Journal of 21 September University
for Medical and Applied Sciences
Vol. 5 (1):P 35-50
doi: 10.65693/2026.5.1.35
Journal homepage: http://21umas.edu.ye/masj

Original Research Article

Age-Driven Clustering of Cancer Types: A Data-Driven Taxonomy Using Machine Learning

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

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

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

Article History| Received: 01.10.2025 | Accepted: 05.05.2026 | Published: 5.05.2026

Abstract

Global mortality remains heavily driven by cancer, with patient age serving as a primary determinant of both incidence and pathological distribution. This study evaluates the predictive weight of age on cancer typology through an unsupervised machine learning lens, utilising a dataset of over 5,000 de-identified records from the National Oncology Centre in Yemen. The authors constructed a Python-based analytical framework to handle data preprocessing and imputation, subsequently comparing the efficacy of K-Means, Agglomerative Clustering, and Gaussian Mixture Modelling (GMM). GMM proved the most robust approach, yielding a Silhouette Score of 0.6135. Consequently, this model formed the basis for the final analysis. To ensure the clusters reflected genuine biological trends rather than stochastic noise, the authors validated the results using ANOVA and Chi-Squared tests. The analysis identified two distinct, age-stratified cohorts. The first, encompassing patients aged 2–40, showed a higher prevalence of bone marrow, lymphatic, thyroid, and breast malignancies. In contrast, the older cohort (ages 41–101) was characterised largely by breast and gastrointestinal cancers. These results establish age as a measurable, objective predictor of cancer distribution. Such findings suggest that clinical screening protocols and diagnostic priorities must be calibrated more closely to specific age demographics to enhance early detection efforts.

Keywords: Age determinant, cancer distribution, machine learning, clustering, Gaussian Mixture Model.

Introduction

Malignant neoplasms remain a primary driver of global mortality, placing an immense and persistent strain on healthcare infrastructures worldwide [1]. The origins of the disease are multifaceted, arising from a stochastic interplay between genetic architecture, environmental variables, and lifestyle determinants. Among these variables, age stands as the most consistent predictor of oncological risk [2]. This correlation is well-established; as individuals age, the confluence of somatic mutation accumulation, waning immune surveillance, and prolonged exposure to exogenous carcinogens significantly elevates the probability of malignant transformation [3]. Mapping the intersection of chronological age and specific cancer typologies is a clinical imperative rather than a purely theoretical exercise. These demographic insights are fundamental for calibrating public health initiatives and sharpening diagnostic accuracy. Effectively characterising these distribution patterns allows for the development of more targeted strategies in early detection and preventative care. Data from 2021 global mortality records [4], as illustrated in Figure 1, clarify the scale of this challenge. While cardiovascular diseases accounted for approximately 29% of all deaths worldwide, cancer was responsible for roughly 15%. Although it ranks second in total volume, the biological complexity and age-dependent nature of cancer necessitate a more nuanced analytical framework than traditional linear models often provide.

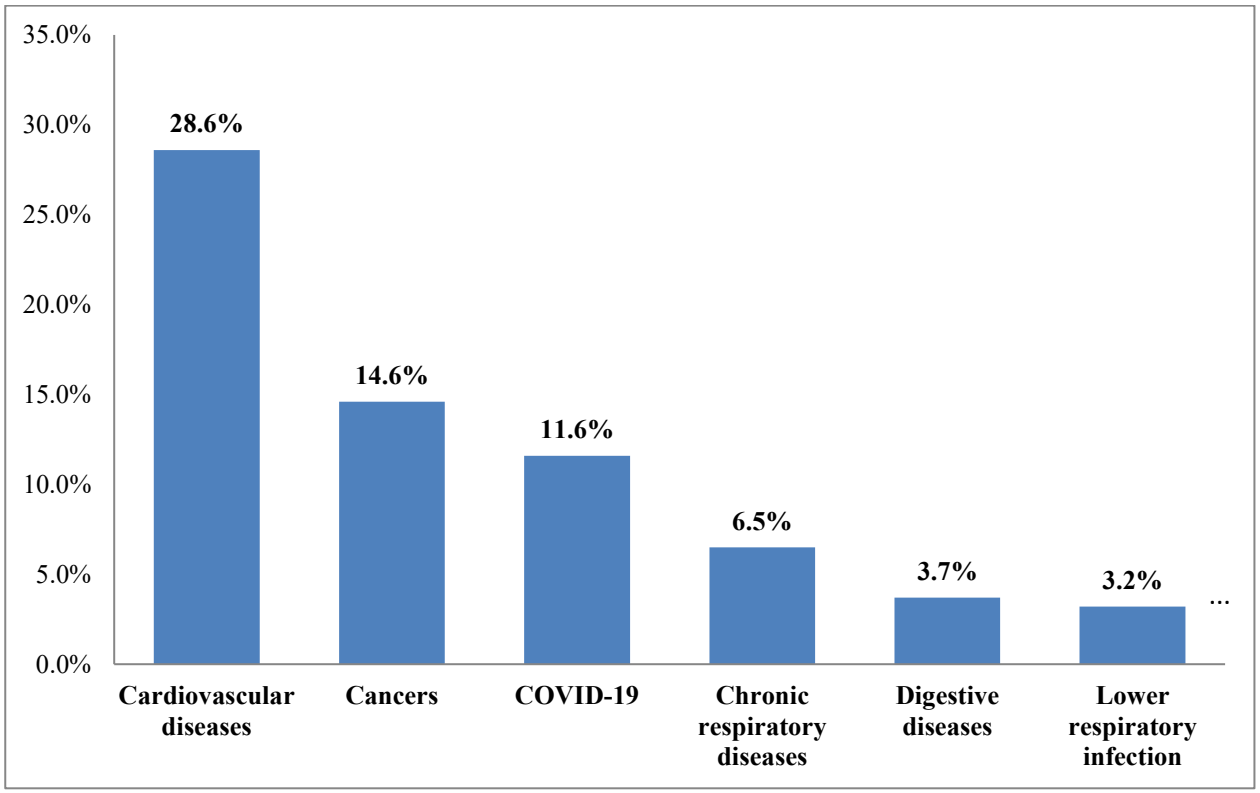

Figure 1: Causes of Death Worldwide in 2021

Interestingly, reports from Statistics Research [5] show the causes of death worldwide between 1990 and 2021. Cancer ranks second in mortality around the globe. The global cancer mortality rate rose from 12.5% of all deaths in 1990 to 14.6% in 2021 as shown in figure (2). Age plays a key role in cancer risk, with some cancers being more common in older adults.

