Classification of urban morphology and its relationship with air pollution using deep learning

MORTEZA AMOUEI

July, 2023

SUPERVISORS:

Dr. Mahdi Farnaghi
Dr. Frank Ostermann

Classification of urban morphology and its relationship with air pollution using deep learning

MORTEZA AMOUEI
Enschede, The Netherlands, July, 2023

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Geoinformatics

SUPERVISORS:
Dr. Mahdi Farnaghi
Dr. Frank Ostermann

THESIS ASSESSMENT BOARD:
Prof. Dr. Menno-Jan Kraak (Chair)
Dr. Nina Schwarz (External Examiner, ITC Faculty, Department of PGM)

DISCLAIMER

This document describes work undertaken as part of a program of study at the Faculty of Geo-Information Science and Earth Observation of the University of Twente. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the Faculty.

ABSTRACT

Air pollution poses a significant threat to public health and the environment, making understanding its causes and implications essential. Several factors, rooted in the underlying processes in urban areas, affect air quality. Urban forms, the subject of study in urban morphology, impact the dispersion of pollutants. Understanding the interactions between urban forms and air quality is crucial for effective urban development and environmental management.

The main problem of this research is the limited access to vector data for urban form studies on air pollution and the lack of consideration for combined measurements, creating different urban patterns. To address this issue, This research aims to analyze, model, and develop the relationship between urban forms and PM2.5 concentration using a deep learning-based model with scene-based comprehension to capture complex interactions applied to earth observation data. The Local Climate Zones (LCZ) framework, a standardized classification system for urban form, is selected for this research. The research objectives include developing an accurate deep learning model for LCZ classification, and training a suitable model to represent the impact of LCZ on PM2.5 distribution, followed by analyzing the sensitivity and feature importance of different LCZ categories.

The study presents a two-stage framework that classifies local climate zones (LCZ) using three supervised convolutional neural networks models, namely the designed CNN by the author, ResNet-50, and EfficientNet models in the first stage and predicts PM2.5 concentration through the regression task of both XGBoost and LSTM models. The methodology involves data acquisition, preparation, and modeling using Sentinel-2 imagery, PM2.5 measurements, meteorological data, and traffic data. The period of the temporal data covers the hourly values between 2021 and 2022. A noteworthy aspect of this research involves citizen science data for air pollution.

The results demonstrate the efficacy of the ResNet-50 model for LCZ classification with an overall accuracy of 87 percent and the LSTM model for PM2.5 prediction, with the R-squared of 0.75 on unseen data. The sensitivity analysis highlights the positive contribution of LCZ to PM2.5 prediction, and the feature importance analysis reveals the varying contributions of different urban form categories, with the significance of the open-highrise type as the most contributor.

Overall, this research provides insights into the relationship between urban morphology and air pollution, facilitating informed urban development decisions and environmental planning.

Keywords: urban morphology, urban form, local climate zones, air pollution, PM2.5 concentration, deep learning, convolutional neural networks, Earth observation data, citizen science.

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to the people who have contributed significantly to the completion of this research. This endeavor would not have been possible without their guidance, support, and encouragement.

First, I would like to extend my deepest appreciation to my first supervisor, Dr. Mahdi Farnaghi. His profound knowledge and invaluable insights have been instrumental throughout this research journey. His guidance and expertise greatly shaped my understanding of the subject and played a crucial role in directing this research. Significantly, his support has pushed me beyond my expectations, building my confidence in coding and data science tasks.

I would also like to express my sincere gratitude to my second supervisor, Dr. Frank Ostermann. His valuable feedback, critical mindset, and constructive suggestions have greatly enhanced the quality of this research. I am grateful for his time, commitment, and mentorship throughout this research.

Lastly, I am grateful to my beloved wife, Kimiya, for her unwavering support and belief in my abilities. Her encouragement and patience have been a constant source of motivation. I am genuinely appreciative of her love, understanding, and sacrifices.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Background

Air pollution is an important phenomenon that can negatively affect public health. Chronic exposure to air pollutants in the environment can result in severe problems for people, e.g., asthma and lung disorders (Li et al., 2019). Moreover, air pollution can cause serious environmental issues, namely acid rain, smog, and haze. Particulate matter, one of the significant contaminants in the air, can result in rainfall shortage due to changes in cloud properties. Plants and crop production can also be affected in air-polluted areas. On a global level, the dramatic increase in the emission of air pollutants contributes to climate change. (Saxena & Srivastava, 2020). These show the importance of good air quality for humans and nature. Thus, scientific communities must investigate air pollution models and identify the influencing factors that increase air pollutants (Speak et al., 2012, as cited in Saxena & Srivastava, 2020).

Human-driven activities that have many manifestations in urban areas exacerbate air pollutant emissions (Speak et al., 2012, as cited in Saxena & Srivastava, 2020). Air quality in cities is negatively impacted by increasing urbanization directly and indirectly. The population growth in urban areas directly worsens air quality through increases in emissions by vehicles and industrial activities. In contrast, different combinations of urban forms, e.g., open spaces and high-rise building blocks, indirectly affect the dispersion of pollutants (Kim & Gim, 2022). Replacing natural landscapes with large areas of roads, buildings, and urban landscapes, intensified by rapid urbanization, is correlated with poor air quality (Yang et al., 2022). The complex urban elements, e.g., land use, transportation, and infrastructure, can considerably affect health by reducing air quality (Ahn et al., 2022). Moreover, Microclimate and air quality are both consequences and prerequisites of urban planning and design (Yuan et al., 2014). Different characteristics of urban forms, such as high-rise urban blocks, street networks, and open spaces also have a direct impact on wind velocity and circulation, which contribute to the accumulation or dispersion of air contaminants (Yang et al., 2020). Particulate matter (PM), an air pollutant commonly observed in industrial and urban settings, is characterized by a complex mixture of solid and liquid particles that remain suspended in the air. This pollutant is divided into three categories for measuring air quality depending on its aerodynamic diameter. PM10 is a particle with a diameter of less than ten µm, and PM2.5, known as fine particulate matter, includes particles with a diameter of fewer than 2.5 µm. The third category is ultrafine particles, defined as a particle with less than 0.2 µm diameter, which has the worst effect on health (Saxena & Srivastava, 2020). The data availability and the high correlation of the PM2.5 index in urban environments led to choose this pollutant as the representative for air pollution in this research.

Several factors, rooted in the underlying processes in urban areas, affect air quality. It is necessary to find measurable proxies to understand them and discover their effect on different problems, such as air pollution. Urban forms, the subject of study in urban morphology, provide such proxies (Barke, 2018). Chen (2014) defines urban morphology as "the study of urban form that focuses on the formation and transformation of urban forms of cities, towns, and villages over time; their spatial patterns at different scales; and physical characteristics to inform appropriate urban interventions to promote sustainable urban development" (Chen, 2014). Another definition states that "urban morphology is the study of physical forms of human settlements, i.e., how the cities, towns, or villages used to be and how it has physically transformed over the years" (Aslam & Rana, 2022). Urban morphology plays an indispensable role in representing spatial urban form parameters. It can help find clues to how different combinations and patterns in a city appear and how they are likely to influence air quality in urban environments. (Shi et al., 2017). As a result, urban forms that include different combinations and three-dimensional structures of urban elements influence air quality in different ways. One of the challenges is finding a proper framework for urban form, which will be addressed in the second chapter of this research, answering one of the research questions.

Numerous cities have undergone dramatic changes in recent decades, creating new types of urban fabrics with complex components, e.g., urban canyons. New urban patterns can have different effects on air pollution. It is crucial to develop new methods for classifying urban forms to keep track of their effects on various phenomena like air pollution (Cai & Chen, 2022). Urban morphology studies traditionally focused on qualitative methods that generally resulted in conceptual urban models as an outcome. In contrast, several studies have recently implemented quantitative methods along with qualitative approaches to analyze urban forms (Moosavi, as cited in Carta, 2022). In recent years, some researchers have concentrated on data-driven approaches using machine learning models for the quantitative study of urban morphology by taking into account the techniques for multi-dimensional aspects of urban form (Cai et al., 2021).

In summary, one of the significant phenomena affected by urban formation processes is air pollution. However, not many studies focus on the effect of urban form on air pollutants. Limited access to data as well as a lack of methodology covering large-scale datasets could be essential reasons (Kim & Gim, 2022). Thanks to the availability of open-source earth observation data, there would be more opportunities to use large-scale datasets with more accurate results. Although several works have been done with new methods for quantifying urban morphology and investigating its relationship with socio-economic and environmental issues, there are many domains to develop this new approach. Few studies used remote sensing data as an input dataset to extract morphological characteristics, especially in the application of air pollution. Huang et al. (2022) also claim that despite studies on the impact of urban morphologic-related measurements on air quality, these results cannot be generalized to actual city blocks due to being modeled in ideal structures (Huang et al., 2022). However, working on earth observation data for examining such a correlation using advanced deep learning techniques, e.g., CNN, could provide a more reliable outcome and higher accuracy. This study aims to use this technique to interpret the effect of urban forms on air pollution. The outcomes of this research can help influential groups on urban development to know what they should not do and how they can act to mitigate the effect of urban forms on air pollution.

1.2. Problem statement

Several studies have focused on quantitative urban form studies using individual measurements for air pollution effects. This approach seems to be problematic in two ways. Firstly, Accessing vector data for such analysis might be limited. In contrast, The possibility of using openly accessible Earth Observation (EO) data with new machine learning methods can provide the opportunity to investigate urban forms directly without creating vector data from raster data. The advantage of extracting morphological information from the raster data such as satellite images is their short update cycle, which can detect changes in the urban environment. This capability cannot be observed using vector data for all places due to the time-consuming process of digitizing and updating the data. Additionally, extracting information about urban forms from open-access EO data could be helpful in underdeveloped and developing countries, where up-to-date vector representation of urban environments is unavailable. To the author's knowledge, only a few studies in urban morphology have investigated the possibility of extracting different types of urban forms from EO data. Chen et al. (2021) worked on a convolution neural network technique to automatically classify street networks, one of the main elements of urban morphology (Chen et al., 2021). In another study, urban form characteristics were delineated in seven classes using an unsupervised deep-learning method in a metropolitan area (Cai and Chen, 2022).

Secondly, several studies have investigated the relationship between urban form and air pollution. However, most of them focused on using measurements related to urban form variables individually. This means they did not consider the combination of measurements that can make different urban patterns in a city. As an example of such an approach, Kim and Gim (2022) explored the correlation between urban form and the PM2.5 index. Using long short-term memory (LSTM) and random forest algorithms, Kim and Gim (2022) predicted the dispersion of PM2.5 using six urban form variables. They discovered the importance of urban form features on PM2.5 concentration (Kim & Gim, 2022). In another study, Huang et al. (2022) obtained nine urban morphological-related indicators of high-density areas as predictors for the level of PM2.5 and PM10 pollutants using mobile monitoring data (Huang et al., 2022). However, by utilizing the scene-based comprehension of deep learning models, this research takes a novel direction in comprehensively capturing the complex interactions and relationships between urban form and PM2.5 concentration.

1.3. Research objectives and questions

This research aims to analyze, model, and develop the relationship between urban forms and PM2.5 (particulate matter with a diameter of 2.5 micrometers or smaller) with severe adverse effects on human health using a deep learning-based model applied to earth observation datasets. To achieve this goal, I need to tackle the following objectives. It is necessary to follow the right direction to accomplish the research goal. Research questions point to the direction of objectives, and answering them can lead to accomplishing the main research goal.

1. **Objective (1):** To achieve a deep learning model with high performance to classify urban forms using open-access Earth observation data.

    1.1. Which classification framework is effective for training deep learning models and utilizing Earth observation data to classify urban form?

    1.2. Which Convolutional Neural Networks architecture from deep learning can provide an acceptable accuracy to predict the urban form classes using EO data?

    1.3. What criteria should be considered for tuning hyperparameters in the CNN model?

2. **Objective (2):** To select and train a suitable model representing urban form classes impact on the spatial distribution of PM2.5, which is applicable in large-scale areas.

    2.1. What modeling techniques are most suitable for examining the impact of urban form classes on the concentration of PM2.5?

    2.2. To what extent does urban form contribute to the concentration of PM2.5?

    2.3. Which types of urban forms have the strongest impact on the dispersion of PM2.5?

1.4. Thesis outline

The rest of the research thesis is organized into five chapters. The next chapter provides a review of the relevant and related work. The third chapter describes the data used in the research. The fourth chapter outlines the research methodology, explaining the methods and required steps to address each research question. The results are presented in the fifth chapter, followed by discussions. The final chapter includes the conclusion, the limitations of the research, and future work.

2. RELEVANT AND RELATED WORK

2.1. Urban morphology and local climate zones (LCZ) classes

For this research, it is necessary to find a suitable framework for urban form categorization.
Several studies focused on finding a framework for categorizing urban forms. Table 1 shows the most relevant framework widely used in recent years.

Table 1: The list of different frameworks for urban form classification

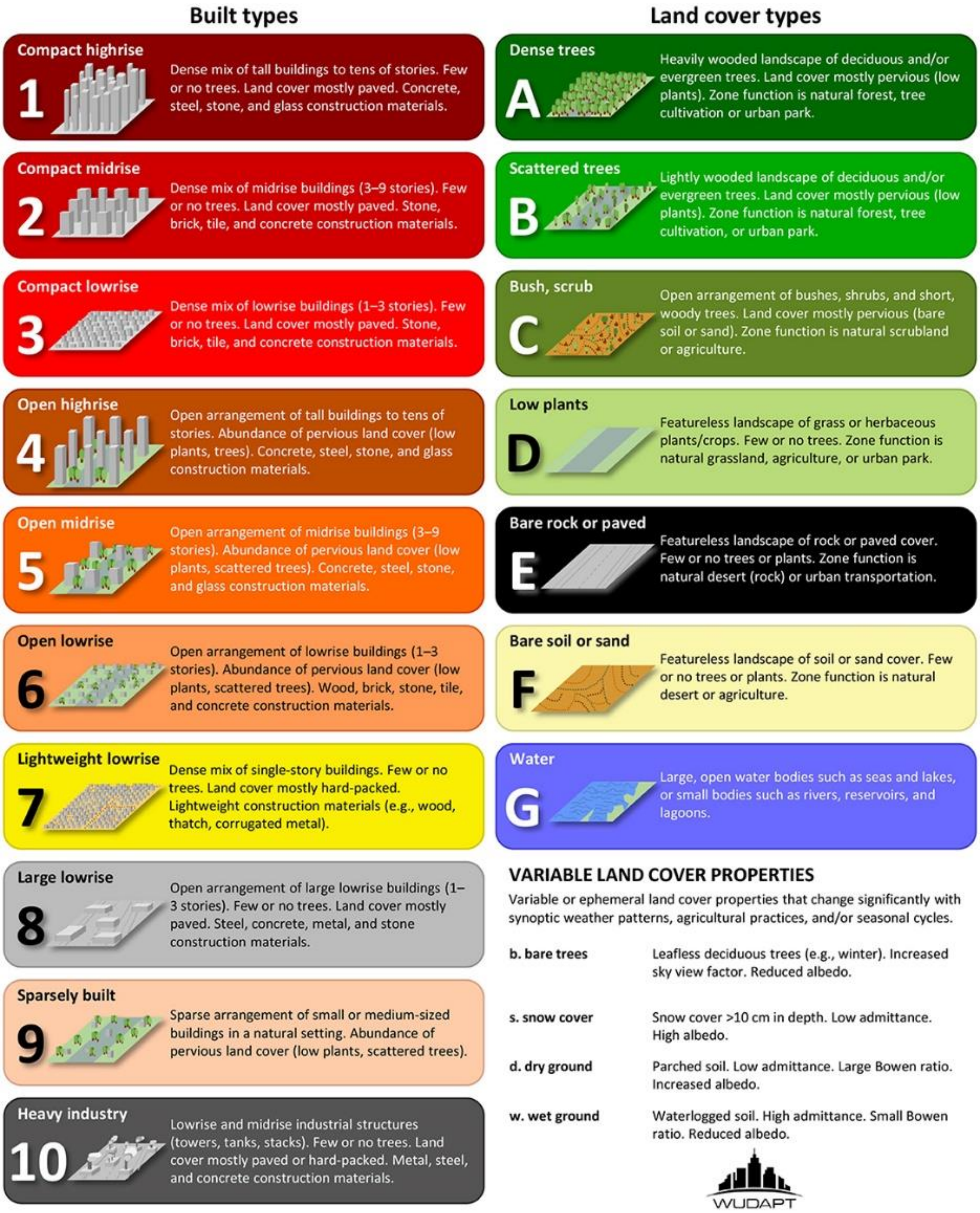
Categories	Source
1. Ground Plan (Streets, Blocks, Buildings) 2. Building form pattern (2D form, 3D form) 3. Land use pattern (land use function, land use intensity)	(Wu et al., 2022)
1. Morphological attributes: Centrality, Density(intensity), Diversity 2. Socioeconomic Livability: Economic Vitality, Accessibility, Affordability, Social Diversity	(Martino et al., 2021)
1. building coverage ratio 2. floor area ratio 3. low building area 4. high building area 5. sources area such as roads and plants 6. green area	(Kim and Gim, 2022)
Two-dimensional variables: (1. Impervious Surface Ratio, 2. Vegetation Ratio, 3. Water Ratio, 4. Soil Ratio, 5. NDVI) Three-dimensional variables (1. Floor Area Ratio, 2. Building Density, 3. Sky View Factor) Distance variables (1. distance to industrial areas, 2. distance to main road, 3. distance to parks, 4. distance to water)	(Gao et al., 2021)
Local climate zones (LCZ)	(Stewart and Oke, 2012),

Wu et al. (2021) divided urban form into three components. The ground plan consists of the road network, block pattern, and building scale, followed by measurements related to 2D and 3D aspects of buildings, such as building area and building height, respectively. Regarding land use patterns, they considered the functions related to land use, such as their proportion (Wu et al., 2022). The other study focused on predicting livability from urban form, considering centrality, density, and diversity as the morphological attributes of a metropolitan area (Martino et al., 2021). Studying the effect of urban form on air pollution, Kim&Gim (2022) considered six individual measures concerning urban density, urban height, and open spaces and roads to represent urban characteristics (Kim and Gim, 2022). Gao et al. (2021) extracted 2D features from satellite imagery and 3D factors from vector data to represent urban morphology on a block scale to discover their impact on urban heat islands (Gao et al., 2021). In addition, several works used the local climate zones (LCZ) framework to represent urban form (Stewart and Oke, 2012, Xu et al., 2019, Demuzere et al., 2019, Bechtel et al., 2017).

Among different ways for urban form classification, the Local Climate Zones (LCZ) will be selected for this research. LCZ is one of the urban form classification systems widely used in climate-based studies of urban environments. This scheme was introduced by Stewart and Oke in 2012. As the authors define it, "The LCZ system is segmented into 10 "built" types (LCZ 1–10) and 7 "natural" types (LCZ A–G), based on the regional landscape patterns. Every class exhibits a distinct urban form, which can be identified by specific spectrums of values for spatial and land cover attributes (Stewart & Oke, 2012). Figure 1 shows all LCZ categories and each class's visual concept.

The local climate zone types not only cover both built and natural environments but also recognize the physical and functional characteristics of an urban area. However, physical characteristics are the core of the classification, and the functionality is used as a supporter to distinguish between similar physical characteristics. LCZ provides a standardized classification system for urban form, resulting in consistent categorization and applicability across different regions. In addition, this scheme offers several measures related to morphological characteristics in one category suitable for scene understanding by deep learning models, and the Earth observation data has proven supervised LCZ mapping effectively based on previous research findings (Bechtel et al., 2015).

Discovering the spatial structures in a city and their effect on climate-based applications is widely conducted by LCZ classes. By utilizing this type of classification, the description of different elements of urban environments can be generalized in urban scientific communities (Aslam & Rana, 2022). Mix forms of urban elements might appear in different urban blocks, representing urban morphological characteristics. The building's compactness and the ratio of open spaces exemplify the measurements, indicating different urban forms. Recently, LCZ maps have been prepared using different techniques to provide input to different urban subjects affected in urban areas, e.g., urban climate, air pollution, and energy (Aslam & Rana, 2022).



Built types

Compact highrise 1 — Dense mix of tall buildings to tens of stories. Few or no trees. Land cover mostly paved. Concrete, steel, stone, and glass construction materials.

Compact midrise 2 — Dense mix of midrise buildings (3–9 stories). Few or no trees. Land cover mostly paved. Stone, brick, tile, and concrete construction materials.

Compact lowrise 3 — Dense mix of lowrise buildings (1–3 stories). Few or no trees. Land cover mostly paved. Stone, brick, tile, and concrete construction materials.

Open highrise 4 — Open arrangement of tall buildings to tens of stories. Abundance of pervious land cover (low plants, trees). Concrete, steel, stone, and glass construction materials.

Open midrise 5 — Open arrangement of midrise buildings (3–9 stories). Abundance of pervious land cover (low plants, scattered trees). Concrete, steel, stone, and glass construction materials.

Open lowrise 6 — Open arrangement of lowrise buildings (1–3 stories). Abundance of pervious land cover (low plants, scattered trees). Wood, brick, stone, tile, and concrete construction materials.

Lightweight lowrise 7 — Dense mix of single-story buildings. Few or no trees. Land cover mostly hard-packed. Lightweight construction materials (e.g., wood, thatch, corrugated metal).

Large lowrise 8 — Open arrangement of large lowrise buildings (1–3 stories). Few or no trees. Land cover mostly paved. Steel, concrete, metal, and stone construction materials.

Sparsely built 9 — Sparse arrangement of small or medium-sized buildings in a natural setting. Abundance of pervious land cover (low plants, scattered trees).

Heavy industry 10 — Lowrise and midrise industrial structures (towers, tanks, stacks). Few or no trees. Land cover mostly paved or hard-packed. Metal, steel, and concrete construction materials.

Land cover types

Dense trees A — Heavily wooded landscape of deciduous and/or evergreen trees. Land cover mostly pervious (low plants). Zone function is natural forest, tree cultivation or urban park.

Scattered trees B — Lightly wooded landscape of deciduous and/or evergreen trees. Land cover mostly pervious (low plants). Zone function is natural forest, tree cultivation, or urban park.

Bush, scrub C — Open arrangement of bushes, shrubs, and short, woody trees. Land cover mostly pervious (bare soil or sand). Zone function is natural scrubland or agriculture.

Low plants D — Featureless landscape of grass or herbaceous plants/crops. Few or no trees. Zone function is natural grassland, agriculture, or urban park.

Bare rock or paved E — Featureless landscape of rock or paved cover. Few or no trees or plants. Zone function is natural desert (rock) or urban transportation.

Bare soil or sand F — Featureless landscape of soil or sand cover. Few or no trees or plants. Zone function is natural desert or agriculture.

Water G — Large, open water bodies such as seas and lakes, or small bodies such as rivers, reservoirs, and lagoons.

