Journal of 21 September University
for Medical and Applied Sciences
Vol. 5 (1): 24-34
doi: 10.65693/JUMAS.2026.5.1.24

Original Research Article

A Machine Learning-Based Epidemiological Analysis of Cancer Distribution in Yemen

Abdulrahman M. H. Obaid[1]*, Gameil S. H. Ali[2], Yousif A. Alhaj[3], Awadh Ali Abdo Mohammed[4].

[1,2,3]Department of Information Technology, 21 September University of Medical and Applied Sciences, Sana'a, Yemen.

[4]Department of Biomedical engineering, 21 September University of Medical and Applied Sciences, Sana'a, Yemen.

[*]**Corresponding Author:** Abdulrahman M. H. Obaid: Email: obaid@21umas.edu.ye

Article History| Received: 01.011.2025 | Accepted: 5.01.2026 | Published: 5.05.2026

Abstract

Cancer represents a severe public health crisis in Yemen. Current literature lacks the regional granularity required for effective policy. To bridge this gap, the authors parsed 5,226 clinical records from the National Oncology Centre, mapping the spatial and demographic dispersion of the disease across all governorates. After scrubbing raw data for inconsistencies, the authors deployed Random Forest classification and hierarchical clustering to quantify patient demographics, temporal shifts, and regional incidence rates. Cases cluster heavily in specific regions. Ibb, Taiz, and Dhamar alone account for over one-third of the total national caseload. Breast cancer is the primary diagnosis. Most patients fall within the middle-adulthood cohort. Hierarchical clustering partitioned the governorates into subgroups with shared oncological profiles, providing a framework for localized intervention. The machine learning models yielded a Top-1 accuracy of 0.192 and a Top-3 accuracy of 0.596. These metrics reflect the high diagnostic heterogeneity and the asymmetric distribution of cases inherent to the Yemeni landscape. This work establishes an empirical baseline for national cancer planning. These results enable evidence-based resource management and prevention programs tailored to regional epidemiological realities.

Keywords: Cancer epidemiology; Yemen; Machine learning; Random Forest; Cancer prevalence; Geographical distribution.

Introduction

Cancer claims millions of lives annually, driven by a volatile mix of genetic predisposition, environmental exposure, and socioeconomic barriers to care [1]. National health strategies fail without a precise map of geographic and demographic disease distribution [2]. In resource-constrained settings, this lack of data is catastrophic.

Yemen faces a massive oncology burden that remains largely invisible to the international community. Fragmented healthcare delivery and the absence of a unified national registry preclude a clear epidemiological census [3]. Currently, lack of fundamental metrics on incidence, spatial clustering, and subtype prevalence across the governorates. Health policy cannot be formulated in a vacuum. Effective screening and resource allocation require high-resolution data to target the areas of greatest clinical need.

Computational advances provide new methods for interpreting messy clinical datasets [4]. Random Forest algorithms allow researchers to classify complex patient profiles and identify regional risk signatures [5]. By applying hierarchical clustering and Principal Component Analysis (PCA), authors can partition geographic regions into subgroups with shared oncological characteristics. This moves public health from broad guesswork to data-driven precision.

This study maps the spatial dispersion of cancer across Yemen. The authors quantified prevalence at the governorate level to isolate high-incidence hotspots. Scrutinised patient age and sex distributions against geographic coordinates to detect temporal shifts in incidence. To predict cancer types from clinical inputs, the authors trained machine learning models and evaluated their reliability using Top-1, Top-2, and Top-3 accuracy metrics. Hierarchical clustering and PCA identified governorates with nearly identical disease signatures. These analyses provide a high-resolution view of regional risk previously unavailable in the Yemeni context.

Integrating predictive modelling with spatial visualisation creates an empirical baseline for future research. The findings offer health authorities the resolution needed for strategic planning and early detection programs. Authors generated uncertainty-aware maps of cancer incidence to identify risk correlates within a severely limited resource environment

The paper proceeds as follows. Section 2 reviews related literature. Section 3 describes data collection and methods. Section 4 presents the results. Section 5 discusses policy implications. Section 6 provides conclusions and recommendations for future research.

