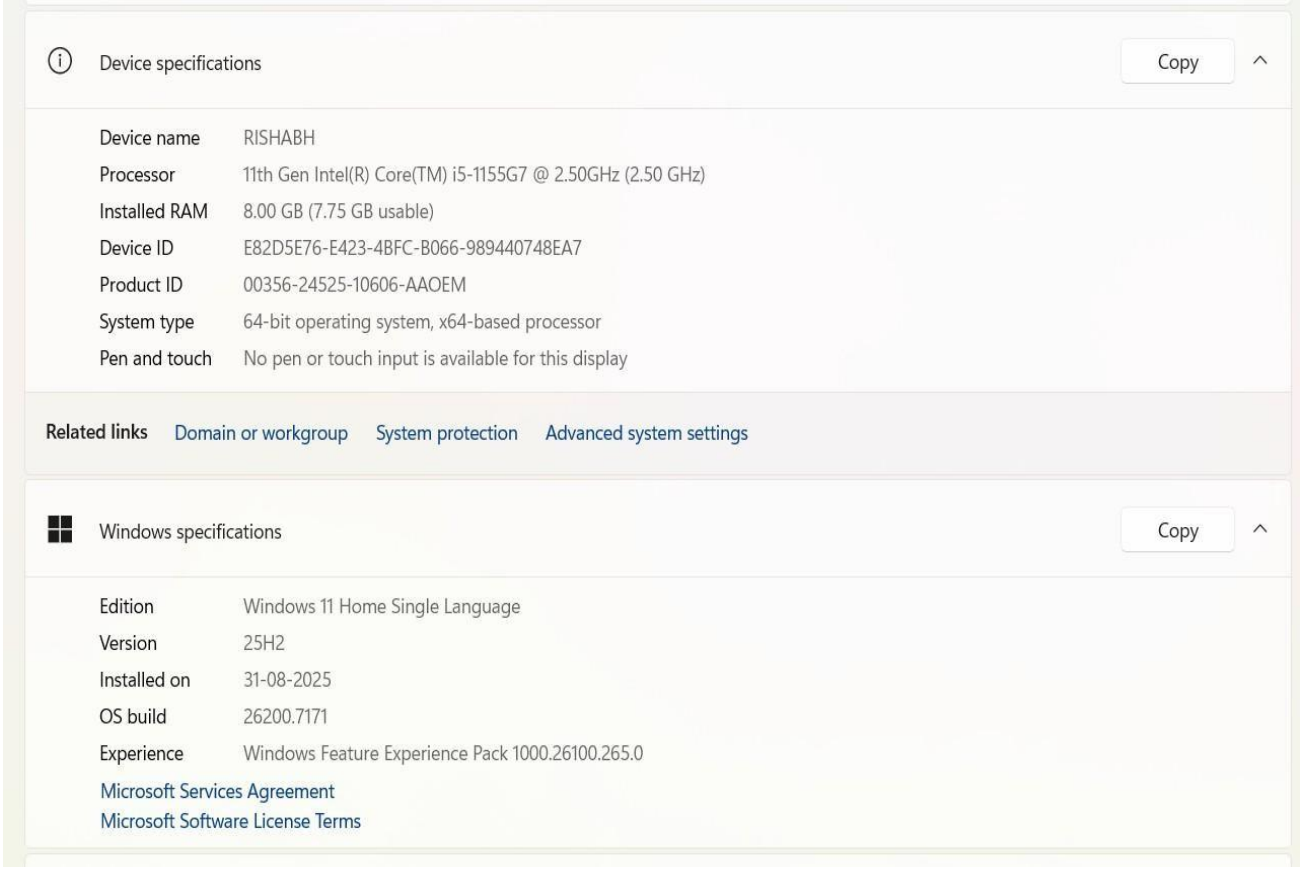
I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of *MSc Research Project in Data Analytics*, is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signature: Rishabh Mishra
Date: 26-01-2026
Student ID: x23350717

Configuration Manual

Rishabh Mishra
x23350717

1 Specifications of Hardware



Hereby, the total hardware specifications for the study conducted to predict the PM2.5 air quality. The computer is equipped with an Intel Core i5 of the 11th Generation, 8GB RAM, and Windows 11 Home as the operating system. The configuration of this hardware allows training of deep learning models due to its computational power.