VARIABLE LAND COVER PROPERTIES

Variable or ephemeral land cover properties that change significantly with synoptic weather patterns, agricultural practices, and/or seasonal cycles.

b. bare trees	Leafless deciduous trees (e.g., winter). Increased sky view factor. Reduced albedo.
s. snow cover	Snow cover >10 cm in depth. Low admittance. High albedo.
d. dry ground	Parched soil. Low admittance. Large Bowen ratio. Increased albedo.
w. wet ground	Waterlogged soil. High admittance. Small Bowen ratio. Reduced albedo.

Figure 1: local climate zone(LCZ) classification types and definitions (Stewart and Oke 2012; Demuzere et al., 2020)

2.2. Deep learning and Earth observation for LCZ classification

Deep learning has gained significant traction in Earth observation for image classification tasks. It includes a computational approach that utilizes neural networks to learn and extract complex patterns automatically from remote sensing data. Deep learning models can efficiently capture the spatial and spectral characteristics of the data, leading to improved classification performance and the ability to handle large-scale and high-dimensional remote sensing datasets (Li et al., 2018).

Among deep learning approaches, convolutional neural networks have demonstrated remarkable success in classifying scenes due to their exceptional capability to acquire knowledge about the composition and contextual details in image scenes (Yao et al., 2022). Convolutional Neural Networks (CNNs) are a widely-used deep learning technique, and it has demonstrated their advantage in automatic detection and the capacity to represent unstructured features (Huang et al., 2021).

CNNs used for image classification typically consist of three main sections: convolutional layers, pooling layers, and fully-connected layers (Li et al., 2018). The convolutional layers employ learnable kernels in the form of filters to capture and represent important patterns and information by convolving the input images. The fusion of local spatial connectivity and spectral bands within the local receptive field enables the generation of high-level image representations and the extraction of valuable features (Kim et al., 2021). Figure 2 illustrates a basic example of CNNs architecture.

In the context of urban form classification, the patch-based convolutional neural network models, which consider smaller parts of an image as a patch for identifying different classes, are more successful than pixel-based classifiers, such as random forest modeling for urban form classification (Yoo et al., 2019). The reason could be the nature of the urban form, which is recognizable in the scale of urban blocks rather than individual buildings.

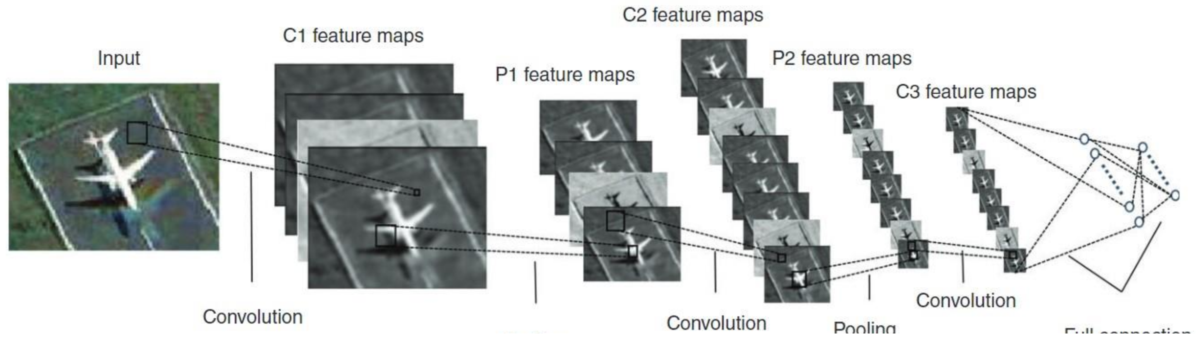


Figure 2: Basic example of convolutional neural networks

Several studies have worked on classifying local climate zones using machine learning techniques. The World Urban Database and Access Portal Tools(WUDPT[1]) is a world urban database for providing LCZ maps globally. This platform also uses ensemble machine learning models, e.g., random forest, for the classification task. In recent years, thanks to developing state-of-the-art CNNs techniques, several studies have paid attention to using convolutional neural networks to classify local climate zones on a scene-based level. They usually use remote sensing or ground-level imagery data and, in some cases, a combination of multiple earth observation data or other datasets, providing higher accuracy in urban form classification.

These studies can be divided into three groups. The first group used transfer learning techniques to implement the advanced models for doing classification tasks on LCZ types. Xu et al. (2019) performed Inception-v3 model of convolutional neural networks for LCZ classification. Their model used ground-

[1] https://www.wudapt.org

level images to consider 3D aspects of urban environments. Their model's accuracy reached 69% (Xu et al., 2019).

In their case study, the second group used transfer learning techniques to develop and modify ongoing procedures to achieve higher accuracy. One of this group's works is mapping LCZ classification over eight German cities. Rosentreter et al. (2020) used supervised convolutional neural networks and Sentinel- 2 satellite imagery. They adapted VGGNet, one of the famous CNN models for image classification, with some modifications in the model architecture, such as adding batch normalization. The result of their model obtained an accuracy of 85%. (Rosentreter et al., 2020). In the other research of this approach, Zhu et al. (2022) adjusted Resnet-50 CNN-based architecture by reducing the number of residual blocks to make the input image patches fit the model. They worked on using both Senl-1 and Senl-2 satellite images for training their model (Zhu et al., 2022).

The last group designed their patch-based CNNs models to classify local climate zones for LCZ classification problems. Huang et al. (2021) designed a light-weight model called LCZ-CNN to classify LCZ maps using multispectral images of Landsat satellite imagery in 32 large cities in China (Huang et al., 2021), operating on Google street view images. They achieved an overall accuracy of 80%. In another study, a multi-scale, multi-level attention network (MSMLA-Net) was introduced for scene-based LCZ classification using deep learning by Kim et al. (2021). They developed advanced computer vision techniques by using sentinel-2 imagery, followed by OSM building data, DSM height, and national land cover map as additional bands for the input data. Implementing their model resulted in an overall accuracy of 87% (Kim et al., 2021).

This research will label LCZ classes from EO data using deep learning models. Urban form analysis traditionally relies on morphological indices to quantify characteristics but often fails to capture the visual patterns that can be intuitively recognized by human observation. However, the rapid advancements in deep learning techniques have empowered machines to develop a human-like understanding of urban form (Chen et al., 2021). It is crucial to consider the combinations of different elements in a city as patterns likely to affect air pollution concentration instead of relying on measurements individually. Moreover, deep learning methods, e.g., Convolutional Neural Networks (CNN) have been proven to reach high accuracy in image recognition tasks. Deep learning models applied to EO data can provide such a perspective. This technique has an excellent performance in classifying scenes because of its exceptional capacity to learn image composition and unstructured information (Yao et al., 2022).

2.3. LCZ classes and air pollution measurement

Several factors from urban morphological characteristics influence the dispersion of PM2.5 in an urban environment. Moreover, meteorological factors, mainly the spatial distribution of the urban wind environment, exhibit significant variation and are heavily influenced by the urban forms features. The presence of traffic emissions in the atmosphere greatly impacts the distribution of air pollution, which is closely linked to urban morphology (Li et al., 2021). Li et al. (2021) investigated the impact of urban form on air pollutants in two urban streets and neighborhood scales. Understanding the neighborhood level, they considered urban density, diversity, and spatial characteristics as the influencing components. They also evaluated the building height level, the opening, and the separation of buildings at the street level. The results of their work show that there is a significant correlation between vertical urban densities and dispersion of air pollution as results in airflow reduction. The street-related factors also directly impact pollutant concentration (Li et al., 2021).

In the other study by Gao et al. (2019), land use categories, climate-based factors, traffic flow, the height of buildings, and road networks were considered as the influencing factor on air pollution. Their research shows that The primary factors influencing the variation in PM2.5 levels were the traffic volume and the heights of buildings (Gao et al., 2019).

The research of Yang et al. (2022) explains that the previous works have extensively investigated the impact of urban landscape composition on air pollution, revealing that industrial and commercial areas are major sources of pollutants, while vegetation and urban afforestation systems act as beneficial sinks. Additionally, limited studies have explored the indirect effects of landscape configuration on air pollution, highlighting the role of microclimate factors in pollutant transport and dispersion. For example, open spaces facilitate improved air circulation and reduced pollutant deposition, whereas compact areas encounter limited air dispersion, resulting in poorer air quality. However, despite these findings, the comprehensive understanding of the impact of urban form on pollutants requires further investigation and examination of the combined influence of composition and configuration (Yang et al., 2022).

The studies highlight the significant urban form indicators and other variables on air pollution dispersion. To address this, the emphasis of the research is on using Local Climate Zones (LCZ) classes to incorporate various individual factors that have been commonly studied. This approach allows for examining the collective impact of each LCZ class, which encompasses multiple measurements, on air pollution concentrations.

For example, the ‘Open Highrise’ LCZ class (class 4) is characterized by an open arrangement of tall buildings spanning multiple stories, surrounded by ample greenery and scattered trees. It encompasses specific properties such as a mean building height exceeding 25 meters, a sky view factor ranging from 0.5 to 0.7, and additional defining characteristics. Additionally, the function of this class is related chiefly to residential with single-unit housing, high-density housing, and commercial, including small retail shops. The class also takes into account certain aspects related to building materials (Stewart, 2011). Therefore, the LCZ classification effectively combines multiple individual indicators within its categories. By employing such an easily understandable and applicable framework for urban form classification, urban planners can facilitate the integration of research findings into urban development strategies focused on sustainability (Yang et al., 2022).

Although the LCZ framework incorporating geometrical, built environment, and human-induced elements have been widely utilized in urban heat island studies, few studies used this scheme to discover the effect of urban forms on air pollutant concentration. In one study, the researchers utilized the Multiple Linear Regression (MLR) and Geographically Weighted Regression (GWR) modeling methods to develop estimation models for PM2.5 concentrations. They employed a set of urban form factors through the LCZ scheme as the foundation for calculating the metrics. These factors were then employed as independent variables to discover the spatial disparities observed in PM2.5 levels. The goal of the research was to analyze the impact of urban form on the concentration of PM2.5 (Shi et al., 2017). Specifically, the researchers in this study aimed to identify landscape categories that significantly influence PM2.5 concentration levels. This research’s findings indicate that with only five LCZ classes, around two-thirds of the dispersion in PM2.5 can be. This highlights the effectiveness of the LCZ framework in the spatial distribution prediction of air pollution. This approach holds significant value in evaluating the air quality of urban areas and cities that lack long-term monitoring data, detailed traffic information, and comprehensive emission inventories (Shi et al., 2019).

New research by Yang et al. (2022) classified LCZ types using random forests. They also created seasonal spatial PM2.5 maps based on air pollution data and other related data, namely wind speed, traffic, land use, and population, using a land use regression model to explore the effect of LCZ classes on PM2.5. The findings of this study indicate that there are notable variations in PM2.5 levels across different LCZ categories, including differences between built and natural classes as well as within the built classes. This suggests that the LCZ scheme can effectively capture the spatial variation of PM2.5 in urban areas. It is consistently observed that the natural category exhibits lower PM2.5 concentrations compared to the built type. Within the built type, there is a general trend of higher PM2.5 concentrations in compact areas compared to open areas and higher concentrations in high-rise areas compared to mid-rise and low-rise areas. These patterns persist throughout the year (Yang et al., 2022).

As a result, Investigating urban environments with LCZ classification provides informative outcomes for urban experts to consider in planning and designing to reduce air pollution concentration. Accordingly, LCZ is helpful for the prediction of air pollutants in urban environments (Shi et al., 2019).

2.4. Predictive modeling and urban form impact analysis

Previous studies assessing the relationship between LCZ classes and PM2.5 pollutants mostly relied on traditional models that assume linearity in their analysis. However, it is essential to note that the contribution of the influential factors on PM2.5, ranging from urban form classes to meteorological data, does not necessarily follow a linear pattern. The emergence of advanced machine learning and deep learning methods has opened up new possibilities for addressing non-linear and complex problems. These methods can uncover more complex interactions among data than deterministic and statistical methods. Due to their powerful nonlinear modeling capabilities, the state-of-the-art artificial intelligence methods performed at the highest level in forecasting air pollution concentrations (Ma et al., 2020). These cutting-edge techniques can achieve more accurate and reliable results and comprehensively understand the complex dynamics and relationships between LCZ classes and PM2.5 concentrations.

The ensemble learning methods, one of the machine learning-based algorithms, combines the predictions of multiple individual models (decision trees) to make a final prediction. This ensemble approach can improve the overall predictive performance of the model and help mitigate overfitting (Breiman, 2001).

Ensemble algorithms with bagging mechanisms use multiple independent models to make predictions. These predictions are then combined through stacking and further improved through boosting. This approach allows these models to make more precise predictions with fewer errors. (Lin et al., 2022).

One of the well-known ensemble learning methods, which can explore the effect of different variables on PM2.5, is the eXtreme Gradient Boosting (XGBoost) model. The framework for gradient boosting developed by Chen and Guestrin (2016) is effectively implemented in XGBoost. By supporting the simultaneous processing of tree building, addressing overfitting, and accelerating the execution, can aid in these tasks (Chen and Guestrin, 2016). It is an adaptable and comprehensive tree-boosing mechanism that covers the entire process from start to finish., and it has received much application and attained cutting-edge performance for regression and classification problems (Zheng et al., 2017). XGBoost is able to discover feature importance among predictors, providing clear outcomes such as feature importance and correlation between variables. This can help find the relationship between urban form and air pollution more straightforwardly and clearly. This analysis helps identify the most relevant variables contributing to PM2.5 concentrations, enabling enhanced comprehension of the factors influencing air pollution in the study area. XGBoost also is known as a high-performance model with acceptable accuracy, especially in cases where the dataset has many features or complex relationships between the input features and the target variable (Lin et al., 2022).

Several studies have been conducted on predicting air pollutants using the XGBoost model. The contributing variables to air pollution measurements were explored by classification task using XGBoost (Nababan et al., 2022). In the other study, the concentration of fine particulate matter was predicted using the regressor of XGBoost. Several factors, including geographical, temporal, meteorological, and topographic features, followed by population and other air pollutants, were considered as predictors. They obtained the R-squared of 0.61 on predicting unseen data (Lin et al., 2022). Ma et al. (2020) used this model to identify the spatial effects of air pollution in order to find the areas that are highly exposed to pollutants (Ma et al., 2020). Finally, Joharestani et al. (2019) predicted PM2.5 values using different models, including XGBoost, using the regression method, resulting in an R-squared of 0.67 (Joharestani et al., 2019).

As discussed, there are several advantages to using an XGBoost technique for the problem of this research, including indicating the relative contribution of each predictors variable in predicting PM2.5 straightforwardly, modeling nonlinear relationships to capture complex interactions between LCZ classes and PM2.5, and the ability of relative robustness to outliers and handling missing data. However, the machine learning models such as XGBoost do not consider the pattern of time stamps in the case of time series data for forecasting. This can negatively affect the performance of predicting model. The results also might not entirely reflect the nature of the trend-based dataset for the observations of PM2.5 in a certain period and with a temporal resolution (Dai et al., 2021).

This research aims to discover the relationship between local climate zone classes and PM2.5, considering hourly values. Therefore, a time series model will also be implemented to use the sequential characteristics of the dataset in the prediction. One of the successful techniques for modeling long-term dependencies of predictors on PM2.5 concentrations is long short-term memory (LSTM) neural networks, which consider the spatiotemporal characteristics of the given dataset for prediction (Li et al., 2017). In 1997, Hochreiter et al. introduced LSTM (Long Short-Term Memory) as an efficient architecture within the domain of recurrent neural networks (RNNs) (Hochreiter and Schmidhuber, 1997). The design of RNNs aims to handle problems where the data changes over time and has non-linear patterns. RNNs have connections that allow information to flow forward and backward, making them well-suited for predicting future values in time series data. They can learn patterns from the sequence of past data to make predictions about what will happen next. However, one drawback of RNNs is the issue of gradient vanishing, where the gradients used for training can become very small and lead the network to stop learning effectively. This limitation makes simple RNNs less suitable for forecasting problems that involve long-term dependencies or relationships between distant events in the time series. (Zheng et al., 2017).

LSTM was therefore designed to tackle the limitation of dealing with long dependencies. As part of the LSTM structure, RNN neurons are provided with input gates, output gates, and forgetting gates to overcome the disappearing gradient issue (Graves, 2012). The novelty in LSTM structure is the memory cell. This block serves as a container unit for important state information. Figure 3 shows the components of an LSTM cell. It consists of several steps as the gates.

1. The forget gate determines what information to remove from the cell state based on the previous hidden state and input.
2. The input gate decides which information should be updated and creates a vector of new candidate values for the next state.
3. The output gate filters the cell state and calculates the desired output based on the updated cell state.

There are two output information for the block. The cell state is a long-term memory output, and the hidden state indicates the short-term memory. These steps involve sigmoid and tanh layers, weight matrices, and bias vectors to compute the necessary activations and transformations within the LSTM model. The purpose of the memory cell is to keep and update relevant information over time (Zheng et al., 2017).

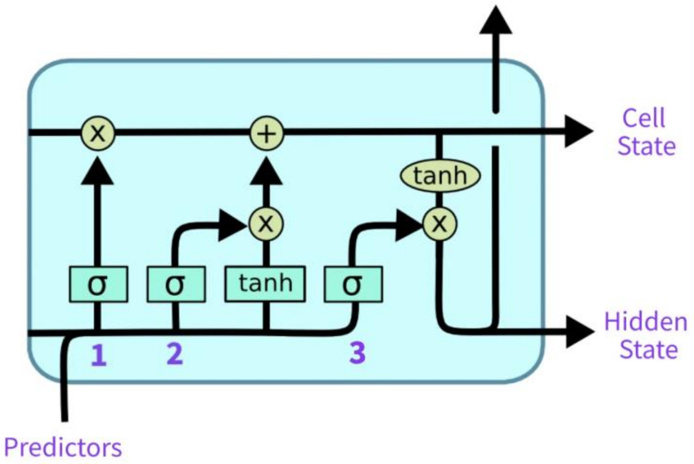


Figure 3: The structure of the LSTM memory block[2]

Reviewing the related work for predicting air pollution using LSTM, Li et al. (2017) proposed a novel model in combination with LSTM for forecasting values of PM2.5 concentrations over a period of two years. They considered weather-related data and time stamps of the month of year and hour of the day as the independent variables in their model. They compared the result of their model using LSTM with other standard techniques in this domain and achieved a more promising outcome (Li et al., 2017). In a further study, the integration of CNN and LSTM was applied to predict the level of fine particulate matter. The past 24 hours of PM2.5 measurements and aggregated wind velocity and rain were selected as the input data to estimate the PM2.5 for the next hour (Huang and Kuo, 2018). Discovering the link between air pollution data and meteorological-related measurements, Zhang et al. (2020) introduced a deep neural networks model based on long short-term memory layers which operate in both forward and backward directions to capture temporal dependencies in the input sequence. The results of their study indicate a significant correlation between the two mentioned variables (Zhang et al., 2020). The other valuable study used both XGBoost and LSTM techniques to estimate PM2.5 volume. In this study, Dai et al. (2021) used climate-related data, followed by other air pollution indexes except for fine particulate matter as the input dataset. Then, using the XGBoost technique, they extracted the most highly correlated variables in space and time with Pearson analysis. Finally, they implemented an LSTM model on time series data for forecasting PM2.5 concentration (Dai et al., 2021).

There are several plus points to using the LSTM model for predicting PM2.5 concentration to discover the contributing factors of LCZ classes and other predictors in this research. Such advantages can be seen in dealing with non-linearity and long-term dependencies in temporal dynamics. However, Interpreting the relationship between variables and within LCZ classes and understanding the importance of features influencing air pollution concentration is challenging with deep-learning models such as LSTM. This difficulty appears because deep-learning models are considered black boxes, meaning their internal workings are not easily interpretable or transparent (Ma et al., 2020). In contrast, the XGBoost model provides built-in functions for interpreting the performance, such as scoring the feature importance (Chen and Guestrin, 2016).

[2] *Source*: *source: https://towardsdatascience.com/how-to-learn-long-term-trends-with-lstm-c992d32d73be*

3. STUDY AREA AND DATA

3.1. Study area

This research is conducted in the city of Amsterdam. The study area for modeling is defined by the bounding box that includes the entire city of Amsterdam. The coordinates of the study area are reported in Table 2 in both the World geodetic system and the Dutch coordinate reference system.

Table 2: The geographic coordinates of the study area

Coordinate Reference System	Geographic Coordinates			
CRS	Minimum Latitude	Maximum Latitude	Minimum Longitude	Maximum Longitude
WGS 84 – EPSG: 4326	52.3038	52.431	4.7288	5.0792
Amersfoort/ RD New – EPSG: 28992	110093.2242	134052.4072	479740.4482	493733.9650

Figure 4 shows the location of the study area in the Netherlands. Covering most classes of local climate zone schemes and the availability of air pollution data was considered as a motivation to select Amsterdam as the study area.

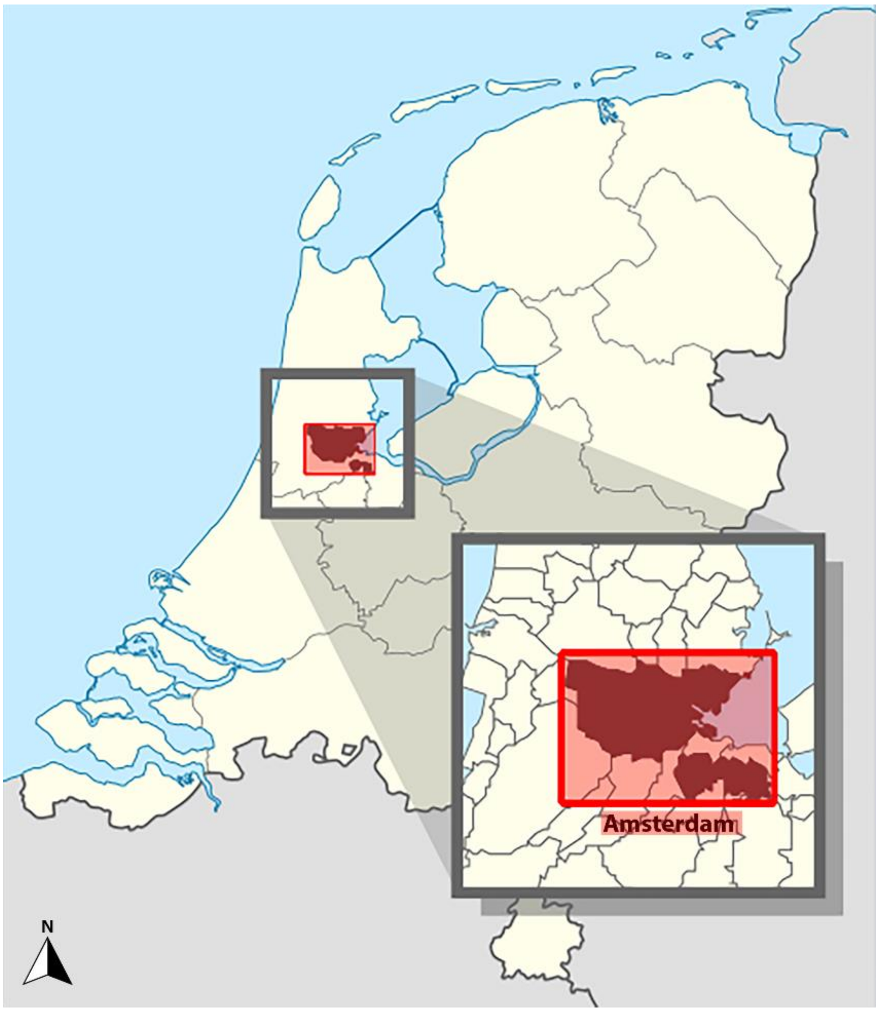


Figure 4: The location of study area

3.2. Data

As explained in the previous chapter, several datasets, including remote sensing, meteorological, and traffic data, are required for this research. The overview and the resource of each dataset will be explained below.

