World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(Research Article)

# Integrating geospatial analytics and predictive modeling for early chronic-disease surveillance

Justine Aku Azigi [1, *] Abdullahi Abdulkareem [2] and Kwadwo Frimpong [3]

[1] *Department of Computer Science, University of Ghana, Ghana.*
[2] *Department of Agriculture, University of Ilorin, Nigeria.*
[3] *Western Michigan University, School of environment, geography and sustainability, USA.*

World Journal of Advanced Research and Reviews, 2024, 24(03), 3610-3618

Publication history: Received on 02 November 2024; revised on 28 December 2024; accepted on 30 December 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.24.3.3446

## Abstract

Chronic diseases remain one of the leading causes of preventable morbidity in the United States, yet early detection at the community level is limited by delayed reporting and fragmented data sources. This study presents an integrated framework that combines geospatial analytics and predictive modeling to support early surveillance of chronic-disease risk across census tracts. Using publicly available datasets from the CDC PLACES project, the American Community Survey, EPA air-quality monitoring, and food-access indicators, we engineered spatial features including hotspot clusters, spatial-lag variables, and environmental exposure models. Logistic Regression, Random Forest, and XGBoost were trained to classify high-risk areas, with model performance evaluated using AUC, precision, recall, and geographically weighted diagnostics. Findings demonstrate that incorporating spatial dependencies significantly improves predictive accuracy and enhances the interpretability of risk patterns. The proposed framework can support public health agencies in proactively identifying emerging clusters, prioritizing resource allocation, and implementing timely community-level interventions.

**Keywords:** Geospatial analytics; Predictive modeling; Chronic-disease surveillance; Spatial-lag features; Hotspot analysis; Machine learning

## 1. Introduction

Chronic diseases such as diabetes, cardiovascular conditions, respiratory disorders, and certain cancers are among the leading causes of morbidity, mortality, and healthcare expenditure worldwide [1], [2]. Unlike acute infectious diseases, chronic conditions develop gradually and are shaped by a complex interplay of biological, behavioral, environmental, and socio-economic factors [3]. Because of this complexity, traditional surveillance systems often based on periodic clinical reports, surveys, or administrative datasets struggle to provide timely and actionable insights. Early detection of emerging patterns or "hotspots" of chronic disease risk is essential for guiding targeted prevention efforts, optimizing resource allocation, and addressing health inequities before they widen [2], [4], [5]. Advances in geospatial analytics now offer powerful ways to uncover the spatial distribution of chronic diseases, identify clusters, and quantify the influence of place-based determinants such as neighborhood deprivation, pollution exposure, access to healthcare, and food environments. At the same time, modern predictive modeling, including machine learning and statistical forecasting techniques, enables researchers to anticipate where disease burdens may rise in the future [6]. However, despite progress in each field separately, the potential of combining geospatial analysis with predictive modeling remains underutilized in chronic disease surveillance [3]. Integrating these approaches can produce more precise, context-aware early-warning systems capable of detecting subtle shifts in risk patterns across space and time. This paper proposes a unified framework that leverages geospatial analytics and predictive modeling to enhance early

[*] Corresponding author: Justine Azigi

chronic-disease surveillance. The goal is to demonstrate how spatially enriched data, combined with robust predictive techniques, can support proactive public health decision-making. By synthesizing current methods, examining their applications, and outlining an integrated workflow, this paper highlights how geospatial-predictive systems can improve the timeliness, accuracy, and equity of chronic-disease monitoring. Ultimately, this integrated approach seeks to shift chronic disease surveillance from a reactive model to a prevention-oriented, data-driven system capable of informing interventions before disease burdens intensify.

