Estimate Particulate Matter PM2.5 Concentration Impact of Wildfires Using Machine Learning in Chiang Mai Province, Thailand

THIWAKORN SENA

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR MASTER DEGREE OF SCIENCE
IN GEOINFORMATICS
FACULTY OF GEOINFORMATICS
BURAPHA UNIVERSITY
2025

การประเมินความเข้มข้นของ PM2.5 จากผลกระทบไฟป่าที่รุนแรงด้วย Machine Learning ในจังหวัดเชียงใหม่ ประเทศไทย

ทิวากรณ์ เสนา

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต
สาขาวิชาภูมิสารสนเทศศาสตร์
คณะภูมิสารสนเทศศาสตร์ มหาวิทยาลัยบูรพา
2568

Estimate Particulate Matter PM2.5 Concentration Impact of Wildfires Using Machine Learning in Chiang Mai Province, Thailand

THIWAKORN SENA

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR MASTER DEGREE OF SCIENCE
IN GEOINFORMATICS
FACULTY OF GEOINFORMATICS
BURAPHA UNIVERSITY
2025

The Thesis of Thiwakorn Sena has been approved by the examining committee to be partial fulfillment of the requirements for the Master Degree of Science in Geoinformatics of Burapha University

Advisory Committee

Principal advisor

.....

(Professor Dr. Zhenfeng Shao)

Co-advisor

.....

(Assistant Professor Dr. Phattraporn Soytong)

Examining Committee

.....

Principal examiner

(Professor Dr. Wolfgang Kainz)

.....

Member

(Professor Dr. Zhenfeng Shao)

.....

Member

(Professor Dr. Timo Balz)

.....

Member

(Professor Dr. Zhangcai Yin)

.....

Dean of the Faculty of Humanities and Social Sciences

(Associate Professor Dr. Suchada Pongkittiwiboon)

.....

This Thesis has been approved by Graduate School Burapha University to be partial fulfillment of the requirements for the Master Degree of Science in Geoinformatics of Burapha University

.....

Dean of Graduate School

(Associate Professor Dr. Witawat Jangiam)

.....

65910029: MAJOR: GEOINFORMATICS; M.Sc. (GEOINFORMATICS)

KEYWORDS: Wildfire, Air pollution, Particulate Matter PM2.5, Remote sensing, Machine learning, Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Convolutional Neural Network (CNN)

THIWAKORN SENA : ESTIMATE PARTICULATE MATTER PM2.5 CONCENTRATION IMPACT OF WILDFIRES USING MACHINE LEARNING IN CHIANG MAI PROVINCE, THAILAND. ADVISORY COMMITTEE: ZHENFENG SHAO, Ph.D. PHATTRAPORN SOYTONG, Ph.D. 2025.

Wildfires are one of the most prominent problems with wide-ranging impacts on terrestrial ecosystems around the world. The important factor that causes air pollution is that the main cause is burning in open areas and large forest areas. Thailand experiences PM2.5 concentrations that are increasing every year, during the winter and dry season from December to April. Most of the concentration is concentrated in the central and northern regions of Thailand, especially Chiang Mai Province. PM2.5 has an effect on the economy and is very dangerous to the health of residents. However, air quality monitoring is often measured with a limited surrounding station.

The insufficient number of monitoring stations is the challenge, rendering the measurement of PM2.5 concentrations less reliable and incongruent with the actual environmental conditions.

This research is directly aimed at assessing PM2.5 concentrations resulting from wildfires using Remote Sensing data.

The main contents of this thesis include:

Evaluate the performance of the developed models in estimating PM2.5 concentrations at on-site scales and different seasons on regional scales within Chiang Mai Province, Thailand, in 2023.

comparing and determining optimal models for estimating PM2.5 concentrations, as well as identifying the primary factors influencing variations in pollution levels in regions impacted by severe wildfires.

Create a map of the spatial distribution of PM2.5 concentrations in areas that do not have ground measurement stations with remote sensing data using machine learning.

The results show that the Random Forest (RF) model demonstrates higher performance than the XGBoost and CNN models in estimating PM2.5 concentrations at on-site scale measuring, as evidenced by determination coefficients (R^2) values of 0.74–0.91, RMSE values of 10.40–30.53 μg/m,^{3} and MAPE values 18.56–36.48 μg/m^3, respectively. And the model demonstrated an average concentration all stations with an R2 value of 0.89, an RMSE of 11.61 μg/m³, and a MAPE of 34.22 μg/m³. Moreover, the RF model estimated the significance of features importance on PM2.5 concentration, including aerosol AOD-MAICA (MCD19A2) at 40%, AOT550 nm from MERRA-2 at 22%, dust mass PM2.5 from MERRA-2 at 12%, and CO from Sentinel-5P TROPOMI at 11%, ranking highest among all chemical components due to origin from combustion, aligning with the hypothesis that wildfires and greenhouse gas emissions significantly impact air quality.

In addition, the RF model was used to predict values and create spatial distribution maps for Chiang Mai Province, Thailand, in 2023. The model's average accuracy had an R^2 of 0.81, RMSE of 14.45 μg/m^3, and MAPE of 21.25 μg/m^3.

The spatial distribution of PM2.5 concentrations reveals distinct patterns, with elevated levels observed in both the north zone and south zone of the study area. This is most consistent with data on wildfire activity. Heat maps of severe wildfires based on ground truth data show the actual heat points from combustion consistent with the same period of the month where the concentration of PM2.5 was high from January to April. This is the clearest confirmation concerning the impact of wildfires on air quality and PM2.5 concentration values, which is in accordance with the assumptions based on the research objectives.

This thesis employs the development of machine learning models in conjunction with the utilization of remote sensing data to efficiently evaluate PM2.5 concentrations, and the results are spatial distribution maps that can be reliable. In future research, it is recommended to undertake further refinement of the RF model's training process, which may entail the exploration of additional features or the fine-tuning of hyperparameters to enhance its predictive capabilities. Integration of additional data is another critical aspect for enhancing the robustness of PM2.5 estimation models.

มหาวิทยาลัยบูรพา
BURAPHA UNIVERSITY

ACKNOWLEDGEMENTS

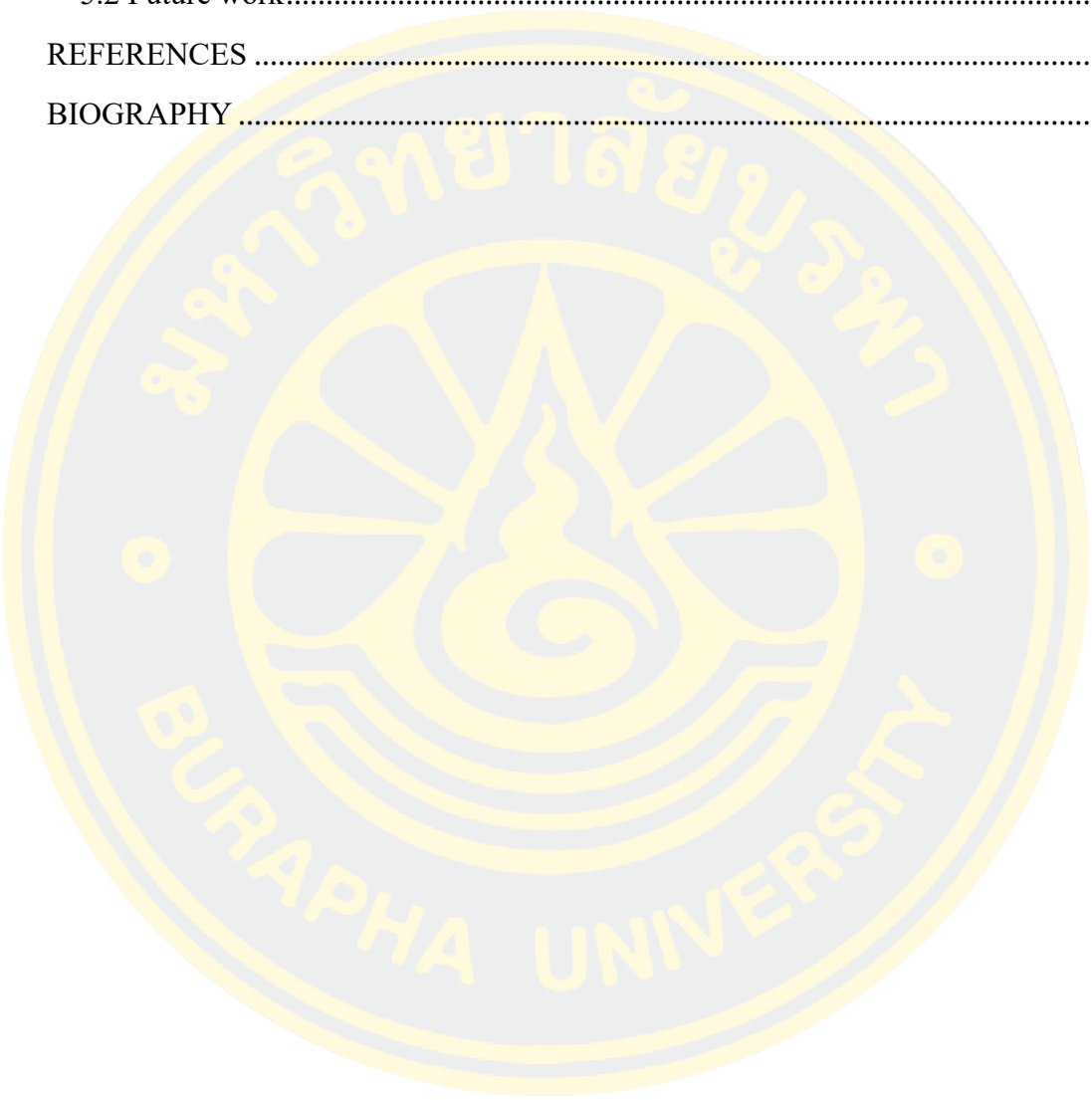
During my master's degree studies, I have to thank my supervisor, Prof. Zhenfeng Shao, for his invaluable guidance, advice, expertise, and experience, which provided insight and direction for my dissertation work. In addition, it is also an inspiration for me in my life of study and further research. I would also like to thank the PhD senior, Dr. Akib Javed, for helping me with advice on writing my thesis throughout. And I would like to thank Asst.Prof. Phattharaphorn Soithong, my co-advisor, for kindly helping me and giving valuable knowledge in studying and living life for me. I would like to thank the Chiang Mai Forestry Department for field survey data on areas burned by wildfires for use as data for training the model. We would also like to thank the Pollution Control Department for supporting air quality measurement data in Chiang Mai Province. For kindly providing PM2.5 information for training and training machine learning models. Thank you to Flying Officer Phongphat Japhichom; he is both my colleague and classmate for helping with the technical methods of doing this research. Thank you, senior. The two SCGI#4 students, Dr. Sunantha ousaha and Flight Lieutenant Sitthisak Thotsaphon, gave their best advice in doing this research. Thank you, my friend SCGI#5 student, who is my fellow countryman and gradually gives strength and assistance in living life Studying for 2 years was very fun and fulfilling. Thank you, lecturer from the Faculty of Geo-Informatics, Burapha University, Thailand, and the State Key Information Engineering Laboratory for Surveying, Mapping, and Remote Sensing (LIESMARS), China, as well as the Geo-Informatics and Space Technology Development Agency of Thailand (GISTDA), for to a great opportunity for me to continue my studies. Thank you to the commander of the Royal Thai Air Force (RTAF) for allowing me this opportunity to enhance my experience. Finally, I would like to thank my father and mother, my elder sister and family. In giving encouragement and pure love, always including everyone behind the scenes.

It will be impossible for me to graduate. Without the help of these people.

Thiwakorn Sena

TABLE OF CONTENTS

Page



LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ACRONYMS AND ABBREVIATIONS

LST	Land surface temperature
NDVI	Normalized differential vegetation index
IDW	Inverse distance weighting
GEE	The Google Earth Engine cloud platform
PM2.5	Particulate Matter <2.5 micron
AOT	Aerosol Optical Thickness
AOD	Aerosol Optical dept
MAIAC	Multi-Angle Implementation of Atmospheric Correction
AAI	Absorbing Aerosol Index
NO2	Nitrogen Dioxide
O3	Ozone
CO	Carbon Monoxide
DUSMASS25	Dust surface mass concentration PM < 2.5 μm
TOTEXTTAU	Total Aerosol Extinction AOT 550 nm
TEMP	Temperature
RH	Relative humidity
WS	Wind speeds
WD	Wind Direction
PRS	Pressure
PRE	Precipitation
TMD	Thai Meteorological Department
R^2	Coefficient of determination
RMSE	Root mean square error
MAPE	Mean Absolute Percentage Error
PCD	Pollution Control Department
LC	land cover
MODIS	Moderate Resolution Imaging Spectroradiometer
MERRA-2	The Modern-Era Retrospective Analysis for Research and Applications, Version 2
TIRS	Thermal infrared sensor
AVHRR	Advanced Very High-Resolution Radiometer
MISR	Multi-angle Imaging SpectroRadiometer
AERONET	AErosol RObotic NETwork
GSI	Grid point statistical interpolation
GEOS	Goddard Earth Observing System
GMAO	Global Modeling and Assimilation Office
QA	Quality assessment
CWV	Column water vapor
VOCs	volatile organic compounds
NOAA	National Oceanographic and Atmospheric Administration

AVHRR	Advanced Very High Resolution Radiometer
RF	Random Forest model
RFR	Random Forest Regression model
XGBoost	eXtreme Gradient Boosting model
CNN	Convolutional neural network model
API	Application Programming Interface
GISTDA	Geo-informatics Technology Development Agency
DNP	Department of National Parks, Wildlife and Plant Conservation
RFD	Royal Forestry Department
ALOS	Advanced Land Observing Satellite

CHAPTER 1
INTRODUCTION

1.1 Background and significance of the problem

Wildfire is one of the dominant disturbances across terrestrial ecosystems globally (Bowman et al., 2009). Fire severity is defined as the loss of above- and below-ground organic matter and is correlated with fire intensity within plant communities with similar vegetation structure (Hammill et al., 2006; Keeley, 2009). Fire severity is an important component of the fire regime, as it influences the post-fire response of plant and animal communities (Smucker et al., 2005). However, large-scale and long-term forest fires, such as those in the Amazon in South America and Australia in 2019 and in California, USA in 2020, have occurred frequently on the planet and still affected not only forest loss but also biodiversity, and ecosystem balance and impacted alters air pollution. These processes result in the emission of dust, smoke, and fine particulate matter into the atmosphere, which subsequently disperses across the region, influenced primarily by meteorological factors. Monitoring wildfire severity is facilitated by the use of satellite imaging products such as the MODIS-derived MCD64A1 (Terra & Aqua). Nevertheless, these products may encounter constraints such as cloud cover or water bodies, leading to potential false alarms and inaccuracies. To address this, we integrated ground-surveyed wildfire area data obtained from the Chiang Mai Forestry Department in 2023, offering crucial insights into the extent of fire damage. This dataset serves as a vital parameter for analysis and modeling, particularly as an independent variable in developing models to estimate PM2.5 concentration levels resulting from the release of fine particles in burned areas impacted by wildfires (Geng et al., 2018).

Environmental problems regarding air pollution occur both domestically and internationally around the world, by the important factors cause air pollution problems are the main reason is due to open burning and large forest areas, transportation industrialization, and the combination of other gases in the atmosphere, etc. Air pollution is the main culprit is the concentration of dust particles no larger than 2.5 microns that are harmful to health. Given its minuscule dimensions, PM2.5

particulate matter exerts detrimental impacts on the respiratory and cardiovascular systems, exacerbating conditions such as asthma, lung cancer, and cardiovascular diseases (Weichenthal et al., 2013).

Thailand experiences elevated annual PM2.5 concentrations, particularly during the dry season from November to April, with recent years witnessing an increase in haze occurrences during this period (PCD, 2023). Numerous studies conducted in various regions of Thailand, notably in the North and Bangkok (BKK), have linked this particulate pollution to industrial and vehicular emissions, biomass burning activities, and meteorological conditions characterized by stagnant air masses induced by cold surges, positioning Thailand as the 34th most polluted nation globally (ChooChuay et al., 2020; Narita et al., 2019; Phairuang et al., 2019). Ambient air pollution is linked to an estimated 40,000 deaths annually in Thailand (Pinichka et al., 2017), with high PM2.5 concentrations attributed to widespread smoke emissions from agricultural burnings and forest fires (Punsompong et al., 2021). The emission of PM2.5 from burning crop residue and forest fires is estimated at 141,000 and 5000 tons per year, respectively, with concentrations predominantly concentrated in central and northern Thailand (Chudnovsky et al., 2013; Junpen et al., 2013; Kanabkaew et al., 2011).

Chiang Mai Province in Thailand is one of the provinces that is ranked among the areas with PM2.5 concentration levels reaching critical levels and causing danger to people in the area because Chiang Mai Province has a problem with forest burning. and becoming more intense every year. The emissions of smoke from intense wildfires wield a pronounced and escalating impact on the comprehensive air quality, driven by the severity of the conflagrations. Furthermore, these emissions engage in substantial interactions with climatic processes, which have undergone amplification in intensity and scope owing to climatic modifications, including the prolongation and intensification of summer heat, thereby elongating the wildfire season. These climate-induced shifts exert profound repercussions on landscapes, notably within the geographical confines of the Chiang Mai province in Thailand. Chiang Mai province has a population of 1,792,474 people living there (Government of Thailand: 2022). It is a city of tourism in the country and has the largest area in the North and is the number 2 in the country. While both wildfires and smog situations have been analyzed by the Geo-Informatics Technology Development Agency (GISTDA) and Thailand's

National Park, Wildlife, and Plant Department, it has been determined that Thailand faces annual challenges stemming from wildfires, and the impacts are quite severe, escalating notably during the dry season, which typically forms in January and peaks in March. According to the statistics of wildfire occurrences in Thailand in 2023, one year, it shows the results of more than 4.7 million rai or approximately 7,520 square kilometers of burned areas (DNP, 2023; GISTDA, 2023).

Utilizing advanced Machine Learning methodologies, this research endeavors to analyze diverse factors contributing to the variability of PM2.5 and Remote Sensing (RS) data, which is considered a serious challenge for the application of satellite data in accurate and reliable estimation of PM2.5 levels. Factors such as weather patterns, wind speed, humidity, temperature fluctuations, seasonal variations, and other parameters influencing PM2.5 concentrations will be examined over the this of the study, Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Convolutional Neural Networks (CNN) in the machine-learning model was used to improve the accuracy relation of between burn area severity of wildfire to particulate matter PM2.5 concentration and prediction with the algorithm. Furthermore, the study seeks to develop sophisticated models or algorithms capable of effectively utilizing ground sensor data for PM2.5, coupled with satellite data (Aerosol Optical Depth (AOD), absorbing Aerosol Index (AAI), NO2, O3, and CO) and integration to enhance the accuracy of the evaluation within the study area, refining the understanding of the relationship between wildfire-related PM2.5 emissions and atmospheric conditions. Including the lack of air quality monitoring stations that cover the entire area. This makes it difficult to obtain air quality measurement data.

From all the problems mentioned see the relationship between the severity of wildfires that affect the concentration of PM2.5 aerosols, including the Estimation of PM2.5 concentration in Chiang Mai province. No air quality measurement station covers the entire area. Given the substantial financial and temporal resources necessary for expanding the network of air quality monitoring stations, employing satellite remote sensing for PM2.5 estimation emerges as a viable alternative to augment spatial coverage, overcoming inherent limitations (Son et al., 2023) and this poses a substantial challenge. Makes it possible to cope with the problem that occurs. Measures or policies can be set in advance for PM2.5 aerosols by applying remote

sensing data and machine learning, Python Application Programming Interface (API) code in a Jupyter notebook environment, which is widely used in remote sensing, especially for wildfire monitoring and detection.

1.2 Research questions

(1) How accurately do the developed models estimate PM2.5 concentrations at the on-site scale and different seasons on regional scale within Chiang Mai Province, Thailand?

(2) Which model proves to be the most optimal for estimating PM2.5 concentrations, and which factors significantly influence variations in pollution levels in areas affected by severe wildfires within Chiang Mai Province, Thailand?

(3) How can remote sensing data and machine learning techniques be effectively utilized to create a map depicting the spatial distribution of PM2.5 concentrations in areas lacking ground measurement stations within Chiang Mai Province, Thailand?

1.3 Research Objectives

(1) To evaluate the performance of the developed models in estimating PM2.5 concentrations at On-site scales and different seasons on regionals scale within Chiang Mai Province, Thailand in 2023.

(2) To compare and determine optimal models for estimating PM2.5 concentrations, as well as identifying the primary factors influencing variations in pollution levels in regions impacted by severe wildfires.

(3) To create a map of the spatial distribution of PM2.5 concentrations in areas that do not have ground measurement stations, with remote sensing data using machine learning.

1.4 Thesis structure

The goal of this research is to estimate the particulate matter PM2.5 concentration impacted by wildfires in Chiang Mai province, which is northern Thailand, that are affected by the smog problem from wildfires every year and tend to become more serious. There are officially main five chapters in this thesis. The first, chapter 1 presents the background and significance of the problem of dust and particulate

matter PM 2.5 caused by wildfires to estimate and monitor PM2.5 concentrations with Machine learning (ML) using remote sensing data, including research questions, research objectives, and thesis structure. The second, chapter 2 illustrated a literature review including the wildfires in Thailand and neighboring countries, foundation knowledge of PM2.5 particulate matter, Air pollution northern in Thailand, wildfires and impact on air quality, satellite image data for modeling, the google earth engine cloud platform, and Machine learning model. The third, chapter 3 show the materials data in use, step of processing, and methodology of this research, including determining the correlation coefficient between PM2.5 concentration and independent variables. The fourth, Chapter 4 demonstrate the analysis experiment and result of estimate PM2.5 concentration from remote sensing data at station measurement using machine learning, and identification of factors affecting of PM2.5 concentration, including evaluate and compare model performance, and finally show an evaluate seasonal and temporal distribution map of PM2.5 concentration. And the finally, chapter 5 is a conclusion and suggestions for future work in the research.

STRUCTURE OF THE THESIS	
CHAPTER 1 INTRODUCTION	• Background and significance of the problem • Research questions • Research Objectives • Structure of thesis
CHAPTER 2 LITERATURE REVIEW	• Wildfire in the Thailand and neighboring countries • Foundation Knowledge of particulate matter • Air pollution at the northern in Thailand • Wildfire and impacted on the air quality • Satellite image • The Google Earth Engine cloud platform • Machine learning models • Relation Research • Summary of this Chapter
CHAPTER 3 MATERIALS AND METHODS	• General background of study area • Data collections • Pre-processing and Processing data • Workflow of research • General statistics model • Prediction model • Model assessment accuracy • Summary of this Chapter
CHAPTER 4 RESULTS AND VALIDATION	• Statistics of Variables • Performance of the model • Feature importance of factors influencing variations • Spatial-Temporal distribution pattern • Summary of Experiment and Result
CHAPTER 5 CONCLUSION AND FUTURE WORK	• Conclusion • Future work

Figure 1 Main five chapter of the structure of the thesis.