Literature Review

Yemeni oncology relies on a patchwork of geographically restricted registries. The Aden Cancer Registry (1997–2011) highlights breast cancer as the dominant female malignancy, while leukaemia and non-Hodgkin lymphoma lead among men [6]. Pediatric data from Hadhramout and Aden echo these trends. Haematological malignancies account for a significant portion of childhood caseloads [7]. These localised datasets fail to capture a national picture. Conflict, service disruption, and systemic underdiagnosis compromise their reliability. GLOBOCAN 2020 estimates a national age-standardised incidence of 97 per 100,000 [8]. This figure likely masks massive under-

reporting driven by a crumbling health infrastructure and broken vital statistics systems.

Spatial epidemiology provides the tools needed to overcome these data voids. Standard Geographic Information System (GIS) methods detect clusters but often succumb to spatial autocorrelation and the modifiable areal unit problem (MAUP) [9]. Bayesian hierarchical frameworks offer more resilience. Models like Besag–York–Mollié (BYM) and Directed Acyclic Graph Autoregressive (DAGAR) structures smooth risk estimates by accounting for latent spatial dependence [10,11]. High-resolution risk surfaces emerge when using point-process models, such as the log-Gaussian Cox process, which fuse spatial covariates with Gaussian random fields [12].

Machine learning (ML) pushes these analytical boundaries. Integrating spatial structures into predictive models enhances both interpretability and predictive power [13]. Geospatial AI, exemplified by the iCAT platform, now maps disparities to direct resource flow [14]. In urban health research, investigators use XGBoost to parse the intersection of built environments, socioeconomic variables, and cancer prevalence [15].

Yemen remains untouched by these methodological shifts. No published literature currently applies hybrid ML-spatial models or Bayesian mapping to Yemeni cancer data. The barriers are steep. Researchers face a total absence of national registries, extreme under-reporting, and a lack of granular covariates. Deploying ML-based spatial frameworks could revolutionise the field in this area. It would allow authors to quantify uncertainty and generate the high-resolution evidence required for survival-critical public health planning.

Methods

Data Source

This study analysed 5,226 cancer patient records from the National Oncology Centre in Yemen in a retrospective design. Each record included demographic information such as age and sex, geographical data including governorate of residence, clinical diagnosis, cancer type, and registration date. The authors used this dataset to examine cancer prevalence, demographic patterns, geographical distribution, and temporal trends across governorates.

Data Preprocessing and Cleaning

Before to analysis, the dataset underwent extensive preprocessing to improve data quality and consistency:

Standardized all column names using case-insensitive mapping to account for variations across datasets.

Converted age values to numeric format and treated any age below zero or above 120 years as missing.

Formatted categorical variables as strings, including sex, governorate, diagnosis, and cancer type, and removed leading or trailing spaces from these fields.

Parsed registration dates into datetime objects and assigned null values to missing or invalid date entries.

Grouped cancer types with fewer than two recorded cases to address class imbalance in the machine learning stage.

These steps produced a clean and consistent dataset suitable for statistical analysis and predictive modelling.

Descriptive Statistical Analysis

Descriptive statistics summarised the dataset and offered an overview of cancer epidemiology in Yemen:

Demographics: Mean and median ages were calculated, and age distributions were reviewed for each cancer type. Sex distributions were assessed for each governorate.

Geographical Distribution: The prevalence of cancer in each governorate was measured as a percentage of total cases. The top ten cancer types were identified and visualized.

Geographical and Clustering Analysis

To study regional patterns, cancer cases were aggregated by governorate. Hierarchical clustering and Principal Component Analysis (PCA) identified groups of governorates with similar distributions of cancer types. Heatmaps, dendrograms, and PCA scatter plots were generated to visualise these relationships.