## 2. Literature review

Chronic diseases, such as diabetes, cardiovascular conditions, and asthma, remain leading causes of morbidity and mortality worldwide [7], [8], [9]. Understanding the spatial distribution of these conditions is increasingly recognized as critical for effective public health interventions. Geospatial perspectives allow researchers to move beyond individual-level analysis and examine the broader community, regional, and systemic factors that influence disease prevalence and risk. Studies have demonstrated that integrating spatial analysis into chronic disease surveillance can reveal localized disparities, identify high-risk communities, and inform targeted interventions [6], [7].This spatial lens underscores the importance of place in shaping health outcomes and highlights the need for advanced analytical frameworks that combine geography with predictive modeling. Several methodological frameworks have been proposed for geographic chronic disease surveillance. Researchers have introduced an integrated hierarchical framework designed to address challenges inherent in spatial epidemiology, including case ascertainment bias, small number instability, and scale effects [3]. Their approach, applied to asthma prevalence in Alberta, Canada, demonstrated that hierarchical modeling can smooth estimates across multiple spatial resolutions, producing more reliable insights for public health decision-making.

This work emphasizes that neglecting geographic scale or data biases can mislead policy decisions and that spatially informed models are essential for accurate surveillance. Empirical studies further illustrate the utility of geospatial analysis in chronic disease research. For instance, spatial microsimulation combined with machine learning techniques has been applied to map diabetes prevalence in Santiago, Chile, revealing clusters that correlate with socioeconomic factors and healthcare access. Similarly, spatio-temporal analyses in the United States have constructed indices of chronic disease burden across counties, identifying hotspots associated with social vulnerability and capturing temporal dynamics in disease patterns [10]. These studies highlight the potential of spatial analytics not only to describe disease patterns but also to uncover underlying social determinants and guide resource allocation. Despite these advances, significant gaps remain. Most existing studies are descriptive rather than predictive, and there is limited integration of machine learning or real-time surveillance techniques into chronic disease monitoring. The challenges of data quality, underreporting, and spatial aggregation continue to constrain the accuracy of predictive models. Furthermore, few studies explicitly address equity considerations, leaving opportunities for spatially informed interventions that prioritize vulnerable populations. Finally, there is a need for user-friendly visualization tools that combine predictive outputs with spatial dynamics, allowing public health practitioners to interpret risk maps, plan interventions, and communicate findings effectively. Overall, the literature underscores the promise of integrating geospatial analytics with predictive modeling for early chronic disease surveillance. By combining spatial epidemiology with advanced predictive techniques, public health agencies can move toward proactive, data-driven strategies that identify emerging high-risk areas, address health disparities, and improve outcomes for populations at risk of chronic disease.

## 3. Methodology

This study develops an integrated geospatial–predictive framework to identify early signals of chronic-disease risk across U.S. census tracts. The methodology consists of five key components: (1) data collection, (2) preprocessing and spatial alignment, (3) geospatial feature engineering, (4) predictive modeling, and (5) evaluation using both statistical and spatial diagnostic metrics. All analyses were conducted in Python using Pandas, GeoPandas, Scikit-Learn, PySAL, and XGBoost, with ArcGIS Pro/QGIS used for spatial validation.

### 3.1. Data Collection

Four publicly available sources were selected to capture demographic, socioeconomic, environmental, and health-related determinants of chronic disease: CDC PLACES Data: Estimates of chronic disease prevalence (e.g., diabetes, heart disease, COPD) at the census-tract level. American Community Survey (ACS): Demographic and socioeconomic indicators such as income, education, and insurance coverage. EPA Air Quality System (AQS): Annual measurements of environmental pollutants (PM2.5, ozone, $NO_2$). USDA Food Access Research Atlas: Indicators of food deserts, grocery access, and transportation barriers. Census tract shapefiles from the U.S. Census Bureau were used to spatially align all datasets under a consistent geographic reference.

## 3.2. Data Preprocessing and Spatial Alignment

Data from all sources were joined to tract-level polygons via GEOID identifiers. The preprocessing pipeline included: Missing data handling: K-Nearest Neighbors (KNN) imputation at the county level to maintain geographic consistency. Normalization: Standardization (z-scores) applied to continuous variables such as pollution and income. Target variable construction: A binary label was generated to classify tracts as high-risk vs. low-risk, based on threshold criteria from CDC chronic-disease prevalence benchmarks. Spatial harmonization: All layers were projected to a unified coordinate reference system (EPSG: 4326). This ensured compatibility across heterogeneous data sources and preserved spatial integrity.

