Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Clustering-Based Knowledge Discovery in Breast Cancer: Insights from a Local Clinical Dataset

Oveis Dehghantanha [1], *Nasser Mehrshad* [*,1], *Roksana Bakhshali* [2], *AhmadReza Sebzari* [3]

[1] *Department of Electrical and Computer Engineering, University of Birjand, Birjand, Iran.*

[2] *Omid Cancer Center, Ahvaz, Iran.*

[3] *Department of Internal Medicine, School of Medicine, Cellular and Molecular Research Center, Valiasr Hospital, Birjand University of Medical Sciences.*

Article Info

Article History:
Received 24 April 2025
Reviewed 29 June 2025
Revised 15 July 2025
Accepted 02 August 2025

Keywords:
Breast cancer
Knowledge discovery
Clustering
K-means clustering
Hierarchical clustering

*Corresponding Author's Email Address:
NMehrshad@Birjand.ac.ir

Abstract

Background and Objectives: Understanding the heterogeneity of breast cancer is crucial for improving treatment strategies. This study investigates the application of K-Means and Hierarchical Clustering to a local dataset of breast cancer patients from Iranmehr Hospital, Birjand, Iran, with the primary goal of identifying potential patient subgroups based on their clinical and treatment characteristics for knowledge discovery. The potential of these subgroups to inform future research on personalized treatment approaches is explored.
Methods: A retrospective dataset comprising pathological and clinical information was analyzed using K-Means and Agglomerative Hierarchical Clustering to identify patient subgroups. The optimal number of clusters was consistently determined to be two (k=2) for both methods based on rigorous internal validation metrics (Elbow Method, Silhouette Analysis, Calinski-Harabasz Index, and Largest Jump Analysis for Hierarchical Clustering). Statistical tests (ANOVA and Chi-squared) were employed to assess significant differences in features across the identified clusters from both K-Means and Hierarchical analyses, providing insights into the key factors differentiating these groups. Internal cluster validity was assessed using Silhouette Score and Calinski-Harabasz Index.
Results: The K-Means analysis identified two clusters exhibiting significant differences in characteristics such as age, chemotherapy session intensity, menopausal status, nodal involvement, and biomarker expression (ER, PR, HER2, Ki67). The Hierarchical Clustering also yielded two clusters with varying characteristics, and a comparison between the two methods highlighted both similarities and differences in the identified patient stratifications. The overall agreement between K-Means and Hierarchical Clustering was quantified by an Adjusted Rand Index (ARI) of 0.4697.
Conclusion: Both K-Means and Hierarchical Clustering effectively revealed potential patient subgroups within the studied dataset, highlighting the heterogeneity of breast cancer presentation and treatment at a local level These clusters exhibited statistically significant differences across key clinical and treatment features. Future research is needed to validate these findings in larger, multi-center studies, explore the clinical significance of these subgroups in terms of treatment outcomes, and compare the effectiveness of different clustering methodologies for this purpose.

How to cite this paper:
O. Dehghantanha, N. Mehrshad, R. Bakhshali, A. R. Sebzari "Clustering-based knowledge discovery in breast cancer: insights from a local clinical dataset," J. Electr. Comput. Eng. Innovations, 14(1): 117-144, 2026.
DOI: 10.22061/jecei.2025.11787.835
URL: https://jecei.sru.ac.ir/article_2392.html

Introduction

Cancer is among the most significant contributors to early death globally, potentially surpassing cardiovascular diseases in terms of prevalence and mortality trends in modern society [1], [2]. Recent data indicates that female breast cancer has overtaken lung cancer as the primary form of cancer globally, with approximately 2.3 million new cases reported in 2020, constituting 11.7% of all cancer diagnoses. It is also the fifth most fatal type of cancer worldwide, with 685,000 deaths. Among women, breast cancer is responsible for one-quarter of all cancer diagnoses and one-sixth of all cancer deaths. It is the leading cause of cancer incidence in most countries (159 out of 185) and cancer mortality in 110 countries [3].

Likewise, several risk factors have been strongly associated with increased breast cancer incidence, including obesity [4], sedentary lifestyles [5], diets high in protein—particularly those involving red meat treated with exogenous hormones or carcinogenic compounds [6]—alcohol consumption [7], tobacco use [8], and the use of oral contraceptives [9].

The different complexities involved in understanding cancer, selecting appropriate treatment methods, estimating survival rates, and predicting recurrence create numerous challenging questions for researchers. Clustering techniques offer a data-driven approach to address these complexities by identifying intrinsic groupings within patient data based on similarities in their features. This study seeks to investigate the application of unsupervised learning techniques, specifically K-Means and Hierarchical Clustering, to a novel local dataset of breast cancer patients from Iranmehr Hospital, Birjand, Iran. The primary goal is to uncover potential patient subgroups based on their clinical and treatment characteristics, thereby facilitating knowledge discovery relevant to this specific population.

This study aims to: 1) Analyze a newly collected local dataset from breast cancer patients at Iranmehr Hospital in Birjand, Iran, using their pathological and clinical information. 2) Employ K-Means and Hierarchical Clustering algorithms to perform knowledge discovery and identify potential patient subgroups within this dataset. The innovation of this research lies in the application of these well-established clustering methods to a unique, local dataset to reveal specific patterns of patient stratification relevant to this Iranian population. This approach can contribute to a better understanding of breast cancer heterogeneity within this context, potentially informing future research on tailored treatment strategies. The main contributions of this study are as follows:

1- Collecting a new local dataset from breast cancer patients using the pathological and clinical information of the patients under treatment from Iranmehr Hospital of Birjand, Khorasan-e-jonoubi, Iran.
2- Performing a knowledge discovery analysis using K-Means and Hierarchical Clustering to extract useful knowledge by identifying potential patient subgroups within the collected dataset.

Related Works

Understanding the complex landscape of breast cancer diagnosis and treatment requires robust analytical tools. This section reviews prior work across seven key themes: clustering methodologies and validation, applications in breast cancer, regional dataset-specific studies, broader machine learning contexts, preprocessing practices, clinical interpretability, and innovation. By synthesizing insights from global and local studies, this review positions the current research within the broader field of unsupervised learning for clinical decision support.

A. Clustering Methodologies and ValidationTechniques

Clustering is a foundational unsupervised learning method used to uncover latent patterns in medical datasets, including breast cancer data. K-Means and Hierarchical Clustering remain the most widely applied due to their simplicity and effectiveness in high-dimensional data contexts [10], [13] and [14]. K-Means is especially valued for its computational efficiency, though it assumes spherical clusters, which may oversimplify real-world data distributions [10]. Hierarchical Clustering, particularly with Ward's linkage method, supports interpretability through dendrogram visualization, making it suitable for subgroup analysis in clinical studies [14].

Validation of clustering results is critical. Pison et al. [15] and Rousseeuw [55] emphasize the need for internal validation indices such as the Silhouette Score and CLUSPLOT, which assesses the cohesion and separation of clusters. However, many breast cancer studies still rely on heuristic methods or visual inspection without rigorous quantitative evaluation.

These considerations guided our use of both K-Means and Hierarchical Clustering, complemented by internal validation using Silhouette Scores to ensure methodological rigor.

B. Applications of Clustering in Breast Cancer

Numerous studies have applied clustering to breast cancer data for classification, subtype discovery, and treatment personalization. For instance, Dubey et al. [16] used K-Means to differentiate subtypes in the Wisconsin Breast Cancer dataset, although their focus was largely diagnostic.

Agrawal et al. [17] proposed an ensemble clustering-classification pipeline to uncover latent patient profiles, while Wang et al. [18] developed a consensus clustering

framework to stratify patients based on molecular features. Yet, these methods often underemphasize treatment variables, and their practical clinical relevance remains limited without outcome validation or interpretability.

This gap informed our focus on treatment-centered clustering and statistical validation to ensure clinical utility and interpretability.

C. Regional and Dataset-Specific Studies

Several Iranian studies have explored breast cancer using local datasets. Sajjadnia et al. [19] examined preprocessing effects on clustering outcomes from Shiraz hospitals but lacked treatment-outcome connections. Ahmadi et al. [19] and Hosseini et al. [21] conducted spatial clustering studies, providing regional incidence insights but not patient-level treatment stratification.

These efforts demonstrate the feasibility of clustering in the Iranian context but underline the scarcity of work involving rich clinical-treatment data and robust algorithmic comparison.

In response, our study leverages a detailed, locally curated dataset with diverse clinical and treatment variables to provide a more comprehensive stratification framework.

D. Machine Learning in Breast Cancer: A Broader Context

Machine learning (ML) and deep learning (DL) methods are extensively used for prediction, classification, and prognosis in breast cancer [22]-[24]. However, most works emphasize diagnostic accuracy and often ignore treatment-specific subgrouping. Radak et al. [23] and Xiao et al. [24] highlighted ML's utility in survival prediction, but clustering is typically peripheral or absent in such analyses. Moreover, these models often lack interpretability and practical guidance for treatment decisions.

Our clustering-based approach addresses this by prioritizing subgroup discovery tied directly to therapeutic features and supporting statistical interpretability.

E. Preprocessing and Mixed Data Clustering

Preprocessing is pivotal in ensuring clustering quality. Studies by Guyon and Elisseeff [25] and Zimek et al. [26] highlighted feature selection and outlier detection as essential steps. Given the mixed-type nature of clinical data, Ahmad and Dey [11], Huang [12], and Dinh et al. [27] have proposed K-Means variants and hybrid techniques to handle numerical and categorical values. Boluki et al. [28] suggested avoiding imputation through model-aware clustering, a technique relevant for incomplete medical records.

Accordingly, we incorporated one-hot encoding and standardized scaling to handle mixed data types and ensure the robustness of our clustering outcomes.

F. Clinical Relevance and Model Interpretability

A critical gap in the literature is the clinical interpretability of clusters. Many studies stop at cluster formation without evaluating their medical implications. The current study addresses this by using ANOVA and Chi-square testing to assess statistically significant differences across treatment-relevant features (e.g., ER, PR, HER2, Ki67, chemotherapy regimen), adding interpretability and clinical value.

This approach ensures that the resulting clusters are not only statistically meaningful but also practically relevant for treatment planning in clinical settings.

G. Innovation and Current Contribution

This study presents a locally curated dataset from Iranmehr Hospital, covering 185 patients with 24 demographic, pathological, and treatment-related features. K-Means and Hierarchical Clustering were employed alongside internal validation using the Silhouette Score [55]. Significant statistical testing (ANOVA, Chi-square) highlighted cluster-driving variables, providing actionable insights for treatment stratification. Importantly, this study proposes future exploration of Federated Learning to enable multi-center collaborations without compromising patient data privacy [29], [30], and intends to incorporate alternative clustering methods like DBSCAN [31] and Gaussian Mixture Models [32] to evaluate robustness.

