IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 27, Issue 3, Ser. 5 (May. – June. 2025), PP 32-36
www.iosrjournals.org

AI for Climate and Health: A Survey of Machine Learning and Big Data Applications in Environmental Risk Analysis

[1]Vani Makula [2]Dr. Akhil Khare
[1]Ph.d Scholar, CSE Dept,Osmania University, Hyderbad, India
[2]Professor, CSE Dept. MVSR Engineering College Hyderabad, India.
Corresponding Author : Vani Makula

Abstract
The escalating impacts of climate change—ranging from extreme weather to deteriorating air quality—have significant implications for human health. In response, the integration of big data analytics and machine learning (ML) techniques has emerged as a powerful approach for forecasting climate patterns and assessing associated health risks. This survey presents a systematic review of 15 recent studies that apply machine learning (ML) models across diverse domains, including air pollution prediction, climate monitoring, water quality assessment, and health sentiment analysis. Key technologies include ensemble models, deep learning frameworks, and cloud-based platforms like Google Earth Engine and Apache Spark. The review also highlights critical research gaps, including the need for real-time predictive systems, explainable AI, multimodal data integration, and global generalizability. Findings emphasize the importance of scalable, transparent, and actionable AI systems for climate-health intelligence. Finally, the paper outlines future directions involving federated learning, benchmark dataset development, and interdisciplinary collaboration to support resilient, ethical, and inclusive climate-health solutions.

Keywords: *Big Data Analytics, Climate Change, Machine Learning, Health Impact Assessment, Explainable AI*

Date of Submission: 20-06-2025 Date of Acceptance: 03-07-2025

I. INTRODUCTION

Climate change has become one of the most pressing global issues of our time, impacting human health, ecosystems, and economies alike. Climate change, particularly the unprecedented frequency and intensity of climate-related phenomena, including heatwaves, floods, wildfires, and air pollution, is now threatening public health systems worldwide both directly and indirectly. Almost all of these environmental stressors are linked to respiratory ailments, cardiovascular diseases, waterborne diseases, and increasing mental health calls, especially in vulnerable groups. Advanced analytical frameworks are needed to manage complex, high-dimensional, and heterogeneous datasets, which can help unpack the multifaceted effects of climate change on health.

Machine learning (ML) has emerged as a valuable predictive tool for climate-health research, thanks to the advances in many big data technologies, including the increasing ease of access to real-time environmental and health data streams. Machine learning techniques enable the power to learn from historical trends, identify outliers, and predict future risks with very high levels of accuracy. More complex models, including Random Forest, XGBoost, deep neural networks, and natural language processing-based architectures like BERT, are increasingly applied to air quality, climate, disease outbreak, and environmental health sentiment analysis, utilizing data such as air and water quality measures.

Platforms like Google Earth Engine and Apache Spark provide the necessary infrastructure for the scalable and efficient processing of large geospatial and sensor-based datasets, which enhances the usability of statistical and machine learning methods for climate analytics. Although significant advances have been made, issues such as model interpretability, real-time deployment, and integrating multimodal data sources remain to be adequately addressed.

It aims to provide an organized and comprehensive summary of developments over the last four years (2019–2023), utilizing ML-based big data analytics to deliver solutions addressing the challenges of climate change and its health effects. A comprehensive literature review across five thematic domains is presented in Section 2. Section 3: Identifying research gaps based on synthesis of the findings. Section 4 summarizes the paper and suggests potential future research directions in this rapidly evolving area.

II. LITERATURE REVIEW

Recent machine learning and big data strategies for examining climate change and its amplifying effects on human health have emerged, which are reviewed in this paper. Agarwal et al. Hindu BusinessLine [1] — Explored air quality prediction models based on machine learning for various metropolitan cities of India. The prediction accuracy of PM2 using ensemble algorithms is greater than 92% (Random Forest and XGBoost, as reported in this study). AQI and five levels of Domain knowledge - The model included environmental data such as temperature, humidity, and vehicle emissions. Proven to have great potential for enabling real-time monitoring in smart city infrastructures. The comparative evaluation of several models emphasizes the crucial need for data preprocessing and hyperparameter tuning. These findings allow the deployment and leveraging of ML-driven forecasting tools to reduce pollution in urban environments and inform the development of effective urban health policies in an anticipatory manner.