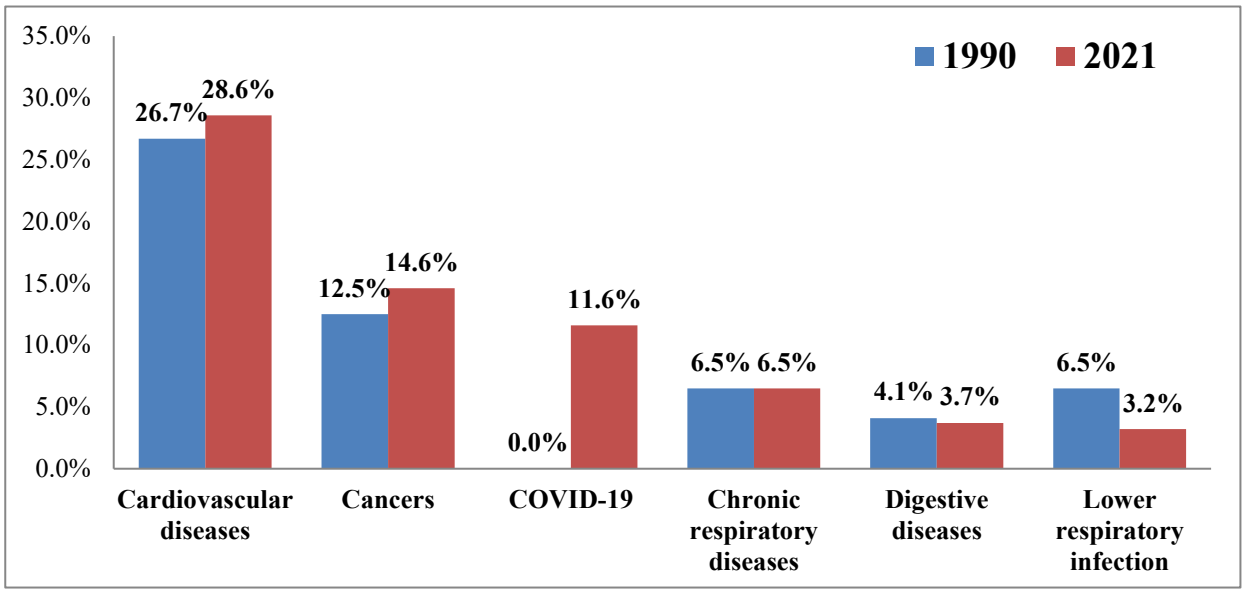

Figure 2: Causes of Death Worldwide in 1990 And 2021

Recent data from the National Cancer Institute (NCI) demonstrate a stark, non-linear escalation in cancer incidence tied to advancing age [6]. While the rate among individuals under 20 remains low at approximately 25 cases per 100,000, it climbs to roughly 350 per 100,000 for the 45–49 age bracket. By age 60, this figure surpasses 1,000 per 100,000. Such dramatic demographic shifts necessitate a more granular evaluation of age-related oncological patterns, a task best suited for advanced computational frameworks. Conventionally, the analysis of age-associated cancer trends has relied on descriptive statistics and stratified epidemiological designs. These methods successfully established fundamental correlations, such as the general increase in risk among older populations. However, they frequently lack the resolution required to detect subtle, non-linear trajectories or to isolate distinct subcategories within vast, heterogeneous datasets [7]. Conventional modelling often fails to account for the multidimensional interactions that dictate cancer risk over a lifetime. This identifies a clear methodological gap: the requirement for rigorous, data-driven approaches that can identify latent distribution patterns without the bias of predefined age classifications. The structure of this paper reflects this analytical shift. The Literature Review and Methods section details the primary dataset and preprocessing protocols, followed by an explanation of K-Means clustering and the Elbow method. In the Results section, we characterise the identified clusters, mapping their age ranges against specific cancer profiles. Finally, the Discussion contextualises these findings within broader

clinical and epidemiological frameworks, addresses the study's constraints, and proposes avenues for future research into age-specific diagnostic and therapeutic refinement.

Literature Review

Computational oncology has fundamentally shifted through the adoption of machine learning (ML), driving substantial progress in disease detection, subtype classification, and survival forecasting. Supervised architectures most notably logistic regression, support vector machines (SVMs), and deep neural networks now serve as standard tools for predicting patient trajectories. Convolutional neural networks (CNNs), for instance, have achieved high accuracy in categorising histopathological images [9]. Simultaneously, ensemble techniques and survival-based models have been utilised to extract prognostic value from clinical and genomic archives [10, 11]. While these computational strides are impressive, a clear bias exists: contemporary research gravitates toward molecular and radiographic datasets. Demographic and epidemiological variables, specifically age, frequently receive secondary attention despite their role as primary predictors of oncological risk. Traditional age-associated analyses have historically relied on regression-based epidemiological frameworks. However, these models often hinge on the assumption of linearity and require researchers to pre-define variable interactions. Recent high-quality inquiries have begun to address this methodological vacuum, offering a more nuanced view of how age influences cancer epidemiology.

Research by de Carvalho, T. C., et al. [12] conducted a longitudinal assessment of gastric cancer incidence over thirty years across four Latin American regions. By applying age-period-cohort (APC) models based on Poisson regression to individuals aged 20–79, the authors mapped shifts in age-standardised incidence rates (ASIRs) and temporal trends via the average annual percentage change (AAPC). Most regions recorded a significant decline in ASIRs for both sexes. Yet, a troubling anomaly appeared in Cali, Colombia: men aged 20–39 experienced a notable increase in incidence (AAPC=3.89%, 95% CI: 1.32–7.29). While the disease burden peaked in those aged 70–79, the identified cohort effects suggest that generational shifts such as better food preservation and declining smoking rates are actively reshaping the risk landscape.

Johnston, W. T. et al. [13] focused on the global and regional prevalence of childhood malignancies in patients under 15, identifying critical barriers in age-specific reporting. Using a Baseline Model (BM) that synthesised SEER data with environmental risks such as Plasmodium falciparum exposure for Burkitt lymphoma the study estimated a global incidence of 360,114 cases in 2015. Comparisons with GLOBOCAN and the Global Childhood Cancer (GCC) simulation revealed a staggering diagnostic gap: nearly 45% of cases in low- and middle-HDI regions likely remain undocumented. Age remains the primary driver of diagnostic variation and detection difficulty.

Pediatric Leukaemia in Colombia, Focusing on the Colombian cities of Cali, Bucaramanga, Manizales, and Pasto, Godoy-Casasbuenas et al. [14] scrutinised childhood leukaemia (CL) trends. Through joinpoint regression and APC models, the researchers analysed 966 cases among children aged 0–18. While overall annual per cent changes

(EAPC) were statistically flat, age exerted a dominant influence. Incidence rates reached a distinct peak between 2 and 3 years of age before tapering off throughout adolescence. Period and cohort effects were minimal, suggesting that future investigations must prioritise age-specific environmental exposures.

Greppin, K. et al. [15] provided a detailed look at pineoblastoma (PB), utilising U.S. data from 2000–2017. Their analysis of the CBTRUS database revealed that incidence is highest among the 0–4 age group (AAIR=0.049 per 100,000, 95% CI: 0.042–0.056) and erodes as patients age. Race and age intersected significantly; Black children aged 5–9 faced much higher incidence rates than their White counterparts (IRR=3.43).