3.2.1. Earth Observation Data

In this study, the free accessible satellite data, Sentinel-2 imagery collection, are selected as the input datasets for urban form modeling to obtain LCZ classes. The satellite missions are employed by the European Space Agency (ESA). Sentinel-2 is an optical satellite producing multispectral images. This dataset is selected based on the related previous studies, which can provide a proper resolution to address the research problem and the data's availability.

3.2.2. Air pollution data

Conventionally, air quality stations implemented and maintained by the official organizations were responsible for collecting air quality data. However, they cannot provide dense coverage. Citizen science activities have rapidly increased recently in many fields (Assumpção et al., 2018), including air pollution monitoring, providing more data and new views to scientists and the public for scientific research (Kullenberg and Kasperowski, 2016). This research will use both citizen scientists' data for air pollution and data collected by governmental sensors.

The air quality data required for this research will be acquired from the National Institute for Health and Environment (RIVM) portal (see http://samenmeten.rivm.nl). This platform provides open access to the air quality data measured by citizen scientists' sensors as well as official measurements, and RIVM investigates the quality of data and presents the data. For this research, the hourly values of PM2.5 concentration for the period of two years, from 2021 to 2022, are selected as the target variable for modeling.

3.2.3. Meteorological data

One type of data required for air pollution modeling in this research is weather-related data sources. The Netherlands' hourly report of the official meteorological measurements is freely available on the Dutch weather service (KNMI) (see https://www.knmi.nl) website. The meteorological data for this research, available in the dataset, includes seven factors: wind speed, wind direction, temperature, precipitation, humidity, cloudiness, and air pressure. Only one sensor covers the study area based on the official stations' location, and using the same values of weather data for all sensors might affect the modeling results. Therefore, in addition to the official measurement, those weather variables (temperature and humidity) available from low-cost sensors in the RIVM portal are also used in air pollution modeling.

3.2.4. Traffic data

The dataset related to the traffic flow was not available for the research period. Therefore, road network data is considered as the basis for creating traffic data based on the distance to different types of roads and time stamps of the dataset as the assumption for traffic flow. The road data is extracted from the open street map (OSM) database, which is freely available.

4. METHODOLOGY

4.1. Method Overview

This research introduces a two-stage model for classifying local climate zones and predicting PM2.5 concentration. In the first phase, through the convolutional neural networks model, a supervised classification task is implemented on the earth observation data and labeled LCZ training data to classify urban local climate zones (LCZ) classes. Sentinel-2 imagery is the input of the first stage, and the image patches of this dataset are prepared concerning the LCZ sample data. Three different CNN models, namely the designed CNN, ResNet-50, and EfficientNet, are then employed for training patch-based classification. The model with the highest performance predicts LCZ categories in the study area. Then, in the second phase, the result of LCZ categories and probabilities, in addition to related spatiotemporal factors, including meteorological data, traffic data, and time stamps, are used to predict the concentration of the PM2.5 index in the study area using the regression task of both XGBoost and LSTM models. Finally, the sensitivity and feature importance analyses are conducted to analyze the effect of LCZ categories on PM2.5. Each stage of the model follows a three-step process, including data acquisition, data preparation, and data modeling and evaluation. There is an additional step in the first stage of the model for applying the trained model on the entire study area for creating the LCZ map and an extra step for analyzing the impact of LCZ on PM2.5 in the second stage. Figure 5 shows the overall view of the methodology.

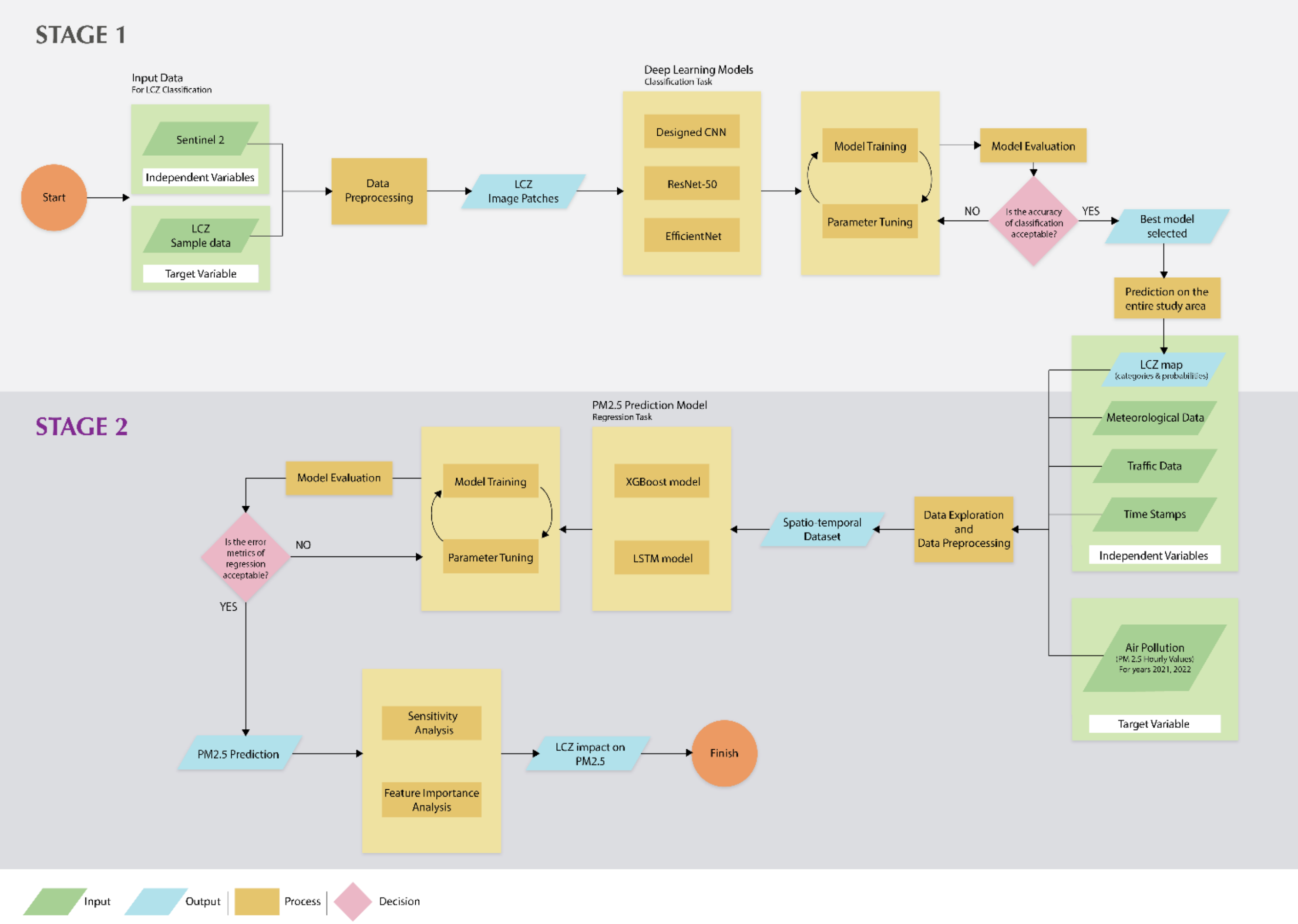


Figure 5: An overall view of the Methodology

4.2. Stage 1: Local Climate Zones Classification

4.2.1. Deep learning model for LCZ classification

Supervised Convolutional neural networks are trained for LCZ classification. The input data in this model are image patches of Sen-2 imagery, and the output will be local climate zone classes per patch. Three CNN model architectures are trained using the prepared image patches. The first CNN model is designed by the author of this research in the TensorFlow environment of Python. The two other models are selected from the state-of-the-art models' architecture available in this package, which is widely used in image classification tasks. These cutting-edge models are often developed and fine-tuned using large-scale datasets and advanced techniques. However, only the architecture of those models is used in this research, and they are trained on the prepared dataset to gain the weights. Regarding the previous study in LCZ classification, also considering the efficiency of the cutting-edge models for classification tasks in computer vision, we selected the two popular models called Resnet-50 (He et al., 2015) and EfficientNet (Tan and Le, 2019) for training. Finally, the model evaluation results are compared to select the best model for predicting LCZ in the study area.

4.2.2. Data acquisition

The data for this stage include Sentinel-2 imagery as remote sensing data and the data required to get a better understanding for preparing training data of LCZ labels, containing elevation level (AHN3-DSM), the land use map, the building footprint and the aerial imagery of 25 cm resolution.

Earth Engine Python API (ee library) is used to extract Sentinel-2 imagery for the study area. The goal is to obtain cloud-free images from the Sentinel-2 collection for 2021 and 2022, with a cloud percentage of less than 5%. Using this library which provides interaction with Google Earth Engine, the procedure outlined below is followed:

- Initializing Google Earth Engine API: Authenticating and initializing the Google Earth Engine Python API using the ee library.
- Defining the Study Area: The study area is determined using a polygon geometry by the list of geographic coordinates.
- Collecting Sentinel-2 Images: The COPERNICUS/S2 collection, the data catalog for sen-2 imagery, is filtered based on the study area and the date range of the research from January 1, 2021, to December 31, 2022. The images are further filtered based on a maximum cloud percentage of 5% to ensure cloud-free imagery.

The related data for creating the training area of LCZ classes are acquired from WFS and/or WMS services, available in the PDOK platform, using QGIS.

4.2.3. Data preparation

Preparing the data for the first stage is done for Sen-2 imagery, the local climate zones sample data, and image patches for training from sen-2 imagery and LCZ sample data.

Sen-2 imagery preparation:

- All the images collection are aggregated by the median value to create a composite image representing the study area as a snapshot.
- Sent-2 imagery contains 13 spectral bands from which 10 bands with 10 and 20 meters resolution are collected. Regarding the previous study using Sen-2 imagery in LCZ classification (Zhu et al., 2022, Kim et al., 2021), the desired spectral bands (B2, B3, B4, B5, B6, B7, B8, B8A, B11, B12) are selected from the composite image. The resolution of the selected bands is close to the resolution of the

desired raster imagery, which is 10 meters. The information on the remaining channels with 60 meters resolution would not be helpful in the problem of this research. Then, the channels with 20 meters resolution are resampled to 10 meters.

- Finally, the output image is exported in GeoTiff format. The exported image (Figure 6) has a dimension of 2399 by 1718 pixels with a spatial resolution of 10m by 10m and is in the Dutch reference system (EPSG:28992).

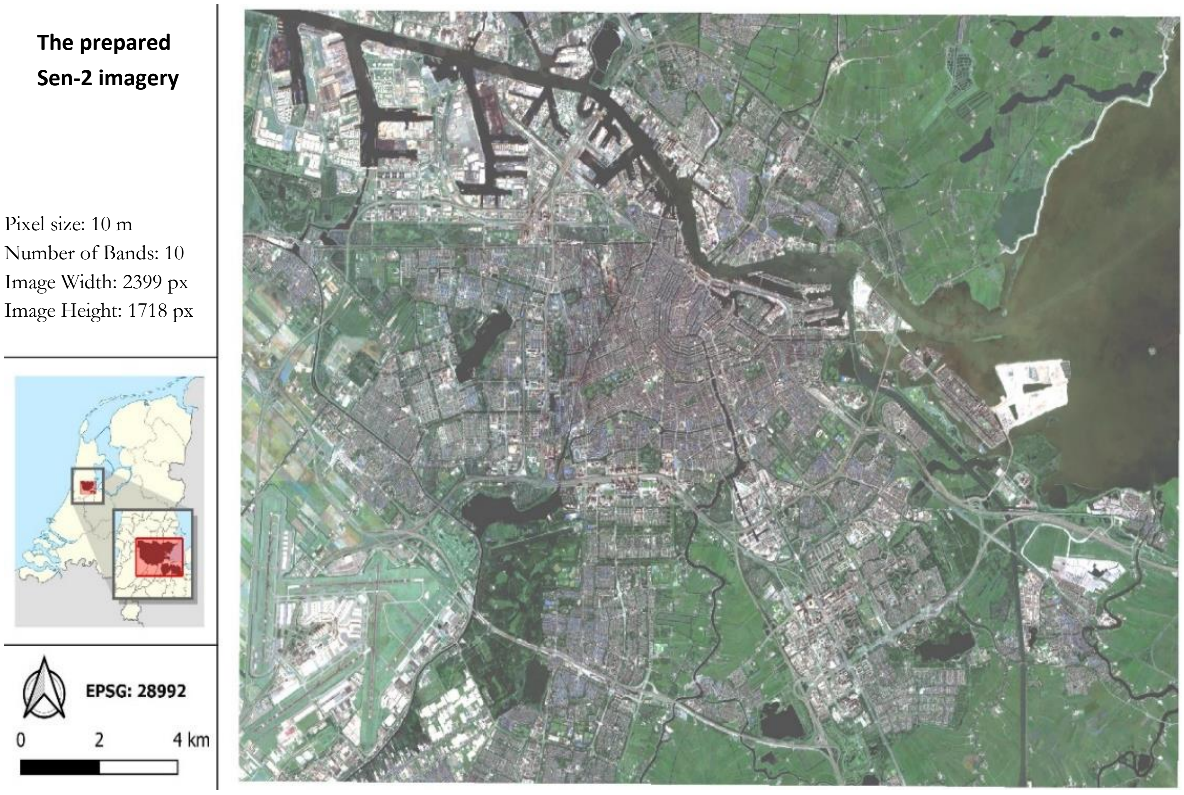


Figure 6: The prepared Sentinel-2 imagery for the research

LCZ sample data preparation

We need sample data to train supervised convolutional neural networks for LCZ classification. There is a reference data resource called So2Sat LCZ 42 project. This project has provided a benchmark dataset for global local climate zone classification. A group of experts labeled LCZ classes for 42 cities worldwide based on Sentinel-1 and Sentinel-2 images. However, filtering the labeled data for a specific location is impossible. As a result, a training-labeled dataset for the study area is prepared manually for this research., The instructions and digitization guidelines provided in the World Urban Database and Access Portal Tools (WUDAPT)[3] portal are used to create the training area.

Regarding the suggestion in the portal, google earth pro is used for drawing LCZ polygons. As the CNN model is trained on image patches of size 32 by 32 pixels, based on the most recent previous studies for LCZ classification, the polygons cover an area of at least 320 m by 320m to ensure that the image patches have the most overlap with its corresponding polygons. It is also essential to keep a reasonable distance between polygons with the same class to avoid overlapping patches created from polygons of different categories.

[3] *https://www.wudapt.org/digitize-training-areas/*

Several related datasets and factors are considered to help achieve more accurate labeled data, including the European LCZ map (Demuzere et al., 2019) and the global LCZ (Demuzere et al., 2022). Both datasets were generated using machine learning models from multiple remote sensing data, and the outputs are raster images with a spatial resolution of 100 meters. The dataset is available in the WUDAPT portal. In addition, the other factors considered in preparing the training area are acquired from the "public services on the map" (PDOK) (see http://pdok.nl/datasets) website, which is a platform for finding geo-datasets in the Netherlands.

Creating training areas is not easy, as there are many similarities between urban form classes, and sometimes cannot be discretized by human inspection. Several factors are used to differentiate the LCZ classes in digitizing steps and reaching more accurate labels. When the polygons are drawn, the LCZ class of the polygon is defined considering the following factors. Figure 7 shows an overview of how LCZ polygons are created.

- Using the available LCZ factsheet[4] (from Stewart, 2011) for the general recognition of categories.
- Control polygons' categories with two referenced datasets: the Global and Europe LCZ maps.
- Control with elevation level (AHN3- DSM).
- Control with 3D terrain in google earth pro and street view.
- Control with the land use map.
- Control the building footprint.
- Control with aerial imagery of 25 cm resolution.

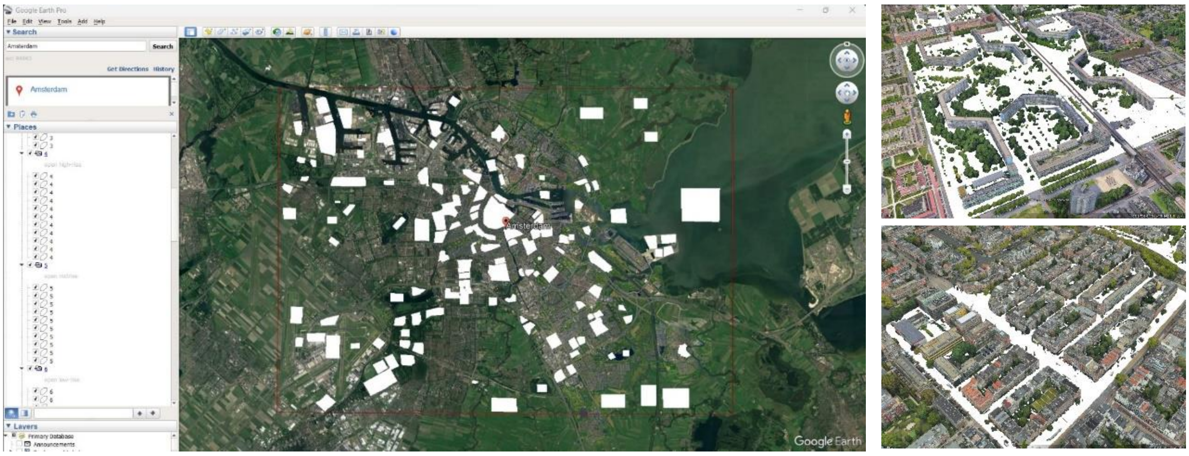


Figure 7: Creating training areas for local climate zone classification using Google Earth Pro

The digitized polygons are stored in separate folders with the corresponding class in kmz format using a template provided in WUDAPT portal. Then, the polygons in the folders of kmz file are combined in a shapefile using merge vector layer tools in QGIS, creating the attribute table for polygons with the name of the classes.

Preparing image patches for training from Sen-2 imagery and LCZ polygons

The dataset to be served as the training set for the CNN model includes image patches of 32 by 32 pixels and a resolution of 10 meters with 10 channels. The area covered by each patch should be good enough to represent an urban form type. Several previous studies, such as (Zhu et al., 2020) and (Rosentreter et al., 2020), utilized image patches measuring 32 by 32 pixels at a spatial resolution of 10 meters, covering an area of 320 by 320 meters on the ground. In a study conducted by Liu and Shi (2020), various patch sizes were compared, and it was determined that larger patch sizes ranging from 32 by 32 to 64 by 64 pixels

[4] https://www.wudapt.org/wp-content/uploads/2021/05/Stewart_PhD_2011_LCZ_Sheets.pdf

(equivalent to patches of 320 meters by 320 meters and 640 meters by 640 meters, respectively) were most effective in enhancing classification accuracy. However, the study found larger patch sizes decreased accuracy (Liu and Shi, 2020). Therefore, the image patches are extracted from Sentinel-2 imagery concerning the labels of the Local Climate Zone (LCZ) polygons.

The procedure described below is followed to prepare the dataset of the image patches for training a CNN model for LCZ classification. A sample of creating image patches of Sen-2 imagery from LCZ polygons can be seen in Figure 8.

- Patch Extraction Algorithm:
  - A function is developed to extract image patches from Sen-2 imagery aligned with the LCZ polygons.
  - The algorithm considers the patch size and the center point coordinates within each polygon.
- Patch Extraction Process:
  - The algorithm iterates through each LCZ polygon.
  - For each polygon, it determines the label associated with the climate zone.
  - The bounding box coordinates of the polygon are obtained.
  - The pixel resolution of the Sentinel-2 imagery is calculated.
  - The number of steps in the x and y directions is determined based on the bounding box and pixel resolution.
  - A grid of points is generated within the polygon, evenly spaced based on the patch size.
  - For each point in the grid, the algorithm checks if the point lies within the polygon.
  - If the point is within the polygon, an image patch is extracted from the Sentinel-2 imagery centered at the point.
  - The extracted image patch is saved to a directory named after the corresponding LCZ label.
- Dataset Generation:
  - The process generates a collection of image patches, each labeled with its corresponding LCZ climate zone.
  - These image patches collectively form the dataset for training the CNN model.

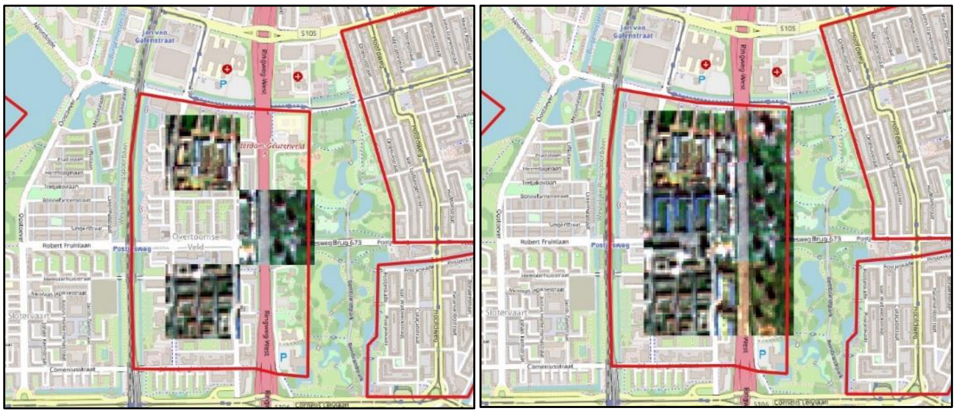


Figure 8: A sample of creating image patches of sen-2 imagery based on the LCZ polygons

4.2.4. Training and evaluation

The architecture of the designed model begins with rescaling input image patches using the Rescaling layer, which normalizes the pixel values by dividing them by the maximum value in the training set. This step ensures that the pixel values are in the range of 0 and 1. The subsequent layers consist of two Conv2D layers, with the first layer having 32 filters and the second layer having 64 filters, both using a 3x3 kernel size and ReLU activation. Using a compact kernel with limited dimensions aligns with the state-of-the-art models and contributes to the optimal use of parameters. BatchNormalization layers are added after each Conv2D layer to normalize the activations. MaxPooling2D layers with a pool size of 2x2 are applied for downsampling. The model then uses a GlobalAveragePooling2D layer to reduce the spatial dimensions and calculate the mean value of each feature map. Next, a Dense layer with 128 units and ReLU activation is added, which is regularized by a Dropout layer with a rate of 0.6 to prevent overfitting. Finally, a Dense layer with 15 units and softmax activation is added to output the class probabilities for the 15 local climate zone classes in the LCZ classification task. Figure 9 shows the CNN model architecture.