CHAPTER 2
LITERATURE REVIEW

In this chapter 2 delves into the extensive impact of wildfires on air quality, particularly in Thailand and its neighboring countries. It examines a decade-long period from 2014 to 2023, revealing a significant number of forest fires, predominantly in Thailand's northern region. The data, drawn from government agency of Thailand sources (Geo-Informatics and Space Technology Development Agency (Public Organization) - GISTDA), highlights key trends and statistics regarding wildfire frequency and scale. Moreover, the chapter explores the link between wildfires and heightened PM2.5 particulate matter, emphasizing adverse effects on public health and the environment. The air Quality Modeling and PM2.5 estimation for determining the relationship between the severity of wildfires and their impact on air quality, The Google Earth Engine (GEE) cloud platform, and including to the global previous research with modern technology and tools, it stresses the imperative of monitoring and mitigating wildfire impact on air quality in regions prone to wildfires, like northern Thailand, as following:

2.1 Wildfire in the Thailand and Neighboring countries

Data sourced from the National Park Department's report on forest fires spanning a decade, from 2014 to 2023 (Fig. 2), reveals significant trends. During this period, a total of 51,265 wildfires were recorded, corresponding to over 81,935 detected hot spots (Fig. 3). The cumulative forest area affected by these fires amounted to 892,309 rai. Notably, the northern region emerged as the epicenter of wildfire occurrences within the country. The details are as follows:

Table 1 Report on actual wildfire areas in each region of Thailand for the 10-year period.

Regions	Number of Province	Total of wildfire occurred (point)	Burn areas (rai)
Northern	16	36,030	612,435
Northeastern	20	10,432	157,311
Central	15	4,055	69,067
Southern	14	748	53,229
Sum Region	**65**	**51,265**	**892,309**

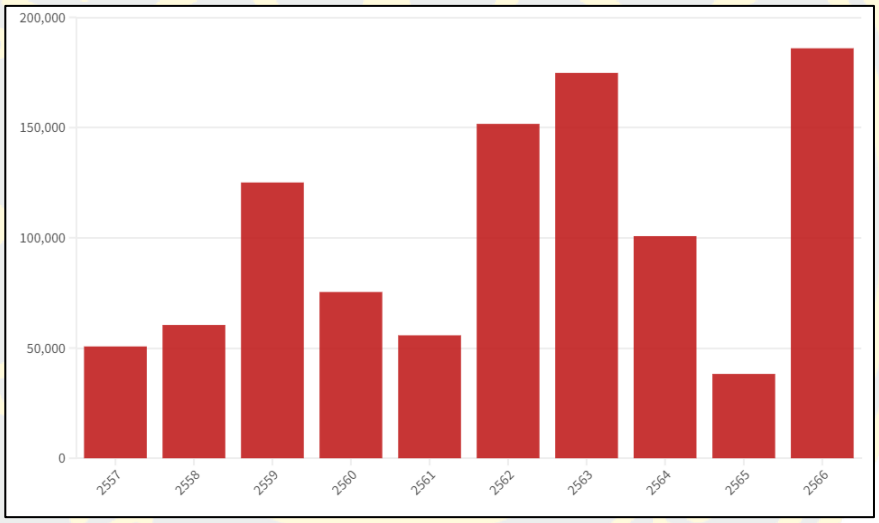


Figure 2 Report on actual wildfire areas in each region of Thailand for the 10-year period.

From reporting data obtained from the Ministry of Higher Education, Science, Research, and Innovation in Thailand, by the Geo-Informatics and Space Technology Development Agency (GISTDA, 2023), revealed a significant presence of high-temperature spots in the northern region, totaling 1,153 points. The highest concentration was observed in the Chiang Mai, with 584 points identified. This information was extracted from observations made utilizing data acquired from the Suomi NPP satellite, Moderate Resolution Imaging Spectroradiometer (MODIS (Terra&Aqua)), and several other satellites, on March 25, 2023.

Analysis of the satellite data indicated that the majority of heat signatures were detected in protected forest areas, accounting for 448 points, followed by national reserved forests with 431 points. Agricultural areas registered 138 points, while communities and other regions recorded 64 points. Additionally, areas under the jurisdiction of the Agricultural Land Reform Office and those along highways reported 62 and 10 points, respectively. Chiang Mai province exhibited the highest number of hot spots, followed by Mae Hong Son and Tak, with 314 and 255 spots, respectively.

In comparing the data on hot spot occurrences across neighboring countries collected from November 1, 2022 to December 15, 2023, it is evident that Burma exhibited the highest frequency of hot spots, with a total of 180,469 detections. Following Burma, Cambodia reported 125,935 spots, while Thailand documented 81,935 spots. Laos and Vietnam recorded 56,195 and 31,963 points, respectively. A detailed information is as follows:

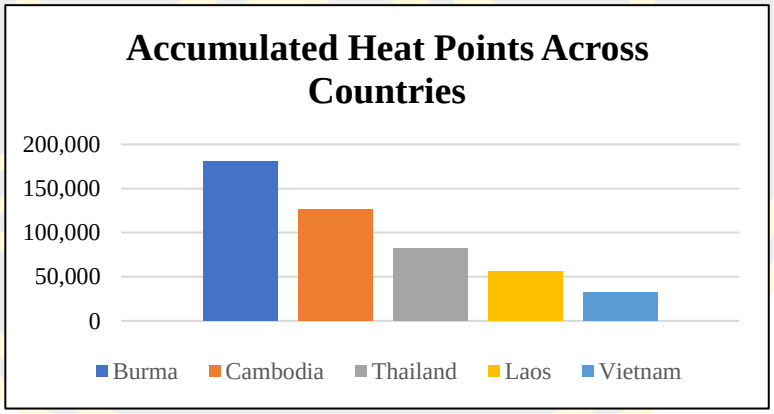


Figure 3 Accumulated Heat Point Across Countries during November 1, 2022 to December 15, 2023

2.2 Foundation knowledge of PM2.5 particulate matter

Particulate Matter PM2.5 refers to suspended particles in the air. The measurement of particulate matter involves separating particles by size to represent a group of pollutants, such as PM2.5, which denotes particles with a diameter smaller than 2.5 micron (in Fig. 4), and PM10, which denotes particles with a diameter smaller than 10 micron (Wilson, 1998). PM2.5, or fine particulate matter, consists of chemical particles generated from combustion processes such as wildfires, agricultural residue

burning (including rice, corn, and sugarcane), transportation, industrial emissions, household fuel use, etc. Some bordering provinces may experience the impacts of cross-border smoke haze.

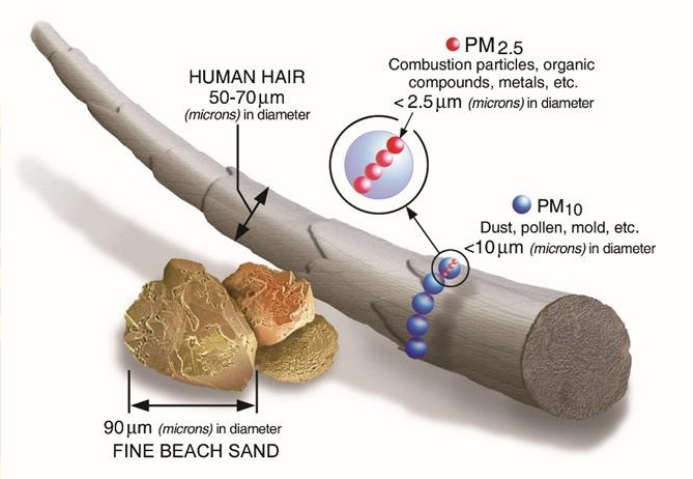


Figure 4 Dust particles no larger than 2.5 microns in diameter are about 1/25th the diameter of a human hair and nose hairs cannot filter. Can float in the air for a long time and up to 1,000 kilometers away.

Due to the issues of smoke haze, wildfires, and PM2.5 particulate matter, which have adverse effects on public health, the environment, and the nation's economy, on February 12, 2019, the government designated measures to address the particulate matter pollution problem. Consequently, in 2020, the government established a network for monitoring and surveillance of Thailand's air quality, overseen by the Pollution Control Department of the Ministry of Natural Resources and Environment. This network employs PM2.5 particle monitoring devices capable of continuous automatic monitoring 24 hours a day, throughout the year. It features a data transmission system via the internet and disseminates information through websites and the Air4Thai application, ensuring rapid data access for relevant agencies and the public. The Pollution Control Department (PCD) has deployed a considerable number of air quality monitoring instruments, with 14 stations in the northern region, 5 in the northeastern region, 3 in the western region, 11 in the eastern region, 7 in the southern region, 7 in the central region, 9 in the metropolitan area, covering a total of 37 provinces nationwide (Tatshakon.Pols, 2022). Shows in Figure 5

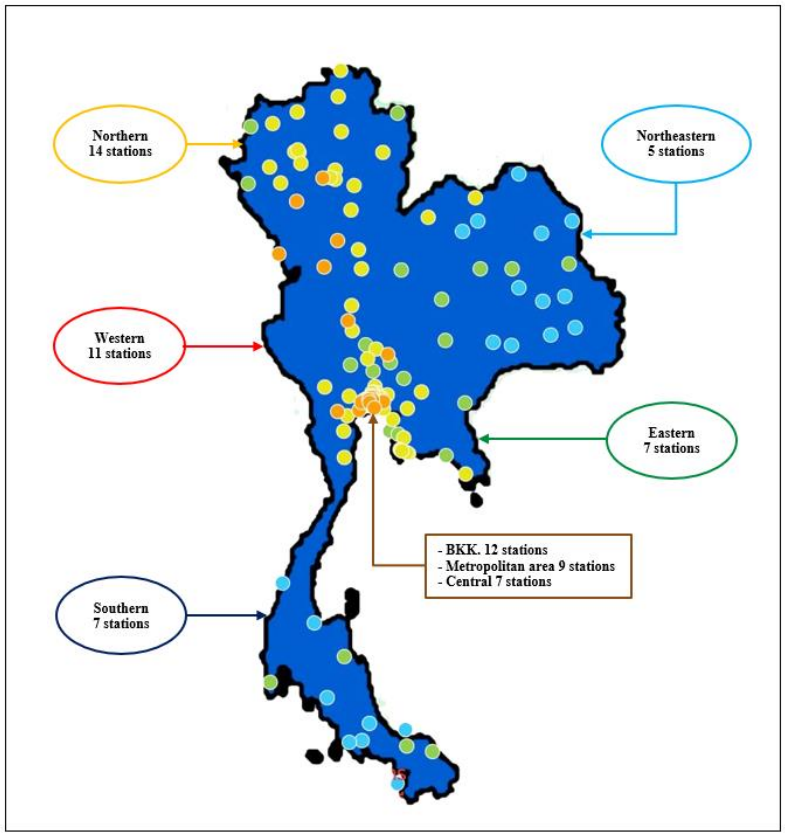


Figure 5 Air quality monitoring and surveillance network by the Pollution Control Department

2.3 Air pollution at the Northern in Thailand

The northern in Thailand faces significant challenges related to air pollution, notably due to several contributing factors. During the dry season, especially from January to May, the region experiences an increase in wildfires and open burning practices. These fires, often set for agricultural purposes like crop residue burning, contribute substantially to the release of particulate matter and other pollutants into the atmosphere and the geographical features of the region, including valleys and mountains, combined with specific weather conditions, contribute to the trapping of pollutants. Stagnant air and atmospheric inversion can lead to the accumulation of pollutants, exacerbating air quality issues. Including to rapid urban growth and industrial activities in cities like Chiang Mai contribute to local emissions. Increased vehicular traffic, industrial emissions, and construction activities are significant sources of air pollutants, including particulate matter (PM2.5).

Air quality in the northern of Thailand, is also affected by pollution from neighboring countries due to smoke of wildfire, especially during dry and thin weather conditions, this could have a significant impact on the air quality in this region. As a result, increased levels of particulate matter have been observed, especially PM2.5, which poses serious health hazards to residents. PM2.5 consist of fine particles with a diameter less than 2.5 micrometers that can penetrate the system respiratory tract is deep, this results in many health complications and including respiratory disease, cardiovascular disease and the aggravation of existing health conditions.

2.4 Wildfires and impacted on the air quality

Wildfires are significant contributors to air pollution, releasing a complex mixture of particulate matter PM2.5(ChooChuay et al., 2022) and chemical composition, including carbon monoxide (CO), nitrogen dioxide (NO2), and ozone (O3), which pose a significant risk to human health. PM2.5, which consists of fine particles with a diameter of 2.5 micrometers or less, is of particular concern due to its adverse impacts on respiratory health and visibility, presenting significant risks to human well-being. (Gupta et al., 2018). Recent studies by (Burke et al., 2021) have highlighted that wildfires alone contribute up to 25% of the global PM2.5 concentration, with localized impacts reaching as high as 50%.

Advancements in assessment and prediction technologies, particularly those leveraging machine learning techniques, have proven instrumental in understanding and mitigating the impacts of modern hazards like wildfires. Remote sensing platforms such as the Sentinel-5P TROPOMI (Atmospheric Airborne Index) and MODIS (Aerosol Optical Depth) missions have revolutionized air quality monitoring on a global scale, providing high-resolution data on pollutant concentrations (Bahadur et al., 2023). Moreover, the characterization of particulate matter, specifically PM2.5 and PM10 (particles with a diameter of 10 micrometers or less), has emerged as a critical aspect of air quality assessment, given their significant implications for human health and environmental quality (Mamić et al.,2023).

Integration of remote sensing data with ground-based measurements offers a comprehensive understanding of pollutant distribution and their sources. Machine

learning algorithms play a pivotal role in synthesizing these datasets, enabling the estimation, prediction, and analysis of pollutant concentrations with high accuracy. By elucidating the relationship between wildfire activity, particulate matter emissions, and ambient air quality, these techniques facilitate informed decision-making and effective mitigation strategies to safeguard public health and the environment.

2.5 Satellite image

Remote sensing data from several sources are used, in this research, divided into 2 components, as follows:

The first part, deals with aerosol properties and chemical composition, especially important for estimating PM2.5 concentration. It includes Sentinel-5P TROPOMI for the monitors atmospheric pollutants such as NO2, O3, CO, and aerosol absorbing index (AAI). Additionally, Aerosol Optical Depth (AOD) data, derived from the Multi-Angle Implementation of Atmospheric Correction (MAIAC) algorithms, specifically the MCD19A2 product from MODIS. The MERRA-2 reanalysis model data are to analyze and estimate Particulate Matter (PM2.5) concentrations due to incompleteness and missing Aerosol Optical Depth (AOD) data. The MERRA-2 reanalysis dataset provides a comprehensive record of aerosol properties from 1980 to the present, encompassing Total AOT550 nm and Dust surface mass PM2.5, which are instrumental in supplementing missing MODIS AOD data.

The second part deals with daily fire detection data and planetary surface temperature data, including elevation information, daily fire detection data, particularly the burn date dataset from the MCD64A1 product of MODIS, which offers fire detection, serving as an important source of PM2.5 distribution patterns. Furthermore, the Land Surface Temperature (LST) of the planet is derived by computing from Landsat 8 image data through the Google Earth Engine (GEE) tool, using from the thermal infrared band. Additionally, the utilization of the Digital Elevation Model (DEM) sourced from Copernicus Digital Elevation Model GLO-30 further enhances in this study.

2.6 The Google Earth Engine cloud platform

The Google Earth Engine (GEE) is considered a powerful platform for analyzing remote sensing data. Provides access to a large repository of time-series satellite imagery and vector data around the world. Backed by state-of-the-art cloud computing capabilities and complex software packages and algorithms tailored for the analysis of such datasets (Gorelick et al., 2017). This repository covers more than four decades of satellite imagery covering the entire Earth's surface. These datasets stem from a diverse array of satellite sources, including the comprehensive Landsat series, the precision of the Moderate Resolution Imaging Spectrometer (MODIS), the detailed observations of the National Oceanographic and Atmospheric Administration (NOAA), and Advanced Very High Resolution Radiometer (AVHRR), as well as the insightful data from Sentinel 1, 2, 3 and 5, and the Advanced Land Observing Satellite (ALOS), among others. A comprehensive compilation of various satellite-based products, ranging from raw sensor bands to pre-processed indices, composites, and high-resolution elevation models. For a thorough exploration of the complete suite of available datasets, interested users are encouraged to consult the portal webpage (https://earthengine.google.com/datasets/).

2.7 Machine Learning Models

Machine learning stands as a pivotal analytical tool within contemporary computational systems, facilitating learning, memory retention, and decision-making processes to address recurring challenges. Notably, machine learning algorithms are adept at simulating scientific expertise. Its pervasive application spans various domains, reflecting its capacity to emulate human cognitive processes and decision-making mechanisms. By leveraging datasets, machine learning empowers computers to autonomously derive informed decisions, thereby achieving results commensurate with human accuracy levels.

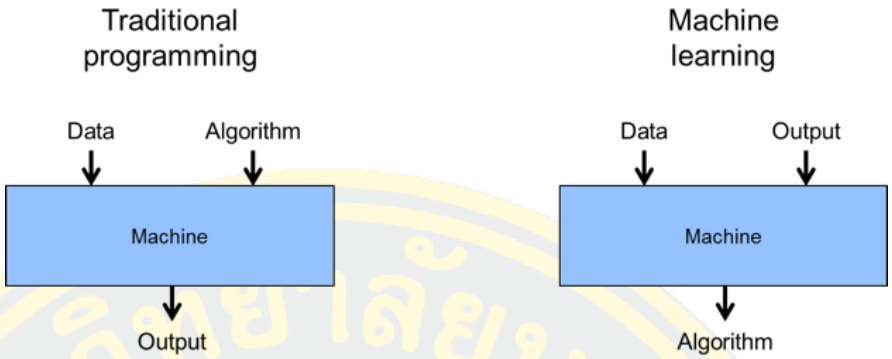


Figure 6 Comparisons between Traditional Programming with Machine Learning

2.7.1 Random Forest Regression

Random Forest Regression is an ensemble learning algorithm used for regression tasks. It creates multiple decision trees during training. Each tree is based on a random subset of features and data (Breiman, 2001; Lyu et al., 2023). Averaging the predictions from these trees gives more stable and accurate results. Randomization was applied through feature subset selection and bootstrapping. This increases the variety and durability of the models (Yang et al., 2022) and this method effectively mitigates overfitting, making it suitable for handling large datasets with high dimensionality. Additionally, Random Forest Regression is less sensitive to noisy data compared to individual decision trees. It excels in capturing complex relationships between variables and is widely used in various fields such as survey, remote sensing, and environmental science. Due to its ability to handle nonlinear relationships and interactions, it is particularly well-suited for modeling real-world phenomena. Overall, Random Forest Regression is a robust model and versatile technique that offers reliable predictions and scalability for regression problems.

2.7.2 eXtreme Gradient Boosting Regression

XGBoost regression, derived from the term eXtreme Gradient Boosting, is a widely praised machine learning algorithm. Known for its excellent performance in regression tasks (Prakash et al., 2023), it functions by iteratively assembling an ensemble of decision trees. Each subsequent diagram aims to correct the mistakes of its predecessor. This makes it a better choice for modeling complex systems (Yin et al., 2023). This iterative process is centered on the minimization of a specified loss

function, thereby refining the model's predictive accuracy at each step. Leveraging a gradient boosting framework, XGBoost harnesses gradients of the loss function to guide the tree-building process(Ma et al., 2023). Additionally, it incorporates regularization techniques such as shrinkage and tree pruning to mitigate overfitting, ensuring robust generalization to unseen data instances. Notably, XGBoost exhibits high scalability and efficiency, making it adept at handling voluminous datasets characterized by high dimensionality. Its inherent flexibility enables seamless integration of custom loss functions and evaluation metrics, thereby accommodating diverse regression scenarios. With its widespread adoption across scientific and industrial domains, XGBoost has firmly entrenched itself as the algorithm of choice for attaining cutting-edge results in regression problems.

2.7.3 Convolutional Neural Network

A Convolutional Neural Network (CNN) represents a specialized class of deep neural networks extensively utilized in image recognition and computer vision applications. It comprises several layers, notably including convolutional layers responsible for feature extraction, pooling layers for spatial dimension reduction, and fully connected layers for classification tasks (Giri et al., 2022; Purwono et al., 2023; Shamsaldin et al., 2019). Within the convolutional layers, filters or kernels are applied to input images to extract intricate features such as edges, textures, and shapes. Subsequently, pooling layers reduce the spatial dimensions of the feature maps generated by the convolutional layers, facilitating efficient computation. Fully connected layers then perform classification tasks based on the extracted features. CNNs are trained using backpropagation and optimization algorithms, notably gradient descent, to iteratively minimize the discrepancy between predicted and actual outputs. Various optimization techniques, including Genetic Algorithm (GA), Particle Swarm Optimization (PSO) (Kalaiarasi et al., 2023), adaptive stochastic gradient descent (aSGD) , beetle antennae search optimization algorithm (D. Chen et al., 2023), as well as several gradient-based optimizers like Adam, Nadam, and AdamW (Nurdiati et al., 2022), are utilized to augment the training of Convolutional Neural Networks (CNNs). The objective of employing these algorithms is to expedite training processes, enhance accuracy, and effectively optimize hyperparameters. Furthermore, CNN architectures offer

flexibility for customization and fine-tuning to cater to specific application requirements, with their performance further optimized through advanced techniques like data augmentation, transfer learning, and ensemble methods.

2.8 Relation Research

Utilizing Sentinel-5P and GEOS Forward Processing dataset, a novel methodology was devised to daily estimate comprehensive 5 km ambient concentrations of PM2.5 and PM10 during the wildfire season across China. The estimation function was derived by integrating data from multiple sources (Sentinel-5P TROPOMI, GEOS Forward Processing, and ground-based stations) through an ensemble machine learning approach, such as the light gradient boosting machine (Wang et al., 2021).

Hourly PM2.5 concentration ground measurements from 2015 to 2020 in Dalian, China, dataset from AOD-MAIAC (1 km resolution) product MODIS data, to better understand the spatio-temporal variations in the mass concentrations of particulate matter less than 2.5 μm (PM2.5). demonstrate that the spatial distributions of PM2.5 and AOD were consistent ($R^2 = 0.922$), with industrial regions having higher PM2.5 values. The test set was cross-divided by year, with PM2.5 serving as the output variable and AOD and climatic factors serving as the input variables. A neural network (BPNN) model the predicted and monitored values showed a consistent trend; the estimation results improved with the addition of meteorological factors; the test set's R^2 and RMSE were 0.01–0.05 μg/ m^3 and 0.663–0.752, respectively (Jilin et al., 2022).

