Enhancing Antenatal Care In South Africa Through Urban Digital Twin Technology: Integrating PM2.5 And Demographic Data For Improved Maternal And Neonatal Outcomes In LMICs.

COM00151M: Independent Research Project

Executive Summary

This project aimed to integrate urban digital twin technology with antenatal care programs to enhance maternal and neonatal health outcomes in South Africa, specifically focusing on the Gert Sibande District in Mpumalanga province. The study developed a predictive model for PM2.5 concentrations, using machine learning and geospatial data, and integrated these predictions into a digital twin framework for healthcare decision-making.

The motivation for this work stems from the urgent need to mitigate the public health risks associated with air pollution, particularly $PM_{2.5}$, which has been identified as a leading contributor to adverse health outcomes globally. In South Africa, where air pollution levels often exceed the World Health Organisation (WHO) guidelines, pregnant women face significant risks, including preterm birth and congenital anomalies such as orofacial cleft lip and palate [1]. Despite this, there has been limited integration of air quality data into healthcare frameworks [33] in low- and middle-income countries (LMICs).

To address this challenge, this study used 2023 as the base year for training the predictive model due to computational requirements. The dataset included meteorological, chemical, and geospatial parameters, derived from ground-based air quality monitoring stations, satellite data, and demographic information. XGBoost, a robust machine learning algorithm, was applied with advanced hyperparameter optimisation techniques via Optuna to ensure high accuracy [33].

The final model achieved a cross-validation average root mean square error (RMSE) of **5.8004** and an R^2 score of **0.9004**, indicating strong predictive performance. Key hyperparameters included a maximum depth of 12, a learning rate of 0.10086, and 680 estimators, among others. Feature importance analysis revealed that factors like boundary

layer height, wind patterns, and chemical pollutants such as NO_2 and SO_2 were critical in predicting $PM_{2.5}$ levels.

Interactive visualisations were implemented using Kepler.gl, showcasing pollution hotspots and temporal patterns. The WHO guidelines on $PM_{2.5}$ were embedded into the digital twin framework to classify air quality into categories: green (safe), amber (moderate), and red (dangerous). The application provided real-time policy recommendations, such as advising pregnant women to stay indoors during high pollution events and wearing masks when outdoors.

Future research included focusing on expanding the dataset beyond 2023 to improve temporal generalisability and predictive accuracy [20]. The integration of real-time air quality monitoring systems with healthcare alerts remains a key area for development, enabling dynamic, proactive interventions. Additional exploration of temporal forecasting models and pollutant interactions would further enhance the digital twin's utility.

Legal, social, and ethical considerations were addressed, including ensuring the use of anonymised and aggregated data to protect individual privacy. Ethical guidelines were followed to avoid misuse of personal health data, and all geospatial data complied with local regulations.

This study demonstrated the feasibility and benefits of integrating digital twin technology with antenatal care programs to reduce air pollution exposure risks. The scalable framework developed here has the potential for adoption across LMICs, enabling targeted interventions to improve maternal and neonatal health outcomes.

Table of Contents

Executive Summary 1

Table of Contents 3

Table of Tables 5

1. Introduction 7

1.1. Research Aims 8

1.2. Objectives 8

1.3. Significance of the Research 9

1.4. Research Questions 9

1.5. Hypotheses 10

2. Literature Review 13

2.1. Environmental Context: Air Quality in South Africa 13

2.2. Air Quality Monitoring in LMICs 14

2.3. Machine Learning (ML) Approaches for Predicting $PM_{2.5}$ 14

2.4. Using Digital Twin to Model $PM_{2.5}$ Levels 16

2.5. Antenatal Care Program and Recommendations 19

2.6. Research Gap 20

3. Research Methodology 21

3.1. Philosophical Approach 21

3.2. Study Area 23

3.3. Computational Approach 24

3.3.1. Data Integration and Preprocessing 24

3.3.2. Feature Engineering 25

3.3.3. Machine Learning and Model Selection 25

3.3.4. Validation and Visualisation 26

3.3.5. Superiority of the Approach 27

4. Methods 28

4.1. Data Collection Methods 28

4.1.1. Air Quality Data 29

4.1.2. ERA5 Hourly Climate Data 30

4.1.3. Population Density Data 30

4.1.4. MODIS/Terra+Aqua MAIAC AOD Data 30

4.1.5. Ward Administration Boundaries 30

4.2. Data Analysis Methods 30

4.2.1. Data Preprocessing 31

4.2.2. Feature Engineering 31

4.2.3. Model Training and Evaluation 31

4.2.4. Spatial and Temporal Modelling 32

4.2.5. Variance Analysis and Policy Recommendations 32

4.3. Integration and Application 33

4.4. Ethical Considerations 35

5. Results and Discussions 37

5.1. Results 37

5.2. Feature Importance 39

5.3. Model Optimisation 42

5.4. Comparative Analysis 43

5.5. Challenges with Dataset 43

5.6. Response to Hypothesis 44

5.7. Policy Recommendations 45

5.8. Limitations 45

6. Conclusions 48

6.1. Future Research 48

7. References 51

Table of Tables

Table 1: PM2.5 estimator datasets used to train the XGBoost model 28

Table 2: Top 15 Features Ranked by F-Score in the PM2.5 Prediction Model. The table lists the most influential features in the XGBoost model, ranked by their importance scores (F-Score) 39

Table of Figures

Fig 1: Secunda, Gert Sibande District, Mpumalanga- a region with significant public health concerns due to elevated PM2.5 levels. 23

Fig 2: Digital twin model- design process flow 24

Fig 3: Digital Twin Model design with integrated data sources. The digital twin model integrates PM2.5 measurements, climate data, population density, and satellite AOD data to predict air quality. Insights from the model are fed into a recommender system..... 34
Fig 4: Density Plot of Predicted vs Actual PM2.5 Concentrations. The chart visualises the relationship between actual and predicted PM2.5 values using a density-based 2D histogram. The colour gradient represents the density of overlapping points..... 38
Fig 5: : PM2.5 Spatial Distribution Visualisation using the Streamlit Application Model. The map highlights PM2.5 concentration polygons for Msukaligwa Local Municipality, classified into five ranges (6.49–113.29 µg/m³)..... 44

1. Introduction

A digital twin (DT) is a dynamic, real-time virtual replica of a city that integrates data from multiple sources to simulate, analyse, and optimise urban systems [1]. Advancements in urban digital twins have facilitated enhanced solid waste management for connected cities and improved air quality management on university campuses [2, 3]. Additionally, Spain utilises urban digital twins to simulate floods under pre- and post-event conditions, aiding in predicting the hydraulic behaviour of channels during large floods [1], while in Barcelona, they are employed to determine optimal locations for climate shelters during extreme weather events [1]. In healthcare, although there have been significant advancements in DT in healthcare planning, and construction management [3, 4], other areas such as urban air quality management have remained siloed. Making it difficult for policy-makers to significantly improve health outcomes by using data. In healthcare, the World Health Organisation (WHO) has reported ambient and household air pollution as the leading cause of death, 6 million in 2022 [5, 6], with particulate matter below 2.5µm ($PM_{2.5}$), as one of the core pollutants contributing to these statistics. Globally, 90% of the population is exposed to unsafe $PM_{2.5}$ levels, and 80% of whom live in low- and middle-income (LMIC) countries. Moreover, approximately 10% of the population has low income (£1.50 per day) and reside in areas with high air pollution, especially in Sub-Saharan Africa [7]. Air pollution levels are particularly high in LMICs, in industrial areas. Studies have shown that air quality significantly affects individuals with comorbidities such as the elderly, infants and pregnant women. For pregnant women in LMICs, ambient air pollutant exposure during the first trimester increased the risk of birth defects including congenital heart disease and cleft lip with or without cleft palate [8]. Previous studies have focused on studying the relationship between air quality and health outcomes with researchers emphasising the need to introduce

effective measures of air pollution control and prenatal care to minimise exposure to air pollution during pregnancy [6, 8].

1.1. Research Aims

This research aims to explore the integration of urban air quality management to existing antenatal care programmes in LMICs. This study will focus on South Africa (SA) within the broader context of LMICs. The study aims to use Geographic Information System (GIS) to recommend interventions for policy-makers in the healthcare space to model $PM_{2.5}$ using urban digital twin technology. Urban Digital twins are increasingly adopted in the healthcare sector to monitor patient health and improve predictive analytics [9]. The study area is the eMbalenhle and Secunda neighbourhoods in Gert Sibande District in the Mpumalanga province of South Africa.

1.2. Objectives

This research introduces the first Digital Twin model for antenatal care management in low- and middle-income countries (LMICs), specifically integrating Landscan population estimates, ERA5 hourly climate reanalysis and $PM_{2.5}$ readings from both satellite aerosol optical depth (AOD) and ground-based stations. The goal is to apply urban digital twin technology to antenatal care (ANC) programs in South Africa, with the aim of contextualising these programs to enhance pregnancy outcomes through the use of environmental and geospatial data [10]. The analysis focuses on women of reproductive age who have been pregnant and reside in communities within South Africa that are affected by poor air quality. This enables a detailed exploration of how environmental factors, such as air pollution, intersect with maternal health, providing valuable insights into the relationship between environmental conditions and health outcomes [3].

1.3. Significance of the Research

Infant mortality in SA remains a significant issue, with an average of 28 deaths per 1,000 live births [11], reflecting broader trends in LMICs. The WHO estimates that around 2.4 million perinatal deaths occur annually, largely in countries with limited access to healthcare services [7]. The effects of industrialisation and inadequate climate change policies have exacerbated environmental health risks in South Africa, where air pollution is a leading contributor to premature deaths globally [6].

1.4. Research Questions

The research questions focus on assessing the impact of $PM_{2.5}$ levels in Mpumalanga and evaluating mitigation strategies to improve maternal and neonatal health outcomes. These questions are crucial due to the widespread and severe impacts of air pollution on human health and the environment [2, 34]. Studies have established a clear link between long-term exposure to particulate matter and increased morbidity and mortality from cardiopulmonary diseases, asthma, and lung cancer, with children and pregnant women being particularly vulnerable [3, 35]. According to the WHO, $PM_{2.5}$ is one of the most dangerous air pollutants, causing systemic health issues and reducing life expectancy. Additionally, air pollution threatens biodiversity, soil, groundwater, and air quality, exacerbating global challenges like climate change and species extinction. Investigating $PM_{2.5}$'s effects in regions with high pollution levels, like Mpumalanga, is essential to understand its localised health and ecological impacts. This study draws on existing research to identify policy interventions, such as improved air quality monitoring and public health messaging, that can reduce these risks and inform future urban and healthcare planning.

1. How can policy-makers in the healthcare space integrate digital twin technology into antenatal care to improve maternal health management in South Africa?
2. How can continuous measurement and prediction of $PM_{2.5}$ levels be used to enhance antenatal care programs?
3. How does integrating environmental and demographic data through digital twin technology improve the personalisation of antenatal care in regions of South Africa affected by poor air quality?

The integration of digital twin technology into antenatal care holds transformative potential for addressing maternal health challenges, particularly in SA, where health inequities and environmental factors significantly impact maternal and neonatal outcomes. According to Katsoulakis et al and Vallee [21, 36], digital twins have already shown promise in healthcare for simulating body systems such as lungs, body functions and the heart, and optimising health operations systems. Their application in antenatal care is timely given the growing evidence linking exposure to PM2.5 during pregnancy with adverse outcomes, including improper immune system, premature birth, and increased neonatal mortality [37]. By enabling continuous measurement and prediction of $PM_{2.5}$ levels, digital twins could enhance antenatal care programs by offering actionable insights for policy-makers to mitigate exposure risks. Furthermore, integrating environmental and demographic data through digital twin models provides a novel pathway for personalising care based on localised air quality conditions and patient-specific health profiles, addressing the socioeconomic and environmental determinants of health [38]. These questions are thus worth investigating as they align with global health priorities and leverage innovative solutions to tackle persistent healthcare challenges.

1.5. Hypotheses

1. H_1: Integrating digital twin technology to predict $PM_{2.5}$ levels, using satellite AOD, land stations, and demographic data, will significantly improve maternal and neonatal health outcomes in SA by offering real-time recommendations for antenatal care. Studies have shown that digital twins can enhance decision-making in healthcare [9] and improve public health interventions through real-time data analysis [4].

   H_0: Integrating digital twin technology for predicting $PM_{2.5}$ levels will have no significant impact on maternal and neonatal health outcomes in SA.

2. H_2: The use of continuous $PM_{2.5}$ monitoring and predictive modelling will enable personalised antenatal care, reducing the risk of adverse pregnancy outcomes in high-risk regions such as the Highveld. Research indicates that continuous air quality monitoring can be linked to improved maternal health outcomes, especially when applied through healthcare frameworks [24].

   H_0: The continuous measurement and prediction of $PM_{2.5}$ will not improve the personalisation of antenatal care or reduce adverse pregnancy outcomes in South Africa.