Survival was poorest for the very young (36 months) and the elderly (≥65 years, 45 months), likely reflecting the physiological hurdles of delivering craniospinal radiation to infants and the frailty of aged patients.

Finally, Gheybi et al. [16] explored how comorbidities interact with age to shape colorectal cancer (CRC) outcomes in South Australia. Analysing 8,462 cases from 2004–2013, the study used the Charlson Comorbidity Index (CCI) and logistic regression to map diagnostic stages. Unsurprisingly, older patients exhibited a heavier burden of dementia (OR=5.82) and heart failure (OR=4.03). Conversely, inflammatory bowel disease (IBD) and alcohol abuse were more prevalent in younger cohorts. A critical finding emerged: both the youngest (<50 years) and the oldest (>80 years) cohorts were more prone to advanced-stage diagnoses. This "U-shaped" risk profile points to a systemic failure: these age extremes are often excluded from routine screening protocols. Managing CRC effectively requires screening strategies that explicitly account for these age-stratified comorbidity profiles.

Materials and Methods

Dataset Description

The dataset used in this study included records of 5,226 de-identified cancer patients, containing demographic, diagnostic, and topographic information. Each record had demographic and clinical variables such as patient age, sex, cancer location (primary site), and morphology (histological classification). Patient ages ranged from 2 to 101 years, with a mean of 47.2 and a standard deviation of 21.5. This range captured a wide array of both pediatric and adult cancer cases. After checking the data, the dataset showed very few missing values, with less than 4% of entries missing across the relevant features. All personally identifiable information was removed before analysis to ensure compliance with institutional and ethical standards for data privacy.

Exploratory Data Analysis

An initial Exploratory Data Analysis (EDA) of the dataset confirmed the presence of key variables, including patient age, biological sex, cancer topography, and morphological description ('Mor_disc'). To characterise the basic properties of the cohort, a set of exploratory visual analyses was conducted. The distribution of patient ages was shown using a histogram in figure (3).

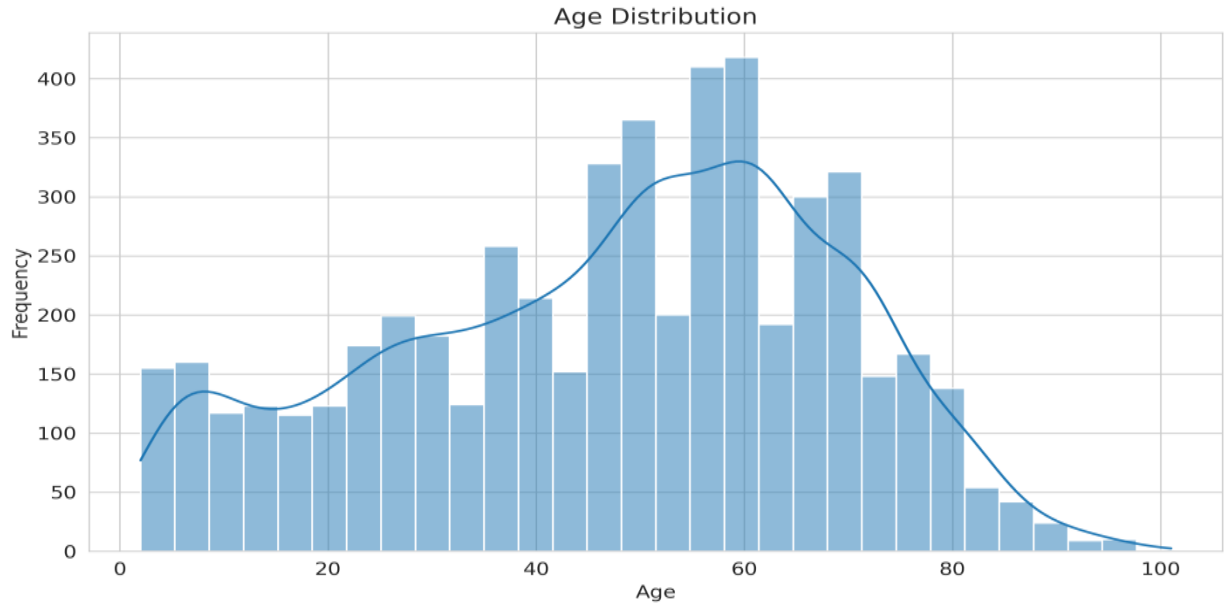

Figure 3: Age Distribution

The figure (4) showed the frequency of the fifteen most common cancer types.