2 Specification of software

Software	Version
Operting System	Windows 11 Home
Python	3.11.5
Anaconda	Installed
Deep Learning	
TensorFlow	2.19.0
Py-Torch	2.7.1+cpu
Keras	3.9.2
Data Science	
NumPy	1.26.4
Pandas	1.5.3
Matplotlib	3.7.2
Seaborn	0.12.2
Scikit-learn	1.6.1
XG-Boost	3.0.0
Deep Learning Tools	
Keras-Tuner	1.4.8
Tensor-Board	2.19.0
NLP Libraries	
NLTK	3.8.1
Transformers	4.32.1
Gensim	4.3.0
System Tools	
Git	2.40.1
Node.js	18.12.1
Total Packages	550

The table above shows the collection of software versions and packages that were installed for the project development. The environment includes Python 3.11.5 with TensorFlow 2.19.0 and Keras 3.9.2 as the main deep learning frameworks. In total, 550 packages were installed to handle tasks related to data processing and visualization in addition to model training.

3 Libraries Import

Importing all the required libraries (NumPy, Pandas, TensorFlow, Scikit-learn etc.) and sets a fixed random seed for the best reproducibility across all runs.

```
import os
import warnings
warnings.filterwarnings('ignore')

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import requests
import time
import joblib
from datetime import datetime
from scipy import stats

import tensorflow as tf
from tensorflow.keras import Sequential
from tensorflow.keras.layers import Dense, LSTM, Conv1D, MaxPooling1D, Flatten, Dropout, BatchNormalization, Bidirectional
from tensorflow.keras.callbacks import EarlyStopping, ReduceLROnPlateau
from tensorflow.keras.optimizers import Adam
import keras_tuner as kt

from sklearn.preprocessing import RobustScaler
from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error
from sklearn.ensemble import RandomForestRegressor
from sklearn.inspection import permutation_importance
import shap
import numpy as np
import matplotlib.pyplot as plt
import joblib
import os
import json
import keras_tuner as kt
from tensorflow.keras.callbacks import EarlyStopping, ReduceLROnPlateau
from keras_tuner import HyperParameters
import time

import shap
import random

RANDOM_SEED = 42
np.random.seed(RANDOM_SEED)
tf.random.set_seed(RANDOM_SEED)
random.seed(RANDOM_SEED)
```

This is a list of all the essential libraries imported at the start of the project with proper configuration. Among the imports are NumPy and Pandas for data manipulation while TensorFlow and Keras are for deep learning with Scikit-learn for preprocessing and evaluation. The random seed is set to 42 to ensure that there is no discrepancy in the results obtained from all experiments.

4 Project Structure

```
FEATURE_COLS = ['temp', 'humidity', 'wind_speed', 'CO', 'NO2', 'O3', 'SO2', 'PM10']

CITIES = {
    'Dublin': {'lat': 53.337758, 'lon': -6.28469, 'weather_file': DUBLIN_WEATHER_FILE},
    'Lahore': {'lat': 31.5620, 'lon': 74.3220, 'weather_file': LAHORE_WEATHER_FILE},
    'Delhi': {'lat': 28.6469, 'lon': 77.3158, 'weather_file': DELHI_WEATHER_FILE}
}

for subdir in ['visualizations', 'data', 'models', 'analysis', 'learning_curves', 'statistics', 'tuner', 'etl', 'tables', 'correlation']:
    os.makedirs(OUTPUT_DIRECTORY + subdir, exist_ok=True)

physical_devices = tf.config.list_physical_devices('GPU')
DEVICE = 'GPU' if len(physical_devices) > 0 else 'CPU'

print("Configuration Summary:")
print(f"TensorFlow version: {tf.__version__}")
print(f"Device: {DEVICE}")
print(f"Random Seed: {RANDOM_SEED}")
print(f"Data Split: 60 Train, 20 Val, 20 Test")
print(f"Sequence Length: {SEQUENCE_LENGTH}")
print(f"Batch Size: {BATCH_SIZE}")
print(f"Max Epochs: {EPOCHS}")
print()
```

```
Configuration Summary:
TensorFlow version: 2.19.0
Device: CPU
Random Seed: 42
Data Split: 60 Train, 20 Val, 20 Test
Sequence Length: 24
Batch Size: 32
Max Epochs: 30
```