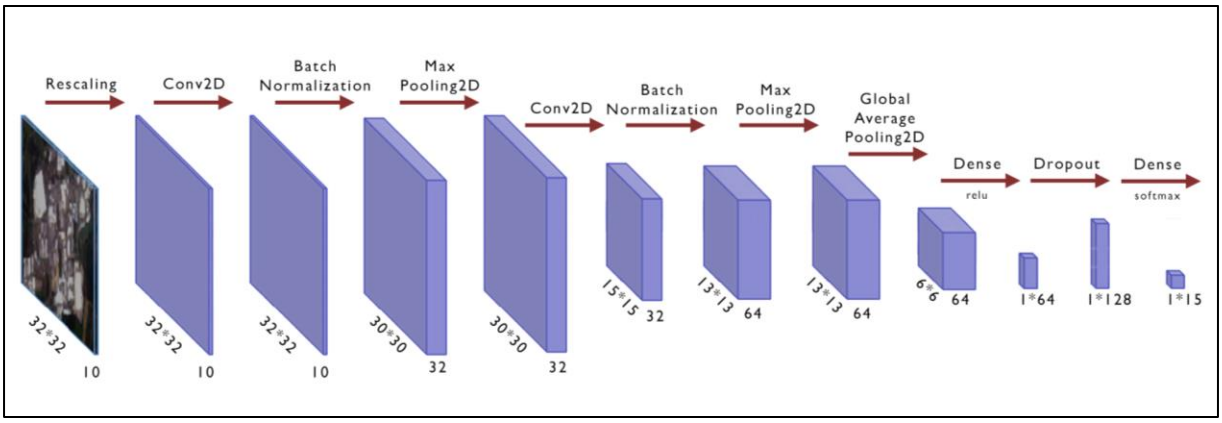


Figure 9: The CNN model architecture for LCZ classification, designed by the author

The following steps are followed to train the CNN model on the image patches dataset for LCZ (Local Climate Zones) classification.

- A function is defined to load the dataset from a specified directory. Then, it reads the images, stores them as a NumPy array, and assigns corresponding labels to another list. The data is then shuffled and split into training, validation, and test sets based on the 70 percent for training, 20 percent for validation, and 10 percent for the testing set.
- After that, the model is compiled with the Adam optimizer, sparse categorical cross-entropy loss function, and the accuracy metrics.
- Two callbacks are defined: `EarlyStopping` and `ModelCheckpoint`. The `EarlyStopping` callback monitors the validation loss and stops training if there is no improvement for a certain number of epochs. The `ModelCheckpoint` callback saves the weights of the best model based on the validation accuracy.
- An instance of the `ImageDataGenerator` class is created to augment the training data. It includes transformations for rotation of 90 degrees, horizontal flip, and vertical flip.
- The generator is fit on the training data using the `fit` method, which applies the data augmentation transformations to generate augmented batches during training.
- The model is trained using the `fit` method. The training data is passed through the data generator, and the validation data is provided for monitoring the model's performance.
- Finally, the trained model is saved for evaluating and predicting unseen images.

The two other state-of-the-art models, Resnet-50 and EfficientNet, are also trained on the prepared dataset. The steps of loading the dataset for training, compiling, and fitting the model follow the same procedure as the first model. Resnet-50 and EfficientNet are imported from TensorFlow library. The top layer in the models works as a mediator by defining the input shape compatible with the LCZ classification dataset. This value is set to (32,32,10), showing the dimension of image patches, followed by the number of bands, and a new input layer is created to add the specified input shape to the model.

The ResNet50 model is loaded, excluding the top layer, and the newly created input layer is used. This allows the ResNet50 model to accept input data with the defined shape. The weights of the pre-trained model are set to None, and the include-top argument is set to False. These adjustments are made to facilitate learning new weights for the LCZ dataset. To achieve the desired output, a GlobalAveragePooling2D layer, followed by a Dense layer with 128 units and ReLU activation, and finally, a Dense layer with 15 units (corresponding to the number of LCZ classes) and softmax activation is added to the top of the loaded ResNet-50 model (Figure 10). The exact process is done for the EfficientNet model as well. The figure below shows the modified model of state-of-the-art models for LCZ classification.

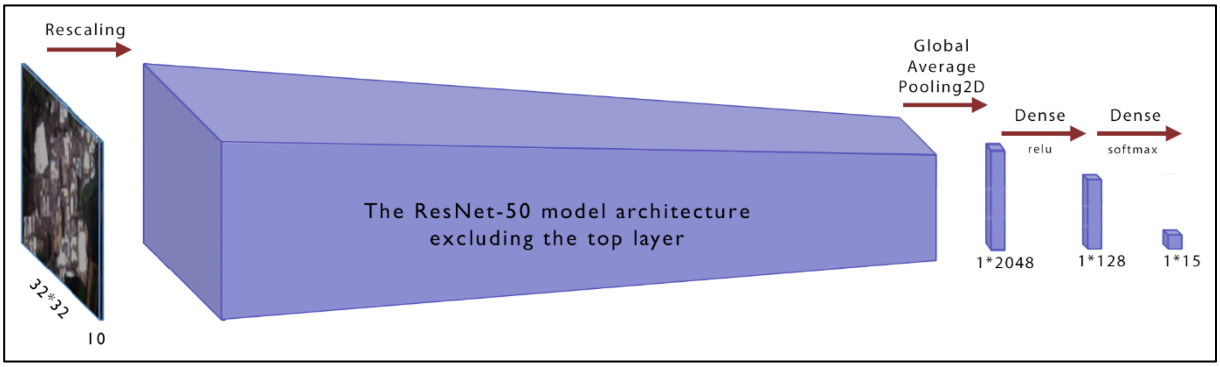


Figure 10: The modified CNN model architecture of ResNet-50 for LCZ classification

The trained three models are then evaluated using train-validation metrics and are fine-tuned by modifying hyperparameters. After achieving an acceptable performance, the models predict the testing set, and the performance of the models is evaluated using the classification evaluation metrics such as Recall, Precision, and F1-score, as well as the confusion matrix to select the best model for the classification of LCZ classes.

4.2.5. Prediction of LCZ classes on the entire study area

After selecting the most efficient models for predicting LCZ classes on the testing set, the trained model is applied to the whole study area. Below is the explanation of obtaining the final output of the first stage of the research, which is the LCZ map in the raster format, containing the LCZ category and the probabilities per pixel within the study area.

Creating Patches: The procedure begins by dividing the large imagery of Sen-2 into patches of size 10 by 10 pixels with a stride of 10, giving the spatial resolution of 100 meters. The purpose of creating patches is to facilitate efficient processing and analysis of smaller sets of large images.

Loading the Model: The algorithm loads a pre-trained LCZ prediction model, then calls the best-performed weights corresponding to the model using TensorFlow and Keras libraries.

Predicting LCZ for Patches: Next, the code accesses the image patches created earlier and predicts the probability of Local Climate Zones classes for each image patch. Additionally, the code determines the predicted LCZ class per patch by selecting the category with the highest probability. The predicted LCZ probabilities and class labels are calculated and stored in separate channels.

Merging Predicted Patches: The code proceeds to merge the individual patch files into a single mosaic TIFF file using the Rasterio library in Python. This step combines the predicted LCZ patches to generate a comprehensive prediction of LCZ for the entire study area. The final output is a tif file containing 16 bands. The first band represents the LCZ class, and the remaining bands inform the probability value of all classes for the corresponding pixel.

4.3. Stage 2: PM2.5 prediction

4.3.1. Data acquisition

In addition to local climate zone data obtained in the model's first stage, several other factors also contribute to the concentration of PM2.5. In order to implement a model for predicting this pollutant, the datasets related to PM2.5 measurements, meteorological data, and traffic data are involved in the model.
The PM2.5 values as the representative of air pollution for this research are acquired using the RIVM portal, including both official and individual air pollutants sensors. It is possible to filter municipalities for extracting the data. Therefore, we focus on the Amsterdam municipality area in the map by considering the sensors within the bounding box of Amsterdam for collecting the data. When choosing a specific municipality, a window shows the number of active sensors at the current time and different types of available measured values. Then, we filter the official sensors as the reference measurements and get their data and the charts of their distribution during the time as a benchmark. After that, the citizen-science sensors are selected considering the calibration and the plausibility of the sensor, the comparison of the distribution of the official measurements, and the availability of the data in the research period.

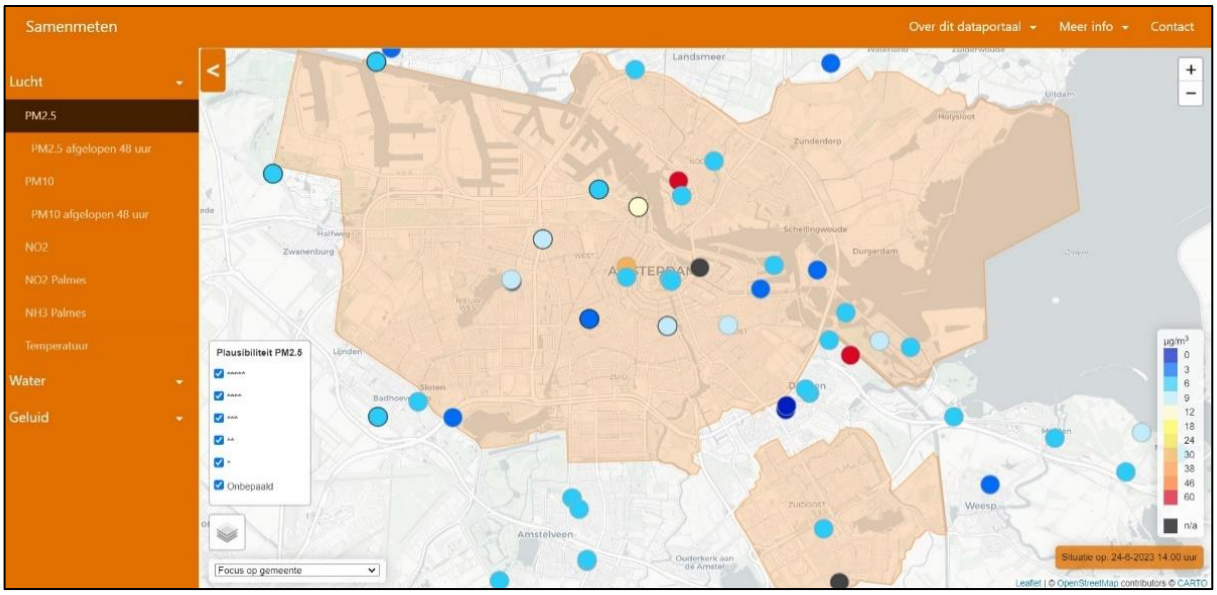


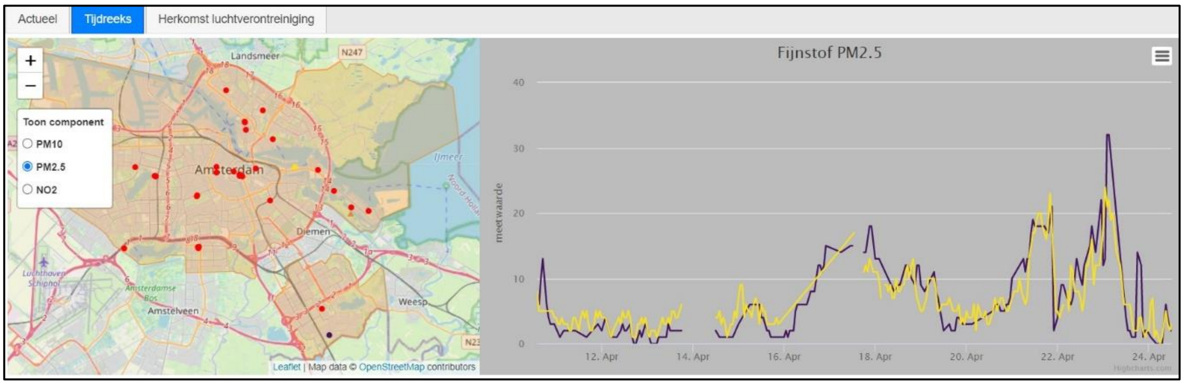
Figure 11: The overview of the extracting air pollutants sensors within the study area
source: https://samenmeten.rivm.nl/dataportaal/

Regarding the information provided on the RIVM website[5], currently, calibration is only implemented for NOVA SDS011 sensors. This is primarily due to the popularity of these sensors, making it convenient to compare a significant number of them with reference measurements. While the Sensirion SPS30 is being used increasingly for PM2.5 measurements, it is not yet as widely utilized. Furthermore, the SPS30 indicates lower sensitivity to humidity compared to the NOVA SDS011. Several studies have proven the reliability of NOVA SDS011 as a low-cost sensor for measuring fine particulate matter (Badura et al., 2018).

However, the reliability and calibration of the sensors can only be demonstrated for the last two weeks, and therefore, it cannot be definitively determined whether a sensor is reliable only based on the plausibility of stars indicating high or low values. Consequently, it is essential to examine the chart of each sensor alongside the official measurements to assess the sensor's reliability for further use. Two options

[5] https://samenmeten.nl/dataportaal/kalibratie-van-fijnstofsensoren

are available for chart analysis. Firstly, when choosing a specific municipality, a window is opened in which we can compare the plot of multiple sensors measurement in the selected city for the last two weeks (See Figure 12). Secondly, The measurement distribution of each sensor for the research period is checked with the distribution of the closest official sensor to the selected sensor in the map. Finally, the sensors' data covering all or most of the research period is downloaded in a CSV format.



Toon component
○ PM10
● PM2.5
○ NO2

Figure 12: The possibility of comparing the sensor's concentration over the last two weeks on the RIVM portal source: https://samenmeten.rivm.nl/dataportaal/

One of the limitations to obtaining sensors' data from this portal in the way described above is that the coordinates of sensors do not exist in the air quality data report. To access the coordinates of the sensors, first, the id number of each sensor is obtained using the following URL through API:
https://api-samenmeten.rivm.nl/v1.0/Things?$filter=startswith(name, 'the name of the sensor e.g., LTD_34577').
After that, by putting the id number in the URL below, the coordinates of the sensor can be reached on the web: https://api-samenmeten.rivm.nl/v1.0/Things(the id number of the sensor, e.g., 2954)/Locations

The information for the sensor's location is required later for assigning the LCZ class to the sensor.

The other required data for this stage is the meteorological data. The official measurement of various variables for the hourly values in the Netherlands is available on the KNMI website[6]. Based on meteorological station locations, the Schiphol station (station 240) is selected as a data source for Amsterdam. The weather-related data from the low-cost sensors are collected with the same procedure as PM2.5 data extraction.
Regarding traffic data, the road data of Amsterdam from the open street map is extracted using the OSMdownloader plugin in QGIS. The data attribute includes the road types based on road hierarchy, which is used for preparing traffic data.

4.3.2. Data preparation

Preparing the data for the second stage is done for PM2.5 measurements which is a target variable of the model, the local climate zones sample data, and image patches for training from sen-2 imagery and LCZ sample data. In this step, pre-processing actions are taken on each dataset. Then, the dataset is merged to create a single tabular data for modeling. Finally, the feature engineering process is conducted to make meaningful features of the prepared variables as the inputs for machine/deep learning algorithms.

[6] *https://www.knmi.nl/nederland-nu/klimatologie/uurgegevens*

Air pollution pre-processing

It is crucial to perform data-cleaning techniques to achieve the dataset's reliability and accuracy for modeling purposes. Exploring the measurements of PM2.5 values within sensors, we observe that some low-cost sensors often report extremely high values, such as 1200. However, when comparing these readings to the measurements from an official sensor located nearby and taken at the same datetime, we find that the official sensor records much lower values. Two approaches are implemented for cleaning the data related to PM2.5 in order to provide a more reliable dataset. First, the measurements with negative values are removed from the dataset. Then, the maximum value of PM2.5 among all official sensors is defined as a threshold for defining the outliers. Official sensors are typically calibrated to meet specific quality standards. By using this value, we establish a reference point for determining outliers within the data obtained from other sensors. Therefore, all measurements higher than this value are removed within low-cost sensors. This action is necessary to ensure that the dataset used for modeling purposes is consistent and free from unreliable measurements. Finally, the PM2.5 files, including the DateTime and PM2.5 values, are combined into a single data frame.

Meteorological data preparation

The pre-processing for official measurements includes modifying the format of the values and extracting the variables defined for this research as weather-related predictors. However, more actions are required for preparing temperature and humidity values from low-cost sensors as the additional independent variables. After ensuring the data's reliability and considering the spatial distribution of the sensors to ensure that they cover the study area, four other sensors for temperature and humidity observations are extracted. The values obtained from the additional temperature and humidity sensors and the official sensor(Schiphol) are assigned to all PM2.5 sensors based on their closest distance. This task is performed using the 'distance to nearest hub' function available in the processing toolbox of QGIS. The closest sensor's values are assigned by calculating the minimum distance between each PM2.5 sensor and the meteorological data sensors, as shown in Figure 13.

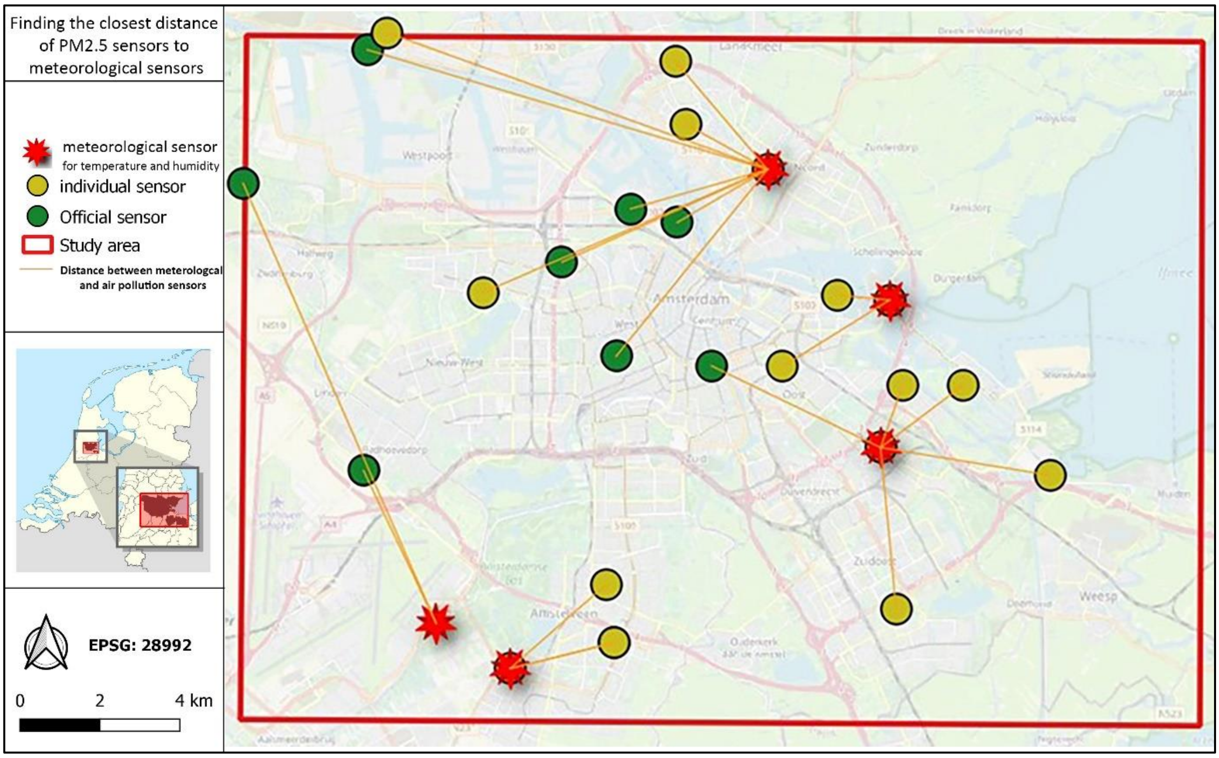

Finding the closest distance of PM2.5 sensors to meteorological sensors
meteorological sensor for temperature and humidity
individual sensor
Official sensor
Study area
Distance between meterologcal and air pollution sensors
EPSG: 28992
0
2
4 km
Amsterdam
Westpoort
Nieuw-West
Zuid
Oost
Amstelveen
Weesp
Zuidoost

Figure 13: spatial analysis for finding the closest distance to meteorological sensors

Two new variables for temperature and humidity are added to the dataset. In the data-cleaning process, if there are missing values, they are removed from the dataset. Additionally, the corresponding values from the Schiphol station (which is an official sensor) are added to the attributes for the missing DateTime values.

Traffic data preparation

As the traffic data is not available for the study area, we assume the combination of the effect of road networks and time of the day, as well as time of the year, would allow the model to capture the effect of traffic. The approach used in creating traffic data from road networks by Ghaemi et al. (2018) is employed to get such a proxy for traffic data. In this approach, the relationship between air pollution caused by traffic and the distance to roads is assumed to be linked. To explore this relationship, a technique called kernel density estimation (KDE) is used. This approach involves creating a raster map that indicates the density of nearby roads, providing a visual representation of the concentration of roads in the surrounding area. Moreover, the type of road based on their importance is considered a weight for KDE (Ghaemi et al., 2018).

Regarding the road types available in the road data, the type of roads related to vehicle transports is filtered as follows: motorway, trunk, primary, secondary, tertiary, unclassified, residential, living street, service, and pedestrian. Then, weights are assigned to road types based on their importance to have an assumption about traffic flow. Therefore, the most important roads are given higher weights, ranging from 10 to 1. Then, using Kernel Distance Estimation in ArcMap, a raster output containing the density of the surrounding road is created. The function considers the weight of roads, and the kernel density is calculated with a maximum distance of 300 meters. Figure 14 shows the road density output raster as the traffic data representative data. Finally, the pixel values surrounding air pollution sensors are assigned as the traffic value in the dataset.