The subsequent sections build upon the introduction to provide a comprehensive analysis of the study. The Literature Review explores prior research on air pollution, focusing on $PM_{2.5}$ and its effects on health and the environment, while the Methodology details the data-driven techniques, machine learning models, and evaluation metrics employed. The Results section highlights the performance of the optimised XGBoost model, key findings, and visual insights, followed by the Analysis and Discussion, which interprets these findings in the context of public health, especially maternal health in Mpumalanga. Lastly, the Legal, Social, Ethical, and Professional Considerations discuss broader implications, and the Conclusion

and Future Work summarises the study, acknowledges limitations, and outlines pathways for further research.

2. Literature Review

2.1. Environmental Context: Air Quality in South Africa

South Africa's air quality issues are rooted in the nation's dependency on coal and industrial activities, with the Mpumalanga province being one of the most polluted regions globally. The Highveld Priority Area (HPA) represents a critical case study for air pollution research due to its unique industrial concentration, including coal-fired power plants and the world's largest coal-to-liquid (CTL) refinery [12] [15] [18].

Air pollution in the region poses significant public health challenges, as highlighted by Lavietes, who reported annual costs of £205.14 million due to heat waves exacerbating cardiovascular issues.. Studies by Millar et al. and Landrigan & Fuller have demonstrated direct correlations between $PM_{2.5}$ exposure and respiratory ailments, cardiovascular diseases, and adverse pregnancy outcomes [15] [14]. For instance, pregnant women exposed to elevated $PM_{2.5}$ levels in Mpumalanga are at increased risk of preterm births, low birth weights, and congenital anomalies such as cleft palate and congenital heart disease [16] [17].

Despite these findings, SA's regulatory response has been inadequate. The country's air quality management policies often lack stringent enforcement mechanisms, allowing pollutants to exceed WHO guidelines. In Mpumalanga, ambient $PM_{2.5}$ levels are frequently recorded at five times the WHO-recommended threshold [6] [13]. This regulatory gap underscores the need for innovative solutions, such as integrating Geographic Information

System (GIS) mapping and digital twin technology, to address these challenges comprehensively [23] [24].

2.2. Air Quality Monitoring in LMICs

Low- and middle-income countries like South Africa face unique challenges in monitoring and mitigating air pollution. Traditional air quality monitoring networks, such as Continuous Air Quality Monitoring Stations (CAAQMS), are sparse due to high installation and maintenance costs [20] [21]. As a result, LMICs rely heavily on complementary climate systems, such as AOD, to estimate $PM_{2.5}$ levels.

Recent advancements have demonstrated the potential of integrating satellite data with machine learning (ML) models. Katoch et al. achieved a cross-validation R^2 of 0.92 for $PM_{2.5}$ predictions in India by combining AOD data with meteorological variables [32]. This method has proven to be a cost-effective solution for regions with limited ground-based observations. However, the reliance on satellite data introduces limitations, including lower temporal resolution and challenges in capturing localised pollution hotspots [14] [22].

In addition, community-based monitoring initiatives, using low-cost sensors, have begun to fill gaps in traditional monitoring networks. These sensors, when combined with geospatial data, provide granular insights into air pollution's spatial distribution, making them invaluable for public health interventions [24] [25]. Despite their promise, challenges such as data accuracy, calibration, and integration with existing systems remain significant barriers [19] [22].

2.3. Machine Learning (ML) Approaches for Predicting $PM_{2.5}$

The rise of machine learning in environmental modelling has provided transformative tools for predicting $PM_{2.5}$ levels, particularly in regions with sparse monitoring infrastructure. By leveraging satellite data, meteorological variables, and population density data, ML approaches can model the complex interactions that influence air quality, offering a cost-effective alternative to traditional methods [20] [31].

Random Forest (RF) and Gradient Boosting algorithms have emerged as prominent methods for air quality prediction. Zhang et al. demonstrated the effectiveness of RF models in South Africa's Highveld region, achieving robust predictions of daily PM2.5 concentrations with a cross-validation R^2 of 0.80 [20]. These models excel in handling non-linear relationships and variable interactions, outperforming linear regression methods that struggle to capture the dynamic nature of pollution patterns [14] [20]. For example, studies have shown that incorporating predictors such as AOD, temperature, humidity, and land-use data significantly enhances the predictive accuracy of RF models [31].

However, the reliance on data-intensive ML models presents challenges in LMICs. Missing data due to sensor malfunctions or power outages is a common issue that undermines the reliability of ML predictions [22] [23]. Abutalip et al. highlighted the importance of developing preprocessing techniques to handle incomplete datasets and ensure data consistency [22]. Furthermore, the computational demands of advanced ML algorithms can limit their application in resource-constrained settings [19].

To address these challenges, hybrid models that integrate statistical and machine learning methods have gained attention. For instance, a two-stage ensemble model combining Generalised Additive Models (GAM) for temporal trend removal with RF for residual predictions demonstrated superior performance in PM2.5 estimation [31]. This approach

leverages the strengths of statistical models in capturing long-term trends while allowing ML algorithms to focus on residual variability, improving overall accuracy [20].

Deep learning approaches, while less frequently applied in LMICs, offer additional potential. Studies have shown that Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks can capture spatial and temporal dependencies in air quality data, providing high-resolution predictions [14] [22]. However, these models require extensive training datasets and computational resources, making them less accessible to LMICs.

The integration of ML models with Geographic Information System (GIS) tools further enhances their applicability. By mapping PM2.5 exposure and identifying pollution hotspots, GIS-enabled ML models can provide actionable insights for policymakers [25] [26]. For example, in South Africa, combining RF predictions with GIS mapping has enabled the identification of high-risk areas in the Highveld region, facilitating targeted interventions [20].

Despite these advancements, more research is needed to adapt ML techniques to the unique challenges of LMICs. The development of cost-effective, scalable models that account for local data constraints remains a critical area for future work. Collaborative efforts involving local researchers, policymakers, and international stakeholders are essential to maximise the potential of ML in improving air quality monitoring and public health outcomes.

2.4. Using Digital Twin to Model $PM_{2.5}$ Levels

Digital Twin technology has emerged as a transformative tool for environmental monitoring, offering dynamic, real-time virtual representations of physical systems [23] [24]. By

integrating data from diverse sources, including low-cost sensors, satellite imagery, and ML models, DTs enable the simulation, analysis, and optimisation of environmental conditions such as air quality. This capability is particularly valuable in LMICs, where traditional air quality monitoring networks are sparse.

DTs have been applied across various sectors, from urban planning to healthcare, demonstrating their versatility in addressing complex, data-driven challenges [9] [22]. For example, Abutalip et al. developed a DT framework for high-resolution $PM_{2.5}$ estimation by fusing sensory data with AOD-derived satellite imagery [22]. The model not only provided accurate air quality predictions but also delivered actionable policy recommendations, underscoring the potential of DTs to bridge the gap between data collection and decision-making.

In regions such as the HPA, where $PM_{2.5}$ pollution regularly exceeds national and WHO standards, DTs can play a critical role. By integrating real-time data from satellite systems like MODIS with geospatial information, DTs can model air quality dynamics and predict pollution trends [20] [23]. Studies by Wu et al. have highlighted the scalability of DTs, emphasising their capacity to simulate multiple environmental scenarios and evaluate the potential impacts of interventions [23].

One of the key strengths of DTs lies in their ability to provide continuous, real-time insights, enabling proactive responses to pollution events. In LMICs, this real-time functionality is particularly crucial given the absence of comprehensive ground-based monitoring systems [21] [23]. However, the successful implementation of DTs requires addressing several

challenges. Data availability and quality remain significant barriers, as incomplete or inaccurate datasets can undermine the reliability of the digital twin model [22] [24].

Ethical considerations are also central to DT deployment, especially in contexts involving sensitive health and demographic data. Wu et al. noted that trustworthiness, privacy, and security are critical for ensuring that DTs adhere to ethical standards while delivering accurate and actionable insights [23]. For LMICs, where data governance frameworks may be underdeveloped, these concerns are particularly pertinent.

DTs have shown particular promise in enhancing air quality monitoring by integrating various data streams into a unified framework. For instance, Abutalip et al. demonstrated how DTs could merge satellite AOD data with low-cost sensor measurements to create a high-resolution map of $PM_{2.5}$ concentrations [22]. This approach not only improved spatial coverage but also allowed for real-time updates, making it an invaluable tool for policy planning and public health interventions.

In South Africa's Highveld region, where coal mining and petrochemical industries dominate, the application of DTs could revolutionise air quality management. By linking $PM_{2.5}$ predictions with Geographic Information System (GIS) tools, DTs can identify pollution hotspots and simulate the effects of proposed mitigation measures, such as stricter emissions controls or urban green infrastructure development [20] [23].

Despite their potential, the application of DTs in LMICs is still in its infancy, with limited studies exploring their integration into public health frameworks. Research by Weil et al. emphasised the need for more interdisciplinary approaches that combine environmental modelling with healthcare interventions [9]. For example, in antenatal care programs, DTs

could provide real-time alerts to pregnant women in high-pollution areas, helping mitigate the adverse effects of $PM_{2.5}$ exposure on maternal and neonatal health [13] [16].

To maximise their impact, DT implementations in LMICs must prioritise scalability, cost-effectiveness, and local capacity-building. Collaborative efforts involving governments, academic institutions, and private sector stakeholders can drive the adoption of DTs, ensuring that they address the unique challenges of resource-constrained environments [22] [23].

2.5. Antenatal Care Program and Recommendations

The integration of vent hoods and indoor air purifiers [24] into antenatal care programs is increasingly recognised as a critical intervention to mitigate the health risks posed by environmental pollution. Zhu et al., provided evidence that $PM_{2.5}$ exposure during pregnancy is associated with low birth weight, preterm birth, and increased rates of infant mortality. In South Africa, where the population is exposed to annual average $PM_{2.5}$ concentrations five times higher than WHO guidelines [5, 13], the need for air quality interventions in antenatal care is urgent .

Research has demonstrated the feasibility of using open-source geospatial data to map air pollution and its health impacts at the community level. For example, Sacks et al., conducted a study using EPA's BenMAP-CE software to quantify the effects of $PM_{2.5}$ pollution on low birth weight [25] while United Nation (UN) Environment [26] linked it to huge child Intelligent Quotient (IQ) loss, underscoring the disproportionate health burden borne by low-income communities. By incorporating air quality data into antenatal care, healthcare

providers can issue timely recommendations, such as advising pregnant women to reduce outdoor exposure during peak pollution hours or to use air purifiers in their homes [22].

2.6. Research Gap

While significant progress has been made in air quality monitoring and its integration into antenatal care, several research gaps remain. First, most studies focus on high-income countries with robust air quality monitoring infrastructure. In LMICs like South Africa, the reliance on satellite data and machine learning models necessitates further research to improve the accuracy and resolution of these models. Additionally, while digital twin technology holds promise for improving air quality management, there is limited research on its application in antenatal care programs, particularly in LMICs. Second, more research is needed to establish the causal relationship between $PM_{2.5}$ exposure and specific health outcomes, such as congenital anomalies and neonatal health. Finally, there is a need for more policy-oriented research that evaluates the effectiveness of integrating air quality monitoring with antenatal care in reducing maternal and neonatal mortality. The literature highlights the critical role of air quality management in improving maternal and neonatal health outcomes. In highly polluted regions such as Secunda and eMbalenhle, integrating urban digital twin technology and geospatial mapping into antenatal care programs can help policymakers make data-driven decisions. This study will provide insights into how environmental data can be used to enhance antenatal care, ultimately reducing the burden of pollution-related health risks in South Africa.

3. Research Methodology

3.1. Philosophical Approach

In the context of this research, the chosen philosophical approach is rooted in positivism [27], aligning with the deductive research methodology [28]. Positivism is a widely recognised philosophy in scientific research that emphasises empirical evidence, objectivity, and the use of measurable phenomena to test hypotheses [27]. It is well-suited for studies that involve quantitative data, such as air quality readings, and seek to uncover objective truths through observation and experimentation.

The positivist philosophy supports the idea that reality is observable and can be studied independently of the researcher's biases or interpretations [29]. In this study, the impact of $PM_{2.5}$ on maternal health outcomes is an objective phenomenon that can be quantified using air quality data and health statistics. The use of digital twin technology to model and predict these outcomes further strengthens the reliance on data-driven, empirical approaches. In this sense, positivism allows the researcher to focus on quantifiable relationships between air pollution and health outcomes, making it possible to apply scientific methods to test the hypothesis [30].

The deductive approach fits within this positivist framework. Deductive reasoning begins with established theories about the negative effects of air pollution on health, specifically the correlation between $PM_{2.5}$ exposure and adverse outcomes like preterm birth and low birth weight. These theories guide the formulation of hypotheses that will be tested against real-world data collected from the Secunda and eMbalenhle regions. By doing so, the study seeks to either confirm or refute the hypothesis that integrating digital twin technology into

antenatal care can improve maternal and neonatal health outcomes by predicting and mitigating the effects of poor air quality.