The Random Forest (RF) and eXtreme Gradient Boosting (RF-XGBoost) model was offer to estimate PM2.5 in Guanzhong Urban Agglomeration (GUA) during the winter of China's regions with the peak of PM2.5 concentrations, by using AOD data, the Multi-Angle Implementation of Atmospheric Correction (MAIAC) from MODIS, along with dense meteorological and topographic conditions, land-use, population density, and air pollution data. Evaluation of the RF-XGBoost model through out-of-sample testing yielded excellent results, with a coefficient of determination (R^2) of 0.93, root-mean-square error (RMSE) of 12.49 μg/ m^3, and mean absolute error (MAE) of 8.42 μg/ m^3. Results from the model indicated severe pollution in GUA

during the winters of 2018 and 2019, attributed to coal burning for heating and unfavorable meteorological conditions (Lin et al., 2022).

Utilized Sentinel-5P TROPOMI and MODIS MOD64A1 by GEE could platform for assessing tropospheric and burned data from India's Kharif/Rabi stubble burning. Agricultural areas were extracted from a generated LULC map using monthly MOD64A1 data. Sentinel-5P TROPOMI evaluated pollutant concentration during stubble burning months. The HYSPLIT model analyzed air mass movement frequency. Findings indicated substantial rises in various pollutant concentrations during specific months. Results were verified using temperature, precipitation, and ground station PM2.5/PM10 data. HYSPLIT revealed higher pollution in specific months yearly. The study affirms Sentinel-5P TROPOMI's effectiveness in tracking contaminants and assessing air quality (Neeraj et al., 2022).

Estimation concentration of PM2.5 and PM10 air pollution caused by burning in the area. Both agriculture and forest burning in the area of Chiang Mai and Lamphun provinces, which is in the northern part of Thailand, that exceeds the standard with sampling data from ground measurement stations. The study utilized a mini-volume air sampler to collect air samples for analysis. Gas analysis was conducted using a gas analyzer (Testo 350 XL) to measure the concentrations of gases such as O2, CO, CO2, NO, NO2, NOx, and SO2.The study also measured PM10 concentrations using a mini-air sampler. The air quality standards for PM2.5 and PM10 were compared with the measurements obtained from the PCD stations. The study referenced air quality standards from Thailand, the WHO, and the US-EPA for comparison. The standard for PM10 (24 hours) was set at 120 mg/ m^3 (Suthini et al., 2018).

The linear regression model using incorporated MODIS (AOD) measurements using the bilinear interpolation technique, other air pollutants, meteorological factors, and greenness indicators (NDVI) to estimate ground-level PM2.5 concentration in the Bangkok Metropolitan Region of Thailand in 2022. The 12-fold cross-validation technique was employed to assess the accuracy of the model's performance.

The measured PM2.5 concentration's yearly mean (standard deviation) was 22.37 (± 12.55) µg/ m^3, while the seasons' mean (summer, winter, and rainy seasons) (standard deviation) was 18.36 (± 7.14) µg/ m^3, 33.60 (± 14.48) µg/ m^3, and 15.30 (± 4.78) µg/ m^3, in that order. the cross-validation produced R^2 values of 0.48, 0.55, 0.21, and 0.52 (Peng-in et al., 2022) .

2.9 Summary of this Chapter

The presents a comprehensive overview of wildfires in Thailand and neighboring countries, focusing on the impact of these fires on air quality, particularly regarding the concentration of PM2.5 particulate matter.

Over a decade from 2014 to 2023, Thailand recorded a significant number of forest fires, with the northern region being the most affected. A total of 51,265 forest fires were recorded during this period, affecting approximately 892,309 rai of forest area. The northern region accounted for the majority of wildfire occurrences, with 36,030 wildfires affecting 612,435 rai of land. The correlation between wildfires and the increase in PM2.5 particulate matter, emphasizing the adverse effects on public health and the environment. Factors contributing to air pollution in northern Thailand include agricultural practices, industrial emissions, and transboundary pollution from neighboring countries. And to address the particulate matter pollution problem, the Thai government established a network for monitoring air quality, employing PM2.5 particle monitoring devices across various regions. Additionally, advancements in assessment and prediction technologies, including machine learning algorithms and remote sensing platforms, have been instrumental in understanding and mitigating the impacts of wildfires on air quality. To assess air quality more accurately, this research focuses on using Machine Learning (ML) Models various algorithms such as Random Forest Regression (RFR), eXtreme Gradient Boosting Regression (XGBoost), and Convolutional Neural Networks (CNN), which are utilized for modeling and predicting air quality parameters, including PM2.5 concentrations. And to demonstrate the viability and accuracy in this study, which several studies are referenced, highlighting the use of satellite data, ground

measurements, and machine learning techniques to assess air quality, track pollutants, and estimate PM2.5 concentrations in different regions, including Thailand and China. Overall, the emphasizes the urgent need for monitoring and mitigating the impact of wildfires on air quality, particularly in regions prone to wildfires like northern Thailand, through the application of advanced technologies and interdisciplinary research approaches.

CHAPTER 3
MATERIALS AND METHODS

The research focuses on Chiang Mai Province, Thailand, selected as the study area due to its critical importance in air quality management, specifically addressing PM2.5 pollution concerns. This region is notable for its persistent challenges with elevated PM2.5 concentrations, largely influenced by factors such as agricultural burning, forest fires, and industrial activities. This chapter provides comprehensive insights into the study area, encompassing detailed discussions on data collection methodologies, data preprocessing techniques, and the employed methodology for PM2.5 concentration estimation. Additionally, it delves into the application of machine learning algorithms for PM2.5 estimation and evaluates the accuracy of the developed models as following:

3.1 General Background of Study Area

The Chiang Mai Province is located at latitude 17.25° and 20.16° N and longitude 98.08° and 99.66° E, has an area of 20,107.057 km^2. General topography area is mountainous. The weather is quite cool most of the year average annual temperature is 25.4 °C, with an average maximum temperature of 31.8 °C and an average minimum temperature of 20.1 °C. There is an average rainfall of 1,100-1,200 mm/year.

Because the terrain is forest and high mountains. This causes during the winter and dry season the air is thin and has lower moisture. This creates a risk of fuel ignition in the forest from the accumulation of vegetation during the rainy season. As a result, forest fires occur every year and their severity increases according to the climate change. The result is air pollution that affects the health of residents and also has an impact on tourism which is the main economy of the province. The study area is show in Figure 7

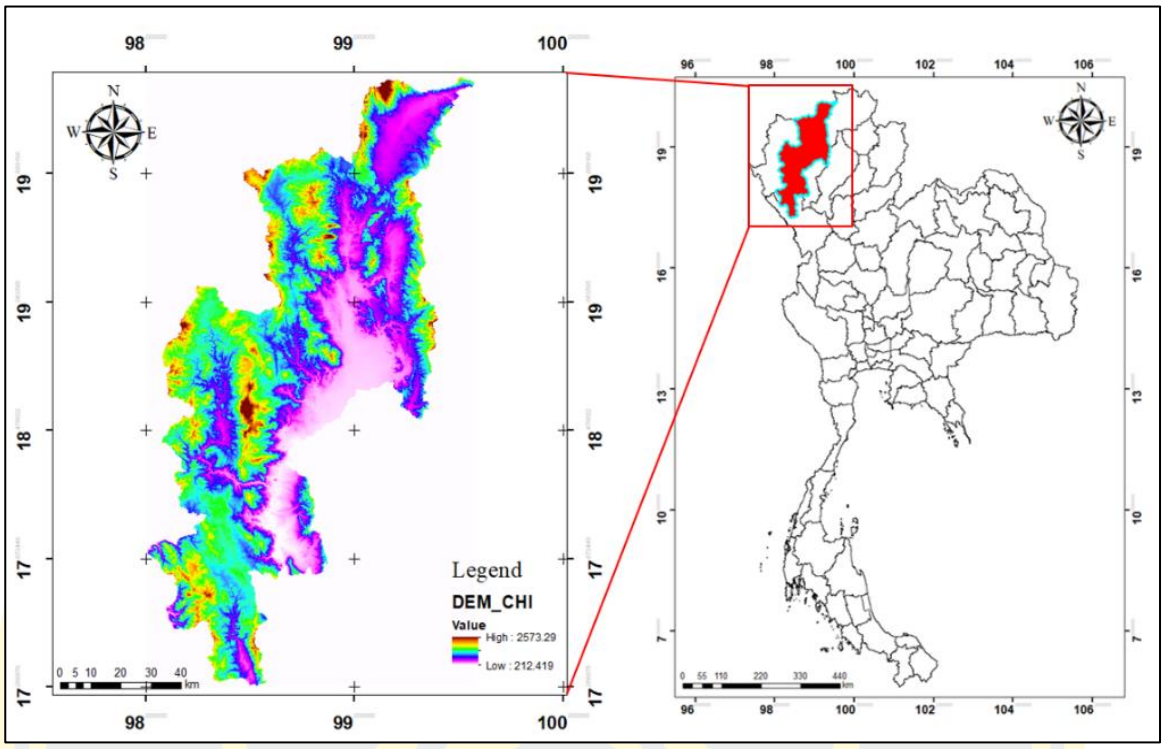


Figure 7 Study area the northeast of Thailand

3.2 Data Collections

3.2.1 Remote Sensing data

3.2.1.1 Sentinel-5P TROPOMI

In the analysis and estimate of Particulate Matter PM2.5. The Sentinel-5P TROPOMI is the Tropospheric Monitoring Instrument observes concentrations of atmospheric pollutants and gases such as NO2 Column number Density, O3 total Atmospheric Column, CO Column number Density, and Aerosol Absorbing Index (AAI) which are emitted into the atmosphere, due to wildfire. It uses the TROPOMI instrument, a multispectral sensor that records the reflectance of wavelengths optimized for measuring the atmospheric concentration of gases at a spatial resolution of 0.01 arc degree (1.11 km). And from study and finding additional information, it was found that Nitrogen dioxide (NO2) and ozone (O3) can has contribute in quantity of Particulate Matter PM2.5 to the increase and affect the concentration including to spread by chemical reaction. Due to NO2 acts as a precursor to the formation of nitrate particles, and Ozone can oxidize volatile organic compounds (VOCs). These

reactions result in the formation of PM2.5 particles, adding to their concentration in the air. Therefore, both the NO2 and O3 quantities which included as parameters to improve the accuracy of concentration estimation and prediction of PM2.5 and to be a parameter for machine learning in the period time of study were retrieved through computing from the Google Earth Engine (GEE) APIs, covering the fire area, in order to understand the spread of Particulate Matter PM 2.5 concentrations by Google Earth Engine.

3.2.1.2 Aerosol Optical Depth product of MCD19A2 from MODIS

Aerosol Optical Depth (AOD) data for the estimation of PM2.5 in this study, obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS) instrument on the Terra and Aqua Collection 6.1 combined Multi-Angle Implementation of Atmospheric Correction (MAIAC) algorithms which is Aerosol Optical Depth (AOD) the MCD19A2 product grid Level 2 product produced daily. The MCD19A2 (AOD) data product includes blue aerosol product (Aerosol optical depth over land at a wavelength of 0.47 μm.), Green bands aerosol product (Aerosol optical depth over land at a wavelength of 0.55 μm.), AOD quality assessment (Purwono et al., 2023), Column water vapor (CWV), Smoke injection height, Fine Mode Fraction, Angstrom exponent 470-780 nm over the ocean, Uncertainty the cosine of solar zenith angle, the cosine of view zenith angle, the relative azimuth angle, the scattering angle, and the glint angle at 5 km, etc.

In this study, used the Aerosol Optical Depth AOD-MAIAC data at 0.55 μm. with a resolution of 1 km values exceeding 4 were systematically excluded from the modeling process (Di et al., 2016), primarily due to factors such as intense surface brightness, cloud interference, and the presence of extreme outliers. Consequently, approximately 39% of the Aerosol Optical Depth (AOD) dataset was deemed missing (Liu, Paciorek, & Koutrakis, 2009), significantly impeding the efficacy of the model. We were extracted and retrieved through computing from the Google Earth Engine (GEE) (https://code.earthengine.google.com (accessed on 01 Jan to 31 Jun 2023)), cloud platform, covering the fire area, in order to understand the spread of Particulate Matter PM 2.5 concentrations, show in the table 2

3.2.1.3 The product of MCD64A1 resolutions 500 m from MODIS

The MCD64A1 is latest version in the MODIS BA product line, offering spatial resolution of 500 m. Its lies in its novel fusion methodology, which integrates the MODIS Terra and Aqua daily surface reflectance products (MOD09GHK/ MYD09GHK) at 500 m. with the respective MODIS active fire datasets (MOD14A1 and MYD14A1) at 1-kilometer resolution, enabling precise identification and characterization of daily fires at 500 meters granularity. Central to this approach is the utilization of the Burn Sensitive Vegetation Index (VI), a metric derived from MODIS atmospherically resolved shortwave infrared in conjunction with surface reflectance bands 5 and 7, encompassing transient surface measurements (Giglio, Boschetti, Roy, Humber, & Justice, 2018). Further insights into the algorithmic intricacies are delineated within the MCD64A1 product user manual. Noteworthy is the substantial improvement observed in MCD64A1 compared to earlier MODIS collections, particularly in the detection of smaller Burned Areas (Junpen et al.), as corroborated by various research findings (Rodrigues et al., 2019). Leveraging the MCD64A1 dataset retrieval through the GEE platform corresponds aptly with the timeline pertinent to the PM2.5 concentration study.

3.2.1.4 Total AOT 550 nm and Dust surface mass PM2.5 from MERRA-2 Reanalysis

MERRA-2, a sophisticated reanalysis product developed by NASA's Global Modeling and Assimilation Office (GMAO) (Gelaro et al., 2017), provides an extensive dataset about of aerosol properties spanning from 1980 to the present day. This reanalysis leverages the atmospheric model from the Goddard Earth Observing System (GEOS) framework (Molod, Takacs, Suarez, & Bacmeister, 2015; Rienecker et al., 2011), employing advanced grid point statistical interpolation (GSI) techniques (Kleist et al., 2009; W.-S. Wu, Purser, & Parrish, 2002) for spatial analyses. The model employs a cubic horizontal discretization with a spatial resolution of approximately 0.5 × 0.625° and integrates 72 vertical pressure layers extending from the Earth's surface to 0.01 hPa. Moreover, MERRA-2 utilizes a sophisticated 3D model data assimilation algorithm with a 6-hour GSI update cycle and incorporates timely first guess estimates. These assimilation techniques are further enhanced using the incremental analysis update method (Bloom, Takacs, Da Silva, & Ledvina, 1996) to refine the background state estimates.

Additionally, MERRA-2 generates a high-resolution 3-hourly global gridded aerosol analysis with bias correction for aerosol optical depth (AOD) at 550 nm, sourcing AOD values from various satellite platforms, including MODIS, Advanced Very High-Resolution Radiometer (AVHRR), Multi-angle Imaging SpectroRadiometer (MISR), and ground-based AErosol RObotic NETwork (AERONET) (Gelaro et al., 2017). However, it's important to note that the assimilation of specific sensor data into MERRA2 varies annually due to data availability constraints. For example, MODIS data assimilation began after 2000, whereas MISR assimilation ceased after 2015(Alcaras et al., 2022). Furthermore, MERRA-2 provides diagnostic data on aerosol species, such as Total Aerosol Extinction AOT 550 nm (TOTEXTTAU) and aerosols (Dust surface mass concentration PM < 2.5 μm (DUSMASS25)) obtained from simulations of the Goddard chemical aerosol radiation and transport model (Chin et al., 2002; Colarco et al., 2010) at a fine global resolution of approximately 50 × 65 km (0.5° latitude and 0.625° longitude). In this study, hourly aerosol diagnostics at the surface and column AOD analyses were by extracting data using the Google Earth Engine (GEE) Cloud platform. Consequently, we employ MERRA-2 reanalysis data as an independent input for our model to replace the missing dataset of AOD-MAIAC aerosol data from MODIS. Details are show in the table 2

Table 2 Type and factor of data

Q0	Temporal Coverage	Resolution	Source
Absorbing Aerosol Index (AAI)	01 Jan - 31 Jun 2023	0.01 arc degrees (1.11 km)	Sentinel-5P TROPOMI
Nitrogen Dioxide (NO2)			
Ozone (O3)			
Carbon Monoxide (CO)			
MCD19A2 (AOD-MAIAC)	01 Jan - 31 Jun 2023	250 m	MODIS (Terra & Aqua)
Dust surface mass concentration PM < 2.5 μm (DUSMASS25)	01 Jan - 31 Jun 2023	0.5 x 0.625°	MERRA-2 Reanalysis
Total Aerosol Extinction AOT 550 nm (TOTEXTTAU)			

3.2.1.5 Land Surface Temperature Using Landsat-8

Land surface temperature (LST) is an important environmental variable. This topic provides a brief overview of how LST is calculated and analyzed using Landsat 8 image data on Google Earth Engine (GEE) instrumentation from the thermal infrared band, also known as the thermal infrared sensor (TIRS band10) for LST estimation using Landsat data processing and LST data acquisition. Calculating land surface temperature (LST) using Planck's law. Formulas are provided as mathematical expressions. It involves using the thermal band (TB) and emissivity (EM) to calculate the LST using the highest and lowest values of the normalized differential vegetation index (NDVI) and then calculating the Proportion of Vegetation (PV), a metric used to quantify the relative abundance of vegetation within a given area. and Emissivity (EM) are important parameters for accurate land surface temperature (LST) calculations. Both PV and EM are used as important parameters for generating LST from Landsat-8 image.

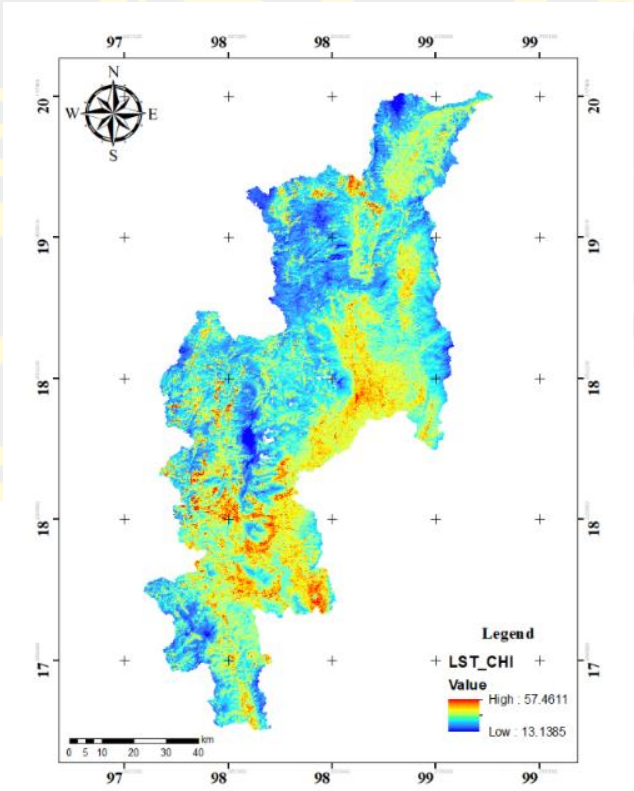


Figure **8** Land Surface Temperature form Landsat 8

3.2.1.6 Type and factor of data

The Copernicus DEM, known as GLO-30, is a Digital Surface Model (DSM) representing Earth's surface, encompassing buildings, infrastructure, and vegetation. Derived from an edited DSM named WorldDEM™, it includes adjustments such as flattening of water bodies and ensuring consistent flow of rivers. Additionally, editing of shorelines, coastlines, and specific features like airports and terrain structures has been implemented. This dataset was released in 2020 and is derived from the WorldDEM data, which itself is based on radar satellite data obtained during the TanDEM-X mission by Airbus. The primary aim of the TanDEM-X mission was to generate a global coverage DEM using InSAR in HRTI-3 standards. The TanDEM-X data acquisition spanned from December 2010 to January 2015. The Copernicus GLO-30 dataset features a grid spacing of 1 arc-second (30 meters) and is standardized to a 1-degree by 1-degree extent. This data was extracted using the Google Earth Engine platform with the code ID (ee.ImageCollection ("COPERNICUS/DEM/GLO30")). The inclusion of this dataset is anticipated to have a significant impact on PM2.5 concentrations.

3.2.2 Ground Station data

3.2.2.1 Thailand's network for monitoring air quality

Daily concentration of air pollution (PM2.5 (μg/ m^3)) with 6 monitoring Ground-base stations, show in the table 3 the Pollution Control Department (PCD) in Thailand, track of the concentration of pollutants on a daily basis. To demonstrate how the severity of the wildfire affects the concentration of particulate matter (PM2.5). This study matched station readings with satellite data from the first 6 months of 2023 to illustrate the period of high concentration and widely spread following the fires, encompassing both pre-fire and post-fire periods. The accuracy of the satellite data was assessed by the use of ground measurement stations.

Table 3 PM2.5 Ground-base stations in the study area

NO.	ID Station	Name Station	Latitude	Longitude	Source
1	35T	Chiang Mai Center Provincial	18.8370549	98.9683173	Pollution Control Department (PCD)
2	36T	Yupparaj Wittayalai School	18.7909216	98.9854743	
3	O27	Muang Na Subdistrict Municipality	19.5932352	98.9591591	
4	O28	Mae Chaem Hospital	18.4980446	98.3767308	
5	O70	Phubing kharajaniwet Palace	18.8053051	98.8978763	
6	O71	PEA Hkod District Branch	18.1565307	98.4093909	

3.2.2.2 Meteorological Factors

Meteorological variables have been utilized as predictive factors affecting the condition, characteristics, and dispersion of Particulate Matter (PM 2.5) concentrations in the atmosphere and are acknowledged to possess distinctive benefits in retrieving past characteristics of PM2.5. These variables include temperature (TEMP (°C)), relative humidity (RH (%)), wind speeds (WS (m/s)), wind direction (WD (m/s)), pressure (PRES (hPa)), and precipitation (PRE(mm)), sourced from the Thai Meteorological Department (TMD). In this study, used mean daily meteorological data points, incorporating parameters such as date, station location, latitude, and longitude, which suggests that ample rainfall and severe cross-ventilation are better suited for the sedimentation and dispersion of PM2.5 particles (Jiao et al., 2016). Additionally, this study utilized corresponding meteorological data to match image-PM2.5 pairs based on location and date, as presented in Table 4.

Table 4 Meteorological Factors data of station in study areas

NO.	Data Type	Description	Source
1	Temperature (TEMP)	327202 and 327501 Station ID.	Thai Meteorological Department (TMD)
2	Relative Humidity (RH)		
3	Wind speeds (WS)		
4	Wind Direction (WD)		
5	Pressure (PRS)		
6	Precipitation (PRE)		

3.2.2.3 Fire count data

In this study, Burn Area Ground Base refers to the ground truth data acquired from the Royal Forestry Department of Chiang Mai, Thailand, for the recent year of 2023, specifically detailing the occurrences of fires. The data was collected using ArcGIS 10.8 to map during the forest fires of 2023. Comprising approximately 1,600 data points, each point provides crucial information regarding the extent of damage caused by the fires and specify the date of the fire.