Machine Learning Analysis

A Random Forest classifier was applied to predict cancer types using patient demographic and clinical features. The workflow included:

- Feature Selection: Age, sex, governorate of residence, and diagnosis were used as predictors.
- Encoding: Categorical variables were converted into numerical formats through label encoding suitable for the model.
- Model Training and Evaluation:
  - The dataset was split into training (75%) and testing (25%) sets with stratification to preserve the original distribution of cancer types.
  - Class imbalance was addressed through class weighting in the Random Forest classifier.
  - Model performance was evaluated with Top-1, Top-2, and Top-3 accuracy metrics, which measured whether the true cancer type appeared among the model's highest-ranked predictions.

Confusion matrices were produced to summarise classification errors and overall prediction quality.

Software and Tools

All analyses were conducted using Python 3.11.The packages included Pandas, NumPy, Matplotlib, Seaborn, scikit-learn, and SciPy. Custom scripts handled data preprocessing and machine learning procedures. All results and figures were saved to a designated output directory to support reproducibility and accessibility.

Results

Demographic and Clinical Characteristics

The authors analysed 5,226 cancer cases. The mean patient age was 47.2 years, and the

median age was 51.0 years, indicating a predominance of middle-aged adults. The dataset included cases from 23 governorates. The overall sex distribution was nearly equal between male and female patients, as shown in Figure 1.

Abdulrahman M. H. Obaid[1]*, Gameil S. H. Ali[2], Yousif A. Alhaj[3], Awadh Ali Abdo Mohammed[4].

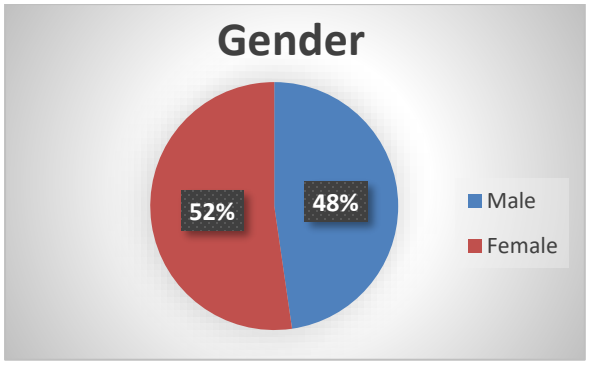

Figure 1: Gender Distribution of Cancer Patients in Yemen

Geographical Distribution of Cancer Cases

Cancer cases were not evenly distributed across the country. Five governorates accounted for the highest proportions: Ibb (14.27%), Taiz (13.62%), Dhamar (9.20%), Alhodeidah (8.90%), and the Capital City (7.75%). Figure 2 presents a bar chart of absolute case counts per governorate.

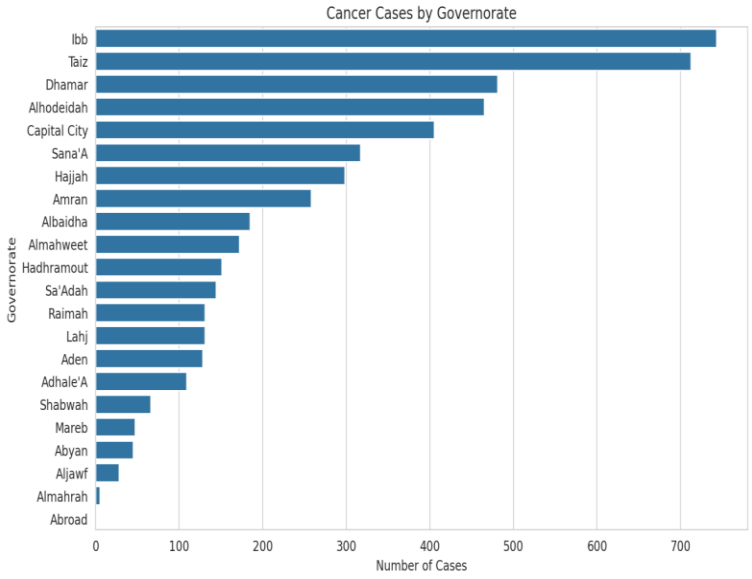

Figure 2: case counts per governorate

A heatmap in Figure 3 shows the prevalence of cancer types by region and reveals notable geographic variation in specific cancer incidences.