Alternative philosophical approaches, such as interpretivism or pragmatism, were considered but deemed less suitable for this study. Interpretivism focuses on understanding subjective experiences and social contexts, often through qualitative methods like interviews or case studies [29, 30]. While valuable in other types of research, interpretivism would not adequately address the objective nature of this study, which relies on measurable environmental data and health outcomes. The goal here is to obtain generalisable findings that can inform policy, rather than exploring individual perceptions or experiences.

Pragmatism, while flexible and outcome-oriented, often combines both quantitative and qualitative methods [29]. It seeks to address practical problems using the most effective tools available. While pragmatism could have been employed if the study included qualitative interviews with healthcare professionals or patients, the research focuses exclusively on quantitative data to test specific hypotheses about air pollution and antenatal health. Thus, positivism provides a more suitable philosophical foundation, as it focuses on objectivity and the ability to generate generalisable results that can contribute to broader public health interventions.

In summary, the positivist philosophy aligns with the deductive methodology employed in this study, focusing on the objective measurement of air pollution's impact on maternal health. By using quantitative data and rigorous hypothesis testing, this research aims to provide scientific evidence that can inform policy-making and enhance antenatal care programs in South Africa and other LMICs facing similar challenges. This philosophy supports the study's goal of contributing to data-driven decision-making in public health, providing

clear and actionable insights into the relationship between environmental health risks and maternal health outcomes.

3.2. Study Area

The study area for this research focuses on Secunda and eMbalenhle in the Gert Sibande District of Mpumalanga [15], South Africa. A map view of the study area is pictured in Fig. 1 below.

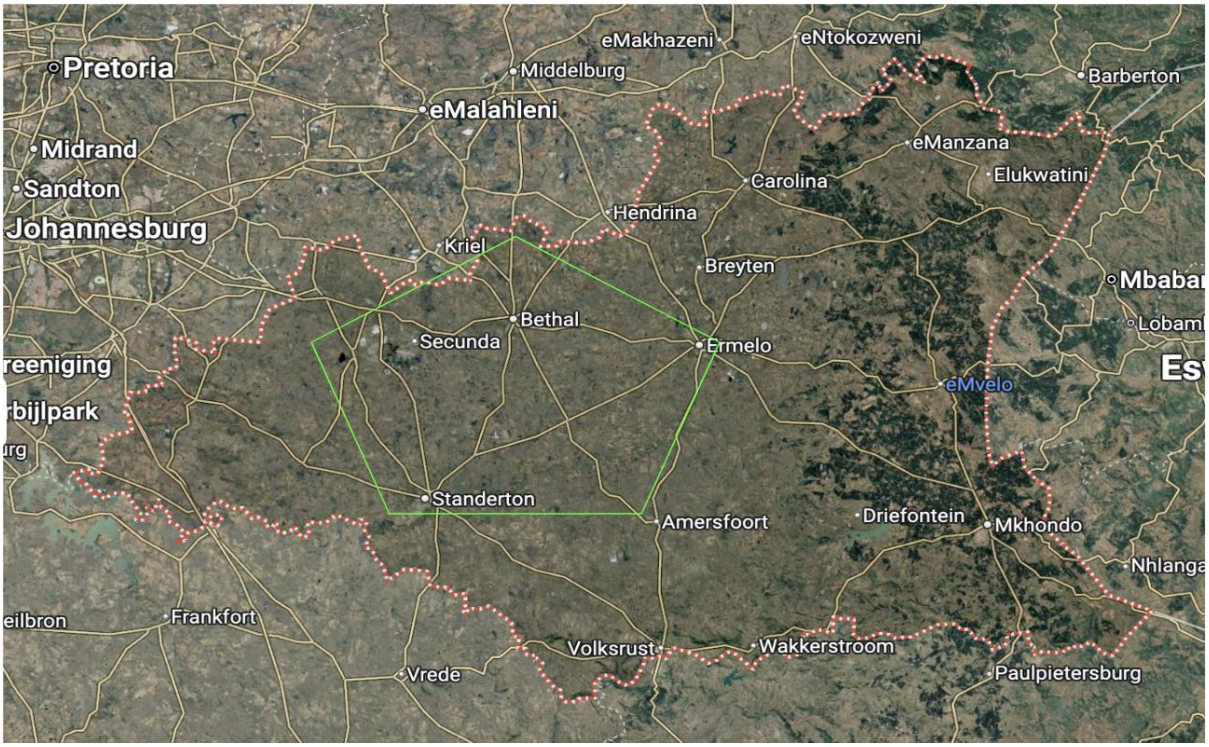

Fig 1: Secunda, Gert Sibande District, Mpumalanga- a region with significant public health concerns due to elevated PM2.5 levels.

Spanning 9.45 km², these industrial towns are part of the HPA, a region severely impacted by air pollution due to the presence of one of the largest CTL plants in the world. This plant significantly contributes to $PM_{2.5}$ levels, affecting public health, particularly pregnant women and children [18].

The research integrates GIS mapping with real-time $PM_{2.5}$ estimates to identify pollution hotspots, offering a comprehensive view of how industrial emissions affect maternal health.

This approach is crucial for informing antenatal care interventions and improving public health outcomes in the region.

3.3. Computational Approach

The approach integrates data pre-processing, harmonisation, and feature engineering to build predictive models using XGBoost [20, 40]. Hyperparameter optimisation ensures model performance, leading to deployment via tools like Streamlit and Kepler.gl for real-time visualisation and policy recommendations [22]. The digital twin system incorporates key performance indicators to enable actionable insights for air quality management [2]. Iterative refinement ensures scalability and accuracy in addressing environmental health challenges.

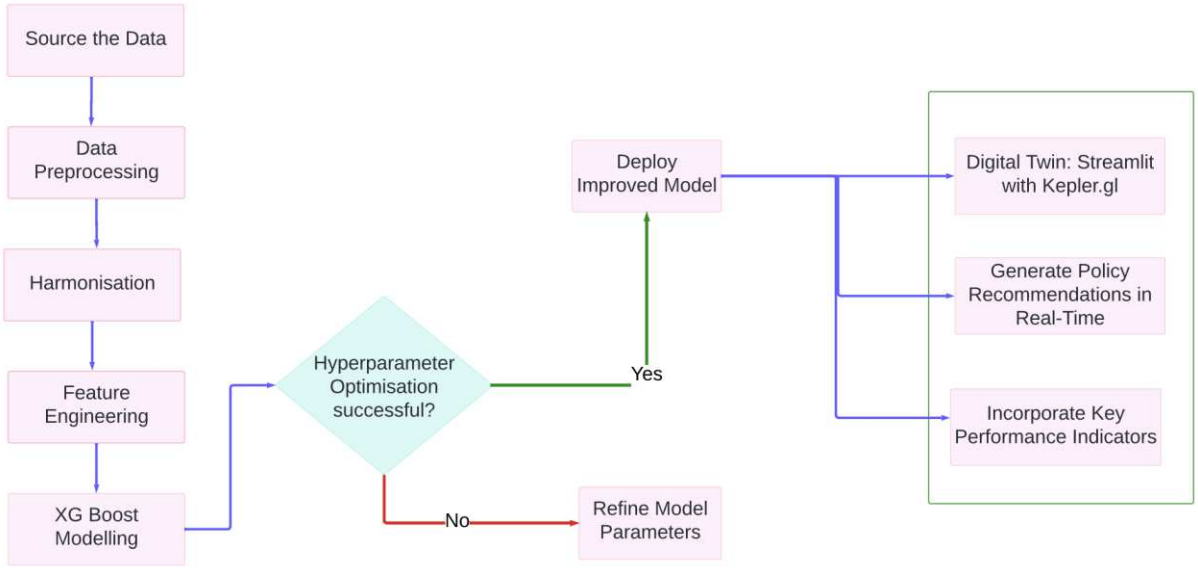

Fig 2: Digital twin model- design process flow

3.3.1. Data Integration and Preprocessing

The integration of satellite data, ground station observations, and climate datasets represents a comprehensive approach to $PM_{2.5}$ prediction. Satellite-derived Aerosol Optical Depth (AOD) data provides broad spatial coverage, making it invaluable in regions like the Highveld Priority Area where ground monitoring networks are sparse [20] [21]. However, satellite data alone is subject to limitations such as reduced accuracy under cloudy conditions

or over heterogeneous terrains [22]. By incorporating ground-based observations, the proposed method enhances reliability through direct pollutant measurements [23]. Climate data, including variables like temperature and wind speed, further enriches the model by accounting for meteorological factors influencing pollutant dispersion [24].

In contrast, approaches that rely solely on ground stations or satellite data often struggle with incomplete spatial or temporal coverage. For instance, regression models using only ground station data fail to account for regional variations, limiting their generalisability [25]. Similarly, satellite-only approaches, while cost-effective, face challenges in capturing fine-scale pollutant dynamics. The multi-source integration in this study mitigates these issues, delivering a more robust dataset for model training and prediction.

3.3.2. Feature Engineering

Feature engineering is pivotal in machine learning models, as it transforms raw data into actionable insights. This study introduces temporal features (e.g., seasonal variability), land-use characteristics, and meteorological parameters to capture the complex interplay of factors affecting $PM_{2.5}$ levels. Research by Zhang et al. highlighted the importance of such multi-faceted features in improving predictive accuracy, particularly in heterogeneous environments like South Africa's Highveld region [26]. The inclusion of these features contrasts with simpler approaches that often overlook temporal or spatial variability, resulting in reduced model performance [27, 41]. By including these features, Zhang's study successfully captured the temporal trends observed in ground measurements and accurately identified central and northern regions as having the highest annual $PM_{2.5}$ concentrations

3.3.3. Machine Learning and Model Selection

Gradient Boosting, implemented using XGBoost, was selected for its ability to handle non-linear relationships, interactions between features, and missing data [28]. This algorithm has consistently outperformed traditional regression models in air quality studies, offering superior accuracy and resilience to overfitting [29]. For instance, a study in urban China demonstrated that Gradient Boosting reduced prediction errors by 15% compared to multiple linear regression models [30]. Furthermore, the use of hyperparameter optimisation via Optuna allows for the fine-tuning of model parameters, enhancing predictive power [31].

Alternative models, such as neural networks, while theoretically more powerful, require significantly larger datasets and computational resources. Deep learning approaches like Convolutional Neural Networks (CNNs) have been employed in air quality modelling but often suffer from overfitting when applied to sparse datasets common in LMICs [32]. Given the resource constraints and data sparsity in this study's context, XGBoost provides an optimal balance between complexity and interpretability.

3.3.4. Validation and Visualisation

Cross-validation ensures model robustness by evaluating performance on multiple data splits, reducing the risk of overfitting. Validation metrics such as Root Mean Squared Error (RMSE) and R^2 offer clear, interpretable measures of model accuracy [33]. These methods are standard in air quality modelling and provide transparency in assessing model performance.

The leveraging of KeplerGL's interactive capabilities enhances the study's utility for policymakers. By presenting data in an accessible format, these tools enable informed

decision-making. In contrast, studies that rely solely on numerical results often fail to bridge the gap between scientific findings and practical implementation [34].

3.3.5. Superiority of the Approach

This computational framework excels by addressing the key limitations of alternative approaches. By integrating diverse data sources, it overcomes spatial and temporal gaps inherent in satellite-only or ground-based methods. The use of Gradient Boosting ensures model robustness [28], while interactive visualisations enhance the accessibility and impact of the findings [34]. These attributes make the approach particularly suited to the Highveld Priority Area, a region characterised by complex pollution dynamics and limited monitoring infrastructure. Ultimately, this methodology not only advances scientific understanding but also supports actionable interventions to mitigate air pollution.

4. Methods

4.1. Data Collection Methods

The data collection process involved integrating multiple datasets to train an XGBoost model which Pan concluded that it outperforms multiple linear regression and random forest [42] for estimating $PM_{2.5}$ concentrations. Table 1 below breaks down how datasets were sourced from ground-based measurements, climate reanalysis data, satellite imagery, administrative boundaries, and population density information.

Table 1: PM2.5 estimator datasets used to train the XGBoost model

Model Dataset	Specifications	File Format	Period	Coordinate System	Source
Land Station Air Quality	Ground-based air quality monitoring stations providing hourly PM2.5 concentrations.	.xlsx	2023	EPSG:4326	SAAQIS
ERA5 Hourly Climate	Hourly reanalysis data including temperature, humidity, wind speed, and pressure to model pollutant dispersion.	.grib	2023	EPSG:4326	Copernicus

South African Municipality Administration Boundary	Polygon shapefile delineating administrative regions for correlating air quality with local governance and policies.	.shp	2023	EPSG:4326	Humanitarian Data Exchange
MODIS/Terra+Aqua MAIAC Land Aerosol Optical Depth Daily L2G (MCD19A2)	Daily aerosol optical depth (AOD) data at 1 km resolution to estimate surface PM2.5 concentrations.	.hdf	2023	EPSG:4326	NASA
Gert Sibande District population density	Spatial, high-resolution approach to disaggregate census counts for the study area.	.tiff	2023		Landscan

4.1.1. Air Quality Data

Hourly SO2, NOx, $PM_{2.5}$, CO were collected from ground-based air quality monitoring stations [43], ensuring accurate representation of real-time pollution levels. This dataset, provided by the South African Air Quality Information System (SAAQIS), was stored in .xlsx format for the year 2023 using the EPSG:4326 coordinate system.