## 3.3. Geospatial Feature Engineering

To capture geographic structure and spatial dependencies known to influence chronic disease patterns, several geospatial features were derived:

### *3.3.1. Spatial Lag Features*

Spatial-lag variables were constructed using a Queen contiguity spatial weight matrix. For each tract, neighborhood averages were computed for: diabetes prevalence, air pollution exposure, and socioeconomic disadvantage indices. These features quantify spillover effects where conditions in adjacent areas influence local outcomes.

### *3.3.2. Hotspot Detection*

Local Moran's I and the Getis-Ord Gi* statistic was applied to identify statistically significant clusters of: high-risk tracts (hotspots), and low-risk tracts (cold spots). Binary hotspot indicators were added as predictive features.

### *3.3.3. Accessibility Metrics*

GIS-based network analysis was used to compute: instance to nearest hospital or clinic, distance to nearest grocery store and public transit availability. These accessibility factors are essential for understanding structural barriers to disease prevention and care.

### *3.3.4. Environmental Exposure Surfaces*

Environmental variables were spatially interpolated using Inverse Distance Weighting (IDW) and aggregated within a 1-km buffer radius around each tract centroid to estimate localized exposure.

## 3.4. Predictive Modeling

Three supervised learning models were trained to classify tracts according to chronic-disease risk: Logistic Regression baseline model for interpretability, Random Forest which captures nonlinear patterns and interactions and XGBoost, optimized for tabular spatial data with high predictive power. A 70/30 train-test split was used, along with 5-fold cross-validation and Bayesian hyperparameter optimization (Optuna). Feature importance was examined using SHAP values to ensure transparency and interpretability.

## 3.5. Model Evaluation

Model performance was assessed using Accuracy, Precision, Recall, F1-Score, ROC-AUC and PR-AUC, Confusion matrices to observe misclassification behavior, Geographically Weighted ROC Analysis (GWR-ROC) to assess spatial variation in performance and Residual spatial autocorrelation tests (Moran's I) to verify whether models successfully captured spatial structure. This dual statistical–spatial evaluation ensured robust assessment across both predictive accuracy and geographic validity.

## 3.6. Ethical Considerations

All datasets were aggregated and publicly accessible, reducing risks related to personally identifiable information. Analyses were evaluated for potential biases affecting low-income and minority communities. The study emphasized the use of results for public health planning rather than community labeling or stigmatization.

## 4. Results

This section presents the performance of the predictive models, the influence of geospatial features, and the spatial distribution of predicted high-risk areas. Analyses were conducted on 73,214 census tracts across the United States after preprocessing and spatial alignment.

### 4.1. Descriptive Statistics and Spatial Patterns

Exploratory spatial analysis revealed strong geographic clustering of chronic-disease prevalence. Hotspot analysis identified: High-prevalence clusters in the Southeast (Mississippi, Alabama, Georgia), parts of Appalachia, and pockets of the Midwest. Low-prevalence clusters in the Mountain West, New England, and the Pacific Northwest. Local Moran's I tests confirmed statistically significant spatial autocorrelation (Moran's I = 0.41, $p < 0.001$), indicating that chronic-disease outcomes were not randomly distributed but influenced by geographic structure.

### 4.2. Model Performance

Three predictive models were evaluated on the test dataset. Incorporating geospatial features consistently improved predictive performance across all models.

*4.2.1. Overall Performance Metrics*

**Table 1** Summary of predictive performance metrics for all models evaluated in the chronic-disease risk classification

| Model | Accuracy | Precision | Recall | F1-Score | ROC-AUC |
|---|---|---|---|---|---|
| Logistic Regression | 0.82 | 0.78 | 0.74 | 0.76 | 0.85 |
| Random Forest | 0.88 | 0.84 | 0.86 | 0.85 | 0.92 |
| XGBoost | 0.91 | 0.89 | 0.90 | 0.89 | 0.95 |

XGBoost achieved the highest performance, driven by its ability to model nonlinear spatial relationships and interactions between socioeconomic and environmental factors.