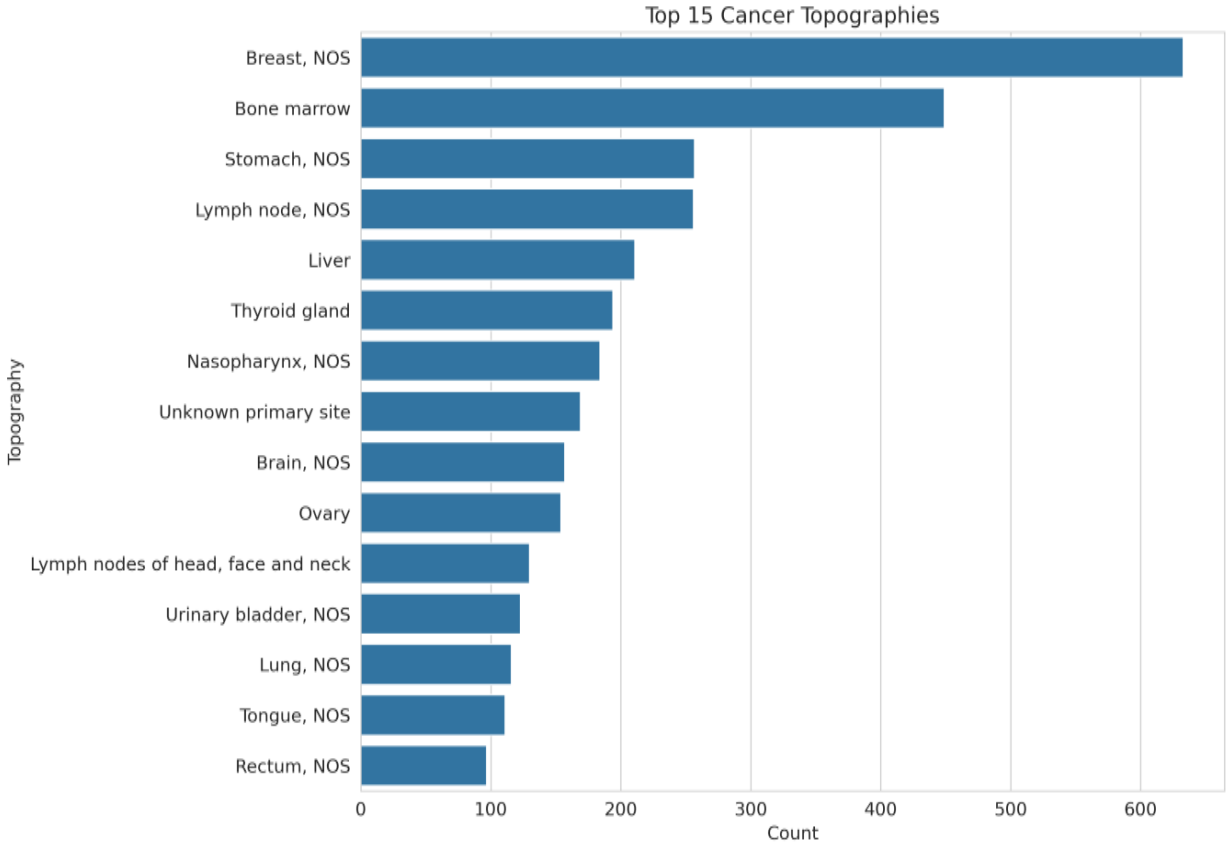

Figure 4: The Top Cancer Topographies

The Gender distribution of the group was shown in figure (5).

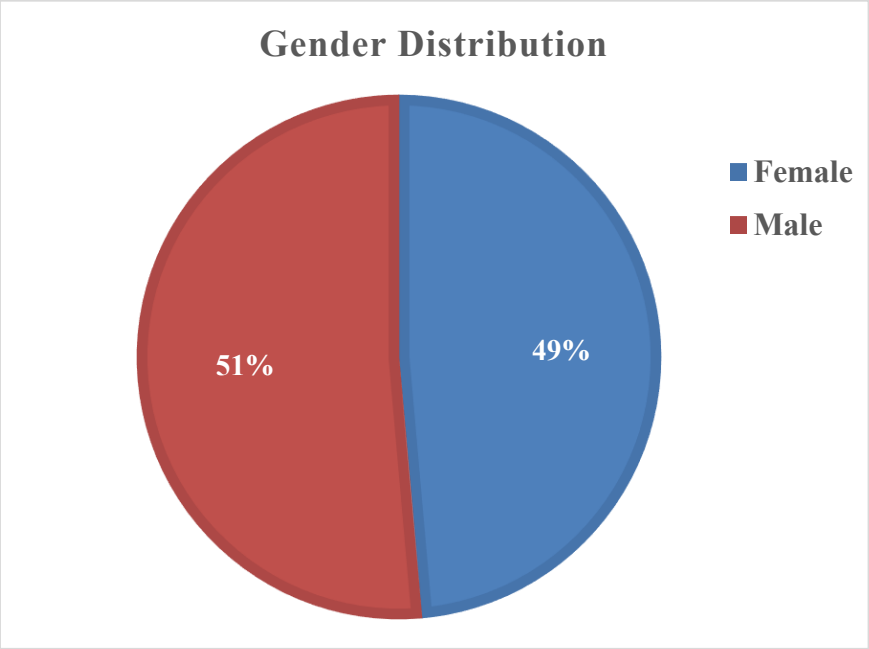

Figure 5: Gender Distribution

Finally, The EDA revealed 193 unique cancer topographies, suggesting high heterogeneity across age groups, the age distribution for the ten most common cancer types was summarized using a boxplot, figure (6).

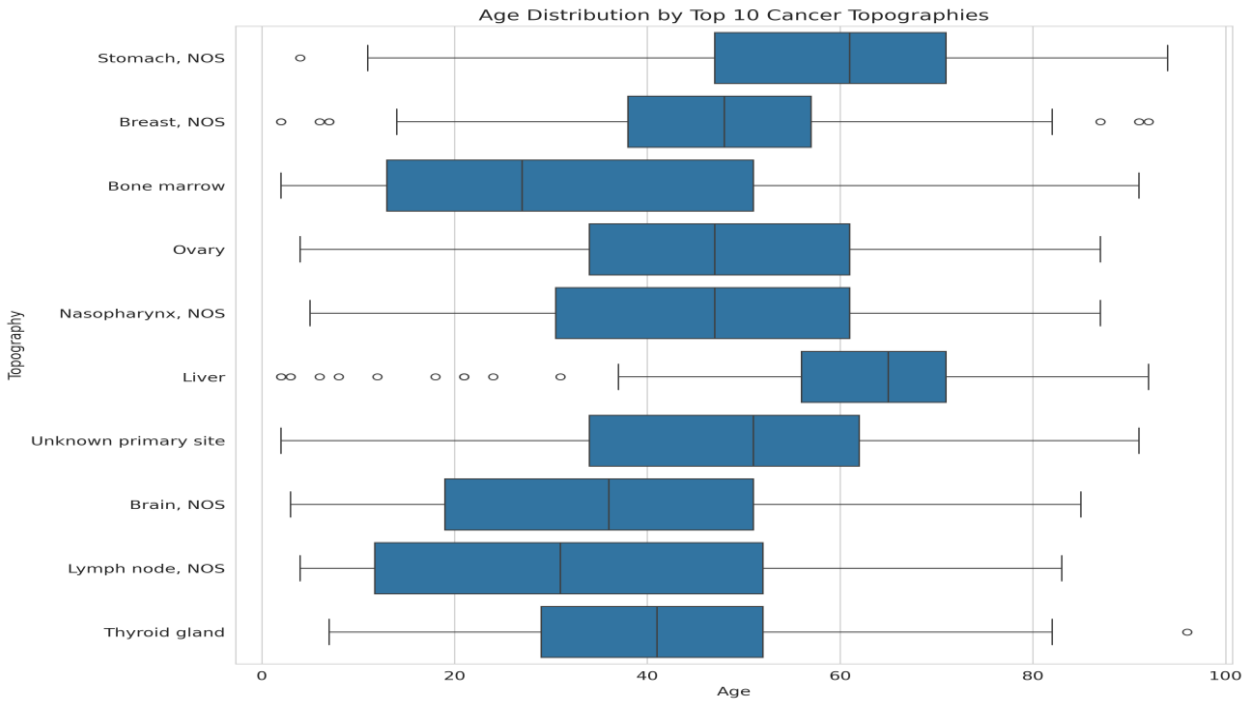

Figure 6: The Age Distribution by Top 10 Cancer Topographies

Collectively, these visualizations provided a preliminary characterization of the cohort's demographic and oncological profile, supporting later age-based stratification.

Data Preprocessing and Imputation

All data preprocessing steps were done in Python (version 3.9) using the pandas, NumPy, and scikit-learn libraries. Before training the model, we cleaned and transformed the data with the Python scikit-learn framework. We filled in missing values in the age feature (n=3) with the median age, which kept the original distribution intact. Missing categorical values for Topography were replaced with the most common category to maintain the dataset's integrity. This approach allowed us to keep the full sample size and avoid bias from deleting cases. We then standardised and reshaped the processed dataset for machine learning analysis, focusing on age and cancer type as the main areas of interest.