Dataset Description

The present study involved the creation of a unique dataset derived from 185 breast cancer patients receiving treatment at Iranmehr Hospital, Khorasan-e-Jonoubi, Birjand, Iran. This dataset, assembled through a collaborative effort with cancer specialists at the institution, includes a unique identification number for each patient and 24 distinct clinical, pathological, and treatment-related features. These features aim to capture the inherent heterogeneity in breast cancer presentation and management within this specific patient population at Iranmehr Hospital. The characteristics, encoding, and clinical relevance of these features are summarized in Table 1, while descriptive statistics for the numerical variables and frequency distributions for the categorical variables are presented in the subsequent "Numerical Features: Descriptive Statistics" and "Categorical Features: Frequency Distributions" subsections, respectively.

The subsequent application and comparison of established clustering algorithms, including k-means, will leverage these data characteristics to identify potentially clinically meaningful patient subgroups relevant to treatment patterns and outcomes within this cohort.

Table 1: Detailed description and clinical relevance of features in the breast cancer patient dataset

Feature Name	Category	Data Type	Range/Categories	Clinical Relevance
Patients ID	Identifier	Integer	1-197	Unique identifier for each patient.
Age	Demographic	Integer	25-80	Patient's age at diagnosis or treatment.
Sex	Demographic	Integer	1=Female 2=Male	Patient's biological sex.
Menopausal	Demographic	Integer	1=Post 2=Pre (male=2)	Menopausal status, relevant for hormonal influences on breast cancer.
Histological Type	Pathological	Integer	1=IDC 2=ILC 3=Tubular 4=Papillary 5=Mucinous 6=Medullary	Microscopic classification of the tumor, influencing prognosis and treatment.
Focality	Pathological	Integer	1=Uni 2=Multi	Number of tumor foci in the breast.
Marginal Surgery	Clinical	Integer	0=Negative 1=Positive	Presence of residual cancer cells after surgery.
T (Tumor size)	Pathological	Integer	1=TX 2=T0 3=Tis 4=T1 5=T2 6=T3 7=T4	Size of the primary tumor, a key factor in staging and prognosis.
N (Nodal involvement)	Pathological	Integer	1=NX 2=N0 3=N1 4=N2 5=N3	Extent of cancer spread to regional lymph nodes, a critical staging component.
Number of dissected nodes	Pathological	Integer	0-30	Number of lymph nodes removed during surgery for pathological assessment.
Node Dissection	Clinical	Integer	0= No dissection 1=SLND 2=ALND	Whether a lymph node dissection was performed.
Type of Surgery	Treatment	Integer	1=BCS 2=Mastectomy	Type of surgical procedure performed.
Surgeon	Clinical	Integer	1=General 2=Oncosurgeon	Specialty of the surgeon who performed the procedure.
ER (Estrogen Receptor)	Pathological	Integer	0=Negative 1=Positive	Status of Estrogen Receptor in tumor cells, guiding endocrine therapy.
PR (Progesterone Receptor)	Pathological	Integer	0=Negative 1=Positive	Status of Progesterone Receptor in tumor cells, also guiding endocrine therapy.
HER2 (Human Epidermal Growth Factor Receptor 2)	Pathological	Integer	0=Negative 1=Positive	Status of Human Epidermal Growth Factor Receptor 2, indicating potential for targeted therapies.
KI67	Pathological	Integer	0=Negative 1=Positive	Marker of cell proliferation, indicating tumor aggressiveness.
Treatment Schedule	Treatment	Integer	1=Sx>ChT>RT>HoT 2=Sx>ChT>RT 3=Sx>ChT>HoT 4=Sx>ChT 5=Sx>HoT 6=Sx>RT>HoT 7=Sx>RT 8=ChT>Sx>RT>HoT 9=ChT>Sx>RT	Sequence of therapeutic modalities used in treatment.
Chemotherapy Regimen	Treatment	Integer	0=No ChT 1=1st Gen 2=2nd Gen 3=3rd Gen	Specific chemotherapy drug combination used.
Trastuzumab	Treatment	Integer	0=Negative 1=Positive	Whether the patient received Trastuzumab, a HER2-targeted therapy.
Radiation dose	Treatment	Integer	0=No RT 1=50-56 Gy 2=42.5 Gy	Total radiation dose administered during radiotherapy.
Radiation Boost Dose	Treatment	Integer	0=Negative 1=Positive	Whether an additional radiation boost was given to the tumor bed.
Chemotherapy Session	Treatment	Integer	Actual sessions administered: 0, 4, 6, 8, or 16 (derived from codes 0-4)	Number of chemotherapy cycles administered.
Hormonetherapy	Treatment	Integer	0=No HoT 1=Tamoxifen 2=Letrozole 3=Tamoxifen-Letrozole	Type of hormonal therapy received.
GnRH Ana.	Treatment	Integer	0=Negative 1=Positive	Status of Gonadotropin-Releasing Hormone Analog use, primarily in pre-menopausal women.

A. Numerical Features: Descriptive Statistics

The numerical features in our Iranmehr Hospital breast cancer patient dataset, namely Age, Number of dissected nodes, and Chemotherapy Session, are continuous variables that provide quantitative information about the patient cohort. Table 2 presents the descriptive statistics for these key numerical variables, including the count, mean, standard deviation, minimum, 25th percentile, median (50th percentile), 75th percentile, and maximum values. These statistics offer an initial understanding of the central tendency and spread of these continuous characteristics within our cohort, which will be used as input for the subsequent clustering analysis using multiple algorithms, including k-means, to identify potential patient subgroups. Key observations from these statistics include:

- **Age:** The cohort exhibited a mean age of 49.16 years, ranging from 25 to 80 years (Standard Deviation = 11.09).
- **Number of dissected nodes:** The patients in the study had an average of 8.19 dissected lymph nodes, with a range from 0 to 30 (Standard Deviation = 5.85).
- **Chemotherapy Session:** The patients in the study received an average of 12.39 chemotherapy sessions, with the number of sessions ranging from 0 to 16 (Standard Deviation = 4.79).

Table 2: Descriptive Statistics of Numerical Variables in the Breast Cancer Patient Dataset

Feature	Count	Mean	Standard Deviation	Minimum	25th Percentile	50th Percentile (Median)	75th Percentile	Maximum
Age	185	49.16	11.09	25	40	48	56	80
Number of dissected nodes	185	8.19	5.85	0	3	8	12	30
Chemotherapy Session	185	12.39	4.79	0	8	16	16	16

B. Categorical Features: Frequency Distributions

The categorical features in our Iranmehr Hospital breast cancer patient dataset encompass a range of demographic, pathological, clinical, and treatment-related characteristics. Table 3 presents the frequency counts and percentages for each category within these nominal and ordinal variables: Sex, Menopausal, Histological Type, Focality, and so forth.

Table 3: Frequency and Percentage Distribution of Categorical Variables in the Breast Cancer Patient Dataset

Feature	Category	Count	Percentage
Sex	Female	181	97.84
	Male	4	2.16
Menopausal	Post	74	40
	Pre	111	60
Histological Type	IDC	167	90.27
	ILC	11	5.95
	Tubular	1	0.54
	Papillary	3	1.62
	Mucinous	2	1.08
	Medullary	1	0.54
Focality	Unifocal	168	90.81
	Multifocal	17	9.19
	Multicentric	0	0
Marginal Surgery	Negative	173	93.51
	Positive	12	6.49
T (Tumor size)	Tx	6	3.24
	T0	0	0
	Tis	0	0
	T1	40	21.62
	T2	116	62.7
	T3	18	9.73
	T4	5	2.7
N (Nodal involvement)	Nx	14	7.57
	N0	66	35.68
	N1	56	30.27
	N2	29	15.68
	N3	20	10.81
Node Dissection	No dissection	8	4.32
	SLND	25	13.51
	ALND	152	82.16
Type of Surgery	BCS	64	34.59
	Mastectomy	121	65.41
Surgeon	General	137	74.05
	Oncosurgeon	48	25.95
ER	Negative	58	31.35
	Positive	127	68.65
PR	Negative	74	40
	Positive	111	60
HER2	Negative	129	69.72
	Positive	56	30.27
KI67	Negative	70	37.84
	Positive	115	62.16
Treatment Schedule	Sx>ChT>RT>HoT	88	47.57
	Sx>ChT>RT	39	21.08
	Sx>ChT>HoT	21	11.35
	Sx>ChT	15	8.11
	Sx>HoT	2	1.08
	Sx>RT>HoT	1	0.54
	Sx>RT	0	0
	ChT>Sx>RT>HoT	11	5.95
	ChT>Sx>RT	8	4.32
Chemotherapy Regimen	No ChT	3	1.62
	1st Gen	23	12.43
	2nd Gen	41	22.16
	3rd Gen	118	63.78
Trastuzumab	Negative	129	69.72
	Positive	56	30.27
Radiation Dose	No RT	35	18.92
	50-56 Gy	147	79.46
	42.5 Gy	3	1.62
Radiation Boost Dose	Negative	98	53
	Positive	87	47
Hormone Therapy	No HoT	55	29.73
	Tamoxifen	72	38.92
	Letrozole	37	20
	Tamoxifen-Letrozole	21	11.35
GnRH Ana.	Negative	154	83.24
	Positive	31	16.76

The distribution of these variables provides crucial insights into the composition of our patient cohort across different subgroups, which will be considered alongside the numerical features in the subsequent clustering analysis using various algorithms, including k-means, to explore potential patient stratifications. The distribution of these variables is detailed below:

- **Sex**: 181 female (97.84%) and 4 male (2.16%) patients, encoded as 1 and 2.
- **Menopausal:** A critical factor in breast cancer risk stratification and therapeutic planning, this variable was classified as post-menopausal (1) or pre-menopausal (2). Notably, male patients (n=4) were assigned a value of 2 (pre-menopausal) for dataset consistency, despite lacking a biological menopausal status. Among the 185 patients, 74 (40%) were post-menopausal, while 111 (60%) were pre-menopausal (including 4 male patients).
- **Histological type:** Histological type categorizes breast cancer based on tumor cell morphology observed microscopically, influencing prognosis and therapeutic strategies. The dataset includes the following subtypes:
  - Invasive Ductal Carcinoma (IDC): 167 patients (90.27% of the cohort).
  - Invasive Lobular Carcinoma (ILC): 11 patients (5.95%).
  - Tubular Carcinoma: 1 patient (0.54%).
  - Papillary Carcinoma: 3 patients (1.62%).
  - Mucinous Carcinoma: 2 patients (1.08%).
  - Medullary Carcinoma: 1 patient (0.54%)

Subtypes were numerically encoded as 1–6 per their listed order.

- **Focality:** Focality categorizes tumors into three groups based on anatomical distribution:
  - Unifocal (i.e., a single tumor focus; encoded as 1): Observed in 168 patients (90.81%).
  - Multifocal (i.e., multiple invasive tumors confined to the same breast quadrant; encoded as 2): Observed in 17 patients (9.19%).
  - Multicentric (i.e., invasive tumors located in distinct breast quadrants; encoded as 3): Observed in 0 patients (0%).

The prognostic significance of multifocal and multicentric tumors remains debated. While some studies associate these classifications with poorer outcomes [10], others report no significant impact on prognosis [34].