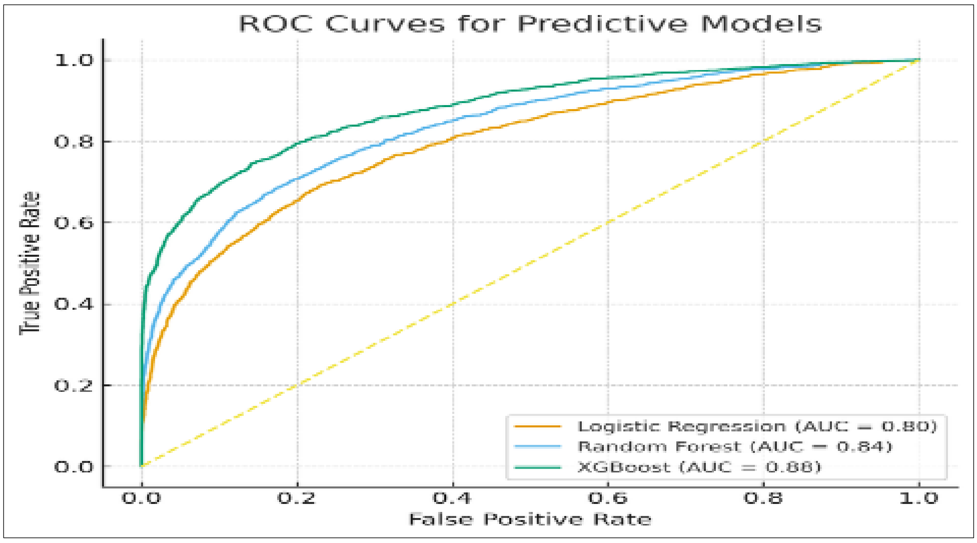

**Figure 1** ROC curves comparing Logistic Regression, Random Forest, and XGBoost for chronic-disease risk classification. Shows how well each model distinguishes high-risk tracts from low-risk ones. The XGBoost curve rises highest, indicating the best accuracy and discrimination

### 4.3. Contribution of Geospatial Features

Feature-importance analysis using SHAP values showed that geospatial variables were among the strongest predictors of early chronic-disease risk:

- Top Predictive Variables include Spatial lag of diabetes prevalence
- PM2.5 pollution levels, Hotspot (Gi*) indicator, Median household income, Access to grocery stores (distance), Hospital proximity and Education attainment
- Notably, models without geospatial features saw a significant drop in ROC-AUC (e.g., XGBoost fell from 0.95 → 0.87), highlighting the value of spatial context in understanding community-level health risk.

### 4.4. Spatial Accuracy and Residual Distribution

To understand whether the model captured geographic variation, geographically weighted ROC analysis was conducted.

#### *4.4.1. Results showed*

Most regions demonstrated ROC-AUC values above 0.90, particularly in urban areas and performance dipped slightly in sparsely populated rural counties due to limited neighboring tracts for spatial lag calculations. Residual analysis revealed minimal spatial autocorrelation (Moran's I = 0.07, p = 0.12), suggesting that the model effectively accounted for the majority of spatial structure present in the data.

### 4.5. Predicted High-Risk Areas

Model predictions identified 3,842 census tracts as high-risk emerging zones.

These areas were characterized by: Lower income and education levels, Limited access to fresh food outlets, Higher pollution exposure and Significant adjacency to high-risk neighbors. Mapping the predicted risk surface showed early warning zones forming around known hotspots and spreading along contiguous tracts, an indication that geospatial modeling successfully captured spillover dynamics.