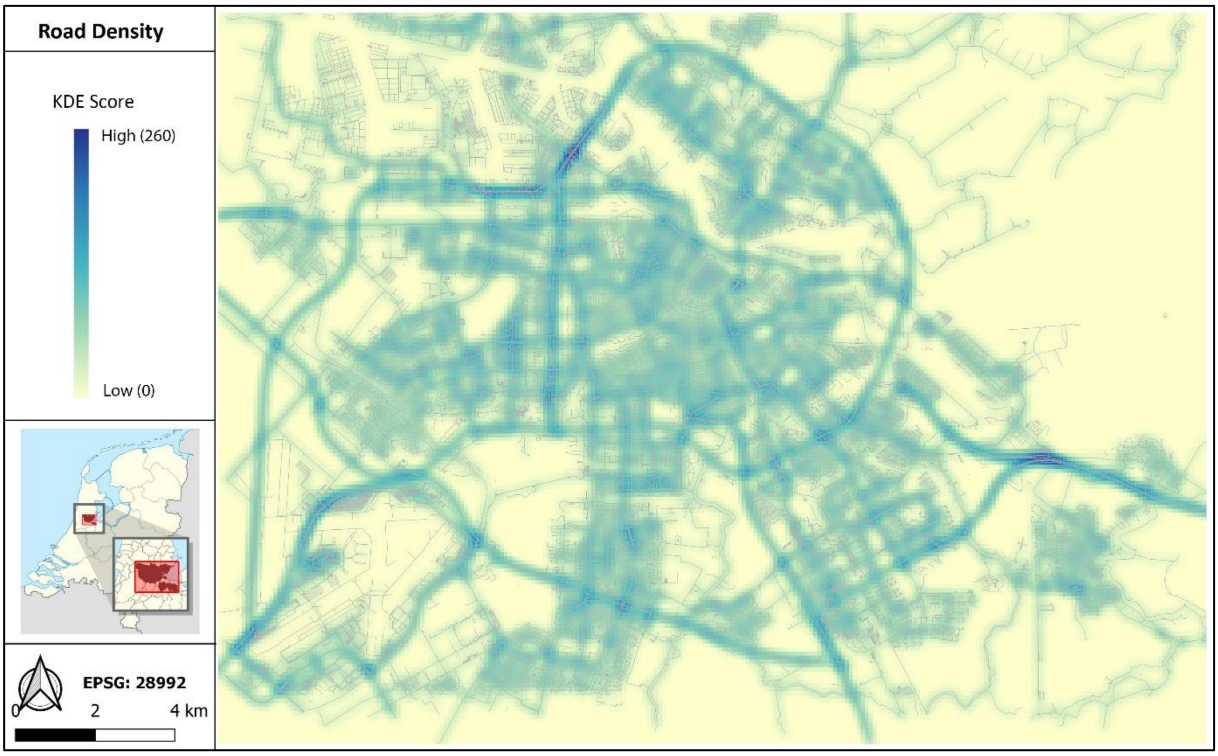


Figure 14: The road density map from KDE function as the representative for traffic data in the study area

Merge the datasets

In the next step of data pre-processing, all the data from variables are merged into a single tabular dataset. This step is required for training machine learning models. To make the dataset for modeling, the dataframe, including the combined PM2.5 measurements, is considered as the basis of the dataset. Then, the meteorological data from the official measurements are added to the dataframe. The DateTime column is used as the common attribute for merging the datasets. The additional temperature and humidity values from low-cost sensors are then added to the dataframe based on DateTime as well as the sensor name. Finally, the traffic data, LCZ category, and the probabilities of all classes are joined to the dataframe using both DateTime and sensor name as the common columns, representing spatial variables. Figure 15 shows the workflow diagram of the procedure for merging the dataset. An overview of the combined dataset for PM2.5 prediction can be seen in Table 3.

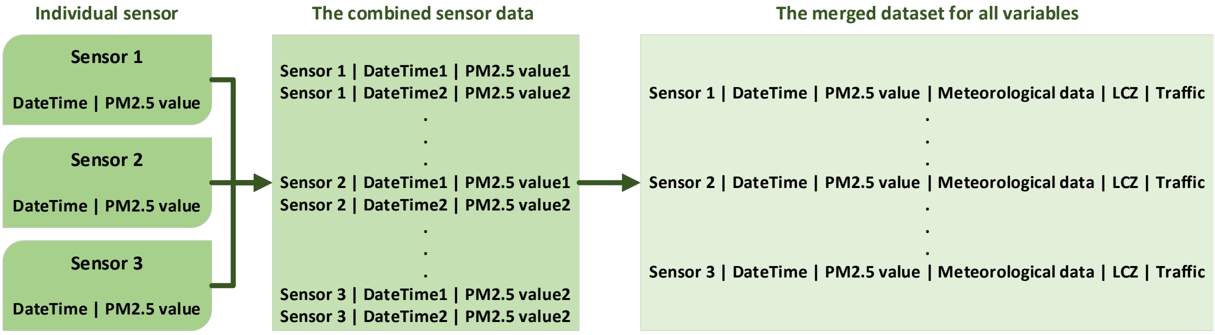
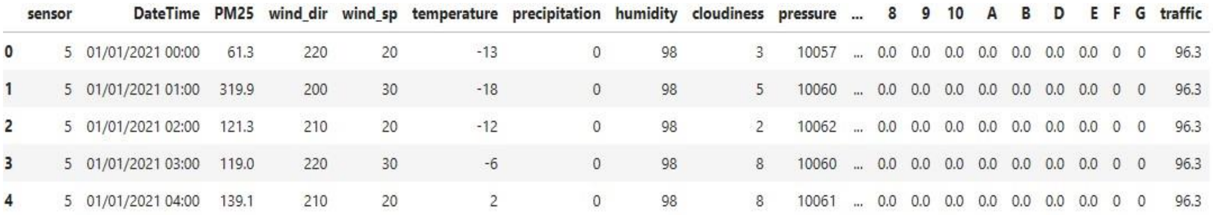

Individual sensor
The combined sensor data
The merged dataset for all variables
Sensor 1
DateTime | PM2.5 value
Sensor 2
DateTime | PM2.5 value
Sensor 3
DateTime | PM2.5 value
Sensor 1 | DateTime1 | PM2.5 value1
Sensor 1 | DateTime2 | PM2.5 value2
Sensor 2 | DateTime1 | PM2.5 value1
Sensor 2 | DateTime2 | PM2.5 value2
Sensor 3 | DateTime1 | PM2.5 value2
Sensor 3 | DateTime2 | PM2.5 value2
Sensor 1 | DateTime | PM2.5 value | Meteorological data | LCZ | Traffic
Sensor 2 | DateTime | PM2.5 value | Meteorological data | LCZ | Traffic
Sensor 3 | DateTime | PM2.5 value | Meteorological data | LCZ | Traffic

Figure 15: The workflow of merging dataset in the preprocessing step for PM2.5 modeling

Table 3: The overview of combined dataset for PM2.5 prediction model



	sensor	DateTime	PM25	wind_dir	wind_sp	temperature	precipitation	humidity	cloudiness	pressure	...	8	9	10	A	B	D	E	F	G	traffic
0	5	01/01/2021 00:00	61.3	220	20	-13	0	98	3	10057	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0	0	96.3
1	5	01/01/2021 01:00	319.9	200	30	-18	0	98	5	10060	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0	0	96.3
2	5	01/01/2021 02:00	121.3	210	20	-12	0	98	2	10062	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0	0	96.3
3	5	01/01/2021 03:00	119.0	220	30	-6	0	98	8	10060	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0	0	96.3
4	5	01/01/2021 04:00	139.1	210	20	2	0	98	8	10061	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0	0	96.3

Feature Engineering

In the final step of data preparation, some actions related to feature engineering are performed for some variables to ensure that the model interprets the data correctly. The recommendations mentioned in the documentation for time series forecasting in TensorFlow[7] are applied for feature engineering. Some modifications are required to make the distribution of the variable more compatible with the model. For example, the wind direction values in degrees (90, 180, 270, etc.) are not a good representation of the model training because the inputs should not be far apart and should avoid any sudden transitions. The existing values for wind direction do not accurately represent the circular nature of wind direction and may not reflect the actual smooth changes in wind patterns. Based on the recommendation, making a vector of wind direction and wind speed makes learning the pattern easier for the model which is visualized in Figure 16.

[7] https://www.tensorflow.org/tutorials/structured_data/time_series

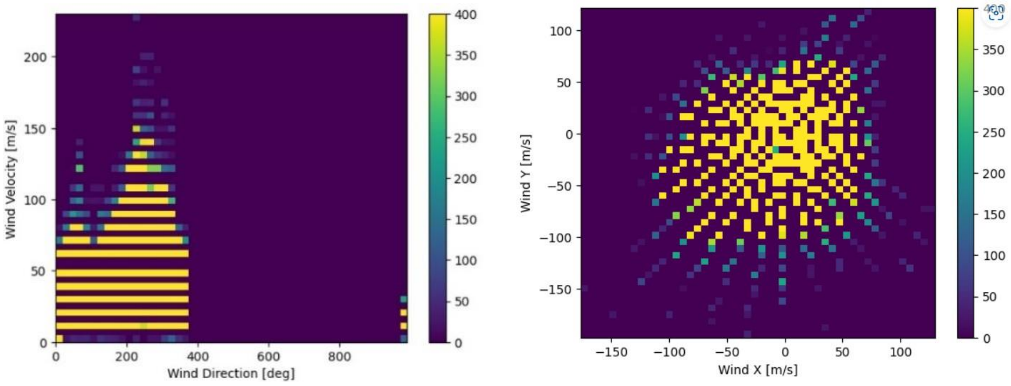


Figure 16: The distribution of wind date before(left) and after(right) feature engineering

Regarding the DateTime variable, which is the core of the temporal information for the model, its current format is not proper for interpretation by the model. Since it represents periodic data, it exhibits distinct daily and yearly patterns. One effective method to handle these frequent patterns involves utilizing sine and cosine transformations to isolate the "Time of day" and "Time of year" signals. This technique allows the model to capture crucial frequency features and obtain meaningful signals for analysis.

4.3.3. Training and evaluation

The prediction of PM2.5 concentration is modeled using the eXtreme Gradient Boosting (XGBoost) model as one of the effective ensemble learning models, as well as the long short-term memory (LSTM) model as the time series model. Then, the model with the more acceptable performance is chosen for further analysis of urban form impact on air pollution.

XGBoost model

A supervised XGBoost model is trained for PM2.5 prediction using a regression task. The independent variables in this model are those prepared in the previous steps: meteorological data, LCZ label and probabilities data, traffic data, and time stamp data of the time of day and time of year. The target variable is PM2.5 concentration.

XGBoost offers three strategies to address overfitting and enhance prediction accuracy effectively. These techniques consist of a regularized objective, shrinkage, and column subsampling. The regularized objective helps discourage complexity, enabling the selection of a more straightforward yet powerful model. Shrinkage reduces the influence of each tree, allowing subsequent trees to make incremental improvements. Regarding efficiency and preventing overfitting, column subsampling proves to be more efficient for using memory than row subsampling (Lin et al., 2022).

The following steps are taken for modeling the dataset with the XGBoost model:

- **Split train/validation and test set:** First, the dataset is split into training/validation and test sets. The two last months of the dataset (The hourly values from 01-11-2021 to 31-12-2022) is cut for the testing set, and the remaining dataset is used for the train/validation set.
- **Split train and validation set:** Then, the dataset for training/validation is split into training and validation based on sensor names. 80 percent of sensors (20 sensors) are used for training, and 20 percent (5 sensors) are used for validation. Random sampling is then used to select the sensors for each set. Next, the input features and output target (PM2.5 values) are extracted from the respective dataframes, and the data is transformed into arrays (X_train, y_train, X_valid, y_valid, X_test, y_test) for training and evaluating the XGBoost model. This methodology allows for random and sensor-based partitioning of the dataset, ensuring diverse sensor representation in each set for robust model training and evaluation.
- **Normalize the data:** After that, all X_columns (input features) from the training, validation, and testing set are normalized by subtracting the mean and dividing by the standard deviation of the

input features in the training set. This normalization process ensures that all input features have a similar scale, which can improve the performance and stability of machine learning models.

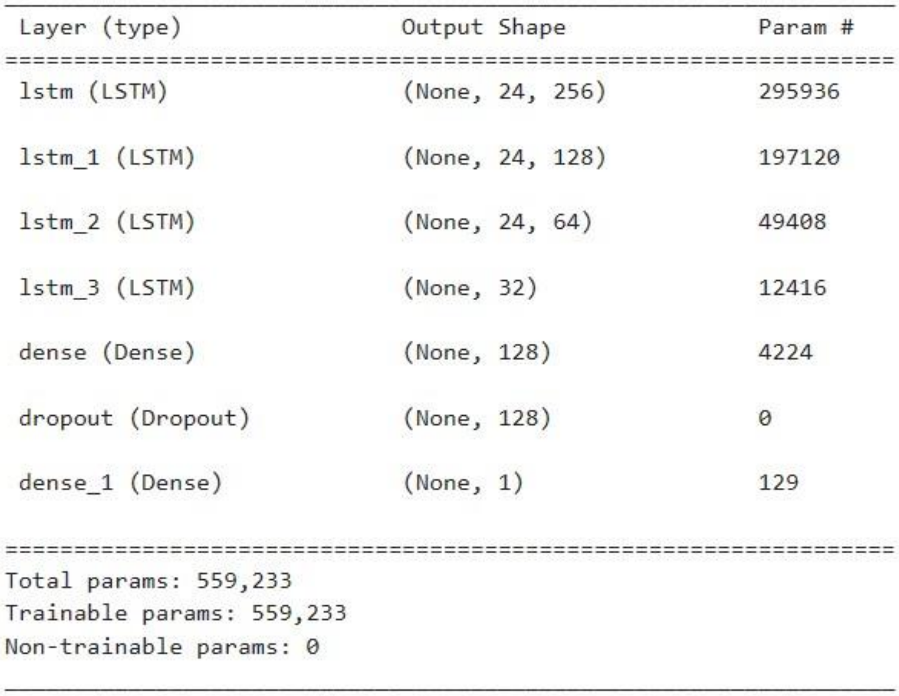
- **Train XGBoost model:** An XGBoost regression model is constructed and trained. The model is defined with specific parameters, including the objective function, the number of estimators, maximum depth, learning rate, and early stopping rounds. The training process includes monitoring the model's performance on the validation set. The model is trained, and the training and validation root mean squared error (RMSE) metrics are extracted from the results.
- **Fine-tune the model hyperparameters:** Finally, a grid search approach is used to optimize the hyperparameters of the XGBoost regression model. The GridSearchCV class from the Scikit-learn library in Python is employed, configured with the model, hyperparameter grid, 5-fold cross-validation, and the negative mean squared error as the scoring metric. The grid search is then fit to the training data. Afterward, the best hyperparameters and model are obtained from the grid search results, which are used for the final training.

LSTM model

In the other method, LSTM neural network model is designed for predicting PM2.5. The input and target variables are the same as XGboost model. The LSTM model is developed in the TensorFlow environment of Python. Several arrangements of layers are examined to achieve the most proper model for the dataset.

The final model architecture for time series forecasting of PM2.5 starts with a series of LSTM layers, which allow the network to learn and capture long-term dependencies in the sequential data. In this case, there are four LSTM layers, each with a decreasing number of units: 256, 128, 64, and 32. The 'return_sequences' parameter of the first three layers of LSTM is set to True, ensuring that the output of each LSTM layer is fed as input to the next LSTM layer, maintaining the sequential nature of the data.

Following the LSTM layers, there are dense layers for prediction. The dense layers provide additional learning capabilities for the model. The first dense layer has 128 units with a ReLU activation function, which introduces non-linearity to the model. A dropout layer with a rate of 0.5 is added to prevent overfitting and improve regularization by randomly dropping out units during training. Finally, the output layer consists of a single neuron with the default activation function, which is the linear function, predicting a continuous output. Figure 17 shows the LSM model architecture for PM2.5 prediction.



```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, 24, 256)           295936

 lstm_1 (LSTM)               (None, 24, 128)           197120

 lstm_2 (LSTM)               (None, 24, 64)            49408

 lstm_3 (LSTM)               (None, 32)                12416

 dense (Dense)               (None, 128)               4224

 dropout (Dropout)           (None, 128)               0

 dense_1 (Dense)             (None, 1)                 129

=================================================================
Total params: 559,233
Trainable params: 559,233
Non-trainable params: 0
_________________________________________________________________
```

Figure 17: The LSTM model architecture for PM2.5 prediction, designed by the author

The following steps are followed to train the LSTM model for forecasting PM2.5:

- **Split train and test set:** Similar to the splitting strategy in the XGBoost model, the two last months of the dataset are considered for the test set, and the remaining data belong to the train set.
- **Normalize the data:** Similar to the normalization approach in the XGBoost model, the input variables of the train and test set are normalized with respect to the mean and standard deviation of predictors in the train set.
- **Windowing:** As we deal with time series prediction, we need to provide the data in the sequential windows for the model. The windowing procedure is an exclusive step for implementing the LSTM model. A 'WindowGenerator' class is responsible for preparing data for training the LSTM model. It takes three important parameters: 'input_width', 'label_width', and 'shift'. These parameters determine the configuration of the input and label windows used for training the LSTM model, which is explained below:
  - input_width: It determines the number of time steps to be used as input features for the model. This parameter defines the length of the input sequence that the LSTM model will process at each time step.
  - label_width: It determines the number of time steps in the future for which the model needs to make predictions. This parameter defines the length of the output sequence that the model is expected to predict.
  - shift: It represents the time shift or the time gap between sequential input and label windows. This parameter determines how much the input and label windows are offset from each other. It allows the model to learn temporal relationships and make predictions at different temporal intervals.

  The WindowGenerator also takes parameters for input train data and test data, label column to define the target variable, followed by the stratification column parameter. When stratification is applied, the data is divided into groups based on the unique values in this column. Each group represents a distinct subset of the data. The separate datasets are then created for each group. The column of sensor names is defined as the stratification column in this model.
- **Split train and validation set:** After applying the windowing function on the train and test set, the train set is split to train and validation set with the proportion of 80 and 20 percent, respectively.
- **Compile and train LSTM model:** Finally, the LSTM model is compiled using Mean Squared Error (MSE) as the loss function and Mean Absolute Error (MAE) as the metrics, followed by setting the optimizer and learning rate. The model is then fit, where the training dataset and validation dataset are provided. The weight of the best epoch performance is saved for evaluation and further analysis.

Model Evaluation and urban form impact analysis

After obtaining the best-trained model of both XGBoost and LSTM approaches, the test set is predicted by both models. Then, the performance of the two models is evaluated using metrics such as R-squared, means square error (MSE), and route means, square error (RMSE). Then, the analysis of the LCZ classes' impact is done by prediction models with the best performance. Two approaches, including Sensitivity analysis and Feature importance analysis, are employed to investigate the effect of LCZ on PM2.5, which are explained below:

Sensitivity analysis: This analysis aims to discover to what extent the LCZ, as the predictor, influences the concentration of PM2.5 in the model. Several alternatives for involving input variables are examined. The results of the performance metrics of the model clarify the contribution of LCZ on the level of PM2.5.

Feature importance analysis: This analysis intends to investigate the significance of each LCZ type on PM2.5 prediction in the model. In this method, the probabilities of LCZ act as the representative for the corresponding category beside the other predictors, including traffic, meteorological data, and time stamp data. The results of the feature importance are reported as the scores. The score represents the proportion or percentage of the overall importance related to that variable within the model. The higher scores indicate the higher importance of the independent variable for predicting PM2.5.

Finally, a heat map is created in the study area to visualize the relative importance of LCZ categories and their potential impact on the concentration of PM2.5 based on the results of the feature importance. This heat map provides a comprehensive overview of areas likely to experience higher levels of PM2.5, as influenced by the different LCZ categories.

5. RESULTS AND DISCUSSION

This chapter is divided into three sections, including local climate zone classification, PM2.5 prediction, and the effect of LCZ on PM2.5. Each section starts by presenting the research results on the related subject. Then, the interpretation of the results, the main reasons for achieving them, and their limitations will be discussed.

5.1. Local climate zones classification

Three CNN models, namely the designed CNN, ResNet-50, and EfficientNet, were trained on the image patches, including the LCZ classes. Based on the provided training area, 15 classes out of 17 classes of the LCZ scheme were found in the study area. Class 7 represents the light-weight low-rise area mostly observed in the slums area, and class C belongs to Bush and Scrub's natural land cover, which does not exist in the study area. The training set includes 1072 samples, followed by 229 and 231 samples for testing and validation sets, respectively. Table 4 shows the number of samples per class for all datasets.

Table 4: the distribution of samples per class within the training, validation, and testing set

Class label	Class name	Train Count	Validation Count	Test Count
Built types				
1	Compact high-rise	33	5	9
2	Compact mid-rise	90	24	22
3	Compact low-rise	17	6	3
4	Open high-rise	67	12	19
5	Open mid-rise	106	17	17
6	Open low-rise	64	8	23
8	Large low-rise	104	21	22
9	Sparsely built	41	7	13
10	Heavy industry	45	12	7
Land cover types				
A	Dense trees	55	14	12
B	Scattered trees	46	11	13
D	Low plants	141	31	25
E	Bare rock or paved	89	17	17
F	Bare soil or sand	26	6	6
G	Water	148	38	23

In deep learning modeling, the number of samples per class is recommended to be in the same range as possible to get the more proper dataset for a model to learn the pattern and also reach more reliable evaluation results. However, providing the labeled samples in this subject is a time-consuming process. Moreover, based on the size of the image patches covering 320 meters by 320 meters, it is hard to find a homogeneous area for the classes that do not cover a considerable area on the ground. For example, the land cover types such as water and dense trees are usually expanded over an area. In contrast, built types classes such as compact high-rise and open high-rise classes are hard to be covered in a polygon with enough size for sampling, especially in the case of Amsterdam.

The designed CNN model LCZ classification was then trained on the training set and was monitored by the validation set. The model was fine-tuned using changes in hyperparameters and redoing the training process several times. Table 5 shows the hyperparameters of the final model.

Table 5: The hyperparameters of fine-tuned CNN model for LCZ classification

Hyperparameter	**Value**
Number of Convolutional Layers	2
Number of Filters	32 and 64
Kernel Size	(3, 3)
Activation Function	ReLU
Pooling Type	MaxPooling
Pooling Size	(2, 2)
Number of Dense Layers	2
Number of Units in Dense Layers	128 and 15
Dropout Rate	0.5
Learning Rate	Adam optimizer default value
Loss Function	Sparse Categorical Crossentropy
Batch Size	32
Number of Epochs	500
Early Stopping Patience	300
Checkpoint Saving Criteria	Validation accuracy (monitor='val_accuracy', mode='max')
Image Data Augmentation	Rotation Range: 90 degrees, Horizontal Flip: True, Vertical Flip: True, Fill Mode: 'nearest'

The two other state-of-the-art models, ResNet-50 and EfficientNet, were also trained with the same values as the designed CNN model for the number of epochs, batch size, loss function, learning rate, and early stopping hyperparameters.