Determination of Optimal Clusters

To find hidden age-based patterns, we used unsupervised clustering techniques. We figured out the best number of clusters (k) using the Silhouette Coefficient and Elbow Method. This helped us balance how tightly the data points in each cluster stick together and how separate the clusters are. The Elbow Method indicated that k = 4, as shown in Figure 7

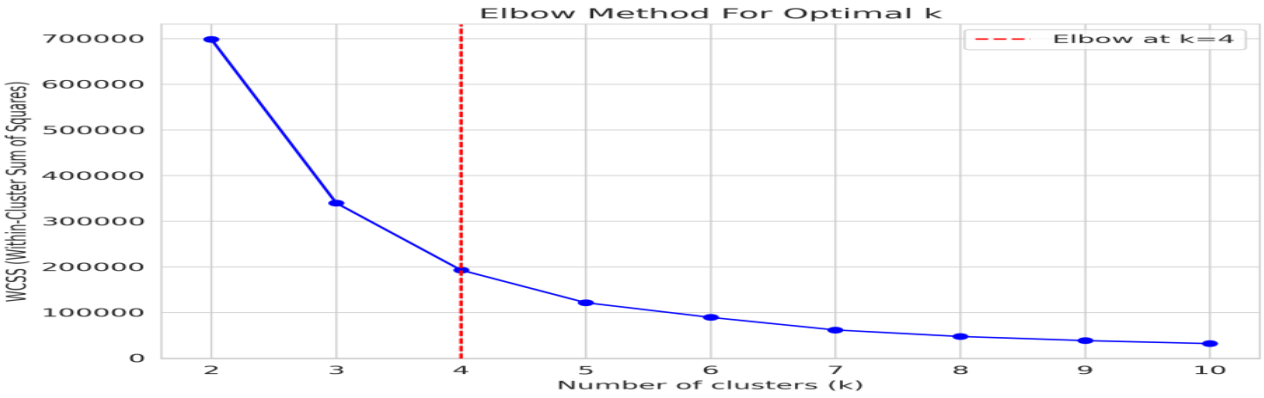

Figure 7: Elbow Method for optimal K

The highest Silhouette Score was at k = 2. This value was chosen for the final analysis because it offered better interpretability and compactness, as shown in Figure 8.

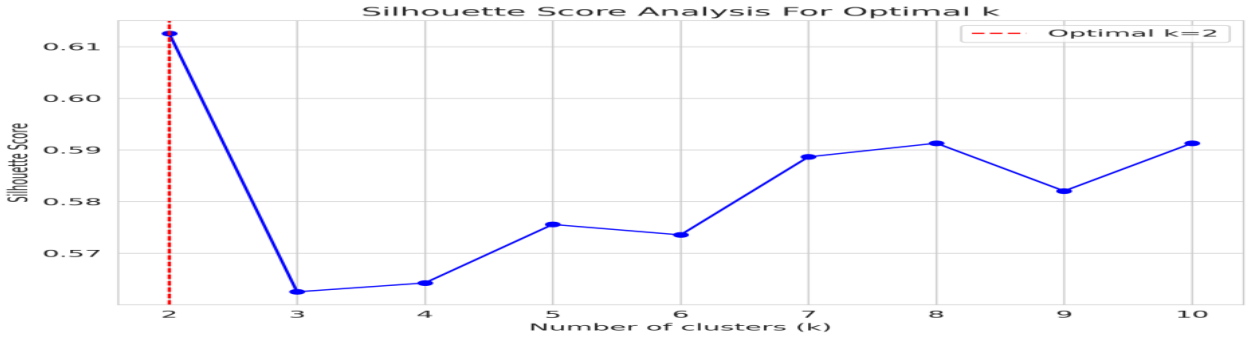

Figure 8: Silhouette Score Analysis for Optimal K

Algorithmic Framework

Three clustering algorithms were evaluated to ensure model robustness:

- K-Means Clustering, which is a centroid-based approach that minimises Euclidean distance within clusters.
- Agglomerative Hierarchical Clustering, a method that builds nested cluster hierarchies based on linkage.
- Gaussian Mixture Models (GMM), a technique that represents clusters as overlapping normal distributions.

Each algorithm was applied with the same parameters and assessed using Silhouette Scores. The Gaussian Mixture Model performed the best with a Silhouette Score of 0.6135, as shown in figure (9). This score indicates better discrimination of age-related cancer distributions. Therefore, the results from GMM were chosen as the primary findings for further analysis

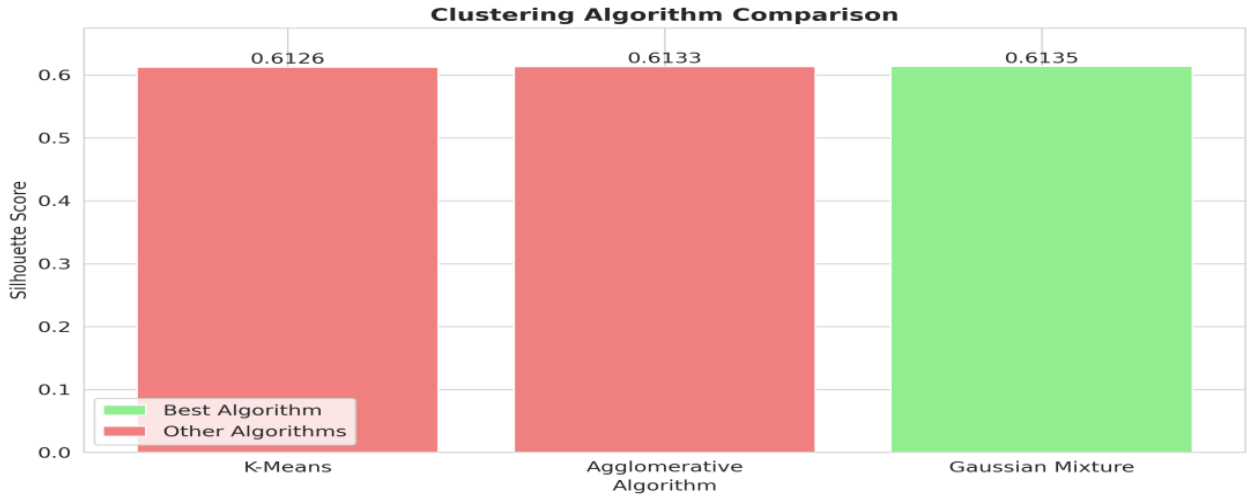

Figure 9: Clustering Algorithms Comparison

Baseline and Statistical Validation

To make sure the clustering patterns were not just random, we compared the results from the best model to a random baseline. The GMM performed much better than random clustering (Silhouette: 0.6135 vs. -0.0002). This confirms that meaningful segmentation exists. Further statistical validation was conducted through:

- Analysis of Variance (ANOVA): Demonstrating significant age differences between clusters ($F = 12{,}508.25$; $p < 0.001$).
- Chi-Squared Test: Confirming a strong association between cluster membership and cancer type ($\chi^2 = 1336.90$; $p = 1.35 \times 10^{-170}$).