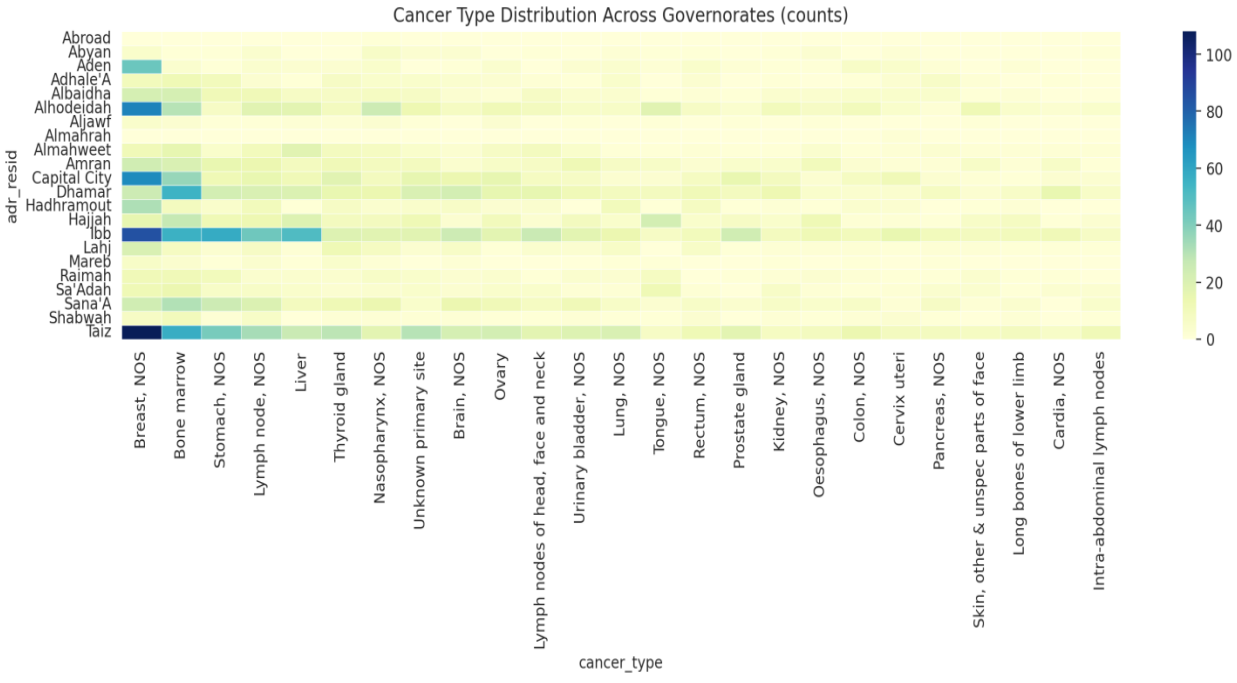

Figure 3: Cancer Type Distribution Across Governorates

Cancer Type Distribution

The ten most common cancer types represented nearly 50% of all cases. Breast cancer was the most frequent (12.11%), followed by bone marrow cancer (8.59%), stomach cancer (4.92%), and lymph node cancer (4.90%). Figure 4 provides a bar chart of the top ten cancer types, and Figure 5 presents their age distributions. Breast and thyroid cancers were more common among middle-aged adults, whereas brain and bone marrow cancers appeared across a wider age range.