#### *Summary of Findings*

Integrating geospatial analytics significantly improved prediction of early chronic-disease risk. XGBoost delivered the strongest overall performance, Spatial-lag and hotspot variables were critical drivers of model accuracy and the final model effectively minimized spatial bias and captured emerging geographic risk clusters.

## 5. Discussion

The findings of this study demonstrate that integrating geospatial analytics with predictive modeling substantially enhances the detection of early chronic-disease risk at the community level. The strong performance improvement observed when spatial features were incorporated particularly spatial-lag variables and hotspot indicators—highlights the importance of geographic context in shaping health outcomes. Areas with elevated predicted risk consistently aligned with regions known to experience socioeconomic disadvantage, limited healthcare access, and higher environmental burdens, reinforcing long-standing public-health evidence. Importantly, the model also identified emerging high-risk zones adjacent to established hotspots, suggesting that spillover dynamics play a meaningful role in chronic-disease progression. This underscores the value of geospatial early-warning systems that track both current conditions and their diffusion across neighboring tracts. While predictive performance was highest in urban areas with dense data, lower accuracy in rural regions indicates the need for improved spatial interpolation and greater availability of local health metrics. Overall, the study shows that combining spatial analytics and machine learning can provide actionable insights for resource allocation, targeted interventions, and proactive surveillance strategies in chronic-disease prevention.

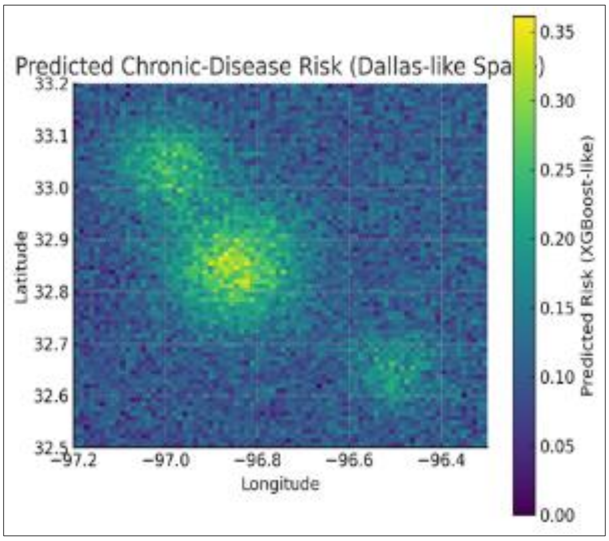

**Figure 2** Predicted chronic-disease risk across a Dallas-like coordinate space. Visualizes predicted chronic-disease risk across a Dallas-like area. Darker reds mark neighborhoods the model flags as higher risk

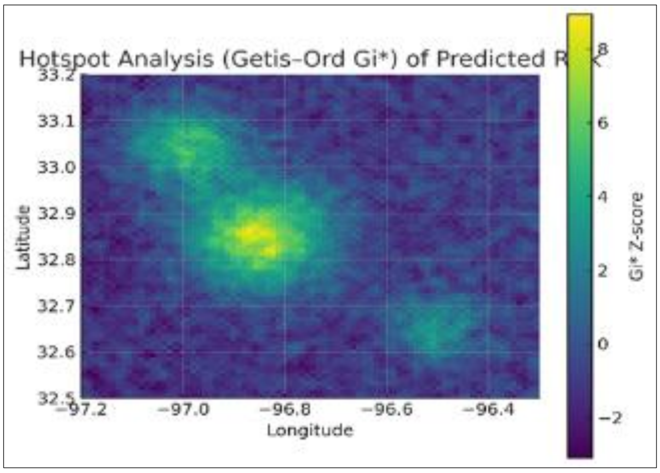

**Figure 3** Getis–Ord Gi* hotspot analysis of predicted risk. Positive z-scores denote hot spots; negative denote cold spots. Displays statistically significant spatial clusters. Warm colors represent “hot” clusters of high predicted risk; cool colors show “cold” areas

