Short Communications

Editorial Process: Submission:26/03/2026 Acceptance:16/10/2026 Published:03/11/2026

AI-based Clinical Phenotyping for Pan-Cancer Stratification Using Unsupervised Learning

Khairunnisa Supardi[1], Soehartati A. Gondhowiardjo[1], Rezzy Eko Caraka[2,3]*, Prana Ugiana Gio[4], Nyityasmono Tri Nugroho[5], Noor Ell Goldameir[6]

Abstract

Objective: The clinical organization of cancer has traditionally been constrained by tumor-centric classifications that inadequately capture the shared phenotypic structure across malignancies. A comprehensive understanding of cancer heterogeneity requires a framework capable of reconstructing the continuous clinical landscape defined by disease burden, demographic factors, and stage at presentation without directly incorporating survival outcomes as model inputs. **Methods:** This study is a retrospective computational analysis utilizing the MLOmics harmonized benchmark dataset derived from The Cancer Genome Atlas (TCGA). A nonlinear artificial intelligence framework was implemented, integrating Uniform Manifold Approximation and Projection (UMAP) for manifold learning and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) for unsupervised phenotype discovery. Clinical variables, including age, gender, and pathological stage, were encoded and standardized prior to embedding. **Results:** The analysis revealed that pan-cancer clinical data are organized within a continuous nonlinear manifold rather than into discrete tumor-specific compartments. Density-based clustering identified multiple stable clinical phenotypes, alongside a subset of noise points representing atypical or heterogeneous cases. The observed clustering structure demonstrated moderate silhouette separation, consistent with the overlapping clinical realities of oncology. Importantly, phenotype formation was not driven by tumor origin but by shared clinical attributes, particularly disease stage and patient demographics. **Conclusion:** In radiation oncology, where treatment decisions are often driven by anatomical extent and disease burden, such phenotype-centric models enable cross-disease insights that support precision treatment planning and resource optimization.

Keywords: Manifold learning, Density-based clustering, Clinical heterogeneity, Phenotype-driven stratification

Asian Pac J Cancer Prev, **27 (10)**, 3867-3877

Introduction

Cancer remains a leading cause of global morbidity and mortality, characterized not only by biological heterogeneity but also by substantial variability in clinical presentation across patients. Traditional oncology frameworks have relied predominantly on tumor origin and histopathological classification to guide diagnosis and treatment. While this paradigm has led to major advances in disease-specific management, it does not fully capture the complexity of how patients present in real-world clinical settings, where age, stage, and systemic condition frequently determine therapeutic strategy more directly than tumor type alone [1, 2].

The increasing availability of large-scale pan-cancer datasets has created an opportunity to move beyond organ-based classification toward more integrative approaches. In particular, datasets derived from The Cancer Genome Atlas (TCGA) and related repositories provide standardized clinical and molecular information across multiple cancer types, enabling cross-disease analysis at an unprecedented scale [3–5] . However, most existing studies focus on survival prediction or molecular subtyping, often neglecting the underlying clinical structure that governs patient management.

From a clinical perspective, especially in radiation oncology, treatment decisions are inherently driven by disease extent, anatomical involvement, and patient condition. These factors frequently transcend tumor origin, suggesting that patients may share common phenotypic

*[1]Department of Radiation Oncology, Faculty of Medicine Universitas Indonesia - Dr Cipto Mangunkusumo National General Hospital, Jakarta 10430, Indonesia. [2]School of Economics and Business Telkom University, Bandung 40257, Indonesia. [3]Research Center for Data and Information Sciences, Research Organization for Electronics and Informatics, National Research and Innovation Agency (BRIN), Bandung 40135, Indonesia. [4]Department of Mathematics, Universitas Sumatera Utara, Medan 20222, Indonesia. [5]Division of Vascular and Endovascular, Department of Surgery, Faculty of Medicine, University of Indonesia, Dr. Cipto Mangunkusumo Centre General Hospital, Jakarta, Indonesia. [6]Department of Mathematics, Statistics Study Program, Riau University, Pekanbaru, Riau 28293, Indonesia. *For Correspondence: rezzy.eko.caraka@brin.go.id*

patterns across different cancers. Identifying such patterns requires analytical approaches capable of capturing nonlinear relationships and continuous variation within heterogeneous datasets [6, 7].

Artificial intelligence, particularly unsupervised learning, provides a framework for uncovering these latent structures without imposing predefined categories. Unlike supervised models, which depend on outcome labels, unsupervised methods can reveal intrinsic organization within clinical data, offering insights into how patients naturally group based on shared characteristics. This is particularly relevant in oncology, where disease progression and treatment response exist along continuous spectra rather than discrete classes [8, 9]. In this study, we applied unsupervised artificial intelligence framework to derive clinically meaningful phenotypes from a large pan-cancer cohort. By integrating nonlinear representation learning with density-based clustering, we aim to map the clinical landscape of cancer patients and identify reproducible patterns of disease presentation. Importantly, this analysis is conducted independently of survival outcomes, focusing instead on the structural organization of clinical features. The objective is not to replace existing tumor-specific classifications, but to complement them by introducing a phenotype-driven perspective that reflects real-world clinical decision-making. Such an approach has potential implications for treatment stratification, particularly in radiation oncology, where planning and delivery are closely aligned with disease burden rather than histological subtype.

Materials and Methods

Conceptual Framework : From Tumor-Centric to Phenotype-Centric Oncology

The analytical framework is designed to reconstruct cancer heterogeneity as a continuous clinical manifold rather than a set of discrete tumor-specific entities. Let the observed clinical space be defined as:

$$\boldsymbol{X} = \{x_1, x_2, \dots, x_n\}, \quad x_i \in \mathbb{R}^{\mathbb{P}}$$

where each x_i represents a patient described by clinical variables including age, gender, and disease stage. Instead of assuming independence across cancer types, the framework models a latent structure

$$\mathcal{M} \subset \mathbb{R}^{\mathbb{P}}$$

which represents the underlying clinical manifold governing patient similarity across malignancies. From an oncological standpoint, this formulation reflects the clinical reality that disease burden, anatomical spread, and host factors often dominate therapeutic decisions more strongly than tumor origin alone. To ensure that each clinical variable contributes proportionally to similarity computations, the data are transformed into a standardized space:

$$\widetilde{x_i} = \left(\frac{x_{i1} - \mu_1}{\sigma_1}, \dots, \frac{x_{ip} - \mu_p}{\sigma_p}\right)$$

This induces a Euclidean metric:

$$d(x_i, x_j) = |\widetilde{x_i} - \widetilde{x_j}|_2$$

which defines the geometric structure of the clinical space. In oncology, this metric implicitly encodes clinical distance. For example, two patients with similar stage and age will lie closer in this space, reflecting comparable disease burden and, consequently, potentially similar treatment pathways.