Sandanam and Uthirapathy [2] explored climate change topics and public sentiment via Twitter data. Methods applied were Latent Dirichlet Allocation (LDA) for topic modeling and BERT for sentiment classification, achieving a classification accuracy of 93.5%. This approach, they say, exposed trends in the way climate-related discourse peaks during climate events. It also demonstrated the ability of transformer-based language models to recognize and address climate concerns and misinformation. Our findings are helpful to policymakers and climate communicators who desire to interpret social responses, either to frame such messages or assess the effectiveness of awareness programs based on real-time social media analytics.

Jovanovska et al. Three presented a comprehensive approach to the literature on air pollution measurement and forecasting, with an emphasis on the integration of an abstraction layer for machine learning in urban areas. In this work, we will examine both the stationary system and the mobile system, and propose using supervised learning models (support vector machines and deep neural networks) to determine the AQI at high resolution. The paper further identifies urban heterogeneity, sensor drift, and lack of data as challenges to this approach. The suggestion includes hybrid models that link sensor information with satellite imagery, providing better spatial and temporal coverage. Such work is critical for developing precision pollution mitigation strategies in innovative city ecosystems.

Misra et al. The application of big data, the Internet of Things (IoT), and machine learning to agriculture and food systems is explored in [4]. In this paper, we have seen how sensor data from farms is being used in ML Models to monitor soil health, predict crop yield, and detect diseases early. Focus on scalable tools to develop effective, real-time analytics systems, such as Keras and TensorFlow. This study explores the potential benefits in terms of sustainability, food safety, and resource efficiency. It also outlines future directions, such as incorporating weather data and autonomous irrigation systems. This foundational work supports the development of innovative agricultural systems in climate-vulnerable regions.

Tamiminia et al. A recent meta-analysis on GEE applications by [5] focused on climate and geospatial applications. The review included studies employing GEE across a range of applications, from land-use classification and vegetation monitoring to climate variability assessments. The authors highlighted the GEE cloud infrastructure, which can tackle petabyte-scale satellite imagery, coupled with machine learning models. The platform demonstrated good applicability in real-time environmental monitoring; however, its memory and function settings were limited. The study highlights the value of GEE in delivering scalable, remote assessments for climate scientists, especially in data-scarce locations with limited local computing infrastructure.

Samal et al. In [6], a TCDA model was proposed for reliable deep learning-based prediction of PM2.5 concentrations in the presence of missing or noisy data. It leveraged real-world urban air quality datasets, which is a step forward in addressing the fundamental accessibility challenges of pollution forecasting, such as incomplete or inconsistent time series. Combining temporal dependencies with autoencoder-based noise reduction, the model outperformed traditional LSTM and ARIMA models. The study highlighted the need for temporal convolution due to the nature of the sequence learning task. It showed that the architecture is robust against non-deterministic sensing environments, making it a good candidate for urban air quality management systems.

Nguyen et al. In the context of big data, it has been shown in [7] that a variety of both machine learning and deep learning frameworks can be applied for data mining on large-scale data. It then proceeded to a critical comparison of platforms like TensorFlow, PyTorch, and Spark MLlib in terms of architecture, scalability, and support for distributed computation. The paper continued by addressing platforms such as MATLAB, Caffe, and MXNet, among others. More specifically, we focused on Apache Spark for parallel machine learning workloads and TensorFlow for at-scale neural network training. Data preprocessing, model portability, and cross-platform deployment were also discussed in the review, among many others. They provide their findings in the form of practical guidance for researchers building scalable machine learning (ML) pipelines for applications such as environmental monitoring and health impact analyses.

Amani et al. The study [8] has explored the application of remote sensing big data analytics using Google Earth Engine (GEE) for climate monitoring and land cover classification. GEE was presented as a cloud-based architecture with a collection of accessible machine learning algorithms, including CART and SVM, through a web interface. Google Earth Engine (GEE) was explicitly mentioned for its role in making satellite imagery more accessible to a broader audience, as well as its integration with APIs for automation. The authors similarly noted the current limitations of memory and computational latencies. This work enables GEE to be deployed at a canonical location in subsequent AI-enabled environmental survey systems and climate change impact analyses.