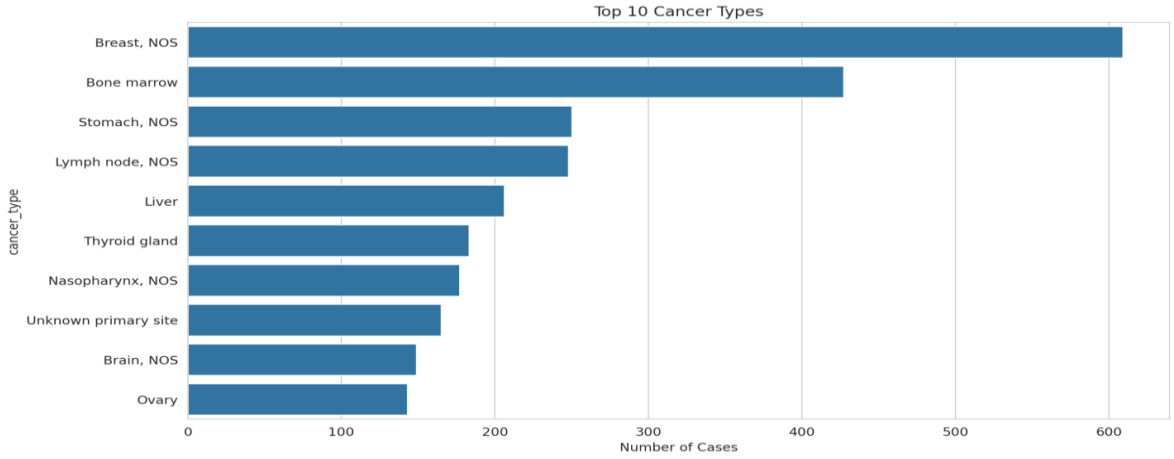

Figure 4: Top 10 cancer types by frequency

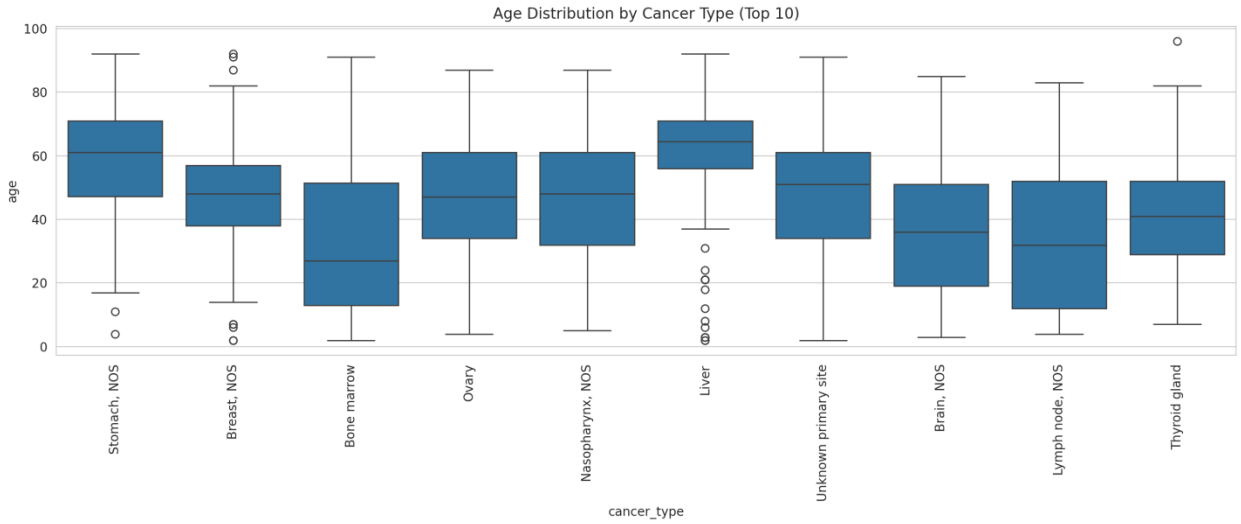

Figure 5: Age Distribution Per Cancer Type

Sex Distribution by Governorate

Sex distributions varied geographically. Figure 6 illustrates male and female case counts by governorate, highlighting regions in which certain cancers were more common in one sex than the other.