- **Marginal Surgery:** In patients undergoing surgical intervention, this feature indicates the status of surgical margins post-tumor excision, distinguishing between negative margins (0: no residual cancer cells at the resection boundary) and positive margins (1: residual cancer cells detected). Among the cohort, 12 patients (6.49%) exhibited positive margins, while 173 (93.51%) had negative margins.
- **T (Tumor size):** Tumor size (T) reflects the largest diameter of the primary breast tumor. While tumor size and nodal involvement are correlated, both independently contribute to prognostic assessment. Notably, in triple-negative breast cancer (TNBC), the relationship between tumor size, nodal status, and prognosis was significantly attenuated [35].

The dataset includes seven T categories:

  - Tx (encoded as 1): 6 patients (3.24%) – Insufficient data to assess the primary tumor.
  - T0 (encoded as 2): 0 patients – No evidence of a primary tumor.
  - Tis (encoded as 3): 0 patients – Carcinoma in situ (non-invasive malignancy confined to ducts/lobules).
  - T1 (encoded as 4): 40 patients (21.62%) – Tumor diameter ≤2 cm.
  - T2 (encoded as 5): 116 patients (62.70%) – Tumor diameter >2 cm but ≤5 cm.
  - T3 (encoded as 6): 18 patients (9.73%) – Tumor diameter >5 cm.
  - T4 (encoded as 7): 5 patients (2.70%) – Tumor invasion into the chest wall or skin, irrespective of size.

- **N (Nodal involvement):** In breast cancer, Nodal involvement (N) reflects the extent of regional lymph node metastasis and is a critical component of the TNM staging system. This feature evaluates whether cancer has spread to axillary lymph nodes (underarm) or internal mammary lymph nodes (near the breastbone) [36]. The most commonly involved lymph nodes are the axillary lymph nodes (located under the arm) and the internal mammary lymph nodes (located near the breastbone). Clinical staging involves lymph node assessment to determine disease progression beyond breast tissue, with nodal metastasis indicating a higher risk of systemic spread. The dataset includes five N categories:
  - Nx: 14 patients (7.57%) – Lymph nodes could not be assessed.
  - N0: 66 patients (35.68%) – No regional lymph node metastasis.
  - N1: 56 patients (30.27%) – Metastasis in 1–3 axillary or internal mammary nodes.
  - N2: 29 patients (15.68%) – Metastasis in 4–9 axillary or internal mammary nodes.
  - N3: 20 patients (10.81%) – Metastasis in ≥10 axillary nodes, infraclavicular nodes, or supraclavicular nodes.

These categories were numerically encoded as integers 1–5 in the dataset, corresponding to the order listed above.

- **Node Dissection:** For patients undergoing surgical intervention, this feature specifies the type of axillary lymph node assessment performed:
  - No dissection (Encoded as 0): No axillary lymph node dissection or sentinel lymph node biopsy was performed on 8 patients (4.32% of the cohort).
  - Sentinel Lymph Node Biopsy (SLND) (Encoded as 1): Identification and removal of the first few lymph nodes to which cancer cells are most likely to spread, performed on 25 patients (13.51%).
  - Axillary Lymph Node Dissection (ALND) (Encoded as 2): Surgical removal of multiple lymph nodes in the armpit, performed on 152 patients (82.16%).
- **Type of Surgery:** For patients undergoing surgical intervention, this feature specifies the surgical procedure performed:
  - Breast-Conserving Surgery (BCS) (Encoded as 1): Partial excision of the tumor with preservation of breast tissue, performed on 64 patients (34.59% of the cohort).
  - Mastectomy (Encoded as 2): Complete removal of the affected breast tissue, performed on 121 patients (65.41%).
- **Surgeon:** This feature identifies the surgical specialist who performed the procedure:
  - General Surgeon (encoded as 1): Performed on 137 patients (74.05% of the cohort).
  - Oncosurgeon (encoded as 2): Specialized in oncologic surgery, performed on 48 patients (25.95%).

While surgeon specialty is not a direct prognostic factor, differences in surgical training (e.g., general vs. oncologic surgery) may reflect variations in technique, institutional protocols, or postoperative care, which could act as confounding variables in outcome analyses.

- **ER (Estrogen Receptor):** ER status, a feature with critical prognostic and therapeutic relevance, was categorized as follows:
  - ER-negative (encoded as 0): 58 patients (31.35%).
  - ER-positive (encoded as 1): 127 patients (68.65%).

ER-positive tumors are more likely to exhibit histological differentiation [37]-[39], lower proliferative activity [40], and diploid DNA content. They are also less frequently associated with high-risk genetic alterations, such as TP53 mutations [41] and [42], HER2/neu amplification [43]-[45], or HER1 (the epidermal growth factor receptor [EGFR]) [46] and [47], which are linked to aggressive tumor behavior and poorer prognosis. Conversely, ER-negative tumors demonstrate higher rates of these molecular aberrations, contributing to their adverse clinical outcomes. ER status remains pivotal in guiding therapeutic decisions, including endocrine therapy for ER-positive cases.

- **PR (Progesterone Receptor):** Progesterone receptor (PR) status, an independent prognostic marker distinct from ER, was categorized as follows:
  - PR-negative (encoded as 0): 74 patients (40.00%).
  - PR-positive (encoded as 1): 111 patients (60.00%).

PR negativity in ER-positive tumors correlates with a more aggressive subtype of hormone receptor-positive breast cancer [48], often classified as the luminal B molecular subtype [49]. These tumors are associated with higher proliferative rates and poorer clinical outcomes compared to ER-positive/PR-positive (luminal A) tumors.

- **HER2 (Human Epidermal Growth Factor Receptor 2):** A protein that promotes cancer growth. HER2-positive cancers are more aggressive but may respond to drugs like trastuzumab (Herceptin). This factor was categorized as follows:
  - HER2-negative (encoded as 0): 129 patients (69.72%).
  - HER2-positive (encoded as 1): 56 patients (30.27%).
- **KI67:** KI67, a nuclear protein marker of cellular proliferation, was categorized as follows:
  - KI67-negative (encoded as 0): 70 patients (37.84%) – Defined as KI67 expression ≤10%.
  - KI67-positive (encoded as 1): 115 patients (62.16%) – Defined as KI67 expression >10%.

Higher KI67 levels correlate with increased tumor aggressiveness and proliferative activity, serving as a prognostic indicator for disease progression and treatment response.

- **Treatment Schedule:** Treatment schedules, representing combinations of therapeutic modalities, were categorized into nine plans:
  - Surgery → Chemotherapy → Radiation → Hormone Therapy (88 patients, 47.57%).
  - Surgery → Chemotherapy → Radiation (39 patients, 21.08%).
  - Surgery → Chemotherapy → Hormone Therapy (21 patients, 11.35%).
  - Surgery → Chemotherapy (15 patients, 8.11%).
  - Surgery → Hormone Therapy (2 patients, 1.08%).
  - Surgery → Radiation → Hormone Therapy (1 patient, 0.54%).
  - Surgery → Radiation (0 patients, 0%).
  - Chemotherapy → Surgery → Radiation → Hormone Therapy (11 patients, 5.95%).
  - Chemotherapy → Surgery → Radiation (8 patients, 4.32%).

These subtypes were encoded as integers (1–9) in the dataset, corresponding to the order listed above.

These schedules reflect clinical decision-making based on tumor biology, stage, and patient-specific factors. The predominance of multimodal therapy (e.g., Plan 1:

47.57%) underscores the integration of adjuvant strategies to mitigate recurrence risk.

- **Chemotherapy Regimen:** The specific combination of chemotherapy drugs used in treatment. Different regimens are selected based on the cancer's characteristics. Four different regimens are used for the patients in this study:
  - No Chemotherapy (encoded as 0): 3 patients (1.62%).
  - First generation (encoded as 1): 23 patients (12.43%).
  - Second generation (encoded as 2): 41 patients (22.16%).
  - Third generation (encoded as 3): 118 patients (63.78%).

The selection of a chemotherapy regimen can be individualized based on several factors, such as the risk of recurrence and the potential benefits of chemotherapy, both relative and absolute. It is also important to consider patient-specific factors like age, comorbidities, and risk tolerance [50]. The decision aids can help patients and caregivers make informed choices about their treatment. Table 4 shows the commonly recommended adjuvant chemotherapy regimens [50].

- **Trastuzumab:** A targeted therapy for HER2-positive breast cancer. This field indicates whether the patient received trastuzumab as part of their treatment. Thus, for 56 patients (30.27%), with HER2-positive breast cancer, trastuzumab was administered.
- **Radiation Dose:** The total amount of radiation given during treatment, typically measured in Gray (Gy). In the proposed dataset, the values used for the radiation dose are categorized into three groups:
  - no radiation dose (encoded as 0): 35 patients (18.92%).
  - 50 Gy, 54 Gy and 56 Gy (encoded as 1): 147 patients (79.46%)
  - 42.5 Gy (encoded as 2): 3 patients (1.62%)
- **Radiation Boost Dose:** The administration of additional radiation to the tumor site following standard radiation therapy serves to minimize the likelihood of cancer recurrence. Among the patients treated with radiotherapy, 87 (47.02%) had received a boost dose.
- **Hormone Therapy:** Indicates whether the patient received hormone therapy (e.g., Tamoxifen or aromatase inhibitors) to block hormone-sensitive cancer growth. This feature has 4 values including:
  - no hormone therapy (encoded as 0): 55 patients (29.73%).
  - Tamoxifen therapy (encoded as 1): 72 patients (38.92).
  - Letrozole therapy (encoded as 2): 37 patients (20%).
  - Tamoxifen and Letrozole therapy (encoded as 3): 21 patients (11.35).

Table 4: Commonly recommended adjuvant chemotherapy regimens [50]

Recurrence risk category and definition	Recommended regimens: ER-positive, HER2 negative	Recommended regimens: ER/PR negative, HER2-negative	Recommended regimens: HER2-positive
Node-Neg, T1a (very low risk)	No chemotherapy	No chemotherapy	No chemotherapy
Node-Neg, T1b (low risk)	Consider second generation Chemotherapy regimen if RS is high	Consider second generation chemotherapy regimen	Consider weekly paclitaxel + H
Node-Neg, T1c (low risk)	Second generation chemotherapy regimen if RS is high (or consider if intermediate)	Second generation chemotherapy regimen	Weekly paclitaxel + H or TCH
Node-Neg, T2 (moderate risk)	Second or third generation chemotherapy regimen if RS intermediate-high	Third generation chemotherapy regimen	AC-T + H or TCH + P
1+ Pos Nodes or T3 (high risk)	Third generation chemotherapy regimen if RS intermediate high (or 4+ positive nodes irrespective of RS)	Third generation chemotherapy regimen	AC-T + H or TCH +P

- **GnRH Ana (Gonadotropin-Releasing Hormone Analog):** GnRH analogs, used to suppress ovarian estrogen production via pituitary gland modulation, were categorized as follows:
  - GnRH Ana-negative (encoded as 0): 154 patients (83.24%).
  - GnRH Ana-positive (encoded as 1): 31 patients (16.76%).

These agents are primarily administered to pre-menopausal women with hormone receptor-positive breast cancer to induce ovarian suppression, thereby depriving tumors of estrogen and slowing disease progression [51].