4.1.2. ERA5 Hourly Climate Data

Climate variables such as temperature, humidity, wind speed, and pressure were extracted from Copernicus ERA5 hourly reanalysis data. This information, stored in .grib format, facilitated modelling pollutant dispersion patterns across the study area which was consistent with Wang et al's conclusions that the reanalysis data performs well in developing communities such as SA [44].

4.1.3. Population Density Data

High-resolution population density estimates for Gert Sibande District were derived from LandScan. This data, in .tiff format, enabled the disaggregation of census counts to refine spatial resolution for the study and was reported by Huang et al as a key for estimating $PM_{2.5}$.

4.1.4. MODIS/Terra+Aqua MAIAC AOD Data

Daily aerosol optical depth (AOD) data at 1km resolution were obtained from NASA's MODIS/Terra+Aqua MAIAC dataset (MCD19A2) in .hdf format. AOD data was included as a crucial predictor to estimate surface-level $PM_{2.5}$ concentrations.

4.1.5. Ward Administration Boundaries

Polygon shapefiles delineating administrative boundaries were sourced from the Humanitarian Data Exchange. These files, in .shp format with the EPSG:4326 coordinate system, provided spatial correlation between air quality and local governance structures.

4.2. Data Analysis Methods

The data analysis method employed a structured approach to estimate $PM_{2.5}$ concentrations using a machine learning-based predictive model. The XGBoost algorithm, known for its efficiency and accuracy in handling large, heterogeneous datasets, was used for this purpose.

The analysis integrated diverse spatial and temporal data sources to enhance model precision.

4.2.1. Data Preprocessing

Raw datasets, including PM2.5 measurements, climate reanalysis data, aerosol optical depth (AOD), administrative boundaries, and population density, were harmonised to a common spatial resolution (EPSG:4326) and temporal format. The ground station air quality data provided baseline hourly PM2.5 concentrations. Missing values and anomalies in the datasets were addressed using interpolation and data imputation techniques. For spatial data, geospatial operations ensured proper alignment of layers [31].

4.2.2. Feature Engineering

Key features were engineered to enhance the model's predictive performance. Aerosol Optical Depth (AOD), derived from satellite imagery, was included as it is strongly correlated with surface-level PM2.5 concentrations. Meteorological variables, including temperature, wind speed, humidity, and atmospheric pressure, were integrated to account for atmospheric dispersion of pollutants. Population density data provided a proxy for anthropogenic emissions, while temporal features such as seasonal indicators captured variations in air quality over time. These features were selected based on their theoretical and empirical relevance to PM2.5 levels.

4.2.3. Model Training and Evaluation

The XGBoost algorithm was employed due to its scalability, handling of missing data, and ability to capture complex feature interactions. The model was trained using the

preprocessed datasets, with PM2.5 concentrations as the dependent variable. A supervised learning framework was utilised, splitting the data into training and testing subsets to evaluate generalisability. The model was assessed using cross-validation to mitigate overfitting, with RMSE illustrated in [48, eq.(1)] and R2 as seen in [48, eq.(2)] to quantify predictive accuracy and model performance [48].

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

$$R^2 = 1 - (\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / \sum_{i=1}^{n} (y_i - \underline{y}_i)^2) \quad (2)$$

4.2.4. Spatial and Temporal Modelling

Predicted PM2.5 concentrations were visualised using geospatial tools, such as Kepler.gl, to identify spatial and temporal patterns. Spatial analysis highlighted hotspots of elevated PM2.5 levels, providing insights into pollution distribution across administrative boundaries. Temporal modelling captured seasonal and diurnal trends, enabling the evaluation of pollution dynamics over time [2, 10]. These analyses supported targeted interventions by identifying regions and periods of concern.

4.2.5. Variance Analysis and Policy Recommendations

The model results were subjected to variance analysis to compute percentage deviations between observed and predicted PM2.5 levels. This analysis was crucial for identifying discrepancies and validating model reliability. High-risk municipalities where predicted PM2.5 exceeded World Health Organisation (WHO) thresholds were flagged for policy recommendations [22]. Recommendations included limiting outdoor exposure as noted by Ha et al. [24], and improving air quality monitoring infrastructure in these regions [25].

4.3. Integration and Application

The digital twin model illustrated in Fig 3. integrates data from multiple disparate sources and offers real-time predictions of $PM_{2.5}$ concentrations. A policy recommendation model is created which policy-makers can incorporate into antenatal care program in the Highveld region, allowing healthcare professionals to provide timely recommendations to pregnant women, and allow for the visualisation of the impact of changes before implementation which El-Agamy et al. highlighted is one of the fundamental applications of digital twins in smart cities [49].

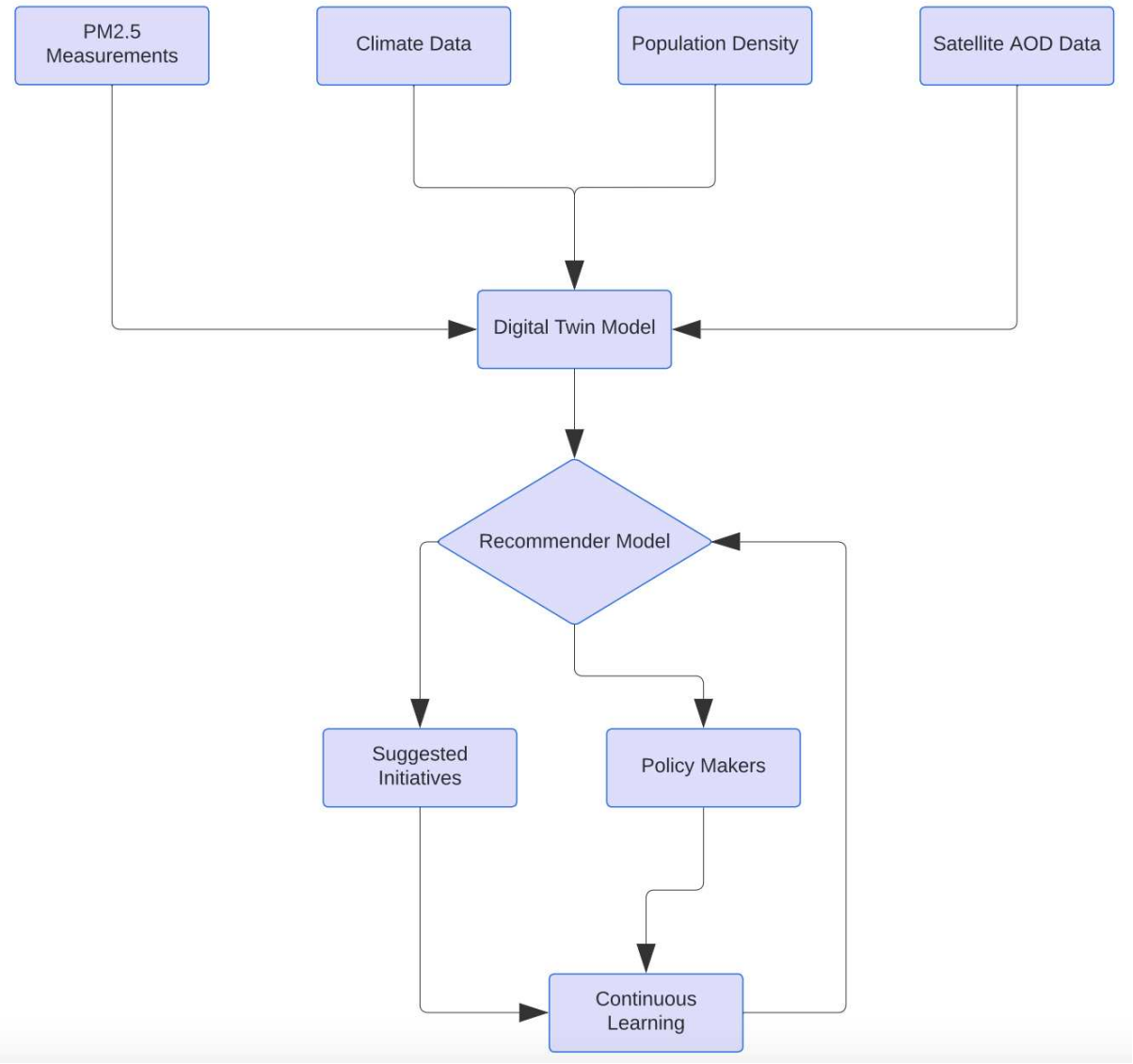

Fig 3: Digital Twin Model design with integrated data sources. The digital twin model integrates PM2.5 measurements, climate data, population density, and satellite AOD data to predict air quality. Insights from the model are fed into a recommender system.

The digital twin will also simulate different pollution scenarios, enabling policymakers to evaluate the effectiveness of air quality regulations and propose new interventions.

By combining machine learning, GIS, and real-time data in a digital twin framework, this study aims to develop a comprehensive tool for air quality management and antenatal care in South Africa. The model's predictions will be tested against real-world health outcomes to ensure accuracy and reliability.

4.4. Ethical Considerations

In this study, high ethical standards were maintained to ensure the responsible use of data while safeguarding privacy and confidentiality. The research focuses exclusively on population-level data related to women of reproductive age within the study area. Importantly, the study avoids any use of sensitive personal health information by relying solely on aggregated and anonymised datasets, which ensures that individual identities remain untraceable. According to international research protocols, such as the General Data Protection Regulation (GDPR) and guidelines from the WHO, anonymisation of data mitigates the risks of violating participant privacy while enabling scientifically meaningful insights [50], [51].

The digital twin model, which is central to this research, does not incorporate personal identifiers or individual-level health records. Instead, the model integrates non-sensitive and publicly available datasets from trusted sources, such as Landscan (for population density data) and SAAQIS (for air quality measurements). These geospatial and environmental datasets facilitate the analysis of $PM_{2.5}$ exposure and its relationship to antenatal outcomes without compromising privacy. As highlighted by Abutalip et al. [22], the ethical use of aggregated data is crucial for maintaining public trust, particularly in digital twin technologies where environmental and demographic data intersect with health outcomes.

To address potential data biases, such as inconsistencies in air quality measurements or missing population statistics, a cross-validation approach was employed. Data from multiple sources, including AOD, ground-based monitoring stations, and geospatial population datasets, were compared and validated to minimise inaccuracies. This approach aligns with

ethical data handling principles outlined by OECD guidelines for research transparency and reproducibility [52]. By mitigating gaps and ensuring data robustness, the study enhances the reliability of its findings while adhering to ethical research standards.

In compliance with institutional and national ethical frameworks, a self-assessment form was completed, and confirmation was sought from the relevant authorities to ensure the ethical use of population density and environmental data. This step further verifies that the study adheres to established ethical protocols and meets the necessary standards for research involving population-level data, even in the absence of personal health information. As digital twin models continue to evolve, addressing such ethical considerations ensures they are developed in ways that balance scientific innovation with privacy protection, as emphasised by Rogers et al. [53].

By prioritising data anonymity, using public and aggregated datasets, and validating data accuracy, this study upholds ethical standards in research. The careful handling of geospatial and environmental data ensures that findings contribute meaningfully to antenatal health modelling while maintaining public trust and privacy.

5. Results and Discussions

5.1. Results

The XGBoost model, trained on 34,738 observations integrating satellite data, ground-based stations, and climate data, achieved robust predictive performance. The scatter plot (Fig. 4) illustrates the predictive performance of the optimised XGBoost model. A strong alignment between predicted and actual PM2.5 values can be observed, particularly for lower and mid-range concentrations. The red dashed line (y=x) serves as the ideal reference, and the majority of data points cluster closely around this line, indicating a high degree of model accuracy.

However, slight deviations at higher PM2.5 concentrations (above 150 μg/m³) suggest that the model slightly underestimates extreme pollution levels. This is a known limitation in machine learning models when handling outliers or regions with limited data points in extreme ranges [30]. Despite this, the model demonstrates robust performance overall, achieving a Root Mean Square Error (RMSE) of 5.8004 and an R_2 score of 90.04%, as previously discussed. Compared to studies by Zhang et al. [20] and Huang et al. [46], where random forest models were applied to estimate $PM_{2.5}$ levels in the South African Highveld and North China, this study highlights both methodological advancements and specific strengths in addressing spatial variability in regions with limited monitoring infrastructure.