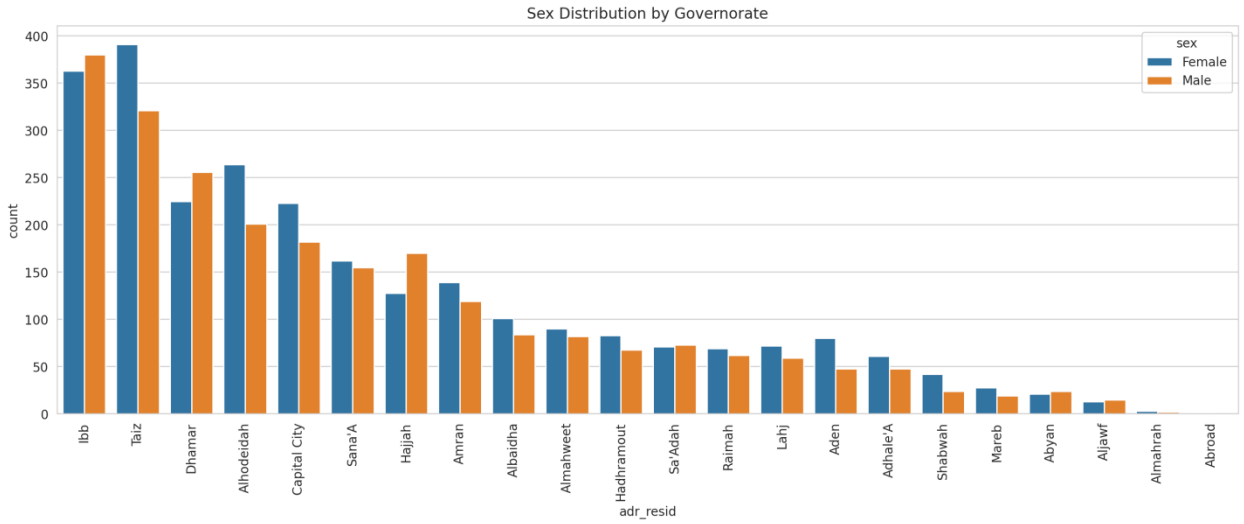

Figure 6: Sex distribution across governorates

Governorate Clustering

Hierarchical clustering grouped governorates with similar cancer profiles. Figure 7 displays the dendrogram of these clusters.

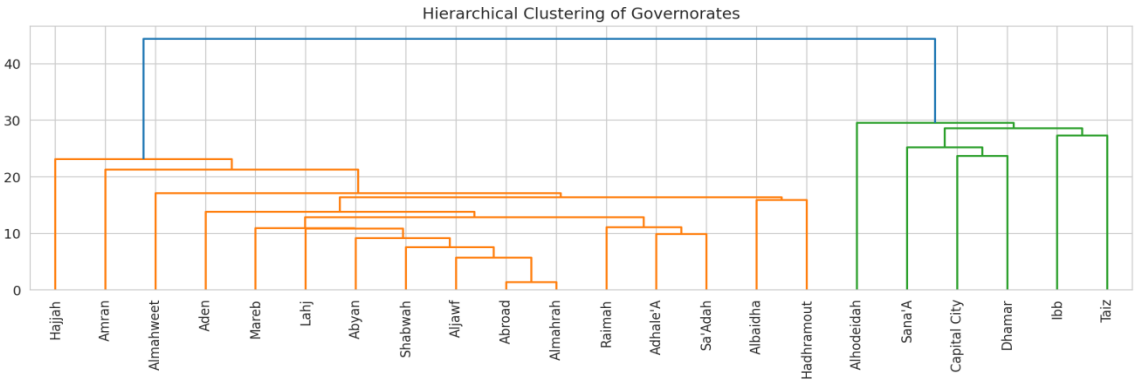

Figure 7: Hierarchical clustering dendrogram of governorates

Key Insights and Findings

Breast cancer is the most frequently occurring cancer type nationwide.

The governorates of Ibb, Taiz, and Dhamar account for a disproportionately high share of total cancer cases.

Cancer incidence rates peak among middle-aged adults.

The machine learning model recorded low Top-1 accuracy, although its Top-2 and Top-3 accuracies provide useful predictive information.

Age distribution patterns confirm the highest incidence in middle-aged adults.

Clustering and PCA results reveal regional patterns that may inform targeted public health interventions and resource allocation.

Discussion

Yemen's oncological landscape is starkly uneven. The authors mapped distinct geographic concentrations where Ibb, Taiz, and Dhamar carry a disproportionate burden, together representing a massive segment of the national caseload. These hotspots likely emerge from a nexus of environmental stressors, demographic shifts, and severe gaps in healthcare access. Hierarchical clustering identified shared risk profiles among governorates. National health policy must mirror this regional heterogeneity.