Kakani et al. Additionally, [9] published a comprehensive review on the introduction of computer vision and artificial intelligence in the food industry, including applications ranging from quality inspection to environmental sustainability. The image processing capabilities of ML can be used to identify spoilage in food or defects in packaging. In contrast, sensor data can be used to predict the future quality of the crop. More specifically, it discussed how deep learning paradigms, such as CNNs and object detection frameworks, are essential building blocks that automate these processes. It also mentioned future trends, describing them to be smart farming and sustainable processing. According to their review, AI and agriculture play critical roles in ensuring food security in the face of climate change.

The review has focused on the application of big data and artificial intelligence techniques in the energy sector, using bibliometric analysis [10]. Among other things, they identified key trends in publications, research clusters, and emerging themes, including renewable energy forecasting, intelligent grid optimization, and carbon footprint reduction (see the full literature review here). They mapped high-impact journals, prevalent keywords, and countries' cooperation using data mining tools such as VOSviewer and Scopus APIs. AI-based energy analytics are at the heart of climate resilience strategies, the paper concluded. The meta-analysis also calls for interdisciplinary collaboration to bring together environmental modeling, data science, and sustainable energy practices in mitigation efforts.

Adamson Oloyede et al. presented a data-driven framework for predicting temperature via big data analysis in [11]. In the paper comparing NASA's satellite temperature data with regional meteorological datasets, such as those from NiMet, they demonstrated robust correlations. Various temporal scales were tested for enhancing forecasting accuracy by implementing regression-based machine learning algorithms. Authors also highlighted the vital contribution of data integration and preprocessing towards achieving more reliable results. Hybrid models that utilize both global and local data sources can together produce the necessary data for better climate projections and public health and agricultural decision-making tools across climate-sensitive areas, with the support of their approach. Table 1: Summary of 15 recent studies that apply ML to climate change and health outcome predictions.

Table 1: Summary of Selected Literature on ML-Based Climate and Health Analytics

Sl. No.	Author(s)	Year	Domain/Focus	Technique(s) Used	Dataset/Source	Key Contribution
1	Agarwal et al.	2023	Air Pollution Forecasting	RF, XGBoost	City air quality data	Achieved 92% accuracy in PM2.5 and AQI prediction using ensemble models.
2	Sandanam & Uthirapathy	2023	Climate Sentiment Analysis	BERT, LDA	Twitter Data	Achieved 93.5% accuracy in classifying climate-related sentiments.
3	Jovanovska et al.	2023	Urban Air Monitoring	SVM, DNN	Stationary + mobile sensors	Proposed hybrid models for high-resolution pollution monitoring.
4	Misra et al.	2020	Agriculture & Food Systems	IoT, DL, ML	Smart farm sensors	Integrated big data and ML for yield prediction and disease detection.
5	Tamiminia et al.	2020	Climate Modeling	GEE + ML	Satellite imagery	Validated Google Earth Engine for scalable geospatial analysis.
6	Samal et al.	2021	PM2.5 Forecasting	TCDA (Autoencoder)	Urban air pollution data	Developed a noise-resistant model for incomplete air quality data.
7	Nguyen et al.	2019	ML Frameworks	TensorFlow, Spark MLlib	Various big data platforms	Compared distributed ML libraries for environmental applications.
8	Amani et al.	2020	Remote Sensing	CART, SVM in GEE	Satellite data	Highlighted benefits and challenges of using GEE in climate analysis.
9	Kakani et al.	2020	Food Quality Inspection	CNNs, DL	Food imaging & sensors	Reviewed CV applications in food safety and crop quality analysis.
10	Hou & Wang	2023	Energy Analytics	Bibliometric & AI tools	Scopus, VOSviewer	Identified key trends in AI-driven energy and climate research.
11	Oloyede et al.	2023	Temperature Prediction	Regression Models	NASA + NiMet	Correlated global and local datasets for regional forecasting.
12	Hatcher & Yu	2018	DL Platforms	TensorFlow, PyTorch	Multi-domain review	Surveyed platforms supporting scalable deep learning models.
13	Bi et al.	2020	Hydrological Modeling	CNN, AI	Porous media data	Demonstrated DL in subsurface flow and geoscience predictions.
14	Fathi et al.	2020	Urban Energy	SVR, DNN,	Building energy	Reviewed ML models for

			Forecasting	GBM	datasets	sustainable energy optimization.
15	Zuo et al.	2019	Water Quality Prediction	RF, SVM, NN	Hydrological datasets	Improved accuracy in water pollution forecasting using ML.