The observed scatter highlights the ability of the model to generalise well across the dataset, capturing key patterns while minimising errors. Future enhancements, such as fine-tuning for extreme values or incorporating additional features like NO_2 and emission inventories, could further improve predictions in high-pollution scenarios [28] [31].

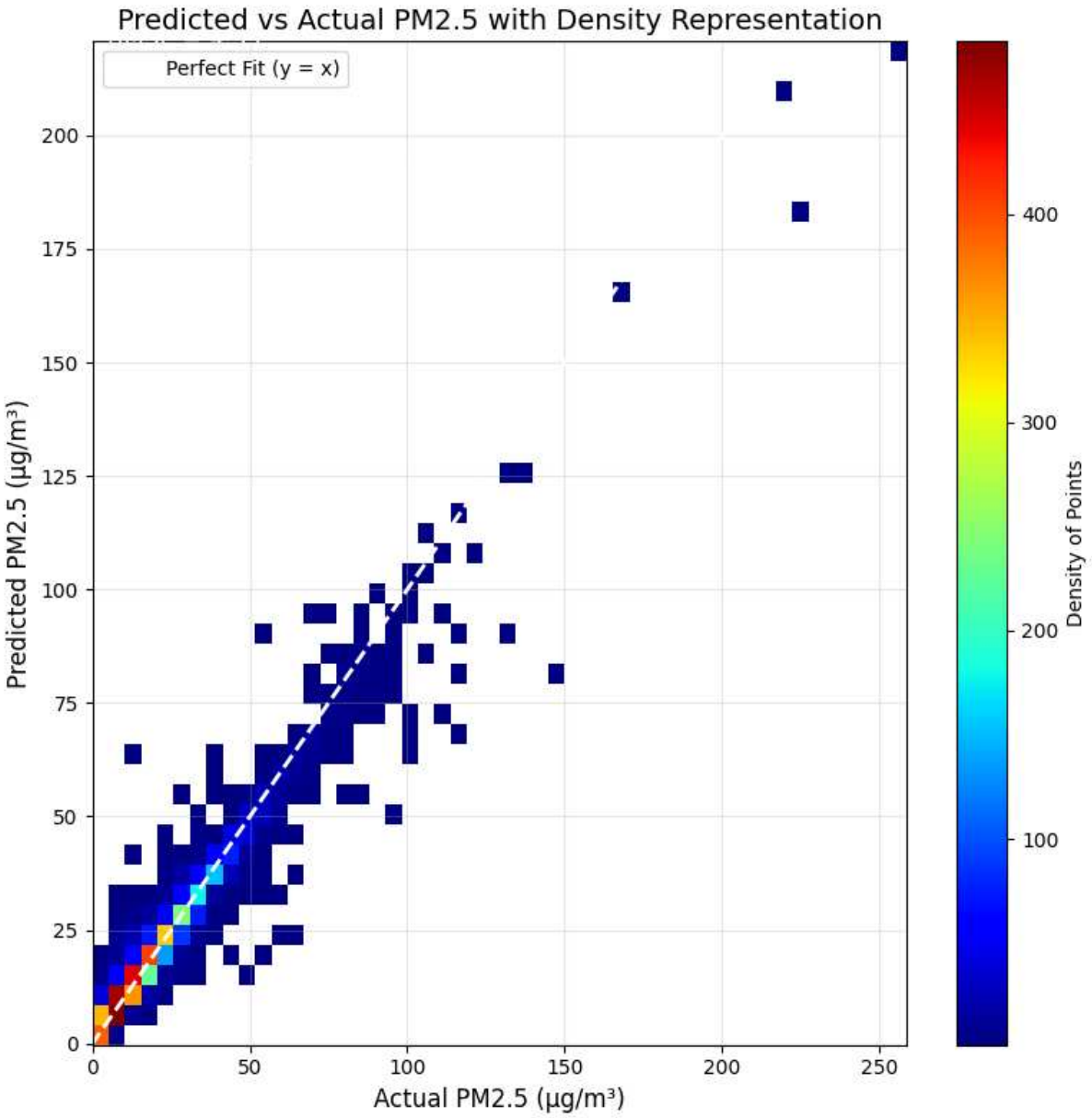

Fig 4: Density Plot of Predicted vs Actual PM2.5 Concentrations. The chart visualises the relationship between actual and predicted PM2.5 values using a density-based 2D histogram. The colour gradient represents the density of overlapping points

The RMSE underscores the model's precision, reflecting the average deviation of predictions from observed values, while the R_2 value highlights its reliability. Zhang et al., focusing on Gauteng Province, achieved an R^2 of 0.80 and RMSE of 9.40 µg/m³, using key predictors like satellite AOD, precipitation, and population density. While effective for spatial estimates, their model showed moderate accuracy.

Huang et al. reconstructed historical $PM_{2.5}$ concentrations in North China Plain, achieving a cross-validation R^2 of 0.88 and an RMSE of 15.06 μg/m³, with temporal validation down to monthly scales. Although their imputation of AOD improved robustness over extended periods, the current study's focus on short-term spatial variability and the inclusion of wind patterns, boundary layer height, and chemical pollutants enabled greater precision.

Despite superior accuracy, this study, like Zhang et al. and Huang et al., showed underestimation of extreme $PM_{2.5}$ values due to data sparsity. Addressing this limitation through temporal generalisation, imputation methods, and additional features like land-use and emission inventories could enhance future performance. This study advances prior work by providing an actionable framework with high spatial accuracy, supporting real-time healthcare interventions.

5.2. Feature Importance

The feature importance analysis highlights the top contributors to the PM2.5 prediction model, ranked based on their F-score, which represents their relative contribution to the XGBoost model's performance. The top 15 features are ranked in Table 2 below.

Table 2: Top 15 Features Ranked by F-Score in the PM2.5 Prediction Model. The table lists the most influential features in the XGBoost model, ranked by their importance scores (F-Score)

Rank	Feature	F-Score
1	v10 (Wind V-Direction)	15793
2	u10 (Wind U-Direction)	14995
3	blh (Boundary Layer Height)	14067

4	tco3 (Total Column Ozone)	13927
5	SO2 (Sulphur Dioxide)	13314
6	O3 (Ground-Level Ozone)	13363
7	t2m (Temperature at 2m)	13130
8	H2S (Hydrogen Sulfide)	12249
9	NO2 (Nitrogen Dioxide)	10863
10	NOX (Nitrogen Oxides)	10362
11	hour_month_interaction	8880
12	hour_day_interaction	8036
13	hour	6553
14	day_month_interaction	5757
15	month	3238

The feature importance chart highlights the key variables used in predicting $PM_{2.5}$ concentrations, ranked by their contribution to the XGBoost model:

1. **Wind V-Direction (v10) and U-Direction (u10)**: Wind patterns play a dominant role in pollutant transport and dispersion. Both horizontal (u10) and vertical (v10) wind components emerged as the most important features, consistent with prior studies on atmospheric dynamics [27, 29]. Wind speeds influence how pollutants like $PM_{2.5}$ accumulate or disperse across regions, particularly in industrial zones such as the Highveld Priority Area.

2. **Boundary Layer Height (BLH)**: BLH, with an F-score of 14,067.0, significantly influences vertical mixing and dispersion of air pollutants. Shallow boundary layers restrict pollutant dispersion, causing PM2.5 concentrations to rise, particularly during calm atmospheric conditions [30].
3. **Total Column Ozone (TCO3) and Ground-Level Ozone (O3)**: Ozone, both in the upper atmosphere (tco3) and near the surface (O3), interacts with PM2.5 as a secondary pollutant, influencing its formation and chemical transformation [28]. These findings align with chemical transport models, which highlight ozone's role in particulate matter formation.
4. **Sulphur Dioxide (SO_2) and Hydrogen Sulfide (H_2S)**: These industrial emissions are key precursors to secondary aerosol formation, particularly in coal-dominated regions like Mpumalanga [31]. The significant importance of SO_2 and H_2S underscores the role of industrial activity in driving $PM_{2.5}$ levels.
5. **Temperature at 2 Meters (T2M)**: Temperature influences atmospheric mixing and the chemical processes that govern particulate matter formation. Elevated temperatures often accelerate secondary pollutant formation, consistent with climate-sensitive air quality studies [32].
6. **Nitrogen Oxides (NO_2 and NOX)**: NO_2 and NOX are critical pollutants emitted from vehicles and industrial processes. These gases contribute to the formation of secondary $PM_{2.5}$ through photochemical reactions [28].
7. **Interaction Features (hour_month, hour_day)**: Interaction terms such as **hour-month** and **hour-day** reflect temporal variability in $PM_{2.5}$ concentrations. This aligns

with studies showing that $PM_{2.5}$ levels fluctuate with daily traffic patterns, industrial emissions, and meteorological changes [26].

8. **Hour and Month**: Temporal variables like **hour** and **month** capture diurnal and seasonal variations in $PM_{2.5}$. These variations are influenced by factors such as heating systems in winter, atmospheric stability, and seasonal wind patterns.

These results align with the research hypothesis that integrating meteorological [20], chemical [20, 47], and geospatial data improves the accuracy of $PM_{2.5}$ predictions [22]. The importance of wind parameters (u10 and v10) underscores the dynamic nature of air pollution [46, 47], while the inclusion of both ground-level and satellite-derived variables ensures a holistic modeling approach. For future work, the integration of additional pollutants, such as NO_2, and enhancing temporal forecasting capabilities could further improve the model's performance.

5.3. Model Optimisation

The optimisation process revealed that the best-performing parameters included a maximum depth of 10, a learning rate of 0.7074, 540 estimators, and colsample_bytree and subsample values of 0.98656 and 0.73614, respectively. These parameters underscore the importance of controlling overfitting through moderate depth and leveraging substantial subsampling for robust generalisation. The improved RMSE of 5.8001 represents a significant enhancement from the initial evaluation metrics, showcasing the efficacy of Bayesian optimisation via Optuna [54].

5.4. Comparative Analysis

The performance of the XGBoost model aligns with findings from similar studies. Zhang et al. [20] demonstrated that gradient boosting methods achieve comparable RMSE values (0.10–0.12) when predicting PM2.5 concentrations in regions with diverse environmental conditions. The model's high R^2 score also mirrors results in high-resolution geospatial studies where machine learning effectively captures the spatial and temporal variability of air pollution [20, 47]. Unlike traditional linear regression approaches, which often underperform with non-linear data, XGBoost capitalises on interactions between variables, thereby aligning with the complex nature of $PM_{2.5}$ dispersion patterns [46].

5.5. Challenges with Dataset

Despite the model's success, significant challenges were encountered during data preparation and training. The dataset contained inconsistencies between satellite AOD readings and ground-level measurements, requiring extensive cleaning and interpolation. Similar issues were flagged by Wang et al. [2], who noted that large environmental datasets are often noisy due to measurement discrepancies and missing data. Handling these limitations required rigorous pre-processing, including feature engineering to account for meteorological variability.

The computational demands of optimising hyperparameters for large datasets were also noteworthy. Bayesian methods, while efficient, required significant computational resources to evaluate hundreds of parameter combinations. This highlights the trade-off between achieving optimal performance and computational feasibility in resource-limited contexts.

5.6. Response to Hypothesis

The Healthcare Air Quality Digital Twin, developed as part of this study, integrates $PM_{2.5}$ predictions with real-time spatial visualisation and health policy recommendations. The tool in Fig. 5 below supports the study's hypothesis that leveraging air quality data through digital twin technology can enhance maternal health interventions, particularly in regions with high pollution levels.

Fig 5: : PM2.5 Spatial Distribution Visualisation using the Streamlit Application Model. The map highlights PM2.5 concentration polygons for Msukaligwa Local Municipality, classified into five ranges (6.49–113.29 µg/m³).

The digital twin provides a dynamic spatial representation of PM2.5 levels across monitored regions, focusing on Govan Mbeki Local Municipality and surrounding areas (Figure 8). The gauge chart highlights an average $PM_{2.5}$ concentration of 23.9 µg/m³, which significantly exceeds the 15 µg/m³ threshold recommended by the World Health Organisation (WHO) for long-term exposure [1].

5.7. Policy Recommendations

Based on the findings, several policy interventions are recommended:

1. **Stricter Air Quality Monitoring and Enforcement:** South Africa's industrial hubs, such as Mpumalanga, should adopt stricter emission control policies. Continuous air quality monitoring, supported by machine learning models, can guide regulatory actions in real-time [12].
2. **Integration of Environmental Data in Healthcare:** Antenatal care programs should incorporate $PM_{2.5}$ exposure data into patient risk assessments. This aligns with recommendations by Bjornsson et al. [38], who emphasised the importance of personalised healthcare through data triangulation.
3. **Investment in Digital Twin Infrastructure:** Governments and healthcare systems should invest in digital twin platforms for real-time air quality monitoring. Abutalip et al. and Zaballos et al.[22, 4] argue that such systems are cost-effective, scalable, and critical for data-driven decision-making in resource-constrained settings.
4. **Public Awareness Campaigns:** Community education on the health risks of air pollution during pregnancy can empower individuals to adopt protective measures, such as avoiding outdoor exposure during peak pollution hours which aligns with the WHO strategies on advocacy and outreach[24].