To facilitate comprehensive analysis, both the ignition date and extinguishment date of the fires were determined. Fire occurrence data is centered around the detection points within a 1 km radius of PM2.5 measurement stations. This dataset is structured in a shapefile format. Within the study timeframe, any day with recorded fire incidents within the designated area is classified as a "fire day," while days lacking such incidents are classified as "non-fire days." This fire count data will be incorporated as an independent variable within the model, serving as an input factor for analysis.

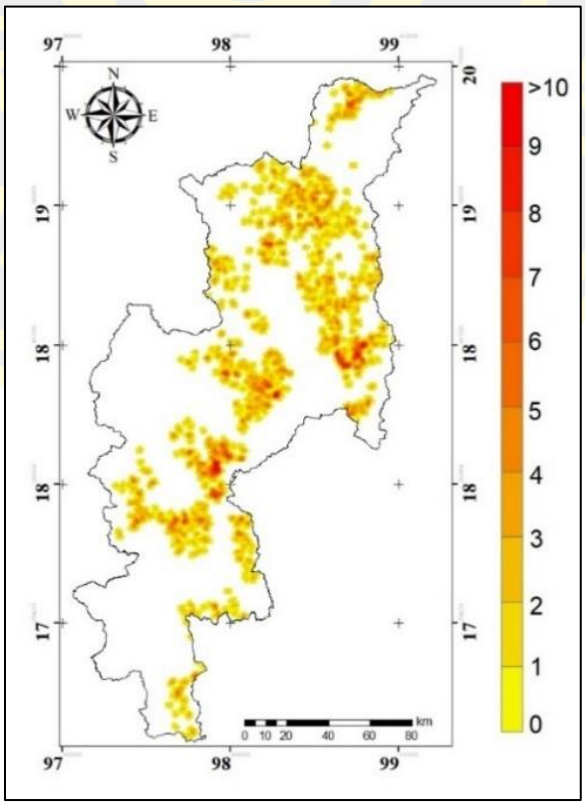


Figure **9** Royal Forestry Department (RFD: 2023)

3.3 Preprocessing and Processing data

Independent all data layers utilized in this study was extracted from Google Earth Engine (GEE) and reprocessed and transformed consistently into the Geodetic Coordinate System (WGS 84) and use all independent data pixels linked to a grid with a spatial resolution to the nearest 1x1 km centered on a designated PM2.5 station for analysis. For example, use a radius of 1 km around the PM2.5 monitoring station using composite bands tool on the ArcToolbox of ArcGIS 10.8., were resampled using a bilinear method and a normal distribution, due to the different spatial and temporal resolutions. This was done to ensure uniformity and compatibility for use as independent variables in the model (in order to corresponding to satellite over pass time 10.30 AM and 13.30 local time). The values of independent variables for each monitoring site were then extracted using the extraction tool (multi-values to points tool) of ArcGIS 10.8. After that, using Python coding was used to convert the vector data to raster data for training set and testing set. And to be utilized for prediction spatial distribution maps based on temporal variations across months and differences seasonal on region scale (see Fig.11).

Moreover, as a preventive measure against potential cloud-induced contamination, all pairs of AOD-PM2.5 possessing fewer than two pixels were excluded from the analysis. Additionally, to reduce the likelihood of combining fake AOD pixels, deviations are subtracted when AOD is greater than 1.5. Including to missing data and outliers due to unavoidable issues such as instrument errors and natural factors were addressed in advance. In cases where ground PM2.5 concentrations did not align completely with the independent variables due to missing data or outliers, these unmatched records were designated as invalid. The theoretical total number of records is 1,270 (multiplying 181 days at 5 sites and 365 days at 1 site in areas where fires occur almost all year round for use in plotting time series graphs in 1 year, according to the framework of the study period in the dry season where severe wildfires often occur). However, the actual number of valid records after matching was determined to be 1,180. The study period encompassed 6 months in 2023, and the number of monitoring sites was six site. For aerosol data from the MCD19A2 (AOD-MAIAC) product from MODIS, more than 39% of the data was found to be missing after

calculation. To address this, backfilling was performed using aerosol products from MERRA-2 (total aerosol extinction AOT 550 nm) and data from previous studies. Additionally, an analysis of total surface mass (PM2.5) emissions resulting from surface-level combustion was conducted using MERRA-2 data. Both datasets were extracted through Google Earth Engine (GEE).

Furthermore, meteorological data underwent cleaning processes in Microsoft Excel, including conversion to average hourly data and then to daily data. Inverse distance weighting (IDW) was employed to resample meteorological variables over each grid cell. The density of fire count data was also calculated use a heat map. Land Surface Temperature (LST) and DEM layer data were processed for each grid cell by measuring the distance between the center of each grid cell and the respective data points at a distance of 1 km to maintain consistency with the specified spatial resolution. The present study utilized a Python project Jupyter notebook for modeling and preprocessing tasks, while ArcGIS 10.8 was employed for analysis and mapping purposes.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
Date	ID_Station	Name Station	Lat	Long	Month	DEM	LST	MCD64A1	Fire count	Wind_Dir	Wind_SD	Temp	Rel_hum	Precipitation	Pressure	NO2	CO	O3	AAI	Total_PM25	MCD19A2	550nm_MERRA2	PM25_G
1/1/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.690252	0.696542	0.418605	0.8125219	111.5	0.88	22.52917	74.64583	0	982.113	1.694	3.782	0.104	0.83	1.83E-05	0.3118293	0.155333333	28
1/2/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.690252	0.547404	0	0.1687926	70.69	1.16	21.08696	74.71667	0	982.979	1.4235		0.109	0.14	1.20E-05	0.2258415	0.095291667	24
1/3/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.690252	0.544622	0	0	75.12	1.24	20.29583	77.15833	0	983.192	1.274		0.11	0.51	1.51E-05	0.2491951	0.176125	24
1/4/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.690252	0.574025	0	0	76.44	1.05	22.90833	75.2375	0	982.221	1.682		0.112	0.29	1.48E-05	0.2749024	0.15625	22
1/5/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.690252	0.664223	0	0	64.81	1.2	23.43478	77.17917	0	981.721	1.667	3.549	0.113	0.96	1.17E-05	0.1824878	0.107041667	21
1/6/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.690252	0.384237	0	0	49.25	1.22	22.50455	69.28333	0	982.883	1.35	3.547	0.108	0.01	1.41E-05	0.1614756	0.116166667	19
1/7/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.690252	0.401188	0	0	45.94	1.1	21.05	62.71667	0	981.863	1.372	3.518	0.106	0.17	1.73E-05	0.1632805	0.117666667	21
1/8/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.690252	0.441052	0	0	48.5	1.22	20.2	72.21667	0	98.8	1.45	3.69	0.107	0.3	1.75E-05	0.186439	0.108666667	30
1/9/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.690252	0.465289	0	0	47	1.1	21.7913	70.25417	0	98.84	1.413	3.695	0.109	0.58	1.75E-05	0.2158415	0.133	30
1/10/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.690252	0.608594	0	0.2987811	75.38	0.88	23.15417	69.73333	0	978.842	1.767		0.111	0.24	1.43E-05	0.298561	0.148916667	25
1/11/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.690252	0.311398	0	0	164.12	0.7	23.90417	73.83333	0.07	976.783	1.691		0.11	0.7	2.56E-05	0.3336829	0.231791667	28
1/12/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.690252	0.383177	0	0	284.56	1.2	23.71739	75.01304	0.05	976.17	1.713	3.596	0.11	0.03	1.55E-05	0.1841341	0.095833333	28
1/13/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.690252	0.540516	0	0	280.81	1.45	22.53182	70.8087	0	977.67	1.648	3.474	0.109	0.27	1.17E-05	0.1443902	0.046458333	25
1/14/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.690252	0.47893	0	0	292.62	1.26	21.16522	70.22917	0	975.679	1.808	3.557	0.109	0.17	1.37E-05	0.1567439	0.0505	22
1/15/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.690252	0.534093	0	1.7095057	322.5	1.18	20.97917	70.36667	0	975.717	1.545		0.108	0.23	2.23E-05	0.1837805	0.099041667	25
1/16/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.690252	0.461713	0	0	155.69	1.08	21.44583	68.8125	0	977.138	1.75	3.639	0.107	0.04	2.61E-05	0.299878	0.146166667	33
1/17/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.690252	0.314311	0	0	130	1.05	21.82917	71.05833	0	978.358	1.58	3.74	0.105	0.41	3.62E-05	0.4075122	0.213916667	34
1/18/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.690252	0.503697	0	0	135.5	1.1	23.7125	72.59167	0	979.933	1.6135	3.771	0.108	0.13	5.90E-05	0.4154268	0.226041667	46
1/19/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.690252	0.58184	0	0	137.62	0.98	23.25417	74.50417	0	979.933	1.466	3.792	0.109	0.31	3.87E-05	0.3461707	0.207458333	43
1/20/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.690252	0.468997	0	0.6168609	130.44	0.99	23.025	72.14167	0	979.78	1.643	3.742	0.11	0.22	2.15E-05	0.3270366	0.140875	37
1/21/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.705336	0.473709	0	0	147.62	0.84	22.80417	72.73333	0	978.679	1.744	3.69	0.11	0.19	2.04E-05	0.2799878	0.109625	35
1/22/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.705336	0.421926	0	0	304.12	0.88	22.94583	69.9625	0	977.479	1.832	3.686	0.108	0	1.63E-05	0.2335366	0.077916667	32
1/23/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.705336	0.560459	0	0	327.5	1.23	22.75833	67.625	0	976.554	1.655	3.64	0.108	0.43	2.27E-05	0.2379878	0.101375	34
1/24/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.705336	0.579134	0	0	101.12	1.11	23.31364	64.85	0	978.396	1.64		0.109	0.36	3.47E-05	0.3586829	0.268916667	36
1/25/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.705336	0.520461	0.476744	0	132.31	1.03	22.875	68.20417	0	978.363	1.775	3.86	0.108	0.35	4.39E-05	0.3761341	0.255	43
1/26/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.705336	0.527348	0	0	230.88	0.7	22.73333	72.49583	0	977.38	1.537	3.745	0.106	0.38	3.42E-05	0.3317073	0.219666667	43
1/27/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.704138	0.336782	0	5.1708984	129.62	1.13	23.6375	73.44167	0	977.679	1.801		0.106	0.26	4.03E-05	0.4323415	0.249041667	48
1/28/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.722068	0.462259	0	0	98.38	1.1	23.73043	70.75652	0	978.413	1.644		0.109	0.56	5.48E-05	0.4721463	0.211791667	49
1/29/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.722068	0.460008	0	0	82.56	1.35	21.52917	77.19167	0	98.58	1.554	3.669	0.106	0.37	5.42E-05	0.4380732	0.206208333	43
1/30/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.722068	0.395379	0	0	80.38	1.12	22.45	67.625	0	98.971	1.641	3.779	0.105	0.06	3.51E-05	0.3443902	0.197208333	38
1/31/2023	35T	Chang Mai Pro	18.83755	98.96832	1	0.722068	0.481595	0	0	193	0.77	22.05217	70.6087	0	977.59	1.867		0.104	0.26	3.24E-05	0.407939	0.192833333	44
2/1/2023	35T	Chang Mai Pro	18.83755	98.96832	2	0.722068	0.507951	0	0	230.38	1.21	23.11739	73.85217	0	975.713	2.17	3.886	0.104	0.28	5.24E-05	0.434439	0.129291667	53
2/2/2023	35T	Chang Mai Pro	18.83755	98.96832	2	0.722068	0.587681	0	0	194.81	1.09	25.2125	71.35	0	976.113	1.926	3.888	0.103	0.23	5.46E-05	0.418061	0.150166667	54
2/3/2023	35T	Chang Mai Pro	18.83755	98.96832	2	0.722068	0.653772	0	0	159.19	0.59	25.94783	69.89167	0	976.46	1.54	3.984	0.105	0.64	4.37E-05	0.4748659	0.132875	63
2/4/2023	35T	Chang Mai Pro	18.83755	98.96832	2	0.722068	0.645295	0	0	226.19	0.91	27.5087	67.30833	0	974.533	1.706	4.027	0.109	0.19	5.77E-05	0.5644634	0.107583333	68
2/5/2023	35T	Chang Mai Pro	18.83755	98.96832	2	0.719491	0.58082	0.476744	0	252.38	1.27	27.71739	67.07083	0	973.942	1.748	3.844	0.109	0.29	3.50E-05	0.3297317	0.165583333	56
2/6/2023	35T	Chang Mai Pro	18.83755	98.96832	2	0.736011	0.36209	0	0	258.38	1.19	27.78333	60.6125	0	974.158	1.835	3.542	0.106	0.27	2.27E-05	0.1467317	0.1185	31

Figure 10 For example, the independent variable (Train dataset) and the dependent variable (Test data set) in CSV file format for the estimated PM2.5 concentration the on-site scale.

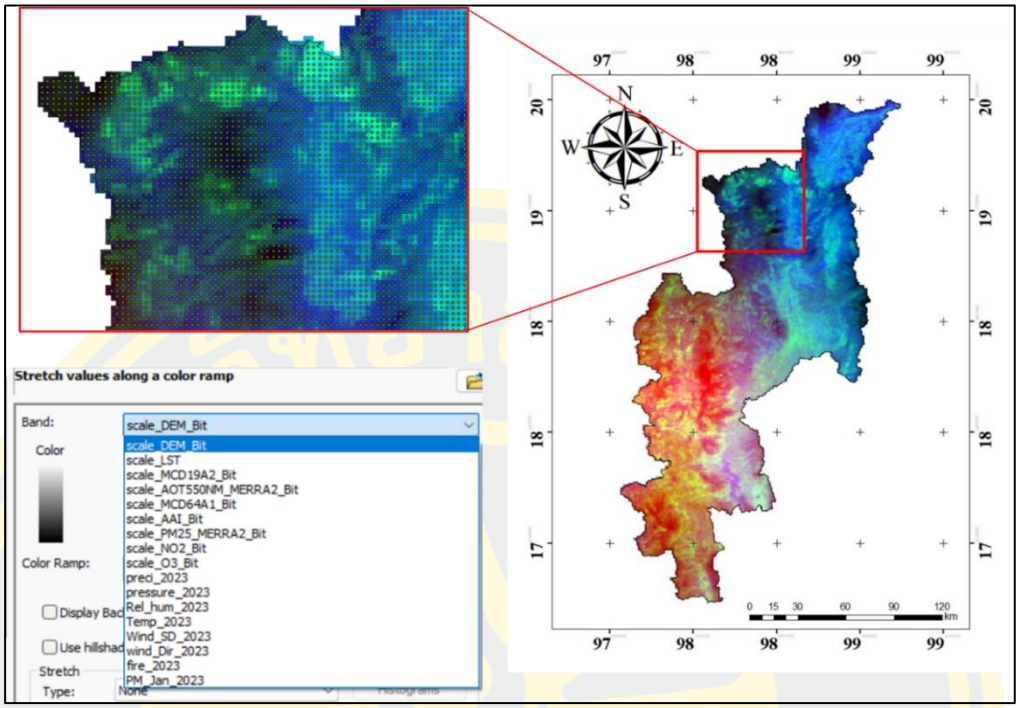


Figure 11 For example, the independent variable (Train dataset) and the dependent variable (Test data set) to the estimated PM2.5 concentration on regionals scale and create a spatial distribution maps with at 1 km resolution in the raster format.

3.4 Workflow of Research

This study was conducted as follows (see Fig.12). The procedures were implemented to achieve an optimal model for predicting ground-level PM2.5 concentrations. The Random Forest (RF) model was employed to assess the significance of features by considering all variables potentially affecting surface PM2.5. This preliminary step involved fitting the model and determining the contribution of each variable. Subsequently, the feature scores were rearranged in descending order, as illustrated in Figure 17. Including to the dataset was separated into two groups - the first group, was prepared as the dataset for modeling, while the subsequent group served as the dataset for prediction purposes. And the data set for modeling was will split into training set and testing set with proportions of 70% and 30%, respectively, by the training set was employed for both training and validating the model, and the testing set was employed to validate the model.

Initially, data MCD19A2 (AOD-MAIAC) from MODIS, Total AOT 550nm and Dust mass PM2.5 from MERRA-2, Chemical composition pollution data of Sentinel-5P TROPOMI, Meteorological data, and Auxiliary datasets such as fire count data, LST,

and DEM datasets were acquired via from the GEE platform and by pre-processing from ArcGIS 10.8 software. These datasets were combined with ground data (PM2.5 concentration data) to pair serve as independent and dependent variables for the model (see Fig.10). Total of 1,180 valid records, comprising dependent variables (PM2.5 concentrations) and independent variables, were utilized to train RF, XGBoost, and CNN models all of this is regression to be used to evaluate PM2.5 concentration at the on-site scale.

Secondly, PM2.5 concentration data obtained from the first step model evaluation, will be combined with PM2.5 concentration data from ground stations at the on-site scale to were processed dataset and create raster format layers representing both the independent and dependent variables across each grid cell with a spatial resolution of 1 km × 1 km (except grid cells at the six measurement stations). This processing is conducted using ArcGIS 10.8 to produce train and test data, will be used to feed the model to predict PM2.5 concentrations on regionals scale throughout the study area to create spatial distribution maps with a 1 km resolution raster format throughout Chiang Mai province, Thailand, in 2023 according to month and difference seasonals such as winter, summer, and rainy (from Jan 01 to Jun 31, 2023).

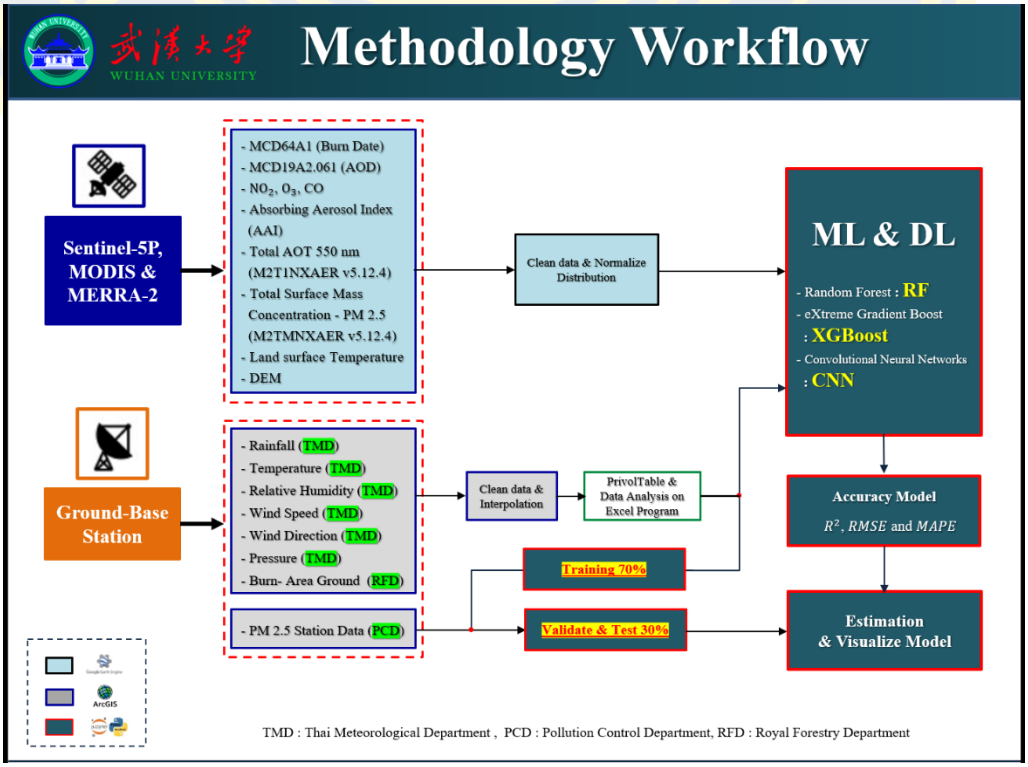


Figure 12 Flowchart of Methodology

3.5 General statistics model

3.5.1 Statistics of parameters used model experiments

Analyzing the statistical parameters employed within the model is of paramount importance, because it must be adjusted to a pattern in which the model can work effectively. The model incorporates diverse factors that have the potential to influence the fluctuations in PM2.5 concentration values, thereby contributing to its refinement. Factors such as Land Surface Temperature (LST), Digital Elevation Model (DEM), and six meteorological variables have been integrated into the model based on previous research indicating an inverse relationship between temperature, relative humidity, and wind speed with PM2.5 concentration levels (Sritong-aon et al., 2021). In addition, parameters such as fire detection data (MCD64A1 and Fire count data) have been included due to the potential impact of severity wildfire on PM2.5 concentration. It is well-established that severe combustion events emitted significant amounts of dust, particles and aerosols into the atmosphere. The model incorporates data concerning aerosols and the chemical composition of daily air pollution, as these substances are linked to PM2.5 concentration fluctuations. For instance, elevated O3 levels during summer, a potent oxidative pollutant, can stimulate the production of secondary particles and elevate PM2.5 levels. Conversely, during winter, high PM2.5 concentrations tend to obstruct solar radiation, resulting in a decrease in O3 concentration and production (Jia et al., 2017). Additionally, the chemical reaction of NO2 in the atmosphere fosters the generation of secondary PM2.5 (Balamurugan et al., 2022; J. Wu et al., 2016).

Descriptive statistics, including mean, median, mode, standard deviation, standard error, range, minimum, and maximum values of the input independent variables used in model regression analysis. Considering that both PM2.5 values and independent variables show vary ranges between minimum and maximum values, normalizing statistical parameters becomes crucial for evaluation to ensure equilibrium. Therefore, the assessment of estimation accuracy relies on metrics such as the coefficient of determination (R^2), Root-Mean-Square Error (RMSE), and Mean Absolute Percentage Error (MAPE), all of which are scaled to a range between 0 and 1 using the normal distribution method. R^2 indicated the explained variance of the model, providing into

the robust of fit a regression model. Conversely, RMSE quantifies the average error magnitude between predicted and observed values, offering a comprehensive of prediction deviations from actual data. the RMSE value are expressed in the same units as the original data, facilitating direct interpret in the context of air quality assessments. As the same time, MAPE assesses the accuracy of prediction models by computing the average percentage difference between actual and predicted values, akin to RMSE. And details of statistical described in Table 5.