Data Visualization

To gain an initial understanding of the characteristics and distributions of the features within the Iranmehr Hospital breast cancer patient dataset, a series of visualizations are presented in this subsection. These visualizations offer insights into the central tendencies, spread, and frequencies of both numerical and categorical variables across the entire cohort of 185 patients. By examining these distributions, we aim to highlight the inherent variability within the dataset, which subsequent clustering analysis will explore to identify potential patient subgroups. The distributions of the numerical and categorical features within the

dataset are illustrated in the following figures:

Fig. 1 illustrates the age distribution of the 185 breast cancer patients in the Iranmehr Hospital cohort. The histogram reveals a range of ages from 25 to 80 years, with a central tendency around the late 40s and early 50s. The mean age of the cohort is 49.16 years (SD = 11.09). The distribution appears slightly right-skewed, indicating a relatively higher frequency of older patients. This broad age range suggests the potential for age to be a differentiating factor in identifying patient subgroups through subsequent clustering analysis.

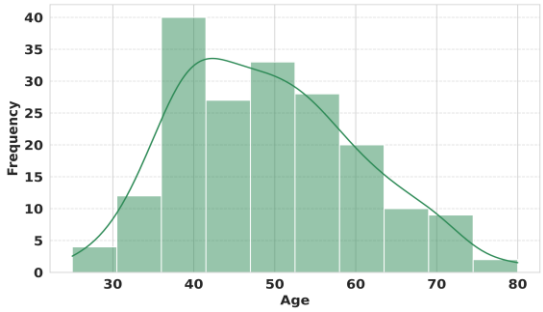

Fig. 1: Distribution of age.

Fig. 2 presents the distribution of the number of dissected lymph nodes, ranging from 0 to 30 (mean = 8.19, SD = 5.85). The distribution is right-skewed, with a higher frequency of patients having fewer dissected nodes. The median was 8. This variability in the extent of nodal assessment might contribute to the heterogeneity observed in the patient population and could potentially be a factor in distinguishing subgroups during clustering.

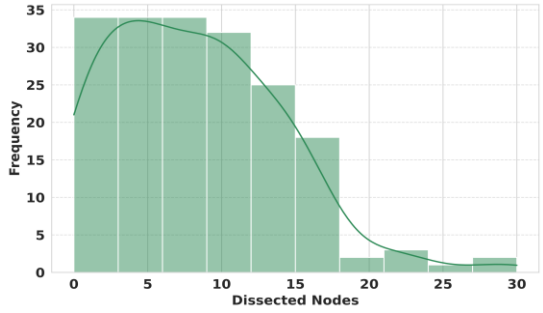

Fig. 2: Distribution of number of dissected nodes.

Fig. 3 illustrates the distribution of the actual number of chemotherapy sessions administered to the 185 patients, based on the current dataset and the defined mapping (where raw codes 0, 1, 2, 3, and 4 correspond to 0, 4, 6, 8, and 16 actual sessions, respectively). The count plot reveals that the majority of patients, 115 (62.2%), underwent 16 actual chemotherapy sessions (code 4). The number of patients receiving 8 actual sessions (code 3) was also notable at 34 (18.4%), followed by 24 patients (13.0%) receiving 6 actual sessions (code 2), and 9 patients (4.9%) receiving 4 actual sessions (code 1). A small subset of 3 patients (1.6%) received no chemotherapy (0 actual sessions; code 0). The mean number of actual chemotherapy sessions for the cohort was 12.39 (SD = 4.79). The median (50th percentile) number of actual sessions was 16, and the 75th percentile was also 16 sessions. This distribution highlights the variability in chemotherapy intensity administered within the cohort, a characteristic that will be considered in subsequent analyses.

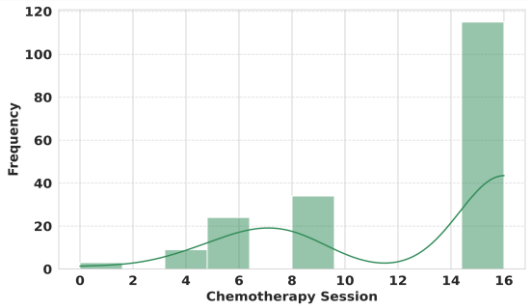

Fig. 3: Distribution of chemotherapy session.

Fig. 4 shows the distribution of sex, with a clear predominance of female patients (97.84%) compared to males (2.16%)

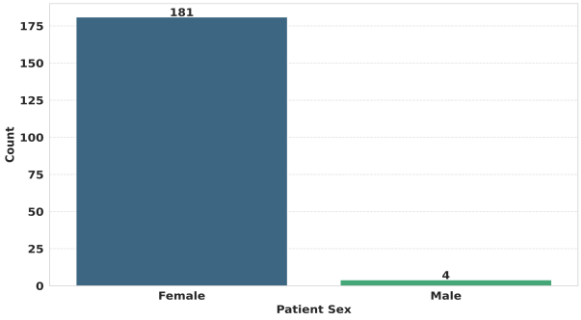

Fig. 4: Distribution of sex.

Fig. 5 illustrates the distribution of menopausal status within the patient cohort. The bar plot shows a higher proportion of pre-menopausal patients (60.0%, 111 individuals) compared to post-menopausal patients (40.0%, 74 individuals). This distribution, where both groups are substantially represented, suggests that menopausal status, with its associated hormonal variations, could be a relevant factor in distinguishing potential patient subgroups in subsequent clustering analyses.

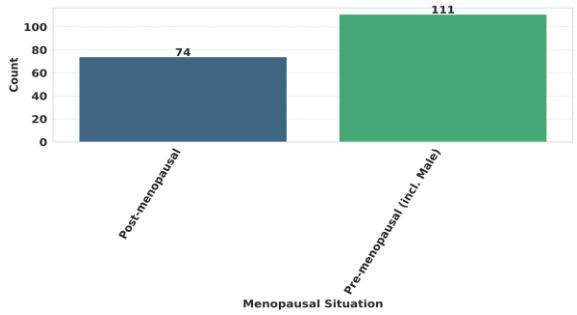

Fig. 5: Distribution of menopausal status.

Fig. 6 illustrates the distribution of histological types, with Invasive Ductal Carcinoma (IDC) being the most common (90.27%). Other types, including ILC (5.95%), Papillary, Mucinous, Medullary, and Tubular, were less frequent.

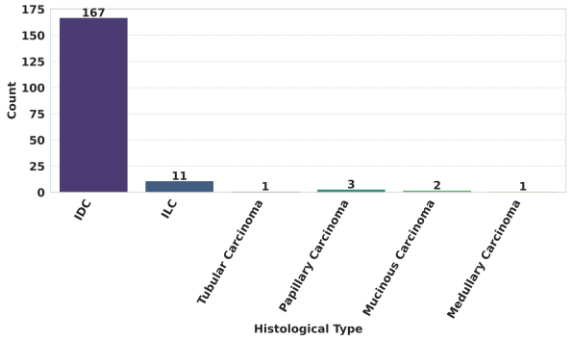

Fig. 6: Distribution of histological types.

Fig. 7 shows that the majority of tumors were unifocal (90.81%), with multifocal tumors being less common (9.19%)

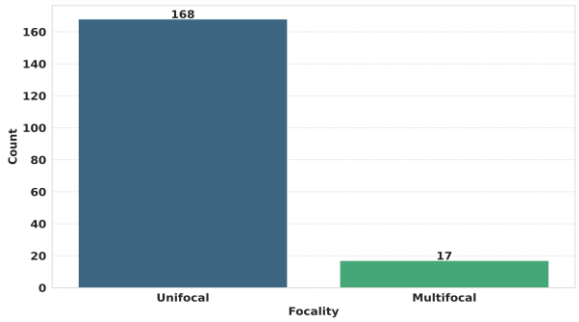

Fig. 7: Distribution of focality.

Fig. 8 displays the distribution of surgical margin status, with most patients having negative margins (93.51%) and a smaller proportion having positive margins (6.49%).

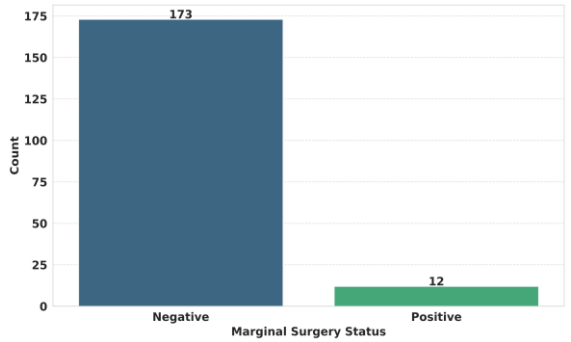

Fig. 8: Distribution of marginal surgery.

Fig. 9 illustrates the distribution of T stages, which categorize the size and extent of the primary tumor, within the patient cohort. The bar plot shows that the most frequent T stage is T2, accounting for 62.70% of the cases. T1 tumors represent the next largest group at 21.62%, followed by T3 at 9.73% and T4 at 2.70%. TX (where the tumor size could not be assessed) and T0 (no evidence of primary tumor) are less common, at 3.24% and 0% respectively.

Tis (carcinoma in situ) also has a frequency of 0%. This distribution highlights the predominance of T2 tumors in this cohort, while also showing the presence of other tumor sizes, which may correlate with disease progression and treatment approaches.

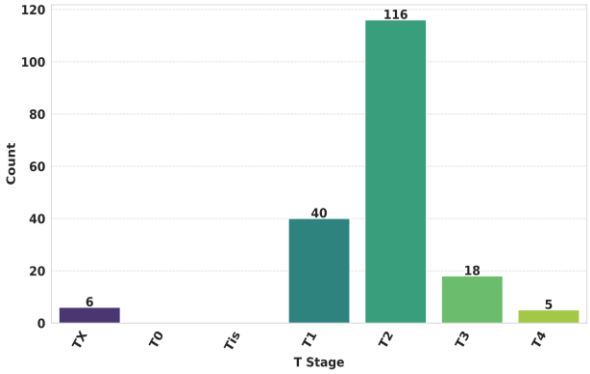

Fig. 9: Distribution of T stage.

Fig. 10 illustrates the distribution of N stages, indicating the extent of regional lymph node involvement, within the patient cohort. The bar plot shows that the most frequent N stage is N0 (code 2), representing no regional lymph node metastasis, which accounts for 35.68% of the cases. N1 (code 3), indicating metastasis to movable ipsilateral axillary lymph nodes, is also common at 30.27%. N2 (code 4), representing metastasis to fixed or matted ipsilateral axillary lymph nodes, occurs in 15.68% of patients, while N3 (code 5), indicating metastasis to infraclavicular or supraclavicular lymph nodes, is seen in 10.81% of cases. NX (code 1), where regional lymph nodes could not be assessed, is the least frequent at 7.57%. This distribution highlights the varying degrees of nodal involvement in this cohort, a critical factor in determining prognosis and treatment strategies.