The complete project directory organization and file structure throughout the study are presented. The structure incorporates distinct folders for raw data, processed data, saved models, and visualization results. This organized method not only ensures easy navigation but also keeps a clear distinction among various project components.

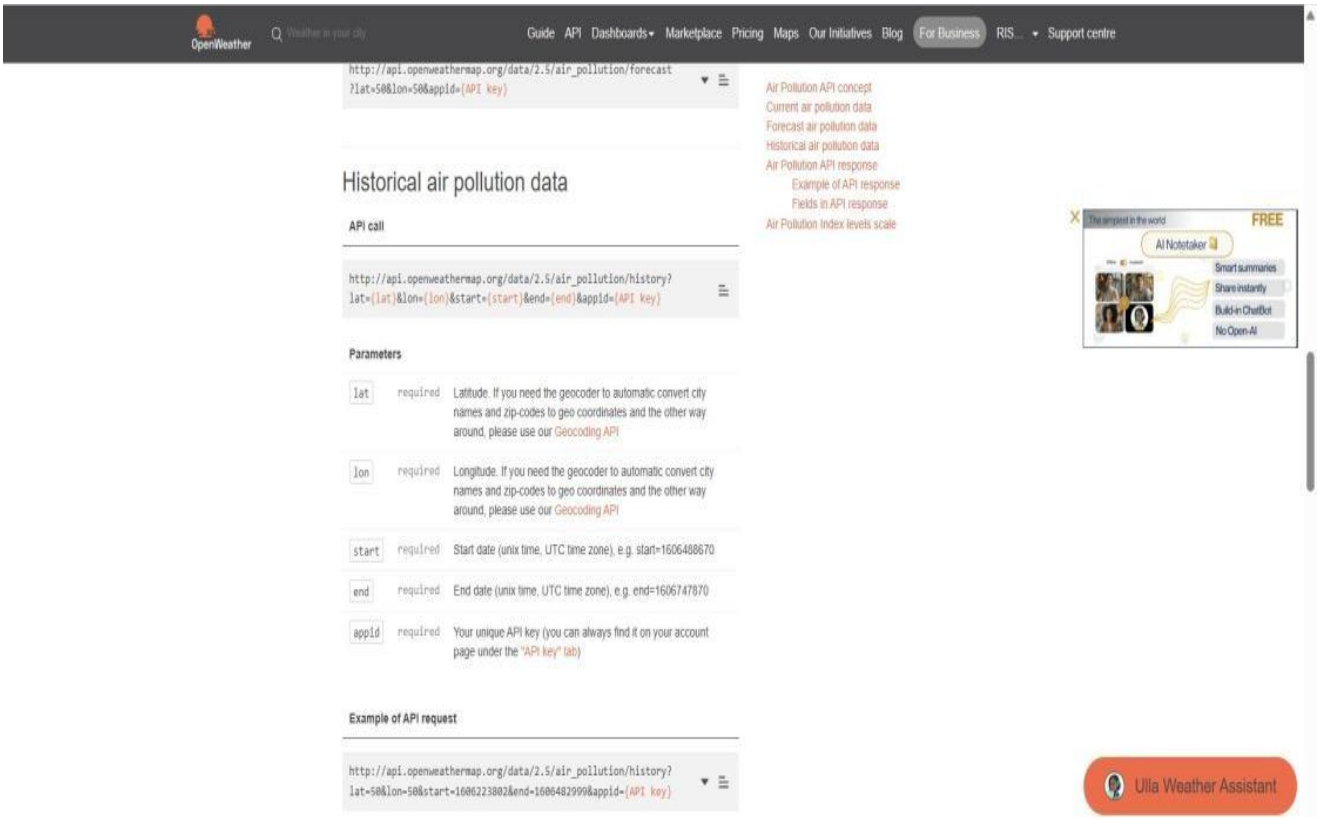
5 Configuration Parameters

```
]:

API_KEY = "9036a2d4a53721afd69878d62ded657d"
START_DATE = "2021-01-01"
END_DATE = "2025-09-27"

DUBLIN_WEATHER_FILE = "C:\\Users\\risha\\Downloads\\temperarure humidity data.csv"
LAHORE_WEATHER_FILE = "C:\\Users\\risha\\Desktop\\research practium\\data\\Samanabad_31_535661_74_298534_1285718400_1759103999_68db1cbf51fde7000792d82d.c
DELHI_WEATHER_FILE = "C:\\Users\\risha\\Desktop\\research practium\\data\\Anand_Vihar_28_653048_77_308242_1285718400_1759103999_68db1cbf51fde7000792d82d.

OUTPUT_DIRECTORY = "pm25_results_corrected/"
SEQUENCE_LENGTH = 24
EPOCHS = 30
BATCH_SIZE = 32
MAX_TRIALS = 10
TUNER_EPOCHS = 8

FEATURE_COLS = ['temp', 'humidity', 'wind_speed', 'CO', 'NO2', 'O3', 'SO2', 'PM10']

CITIES = {
    'Dublin': {'lat': 53.337758, 'lon': -6.28469, 'weather_file': DUBLIN_WEATHER_FILE},
    'Lahore': {'lat': 31.5620, 'lon': 74.3220, 'weather_file': LAHORE_WEATHER_FILE},
    'Delhi': {'lat': 28.6469, 'lon': 77.3158, 'weather_file': DELHI_WEATHER_FILE}
}