Based on the analysis of the train-validation loss and train-validation accuracy metrics, Which is shown in Figure 18, all three models present acceptable metrics during the training. The training loss, which measures the error between the predicted and actual values, gradually decreases as the models are trained over multiple epochs. This indicates that the models are learning and improving their ability to make accurate predictions. However, some sudden jumps are observed in the validation loss during training the ResNet-50 model, followed by a few jumps in loss metrics of the EfficientNet model. These jumps in the validation loss suggest that the models' performance on unseen data temporarily worsens. This can happen because the models become too focused on the training data and struggle to make accurate predictions on new examples, resulting in overfitting. Similar to the pattern of loss metrics, the validation accuracy in the designed CNN model is smoothly increased during the training epochs. However, a sequence of fluctuation is observed during the accuracy metrics of both state-of-the-art models. In particular, the ResNet-50 model exhibits more frequent fluctuations compared to the EfficientNet model. Validation accuracy fluctuations can be due to the considerable number of trainable parameters in the cutting-edge architectures, also complex patterns or outliers in the data, and the model's sensitivity to features and data variations.

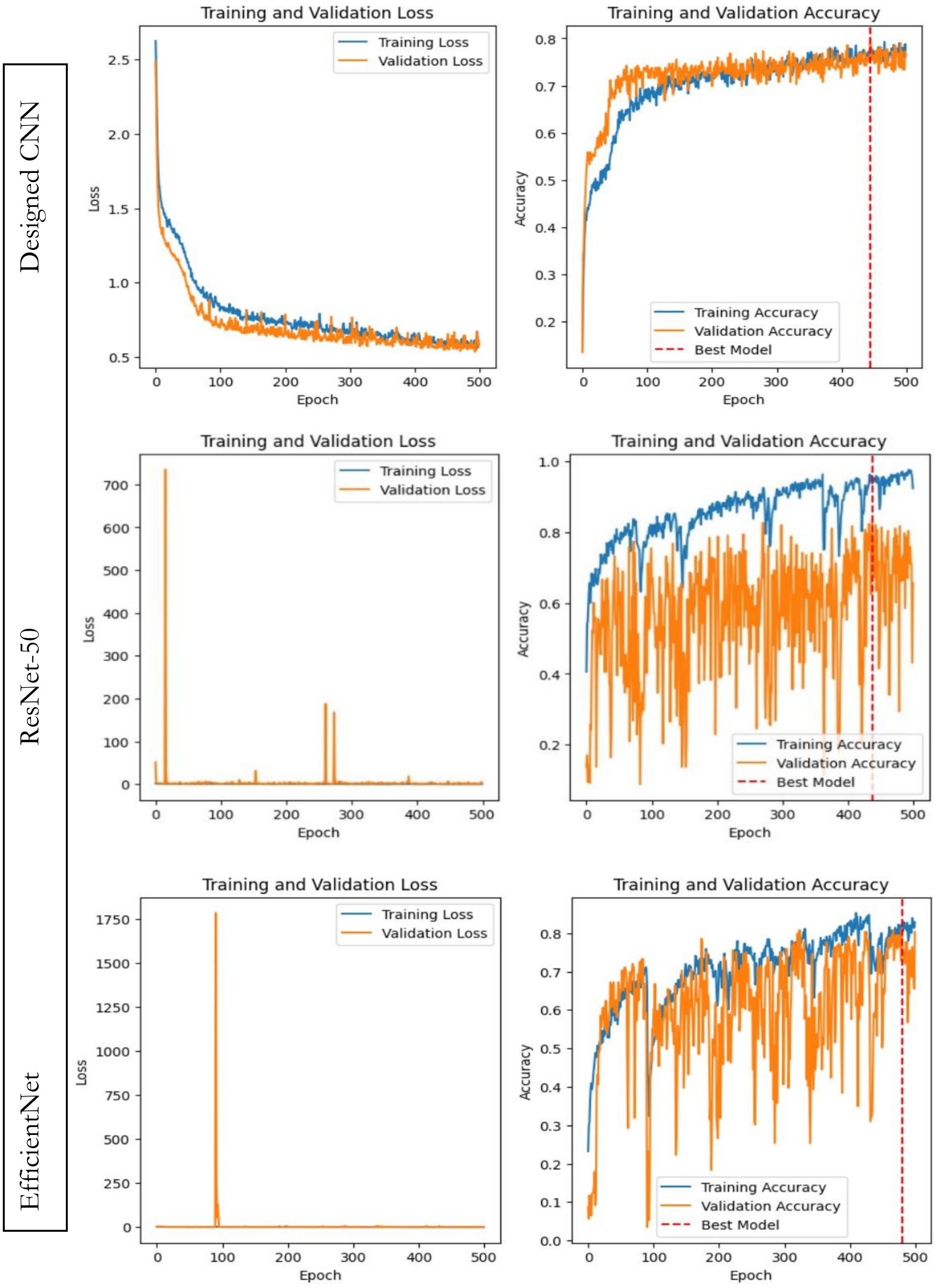


Figure 18: the metrics monitoring of training LCZ classification models

After achieving the best performance of trained models, we used three models to predict the same testing set. As shown in Table 6, the ResNet-50 model earned the highest validation accuracy of 0.8366, outperforming the designed CNN model (0.7948) and the EfficientNet model (0.8122). Compared to the other models, the ResNet-50 model also achieved the highest test accuracy of 0.8766, indicating its superior performance in predicting LCZ on unseen data. The higher accuracy of the Resnet-50 model on the testing set compared to the validation set proves that this model can generalize and predict unseen patterns.

Despite having a significantly smaller number of trainable parameters, the designed CNN model achieved acceptable performance. The model's architecture and fine-tuning actions seem to be effective in capturing relevant features for LCZ classification. The model's relatively simple structure may have contributed to its ability to avoid overfitting and maintain a proper balance between model complexity and generalization performance.

In contrast, although the EfficientNet model has a relatively high number of trainable parameters, its validation and test accuracy performance is slightly lower compared to ResNet-50. This suggests that having a larger number of trainable parameters does not always guarantee better performance.

Table 6: the overall accuracy and loss of three trained CNN models for LCZ classification

Model	Trainable Parameters	Validation Loss	Validation Accuracy	Test Accuracy
Designed CNN	31663	0.5858	0.7948	0.76
ResNet-50	23820751	0.7418	0.8366	0.8766
EfficientNet	64120735	0.7243	0.8122	0.7878

The metrics of the classification evaluation report for predicting the testing set by three models (Table 7) show that ResNet-50 achieved the highest Recall, indicating its ability to identify many positive instances correctly. It also reached a value of 0.89 in the Precision metric, which measures the accuracy of positive predictions. High Precision in this model indicates its ability to minimize false positive examples. The F1-score is the other classification metric, which combines Recall and Precision. It was also the highest for ResNet-50. This implies a balanced performance in terms of identifying positive instances while avoiding false positives. The Designed CNN model achieved slightly lower metrics compared to ResNet-50. EfficientNet shows moderate performance.

Table 7: The evaluation metrics report for LCZ classification by three CNN models

Model	Recall	Precision	F1-score
Designed CNN	0.75	0.76	0.75
ResNet-50	0.92	0.89	0.89
EfficientNet	0.8	0.79	0.78

Finally, the confusion matrix results for three models (Figure 19) are discussed to compare the performance of the models.

Regarding the two types of categories in LCZ classes for built and land cover types, the prediction of land cover types, which are mostly related to natural categories on the ground, is done more accurately compared to the built type in all models. This would be due to the simplicity of the patterns of these classes for recognition by the model. The categories “low plants” (D), “Bare rock or paved” (E), “Bare soil or sand” (F), and water (G) are perfectly classified in the designed CNN and ResNet-50 models, followed by some misclassification between class E and F in Efficientnet model. There is also misclassifications between "Scattered trees" (B) and "Dense trees" (A).

The results of built types categories vary within the models. All classes of built types have a level of accuracy in ResNet-50. However, the class of "Compact low-rise" (3) is not recognized in two other models as the true prediction. In two other models, this category is mostly misclassified as "Open high-rise" (4). When it comes to classifying "Compact high-rise" (1) and "Compact mid-rise" (2), all three models show relatively high accuracy. However, the designed CNN model struggles to distinguish between "Compact high-rise" (1). Regarding categories in the open-built types (4,5,6), "Open mid-rise" (5) are well classified into three models. However, there is some misclassification for "Open high-rise" (4) and "Open low-rise" (6) within these three categories. "large low-rise" (8), and "sparsely-built" (9) are classified mostly truly. In contrast, predicting the "Industrial" (10) category indicates challenges for three models. It is mostly misclassified with "large low-rise" (8), and "Open low-rise" (6). Especially, a considerable amount of misclassification with “Bare rock or paved” (E) is observed in the designed CNN model. The shared features with other categories contribute to the difficulty in distinguishing "Industrial" structures.

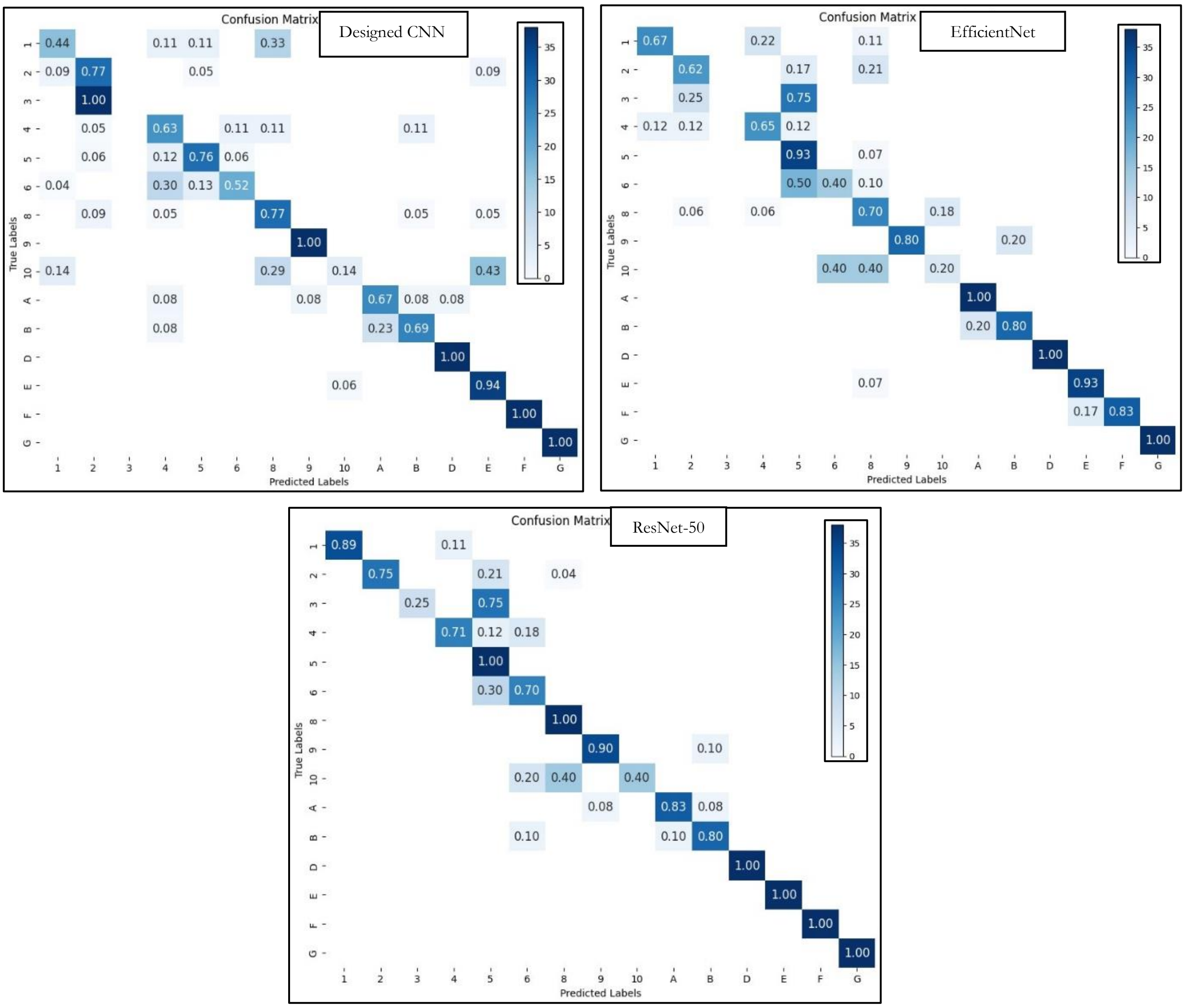


Figure 19: The confusion matrix of LCZ classification by three CNN models

In terms of the effect of LCZ (Local Climate Zone) classes on the accuracy and predictions of the models, we can observe the following patterns:

- Distinctions between different types of high-rise buildings in both compact and open types seem to be challenging for all models. This could be due to similarities in their visual features or variations in the dataset.
- Misclassifications between high-rise and mid-rise buildings are commonly observed, indicating difficulty in discerning subtle differences in building types. This suggests considering elevation-related factors as the supporting variable for training the model.
- The distinction between "Open low-rise" and other low-rise classes seems relatively more straightforward for the models.
- Recognizing between different types of vegetation classes (e.g., "Dense trees" and "Scattered trees") is sometimes challenging for the model. Involving vegetation-related factors such as NDVI in the training process can help face this issue.

After assessing the results of the LCZ classification, the ResNet-50 is selected as the best model for predicting the entire study area. To visually see the prediction of the model, the prediction of one sample per label can be seen in Figure 20.

Finally, the LCZ map for the bounding box of Amsterdam was created by predicting categories with the ResNet-50 model (Figure 21). The output is a raster map with a spatial resolution of 100 meters. The raster has 16 bands, and each pixel of the raster represents the LCZ class stored in the first channel. The probabilities of all 15 classes in a certain pixel are accessible in channel 2 to channel 16, respectively.

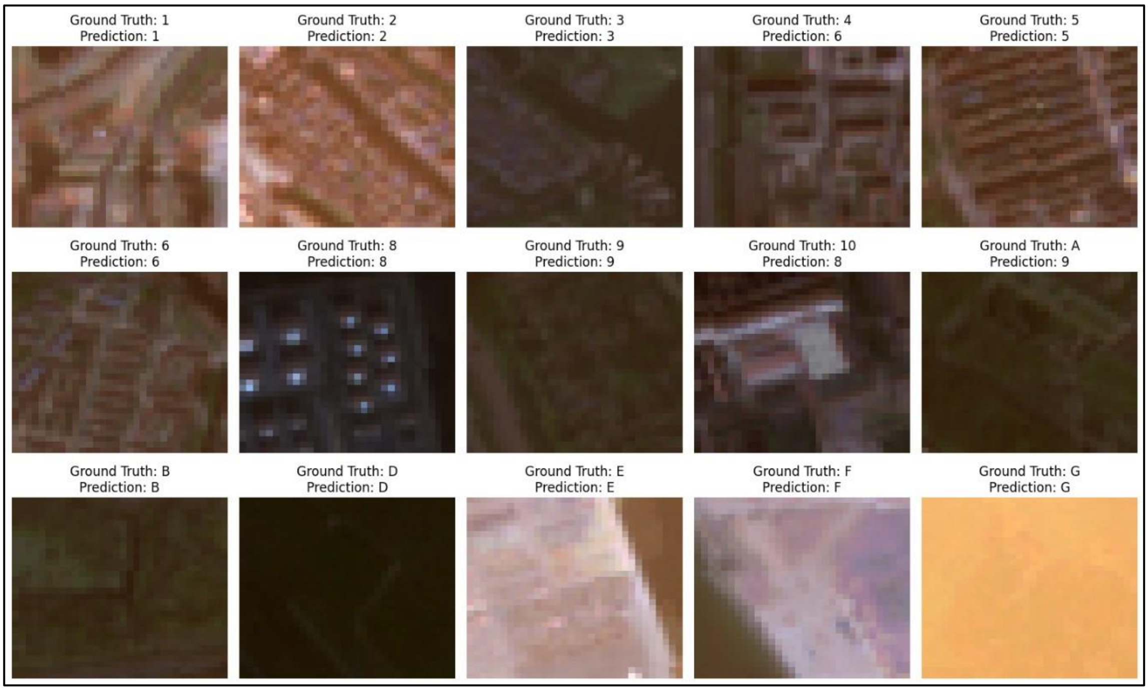


Figure 20: The visualization of predicted samples of LCZ classification by ResNet-50 model

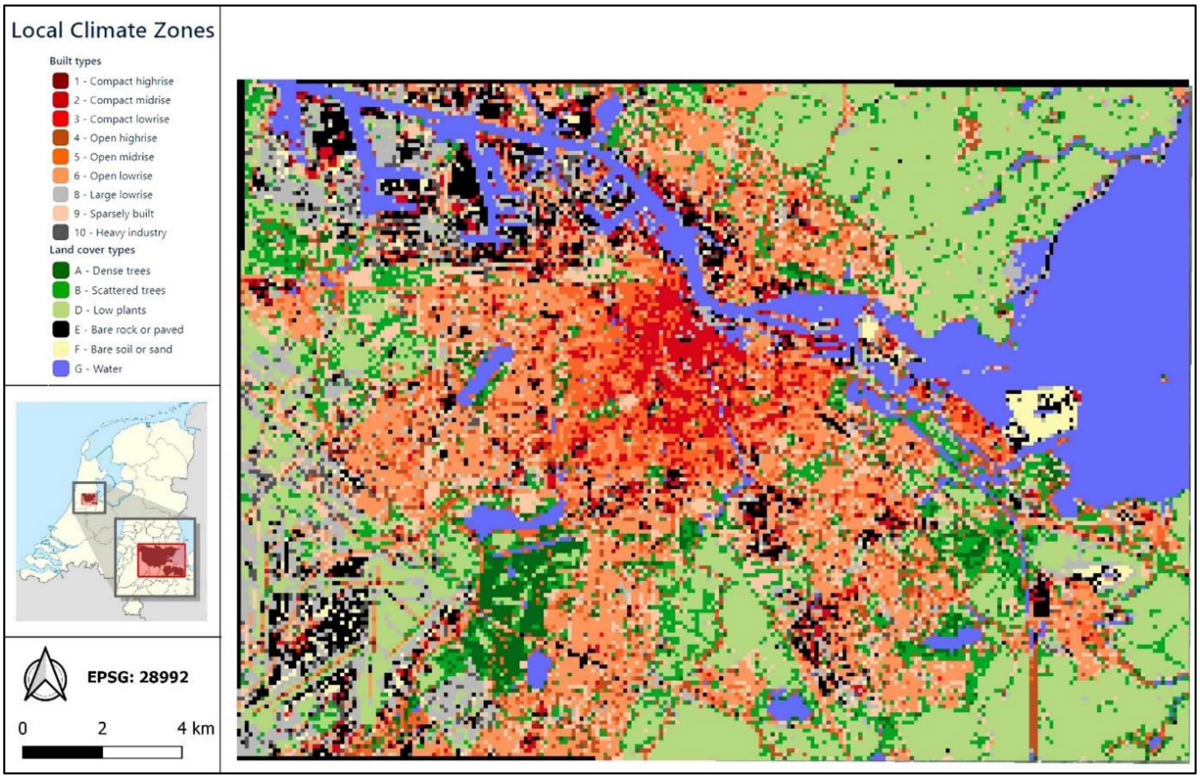


Figure 21: The LCZ classification map of the study area, predicted by ResNet-50 model

5.2. PM2.5 prediction

In this section, before going through the results of the models, the prepared PM2.5 sensors and the dataset for the modeling are explored. In total, the hourly values data of a two-year period (2021-2022) were collected for eight official sensors within the bounding box of Amsterdam. Regarding low-cost sensors, 17 sensors from 37 existing sensors were collected based on the reliability criteria explained in the methodology chapter. Therefore, 25 sensors of PM2.5 observation, including eight official sensors and 17 individual sensors, are involved in the modeling. Figure 22 shows the location of all PM2.5 sensors in the study area with the id of the sensors, which is used in the dataset.

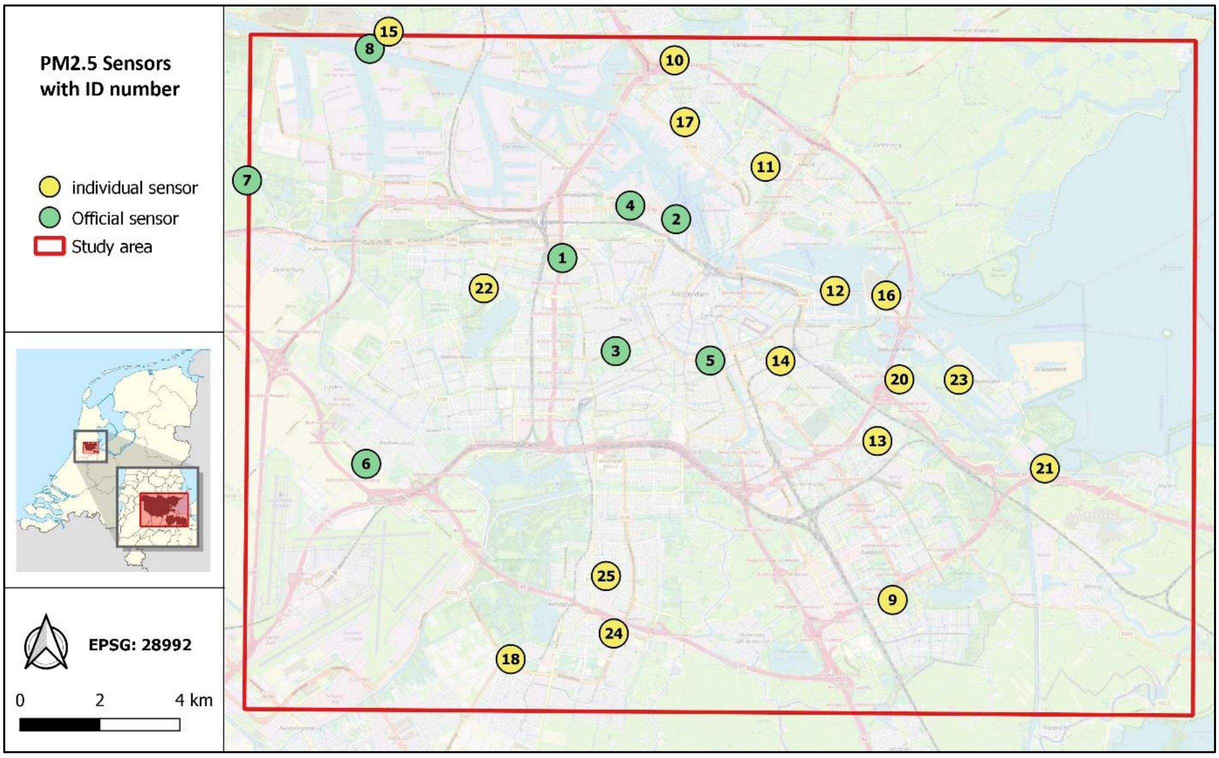


Figure 22: the location of PM 2.5 sensors in the study area

Each sensor was then assigned to its corresponding LCZ category, identifying nine distinct LCZ categories in the proximity of the sensors. These categories include "Compact mid-rise" (2), "Open high-rise" (4), "Open mid-rise" (5), "Open low-rise" (6), "sparsely-built" (9), "Industrial" (10), "Scattered trees" (B), “low plants” (D), and “Bare rock or paved” (E). However, the distribution of LCZ categories within sensors is not the same. However, it is important to note that the distribution of LCZ categories across the sensors was not uniform. Specifically, LCZ 6 was found to cover nine sensors, LCZ 4 covered five sensors, LCZ 5 covered three sensors, LCZ 9 covered three sensors, and the remaining LCZ classes were represented by a single sensor each.