UMAP: Topological Reconstruction of Clinical Manifold

UMAP begins by constructing a weighted graph based on local neighbourhoods. For each point x_i, a local connectivity radius ρ_i is defined as:

$$\rho_i = \min\{\, d(x_i, x_j) \mid j \in \mathcal{N}_k(i), j \neq i\}$$

ensuring that each point is at least connected to its nearest neighbor. A smooth radius σ_i is then determined by solving:

$$\sum_j \exp\left(-\frac{d(x_i, x_j) - \rho_i}{\sigma_i}\right) = \log_2(k)$$

This defines the local connectivity scale, balancing sensitivity to local versus global structure. The resulting membership strength between points i and j is:

$$w_{ij} = \exp\left(-\frac{d(x_i, x_j) - \rho_i}{\sigma_i}\right)$$

These weights define a fuzzy simplicial set, representing the probability that two patients share a local clinical neighborhood.

Low-Dimensional Embedding Optimization

UMAP seeks a low-dimensional representation $Y \epsilon R^{n \times d}$ by minimizing the cross-entropy between high-dimensional and low-dimensional graphs:

$$\mathcal{L} = \sum_{i \neq j} \left[w_{ij} \log \frac{w_{ij}}{w'_{ij}} + (1 - w_{ij}) \log \frac{1 - w_{ij}}{1 - w'_{ij}} \right]$$

where the low-dimensional similarity is modeled as:

$$w'_{ij} = \frac{1}{1 + a|y_i - y_j|^{2b}}$$

with parameters a and b controlling the embedding shape. This optimization preserves neighborhoods of clinically similar patients. In practice, this means that patients with comparable disease stage and demographic characteristics will cluster together in the embedded space, regardless of tumor origin. This reflects a phenotype-centric organization of cancer, aligning with how radiotherapy fields and treatment intent are determined.

HDBSCAN: Density-Based Phenotype Extraction

For each point x_i, the core distance is defined as:

$$\mathrm{core}_k(x_i) = d(x_i, x_{(k)})$$

where x_((k)) is the k-th nearest neighbor. The mutual reachability distance between two points is then:

$$d_{\text{mreach}}(x_i, x_j) = \max\{\text{core}_k(x_i), \text{core}_k(x_j), d(x_i, x_j)\}$$

This transformation stabilizes distances in regions of varying density, preventing fragmentation of clinically meaningful groups. A minimum spanning tree (MST) is constructed over the mutual reachability graph[10, 11]. Clusters are identified by progressively removing edges with increasing distance, forming a hierarchy of connected components. Cluster stability is defined as the persistence of a cluster across density levels:

$$\int_{\lambda_{birth}}^{\lambda_{death}} |C(\lambda)| d$$

Clusters with higher stability are retained, while unstable regions are discarded as noise. This process identifies groups of patients with consistent clinical density. Importantly, it allows for the existence of noise points, which correspond to atypical or rare clinical presentations. In oncology, these may represent unusual disease trajectories, rare cancers, or mixed phenotypes that do not conform to standard treatment pathways.

Continuity of Clinical Phenotypes

Unlike partition-based clustering, HDBSCAN produces clusters embedded within a continuous manifold. Let the embedded space be:

$$\boldsymbol{Y} = \{y_1, \ldots, y_n\}, \quad y_i \in \mathbb{R}^2$$

Clusters $C_1, C_2, \ldots, C_k \subset Y$ are not disjoint in a strict geometric sense but exist as dense regions within a continuum. This reflects the biological and clinical reality that cancer phenotypes are not discrete categories. Patients transition gradually from early-stage, low-burden disease to advanced, high-burden states. The model therefore captures disease progression as a trajectory rather than a classification. Clustering performance is evaluated using the silhouette coefficient [12–14].

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where:
a(i): intra-cluster cohesion
b(i): nearest-cluster separation

Moderate silhouette values are expected and clinically appropriate. In oncology, overlapping phenotypes are common due to shared biological pathways and heterogeneous progression patterns. Therefore, the absence of sharp cluster boundaries reflects realism rather than methodological limitation. To quantify the diversity of cancer types, Shannon entropy is computed:

$$H = -\sum_{i=1}^{K} p_i \log p_i$$

A high entropy value indicates that the dataset captures a wide spectrum of malignancies. This is essential for validating that the derived phenotypes are not driven by a single dominant cancer type but instead represent true pan-cancer structure.

Artificial Intelligence Perspective

The integration of UMAP and HDBSCAN constitutes a unified unsupervised learning pipeline. UMAP performs nonlinear representation learning by approximating the underlying clinical manifold, while HDBSCAN extracts stable density-based structures from this representation. Formally, this can be viewed as a mapping:

$$f: \mathbb{R}^{\mathbb{p}} \to \mathbb{R}^{\mathbb{d}}, \quad g: \mathbb{R}^{\mathbb{d}} \to \{C_1, \ldots, C_k\}$$

where:
f is a nonlinear embedding (UMAP)
g is a clustering operator (HDBSCAN)

This pipeline enables phenotype discovery without reliance on survival outcomes or predefined labels. In clinical oncology, this shifts the paradigm from outcome-driven modeling to structure-driven understanding, allowing clinicians to identify patient groups based on presentation rather than prognosis alone. Such an approach is particularly relevant in radiation oncology, where treatment decisions are determined by spatial and clinical characteristics at diagnosis. By revealing shared phenotypes across cancers, this framework supports the development of cross-disease treatment strategies and more adaptive, patient-centered care models.