Table 5 Descriptive statistics of parameters used model experiments

Descriptive statistics	PM2.5 ($\mu g/m^{-3}$)	DEM	LST	MCD 64A1	Fire count	WD	WS	TEMP	RH	PRE	PRS	NO2	CO	O3	AAI	MCD 19A2	Dust PM2.5 ($\mu g/m^{-3}$)	AOT 550nm
Mean	55.60	828.43	28.29	1.38	0.32	205.50	1.15	27.23	68.02	2.29	803.47	1.56	2.63	0.11	-0.16	0.49	0.06	0.38
Median	38.30	812.60	28.12	0	0	226.93	1.08	27.7	68.90	0	971.40	1.54	3.57	0.11	-0.17	0.43	0.04	0.26
Mode	0	-	-	0	0	262.31	1.05	24.22	66.35	0	0	1.61	0	0.12	-0.12	0.14	0.01	0.10
Standard deviation	53.94	352.50	3.19	8.14	0.95	69.11	0.43	3.71	9.75	5.28	366.01	0.26	1.76	0.01	0.86	0.30	0.05	0.30
Standard error	1.57	12.80	0.11	0.29	0.03	2.01	0.01	0.10	0.28	0.15	10.65	0.007	0.05	0	0.02	0	0.01	0
Range	378.45	2153.22	22.23	84	13.13	289.5	2.94	17.8	53.61	45.73	987.98	2.41	4.87	0.13	5.4	1.68	0.27	1.54
Minimum	2.21	236	16.37	0	0	38	0.48	17.39	42.82	0	0	0.23	0	0	-2.52	0.08	0	0.04
Maximum	378.45	2389.22	38.61	84	13.13	327.5	3.42	35.19	96.44	45.73	987.98	2.64	4.87	0.13	2.90	1.77	0.27	1.58

3.5.2 The correlation coefficients between PM2.5 concentrations and independent variables

The correlation analysis was conducted to examine the relationship between PM2.5 concentrations and various geographical, satellite, and meteorological variables. Geographical variables included land surface temperature (LST), digital elevation model (DEM), and fire count data, while satellite variables comprised NO2, O3, CO, absorbing aerosol index depth (AAI), aerosol optical depth (AOD), Dust surface mass concentration PM2.5 (DUSMASS25), and total aerosol extinction AOT 550 nm (TOTEXTTAU). Meteorological variables encompassed precipitation (PRE), air pressure (PRS), relative humidity (RH), temperature (TEMP), wind direction (WD), and wind speed (WS).

PM2.5 concentration exhibited significant negative correlations with five meteorological factors: PRE, PRS, RH, TEM, and WD, while showing a positive correlation with WS. This indicates that the concentration of PM2.5 will decrease if there is a lot of humidity in the air and wind is an important variable in the increase and decrease of PM2.5 values, showing that changes in meteorological variables have a significant impact on the concentration of PM2.5.

Regarding satellite-derived air quality data (NO2, CO, AAI, AOD, DUSMASS25, and TOTEXTTAU), demonstrated positive associations with PM2.5 concentration (except O3). Additionally, a substantial positive correlation between the frequency of fires (Fire count data) and burn date, as derived from the MODIS product (MCD64A1), indicates increased greenhouse gas emissions and temperatures from intense wildfires in the area. This emphasizes that fires contribute to higher PM2.5 levels. Furthermore, significant negative correlations are observed with variables such as month, DEM, and LST. Consequently, the experiments demonstrate high accuracy in estimating PM2.5 levels, contingent upon the inclusion of these factors. Nevertheless, it remains imperative to thoroughly evaluate the individual impact of each factor on model accuracy. Thus, employing the Pearson correlation coefficient (r) and p-value (refer to Figure 13) is crucial for scrutinizing these associations meticulously. This examination is reiterated through multiple iterations to ensure robustness and reliability.

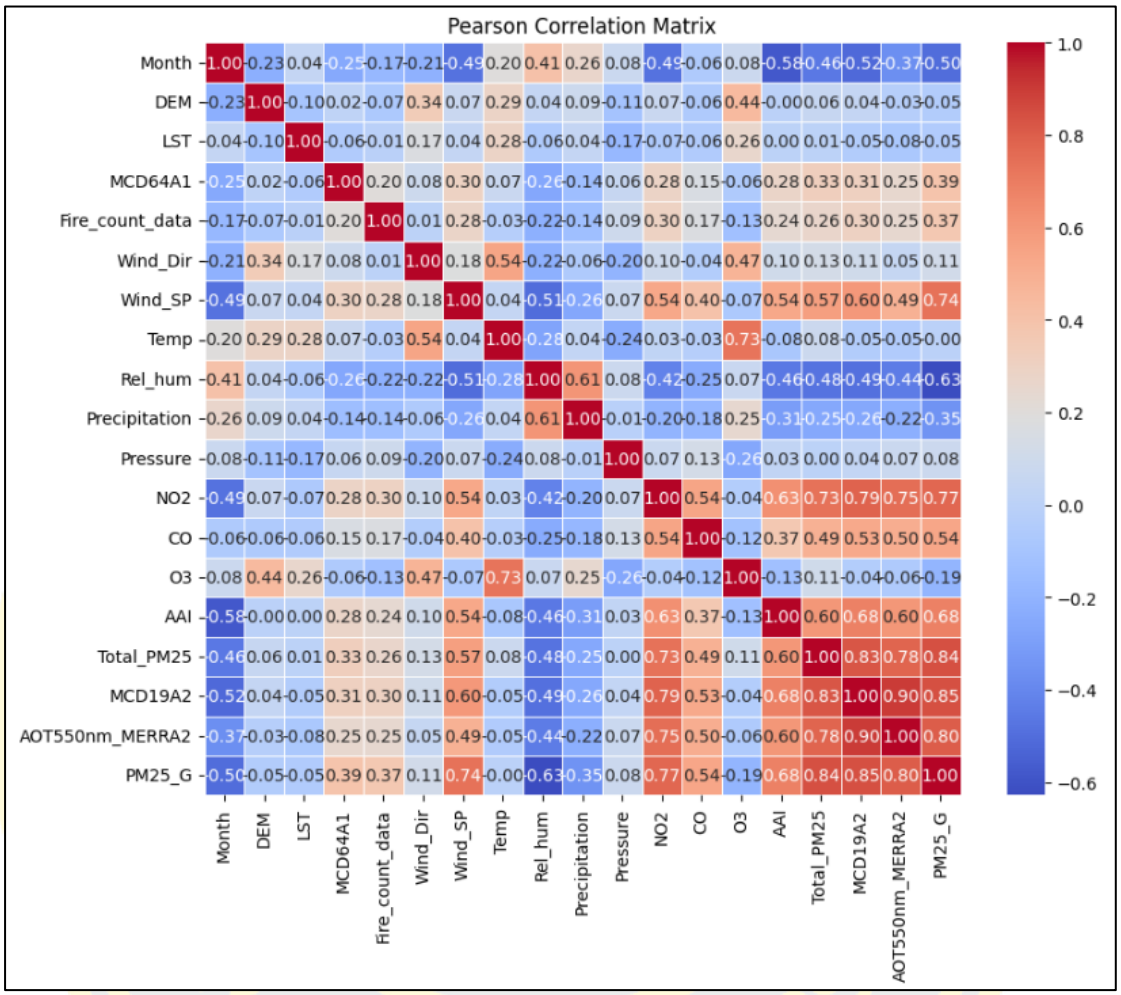


Figure 13 Correlation coefficient matrix between PM2.5 observations and various independent variables.

3.6 Prediction Model

Machine Learning and Deep Learning wase used in the analysis to establish model estimate PM 2.5 concentration impacted by wildfire. In this research, based on previous research both domestically and abroad that is widely. In this study, using 3 algorithms that are guaranteed to be effective in creating models for estimating particulate matter PM2.5 concentrations and generated spatial distribution map. The most important aspect of creating a model for this research is Selecting a methods called GridSearchCV from 'sklearn.model selection' for optimization Model performance and to find the best Hyperparameters.

3.6.1 Random Forest Regression

To provide accurate and fast estimations, machine learning is frequently used in atmospheric research (Li, Y. et al., 2023). This study employs the Random Forest Regression (RFR) machine learning method. The RFR model is a regression model that uses the Random Forest algorithm. Since it can predict continuous functions with numerical values, requires less training, is resistant to overfitting, and can model nonlinear relationships, the RFR method is superior to conventional machine learning methods (Chang et al., 2023; da Silva Chagas et al., 2016). The RFR algorithm uses numerous regression trees, each representing a set of conditions or hierarchical constraints, and then all decision trees are combined to obtain a final prediction (Rodriguez-Galiano et al., 2014; You et al., 2016). Eq. (1) presents the general RFR formula, where $\widehat{f_{RF}^{K}}(\boldsymbol{X})$ denotes the RFR predictors, X denotes the input vector**,** K denotes the number of decision trees, and $\{\boldsymbol{T}(\boldsymbol{X})\}_{1}^{K}$ denotes the constructed decision trees (Rodriguez-Galiano et al., 2014).

$$\widehat{f_{RF}^{K}}(X) = \frac{1}{K}\sum_{k=1}^{K} T(X) \qquad (3-1)$$

In performing Random Forest Regression (RFR), 70% of the data are used as the training sample, while 30% of the data are used for validation. This RFR processing utilizes computing with Python project Jupyter notebook.

Table 6 Hyperparameter tuning of RF algorithms

Parameter	Definition	Tuning
Booststrap	method for sampling data points (with or without replacement)	True
Max_depth	max number of levels in each decision tree	10
Max_features	max number of features considered for splitting a node	auto
Min_samples_leaf	min number of data points allowed in a leaf node	1
Min_samples_split	min number of data points placed in a node before the node is split	2
Min_weight_fraction_leaf	min weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node.	0
n_estimators	number of trees in the forest	550
random_state	Seed of the random number generator used to initialize the random state.	70

3.6.2 eXtreme Gradient Boosting

The eXtreme Gradient Boosting originates from a fusion of gradient descent and boosting, known as Gradient Boosting Machine (GBM). Boosting, an ensemble-learning algorithm, assigns varying weights to the training data distribution at each iteration. In each boosting iteration, weight is added to misclassified error samples and subtracted from correctly classified samples, thereby effectively altering the training data distribution (Bisri et al., 2015). GBM employs second-order gradient statistics to minimize the regularized objectives outlined in equation (2).

$$\mathcal{L}(\varphi) = \sum_{i} l\left(\hat{y}_i, y_i\right) + \sum_{k} \Omega(\mathrm{f_k})$$

where $\Omega(\mathrm{f}) = \gamma \mathrm{T} + \frac{1}{2}\lambda \|w\|^2$ $(3-2)$

The function $\boldsymbol{l}$ represents a differentiable convex loss function, which quantifies the disparity between the prediction $\boldsymbol{yi}$ and the target $\boldsymbol{yi}$, while $\boldsymbol{\Omega}$ penalizes the model's complexity (T. Chen et al., 2016).

To find the optimal parameters created a hyperparameters grid consisting of different hyperparameter combinations. Atmospheric data (AOD, AAI, NO2, CO, and O3) meteorological conditions (WS, RH, PRE, TEMP, and PRS) and other Auxiliary data (LST, Fire count data, DEM) which are identified as useful parameters in Estimating PM2.5, were also included as covariates.

Table 7 Hyperparameter tuning of XGBoost algorithms

Parameter	Definition	Tuning
nrounds	Number of trees to be used in the model	300
eta (η)	Learning rate	0.1
gamma (γ)	Minimum spilt loss reduction	3
Max_depth	Maximum depth of trees	9
Minimum_Child_weight	Minimum observation for a child	1
subsamble	Ratio of random samples to be considered for training	1
Column_sample_tree	Ratio of features (columns) to be used to train each tree	0.8
Column_sample_level	Ratio of features to be used to train each split of the tree	0.8
Lambda (λ)	Regularization term on the weights for the $\boldsymbol{L^2}$norm (Eq.2)	2
Alpha (α)	Regularization term on the weights for the $\boldsymbol{L^1}$ norm (Eq.2)	60

And combined all PM 2.5 ground base data from 6 station. To better capture the seasonality patterns. These factors were randomly divided into 70% and 30% splits for a training set and a testing set. Table 6 presents a comprehensive summary of the ten pivotal parameters of XGBoost, encompassing their respective ranges and tuning values.

3.6.3 Convolutional Neural Networks

The CNN is one of the Deep Neural Networks that can recognize and extract, possesses the capability to identify and extract specific attributes from images, making it extensively employed in visual image analysis. Its architecture comprises two primary components:

(1) The convolution tool for feature extraction, which discerns and segregates image features.

(2) The connected layer for prediction, utilizing the convolutional output to make predictions based on the identified features. The analysis involved constructing a sequential model using a linear stack of layers. This architecture comprised two 2D convolution layers followed by max pooling pairs, culminating in a flatten layer typically serving as a link between convolution and dense layers. The initial layer served as the input image, followed by a max pooling operation in the second layer, which selects the maximum pixel value within a unit's receptive field, effectively subsampling the image. Similarly, the third layer conducted convolution operations with filters, succeeded by a fourth pooling layer. A pivotal aspect of the CNN model is the activation function, employed to capture relationships within the architecture; here, the Rectified Linear Unit (ReLU) activation function was utilized. Subsequently, a flatten layer converted the input image data into a one-dimensional array, leading to the last dense output layer. The data were divided into training and testing sets, with 70% and 30% splits, respectively, along with corresponding PM 2.5 ground-based data allocated for validation purposes.

Table 8 parameter of CNN algorithms by Keras using Sequential API

Layer (type)	Output Shape	Param
Conv1D	0, 11, 64	256
Max_pooling1D	0, 5, 64	0
Conv2D	0, 3, 64	12352
Max_pooling2D	0, 1, 64	0
Flatten	0, 64	0
Dense_1	0, 128	8320
Dense_2	0, 1	129

3.7 Model Assessment Accuracy

The data sets, factors, and parameters were split into a 70% training dataset and a 30% validation dataset by the GridSearchCV functionality in scikit-learn is employed to identify the optimal parameter values for a model through a 5-fold cross-validation method. This method partitions the data into 5 subsets, using 4 subsets for training and 1 subset for testing in each iteration. This process is repeated 5 times, ensuring that each subset is used exactly once for validation. This thorough testing ensures the model's robustness across all of variations split data. Optimizing parameters using GridSearchCV significantly enhances the model's performance and flexibility. Well-tuned parameters enable the model to perform at its best and adapt effectively to new data along with evaluation metrics such as R^2, RMSE, and MAPE, were employed to assess the robustness of all models.

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \qquad (3-3)$$

$$RMSE = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (3-4)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|^2 \qquad (3-5)$$

Where $\boldsymbol{y_i}$ represents the PM2.5 observations, $\boldsymbol{\hat{y}_i}$ stands for the predicted value, $\boldsymbol{\bar{y}}$ denotes the average of the PM2.5 observations, and $\boldsymbol{n}$ indicates the overall sample size.

3.8 Summary of this Chapter

The study area of Chiang Mai Province, Thailand, providing detailed information on its geographical coordinates, topography, climate, and annual rainfall. It highlights the environmental challenges posed by forest fires, exacerbated by seasonal variations and climate change, leading to increased air pollution affecting public health and the economy, and tourism.

Estimation PM2.5 concentration, using remote sensing data, including Sentinel-5P TROPOMI and MODIS, for analyzing and estimating PM2.5 concentrations. It also covers the acquisition of meteorological data, ground station data, and fire count data, emphasizing the comprehensive approach to data collection for modeling purposes.

Preprocessing steps, focusing on standardizing data formats, resampling, and cleaning procedures to ensure consistency and reliability. The workflow of the research is delineated, detailing the methodology employed for model development, including Random Forest Regression (RFR), eXtreme Gradient Boosting (XGBoost), and Convolutional Neural Networks (CNN). The model assessment accuracy evaluation metrics used, such as R^2, RMSE, and MAPE, along with the 5-fold cross-validation technique to validate the models' performance. Additionally, statistical parameters and correlation coefficients between PM2.5 concentrations and independent variables are analyzed to understand the relationship between various factors affecting air quality.

Finally, the prediction model on the machine learning and deep learning algorithms utilized, including Random Forest Regression (RFR), eXtreme Gradient Boosting (XGBoost), and Convolutional Neural Networks (CNN). It provides insights into the architecture and parameters of each model and emphasizes the importance of accuracy assessment for reliable predictions.

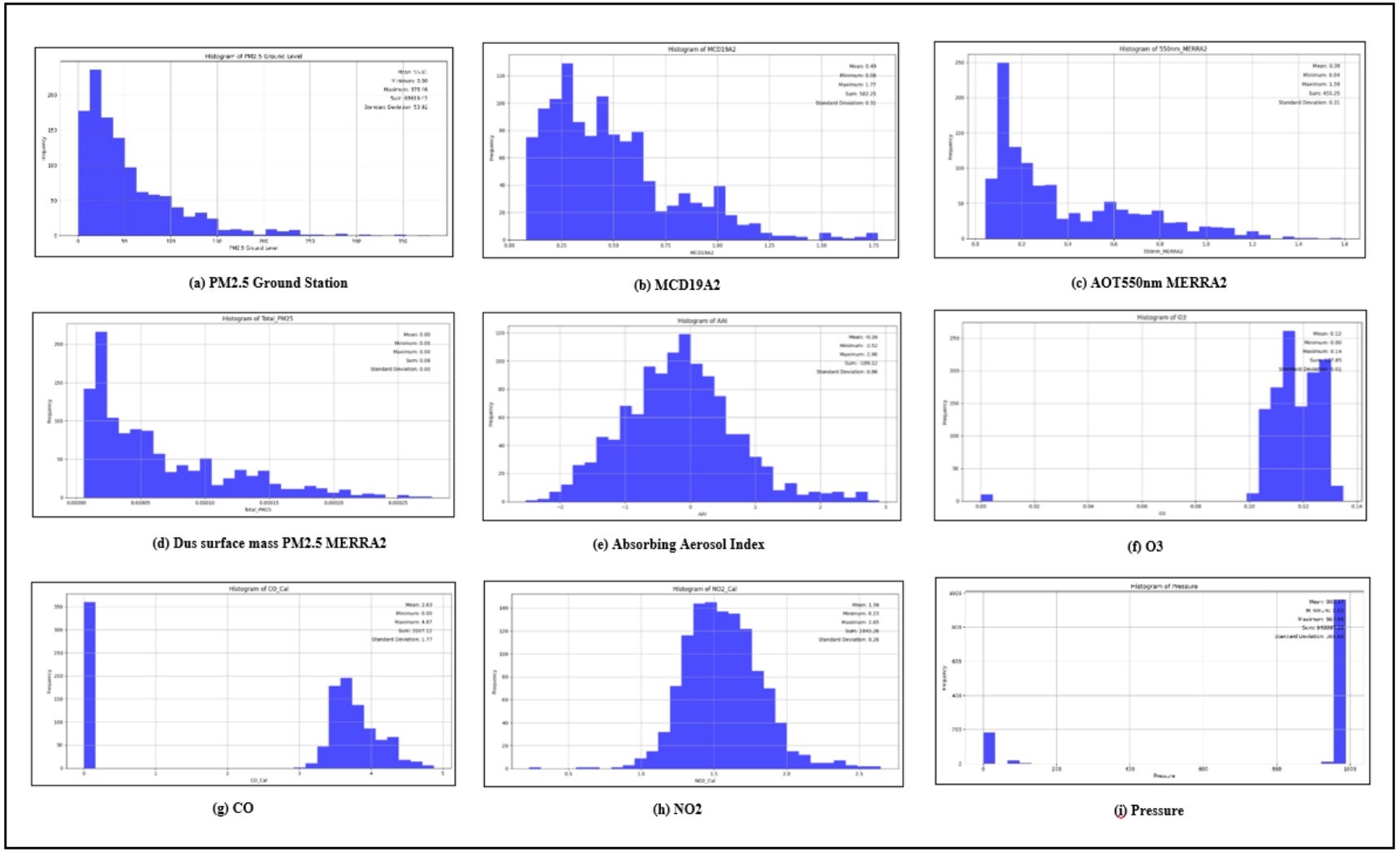
CHAPTER 4
RESULTS AND VALIDATION

In this pivotal chapter, the provided details a comprehensive study on delve into the outcomes of PM2.5 concentration estimation in period of the severe wildfires in Chiang Mai Province, Thailand, using data from six monitoring stations. Including to the experiment and the result. And three machine learning models, including Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Convolutional Neural Network (CNN), were compared for their performance. The RF model outperformed the others, demonstrating the highest R-squared (R^2) of 0.89, the lowest Root Mean Square Error (RMSE) of 11.61 μg/ m^3, and the lowest Mean Absolute Percentage Error (MAPE) of 34.22 μg/ m^3. Additionally, the RF model was utilized to create spatio-temporal distribution maps for PM2.5 concentration, showing seasonal variations and highlighting the impact of factors like agricultural burning and forest fires. The study emphasizes the importance of accurate PM2.5 estimation for understanding air quality dynamics and estimation of the distribution of PM2.5 concentrations in regions without ground measurement stations, suggesting avenues for further research to enhance understanding of PM2.5 pollution in Chiang Mai Province.

4.1 Statistics of Variables

Total data with 1,180 matched samples for 6 ground monitoring stations. The cover 6-month study period of Chiang Mai Province, Thailand, in 2023 was plotted in a histogram with basic statistics including the minimum, maximum, sum, mean, and standard deviation for each variable (Fig. 14). The format of the daily PM2.5 concentration curve data is similar to that of AOD-MAIAC, the MCD19A2 product from MODIS (Wei et al., 2019), and is also similar to the Total Aerosol Extinction AOT 550 nm data. The Dust surface mass concentration PM < 2.5 μm obtained from the MERRA2 reanalysis model. This is something that indicates that Aerosol data from satellite data used in this study. It has a good relationship with the estimation of PM2.5 concentration significantly and greatly contributes to the accuracy of the

model. This is significantly consistent with the results of our Pearson correlation coefficients matrix analysis. The measured mean PM2.5 concentration was 55.60 µg/ m^3 by averaging all ground stations in the study area. The lowest PM2.5 concentration was 2.21 µg/ m^3 in the rainy seasonals and the highest value was 378.45 µg/ m^3 (according to Table 6 and Fig. 14a) in the winter and dry seasonals that had an effect from wildfires that often occur every year and including to the problem of smog wildfires from neighboring countries such as Burma, the average AOD-MAIAC value was 0.49, the lowest value was 0.08, and the highest value was 1.77 respectively, throughout the study period from data extracted from all 6 measurement stations. The average Total AOT550nm value is 0.38, the lowest value is 0.04, and the highest value is 1.58, respectively, which is very similar to the AOD-MAIAC value. And in terms of Dust surface mass concentration PM < 2.5 µm, the average value is 64 µg/m^3, the lowest value is 6 µg/m^3 and the highest value is 277 µg/m^3, respectively, which is very similar to the PM2.5 values obtained from ground station data. Moreover, it can be seen that the curve of carbon monoxide (CO) (in Fig.14g) with the curve of Fire Count Data (Burn scar) are ground survey data received from the Royal Forestry Department (RFD) and the curve of Burn Date analysis data from the MCD64A1 product obtained from MODIS (Fig. 14(o-p)) are very consistent. That means wildfires contributing very important role to the release of air pollutants and particulate matter PM2.5 into the atmosphere. This is consistent with PM2.5 data, both ground station data and remote sensing data.