Fig. 11 illustrates the distribution of the type of nodal assessment performed in the patient cohort. The bar plot reveals that Axillary Lymph Node Dissection (ALND), a more extensive surgical procedure involving the removal of multiple lymph nodes in the armpit, was the most common approach (82.16%). Sentinel Lymph Node Biopsy (SLND), a less invasive procedure to identify and remove only the first few lymph nodes to which cancer cells are most likely to spread, was performed in 13.51% of the patients. A small proportion of patients (4.32%) did not undergo any nodal dissection. The high prevalence of ALND suggests that a comprehensive assessment of axillary lymph nodes was the standard practice for a majority of this cohort, potentially reflecting the clinical stage and risk profiles of the patients.

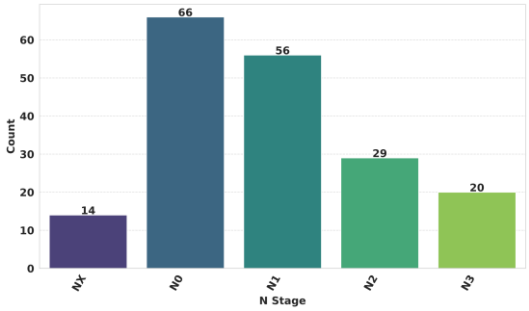

Fig. 10: Distribution of N stage.

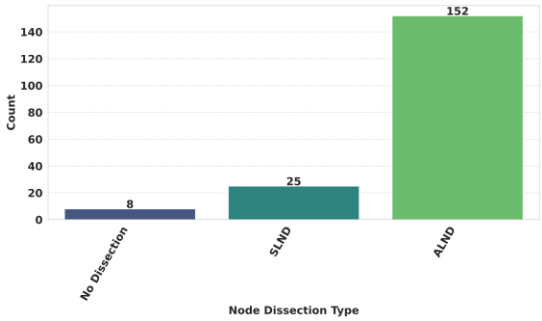

Fig. 11: Distribution of node dissection type.

Fig. 12 illustrates the distribution of the type of surgical procedure performed in the patient cohort. The bar plot reveals that Mastectomy (removal of the entire breast) was the more frequent surgical approach, accounting for 65.41% of the cases. Breast-Conserving Surgery (BCS), which involves the removal of the tumor and some surrounding tissue, was performed in 34.59% of the patients. The significant difference in the frequency of these two surgical types suggests that the extent of surgical intervention varied considerably within the cohort, potentially reflecting differences in tumor size, stage, or patient preference, and could be a relevant factor in distinguishing patient subgroups.

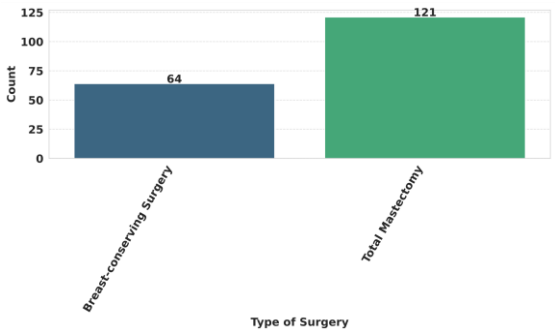

Fig. 12: Distribution of type of surgery.

Fig. 13 illustrates the distribution of the type of surgeon who performed the primary surgical procedure. The majority of surgeries (74.05%) were performed by general surgeons, while oncosurgeons performed 25.95% of the cases. This distribution may reflect the availability of specialists or the complexity of the surgical cases within the cohort.

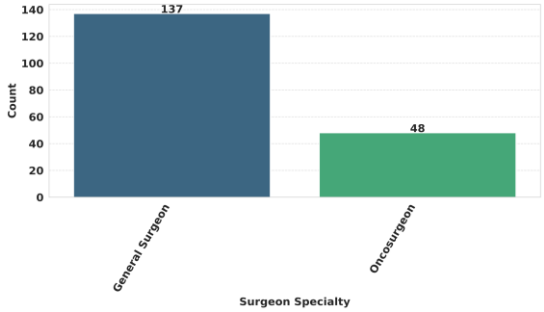

Fig. 13: Distribution of surgeon.

Fig. 14 presents the distribution of Estrogen Receptor (ER) status within the patient cohort. The bar plot shows that the majority of patients (68.65%) had ER-positive tumors, while 31.35% of the tumors were ER-negative. Estrogen Receptor status is a critical biomarker in breast cancer, influencing prognosis and guiding treatment decisions, particularly the use of hormone therapies. The predominance of ER-positive tumors in this cohort suggests that a substantial proportion of patients may be candidates for endocrine treatments, and this feature is likely to be an important factor in defining clinically relevant patient subgroups.

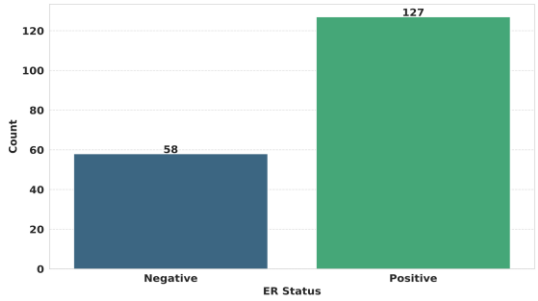

Fig. 14: Distribution of ER status.

Fig. 15 illustrates the distribution of Progesterone Receptor (PR) status within the patient cohort. The bar plot indicates that a majority of the tumors (60.0%) were PR-positive, while 40.0% were PR-negative. Similar to ER, PR status is an important hormone receptor that influences breast cancer biology and response to endocrine therapies. The substantial proportion of PR-positive tumors in this cohort suggests that many patients may benefit from hormonal treatments, and this feature likely contributes to the heterogeneity observed across different patient subgroups.

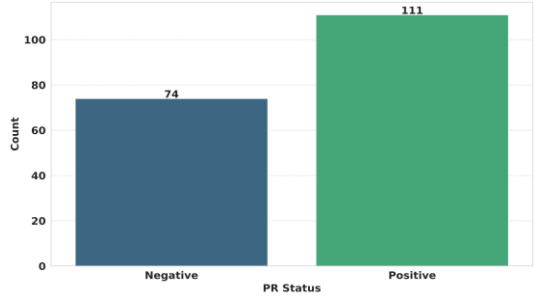

Fig. 15: Distribution of PR status.

Fig. 16 presents the distribution of Human Epidermal Growth Factor Receptor 2 (HER2) status within the patient cohort. The bar plot indicates that a substantial proportion of patients (69.73%) had HER2-negative tumors, while 30.27% of the tumors were HER2-positive. HER2 is a protein that can promote the growth of cancer cells. In breast cancer, HER2 status is a crucial biomarker, impacting treatment strategies, particularly the use of targeted therapies like trastuzumab. The presence of HER2-positive tumors in a notable fraction of the cohort underscores the importance of HER2 testing in guiding personalized treatment approaches.

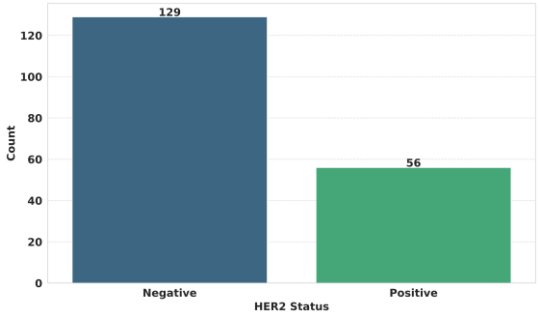

Fig. 16: Distribution of HER2 status.

Fig. 17 illustrates the distribution of KI67 status within the patient cohort. KI67 is a cellular marker associated with cell proliferation, and its expression level is often used to assess tumor aggressiveness. The bar plot shows that the majority of patients (62.16%) had Ki67-positive tumors, indicating a higher level of cell proliferation, while 37.84% had KI67-negative tumors. KI67 status is an important prognostic and predictive factor in breast cancer, often influencing treatment decisions, particularly regarding chemotherapy. The observed distribution suggests a considerable proportion of tumors in this cohort exhibit higher proliferative activity.

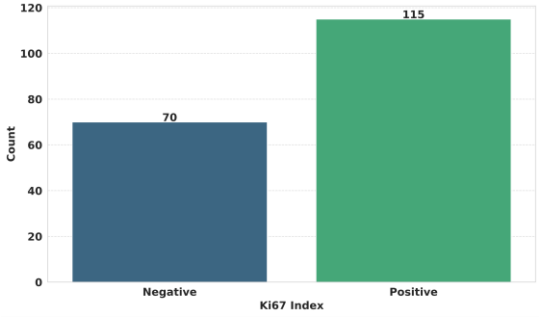

Fig. 17: Distribution of KI67 status.

Fig. 18 illustrates the distribution of different treatment schedules employed in the patient cohort. The most frequent approach was Surgery followed by Chemotherapy, Radiation, and Hormone Therapy (Sx>ChT>RT>HoT), accounting for 47.57% of the patients. The next most common schedules were Surgery followed by Chemotherapy and Radiation (Sx>ChT>RT) at 21.08%, and Surgery followed by Chemotherapy and Hormone Therapy (Sx>ChT>HoT) at 11.35%. Less frequent schedules included Surgery followed by Chemotherapy alone (8.11%), Chemotherapy followed by Surgery, Radiation, and Hormone Therapy (5.95%), and Chemotherapy followed by Surgery and Radiation (4.32%). The remaining schedules, Surgery followed by Hormone Therapy, and Surgery followed by Radiation and Hormone Therapy, were relatively rare. This distribution highlights the variability in treatment strategies, reflecting clinical decision-making based on tumor characteristics, stage, and patient-specific factors.

Fig. 19 illustrates the distribution of different chemotherapy regimens administered to the patient cohort. The most frequently used regimen was the Third Generation chemotherapy, accounting for 63.78% of the patients. Second Generation chemotherapy was the next most common at 22.16%, followed by First Generation chemotherapy at 12.43%. A small subset of patients (1.62%) did not receive any chemotherapy. The variation in chemotherapy regimens likely reflects differences in treatment protocols based on tumor characteristics, stage of disease, and clinical guidelines, and this feature is important for understanding potential differences in treatment response and outcomes across patient subgroups.

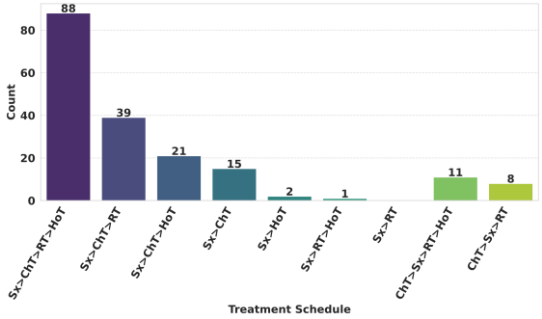

Fig. 18: Distribution of treatment schedules.

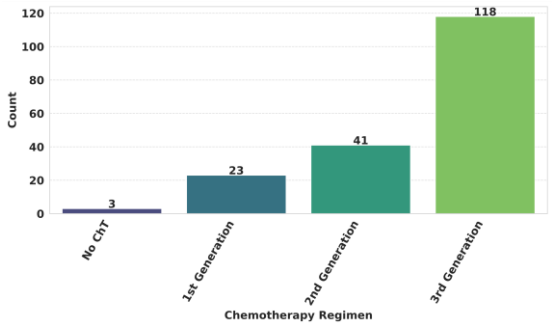

Fig. 19: Distribution of chemotherapy regimen.