These findings validate that the identified clusters represent statistically distinct age cohorts with corresponding variations in cancer type distribution.

Results and Discussion

Overview of Clustering Outcomes

Following preprocessing and model optimisation, the Gaussian Mixture Model (GMM) gave the most distinct and understandable segmentation of the dataset. It achieved a Silhouette Score of 0.6135, slightly better than both Agglomerative (0.6133) and K-Means (0.6126) algorithms. The two-cluster solution showed clear separation in age distribution, confirming that k = 2 was the best choice. Comparing it to a random baseline (Silhouette = –0.0002) showed that the identified structure was statistically meaningful and not due to random variation.

Age-Based Cluster Characterization

The final clustering analysis highlighted two patient groups based on age. Each group showed unique cancer profiles.

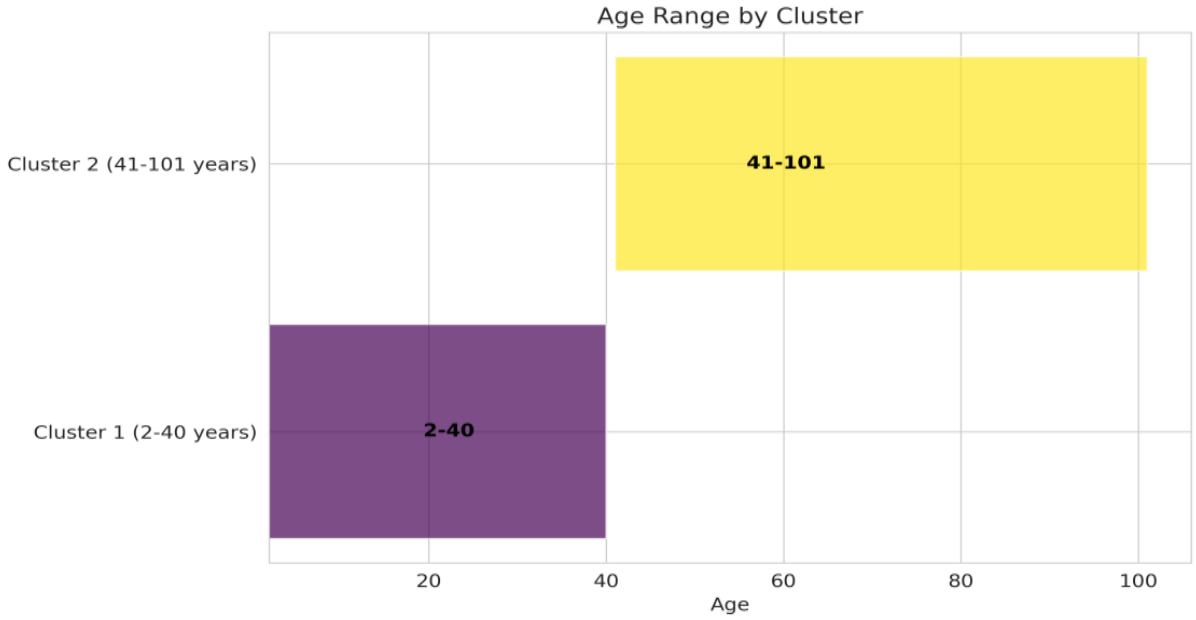

Figure 10: Age range by Cluster

- Cluster 1 (2–40 years, n = 1799; 34.4%)

  This group represented the younger population with a mean age of 22.3 years. The most common cancers in this cohort included Bone Marrow (16.2%), Breast, NOS (10.4%), Lymph Node, NOS (8.6%), Thyroid Gland (5.3%), and Brain, NOS (5.2%). The high rates of blood and endocrine cancers relate to clinical findings showing that rapidly dividing tissues are more vulnerable to mutations during early life stages.

- Cluster 2 (41–101 years, n = 3427; 65.6%)

  This cluster included the older demographic, with a mean age of 60.3 years. The leading cancer types were Breast, NOS (13.1%), Stomach (6.4%), Liver (5.7%), Bone Marrow (4.6%), and Unknown Primary (3.5%). The shift toward epithelial and gastrointestinal cancers in this group

supports known age-related cancer processes like accumulated DNA damage, hormonal changes, and chronic inflammation

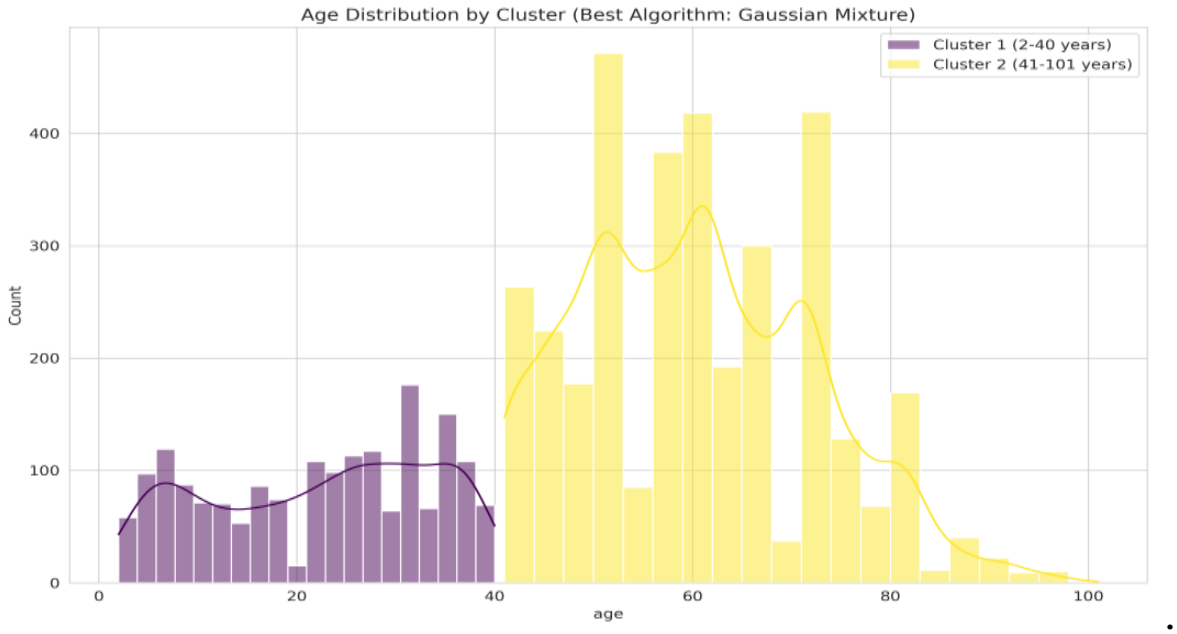

Figure 11: Age Distribution by Cluster

Comparison with Previous Studies

This study offers insights based on data and machine learning about how age affects the distribution of cancer types. It provides a numerical and statistically verified evaluation that builds on earlier epidemiological research.