A wide range of deep learning platforms, applications, and trends, along with their significance in environmental and health applications, has been summarized by Hatcher and Yu [12]. The review encompassed the major frameworks, including PyTorch, TensorFlow, and CNTK, and compared them in terms of performance optimization, community support, and adaptability for large-scale data scenarios. It also highlighted trends such as explainable AI and edge computing. Their findings are highly relevant for researchers using deep models on environmental datasets, especially in resource-constrained contexts where health impact analytics require prompt inferences.

Bi et al. AI and Machine Learning in Porous Media and Geoscience: Numerous articles have discussed the applications of AI and machine learning for hydrological modeling and environmental characterization in porous media and geoscience (e.g., [13]). This approach utilized convolutional neural networks (CNNs) to explain complex subsurface features and flow dynamics. In some geospatial applications, particularly those complemented by large-scale, labeled datasets, the authors have proven that data-driven models surpass the performance of conventional simulation techniques. This is relevant for groundwater evaluation and environmental risk assessment in the context of climate impacts. The authors argue that by combining physical constraints with AI models, a more reliable model can be created, facilitating long-term planning for climate sustainability [N].

Fathi et al. Systematic review of machine learning applications in prediction-based energy performance forecast in buildings at urban scale [14]. In this work, the paper assessed several machine learning models, including Gradient Boosting, Support Vector Regression, and Deep Neural Networks, in terms of their accuracy and efficiency for predicting energy consumption under varying environmental and operational conditions. It also included general feature selection methods, such as principal component analysis (PCA) and recursive elimination, to optimize the input variables. With this diverse dataset and the available tools for big data analytics, their findings help sensitize a data-driven approach to sustainable urban infrastructure, utilizing resources that align with climate action objectives of energy efficiency and emissions reduction in smart cities.

Zuo et al. Different machine learning-based models for forecasting water quality, incorporating large-scale datasets from diverse hydrological sources, have been conducted [15]. This comparative analysis examines the predictive capabilities of random forests, support vector machines, and neural networks for detecting chemical and biological contaminants in surface water. The authors identified critical predictive features, including dissolved oxygen and nitrate. The models performed much more accurately than conventional statistical approaches. Pollution forecasting is a crucial component for early warning systems, the preservation of resources, and ensuring access to potable water amid climate variability. Hence, this study advances the use of data-driven approaches in environmental monitoring.

The literature encompasses various areas, including air pollution forecasting, climate modeling, sentiment analysis for health, and energy prediction. These involve ensemble models, deep learning, autoencoders, and cloud platforms such as Google Earth Engine (GEE) and Apache Spark. The studies provide scalable and accurate insights, but leave room for improvement in deployment, interpretability, and multimodality.

III. RESEARCH GAPS

Although the use of machine learning and extensive data analysis in a climate and health context has advanced considerably, older research gaps remain: most importantly, the integration of multimodal data where health outcome records are entered in a unified framework, environmental data such as air quality and meteorological and remote sensing data are rarely brought together. However, works such as Agarwal et al. [1] and Samal et al. Although these models are used to predict pollutants [6], they provide little direct information that can be used to inform health impacts.

XAI models are part of another gap. Although these models achieve high accuracy in climate discourse [2] and image analysis [9], most are not interpretable, as they are black-box models. This renders the results untrustworthy or unreliable for decision-makers, such as policymakers and healthcare professionals, to use in such important decisions.

Moreover, there is a scarcity of real-time systems capable of supporting early warning and response mechanisms. Though platforms like Google Earth Engine [5][8] and Spark [7] offer scalable computation, few studies demonstrate their deployment in live public health surveillance.

The lack of open benchmark datasets also inhibits consistent model evaluation and comparison across domains. Bi et al. [13] and Zuo et al. [15] emphasize the importance of large-scale, labeled data; however, such resources remain scarce in the context of climate health.

Lastly, existing models show limited geographic generalization. Research efforts are heavily concentrated in specific countries or cities [3], [11], leaving rural or low-resource areas underrepresented, which weakens the global applicability of ML-driven climate-health solutions.