Fig. 20 illustrates the distribution of trastuzumab use within the patient cohort. Trastuzumab is a targeted therapy used in patients with HER2-positive breast cancer. The bar plot shows that the majority of patients (69.73%) did not receive trastuzumab, while 30.27% of

patients were treated with this agent. The use of trastuzumab is directly linked to the HER2 status of the tumor, and its administration in a subset of the cohort reflects the prevalence of HER2-positive disease and the application of targeted therapies in these cases. This feature is crucial for understanding treatment strategies and potential differences in outcomes based on HER2 status.

Fig. 21 illustrates the distribution of radiation doses administered to the patient cohort. The majority of patients (79.46%) received a conventional radiation dose in the range of 50-56 Gy. A notable proportion of patients (18.92%) did not receive radiation therapy (No RT), while a small fraction (1.62%) received a dose of 42.5 Gy. Radiation therapy is a key component of breast cancer treatment for many patients, and the variation in dosage reflects differences in treatment protocols based on tumor stage, location, and other clinical factors. The predominance of the 50-56 Gy range suggests a standard radiation protocol for a large segment of this cohort.

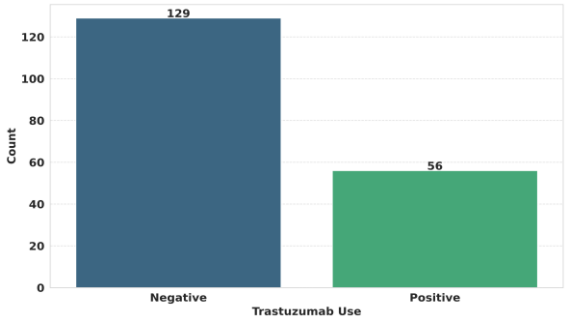

Fig. 20: Distribution of trastuzumab use.

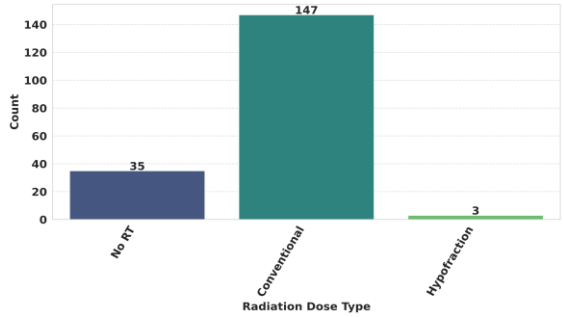

Fig. 21: Distribution of radiotion dose.

Fig. 22 illustrates the distribution of whether patients received a radiation boost dose in addition to their primary radiation therapy. The bar plot shows that slightly more than half of the patients (52.97%) did not receive a boost dose, while 47.03% did. A radiation boost is an additional, focused dose of radiation to the tumor bed after the main course of radiotherapy. The decision to administer a boost depends on various factors, including the size and grade of the original tumor, margin status after surgery, and individual patient risk factors. The near-even distribution suggests that boost radiation was a significant consideration in the treatment protocols for this cohort.

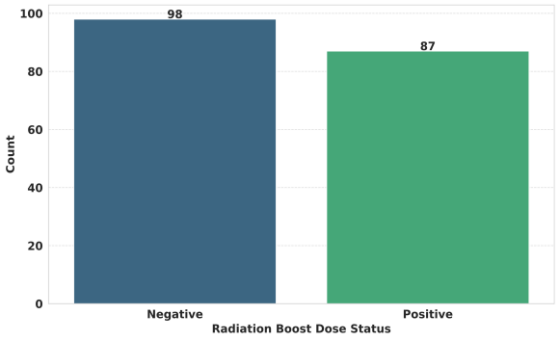

Fig. 22: Distribution of radiation boost dose.

Fig. 23 illustrates the distribution of different hormonetherapy treatments received by the patient cohort. Tamoxifen was the most frequently used agent (38.92%), followed by patients who did not receive hormonetherapy (No HoT, 29.73%). Letrozole was used in 20.00% of the cases, and a combination of Tamoxifen and Letrozole was administered to 11.35% of the patients. Hormonetherapy is a critical adjuvant treatment for hormone-sensitive breast cancers (ER-positive and/or PR-positive), and the distribution of different agents likely reflects clinical guidelines and patient characteristics, such as menopausal status and specific tumor biology. This feature is important for understanding the endocrine treatment landscape within this cohort.

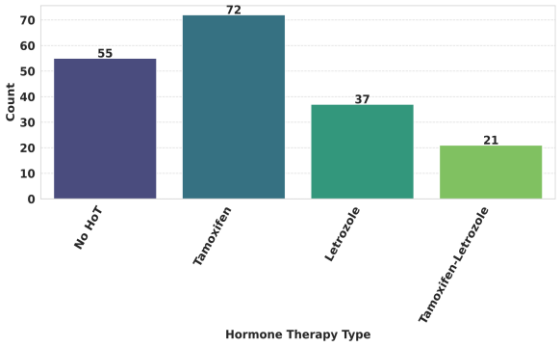

Fig. 23: Distribution of hormonetherapy type.

Fig. 24 illustrates the distribution of Gonadotropin-Releasing Hormone (GnRH) analog use within the patient cohort. GnRH analogs are primarily used in pre-menopausal women with hormone-sensitive breast cancer to suppress ovarian function, thereby reducing estrogen production. The bar plot shows that the majority of patients (83.24%) did not receive GnRH analogs, while 16.76% did. The use of GnRH analogs in a subset of the cohort suggests that these patients were likely pre-menopausal and had hormone-sensitive disease where ovarian suppression was deemed a beneficial treatment strategy. This feature provides insights into the hormonal treatment approaches employed in this specific patient population.

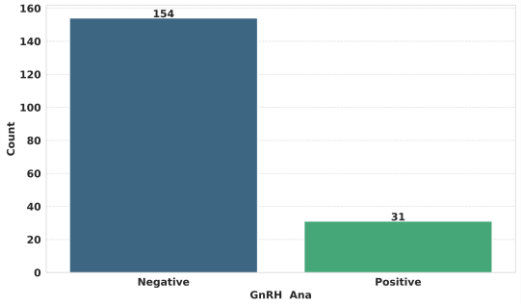

Fig. 24: Distribution of GnRH use.

Summary of Data Characteristics

The dataset for this study comprises 24 distinct clinical, pathological, and treatment-related features collected from 185 breast cancer patients at Iranmehr Hospital. The aim of this data collection was to capture the heterogeneity inherent in breast cancer presentation and management within this specific patient population, thereby enabling the identification of potentially clinically meaningful patient subgroups through subsequent clustering analysis.

The numerical features of the cohort included an age range of 25 to 80 years, with a mean age of 49.16 years. The number of dissected lymph nodes varied from 0 to 30, with an average of 8.19, displaying a right-skewed distribution. Patients received between 0 and 16 chemotherapy sessions, with a mean of 12.39 sessions and a majority receiving 16 sessions.

The categorical features revealed a predominantly female cohort with a notable representation of both pre- and post-menopausal patients. Invasive Ductal Carcinoma was the most common histological subtype, and the majority of tumors were unifocal with negative surgical margins. The distribution of T and N stages indicated a prevalence of T2 and N0 classifications, respectively. Mastectomy was the more frequent surgical procedure. In terms of biomarkers, ER-positive and PR-positive status were more common than negative, while HER2-negative tumors were more frequent than HER2-positive. A majority of tumors exhibited KI67-positive status, indicating higher proliferative activity. The most frequently employed treatment schedule involved surgery followed by chemotherapy, radiation, and hormone therapy, with third generation chemotherapy being the most common regimen. Hormonetherapy most often involved Tamoxifen, and the use of GnRH analogs was relatively infrequent. Axillary Lymph Node Dissection was the predominant type of nodal assessment, and most surgeries were performed by general surgeons.

The variability observed across these numerical and categorical features underscores the heterogeneity within the Iranmehr Hospital breast cancer patient cohort. This inherent diversity provides a strong rationale for employing clustering algorithms to explore the underlying structure of the data and to identify potential patient subgroups that may exhibit distinct patterns in their disease characteristics and treatment approaches. The subsequent sections of this manuscript will detail the application of these clustering methodologies to this dataset.

Methodology

This section details the methodological approach employed to discover potential knowledge and patterns within the breast cancer treatment data from Iranmehr Hospital. The data preprocessing and clustering algorithms were implemented in Python, utilizing libraries such as scikit-learn, and the computational experiments were conducted using Google Colaboratory. It encompasses the steps taken to preprocess the raw data, the implementation of two distinct clustering algorithms – K-means and Hierarchical Clustering – and the methods used to evaluate the resulting clusters.

A. Data Preprocessing

To prepare the breast cancer treatment data for clustering analysis, several preprocessing steps were undertaken.

Handling Missing Values: The initial dataset included 197 patient records. However, a number of these records had incomplete information for certain variables that could not be reliably obtained. To ensure data integrity and avoid potential bias from imputation, a listwise deletion approach was employed, resulting in a final dataset of 185 patients with complete data across all analyzed features.

Feature Classification and Initial Transformation: Prior to further feature transformation, all column names were standardized by stripping leading/trailing spaces, replacing spaces with underscores, and removing specific special characters such as asterisks and periods. Non-breaking spaces were also converted to underscores, and any resulting double underscores were reduced to single underscores to ensure uniformity and facilitate programmatic access (e.g., df.columns.str.strip().str.replace(' ', '_').str.replace('*', '').str.replace('.', '').str.replace(' ', '_').str.replace('__', '_')). The 'patients_ID' column, serving as a unique identifier wes excluded from the feature set used for clustering.

The remaining 24 features were systematically classified and then transformed based on their inherent data types:

- **True Numerical Features (Continuous or Discrete Count):** These features represent measurable quantities with inherent order and meaningful distances between values. This category included:
  - Age

- Dissected_Nodes
- Chemotherapy_Session

- **Ordinal Categorical Features:** These features represent categories with a clear, inherent order, even if the numerical difference between categories is not uniform. For these features, explicit re-mapping was performed to preserve their ordinality:
  - T (Tumor size and extent): Original codes 1-7 (representing TX, T0, Tis, T1, T2, T3, T4) were re-mapped to a sequential numerical scale as follows: 1=TX to 0, 2=T0 to 1, 3=Tis to 2, 4=T1 to 3, 5=T2 to 4, 6=T3 to 5, 7=T4 to 6.
  - N (Nodal involvement): Original codes 1-5 (representing NX, N0, N1, N2, N3) were re-mapped as: 1=NX to 0, 2=N0 to 1, 3=N1 to 2, 4=N2 to 3, 5=N3 to 4.
  - Node_Dissection: Original codes 0-2 (representing No dissection, SLND, ALND) were re-mapped as: 0=No dissection to 0, 1=SLND to 1, 2=ALND to 2.
  - Chemotherapy_Regimen: Original codes 0-3 (representing No ChT, 1st Generation, 2nd Generation, 3rd Generation) already possessed an appropriate sequential order (0, 1, 2, 3) for direct use as numerical values.
- **Nominal Categorical Features:** These features represent categories without any inherent order or ranking. This category included:
  - Sex
  - Menopausal_Situation
  - Histological_Type
  - Focality
  - Marginal_Surgery
  - Type_of_Surgery
  - Surgeon
  - ER
  - PR
  - HER2
  - Ki67
  - Trastuzumab
  - Treatment_Schedule
  - Radiation_dose
  - Radiation_Boost_Dose
  - Hormonetherapy
  - GnRH_Ana

Feature Scaling: All True Numerical features and the re-mapped Ordinal Categorical features were subjected to Feature Scaling using the StandardScaler from the scikit-learn library. This standardization method transforms these features to have a mean of zero and a standard deviation of one, ensuring that they contribute equally to the distance calculations performed by the clustering algorithms and preventing features with larger scales from dominating the results.