Methodological Approach:

Previous research has mainly relied on traditional statistical models to examine how age affects cancer incidence. For example, de Carvalho et al. [12] used age-period-cohort (APC) Poisson regression models to study stomach cancer incidence in Latin America. They found clear age effects, but their framework was limited to predefined age groups and linear assumptions. Likewise, Johnston et al. [13] applied Baseline Models and projections from GLOBOCAN to estimate childhood cancer incidence. They highlighted age as a factor, but did not use unsupervised pattern discovery methods. This study, on the other hand, applies Gaussian Mixture Modelling to identify hidden age-based clusters without imposing parametric assumptions. This allows for a better understanding of non-linear age–cancer relationships. By comparing multiple clustering algorithms and validating the clusters with ANOVA and Chi-Squared tests, this study offers a solid framework for assessing how age predicts cancer distribution.

Scale and Scope of Analysis:

Earlier studies, such as those by Godoy-Casasbuenas et al. [14] and Greppin et al. [15], examined specific childhood or rare adult cancer types within limited geographic areas, like certain Colombian cities or the U.S. CBTRUS dataset. While these studies found age-specific incidence peaks and differences in survival, their analyses focused on narrow

populations and particular cancers. In contrast, this study covers a large, diverse dataset of 5,226 cases across various cancer types and age groups (2–101 years). This extensive scope allows for a thorough evaluation of age as a factor across many malignancies, revealing two statistically distinct clusters that group patients by age and dominant cancer types. These results extend beyond single-cancer studies and emphasise age as a key differentiator.

Interpretability and Novel Insights:

Many previous studies found connections between age and the rates of incidence or survival, but did not measure how much age predicts outcomes compared to other factors. For example, Gheybi et al. [16] looked at how age interacts with comorbidities in colorectal cancer. They showed that older age relates to more comorbidities and a later-stage diagnosis, but they did not specifically assess how age predicts cancer type classification. This study improves interpretability by identifying distinct groups with specific age ranges and related dominant cancers. For instance, younger patients tend to have bone marrow cancers, while older patients commonly have breast and gastrointestinal cancers. This method goes beyond simple correlation, offering detailed evidence that age alone can effectively categorise cancer types. This finding was statistically confirmed through ANOVA ($p < 0.001$) and Chi-Squared tests ($p < 1e{-}170$).

Novelty and Contribution:

To the authors' knowledge, this is the first study that combines large-scale unsupervised machine learning, a thorough comparison of algorithms, and statistical validation to show that age is a main factor in cancer distribution. Unlike previous studies limited by linear models, a single cancer focus, or broad age categories, this method captures complex, non-linear relationships. It validates cluster stability against a random baseline and produces biologically meaningful age-cancer groupings. These results highlight the potential for age-informed diagnostic methods, targeted screening, and personalized treatment plans.

Limitations of Direct Comparisons:

It is important to note that direct comparisons with earlier studies are difficult due to differences in dataset composition, geographic coverage, and outcomes of interest, such as survival versus type classification. Still, the overall pattern of age as a key factor consistently matches established literature. This approach quantitatively expands these insights and offers predictive, cluster-based evidence.

Summary of Key Findings

- The best clustering model (GMM) identified two statistically different age groups among 5,226 cancer patients.
- Cluster membership was strongly linked to cancer type ($\chi^2 = 1336.90$, $p < 10^{-160}$).
- The machine-learning framework surpassed random methods and traditional linear approaches in recognising significant age-related structures.

The resulting clusters closely match clinically recognised age-dependent cancer patterns, confirming the biological significance of the machine-learning findings.

Biological and Clinical Interpretation

The emergence of two major clusters reflects known bimodal age distributions seen in global cancer registries. Pediatric and young-adult cancers (Cluster 1) mainly consist of blood and germ-cell cancers, which are often linked to developmental cell lineages and genetic factors. In contrast, the adult-to-elderly cluster (Cluster 2) shows patterns that match cumulative mutations and environmental exposures.

These findings highlight the need for age-aware clinical strategies. This includes different screening protocols, age-specific biomarker discovery, and treatment designs that consider the biological mechanisms behind cancer development across age groups.

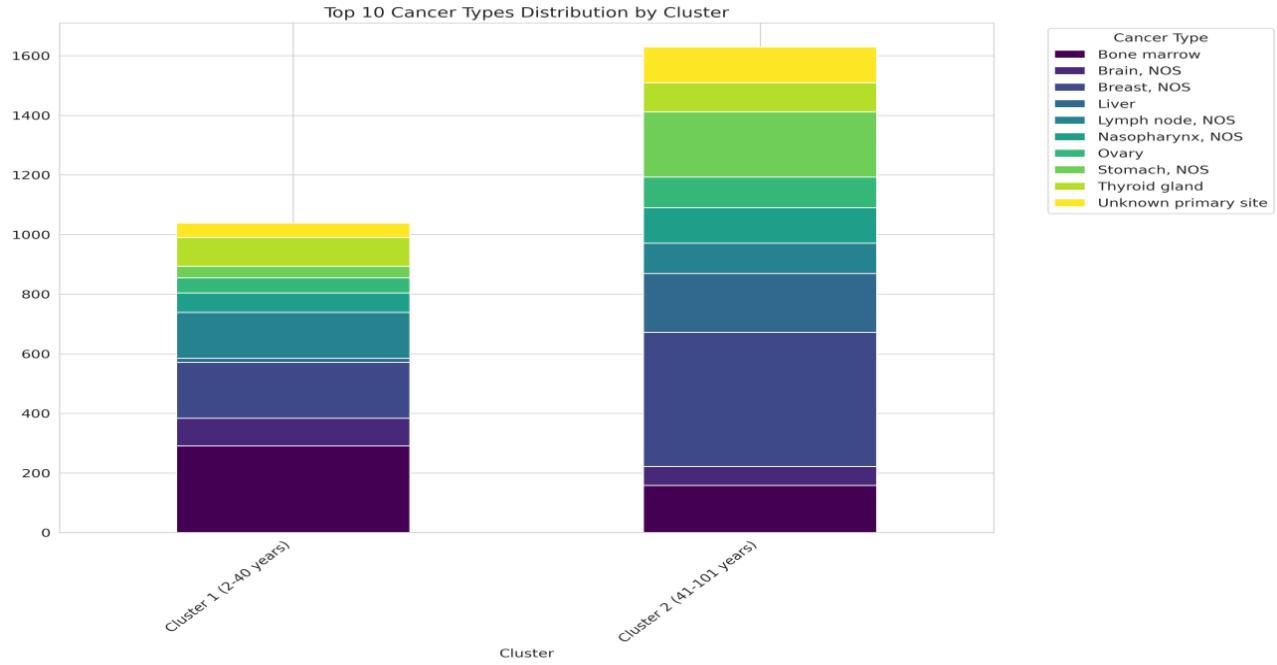

Figure 12: Top 10 cancer typed Distribution cy cluster

Visualisation and Interpretability

All analytical visualisations, including age distribution histograms, cluster assignment plots, and stacked bar charts of cancer-type frequencies, supported the quantitative results. The cluster-age distribution figure clearly showed the split around the 40-year mark. The cancer-type by cluster plot highlighted the shift from blood-related cancers to epithelial cancers as age increased. These graphical outputs aid in interpreting the machine-learning results, which are vital for applying this research in biomedicine.