5.8. Limitations

While the study successfully demonstrates the utility of a digital twin for $PM_{2.5}$ prediction and maternal health interventions, several limitations must be acknowledged.

The reliance on $PM_{2.5}$ as the sole pollutant metric restricts the comprehensiveness of the model. Although $PM_{2.5}$ is widely regarded as a key determinant of air quality and adverse health outcomes, other pollutants, such as sulphur dioxide (SO_2), nitrogen dioxide (NO_2), and volatile organic compounds (VOCs), also play significant roles in health risks [55]. For instance, Wu et al. demonstrated that NO_2 and SO_2 are critical contributors to respiratory illnesses and congenital anomalies when combined with $PM_{2.5}$ exposure [23] while Lai et al [59]. emphasised the importance of VOCs in determining air quality. The exclusion of these pollutants may lead to an underestimation of the cumulative effects of air pollution, particularly in industrial regions like Mpumalanga.

The geographic focus on Mpumalanga introduces limitations in the generalisability of the findings. Mpumalanga, as an industrial hub, is characterised by unique pollution patterns dominated by coal mining and petrochemical industries. This specificity may reduce the model's applicability to other geographic regions with differing pollution sources and environmental conditions, such as urban centres dominated by vehicular emissions or rural areas reliant on biomass burning. Validation across diverse geographic and socioeconomic contexts is necessary to ensure robustness and scalability [46, 47].

The absence of temporal forecasting capabilities in the current model constrains its ability to enable proactive healthcare planning. Predictive models that forecast $PM_{2.5}$ levels over short- and long-term periods are critical for issuing timely interventions, such as alerting vulnerable populations to stay indoors during peak pollution hours. Hao et al. highlighted the importance of temporal forecasts in enabling adaptive healthcare strategies, particularly for pregnant women and children [17]. The current static approach limits the digital twin's ability to anticipate pollution trends and inform decision-making ahead of time.

The study faced computational constraints during data preprocessing and model training. Processing large, multi-source datasets (e.g., satellite AOD, ground-based measurements, and meteorological features) required significant computational resources, particularly for coordinate transformation, spatial joins, and null imputation steps. These intensive preprocessing requirements limited the ability to train the model on the full dataset, necessitating the use of a smaller subset of 34,738 observations for optimisation. While the final model demonstrated strong performance, training on larger datasets could potentially improve accuracy and generalisability by capturing more complex relationships and edge cases [20].

6. Conclusions

This research demonstrates the potential of integrating machine learning and digital twin technology to address urban air quality management challenges, particularly in antenatal care programs. The high accuracy of the XGBoost model (RMSE = 5.8004, R^2=90.04) in predicting $PM_{2.5}$ concentrations validates the hypothesis and underscores the feasibility of incorporating environmental data into healthcare decision-making frameworks. By addressing spatial and temporal variability, the findings offer actionable insights for healthcare providers and policymakers in LMICs, aligning with previous research on digital twins for environmental health monitoring [22, 23].

The study identified critical features such as wind parameters, boundary layer height, and chemical pollutants, consistent with prior findings on air quality modelling [20]. The interactive digital twin model visualised pollution hotspots, enabling real-time recommendations to mitigate exposure risks. However, limitations such as the reliance on $PM_{2.5}$ as the sole pollutant, the absence of temporal forecasting, and geographic constraints highlight areas for future research, including IoT-enabled monitoring and validation across diverse contexts [17].

By integrating real-time data with predictive analytics, this research supports the use of digital twins to improve maternal health outcomes and informs policy interventions for mitigating air pollution impacts in LMICs.

6.1. Future Research

The study demonstrates a robust digital twin framework for monitoring $PM_{2.5}$ levels and informing maternal healthcare interventions; several avenues for future research and

development remain. Addressing these areas will significantly improve the comprehensiveness, scalability, and predictive utility of the model.

The current model provides a static representation of $PM_{2.5}$ levels, limiting its ability to enable proactive healthcare interventions. Future research should focus on integrating temporal forecasting techniques to predict pollution levels over short- and long-term periods. Methods such as recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) networks have shown promise in capturing temporal dependencies in air quality data [57]. Predictive capabilities would allow policymakers and healthcare providers to issue early warnings and mitigate exposure risks for vulnerable populations, such as pregnant women, during periods of peak pollution.

The geographic focus on Mpumalanga limits the generalisability of the current findings to regions with differing pollution sources, climate conditions, and socioeconomic factors. Future work should validate the model across diverse settings, including urban areas dominated by vehicular emissions, rural Areas reliant on biomass burning, and industrial regions with varying pollutant compositions.

Studies such as Abutalip et al. [22] and Zhang et al.[20] highlight the importance of validating air quality models in multiple geographic and socioeconomic contexts to ensure robustness and scalability. Such validation will also provide insights into regional differences in pollution dynamics and their health impacts.

The inclusion of IoT-enabled real-time sensors can enhance the accuracy, responsiveness, and granularity of the digital twin platform. Low-cost IoT devices deployed in targeted locations can provide continuous, real-time measurements of PM2.5 and other pollutants, addressing gaps in spatial and temporal coverage. Research by Cárdenas-León et al. and Li

et al. [2, 54] has shown that combining IoT data streams with machine learning models significantly improves real-time air quality predictions. By integrating IoT devices, the digital twin can adapt dynamically to real-world conditions, increasing its utility for decision-making and public health interventions.

Future work should improve the usability and accessibility of the digital twin by developing notification systems for vulnerable populations (e.g., pregnant women) via SMS or mobile applications during high pollution events, and integration of additional maternal and neonatal health indicators, enabling a more holistic risk assessment.

Such enhancements will ensure the digital twin's outputs are actionable, user-friendly, and accessible to both policymakers and non-technical stakeholders, particularly in low- and middle-income countries (LMICs).

7. References

[1] J. M. Diaz-Sarachaga, "May urban digital twins spur the New Urban Agenda? The Spanish case study," *Sustainable Cities and Society*, vol. 114, pp. 105788–105788, Aug. 2024, doi: https://doi.org/10.1016/j.scs.2024.105788/

[2] I. Cárdenas-León, M. Koeva, P. Nourian, and C. Davey, "Urban digital twin-based solution using geospatial information for solid waste management," *Sustainable Cities and Society*, vol. 115, p. 105798, Nov. 2024, doi: https://doi.org/10.1016/j.scs.2024.105798/

[3] A. Zaballos, A. Briones, A. Massa, P. Centelles, and V. Caballero, "A Smart Campus' Digital Twin for Sustainable Comfort Monitoring," *Sustainability*, vol. 12, no. 21, p. 9196, Nov. 2020, doi: https://doi.org/10.3390/su12219196/

[4] Z. Liang, Y. Jin, J. Singh, and A. Khan, "Demo: BuildTwin: Towards Real-Time High-Fidelity Digital Twin for Smart Building Management," Oct. 2023, doi: https://doi.org/10.1109/icnp59255.2023.10355596/

[5] H. Ritchie, "Deaths from air pollution are high, but the data contains hope," *Clean Air Fund*, Jan. 29, 2024. https://www.cleanairfund.org/news-item/deaths-air-pollution-data-hope/#:~:text=The%20World%20Health%20Organization%20estimates/

[6] T. Wainstock, I. Yoles, R. Sergienko, I. Kloog, and E. Sheiner, "Prenatal particulate matter exposure and Intrauterine Fetal Death," *International Journal of Hygiene and Environmental Health*, vol. 234, p. 113720, May 2021, doi: https://doi.org/10.1016/j.ijheh.2021.113720/

[7] J. Rentschler and N. Leonova, "Global air pollution exposure and poverty," *Nature Communications*, vol. 14, no. 1, p. 4432, Jul. 2023, doi: https://doi.org/10.1038/s41467-023-39797-4/

[8] N. Zhu, X. Ji, X. Geng, H. Yue, G. Li, and N. Sang, "Maternal PM2.5 exposure and abnormal placental nutrient transport," *Ecotoxicology and Environmental Safety*, vol. 207, p. 111281, Jan. 2021, doi: https://doi.org/10.1016/j.ecoenv.2020.111281/

[9] C. Weil , S. E. Bibri , R. Longchamp, F. Golay, and A. Alahi, "Urban Digital Twin Challenges: A Systematic Review and Perspectives for Sustainable Smart Cities," *Sustainable Cities and Society*, vol. 99, p. 104862, Dec. 2023, doi: https://doi.org/10.1016/j.scs.2023.104862/

[10] J. J. Frostad *et al.*, "Mapping development and health effects of cooking with solid fuels in low-income and middle-income countries, 2000–18: a geospatial modelling study," *The Lancet Global Health*, vol. 10, no. 10, pp. e1395–e1411, Oct. 2022, doi: https://doi.org/10.1016/s2214-109x(22)00332-1/

[11] M. H. L. Mabaso, T. Ndaba, and Z. L. Mkhize-Kwitshana, "Overview of Maternal, Neonatal and Child Deaths in South Africa: Challenges, Opportunities, Progress and Future Prospects," *International journal of MCH and AIDS*, vol. 2, no. 2, pp. 182–9, 2014, Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4948143/

[12] Clean Air Fund, "South Africa," *Clean Air Fund*. https://www.cleanairfund.org/geography/south-africa/

[13] Unicef, "Air pollution accounted for some 3,365 deaths of children under five years across South Africa in 2021," *Unicef.org*, 2021. https://www.unicef.org/southafrica/press-releases/air-pollution-accounted-some-3365-deaths-children-under-five-years-across-south/

[14] P. J. Landrigan and R. A. Fuller, "Environmental pollution: An enormous and invisible burden on health systems in low- and middle-income counties.," *World hospitals and health services*, vol. 50, no. 4, pp. 35–40, Jan. 2014.

[15] D. A. Millar *et al.*, "Respiratory health among adolescents living in the Highveld Air Pollution Priority Area in South Africa," *BMC Public Health*, vol. 22, no. 1, Nov. 2022, doi: https://doi.org/10.1186/s12889-022-14497-8/

[16] X. Xu *et al.*, "Maternal PM2.5 exposure during gestation and offspring neurodevelopment: Findings from a prospective birth cohort study," *Science of The Total Environment*, vol. 842, p. 156778, Oct. 2022, doi: https://doi.org/10.1016/j.scitotenv.2022.156778/

[17] H. Hao *et al.*, "Effects of air pollution on adverse birth outcomes and pregnancy complications in the U.S. state of Kansas (2000–2015)," *Scientific Reports*, vol. 13, no. 1, p. 21476, Dec. 2023, doi: https://doi.org/10.1038/s41598-023-48329-5/

[18] R. Garland, "The air in South African Highveld cities smells foul in the winter: here's why," *The Conversation*, Jun. 19, 2022. https://theconversation.com/the-air-in-south-african-highveld-cities-smells-foul-in-the-winter-heres-why-184884/ (accessed Oct. 16, 2024).