The dataset was prepared based on the procedure described in the methodology. The dataset presents 28 independent variables, followed by PM2.5 as the target variable. The predictors include nine meteorological-related variables, 14 variables related to LCZ (LCZ category and probabilities), one for traffic, and four for periodicity for the time of day and the time of year. Table 8 shows the details of the variables in the dataset.

Table 8: The details of variables of the dataset for PM2.5 prediction

variable column	Description	unit	Spatial/temporal
PM25	PM2.5 observation	µg/m³	-
temperature	Temperature of official sensor	0.1 degrees Celsius	Temporal
precipitation	Precipitation of official sensor	(in 0.1 mm) (-1 for <0.05 mm)	Temporal
humidity	Humidity of official sensor	%	Temporal
cloudiness	Cloud cover of official sensor	octant	Temporal
pressure	Air pressure of official sensor	0.1 hPa	Temporal
Wx	The vector of wind speed-direction in X dimension	-	Temporal
Wy	The vector of wind speed-direction in Y dimension	-	Temporal
temperature1	Temperature of low-cost sensors	0.1 degrees Celsius	Spatio-temporal
humidity1	Humidity of low-cost sensors	%	Spatio-temporal
lcz	LCZ category	-	Spatial
1,2,3,4,5,6,8, 9,10,A,B,D,E	Including 13 columns for LCZ probabilities	-	Spatial
traffic	The value of kernel density estimation	-	Spatial
Day sin	Sine transformation for time of the day	-	Temporal
Day cos	Cosine transformation for time of the day	-	Temporal
Year sin	Sine transformation for time of the year	-	Temporal
Year cos	Cosine transformation for time of the year	-	temporal

Before modeling, to get an overview of variables, we do some data explanatory by visualizing the distribution of the variables over two years of the dataset (Figure 23). Certain variables, such as temperature and humidity, exhibit a distinct periodic pattern characterized by a symmetrical arrangement of data points between the two halves of the dataset representing 2021 and 2022. Conversely, variables like pressure do not display a discernible temporal pattern or exhibit any regularity in their fluctuations over time.

According to the strategy described in the methodology section for splitting the data into training, validation, and testing subset sets, the training set has 271752 hours of data, followed by 67938 hours in the validation set and 33578 hours in the testing set.

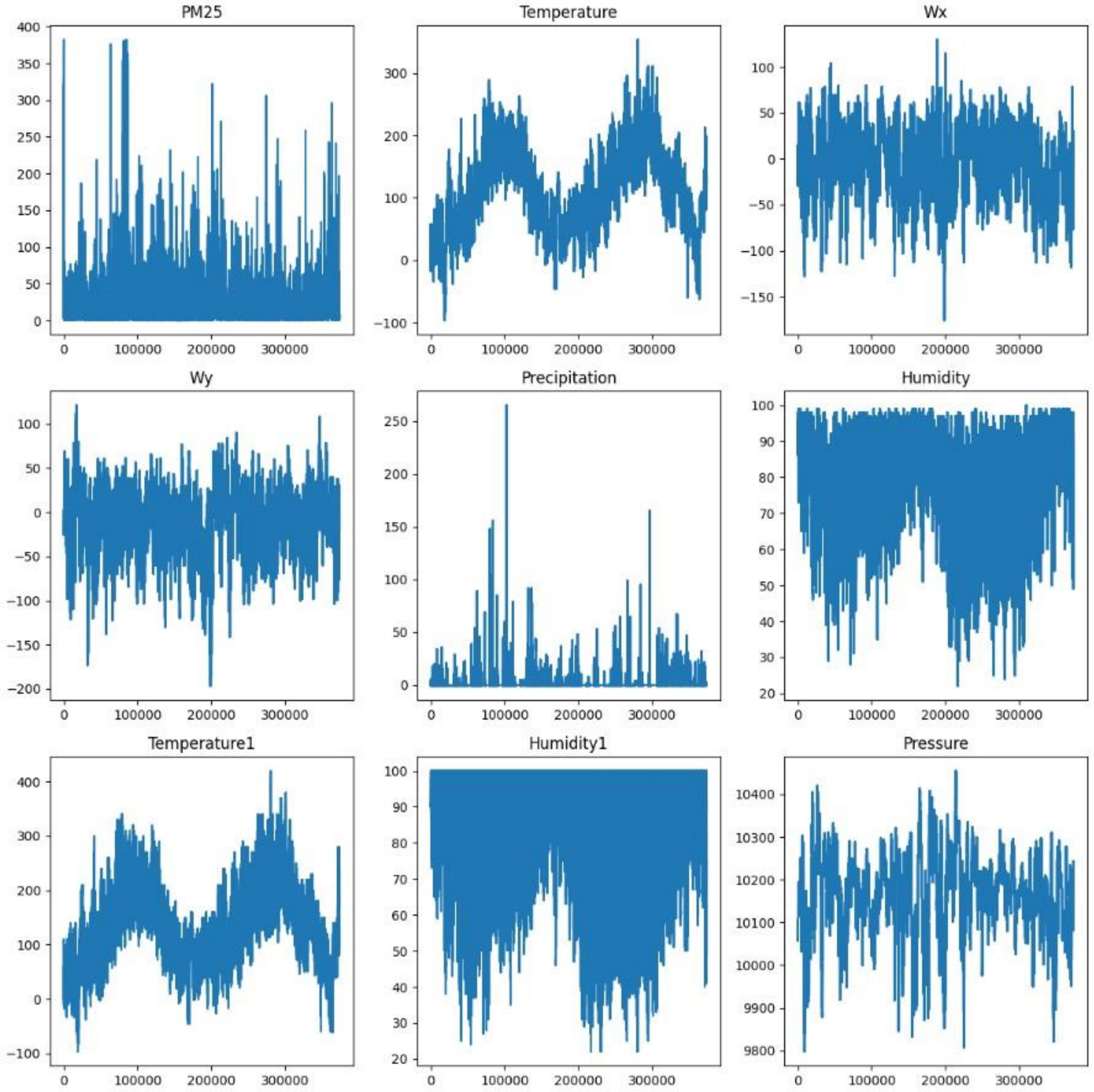


Figure 23: The distribution of variables related to PM2.5 prediction over the time period of the dataset

Regarding implementing the XGBoost model on the dataset, the initial hyperparameters for max_depth and learning rate were set, then the model was fine-tuned using a grid search technique with defining multiple values for these parameters. A ten-fold cross-validation strategy, considering negative mean square error scoring, was employed to achieve the optimal hyperparameters. The XGBoost model was then trained using the optimal parameter values, shown in Table 9.

Table 9: The hyperparameters of fine-tuned XGBoost model for PM2.5 prediction

Parameter	**Value**
Objective (Loss function)	Mean square error
n_estimators	1000
max_depth	3
learning_rate	0.1
early_stopping_rounds	250

During the training of the LSTM model, the dataset was first split into training and testing sets, followed by data normalization. The windowing technique was applied with specific parameters to create sequential data for training. The input_width was set to 24, indicating the number of input time steps. The label_width was set to 1, representing the number of output time steps. The shift value of 1 determined the time shift between consecutive windows. The hyperparameters presented in Table 10 were obtained through a process of fine-tuning the LSTM model. As the model's training process was computationally intensive, a batch size of 128 was chosen to reduce training time. However, such a batch size may require a more powerful computing environment for efficient training.

Table 10: The hyperparameters of fine-tuned LSTM model for PM2.5 prediction

Hyperparameter	**Value**
Loss Function	MeanSquaredError
Optimizer	Adam
Learning rate	0.001
Batch size	128
Epochs	200
Monitor	val_loss
Dropout rate	0.5

The results of training and evaluation of the dataset using the XGBoost and LSTM models (Table 11) shows that the LSTM model outperforms XGBoost for predicting PM2.5 values. It achieved a validation RMSE of 1.8057 and a validation R-squared of 0.9589, indicating a highly accurate fit to the training data.
While the LSTM model indicates impressive results during the training, its performance on the testing set resulted in a higher RMSE of 5.1270 and a lower R-squared of 0.7521. This suggests that the model might have to overfit the training data. However, the LSTM model still outperformed the XGBoost model regarding prediction accuracy, showcasing its ability to capture the temporal patterns in PM2.5 pollutant concentrations.

Table 11: The results of the evaluation metrics of XGBoost and LSTM models for PM2.5 prediction

Model	**trainable parameters**	**Validation set**			**Test set**		
		MAE	**RMSE**	**R-squared**	**MAE**	**RMSE**	**R-squared**
XGBoost	-	3.80	5.87	0.61	4.32	5.93	0.57
LSTM	559233	1.1383	1.8057	0.9589	3.19	5.127	0.7521

Figure 24 shows the performance of two models on the comparison between the actual and the predicted values. According to the plots of actual data and predicted, we see that the predicted values are closer to their corresponding actual values LSTM model, as they are closer to the reference line (the red line), representing the scenario where the predicted values perfectly match the actual values.

The superior performance of the LSTM model in the dataset can be rooted in its capability to handle sequential data. By considering the sequential nature of the predictors and the target variable, the LSTM model can effectively capture time-dependent patterns and correlations. This is especially beneficial for predicting PM2.5 levels, as they are influenced by factors that change over time, such as weather, traffic, and seasons. In comparison to the LSTM model, the XGBoost model also demonstrated acceptable performance in predicting PM2.5 pollutant levels, achieving a test RMSE of 5.87 and a validation R-squared of 0.57. One of the strengths of the XGBoost model lies in its ability to handle a wide range of

predictors and capture complex interactions among them. However, compared to the LSTM model, the XGBoost model is not able to learn the temporal dependencies present in the nature of the dataset.
The LSTM model, which proved a better performance, is used for analyzing the possibility of the LCZ impact on PM2.5 concentrations as a more reliable outcome. However, the analysis of feature importance for exploring to what extent different LCZ classes affect the PM2.5 levels are conducted using the XGBoost model. While the performance of the XGBoost model may be comparatively less reliable, it offers a more straightforward interpretation of variable importance on PM2.5 prediction.

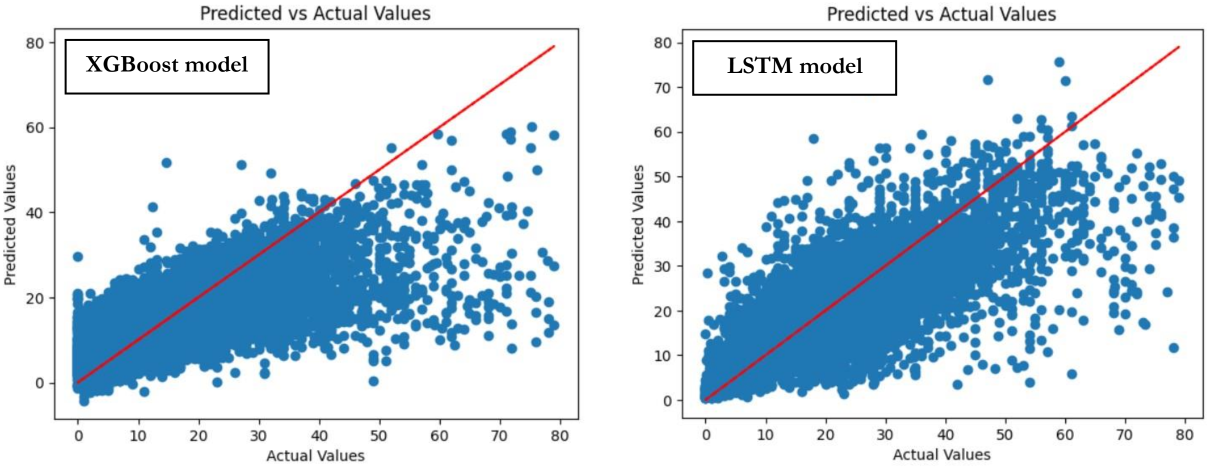


Figure 24: The plot of predicted vs actual PM2.5 values of XGBoost and LSTM models on testing set

5.3. Effect of LCZ on PM2.5

5.3.1. Sensitivity analysis

The LSTM model was employed for sensitivity analysis on PM2.5 prediction to explore if there is a significant impact of LCZ on PM2.5 prediction. Six scenarios involving predictors in training were examined to compare the performance results on PM2.5 prediction. First, the model was trained by all variables, and the evaluation metrics were considered as the benchmark for comparison. The next experiment considered all variables except the LCZ class variable as the input variables for training. In the third condition, all variables related to LCZ, including LCZ and its probabilities, were excluded from predictors. Forth scenario trained the model by putting aside all spatial variables, consisting of LCZ classes, LCZ probabilities, and Traffic. In addition, two other approaches with the same involved predictors in the first and second scenarios were trained, considering the observations related to all PM2.5 concentrations, including the values that were removed previously as the outlier.

Interpreting the results of training metrics, shown in Table 12, we see a slight drop in the performance of the model by removing the variables related to LCZ from R-squared 0.7521 on the test set to 0.7399. This proves the existence of LCZ impact on PM2.5 predictions. The model performance difference was even larger when the model was trained without any spatial variables (LCZ and Traffic). The greater decrease in the model's performance, the higher the importance of the removed variables from predictors on the PM2.5 prediction. This shows both LCZ and traffic are sensitive variables for forecasting the PM2.5 pollutant levels.

Table 12: The results of sensitivity analysis for PM2.5 prediction

Scenario	Predictors	Validation set			Test set	
		MSE	RMSE	R-squared	RMSE	R-squared
1	All variables	2,4199	1,8057	0,9589	5,1270	0,7521
2	Remove LCZ probabilities	2,4487	1,5670	0,9700	5,3099	0,7341
3	Remove LCZ class and probabilities	2,4839	1,5845	0,9652	5,25187	0,7399
4	Remove LCZ class, LCZ probabilities, and traffic	3,2002	1,9487	0,9524	5,2355	0,7415
5	All variables (including outliers of PM2.5 for records)	6,8408	2,8207	0,9116	6,77	0,6857
6	Remove LCZ class and probabilities (including outliers of PM2.5 for records)	6,075	2,6531	0,9275	6,873	0,6761

Regarding the implementation of the model on all records, including the outlier samples, we can see that generally, the performance of the model has become worsen. However, In the case of excluding LCZ-related variables, the decrease in model performance shows a proportional similarity to the drop observed when training the model in the first and second scenarios, using the dataset without outliers. This indicates that Local climate zones (LCZ) types are not that much sensitive to the prediction of higher values of PM2.5 within sensors, and the other variables should have caused it.

The comparison between the second and third scenarios demonstrates that although the model achieves better results during training(R-squared of 0.97), its performance on unseen data deteriorates. This highlights the significance of incorporating the probabilities of LCZ to enhance the model's ability for generalization.

A sample of daily and weekly predictions of PM2.5 are visualized in Figure 25 to get a more precise understanding of the performance of the models in different scenarios (Scenarios 1, 3, and 4).

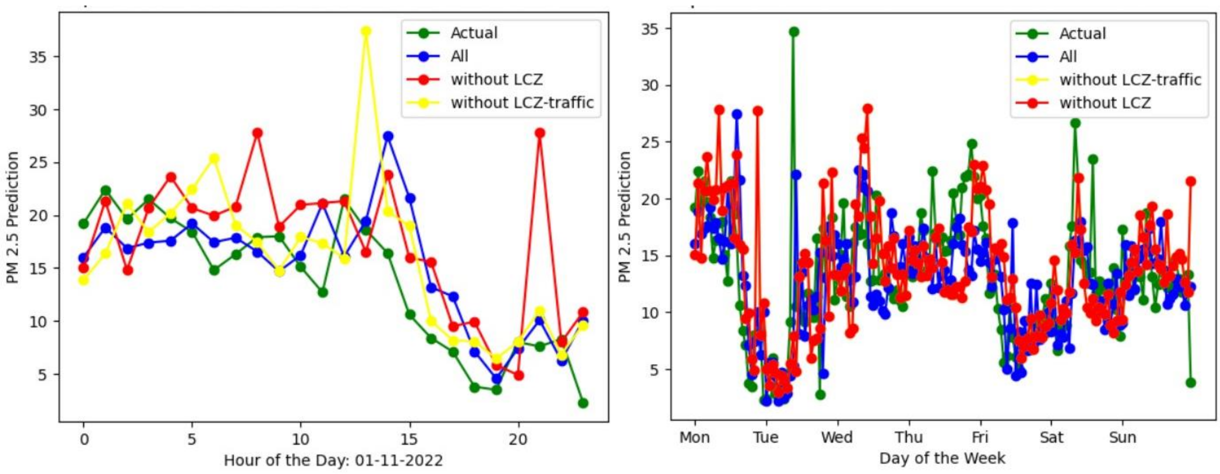


Figure 25: Comparison of predicted and actual PM2.5 values for 24 hours and one week in different scenarios

5.3.2. Feature importance analysis

In the feature importance analysis, we discover to what extent different LCZ categories affect PM2.5 concentration. The results of the feature importance function of the XGBoost model on predicting PM2.5 value are shown in Figure 26. Regarding LCZ classes' contribution to PM2.5, the LCZ class labeled “Open highrise” (4) has the most effect on prediction among all predictors, scoring approximately 0.15. It means that this category contributes 15% to the overall importance of the model in predicting PM2.5 concentration. The traffic variable demonstrates the second highest level of influence on PM2.5, with a score of 0.13 . Among the remaining LCZ categories, four LCZ, namely "large low-rise" (8), "Compact low-rise" (3), "Open low-rise" (6), and "Compact high-rise" (1), show the next levels of importance among variables, ranging from scores of 0.10 to 0.04 respectively. Other LCZ classes, including "Industrial" (10), "Compact mid-rise" (2), and "sparsely-built" (9), indicate minor importance for predicting fine particulate matter levels.

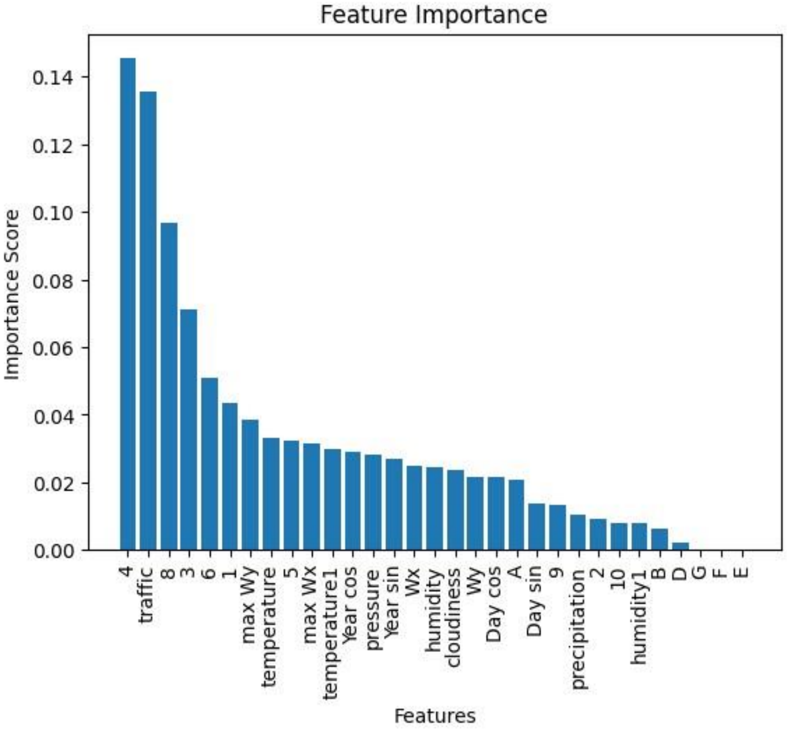


Figure 26: The feature importance of independent variables in predicting PM2.5 in the XGBoost model

Although the results of the feature importance and the reliability of feature importance results derived from the XGBoost model may not be completely reliable due to the performance level of the XGBoost model, it can still provide an overview of the contribution of LCZ on PM2.5 concentration. Another factor that could have impacted the reliability of these scores is the insufficient availability of actual LCZ classes on PM2.5 sensors, resulting in the use of class probabilities instead. The example showing the effect of this limitation can be seen in the industrial category, which is typically expected to have a significant influence on PM2.5 levels but does not exhibit high importance according to the results of this model.

Finally, the map of the importance of LCZ categories on PM2.5 (Figure 27) was created to highlight the potential areas affected by different types of LCZ. According to the map, the areas with stronger red have more influence on the concentration of PM2.5. This can be proven by the location of some industrial areas in the north-western part of Amsterdam, followed by the area covered by Schiphol airport in the southwest of the city. Some areas with moderate impact are detected in the western and southeast parts of the city, belonging to light industry and commercial areas. The central area of the city has a different range of influence on PM2.5.

This map can be a guideline for urban management to mitigate the concentrations of PM2.5, which are caused by urban forms.