Conclusion

This research establishes age as a primary driver of cancer distribution across both biological and statistical dimensions. While traditional linear models often struggle to map the complexities of epidemiological shifts, machine learning provides a flexible, data-driven lens to clarify these demographic patterns. Through the Gaussian Mixture

Model (GMM), we identified two distinct age-based clusters. The younger cohort (ages 2–40) was defined largely by hematologic and endocrine malignancies. By contrast, the older demographic (ages 41–101) presented a higher frequency of epithelial and gastrointestinal cancers. Rigorous validation against a random baseline and alternative clustering algorithms confirmed that these findings reflect genuine biological trends rather than computational artefacts. Statistical significance was confirmed via ANOVA and Chi-Squared tests ($p<0.001$), effectively ruling out stochastic variation. By capturing the non-linear relationship between patient maturity and disease type, the model addresses a significant gap in traditional trend analysis. These results carry immediate practical weight for cancer informatics. Integrating age-aware modelling into clinical systems allows for more nuanced screening protocols and better-informed resource allocation. The proposed framework is not static. Its scalability allows for the future inclusion of molecular or additional demographic variables—an adaptability essential for tracking cancer epidemiology in evolving populations. Ultimately, this study positions age as an essential metric in understanding oncological distribution, effectively synthesising epidemiological observation with computational rigour. This approach advances the utility of public health informatics and provides a clear path for future investigations into the non-linear determinants of cancer risk.

Future Work

Based on the results of this analysis, subsequent research will pursue several distinct avenues to refine the predictive utility of the framework:

- Multivariate Expansion: Future iterations will incorporate a broader array of covariates including sex, lifestyle factors, and genomic signatures to construct high-dimensional clustering models that capture more granular patient profiles.
- Temporal and Survival Modelling: Transitioning to longitudinal datasets will enable the assessment of age-sensitive trajectories in cancer progression, recurrence, and survival outcomes.
- Explainable AI (XAI) Integration: To move beyond "black box" associations, we will employ interpretable machine learning techniques, such as SHAP or LIME. This will clarify the specific feature importance driving the segregation of distinct clusters.
- Cross-Regional Validation: Evaluating the framework against international cancer registries is essential to determine its generalizability across varied populations and disparate healthcare infrastructures.
- Clinical Translation: Effective implementation requires direct collaboration with oncological centres to embed these age-based categorisations into real-world diagnostic decision-support systems.

This study establishes a necessary foundation for understanding the influence of age on cancer distribution. By synthesising epidemiological observation with the analytical rigour of machine learning, the research advances both computational oncology and public health informatics. The

result is a reproducible pathway for investigating the complex, non-linear variables that define global cancer risk.

Acknowledgements
The authors thank the National Oncology Centre (NOC) in Yemen for providing the data that enabled this research. We 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, we 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. Sung, H., et al. (2020). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians.
2. Smetana, K., Lacina, L., Szabo, P., Dvořánková, B., Brož, P., & Šedo, A. (2016). Ageing as an important risk factor for cancer. Anticancer research, 36(10), 5009-5017.
3. Terracina, S., Ferraguti, G., Petrella, C., Bruno, S. M., Blaconà, G., Di Certo, M. G., ... & Fiore, M. (2023). Characteristic hallmarks of aging and the impact on carcinogenesis. Current Cancer Drug Targets, 23(2), 87-102.
4. OWID. (September 16, 2024). Distribution of causes of death worldwide in 2021. In Statista. Retrieved March 26, 2025, from https://www.statista.com/statistics/1493203/share-of-deaths-worldwide-by-cause-2021/
5. Wu, Z., Xia, F., & Lin, R. (2024). Global burden of cancer and associated risk factors in 204 countries and territories, 1980-2021: a systematic analysis for the GBD 2021. Journal of hematology & oncology, 17(1), 119. https://doi.org/10.1186/s13045-024-01640-8
6. National Cancer Institute. (n.d.). Age and cancer risk. U.S. Department of Health and Human Services. https://www.cancer.gov/about-cancer/causes-prevention/risk/age
7. Woodman, R. J., & Mangoni, A. A. (2023). A comprehensive review of machine learning algorithms and their application in geriatric medicine: present and future. Aging Clinical and

Experimental Research, 35(11), 2363-2397.

8. Ballard, J. L., Wang, Z., Li, W., Shen, L., & Long, Q. (2024). Deep learning-based approaches for multi-omics data integration and analysis. BioData Mining, 17(1), 38.
9. Gab Allah, A. M., Gaballa, M., Elshennawy, N. M., & Elkholy, A. (2025). Edge-supervised convolutional neural network for histopathological classification of oral cancer images. Neural Computing and Applications, 1-20.
10. Buyrukoğlu, G. (2024). Survival analysis in breast cancer: evaluating ensemble learning techniques for prediction. PeerJ Computer Science, 10, e2147.
11. Zolfaghari, B., Mirsadeghi, L., Bibak, K., & Kavousi, K. (2023). Cancer prognosis and diagnosis methods based on ensemble learning. ACM Computing Surveys, 55(12), 1-34.
12. de Carvalho, T. C., da Mota Borges, A. K., & da Silva, I. F. (2023). Stomach cancer incidence trends in selected Latin America countries: Age, period, and birth-cohort effects. Cancer Epidemiology, 85, 102392
13. Johnston, W. T., Erdmann, F., Newton, R., Steliarova-Foucher, E., Schüz, J., & Roman, E. (2021). Childhood cancer: Estimating regional and global incidence. Cancer epidemiology, 71, 101662.
14. Godoy-Casasbuenas, N., Rincón, C. J., Gil, F., Arias, N., Pérez, C. U., Yépez, M. C., & de Vries, E. (2024). Age-period-cohort effects on incidence trends of childhood leukemia from four population-based cancer registries in Colombia. Cancer Epidemiology, 89, 102548.
15. Greppin, K., Cioffi, G., Waite, K. A., Ostrom, Q. T., Landi, D., Takaoka, K., ... & Barnholtz-Sloan, J. S. (2022). Epidemiology of pineoblastoma in the United States, 2000–2017. Neuro-Oncology Practice, 9(2), 149-157.
16. Gheybi, K., Buckley, E., Vitry, A., & Roder, D. (2022). Occurrence of comorbidity with colorectal cancer and variations by age and stage at diagnosis. Cancer Epidemiology, 80, 102246.