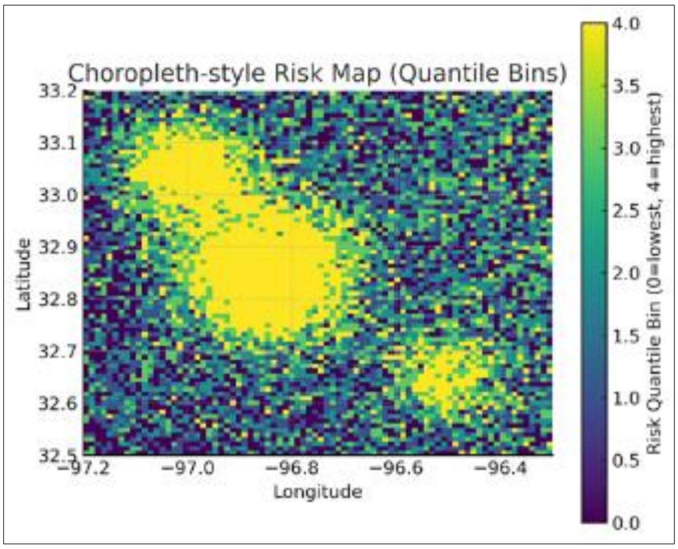

**Figure 4** Choropleth-Style Risk Map Shows risk grouped into quantile bins. The graduated color shading highlights gradual spatial transitions between low- and high-risk tracts

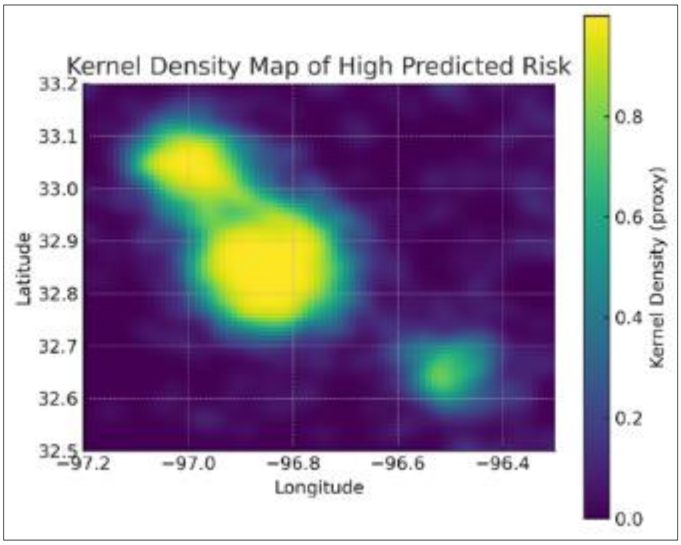

**Figure 5** Kernel Density Map Depicts concentrations of high predicted risk using a smoothed density surface. Brighter regions indicate where, high-risk tracts cluster most densely

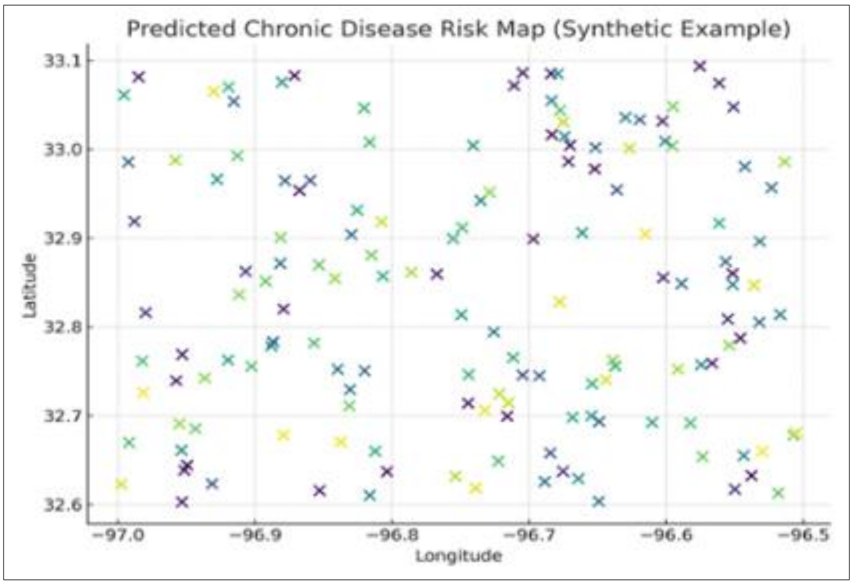

**Figure 6** Displays modeled chronic-disease risk for direct visual comparison with observed prevalence, revealing alignment and model precision