Dataset

The dataset used in this study was obtained from the publicly available Cancer Multi-Omics Benchmark repository on GitHub (https://github.com/chenzRG/Cancer-Multi-Omics-Benchmark). This resource provides a standardized, machine learning–ready compilation of pan-cancer data primarily derived from The Cancer Genome Atlas (TCGA). The dataset comprises 8,314 patient samples across 32 cancer types, representing a broad spectrum of malignancies, including breast, lung, prostate, colorectal, and head and neck cancers. It integrates multi-omics layers mRNA, miRNA, DNA methylation, and copy number variation alongside associated clinical information. For the purpose of this study, the analysis focuses on a clinical subset consisting of age at diagnosis, gender, pathological stage, survival status, and follow-up time. After data cleaning and removal of incomplete records, a total of 8,183 patients were included in the final cohort. This dataset is particularly suitable for pan-cancer analysis due to its standardized preprocessing and cross-cancer consistency, enabling the identification of shared clinical phenotypes beyond tumor-specific boundaries.

Results

Latent Clinical Topology of Pan-Cancer Cohort

The central clinical issue in this section is not merely heterogeneity, but rather how this heterogeneity translates into real patient burden in daily oncology practice.

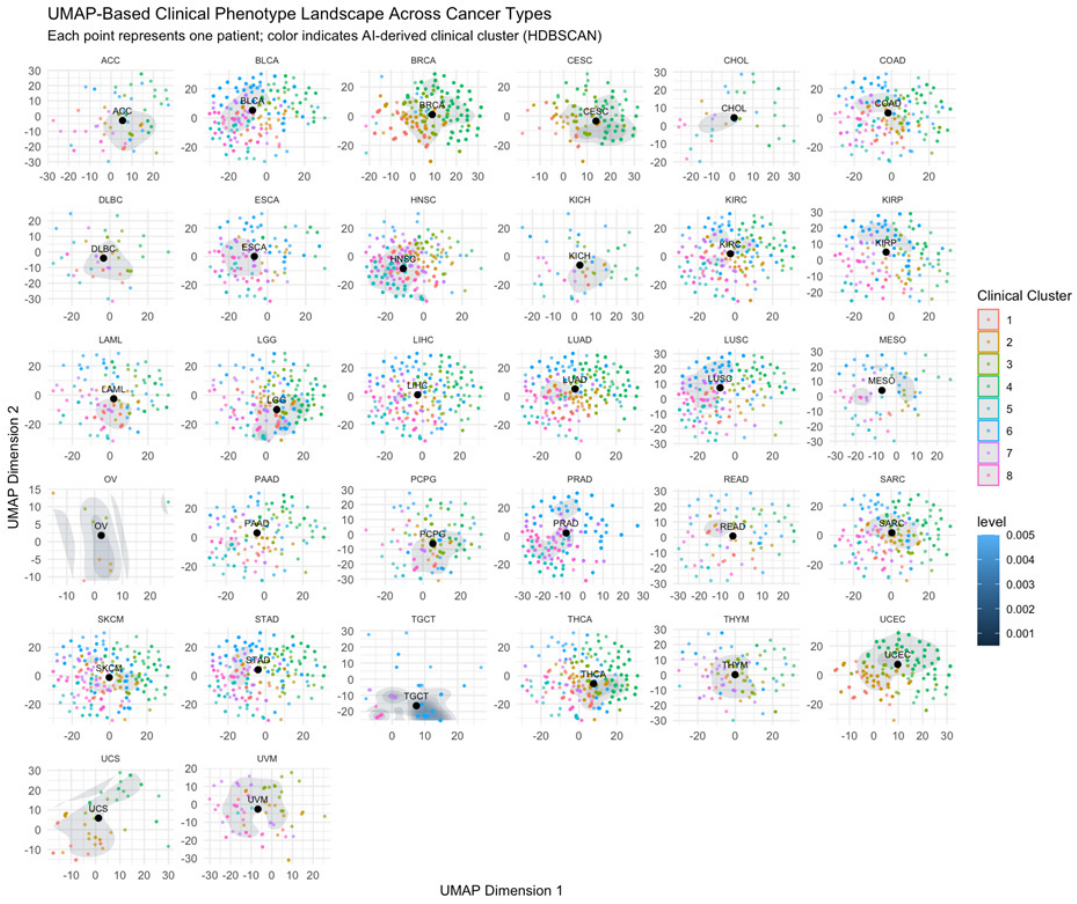


Figure 1. UMAP-Based Clinical Phenotype Landscape Across Cancer Types

When examining 8,183 patients across more than 30 malignancies, what emerges is not a fragmented dataset but a structured clinical ecosystem shaped primarily by stage at diagnosis and demographic context.

From a radiation oncology perspective, stage is not just a variable; it defines intent. A patient in Stage I is typically approached with curative intent using surgery or radiotherapy, whereas Stage IV often implies systemic disease requiring palliative strategies. When we observe that 3,498 patients present at Stage I compared to 1,472 at Stage IV, this immediately signals that a substantial proportion of the cohort remains within a potentially curable window. However, this global number conceals profound inter-cancer disparities.The distribution becomes more clinically striking when examined at the tumor level (Table 1). Breast cancer (BRCA), for instance, shows 338 patients diagnosed at Stage I versus 126 at Stage IV. This ratio reflects the success of early detection programs and screening awareness. In contrast, head and neck squamous cell carcinoma (HNSC) presents a fundamentally different story: only 24 patients at Stage I, but 281 already at Stage IV. This nearly tenfold difference illustrates a systemic delay in diagnosis, where patients often present only when symptoms become severe. In practical radiotherapy settings, this translates into a higher reliance on combined modality treatment, including chemoradiation, rather than definitive single-modality therapy.

A similar contrast is seen when comparing indolent versus aggressive tumor profiles. Thyroid cancer (THCA) maintains a relatively balanced distribution across all stages (210 Stage I vs 94 Stage IV), suggesting a slower disease trajectory and effective incidental detection. On the other hand, pancreatic cancer (PAAD), with 69 Stage I and 39 Stage IV cases, shows a narrower gap, indicating that a significant proportion of patients are already approaching advanced disease at diagnosis. Clinically, this aligns with the well-known difficulty of early detection in pancreatic malignancies and explains why curative radiotherapy is less frequently applicable.