Patients averaged 47.2 years. Malignancy in Yemen primarily targets the economically active, middle-aged cohort. The authors scrutinised sex-specific variances across the governorates; these fluctuations likely reflect local disparities in diagnostic saturation or cultural barriers to health-seeking behaviour. Breast cancer dominated the results. Targeted screening remains non-negotiable. These metrics are showing Random Forest classification returned a Top-1 accuracy of 19.2%, scaling to 39.2% for Top-2 and 59.6% at the Top-3 level. While the model struggles with definitive first-rank classification, it effectively narrows the differential diagnosis. The authors view this as a clinical triage aid rather than an automated diagnostic replacement. It helps clinicians prioritise high-probability malignancies in data-sparse environments where specialist oversight is rare.Partitioning governorates by oncological signature offers a roadmap for intervention. Governorates within the same cluster require harmonised screening protocols and shared resource pools. Strategic planning should follow these data-driven groupings rather than arbitrary administrative boundaries.

The authors fused machine learning with spatial visualisation to parse a notoriously difficult and fragmented dataset. Significant limitations persist. Pervasive underreporting, particularly in conflict-affected regions with zero diagnostic infrastructure, likely skews the true incidence rates. Future iterations require stage-at-diagnosis data and longitudinal treatment histories to refine the predictive surface.

The findings dictate a shift in Yemeni health policy. High-prevalence governorates require immediate capacity expansion for both screening and palliative care. Integrating these predictive outputs into clinical workflows could optimise the utilisation of strained resources. Data, not intuition, must drive the national cancer response.

Conclusion

Developing nations face a surging cancer crisis that outpaces current diagnostic capacity. Mitigating this threat demands computational frameworks that can parse sparse clinical records to reveal the true geographic and demographic footprint of the disease. Where national registries are broken or fragmented, the authors must integrate statistical inference and machine learning to isolate high-risk populations and direct limited resources toward the most impacted governorates. Within this context, the authors mapped 5,226 clinical records from the National Oncology Centre, exposing sharp geographic fractures in Yemen's cancer landscape. Ibb, Taiz, and Dhamar represent a massive, disproportionate share of the national caseload. Malignancy peaks in the middle-adulthood cohort. Breast cancer remains the primary diagnostic challenge nationwide.

Random Forest classification struggled with Top-1 accuracy, a direct consequence of diagnostic heterogeneity and the inherent class imbalance within the dataset. Reliability climbed significantly at the Top-3 level. This performance profile suggests the model serves best as a clinical triage aid to narrow differential diagnoses in data-scarce environments. The authors partitioned governorates into distinct subgroups through Principal Component Analysis (PCA) and hierarchical clustering. These groupings move the discourse beyond arbitrary administrative boundaries toward an empirical, profile-based health strategy.

Public Health Recommendations

- Deploy localised screening and early detection programs in identified hotspots like Ibb and Taiz.
- Prioritise breast cancer diagnostics and treatment infrastructure as a national mandate.
- Overhaul national registry protocols to mitigate systemic under-reporting and data lag.
    - Align resource distribution with the specific oncological signatures of each governorate cluster.
    - Implement predictive tools to support primary care clinicians in remote, specialist-deprived regions.

Future Research Directions

- Scrub and integrate genetic, environmental, and socio-economic variables to refine predictive surfaces.
- Investigate the specific barriers to healthcare access driving regional disparity.
- Apply advanced ensemble learning and cost-sensitive frameworks to improve classification thresholds.
- Benchmark Yemeni epidemiological patterns against neighbouring regional data to isolate shared risk correlates.

This study establishes a rigorous evidentiary baseline for Yemeni oncology. The Authors validated the use of machine learning to navigate data-scarce, resource-limited environments. Policy must now shift toward data-driven allocation to optimise survival outcomes. Data, not intuition, must drive the national cancer response.