[19] S. Janjua, P. Powell, R. Atkinson, E. Stovold, and R. Fortescue, "Individual-level interventions to reduce personal exposure to outdoor air pollution and their effects on people with long-term respiratory conditions," *Cochrane Database of Systematic Reviews*, vol. 2021, no. 8, Aug. 2021, doi: https://doi.org/10.1002/14651858.cd013441.pub2/

[20] D. Zhang *et al.*, "A machine learning model to estimate ambient PM2.5 concentrations in industrialized highveld region of South Africa," *Remote Sensing of Environment*, vol. 266, p. 112713, Dec. 2021, doi: https://doi.org/10.1016/j.rse.2021.112713/

[21] E. Katsoulakis *et al.*, "Digital twins for health: a scoping review," *npj Digital Medicine*, vol. 7, no. 1, pp. 1–11, Mar. 2024, doi: https://doi.org/10.1038/s41746-024-01073-0/

[22] K. Abutalip, A. Al-lahham, and A. E. Saddik, "Digital Twin of Atmospheric Environment: Sensory Data Fusion for High-Resolution $PM_{2.5}$ Estimation and Action Policies Recommendation," *IEEE Access*, pp. 1–1, 2023, doi: https://doi.org/10.1109/access.2023.3236414/

[23] H. Wu, P. Ji, H. Ma, and L. Xing, "A Comprehensive Review of Digital Twin from the Perspective of Total Process: Data, Models, Networks and Applications," *Sensors*, vol. 23, no. 19, p. 8306, Jan. 2023, doi: https://doi.org/10.3390/s23198306/

[24] S. Ha *et al.*, "Air Pollution Exposure Monitoring among Pregnant Women with and without Asthma," *International Journal of Environmental Research and Public Health*, vol. 17, no. 13, p. 4888, Jul. 2020, doi: https://doi.org/10.3390/ijerph17134888/

[25] J. Sacks, N. Fann, S. Gumy, I. Kim, G. Ruggeri, and P. Mudu, "Quantifying the Public Health Benefits of Reducing Air Pollution: Critically Assessing the Features and Capabilities of WHO's AirQ+ and U.S. EPA's Environmental Benefits Mapping and Analysis Program—Community Edition (BenMAP—CE)," *Atmosphere*, vol. 11, no. 5, p. 516, May 2020, doi: https://doi.org/10.3390/atmos11050516/

[26] UN Environment Programme, "Air pollution linked to 'huge' reduction in intelligence," *UNEP*, Oct. 11, 2018. https://www.unep.org/news-and-stories/story/air-pollution-linked-huge-reduction-intelligence/

[27] Y. S. Park, L. Konge, and A. R. Artino, "The Positivism Paradigm of Research," *Academic Medicine*, vol. 95, no. 5, pp. 690–694, 2020, doi: https://doi.org/10.1097/ACM.0000000000003093/

[28] J. Bergmann, "Research Philosophy, Methodological Implications, and Research Design," *Studien zur Migrations- und Integrationspolitik*, pp. 57–89, Oct. 2023, doi: https://doi.org/10.1007/978-3-658-42298-1_3/

[29] S. M. Kim, "Inductive or deductive? Research by maxillofacial surgeons," *Journal of the Korean Association of Oral and Maxillofacial Surgeons*, vol. 47, no. 3, pp. 151–152, 2021, doi: https://doi.org/10.5125/jkaoms.2021.47.3.151/

[30] V. Kaushik and C. A. Walsh, "Pragmatism as a Research Paradigm and Its Implications for Social Work Research," *Social Sciences*, vol. 8, no. 9, pp. 1–17, 2019, doi: https://doi.org/10.3390/socsci8090255/

[31] C.-C. Chen, Yin Ru Wang, Hung Yi Yeh, Tang Huang Lin, Chun Sheng Huang, and Chang Fu Wu, "Estimating monthly PM2.5 concentrations from satellite remote sensing data, meteorological variables, and land use data using ensemble statistical modeling and a random forest approach," *Environmental Pollution*, vol. 291, pp. 118159–118159, Dec. 2021, doi: https://doi.org/10.1016/j.envpol.2021.118159/

[32] V. Katoch *et al.*, "Addressing Biases in Ambient $PM_{2.5}$ Exposure and Associated Health Burden Estimates by Filling Satellite AOD Retrieval Gaps over India," *Environmental Science & Technology*, vol. 57, no. 48, pp. 19190–19201, Nov. 2023, doi: https://doi.org/10.1021/acs.est.3c03355/

[33] En Xin Neo *et al.*, "Towards Integrated Air Pollution Monitoring and Health Impact Assessment Using Federated Learning: A Systematic Review," *Frontiers in Public Health*, vol. 10, May 2022, doi: https://doi.org/10.3389/fpubh.2022.851553/

[34] M. Balali-Mood, A. Ghorani-Azam, and B. Riahi-Zanjani, "Effects of air pollution on human health and practical measures for prevention in Iran," *Journal of Research in Medical Sciences*, vol. 21, no. 1, p. 65, 2016, doi: https://doi.org/10.4103/1735-1995.189646/

[35] M. Lippmann, *Environmental toxicants : human exposures and their health effects*. Hoboken, N.J.: John Wiley & Sons, 2009.

[36] A. Vallée, "Digital twin for healthcare systems," *Frontiers in digital health*, vol. 5, Sep. 2023, doi: https://doi.org/10.3389/fdgth.2023.1253050/

[37] P. Rani and A. Dhok, "Effects of Pollution on Pregnancy and Infants," *Cureus*, vol. 15, no. 1, Jan. 2023, doi: https://doi.org/10.7759/cureus.33906/

[38] B. Björnsson *et al.*, "Digital twins to personalize medicine," *Genome Medicine*, vol. 12, no. 4, Dec. 2019, doi: https://doi.org/10.1186/s13073-019-0701-3/

[39] M. Lavietes, "The price of air pollution on American's healthcare," *World Economic Forum*, Jun. 2021. https://www.weforum.org/stories/2021/06/air-pollution-cost-america-healthcare-study/

[40] T. Chen and C. Guestrin, "XGBoost: a Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, pp. 785–794, 2016, doi: https://doi.org/10.1145/2939672.2939785/

[41] Q. Di *et al.*, "An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution," *Environment International*, vol. 130, p. 104909, Sep. 2019, doi: https://doi.org/10.1016/j.envint.2019.104909/

[42] B. Pan, "Application of XGBoost algorithm in hourly PM2.5 concentration prediction," *IOP Conference Series: Earth and Environmental Science*, vol. 113, p. 012127, Feb. 2018, doi: https://doi.org/10.1088/1755-1315/113/1/012127/

[43] SAAQIS, "Ambient air quality monitoring and data availability At a glance," 2024. Available: https://saaqis.environment.gov.za/Pagesfiles/Chapter%206%20Ambient%20Air%20Quality%20Monitoring.pdf/

[44] Z. Wang, P. Chen, R. Wang, Z. An, and L. Qiu, "Estimation of PM2.5 concentrations with high spatiotemporal resolution in Beijing using the ERA5 dataset and machine learning models," *Advances in Space Research*, vol. 71, no. 8, pp. 3150–3165, Apr. 2023, doi: https://doi.org/10.1016/j.asr.2022.12.016/

[45] X. Hu *et al.*, "Estimating ground-level PM2.5 concentrations in the Southeastern United States using MAIAC AOD retrievals and a two-stage model," *Remote Sensing of Environment*, vol. 140, pp. 220–232, Jan. 2014, doi: https://doi.org/10.1016/j.rse.2013.08.032/

[46] K. Huang *et al.*, "Predicting monthly high-resolution PM2.5 concentrations with random forest model in the North China Plain," *Environmental Pollution*, vol. 242, pp. 675–683, Nov. 2018, doi: https://doi.org/10.1016/j.envpol.2018.07.016/

[47] Z. Wang, Y. Zhou, R. Zhao, N. Wang, A. Biswas, and Z. Shi, "High-resolution prediction of the spatial distribution of PM2.5 concentrations in China using a long short-term memory model," *Journal of Cleaner Production*, vol. 297, p. 126493, May 2021, doi: https://doi.org/10.1016/j.jclepro.2021.126493/

[48] W. Y. Hong, D. Koh, and L. E. Yu, "Development and Evaluation of Statistical Models Based on Machine Learning Techniques for Estimating Particulate Matter ($PM_{2.5}$ and PM_{10}) Concentrations," *International journal of environmental research and public health*, vol. 19, no. 13, p. 7728, Autumn 2022, doi: https://doi.org/10.3390/ijerph19137728/

[49] R. F. El-Agamy, H. A. Sayed, A. Akhatatneh, Mansourah Aljohani, and Mostafa Elhosseini, "Comprehensive analysis of digital twins in smart cities: a 4200-paper bibliometric

study," *Artificial intelligence review*, vol. 57, no. 6, May 2024, doi: https://doi.org/10.1007/s10462-024-10781-8/

[50] C. J. Hoofnagle, B. V. D. Sloot, and F. Z. Borgesius, "The European Union general data protection regulation: what it is and what it means," *Information & Communications Technology Law*, vol. 28, no. 1, pp. 65–98, Feb. 2019, doi: https://doi.org/10.1080/13600834.2019.1573501/

[51] World Health organisation (WHO), "The protection of personal data in health information systems -principles and processes for public health," 2020. Available: https://iris.who.int/bitstream/handle/10665/341374/WHO-EURO-2021-1994-41749-57154-eng.pdf/

[52] D. Pilat and Y. Fukasaku, "OECD Principles and Guidelines for Access to Research Data from Public Funding," *Data Science Journal*, vol. 6, pp. OD4–OD11, 2007, doi: https://doi.org/10.2481/dsj.6.od4/

[53] E. Illman, "Avoiding growing pains in the development and use of digital twins," *Reuters*, Aug. 20, 2024. https://www.reuters.com/legal/legalindustry/avoiding-growing-pains-development-use-digital-twins-2024-08-20/

[54] J. Li *et al.*, "Application of XGBoost algorithm in the optimization of pollutant concentration," *Atmospheric Research*, vol. 276, pp. 106238–106238, Oct. 2022, doi: https://doi.org/10.1016/j.atmosres.2022.106238/

[55] E. David and V.-C. Niculescu, "Volatile Organic Compounds (VOCs) as Environmental Pollutants: Occurrence and Mitigation Using Nanomaterials," *International Journal of Environmental Research and Public Health*, vol. 18, no. 24, p. 13147, Dec. 2021, doi: https://doi.org/10.3390/ijerph182413147/

[56] Q. Chen, K. Shao, and S. Zhang, “Enhanced PM2.5 estimation across China: An AOD-independent two-stage approach incorporating improved spatiotemporal heterogeneity representations,” *Journal of Environmental Management*, vol. 368, p. 122107, Aug. 2024, doi: https://doi.org/10.1016/j.jenvman.2024.122107/

[57] X. Li *et al.*, “Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation,” *Environmental Pollution*, vol. 231, pp. 997–1004, Dec. 2017, doi: https://doi.org/10.1016/j.envpol.2017.08.114/

[58] X. Hu, L. A. Waller, A. Lyapustin, Y. Wang, and Y. Liu, “Improving satellite-driven PM2.5models with Moderate Resolution Imaging Spectroradiometer fire counts in the southeastern U.S.,” *Journal of Geophysical Research: Atmospheres*, vol. 119, no. 19, pp. 11, 375–11, 386, Oct. 2014, doi: https://doi.org/10.1002/2014jd021920/

[59] H. K. Lai *et al.*, “Personal exposures and microenvironment concentrations of PM2.5, VOC, NO2 and CO in Oxford, UK,” *Atmospheric Environment*, vol. 38, no. 37, pp. 6399–6410, Dec. 2004, doi: https://doi.org/10.1016/j.atmosenv.2004.07.013/

[60] V. Lebakula et al., LandScan Silver Edition [Data set], Oak Ridge National Laboratory, 2024. [Online]. Available: https://doi.org/10.48690/1531770/

[61] NASA, “MODIS/Terra+Aqua Land Aerosol Optical Depth Daily L2G Global 1km SIN Grid - LAADS DAAC,” NASA.gov, 2015. [Online]. Available: https://ladsweb.modaps.eosdis.nasa.gov/missions-and-measurements/products/MCD19A2#product-information/ . [Accessed: Dec. 10, 2024]

[62] SAAQIS, “SAAQIS,” saaqis.environment.gov.za, 2024. [Online]. Available: https://saaqis.environment.gov.za/. [Accessed: Dec. 10, 2024]

[63] H. Hersbach et al., “ERA5 hourly data on single levels from 1940 to present,” Copernicus Climate Change Service (C3S) Climate Data Store (CDS), 2023. [Online]. Available: https://doi.org/10.24381/cds.adbb2d47. [Accessed: Dec. 10, 2024].

[64] MAPOG, “Map Story,” Mapog.com, 2024. https://story.mapog.com/app/gisdata/south%20africa/Ward%20level%204 . [Accessed Dec. 10, 2024].