The UMAP representation (Figure 1) translates these numerical imbalances into spatial relationships. Patients do not cluster strictly by cancer type; instead, they organize along gradients of clinical severity. Regions with dense accumulation correspond to common phenotypic profiles typically mid-to-late stage patients within the age range of 50–70 years, which represents the bulk of oncology workload globally. Sparse regions, in contrast, represent either rare cancers or atypical clinical presentations.

The UMAP projection provides a low-dimensional representation of patient-level clinical similarity derived from age, sex, and pathological stage. Each point corresponds to an individual patient, while spatial proximity reflects shared clinical characteristics rather than tumor histology alone. A key observation is that the distribution of patients does not form sharply segregated clusters by cancer type. Instead, multiple malignancies occupy overlapping regions within the UMAP space, indicating that clinical phenotypes are partially conserved

across traditionally distinct tumour entities. This pattern is particularly evident in high-prevalence cancers such as breast (BRCA), thyroid (THCA), and prostate (PRAD), which display broad dispersion rather than tight, isolated clusters.

From a clinical standpoint, this suggests that age and stage-driven disease burden may transcend histological classification, reinforcing the notion that patient condition at presentation rather than tumour origin alone plays a dominant role in defining clinical phenotype. For example, early-stage disease across multiple cancers converges within similar regions of the embedding, reflecting comparable systemic impact and clinical management pathways. The integration of HDBSCAN clustering further refines this interpretation by identifying nine distinct clinical phenotypes independent of cancer labels. These clusters are not cancer-specific; instead, they are distributed across multiple tumour types, indicating shared clinical archetypes across cancers. Certain clusters are enriched for advanced-stage presentations, while others are dominated by early-stage disease, effectively recapitulating disease severity gradients without explicit supervision.

Importantly, the presence of a noise cluster (cluster 0), subsequently excluded from the main visualization, highlights a subset of patients with atypical or heterogeneous clinical profiles. In an oncological context, this group may represent rare presentations, mixed-stage disease, or patients whose clinical trajectories do not conform to dominant patterns. Such cases are often clinically challenging and may require individualized management strategies. The density contours overlaid on each facet reveal the internal structure of each cancer type. Rather than uniform distributions, most cancers exhibit multi-modal density patterns, suggesting intra-tumour heterogeneity at the clinical level. For instance, cancers such as head and neck squamous cell carcinoma (HNSC) and lung cancers (LUAD, LUSC) demonstrate dispersed and fragmented distributions, consistent with their known biological and clinical variability. In contrast, malignancies such as thyroid cancer (THCA) show more compact distributions, reflecting relatively homogeneous clinical presentation.

The centroid markers provide an approximate "clinical center of gravity" for each cancer type. Notably, these centroids often lie within shared regions rather than isolated zones, further supporting the concept that clinical similarity across cancers is substantial and non-trivial. This challenges the conventional siloed approach to oncology and suggests that cross-cancer stratification strategies may be feasible when grounded in patient-level clinical features. Collectively, this figure demonstrates that pan-cancer clinical phenotyping using unsupervised AI reveals a continuum of disease states rather than discrete categories. This continuum is shaped more strongly by stage and patient characteristics than by tumour origin alone.

Such insights are highly relevant for clinical trial design, risk stratification, and resource allocation, particularly in settings where treatment decisions are increasingly guided by patient condition rather than histology alone. Notably, different cancer types occupy overlapping regions within the embedding space. While this does not imply clinical equivalence, it suggests that patients with distinct primary tumors may share similar patterns of disease burden and presentation