Acknowledgements

The authors thank the National Oncology Centre (NOC) in Yemen for providing the data that enabled this research. Authors also appreciate the dedicated staff of the NOC registry for their careful work in collecting and curating the data.

Data Availability Statement

The dataset for this article came from the National Oncology Centre (NOC), Yemen, under a data use agreement for this study. It is not publicly available to protect patient confidentiality. Anonymised data may be provided by the corresponding author upon reasonable request and with permission from the NOC institutional review board.

Funding Statement

This research did not receive any specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest Disclosure

The authors declare there are no conflicts of interest related to the publication of this paper.

Patient Consent Statement

For this retrospective study, authors used pre-existing, fully de-identified patient data.

Permission To Reproduce Material From Other Sources

This manuscript does not include any material reproduced from other sources.

Clinical Trial Registration

This study is not a clinical trial and was not registered.

References

1. Ajayi, R. O., & Ogunjobi, T. T. (2025). Environmental exposures and cancer risk: a comprehensive review. Medinformatics, 2(2), 80-92.
2. Setyawati, R., Astuti, A., Utami, T. P., Adiwijaya, S., & Hasyim, D. M. (2024). The importance of early detection in disease management. Journal of World Future Medicine, Health and Nursing, 2(1), 51-63.
3. Mansour, R., Abdel-Razeq, H., Al-Hussaini, M., Shamieh, O., Al-Ibraheem, A., Al-Omari, A., & Mansour, A. H. (2024). Systemic barriers to optimal cancer care in resource-limited countries: Jordanian healthcare as an example. Cancers, 16(6), 1117.
4. Arora, A., & Basu, N. (2023). Machine learning in modern healthcare. International Journal of Advanced Medical Sciences and Technology, 3(4), 12-18.
5. Wiratama, R. K. P., Cahyadi, E. S., Meshcherekov, D., & Purwitasari, D. (2024, June). Random forest based risk factor analysis for lung cancer prediction. In 2024 International Conference on Smart Computing, IoT and Machine Learning (SIML) (pp. 62-67). IEEE.
6. A. A. Bawazir, "Cancer incidence in Yemen from 1997 to 2011: A report from the Aden cancer registry," BMC Cancer, vol. 18, p. 540, 2018.
7. M. A. Jawass et al., "Pattern of malignancies in children <15 years of age reported in the Hadhramout cancer registry, Yemen between 2002 and 2014," Saudi Medical Journal, vol. 37, no. 5, pp. 513–520, 2016.
8. A. Ibrahim et al., "Cancer statistics in Yemen: Incidence and mortality, in 2020," BMC Public Health, vol. 24, p. 962, 2024.
9. J. W. Bauer et al., "Geographic Information Systems and spatial analysis in cancer epidemiology," Cancer Epidemiology, Biomarkers & Prevention, vol. 28, no. 3, pp. 333–339, 2019.
10. L. Gao, A. Datta, and S. Banerjee, "Spatial modeling for correlated cancers using bivariate directed graphs," arXiv preprint arXiv:1911.11342, 2019.
11. Y. Wang, X. Chen, and F. Xue, "A review of Bayesian spatiotemporal models in spatial epidemiology," ISPRS International Journal of Geo-Information, vol. 13, no. 3, p. 97, 2024.
12. F. Palmi-Perales et al., "Approximate Bayesian inference for multivariate point pattern analysis in disease mapping," arXiv preprint arXiv:1903.11647, 2019.
13. N. Kianfar, B. Sartorius, C. L. Lau, R. Bergquist, and B. Kiani, "The future of spatial epidemiology in the AI era: Enhancing machine learning approaches with explicit spatial structure," Geospatial Health, 2025.
14. R. K. McIntire et al., "Utilizing geospatial artificial intelligence to

map cancer disparities across health regions: The iCAT tool," International Journal of Environmental Research and Public Health, vol. 21, no. 4, 2024.

15. C. Liu and A. Mostafavi, "Decoding Urban-Health Nexus: Interpretable machine learning illuminates cancer prevalence based on intertwined city features," arXiv preprint arXiv:2306.11847, 2023.