(a) PM2.5 Ground Station

(b) MCD19A2

(c) AOT550nm MERRA2

(d) Dus surface mass PM2.5 MERRA2

(e) Absorbing Aerosol Index

(f) O3

(g) CO

(h) NO2

(i) Pressure

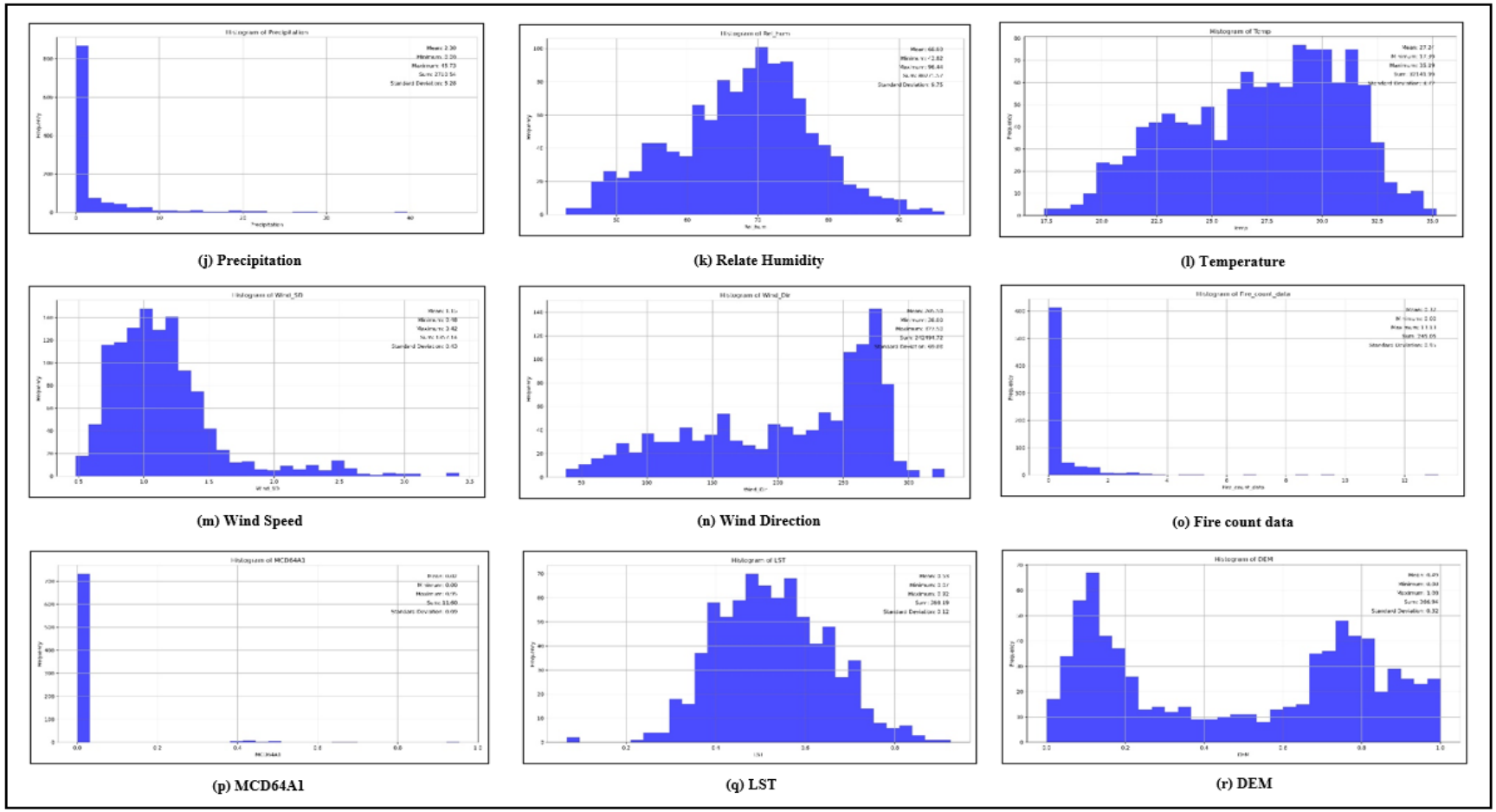


Figure 14 Shows histograms and descriptive statistics (minimun, maximum, sum, mean, and standard deviation) for PM2.5 concentration ground station and independent variables used for modeling. Data are six month over Chiang Mai province, Thailand. The number of samples is 1,180.

4.2 Performance of the Model

4.2.1 On-site scale Model performance

The accuracy of the verification results for PM2.5 concentration estimates in the study area of Chiang Mai Province, Thailand, is assessed through the analysis of data were combined from all 6 stations distributed across the study domains (refer to Table 3). Three key evaluation indicators are employed. The Random Forest (RF) model demonstrates excellent performance, achieving value of 0.89 of R-squared (R^2), 11.61 of root mean square error (RMSE), and 34.22 of mean absolute percentage error (MAPE) and the eXtreme Gradient Boosting (XGBoost) model shows inferior performance, achieving 0.81 of R-squared (R^2), 12.51 of root mean square error (RMSE), and 37.38 of mean absolute percentage error (MAPE), respectively. And finally, The Convolution Natural Network (CNN) model showed the poorest performance of the three models, is 0.67 of the R-squared (R^2), 14.70 of the Root Mean Square Error (RMSE), and 37.78 of the Mean Absolute Percentage Error. (MAPE), respectively, as shown in Table 8. Comparative analysis against other cutting-edge machine learning algorithms underscores the of the RF model, particularly evident in its superior R^2 and RMSE scores, indicative of a stronger linear relationship between observed and estimated PM2.5 concentrations. The PM2.5 scatter plot shows the relationship and accuracy between the training and testing generated by the machine learning, RF model for each and every station in 2023. Details are shown in Fig. 15

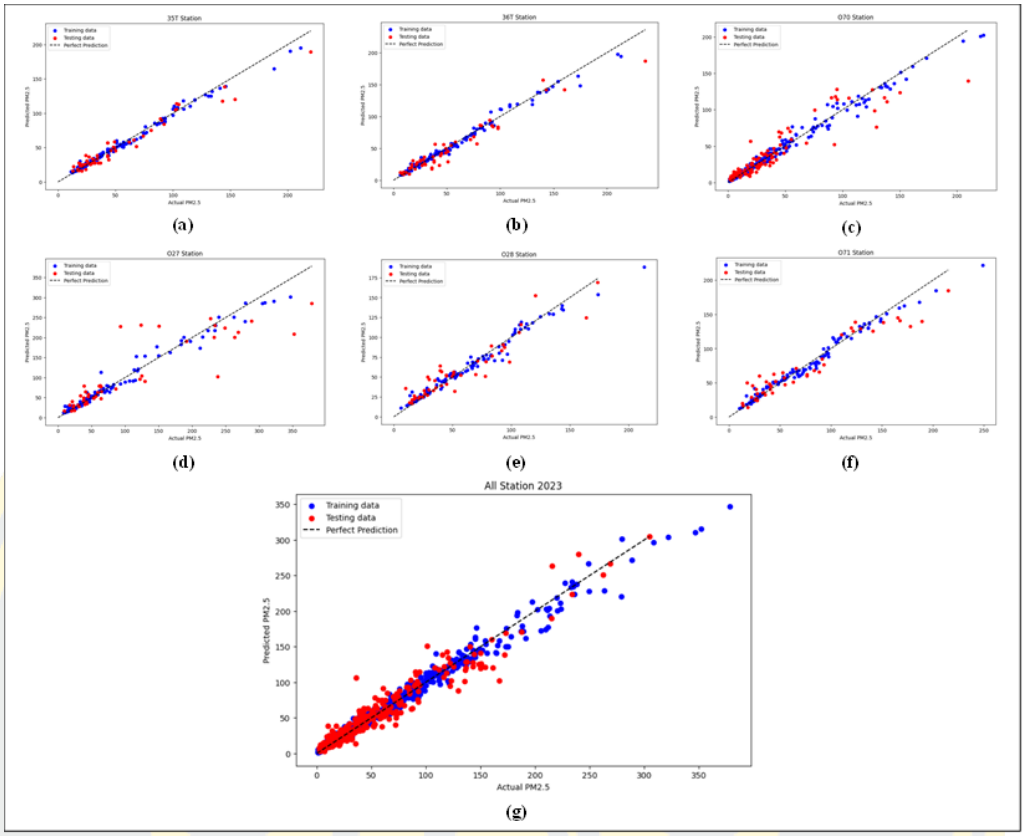


Figure 15 Shows a scatter plot of PM2.5, illustrating the correlation and accuracy between the training and the testing generated by the machine learning model of each station and all station 2023.

Table 9 Comparison of model performance on the testing dataset at on-site scale.

Station	RF			XGBoost			CNN		
	R^2	*RMSE*	*MAPE*	R^2	*RMSE*	*MAPE*	R^2	*RMSE*	*MAPE*
35T	0.91	10.40	15.58	0.71	11.95	14.53	0.82	15.16	16.23
36T	0.88	14.05	18.56	0.86	15.87	19.66	0.58	23.37	22.55
O27	0.81	30.53	25.67	0.73	34.15	29.88	0.49	32.79	31.85
O28	0.74	16.91	24.19	0.67	18.51	27.26	0.47	26.41	27.57
O70	0.81	14.83	36.48	0.80	16.36	39.83	0.61	22.66	38.97
O71	0.78	21.03	20.63	0.77	19.40	21.62	0.36	39.23	24.63
All Station	**0.89**	**11.61**	**34.22**	**0.81**	**12.51**	**37.38**	**0.67**	**14.70**	**37.78**

The comprehensive analysis of prediction performance across various criteria, particularly focusing on elevated PM2.5 concentrations (Fig 16). Notably, the Random Forest (RF) model consistently exhibits superior performance compared to

other machine learning (ML) models. This is evidenced by its attainment of the highest R-squared (R^2) score alongside relatively low Root Mean Square Error (RMSE), as depicted in Figure 16a and b respectively. Moreover, RF demonstrates the lowest Mean Absolute Percentage Error (MAPE) for concentrations surpassing 20 $\mu g/m^3$, as illustrated in Fig 16c. As PM2.5 pollution levels escalate, the performance disparity between RF and other ML models becomes more conspicuous. Furthermore, RF illustrate a remarkable capability in the issue of PM2.5 concentration overestimation, further affirming its superior predictive prowess in this domain.

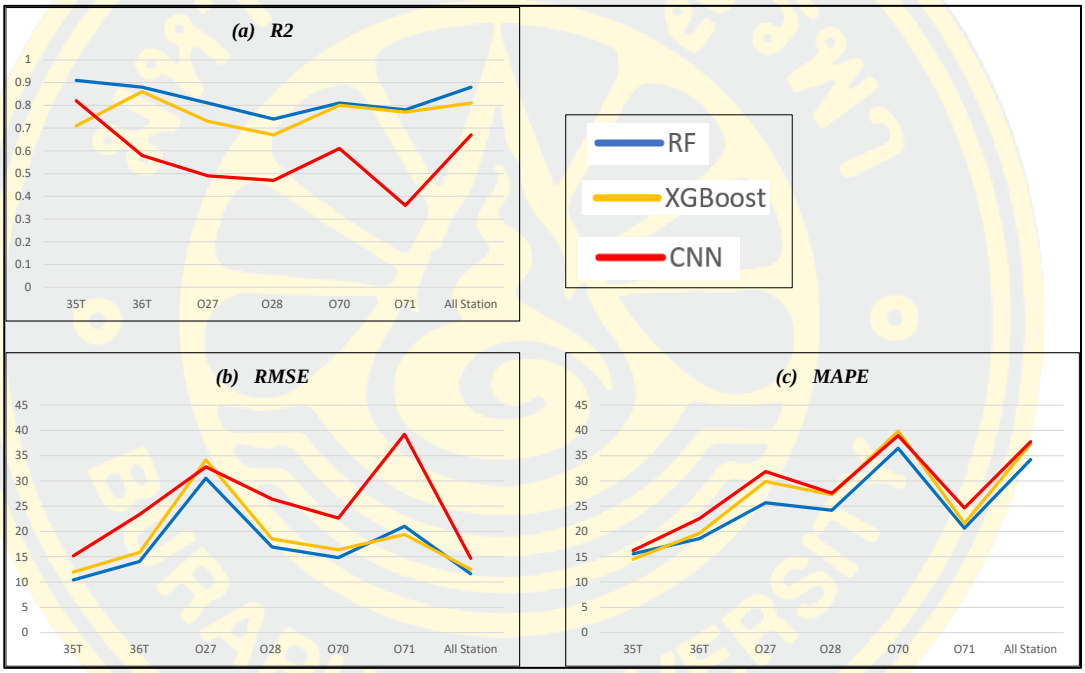


Figure 16 Model comparison across various thresholds of high PM2.5 concentrations: (a) R^2, RMSE, and MAPE

When scrutinizing the evaluation indicators across the six monitoring stations (refer to Table 8), notable discrepancies emerge in the degree of bias, particularly evident in stations O27, O28, O70, and O71. These stations exhibit markedly higher values compared to the two stations (35T and 36T) situated within urban areas. This disparity is reflected RMSE (30.53, 16.91, 14.83, and 21.03 $\mu g/m^3$) and MAPE (25.67, 24.19, 36.48, and 20.63 $\mu g/m^3$) values, respectively, within the RF model. Such disparities can be attributed to the wider range of PM2.5 concentrations observed, exceeding 295 $\mu g/m^3$, notably the maximum value recorded and this results in such drastically different bias values (Son et al., 2023). These elevated concentrations are primarily

attributed to agricultural burning and intense forest fires during the summer, as elucidated by (Punsompong et al., 2021). Such extreme values often lead to substantial errors, either stemming from the model's underestimation or the inaccuracies of certain independent parameters, particularly notable in instances of exceptionally high PM2.5 concentrations. However, it is noteworthy that this considerable deviation is largely attributable to the pronounced concentration resulting from wildfires, aligning with the second objectives and guiding principles of in this study.

4.2.1 Month and Seasonal on regional scale Model performance

In evaluating PM2.5 at the inspection on-site scale has been proved that the RF model has the highest potential in this research's PM2.5 concentration estimation experiment. Therefore, it has been used to predict PM2.5 concentration values to create spatial distribution maps based on time of month and season on regional scale of Chiang Mai Province, Thailand for the year 2023.

Data sets for building prediction model divided according to the time of the month and according to different seasons. It divides each season's sample into training and testing samples. The test accuracy of the monthly model and averaged over the entire study period has R^2, RMSE, and MAPE values as shown in Table 10. The model is based on winter seasonals months (January - February) and dry season months (March - April) models perform well, with R^2 of 0.79 - 0.85, RMSE of 13.51 – 19.31 $\mu g/m^3$, and MAPE of 22.14 - 25.54 $\mu g/m^3$, respectively, which is higher than the model for the rainy season months (May - June) R^2 of 0.74 - 0.81, RMSE of 12.43 - 21.02 $\mu g/m^3$, and MAPE of 20.29 - 20.37 $\mu g/m^3$, respectively, due to air pollution during winter and summer had significantly higher PM2.5 concentration densities than during the rainy season. This is because the study area is located in the tropical monsoon region. In the winter and summer are thin and dry. And during the rainy season, it is caused by precipitation and high relative humidity. Resulting in a decrease in PM2.5 concentration. The test accuracy of the model averaged over the 6-month study period was an R^2 of 0.81, an RMSE of 14.45 $\mu g/m^3$, and MAPE of 21.25 $\mu g/m^3$, respectively.

Table 10 RF model performance on the testing dataset month and difference seasonals on regional scale.

Month	RF		
	R^2	*RMSE*	*MAPE*
Jan	0.85	18.25	23.71
Feb	0.81	16.03	22.14
Mar	0.79	13.51	22.76
Apr	0.80	19.31	25.54
May	0.81	12.43	20.37
Jun	0.74	21.02	20.29
Average periods	**0.81**	**14.45**	**21.25**

4.3 Feature importance of factors influencing variations

The Random Forest (RF) model was used to calculate and assess the significance of features by using all variables with potential influence on PM2.5 concentration at the surface level. After operation, evaluate the contribution of each feature, which was used for analysis throughout the 6-month study period. In summary, the observed positive participation ratios affecting PM evaluations were as follows: aerosol AOD-MAICA variable (MCD19A2), which is the highest MODIS product at 40%, 550 nm aerosols from the MERRA-2 reanalysis at 22%, total surface concentration of PM2.5 from the MERRA-2 reanalysis at 12%, and carbon monoxide (CO) obtained from Sentinel-5P TROPOMI at 11%, respectively. The remaining factors show an inverse contribution ratio to the assessment and although Fire Count Data and Burn Date data from MODIS (MCD64A1) were underestimated for feature importance value in estimating low PM concentrations, but the resulting spatial distribution maps (Fig. 19) can indicate that Fire is the most obvious contributor to pollution. Due to the spatial distribution map the displayed results of model (Fig. 19a-d) and ground data received from the Royal Forestry Department, Chiang Mai Province (Fig. 21a-d), there is consistency during that time significantly. This may be a result of the model being trained in the direction of PM estimation rather than focusing on fire analysis. In addition, PM concentrations may be transported through the air from combustion in the country surrounding and brought in by the wind, this phenomenon may not solely stem from wildfires within the study area. Including to influenced by factors such as

wind direction and wind speed, which affect the distribution of these concentrations. Among geographic features, LST shows a low but still higher correlation than DEM. In terms of chemical composition, AAI ranks lowest among all features. NO2 and O3 show low values and are commonly known as precursors to the secondary formation of PM2.5 (Baker et al., 2007; Tucker, 2000). One contributing factor to carbon monoxide (CO) attaining the highest rank value among chemical elements is as a combustion byproduct arising from the burning, which liberates carbon. This carbon may originate from the combustion of agricultural land or the occurrence of severe wildfires within the study area, as well as wildfires in neighboring countries. These findings align with the hypothesis that emissions of greenhouse gases and wildfires exert a more pronounced influence on air quality fluctuations in Chiang Mai Province, Thailand, compared to industrial greenhouse gas emissions (ChooChuay et al., 2020). Further elucidation of these dynamics is presented in Figure 17.

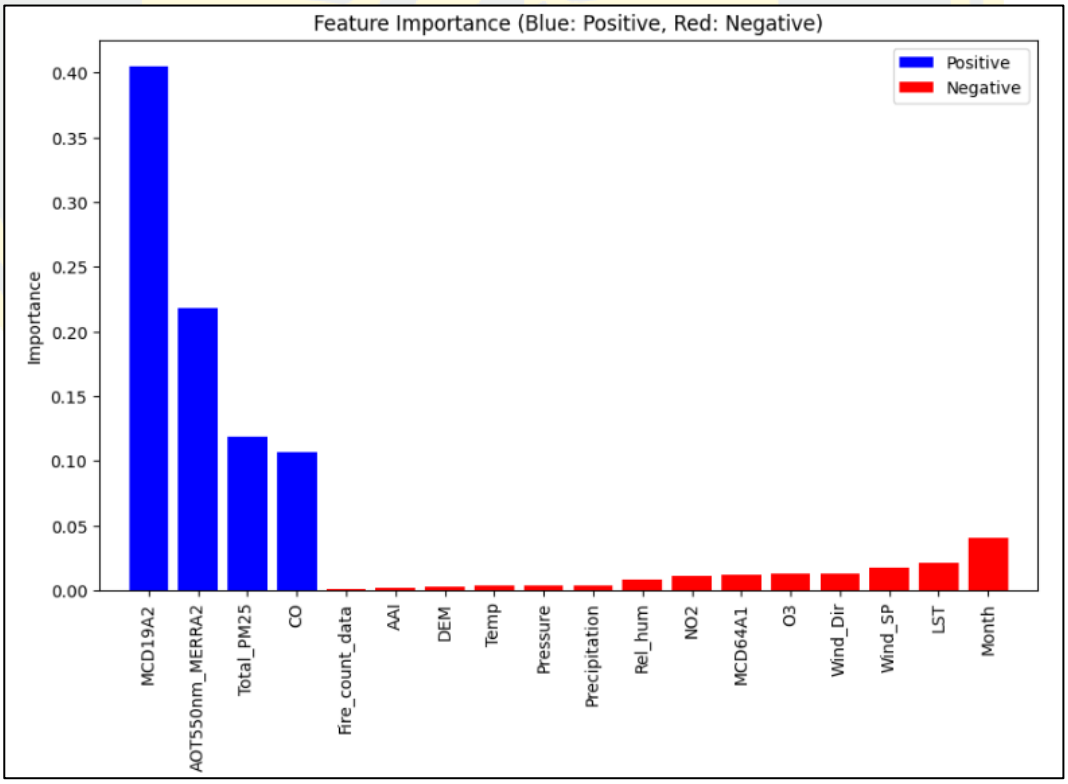


Figure 17 Feature importance for the RF model.

4.4 Spatial-Temporal Distribution pattern

4.4.1 Estimation PM2.5 concentration at On-site scale measurement

Show the spatial distribution of PM2.5 concentration was investigated by computing the average concentrations from January 01 to June 31, 2023, at the six designated sites within the study area. Any missing data were excluded in the calculation process. The results are visualized in Fig. 18a. It is apparent that PM2.5 concentrations exhibited high values in both the north station (O27 station) (ranging from 60 $\mu g/m^3$ to over 90 $\mu g/m^3$) and south station (O71 station) regions of the study area. Conversely, lower PM2.5 concentrations were observed at central stations (ranging around 35 $\mu g/m^3$ or lower) (35T, 36T, O28, and O70 stations) within the study region. Figure 18b represent the spatial distribution of station-level PM2.5 concentrations estimated using our optimized Random Forest (RF) model, leveraging all independent variable data through Javascript extraction on the Google Earth Engine (GEE) cloud platform across the six site of the study area, monitoring stations. The findings reveal that PM2.5 concentrations derived from the assessment exhibit the highest values in the north station and southern station areas of the study region, surpassing 60 $\mu g/m^3$. Conversely, lower PM2.5 concentrations are observed at the central station, ranging approximately between 40 $\mu g/m^3$ to 50 $\mu g/m^3$ or less (Stations 35T, 36T, O28, and O70), within the study region This shows the value Concentrations obtained from observations and concentration values obtained from the evaluation of the model. There was significant consistency at the site level along measurement stations.