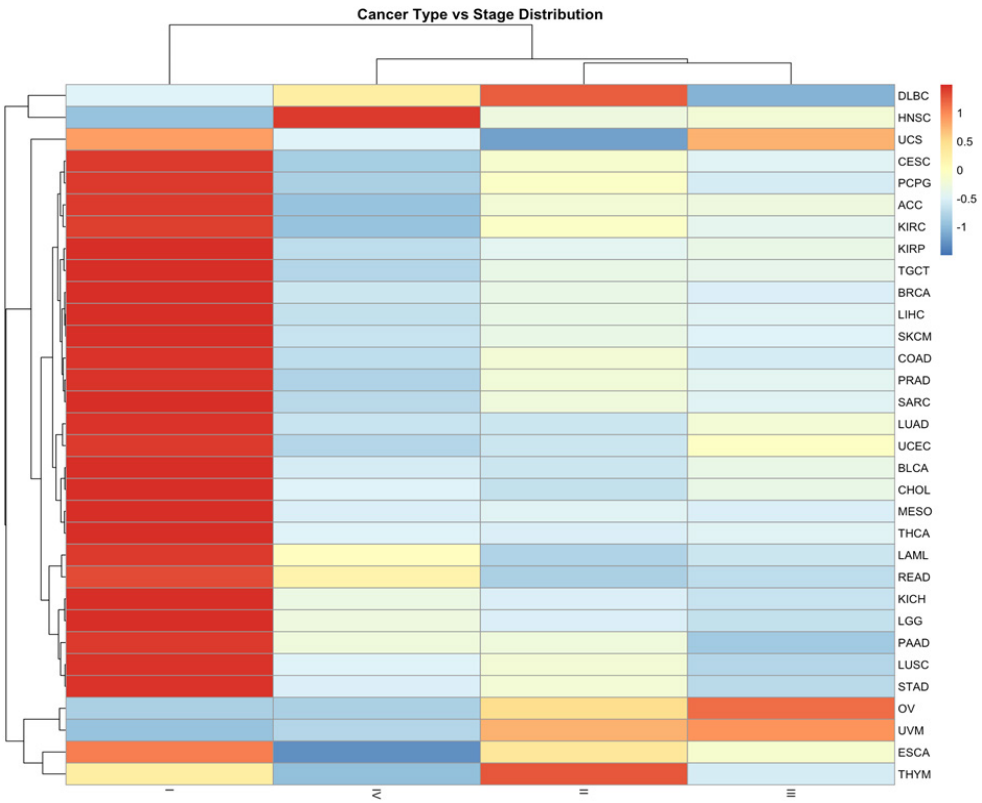


Figure 2. Heatmap Cancer Type vs Stage Distribution

Table 1. Pan Cancer Composition

Cancer Type	Full Description	Number of Patients (N)	Percentage (%)
BRCA	Breast invasive carcinoma	755	9,23
HNSC	Head and Neck squamous cell carcinoma	505	6,17
LGG	Brain Lower Grade Glioma	502	6,13
THCA	Thyroid carcinoma	493	6,02
PRAD	Prostate adenocarcinoma	484	5,91
LUAD	Lung adenocarcinoma	427	5,22
UCEC	Uterine Corpus Endometrial Carcinoma	407	4,97
BLCA	Bladder Urothelial Carcinoma	399	4,88
STAD	Stomach adenocarcinoma	355	4,34
LIHC	Liver hepatocellular carcinoma	351	4,29
LUSC	Lung squamous cell carcinoma	345	4,22
SKCM	Skin Cutaneous Melanoma	336	4,11
KIRC	Kidney renal clear cell carcinoma	304	3,72
COAD	Colon adenocarcinoma	281	3,43
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	278	3,4
KIRP	Kidney renal papillary cell carcinoma	263	3,21
SARC	Sarcoma	250	3,06
ESCA	Esophageal carcinoma	180	2,2
PAAD	Pancreatic adenocarcinoma	175	2,14
PCPG	Pheochromocytoma and Paraganglioma	161	1,97
LAML	Acute Myeloid Leukemia	141	1,72
TGCT	Testicular Germ Cell Tumors	133	1,63
THYM	Thymoma	118	1,44
READ	Rectum adenocarcinoma	89	1,09
MESO	Mesothelioma	85	1,04
UVM	Uveal Melanoma	80	0,98
ACC	Adrenocortical carcinoma	76	0,93
KICH	Kidney Chromophobe	65	0,79
UCS	Uterine Carcinosarcoma	54	0,66
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	46	0,56
CHOL	Cholangiocarcinoma	36	0,44
OV	Ovarian serous cystadenocarcinoma	9	0,11

as captured by the underlying clinical variables. For example, advanced-stage lung cancer (LUAD/LUSC) and head–neck cancer frequently appear in proximate regions, potentially reflecting shared etiological factors such as tobacco exposure and comparable patterns of locoregional progression. In radiotherapy contexts, these similarities may translate into analogous planning challenges, including large target volumes and proximity to critical organs.

The heatmap of stage distribution (Figure 2) further clarifies why this overlap occurs. Cancers with broad stage distributions, such as lung and colorectal cancers, contribute to wide dispersion across the clinical space. In contrast, cancers with tightly constrained stage profiles, such as testicular germ cell tumors (TGCT), which show 98 cases at Stage I and none at Stage IV, occupy more defined regions. This reinforces the notion that variability in stage at diagnosis is a primary driver of clinical heterogeneity.

Taken together, these findings redefine how heterogeneity should be interpreted in oncology. The variability observed is not simply a reflection of different cancer types, but a manifestation of when and how patients enter the healthcare system. Early-stage dominance in some cancers reflects successful screening and awareness, while late-stage dominance in others highlights persistent diagnostic gaps.

From a clinical decision-making standpoint, this section establishes a crucial foundation. It demonstrates that patient similarity is not dictated solely by tumor origin, but by a multidimensional interaction of stage, age, and disease progression. This insight is particularly relevant for radiation oncology, where treatment planning often depends more on anatomical extent and disease burden than on histological label alone.

This structured heterogeneity provides the rationale for

Table 2. Integrated Distribution of Pathologic Stage Across Pan-Cancer Cohort

Cancer	Stage I	Stage II	Stage III	Stage IV
ACC	34	17	16	9
BLCA	179	67	82	71
BRCA	338	155	136	126
CESC	157	62	41	18
CHOL	18	5	7	6
COAD	129	62	48	42
DLBC	9	18	6	13
ESCA	62	51	43	24
HNSC	24	96	104	281
KICH	31	11	10	13
KIRC	119	74	64	47
KIRP	158	38	45	22
LAML	59	22	25	35
LGG	220	92	82	108
LIHC	153	73	67	58
LUAD	187	73	96	71
LUSC	160	77	47	61
MESO	42	15	14	14
OV	1	3	4	1
PAAD	69	39	28	39
PCPG	67	39	30	25
PRAD	209	107	95	73
READ	34	15	16	24
SARC	109	54	48	39
SKCM	157	67	60	52
STAD	164	78	51	62
TGCT	98	18	17	0
THCA	210	93	96	94
THYM	36	60	16	6
UCEC	245	41	97	24
UCS	20	5	19	10
UVM	0	36	40	4

transitioning into AI-driven clustering in the next section, where the objective is to formalize these gradients into reproducible clinical phenotypes rather than relying on conventional categorical classifications.

AI-Derived Clinical Phenotypes as Density-Driven Disease States

The transition from structural visualization to phenotypic definition requires a shift in perspective. The objective is no longer to identify whether patients differ, but to determine how these differences stabilize into reproducible clinical states that can be interpreted in practice. The integration of UMAP with HDBSCAN provides this bridge by transforming continuous clinical variation into density-based groupings that reflect how patients aggregate in real-world settings.

Unlike partition-based clustering, HDBSCAN identifies regions where patient density is sufficiently high to form stable phenotypes, while allowing low-density observations to remain unassigned. This distinction is clinically essential. In oncology, not every patient belongs to a well-defined category. Rare tumors, atypical staging patterns, and incomplete clinical trajectories naturally exist at the margins. The identification of a noise component is therefore not an analytical limitation, but a representation of clinical reality.

After excluding noise, the analysis yields nine stable clusters embedded within the UMAP space. These clusters are not isolated compartments but density peaks within a continuous landscape (Figure 1). Each peak represents a region where patients share similar combinations of age, stage, and demographic structure. The moderate silhouette value reflects this continuity, confirming that phenotypes overlap rather than separate sharply, which is consistent with how disease manifests clinically.