One-Hot Encoding: For the Nominal Categorical features, One-Hot Encoding was applied. This process converts each categorical variable into new binary (0 or 1) columns, one for each unique category. This transformation is crucial to prevent the algorithms from misinterpreting arbitrary numerical labels (e.g., 1, 2, 3) as ordinal relationships, which would distort distance calculations.

Handling Outliers: During the initial data exploration, some data points appeared as potential outliers based on the distribution of certain numerical features. However, upon further review and considering the clinical context of the data, it was determined that these extreme values represented genuine variations within the patient cohort and were not due to measurement errors or anomalies. Therefore, these potential outliers were retained in the dataset to ensure a comprehensive representation of the patient population. This decision acknowledges that the heterogeneity inherent in clinical data may result in values that appear statistically distant from the mean but are nonetheless valid observations.

Feature Selection: For this exploratory study, all 24 available clinical and treatment-related features were initially included as input for both the K-Means and Hierarchical Clustering algorithms. The rationale for this comprehensive inclusion was to provide an unbiased view of the patient characteristics and treatment modalities, allowing the algorithms to identify potential subgroups based on the entirety of the available information without imposing premature assumptions on feature importance. While this approach maximizes the breadth of initial knowledge discovery, it is acknowledged that no explicit feature selection or dimensionality reduction techniques were applied at this stage to specifically optimize cluster separability. The implications of this approach, particularly in relation to the observed internal validity scores, are further discussed in the 'Cluster Evaluation (Internal Validity)' subsection.

B. Clustering-based Knowledge Discovery Approach

To identify potential patient subgroups within the breast cancer treatment data, two distinct clustering algorithms were employed: K-Means and Agglomerative Hierarchical Clustering. K-Means, a widely used partitional clustering technique, was chosen for its efficiency and ability to handle relatively large datasets, making it suitable for the exploratory nature of this study [52]. Hierarchical Clustering, on the other hand, was utilized to explore the inherent hierarchical structure of the data and to provide a different perspective on potential patient groupings [53].

K-Means Clustering Implementation: To determine the optimal number of clusters (k) for the K-Means algorithm, three common internal validation methods were employed: the Elbow method [54], Silhouette

Analysis [55], and Calinski-Harabasz Index [56]. The performance of the clustering was evaluated across a range of k values from 2 to 10.

The Elbow method (Fig. 25) visually plots the Within-Cluster Sum of Squares (WCSS) against the number of clusters, aiming to identify a point where the rate of decrease in WCSS significantly diminishes, resembling an "elbow." Quantitative analysis of the steepest drops in WCSS indicated the most significant decreases in cluster heterogeneity. The largest decrease in WCSS was observed from K=2 to K=3 (Drop = 203.51), followed by a notable drop from K=3 to K=4 (Drop = 147.12), and then from K=6 to K=7 (Drop = 111.80).

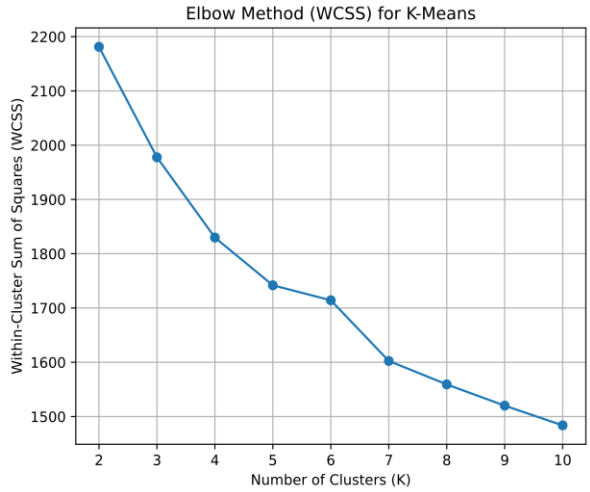

Fig. 25: Elbow method (wcss) for K-Means.

Silhouette Analysis (Fig. 26) calculates the average silhouette score, which measures how similar an object is to its own cluster compared to other clusters. Higher scores indicate better-defined and more separated clusters. The analysis revealed that the highest average silhouette score was consistently achieved at K=2 (0.1475). Other notable scores included K=5 (0.1227) and K=4 (0.1209).

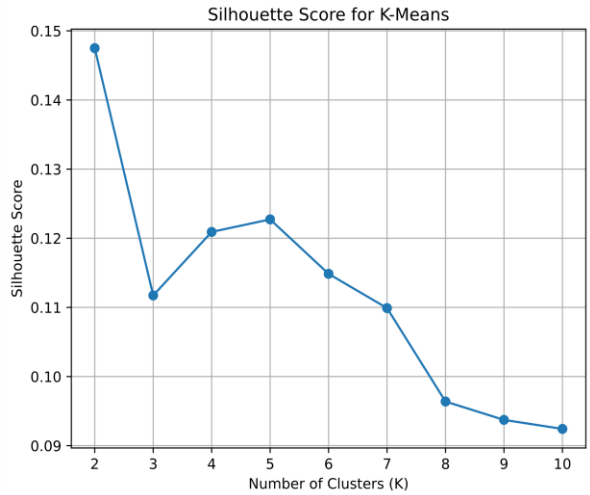

Fig. 26: Silhouette score for K-Means.

The Calinski-Harabasz Index (Fig. 27) quantifies the ratio of between-cluster dispersion to within-cluster dispersion, with higher values typically indicating more dense and well-separated clusters. The results showed that the highest Calinski-Harabasz Index was also consistently achieved at K=2 (27.6996). Other high scores were observed at K=3 (24.5557) and K=4 (22.4737).

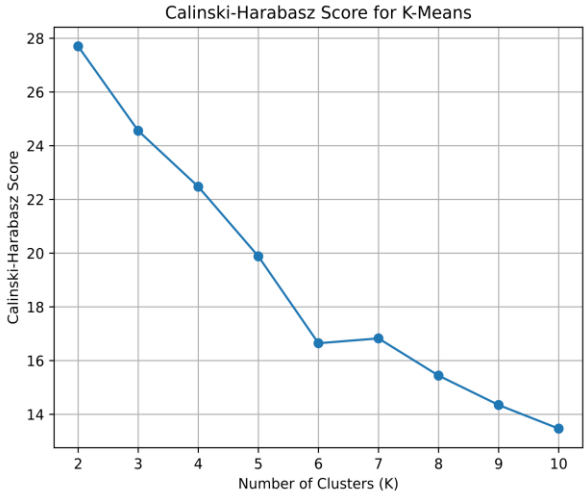

Fig. 27: Calinski Harabasz score for K-Means.

Considering the combined evidence from all three internal validation metrics: the highest Silhouette Score (0.1475) and the highest Calinski-Harabasz Index (27.6996) are both consistently observed at K=2. Furthermore, the Elbow Method's steepest WCSS drop from K=2 to K=3 suggests a strong partitioning at K=2 or K=3. The overall consensus across these robust metrics indicates that K=2 provides the most optimal balance between cluster cohesion and separation for this dataset, supporting a parsimonious and clinically interpretable solution.

The K-Means algorithm was implemented using the K-Means class from the scikit-learn library with the number of clusters set to **2** (n_clusters=2). The n_init parameter was set to 'auto', which intelligently determines the number of initializations to perform, and a random_state of 42 was used to ensure reproducibility of the clustering results. The K-Means algorithm for K=2 resulted in an average Silhouette Score of 0.1475 and a Calinski-Harabasz Index of 27.6996.

Hierarchical Clustering Implementation: Agglomerative Hierarchical Clustering, a bottom-up approach that iteratively merges data points into clusters based on their similarity [57], was also employed to identify potential patient subgroups and to compare the results with the K-Means algorithm.

The algorithm was implemented using the AgglomerativeClustering class from the scikit-learn library [58]. The preprocessed and scaled data (X_processed), as described in Data Preprocessing

Subsection, was used as input for the hierarchical clustering algorithm.

To determine the optimal number of clusters (k) for Hierarchical Clustering, similar internal validation methods were applied across a range of k values (2 to 10), complementing the visual interpretation of the dendrogram (Fig. 28).

The Hierarchical Elbow Method (Fig. 29) plots the WCSS against k. Quantitative analysis of the steepest drops in WCSS indicated the most significant decreases in cluster heterogeneity. The analysis revealed that the largest decrease in WCSS was observed from K=2 to K=3 (Drop = 209.83), followed by K=3 to K=4 (Drop = 133.75), and then from K=4 to K=5 (Drop = 119.43).

Hierarchical Silhouette Analysis (Fig. 30) calculates the average silhouette score. The analysis indicated that the highest average silhouette score was achieved at K=2 (0.1364). Other notable scores included K=5 (0.1250) and K=4 (0.1099).

The Hierarchical Calinski-Harabasz Index (Fig. 31) quantifies the ratio of between-cluster dispersion to within-cluster dispersion. The results showed that the highest Calinski-Harabasz Index was achieved at K=3 (20.5170). Other high scores were observed at K=2 (20.4309) and K=4 (18.7656).

The dendrogram (Fig. 28) visually displays the hierarchical merging of patient data points based on their feature similarity. The height of the vertical branches indicates the distance at which clusters were merged, providing insights into the structure of the underlying groupings. A quantitative analysis of the "Largest Jump" in merge distances from the dendrogram (which can be computed from the linked matrix, see Fig. 28) further supports specific K values. The largest jump in merge distance was observed at a specific merge point (Jump Value: 4.13 at merge index 181), which suggests an Optimal K = 2. Other significant jumps included a Jump Value of 2.68 (at merge index 178) suggesting Optimal K = 5, and a Jump Value of 1.97 (at merge index 182) suggesting Optimal K = 1.

Considering the combined evidence from all hierarchical internal validation metrics and the dendrogram analysis: The highest Silhouette Score (0.1364) is at K=2, and while the highest Calinski-Harabasz Index (20.5170) is at K=3 (though very close to K=2), the Elbow Method shows its steepest drop from K=2 to K=3, and the Largest Jump Analysis for the dendrogram also primarily suggests K=2. The linkage criterion for Hierarchical Clustering was set to 'ward', which aims to minimize the variance within each cluster being merged.