for subdir in ['visualizations', 'data', 'models', 'analysis', 'learning_curves', 'statistics', 'tuner', 'etl', 'tables', 'correlation']:
    os.makedirs(OUTPUT_DIRECTORY + subdir, exist_ok=True)

physical_devices = tf.config.list_physical_devices('GPU')
DEVICE = 'GPU' if len(physical_devices) > 0 else 'CPU'

print("Configuration Summary:")
print(f"TensorFlow version: {tf.__version__}")
print(f"Device: {DEVICE}")
print(f"Random Seed: {RANDOM_SEED}")
print(f"Data Split: 60 Train, 20 Val, 20 Test")
print(f"Sequence Length: {SEQUENCE_LENGTH}")
print(f"Batch Size: {BATCH_SIZE}")
print(f"Max Epochs: {EPOCHS}")
print()
```

Here are the feature configuration parameters and global settings for model training. The configuration consists of random seed = 42, sequence length =24.

The study uses 24-hour data with a batch size of 32 and an epoch duration of 30. The study uses temperature and humidity and wind speed and pollutant concentrations which include CO and NO2 and O3 and SO2 as feature columns while PM10 was excluded through VIF analysis.

6 API Configuration



The API documentation that was used for collecting air quality data from three cities is shown here also a link to the website provided below. Historical air pollution data, including PM2.5, PM10, and various pollutant gas concentrations provided by this API. This source of real-time data enables the monitoring and prediction of air pollution levels continuously.

(Link to the website-Weather API - OpenWeatherMap)