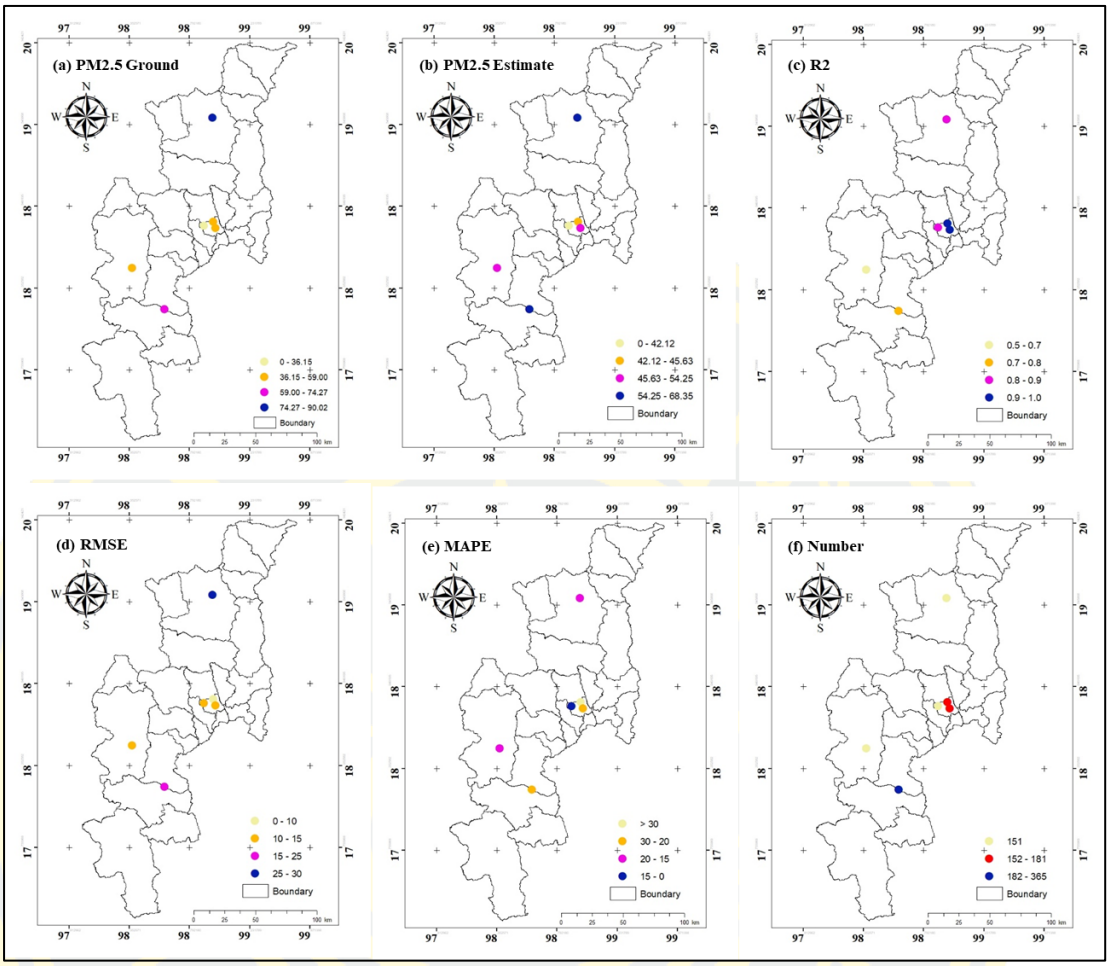


Figure 18 Spatial distribution of (a) PM2.5 ground station and (b) Estimate PM2.5 concentrations (μg/m^3) (c) is R^2 (d) RMSE (e) MAPE and (f) Number of

4.4.2 Estimation and Mapping the spatial distribution of PM2.5 concentrations across different months and seasons on regional scale.

The model was run to map PM2.5 concentrations with a resolution of 1 km, derived from the best model, the RF model, which shows temporal and spatial heterogeneity and integration area projected annual PM2.5 concentrations are spatially consistent with in situ measurements. Spatial coverage of PM2.5 measurements Fig. 19 shows monthly averages of PM2.5 estimates for 2023. The mapping results (Fig. 19a–h) indicate that PM2.5 concentrations are change according to the season winter brings with it the worst pollution. In January and February, the air is dry and thin, including the beginning of setting fires in agricultural areas to clear the area. As a result, PM2.5 concentrations began to reach high levels (>120 μg/m^3), as can be seen in Fig. 19a-b,

which is consistent with the graph plot time series line chart of PM2.5 estimation at the Hkod O71 site in the south zone (Fig. 20c). The concentration increased significantly since January and peaked (>150 μg/m^3) in late February and for most of March. It was found in the north zone of the study area (Fig. 19c–d) because the area is a high mountainous area according to its topography, and there are forest fires that occur in this area every year. Variation in peak periods of air pollution and this is due to seasonal differences in the severity of annual forest fires in the study area, especially in 2023, which will have the most severe wildfires in 10 years, and agricultural activities such as harvesting and burning agricultural (Kanabkaew et al., 2011). Moreover, it is well known that the transport of PM from forest fires in neighboring countries such as Burma has a significant impact on causing smog problems in those areas, which is consistent with data on meteorological factors showing our correlation coefficient that shows the value of wind speed is positive. And that when entering the rainy season at the end of May, PM2.5 concentrations are relatively low in most areas except for some areas in the city clearly (Fig.19e-g). Therefore, in this section, we conclude that PM2.5 in urban areas is not caused by wildfires or agricultural land burning, but rather by the various activities of humans themselves. Figure 19h shows the PM2.5 concentration values throughout the 6-month study period, indicating that concentrations are highest in the southern and northern zones of the study area due to severe wildfires. This is consistent with fire count data from the Royal Forestry Department (RFD) 2023. However, our dataset is limited to only 6 stations, most of which are not evenly distributed throughout the area. This limitation appears to hinder capturing and understanding the underlying mechanisms involved in estimating PM concentration.

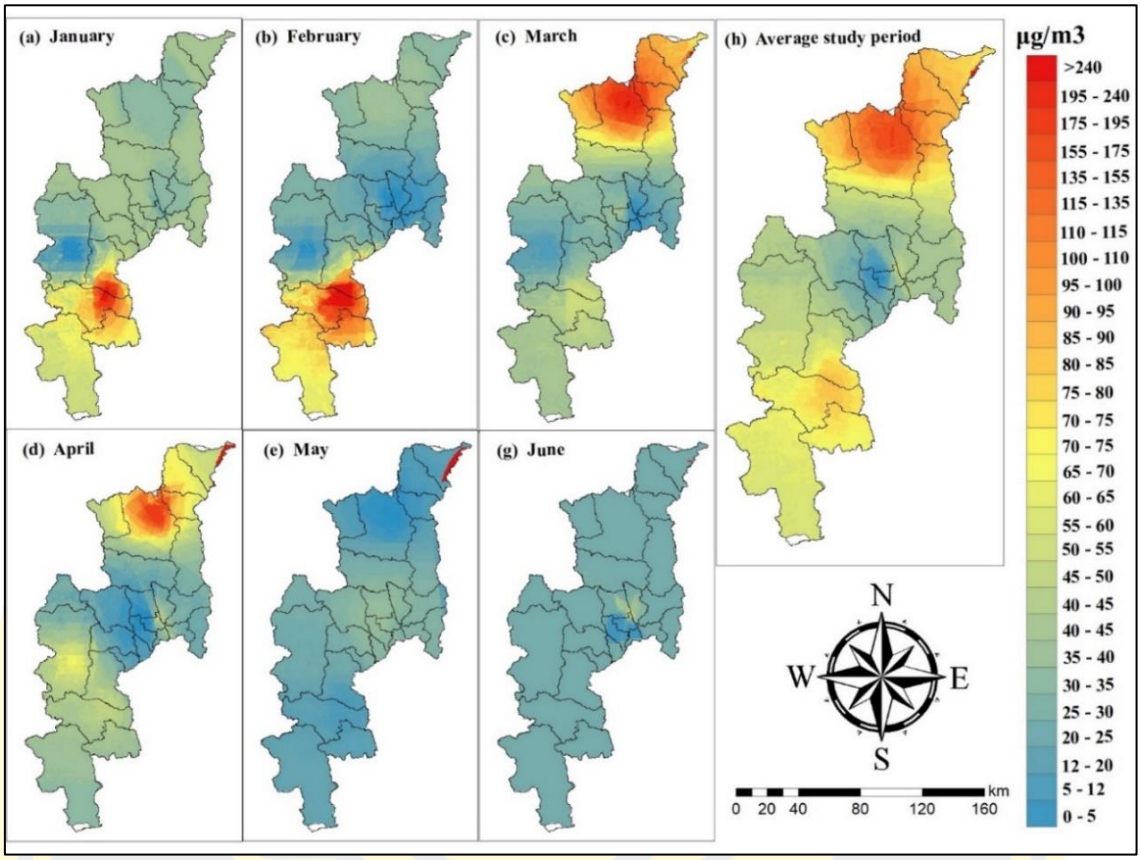


Figure 19 (a–h) Monthly averaged spatial-temporal distributions of the estimated PM2.5 concentrations by the RF Model for the year 2023.

To assess the temporal variation of PM2.5, we compared the daily and month variation of observed PM2.5 values from the six ground stations and PM2.5 estimates from models, across the study area (Fig. 20a-f) and throughout the year (Fig. 20g) shows that the concentration values begin in January and increase and highest in March, this shows the value of our model in estimating PM2.5 concentration values in accordance with observed data. Plotted times series line chart this compare show that Remote Sensing data obtained PM2.5 estimates from models can be applied to estimate PM2.5 concentrations with high efficiency and reliability.

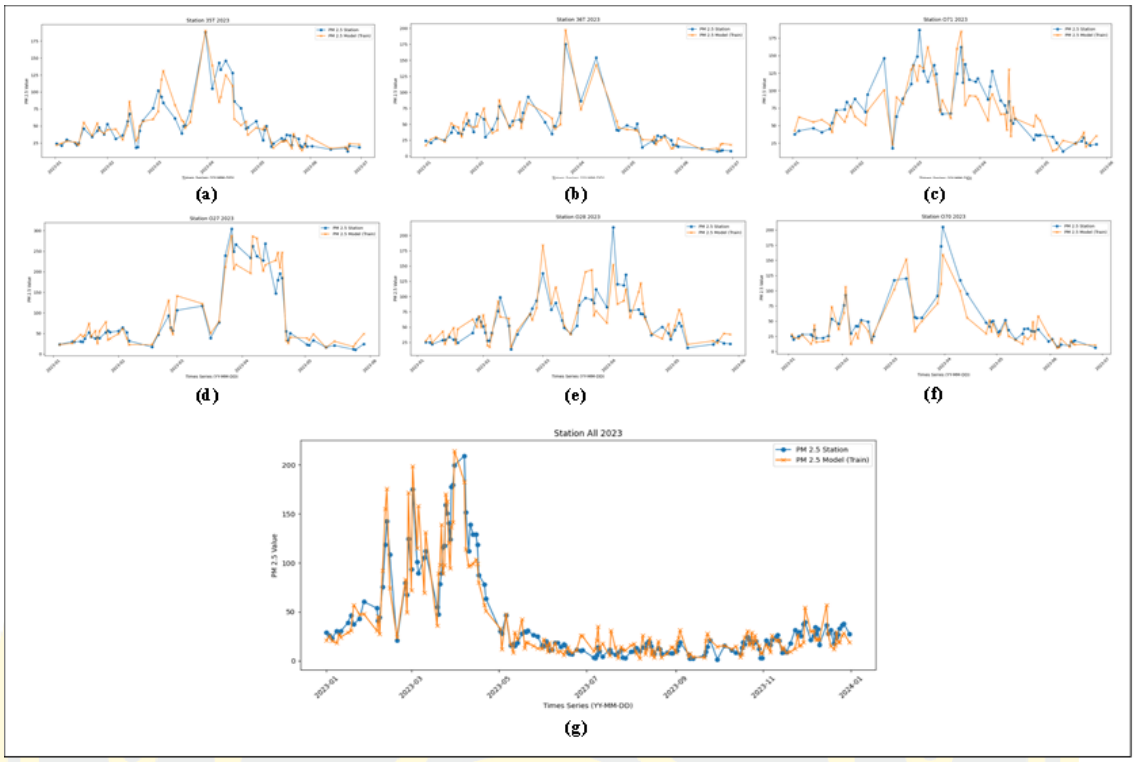


Figure 20 Shows plot times series line chart of PM2.5 concentration daily and month variations of observation (blue) compare with estimation PM2.5 value of the model (orange)

4.5 Impacts of Wildfire on PM2.5 concentration in Chiang Mai Province, Thailand

To investigate the impact of fire on the air quality in Chiang Mai province, Thailand, in 2023, reveal significant seasonal and spatial variations in air pollution levels. The analysis utilizes high-resolution mapping of PM2.5 concentrations, with data derived from a Random Forest (RF) model, showing a close correlation with in situ measurements and highlighting the heterogeneous distribution of pollution across the region.

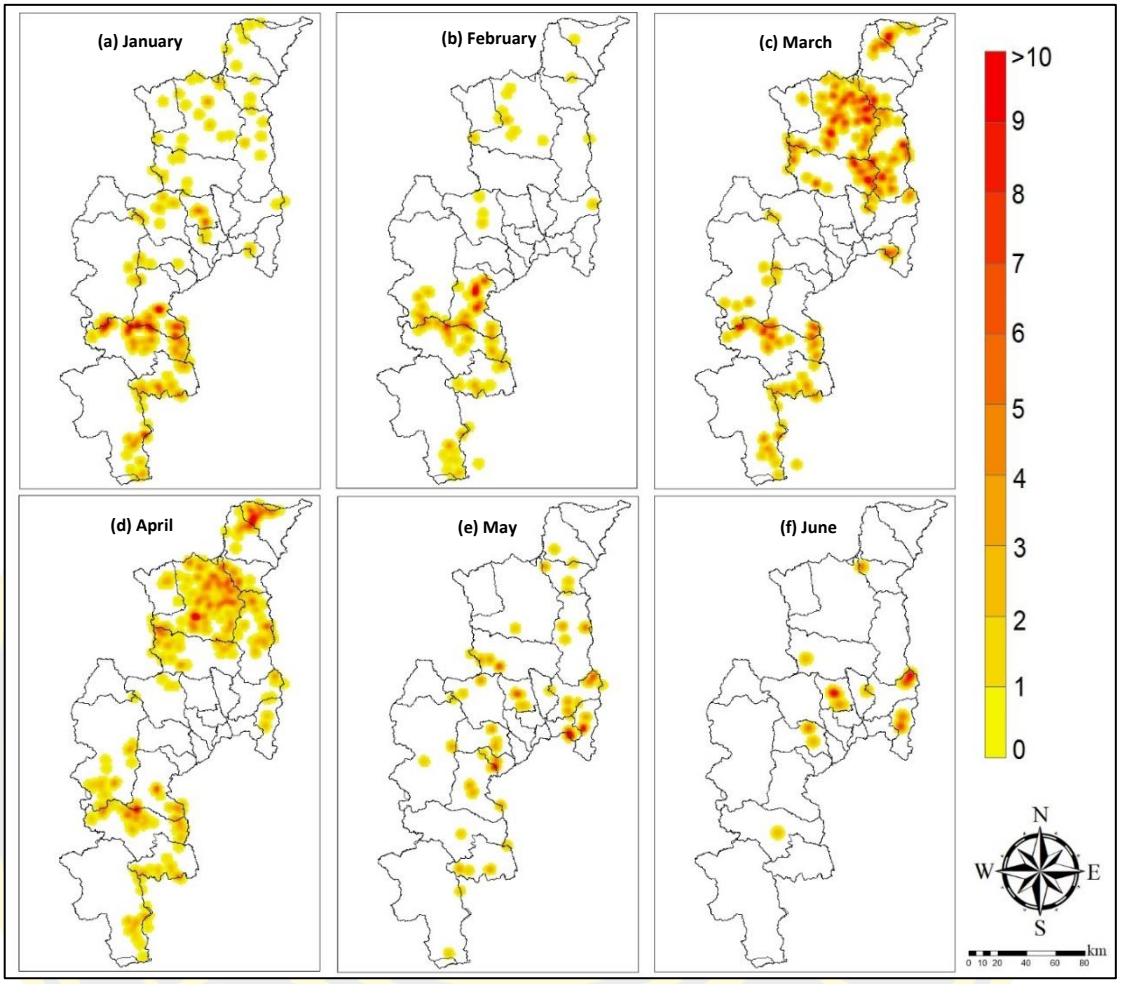


Figure 21 Heat map of severe wildfires based on ground truth data for 1,600 fire incident areas derived from the Forestry Department of Chiang Mai province from 1 Jan. to 30 Jun, in 2023.

In this study, used analyze spatial distribution of Heat maps of severe wildfires based on ground truth data for 1,600 fire incident areas derived from the Forestry Department of Chiang Mai province from 1 Jan. to 30 Jun, in 2023 (Fig. 21) The heat level will be higher and distributed over a wide area. Especially in the southern zone from January to February (in Fig. 21a-b) and in the northern zone from March to April (in Fig. 21c-d). Due to during the period from February 20th to April 10th, PM2.5 concentrations in Chiang Mai Province, Thailand, are particularly high due to the peak of wildfire activity and agricultural burning. This time frame falls within the dry season, characterized by low humidity and stable atmospheric conditions, which exacerbate the accumulation of particulate matter in the air.

The presence of significant correlation between Heat map burn area and PM2.5 suggests that fires play a critical role in explaining air quality variations throughout study area (Figs. 19 and 21). Nevertheless, it is important to acknowledge with care that the occurrence of fires alone cannot be regarded as the sole factor responsible for high PM2.5 concentrations. This is evident from the statistical correlation analysis of the feature importance values derived from the model outcomes.

However, The influence of wildfires is particularly pronounced, contributing significantly to the elevated PM2.5 levels. The data indicate that the year 2023 experienced the most severe wildfires in a decade, exacerbating the pollution levels. Moreover, transboundary pollution from forest fires in neighboring Burma also plays a role, as indicated by the correlation with meteorological data showing positive wind speed values transporting PM2.5 across borders.

And although wind components and wind speed show relatively smaller importance (Fig. 17), wind system across Southeast Asia also plays a crucial role on the pattern of emission propagation. The northeasterly wind is generally dominant in Thailand under the influence of the trade winds. However, two different monsoon systems in South Asia, such as the northeast and the southwest monsoon, cause seasonal changes in the wind direction and speed (Inthacha, 2011). Also, their variabilities are known to be modulated by the El Niño Southern Oscillation (Kirtphaiboon et al., 2014).

4.6 Summary of Experiment and Result

The study conducted an evaluation of model performance for estimating PM2.5 concentrations in Chiang Mai Province, Thailand, using data from six monitoring stations. Three machine learning models were compared are Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Convolutional Neural Network (CNN). To evaluate the concentration of PM2.5 at the station location level the result show RF outperformed the other models with the highest R-squared (R^2) with 0.89, lowest Root Mean Square Error (RMSE) with 11.61 $\mu g/m^3$, and lowest Mean Absolute Percentage Error (MAPE) with 34.22 $\mu g/m^3$. And the RF model has proven potential in estimating PM2.5 concentrations more than other models, Therefore it is used for modeling spatial-temporal distribution maps for estimate of PM2.5 concentration according to time of month and seasonal differences on regional scale (according to

January - June) is R^2 of 0.85, 0.81, 0.79, 0.80, 0.81, and 0.74 respectively, lowest Root Mean Square Error (RMSE) with 18.25, 16.03, 13.51, 19.31 ,12.13, and 21.02 $\mu g/m^3$, respectively, and have lowest Mean Absolute Percentage Error (MAPE) with 23.71, 22.14, 22.76 ,25.54, 20.37, and 20.29 $\mu g/m^3$, respectively. And finally, show outcome R^2, RMSE, and MAPE in regional estimate (across different months and seasons) overall throughout the study period was 0.81, 14.45 $\mu g/m^3$, and 21.25 $\mu g/m^3$, respectively.

The RF showed superior performance, particularly in areas with elevated PM2.5 concentrations attributed to agricultural burning and forest fires. Spatial patterns of PM2.5 concentrations were analyzed, revealing higher values in the north and south zone compared to central zone. Correlation analysis identified relationships between PM2.5 concentrations and various geographic, satellite data, and meteorological variables. PM2.5 concentration showed negative correlations with meteorological factors like precipitation, air pressure, and relative humidity, while exhibiting positive correlations with wind speed and satellite-derived air quality data. Feature importance analysis using the RF model highlighted significant contributions from variables such as aerosol optical depth (AOD), Total Aerosol Extinction AOT 550 nm (TOTEXTTAU), Dust surface mass concentration PM < 2.5 μm (DUSMASS25)), and carbon monoxide (CO). Spatial-temporal distributions of estimated PM2.5 concentrations showed seasonal variations, with peak pollution levels occurring during the dry season due to agricultural burning and forest fires.

The study's findings underscored the importance of accurately estimating PM2.5 concentrations for understanding air quality dynamics in the region, particularly in addressing issues related to agricultural practices and forest fires. However, limitations such as uneven distribution of monitoring stations were noted, highlighting the need for further research to improve the understanding of PM2.5 pollution in Chiang Mai Province.

CHAPTER 5
CONCLUSION AND FUTURE WORK

5.1 Conclusion

Wildfires are a prevalent disturbance globally, impacting ecosystems and contributing to air pollution. Thailand, including Chiang Mai Province, faces challenges exacerbating air pollution that occur from wildfires and health risks. Advanced machine learning techniques aim to improve PM2.5 estimation accuracy, utilizing remote sensing data to address monitoring limitations, these poses a serious challenge for accuracy and reliability by applying data from using satellite data modeling PM2.5 concentrations in Chiang Mai Province, Thailand, presents intriguing insights into the efficacy of different machine learning algorithms. The comparison between Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Convolutional Neural Network (CNN) models reveals notable differences in their predictive capabilities. Firstly, the RF model emerges as having the most robust performance among the three from estimation at on-site scale, consistently exhibiting superior results across various evaluation metrics with R^2, RMSE, and MAPE values of 0.89, 11.61 $\mu g/m^3$, and 34.22 $\mu g/m^3$, respectively, the RF model demonstrates its effectiveness in accurately estimating PM2.5 concentrations. In contrast, the XGBoost model, while still yielding respectable results, falls short of the RF model's performance with R^2, RMSE, and MAPE values of 0.81, 12.51 $\mu g/m^3$, and 37.38 $\mu g/m^3$, respectively, the XGBoost model shows inferior predictive compared to RF. Furthermore, the CNN model exhibits the poorest performance among the three models evaluated with R^2, RMSE, and MAPE values of 0.67, 14.70 $\mu g/m^3$, and 37.78 $\mu g/m^3$, respectively. And The results of estimation PM2.5 concentration at the regional scale, using the RF model show good performance, the test accuracy of the monthly model, averaged over the entire study period, on winter and dry season perform well, with R^2 of 0.79-0.85, RMSE of 13.51-19.31 $\mu g/m^3$, and MAPE of 22.14-25.54 $\mu g/m^3$, higher than the rainy season model with R^2 of 0.74-0.81, RMSE of 12.43-21.02 $\mu g/m^3$, and MAPE of 20.29-20.37 $\mu g/m^3$. This is due to higher PM2.5 concentrations during winter and dry seasons. Averaged over the 6-month study period, the model achieved an R^2 of 0.81, RMSE of 14.45 $\mu g/m^3$, and MAPE of 21.25 $\mu g/m^3$.