What distinguishes this model from conventional classification is that clustering is not driven by tumor origin. Instead, it is governed by shared clinical burden. This is evident in Table 2, where each cluster contains multiple cancer types rather than a single dominant entity. The implication is that patients are grouped by how disease presents, not where it originates.

The largest cluster, comprising 1,771 patients, represents a high-density clinical state dominated by breast, thyroid, and endometrial cancers. The numerical weight of this cluster is not incidental. It reflects a convergence of diseases that are frequently detected across multiple stages, often through structured diagnostic pathways. The density of this cluster in UMAP space indicates stability, suggesting that these patients share consistent clinical characteristics. In radiation oncology, this phenotype aligns with cases where treatment intent

Table 3. Integrated Distribution of Cancer Types Across AI-Derived Clinical Phenotypes

Cluster	Dominant Composition (Selected High-Frequency Signals)	Clinical Density Interpretation
1	BRCA, HNSC, LGG with moderate spread	Early-to-intermediate disease mixture
2	BRCA, CESC, THCA increasing	Expanding screening-related population
3	BRCA with gastrointestinal cancers (COAD, STAD)	Transitional burden phenotype
4	BRCA (333), THCA (156), UCEC (245)	High-density stable phenotype (largest cluster)
5	HNSC (218), ESCA, advanced-stage signals	Locally advanced disease concentration
6	PRAD (209), KIRP, KIRC, LGG	Indolent, age-driven phenotype
7	Mixed HNSC, GI cancers	Intermediate burden phenotype
8	Balanced multi-cancer distribution	Overlapping clinical spectrum
0	Sparse across all cancers	Clinically heterogeneous outliers

Table 4. Clinical Interpretation of Phenotypic Continuum and Radiotherapy Implications

Phenotypic Region	Dominant Characteristics	Representative Signals	Clinical Interpretation	Radiotherapy Implication
Stable high-density phenotype	Broad stage distribution, structured detection	BRCA, THCA, UCEC	Predictable disease course	Curative or adjuvant RT with standard planning
Transitional phenotypes	Mixed stage and demographic features	Mixed cancers across clusters 1–3, 7–8	Heterogeneous presentation	Individualized RT planning
Advanced disease phenotype	High stage concentration, anatomical complexity	HNSC, ESCA	High disease burden	Chemoradiation and complex RT techniques
Indolent trajectory phenotype	Older age, slower progression	PRAD, KIRC, LGG	Chronic disease course	Selective or palliative RT
Outlier phenotype	Sparse and heterogeneous	Rare cancers	Non-standard patterns	Case-by-case decision

is predominantly curative, and where planning strategies are relatively standardized.

In contrast, the cluster enriched with head and neck cancer demonstrates a markedly different density profile. The concentration of 218 cases within a single cluster reflects a strong convergence of advanced-stage disease. This is not simply a distributional artifact. Head and neck malignancies often present with extensive locoregional involvement, and when combined with other cancers exhibiting similar stage burdens, they form a dense region representing high clinical complexity. From a therapeutic standpoint, this phenotype corresponds to patients requiring aggressive multimodal treatment, where radiotherapy is delivered with concurrent systemic therapy and significant anatomical constraints.

A third distinct density peak emerges from prostate-dominant clustering, where more than 200 cases co-localize with kidney cancers and gliomas. At first glance, this grouping appears biologically heterogeneous. However, the clustering is driven by shared clinical features rather than histology. These patients are typically older, with disease trajectories that may be prolonged and less acutely aggressive. The density structure suggests a phenotype characterized by chronicity rather than rapid progression. In practice, radiotherapy in this group is often selective, balancing disease control with long-term quality of life.

The remaining clusters do not exhibit singular dominance but instead form intermediate density regions. These represent transitional states where patients share overlapping characteristics without a clear defining feature. Their presence is critical for interpretation. Oncology practice is not limited to archetypal cases; most patients fall within these intermediate zones, where treatment decisions require integration of multiple factors rather than adherence to a single pathway.

The spatial organization of these clusters within the UMAP framework reinforces a key clinical principle. Disease phenotypes exist along gradients of severity and progression. Dense clusters correspond to stable and reproducible clinical states, while diffuse regions indicate variability and transition. This structure cannot be captured by linear models or rigid classifications, which explains why nonlinear embedding is necessary in this context.

From a translational perspective, the findings redefine how patient grouping can be approached in oncology. Instead of relying exclusively on tumor type, patients can be stratified based on shared clinical burden. This is particularly relevant in advanced disease, where treatment strategies often converge across different cancers. Radiotherapy planning, for example, is influenced more by tumor extent and anatomical considerations than by histological subtype alone.The model therefore introduces a practical framework in which phenotypes are interpreted as density-defined disease states. Each state reflects a combination of presentation, progression, and treatment complexity. This approach does not replace traditional classification but augments it by revealing patterns that are otherwise obscured when cancers are analyzed in isolation.

In essence, the clustering results demonstrate that pan-cancer heterogeneity can be organized into a limited number of clinically meaningful density regions. These regions correspond to how patients actually present in clinical settings, providing a foundation for phenotype-driven decision support that extends beyond conventional disease boundaries.

Clinical Interpretation of the Phenotypic Continuum and Implications for Oncology Practice

The phenotypic structure identified in this study does not function as a static classification system, but rather as a dynamic continuum that reflects how disease manifests and evolves in real patients. The transition from density-defined clusters to clinical meaning requires interpreting these phenotypes not as isolated groups, but as positions along a trajectory of disease burden, progression, and therapeutic complexity.

What emerges from the UMAP landscape is a gradient that moves from stable, screening-influenced phenotypes toward increasingly complex and advanced disease states. This gradient is not imposed by the model; it is intrinsically derived from how patients cluster based on age, stage, and demographic structure. The implication is that clinical reality is inherently continuous, and any attempt to discretize it must preserve this continuity to remain meaningful.

At one end of this continuum lies the high-density phenotype dominated by breast, thyroid, and endometrial cancers. The numerical concentration observed in Table 2 is clinically consistent with diseases that benefit from structured detection pathways. The distribution across stages, combined with the relative stability of

this cluster in the UMAP space, reflects a population entering care earlier and more predictably. In radiation oncology, this translates into treatment scenarios where intent is predominantly curative, planning volumes are well-defined, and outcomes are more controllable. The predictability of this phenotype reduces uncertainty in both target delineation and dose optimization.