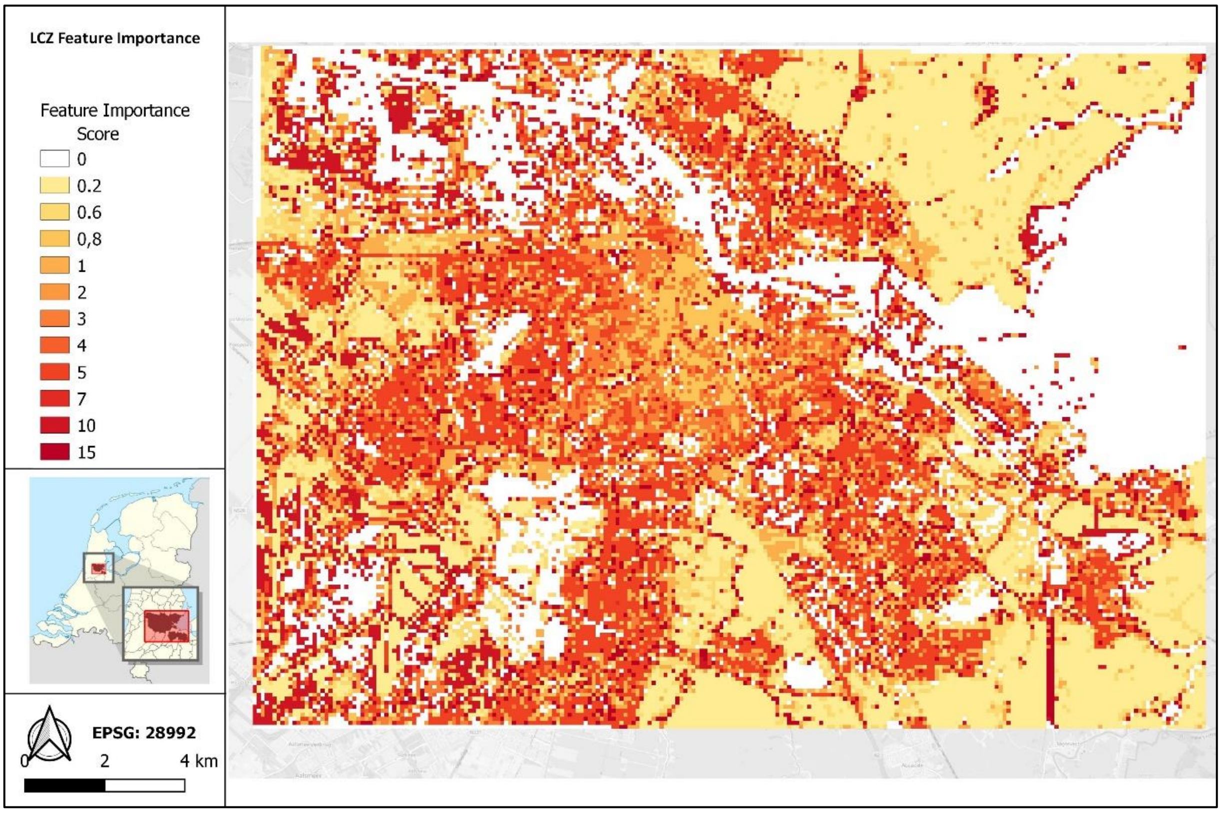


Figure 27: The map of the level of LCZ importance on PM2.5 concentration

6. CONCLUSION AND FUTURE WORK

This research developed a two-stage framework to investigate the impact of urban form on air pollution, focusing on modeling, evaluating, and analyzing the problem to accomplish the goal. Local climate zones(LCZ) were employed as a comprehensive scheme for introducing physical and functional characteristics of an urban environment to represent different types of urban forms. Concerning air pollution, PM2.5 was selected as the pollutant index, representing the most influential factors caused by urban environments and the data's availability. In previous studies, the analysis of urban form's impact on air pollution primarily focused on considering individual measurements of urban form as influencing factors. However, this research takes a different approach by designing a comprehensive framework that incorporates urban form categories and probabilities derived from Local Climate Zones (LCZ). This two-directional approach allows for a more thorough investigation of urban form and its relationship to air pollution. Additionally, a noteworthy aspect of this research involves citizen science data for air pollution, which sets it apart from previous works in the field.

Two objectives were defined regarding completing each stage of the research framework, followed by introducing the related research questions. The answers to the questions are discussed as follows:

1.1. Which classification framework is effective for training deep learning models and utilizing Earth observation data to classify urban form?

The initial intention of this research is to find a classification scheme for urban forms to be used in deep learning models. The selected system is LCZ. Compared to other frameworks, LCZ integrates multiple factors such as physical properties, land cover, and land use, providing a comprehensive representation of urban form. This makes this framework more compatible with those deep learning models such as CNN, which learn patterns on scene-based understanding for classification. In addition, LCZ provides a standardized and widely accepted framework for classifying urban areas that researchers have adopted globally. LCZ classification, which can be derived from remote sensing data, is particularly advantageous in addressing the limitations of using direct measurements to characterize urban form.

1.2. Which Convolutional Neural Networks architecture from deep learning can provide an acceptable accuracy to predict the urban form classes using EO data?

This research implemented three different CNN models for the LCZ classification task. The author developed the first model, and the two other model architectures from well-known state-of-the-art models, namely ResNet-50 and EfficienNet, were adjusted to the problem of the research. Although all models resulted in an acceptable accuracy, the ResNet-50 model indicated the highest performance with an overall accuracy of 87 percent on unseen data. Regarding the evaluation metrics, this model obtained 92 percent for Recall and 89 percent for both Precision and F1-score. However, in all models, some categories could not be predicted with high accuracy. This issue can be seen in class 3 and class 10, representing compact low-rise and industrial areas, respectively. The lack of enough sample data and the ability to distinguish the pattern of these classes from the data of single satellite imagery might be the main reasons for such a performance.

1.3. What criteria should be considered for tuning hyperparameters in LCZ classification using the CNN model?

Several hyperparameters configuration should be considered in this model to make the optimal performance. In this research, the criteria focused on the efficiency of the model on generalization also make the model the model to learn the patterns of classes comprehensively. The hyperparameters related to compiling the model, including the number of epochs (set to 500), batch size (32), and early stopping (300) were set to the values in which the model is able to capture the image patterns considering the efficient computational tasks and the existing capacity. Data augmentation technique is also applied to increase the number of samples in the training set. Regarding the parameters in designing the model architecture, Preventing overfitting and underfitting the data was considered the most crucial factor in defining convolutional layers, dropout layers, and regularization techniques.

2.1. What modeling techniques are most suitable for examining the impact of urban form classes on the concentration of PM2.5 pollutant?

In this research, we predicted the concentration of PM2.5 using two models, considering several predictors such as LCZ classes and probabilities, meteorological factors, traffic, and time stamp data. The first approach focused on implementing the XGBoost model from ensemble learning techniques, and the second one employed an LSTM model from neural networks. The results of the evaluation metrics indicate that the model that considers the sequential trend of the dataset and involves spatiotemporal characteristics in prediction shows more precise performance, which can be present in the LSTM model. This is proved by the R-squared, RMSE, and MAE of 0.96, 1.80, and 1.14 on the validation set and 0.75, 5.12, and 3.19 on unseen data.

2.2. To what extent does urban form contribute to the concentration of PM2.5?

Using the designed LSTM model for pM2.5 prediction, we applied sensitivity analysis to discover the possibility of the impact of the LCZ variable as the representative of urban form on PM2.5 by training the model on different alternatives involving predictors and records. This analysis shows that removing LCZ categories and probabilities from independent variables resulted in a drop in the model performance of around 2.30 percent in R-squared. This proves the positive contribution of urban form on PM2.5 prediction. In addition, the similar performance of the model on all records and the records without outliers show that the prediction of much higher values of PM2.5 out of the range of the observations are rooted in other predictors, or the errors in the sensors measurements might cause it.

2.3. Which types of urban forms have the strongest impact on the distribution of PM2.5?

The task of feature importance analysis on the XGBoost model determined the effect of LCZ classes as the representative of urban form on PM2.5 concentrations. The results show the significance of the open-highrise category at the highest level, with a score of around 14 percent, the highest among all predictors in the model. This LCZ type consists of height buildings with more than eight floors covered by scattered trees in the open spaces, and it typically has a residential function. The large low-rise, the compact low-rise, and the open low-rise types received lower levels of importance among LCZ classes. The light industry, transportation hubs and commercial, and the city center area highlight the urban functions within these categories.

6.1. Limitations

This research selected one urban area, Amsterdam, as the study area. Although several factors were considered in choosing a location that covers most LCZ classes and the availability of air pollution sensors, we faced some limitations during the research procedure. The most critical restrictions were related to providing the training area for LCZ classification, also covering all built types by PM2.5 sensors. The former was addressed by using augmented techniques in the training set, and the latter was handled by involving probabilities of LCZ in addition to the category in the location of each sensor. However, using more number of cities as the study area might give more promising data to manage imbalance in the input categories for LCZ classification. Moreover, more air pollution sensors can cover all LCZ categories and provide a more reliable dataset for PM2.5 observations.

6.2. Future work and recommendations

This research adopted a two-stage approach: the first stage for classifying urban forms using deep learning and the second stage for predicting PM2.5. However, an alternative can be considered where all procedures are conducted within a single-stage model. It means that the input data in this one-stage approach would be both data related to urban form classification, such as the EO data and sample data (in this case, LCZ training data), followed by other related data contributing to predicting air pollution. The output data would also be both classification results of urban form and PM2.5 prediction. By integrating these diverse datasets into an integrated framework, the model can benefit from the combined information to make simultaneous predictions for both urban form classification and PM2.5 levels. This might improve accuracy and predictive performance. However, It may bring more complexities regarding data integration, model design, computational requirements, and, more importantly, the interpretation of the results. Another recommendation for improving the accuracy of the classification task of urban form is to fuse additional datasets that give information about the elevation and visual characteristics, such as street view imagery, for training a deep learning model.

All reproducibility materials for this research can be found on:

https://github.com/Morteza-Amouei/MSc-Thesis

LIST OF REFERENCES

Ahn, H., Lee, J., Hong, A., 2022. Urban form and air pollution: Clustering patterns of urban form factors related to particulate matter in Seoul, Korea. https://doi.org/10.1016/j.scs.2022.103859

Aslam, A., Rana, I.A., 2022. The use of local climate zones in the urban environment: A systematic review of data sources, methods, and themes. Urban Clim. 42, 101120. https://doi.org/10.1016/J.UCLIM.2022.101120

Badura, M., Batog, P., Drzeniecka-Osiadacz, A., Modzel, P., 2018. Optical particulate matter sensors in PM2.5 measurements in atmospheric air. E3S Web Conf. 44, 00006. https://doi.org/10.1051/E3SCONF/20184400006

Barke, M., 2018. The importance of urban form as an object of study. Urban B. Ser. 11–30. https://doi.org/10.1007/978-3-319-76126-8_2

Bechtel, B., Alexander, P.J., Böhner, J., Ching, J., Conrad, O., Feddema, J., Mills, G., See, L., Stewart, I., 2015. Mapping local climate zones for a worldwide database of the form and function of cities. ISPRS Int. J. Geo-Information 4, 199–219. https://doi.org/10.3390/IJGI4010199

Bechtel, B., Demuzere, M., Sismanidis, P., Fenner, D., Brousse, O., Beck, C., Van Coillie, F., Conrad, O., Keramitsoglou, I., Middel, A., Mills, G., Niyogi, D., Otto, M., See, L., Verdonck, M.-L., Wentz, E., Conrow, L., Fischer, H., 2017, U.S., 2017. Quality of Crowdsourced Data on Urban Morphology—The Human Influence Experiment (HUMINEX). Urban Sci. 2017, Vol. 1, Page 15 1, 15. https://doi.org/10.3390/URBANSCI1020015

Breiman, L., 2001. Random Forests. Mach. Learn. 2001 451 45, 5–32. https://doi.org/10.1023/A:1010933404324

Cai, C., Guo, Z., Zhang, B., Wang, X., Li, B., Tang, P., 2021. Urban Morphological Feature Extraction and Multi-Dimensional Similarity Analysis Based on Deep Learning Approaches. Sustain. 2021, Vol. 13, Page 6859 13, 6859. https://doi.org/10.3390/SU13126859

Cai, J., Chen, Y., 2022a. A novel unsupervised deep learning method for the generalization of urban form. http://www.tandfonline.com/action/journalInformation?show=aimsScope&journalCode=tgsi20#.VsXpLiCLRhE 1–20. https://doi.org/10.1080/10095020.2022.2068384

Cai, J., Chen, Y., 2022b. A novel unsupervised deep learning method for the generalization of urban form. Geo-spatial Inf. Sci. 1–20. https://doi.org/10.1080/10095020.2022.2068384

Carta, S., n.d. Machine Learning and the City: Applications in Architecture and Urban Design - Google Books [WWW Document]. URL https://books.google.nl/books?id=phh1EAAAQBAJ&pg=PT446&lpg=PT446&dq=Urban+morphology+meets+deep+learning:+Exploring+urban+forms+in+one+million+cities,+town+and+villages+across+the+planet&source=bl&ots=mj09VUqg6a&sig=ACfU3U0YiJ9pSLOvieveYnW4aOvgfuzSNA&hl=e (accessed 6.28.22).

Chen, F., 2014. Urban Morphology and Citizens' Life. Encycl. Qual. Life Well-Being Res. 6850–6855. https://doi.org/10.1007/978-94-007-0753-5_4080

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. 13-17-Augu, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, W., Wu, A.N., Biljecki, F., 2021. Classification of urban morphology with deep learning: Application on urban vitality. Comput. Environ. Urban Syst. 90. https://doi.org/10.1016/j.compenvurbsys.2021.101706

Dai, H., Huang, G., Zeng, H., Yang, F., 2021. PM2.5 Concentration Prediction Based on Spatiotemporal Feature Selection Using XGBoost-MSCNN-GA-LSTM. Sustain. 2021, Vol. 13, Page 12071 13, 12071. https://doi.org/10.3390/SU132112071

Demuzere, M., Bechtel, B., Middel, A., Mills, G., 2019. Mapping Europe into local climate zones. PLoS One 14. https://doi.org/10.1371/JOURNAL.PONE.0214474

Demuzere, M., Hankey, S., Mills, G., Zhang, W., Lu, T., Bechtel, B., 2020. Combining expert and crowd-sourced training data to map urban form and functions for the continental US. Sci. Data 2020 71 7, 1–13. https://doi.org/10.1038/s41597-020-00605-z

Demuzere, M., Kittner, J., Martilli, A., Mills, G., Moede, C., Stewart, I.D., Van Vliet, J., Bechtel, B., 2022. A global map of local climate zones to support earth system modelling and urban-scale environmental science. Earth Syst. Sci. Data 14, 3835–3873. https://doi.org/10.5194/ESSD-14-3835-2022

Gao, Y., Wang, Z., Liu, C., Peng, Z.R., 2019. Assessing neighborhood air pollution exposure and its relationship with the urban form. Build. Environ. 155, 15–24. https://doi.org/10.1016/J.BUILDENV.2018.12.044

Gao, Y., Zhao, J., Han, L., 2021. Exploring the spatial heterogeneity of urban heat island effect and its relationship to block morphology with the geographically weighted regression model. https://doi.org/10.1016/j.scs.2021.103431

Ghaemi, Z., Alimohammadi, A., Farnaghi, M., 2018. LaSVM-based big data learning system for dynamic prediction of air pollution in Tehran. Environ. Monit. Assess. 190, 1–17. https://doi.org/10.1007/S10661-018-6659-6/FIGURES/13

Graves, A., 2012. Supervised Sequence Labelling with Recurrent Neural Networks. Studies in Computational Intelligence 385. https://doi.org/10.1007/978-3-642-24797-2

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep Residual Learning for Image Recognition. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2016-December, 770–778. https://doi.org/10.1109/CVPR.2016.90

Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory. Neural Comput. 9, 1735–1780. https://doi.org/10.1162/NECO.1997.9.8.1735

Huang, C., Hu, T., Duan, Y., Li, Q., Chen, N., Wang, Q., Zhou, M., Rao, P., 2022. Effect of urban morphology on air pollution distribution in high-density urban blocks based on mobile monitoring and machine learning. Build. Environ. 219, 109173. https://doi.org/10.1016/j.buildenv.2022.109173

Huang, C.J., Kuo, P.H., 2018. A Deep CNN-LSTM Model for Particulate Matter (PM2.5) Forecasting in Smart Cities. Sensors 2018, Vol. 18, Page 2220 18, 2220. https://doi.org/10.3390/S18072220

Huang, X., Liu, A., Li, J., 2021. Mapping and analyzing the local climate zones in China's 32 major cities using Landsat imagery based on a novel convolutional neural network. Geo-Spatial Inf. Sci. 24, 528–557. https://doi.org/10.1080/10095020.2021.1892459

Kim, M., Jeong, D., Kim, Y., 2021. Local climate zone classification using a multi-scale, multi-level attention network. ISPRS J. Photogramm. Remote Sens. 181, 345–366. https://doi.org/10.1016/J.ISPRSJPRS.2021.09.015

Kim, M.-H., Gim, T.-H.T., 2022. Deep learning-based investigation of the impact of urban form on the particulate matter concentration on a neighborhood scale 0, 1–16. https://doi.org/10.1177/23998083221111162

Kullenberg, C., Kasperowski, D., 2016. What Is Citizen Science? – A Scientometric Meta-Analysis. PLoS One 11, e0147152. https://doi.org/10.1371/JOURNAL.PONE.0147152

Li, C., Wang, Z., Li, B., Peng, Z.R., Fu, Q., 2019. Investigating the relationship between air pollution variation and urban form. Build. Environ. 147, 559–568. https://doi.org/10.1016/J.BUILDENV.2018.06.038

Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., Chi, T., 2017. Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. Environ. Pollut. 231, 997–1004. https://doi.org/10.1016/J.ENVPOL.2017.08.114

Li, Y., Zhang, H., Xue, X., Jiang, Y., Shen, Q., 2018. Deep learning for remote sensing image classification: A survey. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 8, 1–17. https://doi.org/10.1002/widm.1264

Li, Z., Ming, T., Liu, S., Peng, C., de Richter, R., Li, W., Zhang, H., Wen, C.Y., 2021. Review on pollutant dispersion in urban areas-part A: Effects of mechanical factors and urban morphology. Build. Environ. 190, 107534. https://doi.org/10.1016/J.BUILDENV.2020.107534

Lin, L., Liang, Y., Liu, L., Zhang, Y., Xie, D., Yin, F., Ashraf, T., 2022. Estimating PM2.5 Concentrations Using the Machine Learning RF-XGBoost Model in Guanzhong Urban Agglomeration, China. Remote Sens. 14, 5239. https://doi.org/10.3390/rs14205239

Liu, S., Shi, Q., 2020. ISPRS Journal of Photogrammetry and Remote Sensing Local climate zone mapping as remote sensing scene classification using deep learning: A case study of metropolitan China. https://doi.org/10.1016/j.isprsjprs.2020.04.008

Ma, J., Cheng, J.C.P., Xu, Z., Chen, K., Lin, C., Jiang, F., 2020. Identification of the most influential areas for air pollution control using XGBoost and Grid Importance Rank. J. Clean. Prod. 274, 122835. https://doi.org/10.1016/J.JCLEPRO.2020.122835

Martino, N., Girling, C., Lu, Y., 2021. Urban form and livability: socioeconomic and built environment indicators. Build. Cities 2, 220–243. https://doi.org/10.5334/BC.82

Nababan, A.A., Sutarman, Zarlis, M., Nababan, E.B., 2022. Air Quality Prediction Based on Air Pollution Emissions in the City Environment Using XGBoost with SMOTE. ICOSNIKOM 2022 - 2022

IEEE Int. Conf. Comput. Sci. Inf. Technol. Bound. Free Prep. Indones. Metaverse Soc. https://doi.org/10.1109/ICOSNIKOM56551.2022.10034887

Rosentreter, J., Hagensieker, R., Waske, B., 2020. Towards large-scale mapping of local climate zones using multitemporal Sentinel 2 data and convolutional neural networks. Remote Sens. Environ. 237, 111472. https://doi.org/10.1016/J.RSE.2019.111472

Saxena, P., Srivastava, A. (Eds.), 2020. Air Pollution and Environmental Health. Environmental Chemistry for a Sustainable World 20. https://doi.org/10.1007/978-981-15-3481-2

Shi, Y., Ren, C., Lau, K.K.L., Ng, E., 2019. Investigating the influence of urban land use and landscape pattern on PM2.5 spatial variation using mobile monitoring and WUDAPT. Landsc. Urban Plan. 189, 15–26. https://doi.org/10.1016/J.LANDURBPLAN.2019.04.004

Shi, Y., Xie, X., Chi-Hung Fung, J., Ng, E., 2017. Identifying critical building morphological design factors of street-level air pollution dispersion in high-density built environment using mobile monitoring. https://doi.org/10.1016/j.buildenv.2017.11.043

Stewart, I.D., 2011. Redefining the urban heat island. https://doi.org/10.14288/1.0072360

Stewart, I.D., Oke, T.R., 2012. Local Climate Zones for Urban Temperature Studies. Bull. Am. Meteorol. Soc. 93, 1879–1900. https://doi.org/10.1175/BAMS-D-11-00019.1

Tan, M., Le, Q. V., 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. 36th Int. Conf. Mach. Learn. ICML 2019 2019-June, 10691–10700.

Wu, J., Lu, Y., Gao, H., Wang, M., 2022. Cultivating historical heritage area vitality using urban morphology approach based on big data and machine learning. Comput. Environ. Urban Syst. 91, 101716. https://doi.org/10.1016/J.COMPENVURBSYS.2021.101716

Xu, G., Zhu, X., Tapper, N., Bechtel, B., 2019. Urban climate zone classification using convolutional neural network and ground-level images. https://doi-org.ezproxy2.utwente.nl/10.1177/0309133319837711 43, 410–424. https://doi.org/10.1177/0309133319837711

Yang, H., Leng, Q., Xiao, Y., Chen, W., 2022. Investigating the impact of urban landscape composition and configuration on PM2.5 concentration under the LCZ scheme: A case study in Nanchang, China. Sustain. Cities Soc. 84, 104006. https://doi.org/10.1016/J.SCS.2022.104006

Yang, J., Shi, B., Shi, Y., Marvin, S., Zheng, Y., Xia, G., 2020. Air pollution dispersal in high density urban areas: Research on the triadic relation of wind, air pollution, and urban form. Sustain. Cities Soc. 54, 101941. https://doi.org/10.1016/J.SCS.2019.101941

Yao, Q., Li, H., Gao, P., Guo, H., Zhong, C., 2022. Mapping Irregular Local Climate Zones from Sentinel-2 Images Using Deep Learning with Sequential Virtual Scenes. Remote Sens. 14, 5564. https://doi.org/10.3390/RS14215564

Yoo, C., Han, D., Im, J., Bechtel, B., 2019. Comparison between convolutional neural networks and random forest for local climate zone classification in mega urban areas using Landsat images. ISPRS J. Photogramm. Remote Sens. 157, 155–170. https://doi.org/10.1016/J.ISPRSJPRS.2019.09.009

Yuan, C., Ng, E., Norford, L.K., 2014. Improving air quality in high-density cities by understanding the relationship between air pollutant dispersion and urban morphologies. Build. Environ. 71, 245–258. https://doi.org/10.1016/J.BUILDENV.2013.10.008

Zhang, B., Zhang, H., Zhao, G., Lian, J., 2020. Constructing a PM2.5 concentration prediction model by combining auto-encoder with Bi-LSTM neural networks. Environ. Model. Softw. 124, 104600. https://doi.org/10.1016/J.ENVSOFT.2019.104600

Zheng, H., Yuan, J., Chen, L., 2017. Short-Term Load Forecasting Using EMD-LSTM Neural Networks with a Xgboost Algorithm for Feature Importance Evaluation. Energies 2017, Vol. 10, Page 1168 10, 1168. https://doi.org/10.3390/EN10081168

Zhu, X.X., Hu, J., Qiu, C., Shi, Y., Kang, J., Mou, L., Bagheri, H., Haberle, M., Hua, Y., Huang, R., Hughes, L., Li, H., Sun, Y., Zhang, G., Han, S., Schmitt, M., Wang, Y., 2020. So2Sat LCZ42: A Benchmark Data Set for the Classification of Global Local Climate Zones [Software and Data Sets]. IEEE Geosci. Remote Sens. Mag. 8, 76–89. https://doi.org/10.1109/MGRS.2020.2964708

Zhu, X.X., Qiu, C., Hu, J., Shi, Y., Wang, Y., Schmitt, M., Taubenböck, H., 2022. The urban morphology on our planet – Global perspectives from space. Remote Sens. Environ. 269, 112794. https://doi.org/10.1016/J.RSE.2021.112794