These gaps underscore the urgent need for interdisciplinary collaboration, ethical AI practices, and the development of standardized, explainable, and context-aware frameworks to facilitate meaningful progress in this domain.

IV. CONCLUSION AND FUTURE SCOPE

We synthesized evidence of recent progress that has employed machine learning and big data analytics to study climate change and its health effects through a systematic review of the literature. Using 15 recent works as examples, we examined various applications, including air pollution forecasting, climate modeling, water quality forecasting, and urban energy analysis. The results highlight an increasing dependence on scalable platforms (e.g., Google Earth Engine 5, spark [7]) and more complex models (e.g., BERT [2], CNNs [9], autoencoders [6]) for processing complex datasets. Altogether, these studies show the potential of AI to identify patterns that not only shape environmental policy but also public health. While these accomplishments are commendable, the survey also identified significant research limitations in terms of explainability, data integration, real-time deployment, and geographical generalization. In the future, new work should focus on extractable models and provide transparency to facilitate the adoption of policies. The fusion of multimodal data, particularly the integration of health and natural environmental datasets, will improve our capabilities in prediction and translation. Climate-sensitive regions are perceived to severely lag behind non-sensitive areas when it comes to early warning efforts that could support the acceleration of decision-making – this could change with the application of edge AI and federated learning-powered real-time early warning systems. Third, we advocate for open-access benchmark datasets to provide a consistent method for evaluating models and facilitate progress. Such directions will set the stage for improved climate–health information systems that are more equitable, usable, and resilient.

References

[1]. Aryan Agarwal, Pratik Dighole, Abhishek Sabnis, Dhananjay Thosar, Madhuri Mane. (2023). Detection and Predicting Air Pollution Level in a Specific City Using Machine Learning Models. *International Journal of Creative Research Thoughts (IJCRT)*, 11(3), pp.158–167.

[2]. Samson Ebenezar Uthirapathy and Domnic Sandanam. (2023). Topic Modelling and Opinion Analysis On Climate Change Twitter Data Using LDA And BERT Model. *Elsevier*, pp.908–917.

[3]. Elena Mitreska Jovanovska et al. (2023). Methods for Urban Air Pollution Measurement and Forecasting. *MDPI*, pp.1–25.

[4]. Misra, N. N., et al. (2020). IoT, big data and artificial intelligence in agriculture and food industry. *IEEE Internet of Things Journal*, 1–1. http://doi.org/10.1109/JIOT.2020.2998584

[5]. Tamiminia, H., et al. (2020). Google Earth Engine for geo-big data applications: A meta-analysis and systematic review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 164, 152–170.

[6]. Samal, K. K. R., et al. (2021). Temporal convolutional denoising autoencoder for air pollution prediction. *Urban Climate*, 100872.

[7]. Nguyen, G., et al. (2019). ML and DL frameworks and libraries for large-scale data mining: a survey. *Artificial Intelligence Review*. http://doi.org/10.1007/s10462-018-09679-z

[8]. Amani, M., et al. (2020). Google Earth Engine Cloud Platform for Remote Sensing Big Data Applications: A Comprehensive Review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 1–1.

[9]. Kakani, V., et al. (2020). A critical review on computer vision and artificial intelligence in food industry. *Journal of Agriculture and Food Research*, 2, 100033.

[10]. Hou, Y., & Wang, Q. (2023). Big data and artificial intelligence application in energy field: a bibliometric analysis. *Springer*, pp.1–27.

[11]. Adamson Oloyede et al. (2023). Data-driven techniques for temperature data prediction: big data analytics approach. *Springer*, pp.1–21. https://doi.org/10.1007/s10661-023-10961-z

[12]. Hatcher, W. G., & Yu, W. (2018). A Survey of Deep Learning: Platforms, Applications and Emerging Research Trends. *IEEE Access*, 1–1.

[13]. Bi, Y., et al. (2020). AI and ML in porous media and geoscience for hydrological modeling. *Advances in Water Resources*, 103619.

[14]. Fathi, S., et al. (2020). Machine learning applications in urban building energy performance forecasting: A systematic review. *Renewable and Sustainable Energy Reviews*, 133, 110287.

[15]. Zuo, M., et al. (2019). Surface water quality prediction using ML models based on big data. *Water Research*, 115454. http://doi.org/10.1016/j.watres.2019.115454