Discussion

Progressing along the continuum, the phenotypes become less stable and more heterogeneous. Transitional clusters, which collectively represent a substantial proportion of patients, reflect the complexity of real-world oncology. These patients do not conform to a single dominant pattern of disease. Instead, they exhibit overlapping characteristics that require individualized interpretation. The absence of a clear density peak in these regions is clinically meaningful. It indicates variability in disease presentation, differences in access to care, and potential heterogeneity in tumor biology. For the radiation oncologist, this is the domain where standardized protocols become less sufficient, and treatment planning must adapt to nuanced clinical scenarios.

At the opposite end of the continuum, the phenotype enriched with advanced-stage head and neck cancer represents a distinct clinical reality. The concentration of late-stage cases creates a dense region associated with high disease burden and anatomical complexity. This is not merely a statistical observation; it reflects the challenges encountered in daily practice. Large tumor volumes, proximity to critical organs, and the frequent need for concurrent systemic therapy define this group. Radiotherapy in this context becomes technically demanding, requiring advanced techniques such as intensity-modulated radiotherapy and adaptive planning. The clustering of these patients into a coherent density region confirms that advanced disease states are not random but follow reproducible patterns.

A separate dimension of the continuum is represented by the phenotype dominated by prostate cancer and related malignancies. The numerical dominance of prostate cases within this cluster, combined with contributions from kidney cancers and gliomas, reflects a population characterized by longer disease trajectories. These patients often present with slower progression and extended survival, shifting the clinical focus from immediate disease eradication to long-term management. Radiotherapy in this setting is frequently used selectively, either for local control or palliation, and must be integrated into a broader treatment strategy that considers quality of life over time.

The presence of a noise component further refines this interpretation. Patients classified outside stable clusters do not conform to dominant phenotypic patterns. Clinically, these cases often correspond to rare malignancies, atypical staging combinations, or incomplete clinical information. Their exclusion from structured clusters prevents distortion of the overall phenotypic landscape and preserves the integrity of the dominant patterns. From a practical standpoint, these patients require individualized evaluation and cannot be reliably managed through generalized frameworks. To consolidate these observations into a clinically actionable format, Table 3 integrates phenotypic structure with therapeutic implications (Table 4).

The integration of these phenotypes into clinical practice suggests a shift from tumor-centric to patient-centric stratification. Traditional oncology relies heavily on anatomical classification, yet the present findings demonstrate that patients reorganize into patterns defined by disease burden and progression rather than origin alone. This does not diminish the importance of tumor biology, but it highlights an additional layer of organization that is directly relevant to treatment delivery.In radiation oncology, this perspective is particularly valuable. Treatment planning inherently depends on spatial extent, organ involvement, and patient condition. These factors align more closely with phenotypic clustering than with tumor type alone. As a result, phenotype-based stratification can enhance anticipation of treatment complexity, resource allocation, and potential toxicity.

Beyond individual patient care, the phenotypic continuum also carries implications at the system level. High-density early-stage phenotypes reflect effective screening and diagnostic infrastructure, while clusters dominated by advanced disease indicate gaps in early detection. The distribution of patients across this continuum therefore provides indirect insight into healthcare performance and accessibility. In summary, the clinical value of the identified phenotypes lies not in their ability to categorize patients rigidly, but in their capacity to represent the continuous nature of disease. By aligning data-driven clustering with real-world clinical reasoning, this framework transforms abstract computational outputs into meaningful guidance for oncology practice.

Our findings highlight a persistent limitation in oncology the reliance on tumor-centric classification systems that often fail to capture patient-level heterogeneity. Using a large pan-cancer cohort and routinely available clinical variables, we demonstrate that meaningful and reproducible phenotypes can be derived without dependence on survival endpoints or high-dimensional molecular data.

The findings support the view that patient organization may follow a structured continuum rather than strictly discrete, tumor-bound categories. This perspective is consistent with prior developments in oncology, including pan-cancer initiatives such as The Cancer Genome Atlas, which revealed molecular commonalities across tumor types. Rather than introducing a fundamentally new concept, the present study extends this paradigm into the clinical domain. While TCGA emphasized genomic convergence, our results suggest that similar patterns may also be observable in routinely collected clinical variables, including age, stage, and demographic context. This distinction is clinically relevant, as such variables remain central to treatment decision-making in many healthcare settings.

A key methodological insight emerges when comparing approaches. The earlier network-based strategy using proximity matrices from Random Survival Forests resulted in excessive fragmentation, producing an

implausible number of clusters. From a clinical standpoint, such fragmentation lacks interpretability and cannot be translated into practice. In contrast, the combination of UMAP and HDBSCAN produces a constrained number of phenotypes, reflecting density-based stability rather than algorithmic overfitting. The emergence of nine clusters, alongside a small proportion of noise, is consistent with the expectation that real-world oncology data contains both structured patterns and irreducible variability.

Clinically, the phenotypes identified in this study reflect three dominant axes of patient organization. The first axis is disease burden, primarily captured by stage. This variable consistently acts as the strongest determinant of cluster formation, reinforcing its central role in treatment planning. The second axis is demographic context, particularly age, which differentiates indolent, long-course malignancies from more aggressive disease trajectories. The third axis is detection pathway, which distinguishes cancers identified through screening or incidental findings from those presenting symptomatically at advanced stages.

These axes are not theoretical constructs; they directly map onto clinical workflows. For example, patients within early-stage dominant phenotypes are more likely to undergo curative-intent treatment, including surgery and radiotherapy with well-defined margins. In contrast, advanced-stage phenotypes correspond to complex cases requiring multimodal therapy, often with significant toxicity considerations. Transitional phenotypes, which comprise a substantial proportion of the cohort, represent the clinical gray zone where individualized decision-making becomes critical.