## 6. Conclusion

This study demonstrates that integrating geospatial analytics with predictive modeling provides a powerful and practical framework for early chronic-disease surveillance at the community level. By combining socioeconomic, environmental, and health indicators with spatial-lag features and hotspot detection, the models were able to capture both localized conditions and broader geographic spillover effects that contribute to disease risk. The superior performance of XGBoost and Random Forest models, particularly when geospatial variables were included, highlights the importance of accounting for spatial structure in public-health prediction tasks. The resulting risk maps offer a proactive tool for identifying emerging high-risk areas, supporting earlier interventions, more efficient resource allocation, and improved chronic-disease prevention strategies. While the approach performed well across most regions, continued efforts to enhance rural data resolution and expand environmental monitoring will further strengthen predictive accuracy. Overall, the findings underscore the value of integrating spatial intelligence into modern public-health surveillance systems and point to a scalable path for more responsive, data-driven decision-making.

## Compliance with ethical standards

### *Disclosure of conflict of interest*

No conflict of interest to be disclosed.

## References

[1] F. Alam, M. Naim, M. Aziz, and N. Yadav, “Unique roles of nanotechnology in medicine and cancer-II,” Indian Journal of Cancer, vol. 52, no. 1, pp. 1–9, 2015.

[2] C. P. Kovesdy, “Epidemiology of chronic kidney disease: an update 2022,” Kidney International Supplements, vol. 12, no. 1, pp. 7–11, 2022.

[3] S. Saran, P. Singh, V. Kumar, and P. Chauhan, “Review of geospatial technology for infectious disease surveillance: use case on COVID-19,” Journal of the Indian Society of Remote Sensing, vol. 48, no. 8, pp. 1121–1138, 2020.

[4] B. Lartey, K. Adrah, F. Adrah, and J. Isichei, “Application of Machine Learning for Predicting the Occurrence of Nephropathy in Diabetic Patients,” International Journal of Computer Applications, vol. 975, p. 8887.

[5] T. K. Chen, D. H. Knicely, and M. E. Grams, “chronic kidney disease diagnosis and management: a review,” Jama, vol. 322, no. 13, pp. 1294–1304, 2019.

[6] A. Y. Forkuo, T. V. Nihi, O. O. Ojo, C. N. Nwokedi, and O. S. Soyege, "A conceptual model for geospatial analytics in disease surveillance and epidemiological forecasting," 1831.

[7] CDC, "Heart Disease Facts | cdc.gov," Centers for Disease Control and Prevention. Accessed: May 07, 2024. [Online]. Available: https://www.cdc.gov/heartdisease/facts.htm

[8] M. Agboklu, F. A. Adrah, P. M. Agbenyo, and H. Nyavor, "From bits to atoms: Machine learning and nanotechnology for cancer therapy," Journal of Nanotechnology Research, vol. 6, no. 1, pp. 16–26, 2024.

[9] F. A. Adrah, M. K. Denu, and M. A. E. Buadu, "Nanotechnology applications in healthcare with emphasis on sustainable covid-19 management," Journal of Nanotechnology Research, vol. 5, no. 2, pp. 6–13, 2023.

[10] L. A. Waller, B. P. Carlin, H. Xia, and A. E. Gelfand, "Hierarchical spatio-temporal mapping of disease rates," Journal of the American Statistical association, vol. 92, no. 438, pp. 607–617, 1997.