8. Appendices

1. Artefacts

Codebase

- Python code of model (XGBoost, digital twin integration, streamlit application)
- Readme file
- Comments for key functions

Datasets

- Zipped file of raw Landscan, SAAQIS, NASA, Copernicus, AOD (share the pre-processed version

Project Planning gantt chart
Streamlit app deployment Loom demo
Ethics Approval

- Self-assessment form
- Email responses from data stewards

Policy Recommendation

Table 1: Data Sources

Model Dataset	Specifications	Fields Extracted	File Format	Period	Coordinate System	Source (Link)
Land Station Air Quality	Ground-based air quality monitoring stations providing hourly PM2.5 concentrations.	Station name, PM2.5, SO_2, NO_2, NO, Nox, H2S Timestamp	.xlsx	2023	EPSG:4326	https://www.saaqis.org.za/

ERA5 Hourly Climate	Hourly reanalysis data including temperature, humidity, wind speed, and pressure to model pollutant dispersion.	Temperature, Wind Speed (U, V), Boundary Layer Height, Precipitation	.grib	2023	EPSG:4326	https://cds.climate.copernicus.eu/d levels?tab=download
South African Municipality Administration Boundary	Polygon shapefile delineating administrative regions for correlating air quality with local governance and policies.	Municipality Name, Administrative Boundaries geometry, longitude, latitude	.shp	2023	EPSG:4326	https://data.humdata.org/dataset/c
MODIS/Terra+Aqua MAIAC Land Aerosol Optical Depth Daily L2G (MCD19A2)	Daily aerosol optical depth (AOD) data at 1 km resolution to estimate surface PM2.5 concentrations.	AOD, Latitude, Longitude	.hdf	2023	EPSG:4326	https://ladsweb.modaps.eosdis.nasa -61
Gert Sibande District Population Density	Spatial, high-resolution approach to disaggregate census counts for the study area.	Population Count, Population Density, Coordinates	.tiff	2023	EPSG:4326	https://landscan.ornl.gov/

```python
def extract_aod_and_coords(hdf_file, grid_name='GRID_1', sds_name='Optical_Depth_055'
    """
    Extract AOD data along with latitude and longitude coordinates.
    """
    hdf = SD(hdf_file, SDC.READ)
    metadata_raw = hdf.attributes()['StructMetadata.0']
    grid_metadata = parse_metadata(metadata_raw, grid_name)

    lon, lat = construct_coords(grid_metadata)

    sds = hdf.select(sds_name)
    aod_data = sds.get()

    if aod_data.ndim == 3:
        print(f"Processing multiple orbits: {aod_data.shape[0]} orbits found.")
        aod_data = np.mean(aod_data, axis=0)

    if aod_data.shape != lon.shape:
        raise ValueError(f"Shape mismatch: AOD shape {aod_data.shape}, Latitude/Longi

    filename = hdf_file.split("/")[-1]
    parts = filename.split(".")
    dataset_name = parts[0]
    acquisition_date = datetime.strptime(parts[1][1:], "%Y%j").strftime("%Y-%m-%d")
    production_datetime = datetime.strptime(parts[4], "%Y%j%H%M%S")

    flat_aod = aod_data.ravel()
    flat_lat = lat.ravel()
    flat_lon = lon.ravel()
```

Fig 1: Extract Optical Depth 055 and coordinates for AOD measurements

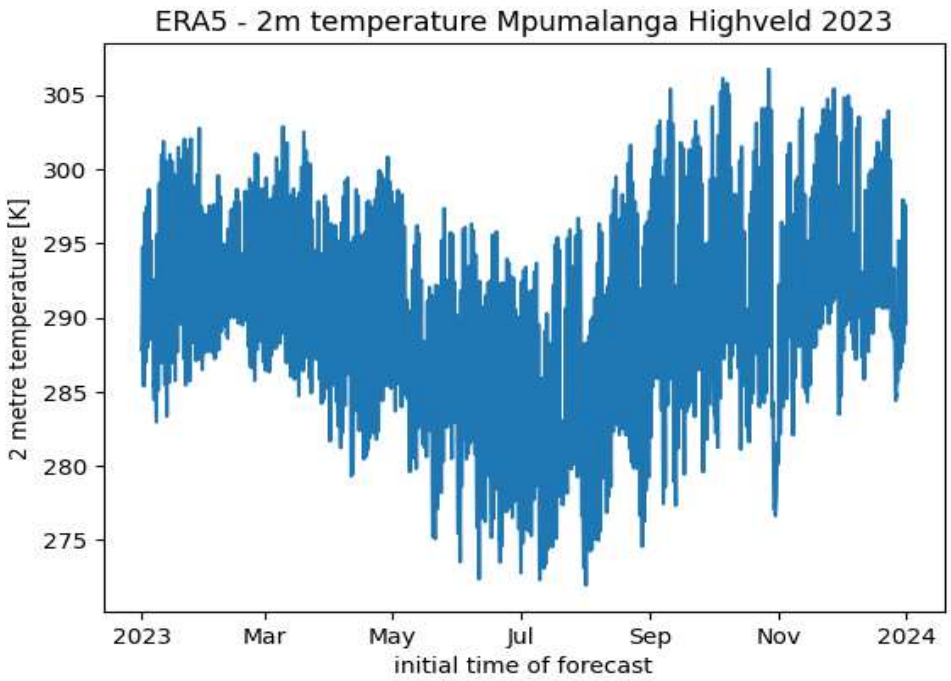

Fig 2: 2M Temperature data (in K) for a single coordinates pair (longitude and latitude)

```python
def combine_batches_with_spark(temp_folder, output_file):
    """
    Combine all batch files in the temp folder using Spark and save the result as a s
    """
    # List all batch files
    temp_files = [os.path.join(temp_folder, f) for f in os.listdir(temp_folder) if f.

    if not temp_files:
        print("No batch files found in the temp folder.")
        return

    print(f"Found {len(temp_files)} batch files. Starting the union process.")

    # Read all batch files into a single Spark DataFrame
    df_union = spark.read.parquet(temp_files[0])
    for temp_file in temp_files[1:]:
        temp_df = spark.read.parquet(temp_file)
        df_union = df_union.union(temp_df)

    # Write the combined DataFrame to a single Parquet file
    df_union.write.mode("overwrite").parquet(output_file)
    print(f"All batches combined and saved to {output_file}")

# Combine the batch files
combine_batches_with_spark(TEMP_FOLDER, FINAL_OUTPUT)

# Verify the output (optional)
final_df = spark.read.parquet(FINAL_OUTPUT)
print(f"Combined DataFrame contains {final_df.count()} rows and {len(final_df.columns
```

Fig 3: Big Data Processing with Spark

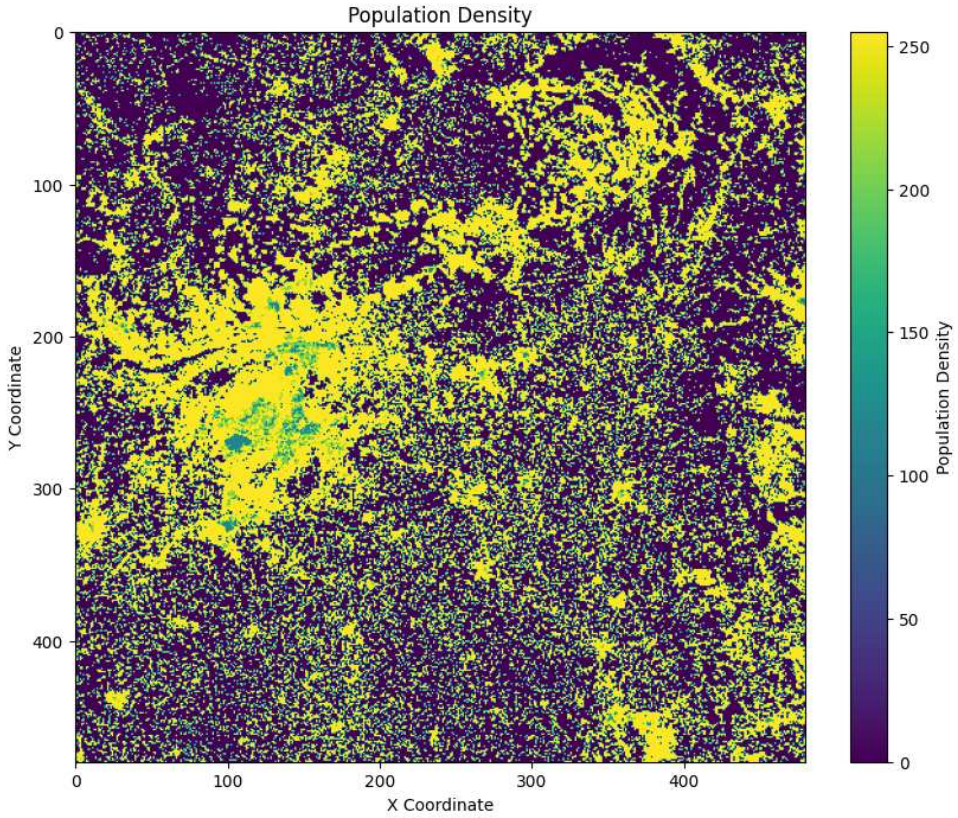

Fig 4: Study Area population density

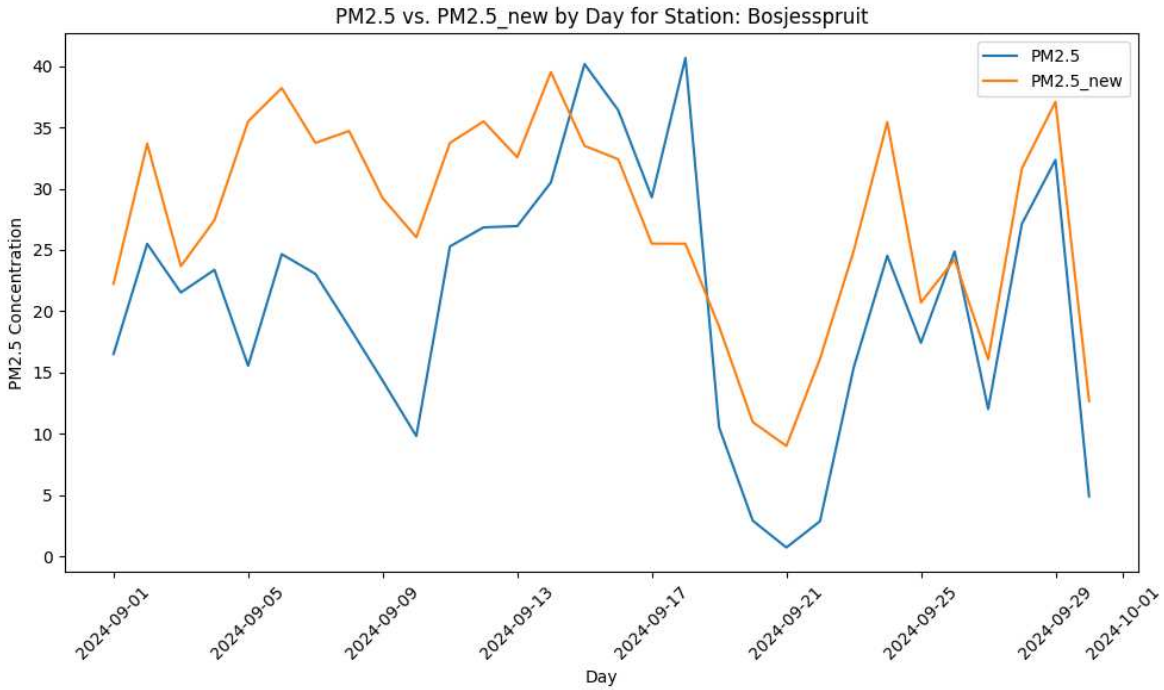

Fig 5: Actual vs Predicted PM2.5 values for wards within 50km radius from Bosjesspruit Land Station

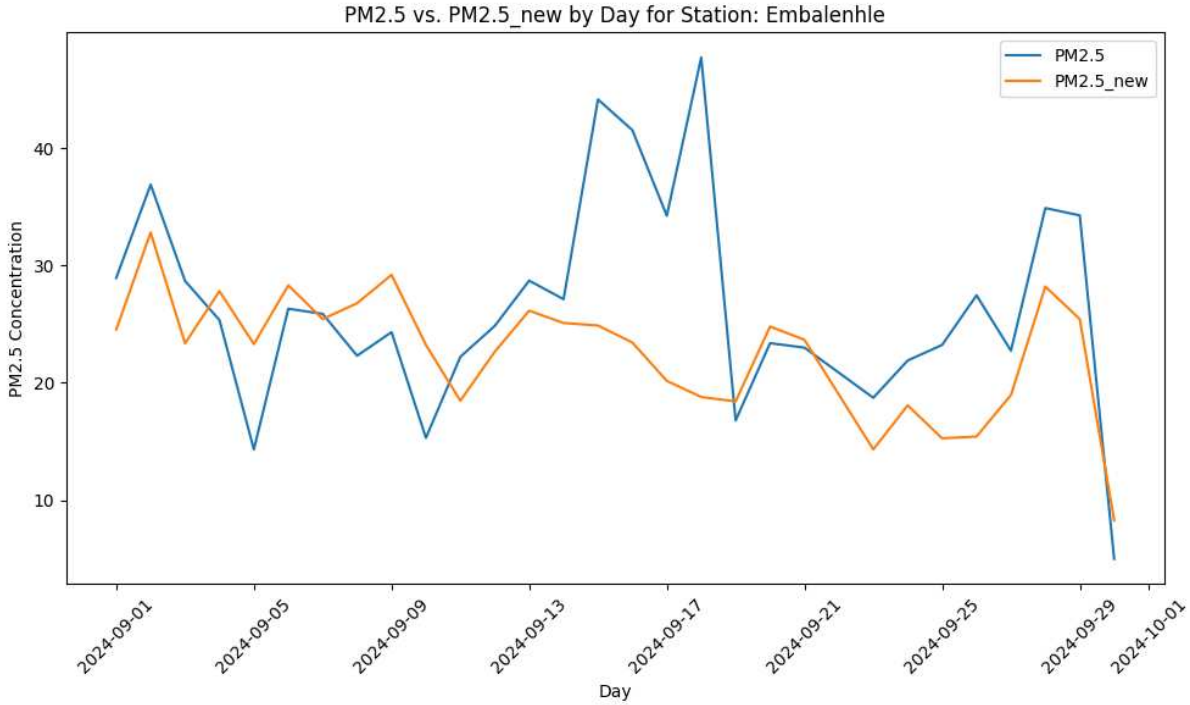

Fig 6: Actual vs Predicted PM2.5 values for eMbalenhle Land Station