From a radiation oncology perspective, the implications are particularly relevant. Treatment planning is inherently dependent on anatomical extent and disease burden rather than tumor origin alone. The observation that patients from different cancers cluster together based on shared clinical features supports a more unified framework for treatment stratification. For instance, advanced-stage lung and head–neck cancers, despite differing anatomically, often require similar radiotherapeutic considerations, including large target volumes and organ-at-risk constraints. The clustering results provide empirical support for this convergence.

The study also carries implications for health systems. Phenotypes enriched with late-stage disease highlight areas where early detection remains inadequate. Conversely, clusters dominated by early-stage cancers reflect successful screening and diagnostic pathways. This duality suggests that phenotypic analysis could serve as a population-level tool for identifying gaps in cancer control strategies.

In conclusion, this study provides a proof of concept that clinical phenotyping can complement traditional tumor-based classification. Rather than replacing existing frameworks, this approach adds a clinically grounded layer of stratification aligned with patient presentation and treatment decision-making. In conclusion, the findings support a shift toward phenotype-oriented oncology, where patients are grouped not only by tumor type but by how the disease manifests clinically. This perspective has the potential to improve treatment personalization, optimize resource allocation, and enhance patient outcomes across diverse healthcare settings.

Author Contribution Statement

Khairunnisa Supardi, Soehartati A. Gondhowiardjo, and Rezzy Eko Caraka contributed equally to the conception, design, analysis, interpretation of the study, and preparation of the manuscript. Khairunnisa Supardi, Soehartati A. Gondhowiardjo, Rezzy Eko Caraka, Prana Ugiana Gio, Nyityasmono Tri Nugroho reviewed and approved the final version of the manuscript.

Acknowledgements

General

This study was supported by the Indonesia Endowment Fund for Education (LPDP). The authors would like to express their gratitude to all parties who contributed to the completion of this research.

Funding Statement

This research was fully funded by the Indonesia Endowment Fund for Education (LPDP). The funding body had no role in the design of the study, data collection, analysis, interpretation of data, or in writing the manuscript.

Approval

Not applicable (this study does not involve an institutional or organizational approval as a prerequisite beyond standard ethical clearance procedures).

Data Availability

The datasets analyzed during this study are publicly available and entirely open-source. The harmonized clinical and multi-omics data can be freely accessed and downloaded from the MLOmics: Cancer Multi-Omics Database for Machine Learning via its public GitHub repository (https://github.com/chenzRG/Cancer-Multi-Omics-Benchmark). The foundational raw sequencing data and original clinical metadata are also publicly accessible through the National Cancer Institute's The Cancer Genome Atlas (TCGA) portal. No special access privileges or author permissions are required to replicate the findings of this study.

Conflict of Interest

Ethical Declaration

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

1. Valentini V, Boldrini L, Mariani S, Massaccesi M. Role of radiation oncology in modern multidisciplinary cancer treatment. Mol Oncol. 2020;14(7):1431-41. https://doi.org/10.1002/1878-0261.12712.
2. Kimman M, Norman R, Jan S, Kingston D, Woodward M. The burden of cancer in member countries of the

Association of Southeast Asian Nations (ASEAN). Asian Pac J Cancer Prev. 2012;13(2):411-20. https://doi.org/10.7314/apjcp.2012.13.2.411.
3. Hou C, Huang T, Hu K, Ye Z, Guo J, Zhou H. Artificial intelligence-assisted multimodal imaging for the clinical applications of breast cancer: A bibliometric analysis. Discov Oncol. 2025;16(1):537. https://doi.org/10.1007/s12672-025-02329-1.
4. Caraka RE, Supardi K, Gondhowiardjo SA, Isnaniawardhani V, Gio PU, Chen RC, et al. Artificial intelligence in cancer oncology through comprehensive bibliometric mapping of global trends, impact, and conceptual structures. J Healthc Leadersh. 2025;17:591-621. https://doi.org/10.2147/jhl.S550933.
5. Zubizarreta EH, Fidarova E, Healy B, Rosenblatt E. Need for radiotherapy in low and middle income countries – the silent crisis continues. Clin Oncol (R Coll Radiol). 2015;27(2):107-14. https://doi.org/10.1016/j.clon.2014.10.006.
6. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74(3):229-63. https://doi.org/10.3322/caac.21834.
7. Baeza SM, Lam S. Lung cancer screening—contemporary issues: A narrative review. Curr Chall Thorac Surg. 2023;5.
8. Atun R, Jaffray DA, Barton MB, Bray F, Baumann M, Vikram B, et al. Expanding global access to radiotherapy. Lancet Oncol. 2015;16(10):1153-86. https://doi.org/10.1016/s1470-2045(15)00222-3.
9. Han C, Burd C, Plascak J, Tounkara F, Rosko A, Noonan A, et al. AI-driven chemotoxicity prediction in colorectal cancer: Impact of race, SDOH, and biological aging. BMC Cancer. 2025;25(1):1513. https://doi.org/10.1186/s12885-025-14831-4.
10. Anderson TK. Kernel density estimation and k-means clustering to profile road accident hotspots. Accid Anal Prev. 2009;41(3):359-64. https://doi.org/10.1016/j.aap.2008.12.014.
11. Caraka R, Lee Y, Chen R-C, Toharudin T, Gio P, Kurniawan R, et al. Cluster around latent variable for vulnerability towards natural hazards, non-natural hazards, social hazards in West Papua. IEEE Access. 2021;9:1972-86. https://doi.org/10.1109/ACCESS.2020.3038883.
12. Son LH, Cuong BC, Lanzi PL, Thong NT. A novel intuitionistic fuzzy clustering method for geo-demographic analysis. Expert Syst Appl. 2012;39(10):9848–59. https://doi.org/10.1016/j.eswa.2012.02.167.
13. Pham TD. Clustering data with spatial continuity. 10th IEEE International Conference on Fuzzy Systems. (Cat. No. 01CH37297); 2001 2-5 Dec. 2001.
14. Caraka RE, Supardi K, Gio PU, Isnaniawardhani V, Chen RC, Djatmiko B, et al. Big data analytics for uncovering voxel connectivity patterns in attention-deficit/hyperactivity disorder. J Multidiscip Healthc. 2025;18:4411-30. https://doi.org/10.2147/jmdh.S523137.

This work is licensed under a Creative Commons Attribution-Non Commercial 4.0 International License.