The AOD-MAIAC, boasting a spatial resolution of 1 km × 1 km, has emerged as a pivotal tool for estimating ground-level PM2.5 concentrations. However, its efficacy is hampered by missing data arising from cloud cover and high surface reflectivity, which substantially curtails further enhancements in model performance. To address this limitation, we integrate the Total Aerosol Extinction AOT 550nm aerosol reanalysis from MERRA-2 at the station level, supplementing it with chemical composition data from the Sentinel-5P TROPOMI satellite to bolster the analysis. Additionally, we incorporate a model reanalysis of MERRA-2 surface-level Total Surface Mass PM2.5, which significantly enriches the model's training capacity. Leveraging the RF regression algorithm, we assess feature importance, select pertinent variables, and validate the relationship between PM2.5 and aerosol data, encompassing on atmospheric gas composition from satellites, meteorological factors, and geographical data. Calibration of the model's performance is executed using the 5-fold cross-validation method was used to calibrated the model's performance.

A critical analysis of the evaluation indicators across six monitoring stations highlights notable discrepancies, particularly in stations O27, O28, O70, and O71, which display markedly higher RMSE and MAPE values. These disparities can be attributed to the wider range of PM2.5 concentrations observed in these areas, primarily due to agricultural burning and intense forest fires during the summer months.

The spatial distribution of PM2.5 concentrations reveals distinct patterns, with elevated levels observed in both the north zone and south zone of the study area. This is most consistent with data on wildfire activity. Heat maps occurred severe wildfire based on ground truth data shows the actual heat points from combustion consistent with the same period of the month where the concentration of PM2.5 was high from January to April. This is the clearest confirmation concerning the impact of wildfires on air quality and PM2.5 concentration values, which is in accordance with the assumptions based on the research objectives. Conversely, lower concentrations are noted at central zone stations within the study. This spatial variation underscores the influence of geographical factors and human activities on PM2.5 pollution levels. Correlation analysis identifies significant relationships between PM2.5 concentrations and various geographic, satellite, and meteorological variables. Factors such as precipitation, air pressure, and relative humidity show negative correlations with

PM2.5 concentrations, while wind speed exhibits a positive correlation. Additionally, satellite-derived air quality data and land surface temperature demonstrate positive associations with PM2.5 concentrations. Feature importance analysis highlights the significant contributions of variables such as aerosol optical depth, total surface concentration of PM2.5, carbon monoxide, and fire count data in estimating PM2.5 concentrations. These findings underscore the importance of incorporating diverse datasets and variables in PM2.5 prediction models to enhance accuracy and reliability. Overall, the study's comprehensive evaluation and analysis provide valuable insights into the dynamics of PM2.5 pollution in Chiang Mai Province, Thailand. The superior performance of the RF model, coupled with its ability to accurately estimate PM2.5 concentrations across various spatial and temporal scales, suggests its potential utility in environmental protection and epidemiological research.

5.2 Future work

Based on the comprehensive evaluation of PM2.5 concentration estimation models conducted in Chiang Mai Province, Thailand, there are several avenues for improving the scientific rigor and practical implications of our research:

Firstly, Given the demonstrably superior performance of the Random Forest (RF) model compared to the eXtreme Gradient Boosting (XGBoost) and Convolutional Neural Network (CNN) models, it is recommended to undertake further refinement of the RF model's training process. This refinement process may entail the exploration of additional features or the fine-tuning of hyperparameters to enhance its predictive capabilities. Such efforts are crucial for optimizing accuracy, especially in capturing the intricate dynamics of PM2.5 pollution.

Secondly, integrating pattern analysis can enhance the robustness of PM2.5 estimation models. Consider using in-depth data analysis methods such as Deep Learning (DL) to maximize efficiency and understand PM2.5 pollution dynamics. This is particularly useful in regions with high pollution variability due to agricultural burning and wildfires, where researchers can glean valuable insights into PM2.5 patterns.

By incorporating these recommendations into future research endeavors, we can advance our understanding of PM2.5 pollution dynamics in Chiang Mai Province and contribute to the development of effective strategies for air quality management and public health protection.

REFERENCES

Alcaras, E., Costantino, D., Guastaferro, F., Parente, C., & Pepe, M. (2022). Normalized Burn Ratio Plus (NBR+): A New Index for Sentinel-2 Imagery. *Remote Sensing, 14*. doi:10.3390/rs14071727

Bahadur, F., Shah, S., & Nidamanuri, R. (2023). Air Pollution Monitoring, and Modelling: An Overview. *Environmental Forensics*. doi:10.1080/15275922.2023.2297437

Baker, K., & Scheff, P. (2007). Photochemical model performance for PM2. 5 sulfate, nitrate, ammonium, and precursor species SO2, HNO3, and NH3 at background monitor locations in the central and eastern United States. *Atmospheric Environment, 41*(29), 6185-6195.

Balamurugan, V., Chen, J., Qu, Z., Bi, X., & Keutsch, F. N. (2022). Secondary PM< sub> 2.5</sub> decreases significantly less than NO< sub> 2</sub> emission reductions during COVID lockdown in Germany. *Atmospheric Chemistry and Physics, 22*(11), 7105-7129.

Bisri, A. M., & Wahono, R. S. (2015). Penerapan Adaboost untuk Penyelesaian Ketidakseimbangan Kelas pada Penentuan Kelulusan Mahasiswa dengan Metode Decision Tree. *Journal of intelligent systems, 1*, 27-32.

Bloom, S., Takacs, L., Da Silva, A., & Ledvina, D. (1996). Data assimilation using incremental analysis updates. *Monthly Weather Review, 124*(6), 1256-1271.

Bowman, D. M. J. S., Balch, J. K., Artaxo, P., Bond, W. J., Carlson, J. M., Cochrane, M. A., . . . Pyne, S. J. (2009). Fire in the Earth System. *Science, 324*(5926), 481-484. doi:doi:10.1126/science.1163886

Breiman, L. (2001). Random Forests. *Machine Learning, 45*, 5-32. doi:10.1023/A:1010950718922

Burke, M., Driscoll, A., Heft-Neal, S., Xue, J., Burney, J., & Wara, M. (2021). The changing risk and burden of wildfire in the United States. *Proceedings of the National Academy of Sciences, 118*(2), e2011048118. doi:doi:10.1073/pnas.2011048118

Chang, X., Xing, Y., Gong, W., Yang, C., Guo, Z., Wang, D., . . . Yang, S. (2023). Evaluating gross primary productivity over 9 ChinaFlux sites based on random forest regression models, remote sensing, and eddy covariance data. *Science of The Total Environment, 875*, 162601.

Chen, D., Li, X., & Li, S. (2023). A Novel Convolutional Neural Network Model Based on Beetle Antennae Search Optimization Algorithm for Computerized Tomography Diagnosis. *IEEE Transactions on Neural Networks and Learning Systems, 34*(3), 1418-1429. doi:10.1109/TNNLS.2021.3105384

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. Paper presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA. https://doi.org/10.1145/2939672.2939785

Chin, M., Ginoux, P., Kinne, S., Torres, O., Holben, B. N., Duncan, B. N., . . . Nakajima, T. (2002). Tropospheric aerosol optical thickness from the GOCART model and comparisons with satellite and Sun photometer measurements. *Journal of the*

atmospheric sciences, 59(3), 461-483.

ChooChuay, C., Pongpiachan, S., Tipmanee, D., Suttinun, O., Deelaman, W., Wang, Q., . . . Palakun, J. (2020). Impacts of PM2. 5 sources on variations in particulate chemical compounds in ambient air of Bangkok, Thailand. *Atmospheric Pollution Research, 11*(9), 1657-1667.

Chudnovsky, A., Tang, C., A, L., Y, W., J, S., & P, K. (2013). A critical assessment of high resolution aerosol optical depth (AOD) retrievals for fine particulate matter (PM) predictions.

Colarco, P., da Silva, A., Chin, M., & Diehl, T. (2010). Online simulations of global aerosol distributions in the NASA GEOS‐4 model and comparisons to satellite and ground‐based aerosol optical depth. *Journal of Geophysical Research: Atmospheres, 115*(D14).

da Silva Chagas, C., de Carvalho Junior, W., Bhering, S. B., & Calderano Filho, B. (2016). Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. *Catena, 139*, 232-240.

Di, Q., Kloog, I., Koutrakis, P., Lyapustin, A., Wang, Y., & Schwartz, J. (2016). Assessing PM2. 5 exposures with high spatiotemporal resolution across the continental United States. *Environmental science & technology, 50*(9), 4712-4721.

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., . . . Reichle, R. (2017). The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). *Journal of climate, 30*(14), 5419-5454.

Geng, G., Murray, N. L., Tong, D., Fu, J. S., Hu, X., Lee, P., . . . Liu, Y. (2018). Satellite-Based Daily PM(2.5) Estimates During Fire Seasons in Colorado. *J Geophys Res Atmos, 123*(15), 8159-8171. doi:10.1029/2018jd028573

Giglio, L., Boschetti, L., Roy, D. P., Humber, M. L., & Justice, C. O. (2018). The Collection 6 MODIS burned area mapping algorithm and product. *Remote sensing of environment, 217*, 72-85.

Giri, R. N., Janghel, R. R., Govil, H., & Pandey, S. K. (2022, 8-9 Oct. 2022). *Spatial Feature Extraction using Pretrained Convolutional Neural network for Hyperspectral Image Classification.* Paper presented at the 2022 IEEE 4th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA).

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment, 202*, 18-27. doi:https://doi.org/10.1016/j.rse.2017.06.031

Gupta, P., Doraiswamy, P., Levy, R., Pikelnaya, O., Maibach, J., Feenstra, B., . . . Mills, K. (2018). Impact of California Fires on Local and Regional Air Quality: The Role of a Low-Cost Sensor Network and Satellite Observations. *GeoHealth, 2*. doi:10.1029/2018GH000136

Hammill, K., & Bradstock, R. (2006). Remote sensing of fire severity in the Blue Mountains: Influence of vegetation type and inferring fire intensity. *International Journal of Wildland Fire - INT J WILDLAND FIRE, 15*. doi:10.1071/WF05051

Inthacha, S. (2011). *The climatology of Thailand and future climate change projections using the regional climate model PRECIS.* University of East Anglia,

Jia, M., Zhao, T., Cheng, X., Gong, S., Zhang, X., Tang, L., . . . Chen, Y. (2017). Inverse relations of PM2. 5 and O3 in air compound pollution between cold and hot seasons over an urban area of east China. *Atmosphere, 8*(3), 59.

Jiao, L., Zhang, B., Xu, G., & Zhao, S. (2016). Spatio-temporal variability of correlation between aerosol optical depth and PM 2.5 concentration. *J. Arid Land Resour. Environ, 30*, 34-39.

Jilin, G., Yiwei, W., Ji, M., Yaoqi, L., Shaohua, W., & Xueming, L. (2022). An Estimation Method for PM2.5 Based on Aerosol Optical Depth Obtained from Remote Sensing Image Processing and Meteorological Factors.

Junpen, A., Garivait, S., & Bonnet, S. (2013). Estimating emissions from forest fires in Thailand using MODIS active fire product and country specific data. *Asia-Pacific Journal of Atmospheric Sciences, 49*, 389-400.

Kalaiarasi, K., & P, E. (2023). A Novel Approach for Optimization of Convolution Neural Network with Particle Swarm Optimization and Genetic Algorithm for Face Recognition. *International Journal on Recent and Innovation Trends in Computing and Communication, 11*, 215-223. doi:10.17762/ijritcc.v11i4s.6531

Kanabkaew, T., & Kim Oanh, N. T. (2011). Development of spatial and temporal emission inventory for crop residue field burning. *Environmental Modeling & Assessment, 16*, 453-464.

Keeley, J. (2009). Fire intensity, fire severity and burn severity: A brief review and suggested usage. *International Journal of Wildland Fire, 18*, 116-126. doi:10.1071/WF07049

Kirtphaiboon, S., Wongwises, P., Limsakul, A., Sooktawee, S., & Humphries, U. (2014). Rainfall variability over Thailand related to the El nino-southern oscillation (ENSO). *J. Sustain. Energy Environ, 5*, 37-42.

Kleist, D. T., Parrish, D. F., Derber, J. C., Treadon, R., Errico, R. M., & Yang, R. (2009). Improving incremental balance in the GSI 3DVAR analysis system. *Monthly Weather Review, 137*(3), 1046-1060.

Li, Y., Sha, Z., Tang, A., Goulding, K., & Liu, X. (2023). The application of machine learning to air pollution research: A bibliometric analysis. *Ecotoxicology and Environmental Safety, 257*, 114911.

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 Sensing, 14*(20), 5239. Retrieved from https://www.mdpi.com/2072-4292/14/20/5239

Liu, Y., Paciorek, C. J., & Koutrakis, P. (2009). Estimating regional spatial and temporal variability of PM2. 5 concentrations using satellite data, meteorology, and land use information. *Environmental health perspectives, 117*(6), 886-892.

Lyu, Y., Wang, Y., Jiang, C., Ding, C., Zhai, M., Xu, K., . . . Wang, J. (2023). Random forest regression on joint role of meteorological variables, demographic factors, and policy response measures in COVID-19 daily cases: global analysis in different climate zones. *Environmental Science and Pollution Research, 30*(32), 79512-79524. doi:10.1007/s11356-023-27320-7

Ma, C., Qiu, X., Beutel, D., & Lane, N. (2023). *Gradient-less Federated Gradient Boosting Tree with Learnable Learning Rates*. Paper presented at the Proceedings of the 3rd Workshop on Machine Learning and Systems, Rome, Italy. https://doi.org/10.1145/3578356.3592579

Mamić, L., Gasparovic, M., & Kaplan, G. (2023). Developing PM2.5 and PM10 Prediction Models on National and Regional Scale Using Open-source Remote Sensing Data. *Environmental Monitoring and Assessment, 195*. doi:10.1007/s10661-023-11212-x

Molod, A., Takacs, L., Suarez, M., & Bacmeister, J. (2015). Development of the GEOS-5 atmospheric general circulation model: Evolution from MERRA to MERRA2. *Geoscientific Model Development, 8*(5), 1339-1356.

Narita, D., Oanh, N. T. K., Sato, K., Huo, M., Permadi, D. A., Chi, N. N. H., . . . Pawarmart, I. (2019). Pollution Characteristics and Policy Actions on Fine Particulate Matter in a Growing Asian Economy: The Case of Bangkok Metropolitan Region. *Atmosphere, 10*(5), 227. Retrieved from https://www.mdpi.com/2073-4433/10/5/227

Neeraj, K. M., Prem, C. P., Subhadip, S., Rajesh, K., & Prashant, K. S. (2022). Spatio-Temporal Monitoring of Atmospheric Pollutants Using Earth Observation Sentinel 5P TROPOMI Data: Impact of Stubble Burning a Case Study.

Nurdiati, S., Najib, M. K., Bukhari, F., Revina, R., & Salsabila, F. N. (2022). Performance Comparison of Gradient-Based Convolutional Neural Network Optimizers For Facial Expression Recognition. *BAREKENG: Jurnal Ilmu Matematika dan Terapan, 16*(3), 927-938.

Peng-in, B., Sanitluea, P., Monjatturat, P., Boonkerd, P., & Phosri, A. (2022). Estimating ground-level PM2.5 over Bangkok Metropolitan Region in Thailand using aerosol optical depth retrieved by MODIS. *Air Quality, Atmosphere & Health, 15*(11), 2091-2102. doi:10.1007/s11869-022-01238-4

Phairuang, W., Suwattiga, P., Chetiyanukornkul, T., Hongtieab, S., Limpaseni, W., Ikemori, F., . . . Furuuchi, M. (2019). The influence of the open burning of agricultural biomass and forest fires in Thailand on the carbonaceous components in size-fractionated particles. *Environmental Pollution, 247*, 238-247.

Pinichka, C., Makka, N., Sukkumnoed, D., Chariyalertsak, S., Inchai, P., & Bundhamcharoen, K. (2017). Burden of disease attributed to ambient air pollution in Thailand: A GIS-based approach. *PloS one, 12*(12), e0189909.

Prakash, A., Thangaraj, J., Roy, S., Srivastav, S., & Mishra, J. K. (2023). Model-Aware XGBoost Method Towards Optimum Performance of Flexible Distributed Raman Amplifier. *IEEE Photonics Journal, 15*(4), 1-10. doi:10.1109/JPHOT.2023.3286272

Punsompong, P., Pani, S. K., Wang, S.-H., & Pham, T. T. B. (2021). Assessment of biomass-burning types and transport over Thailand and the associated health risks. *Atmospheric Environment, 247*, 118176.

Purwono, P., Ma'arif, A., Rahmaniar, W., Imam, H., Fathurrahman, H. I. K., Frisky, A., & Haq, Q. M. U. (2023). Understanding of Convolutional Neural Network (CNN): A Review. *2*, 739-748. doi:10.31763/ijrcs.v2i4.888

Rienecker, M. M., Suarez, M. J., Gelaro, R., Todling, R., Bacmeister, J., Liu, E., . . . Kim, G.-K. (2011). MERRA: NASA's modern-era retrospective analysis for research and applications. *Journal of climate, 24*(14), 3624-3648.

Rodrigues, J. A., Libonati, R., Pereira, A. A., Nogueira, J. M., Santos, F. L., Peres, L. F., . . . Giglio, L. (2019). How well do global burned area products represent fire patterns in the Brazilian Savannas biome? An accuracy assessment of the

MCD64 collections. *International Journal of Applied Earth Observation and Geoinformation, 78*, 318-331.

Rodriguez-Galiano, V., Mendes, M. P., Garcia-Soldado, M. J., Chica-Olmo, M., & Ribeiro, L. (2014). Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (Southern Spain). *Science of the Total Environment, 476*, 189-206.

Shamsaldin, A., Fattah, P., Rashid, T., & Al-Salihi, N. (2019). The Study of The Convolutional Neural Networks Applications. *UKH Journal of Science and Engineering, 3*, 31-40. doi:10.25079/ukhjse.v3n2y2019.pp31-40

Smucker, K. M., Hutto, R., & Steele, B. (2005). Changes in bird abundance after wildfire: Importance of fire severity and time since fire. *Ecological Applications, 15*, 1535-1549.

Son, R., Stratoulias, D., Kim, H. C., & Yoon, J.-H. (2023). Estimation of surface PM2.5 concentrations from atmospheric gas species retrieved from TROPOMI using deep learning: Impacts of fire on air pollution over Thailand. *Atmospheric Pollution Research, 14*(10), 101875. doi:https://doi.org/10.1016/j.apr.2023.101875

Sritong-aon, C., Thomya, J., Kertpromphan, C., & Phosri, A. (2021). Estimated effects of meteorological factors and fire hotspots on ambient particulate matter in the northern region of Thailand. *Air Quality, Atmosphere & Health, 14*, 1857-1868.

Suthini, D., Thopawan, P., Somporn, C., & Lawa, P.-C. (2018). Smog in the North: severity, effects, causes and solutions.

Tatshakon.Pols. (2022). Efficiency Study of Pm2.5 Forecasting In Bangkok By The Hybrid Model With Weighted Values.

Tucker, W. G. (2000). An overview of PM2. 5 sources and control strategies. *Fuel Processing Technology, 65*, 379-392.

Wang, Y., Yuan, Q., Li, T., Tan, S., & Zhang, L. (2021). Full-coverage spatiotemporal mapping of ambient PM2.5 and PM10 over China from Sentinel-5P and assimilated datasets: Considering the precursors and chemical compositions. *Science of The Total Environment, 793*, 148535. doi:https://doi.org/10.1016/j.scitotenv.2021.148535

Wei, J., Li, Z., Guo, J., Sun, L., Huang, W., Xue, W., . . . Cribb, M. (2019). Satellite-derived 1-km-resolution PM1 concentrations from 2014 to 2018 across China. *Environmental Science & Technology, 53*(22), 13265-13274.

Weichenthal, S. A., Godri Pollitt, K., & Villeneuve, P. J. (2013). PM 2.5, oxidant defence and cardiorespiratory health: a review. *Environmental Health, 12*, 1-8.

Wilson, W. E. (1998). Fine and coarse particles: Chemical and physical properties important for the standard-setting process. In T. Schneider (Ed.), *Studies in Environmental Science* (Vol. 72, pp. 87-115): Elsevier.

Wu, J., Xu, C., Wang, Q., & Cheng, W. (2016). Potential sources and formations of the PM2. 5 pollution in urban Hangzhou. *Atmosphere, 7*(8), 100.

Wu, W.-S., Purser, R. J., & Parrish, D. F. (2002). Three-dimensional variational analysis with spatially inhomogeneous covariances. *Monthly Weather Review, 130*(12), 2905-2916.

Yang, F., Chen, Z., & Gangopadhyay, A. (2022). Using Randomness to Improve Robustness of Tree-Based Models Against Evasion Attacks. *IEEE Transactions*

on Knowledge and Data Engineering, 34(2), 969-982. doi:10.1109/TKDE.2020.2987299

Yin, X., Fallah-Shorshani, M., McConnell, R., Fruin, S., Chiang, Y.-Y., & Franklin, M. (2023). Quantile Extreme Gradient Boosting for Uncertainty Quantification. *arXiv preprint arXiv:2304.11732*.

You, W., Zang, Z., Zhang, L., Li, Y., Pan, X., & Wang, W. (2016). National-Scale Estimates of Ground-Level PM2.5 Concentration in China Using Geographically Weighted Regression Based on 3 km Resolution MODIS AOD. *Remote Sensing, 8*(3), 184. Retrieved from https://www.mdpi.com/2072-4292/8/3/184

Estimate Particulate Matter (PM2.5) Concentrations impact of severity wildfire using Machine learning in Chiang Mai province, Thailand

BIOGRAPHY

NAME	SENA THIWAKORN
DATE OF BIRTH	10 Junly 1993
PLACE OF BIRTH	Mueang sisaket district, Sisaket province, Thailand.
PRESENT ADDRESS	68/321 Rachatani9 saimai Rd. saimai District bangkok 10210
POSITION HELD	Officer of Training Program, Department of Training Program, Intelligence, Surveillance, and Reconnaissance Center, Air Operations Control Command, Royal Thai Air Force (RTAF).
EDUCATION	Bachelor's degree in Political Science, Faculty of Political Science, Ramkhamhaeng University, Thailand.