The resulting dendrogram (Fig. 28) displays the hierarchical merging of patient data points. Individual patient labels were omitted from the dendrogram for visual clarity, with the x-axis representing the data points. The color_threshold parameter was set to 6 to visually distinguish clusters at a specific level of dissimilarity. For K=2, the hierarchical clustering yielded an average Silhouette Score of 0.1364 and a Calinski-Harabasz Index of 20.4309.

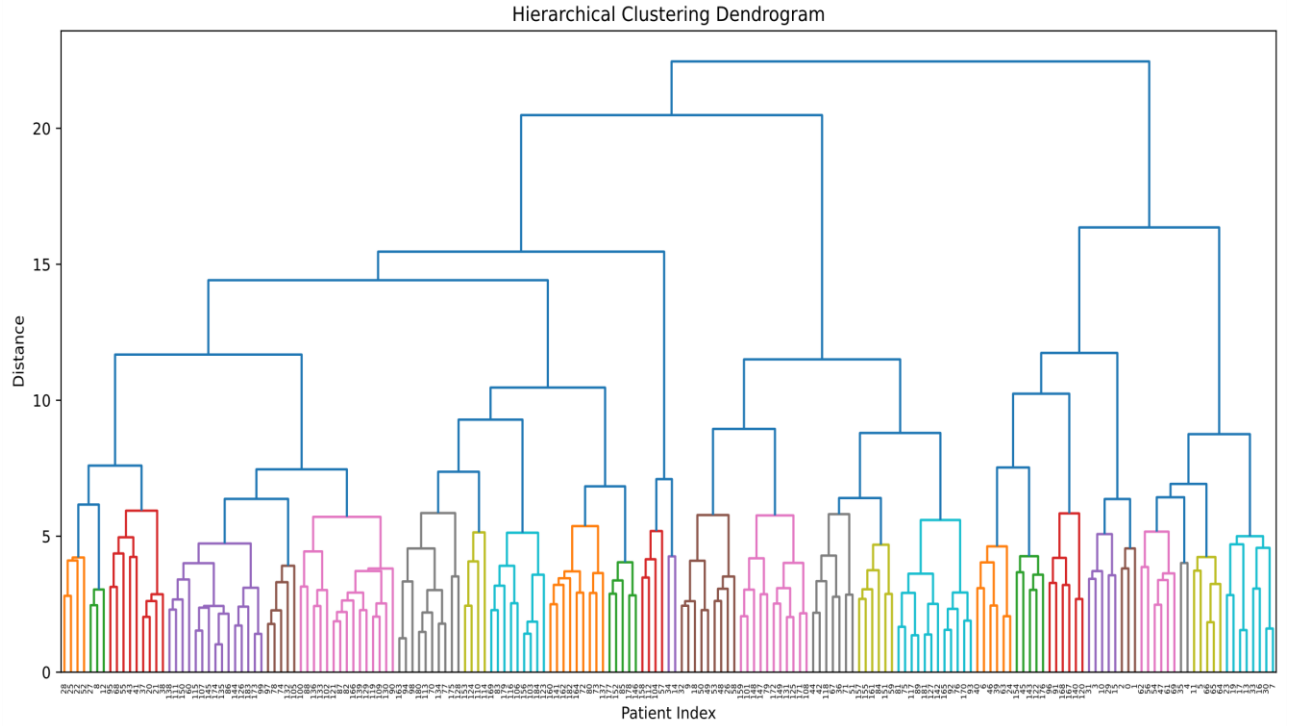

Fig. 27: Generated dendrogram using the linkage and dendrogram functions.

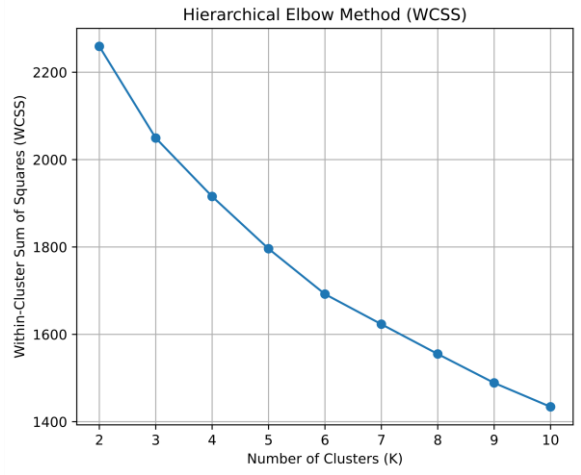

Fig. 29: Hierachical Elbow Method (WCSS).

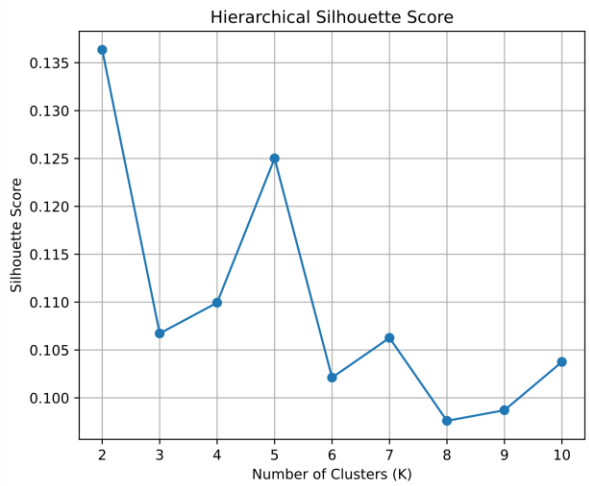

Fig. 30: Hierachical silhouette score.

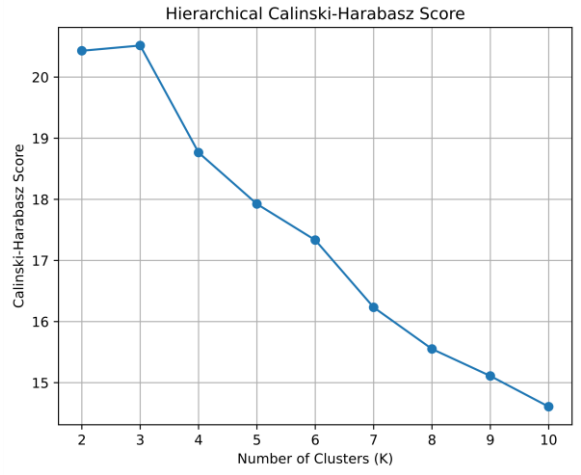

Fig. 31: Hierachical Calinski-Harabasz index.

C. Cluster Evaluation (Internal Validity)

The internal validity of the clusters obtained from both the K-Means and Hierarchical Clustering algorithms was assessed using the Silhouette Score [55] and Calinski-Harabasz Index [56]. The Silhouette Score measures how well each data point fits within its assigned cluster compared to other clusters, with values ranging from -1 to 1; higher scores indicate better-defined and more separated clusters. The Calinski-Harabasz Index quantifies the ratio of between-cluster dispersion to within-cluster dispersion, with higher values typically indicating better clustering.

For the K-Means clustering with two clusters (k=2), the average Silhouette Score obtained was 0.1475, and the Calinski-Harabasz Index was 27.6996. This suggests that the K-Means clusters, while not perfectly distinct, exhibit a reasonable structure and a moderate degree of separation, particularly in the context of complex clinical data.

The Hierarchical Clustering, also evaluated with two clusters (k=2) extracted from the dendrogram using Ward's linkage, yielded an average Silhouette Score of 0.1364 and a Calinski-Harabasz Index of 20.4309. Similar to the K-Means result, these scores indicate clusters that, while showing some internal structure, still exhibit a degree of overlap and moderate separation. Notably, the K-Means clustering demonstrated slightly higher Silhouette and Calinski-Harabasz scores, indicating marginally better internal consistency, compactness, and separation in its two-cluster solution compared to Hierarchical Clustering at k=2 for this dataset.

These internal validity scores, while confirming that both algorithms identified potential groupings within the patient data, also strongly suggest that the resulting clusters exhibit a level of internal overlap and may not be distinctly separated. This is a common and often expected characteristic when clustering complex, high-dimensional real-world clinical datasets. The inherent heterogeneity of patient populations, the continuous nature of many clinical features, and the nuanced, non-discrete boundaries often found in biological phenomena mean that subgroups rarely possess perfectly crisp or geometrically isolated boundaries that maximize mathematical separation. Consequently, Silhouette Scores, which primarily measure compactness and separation, can often appear lower in such contexts.

It is crucial to interpret these scores critically: while a lower Silhouette score mathematically indicates some overlap, it does not necessarily negate the clinical utility or inherent validity of the identified subgroups. Instead, the significance of these clusters for 'knowledge discovery' primarily stems from the statistically significant differences in feature profiles observed across the groups (as validated by ANOVA for numerical features and Chi-squared tests for categorical features in the Results section). Furthermore, the identification of these potential subgroups, despite mathematical overlap, aligns with the understanding and observations of the oncology specialists who collaborated on the dataset's collection and interpretation, thereby

enhancing their perceived clinical relevance and practical utility. These robust and statistically significant differences, supported by clinical expert insight, reveal meaningful patterns of patient characteristics and treatment responses that can inform clinical hypotheses and guide future research.

Results

This section presents the outcomes of applying the K-Means and Hierarchical Clustering algorithms to the preprocessed breast cancer treatment data from Iranmehr Hospital. The results for each clustering method, including the determined number of clusters, their sizes, and characteristic profiles based on the analyzed features, are detailed in the following subsections. Additionally, the internal validity of the obtained clusterings, as assessed by the Silhouette Score, is summarized.

A. Clustering-based Knowledge Discovery in Breast Cancer Treatment Data from Iranmehr Hospital

As described in the Methodology section, both K-Means and Hierarchical Clustering algorithms were implemented to identify potential patient subgroups. Based on the rigorous evaluation using Elbow Method, Silhouette Analysis, Calinski-Harabasz Index, and for Hierarchical Clustering, the Largest Jump Analysis from the dendrogram, the optimal number of clusters was consistently identified as two (k=2) for both methods. This choice aims to provide a robust and clinically interpretable stratification of the patient cohort.

B. K-Means Clustering Results

The K-Means algorithm, with the number of clusters set to two, partitioned the patient population into two distinct clusters with varying sizes. The distribution of patients across the two clusters was as follows: Cluster 0 (n=62) and Cluster 1 (n=123).

An analysis of the numerical and categorical features within each cluster revealed differentiating characteristic profiles:

- **Cluster 0 (n=62):** This cluster, comprising approximately 33.5% of the cohort, is primarily characterized by patients with a lower mean number of dissected lymph nodes (Mean = 4.77 ± 5.11) and fewer chemotherapy sessions (Mean = 7.16 ± 3.70). These patients also tended to have lower ordinal N-stages (Mean N = 1.34 ± 0.99) and less extensive node dissection (Mean Node_Dissection = 1.47 ± 0.72). In terms of nominal features, this cluster is entirely Female (100.0%). Menopausal Situation is balanced (50.0% Post-menopausal vs. 50.0% Pre-menopausal). Histological Type is predominantly IDC (83.87%), with ILC accounting for 11.29%. The majority underwent Total Mastectomy (66.13%) performed by General Surgeons (75.81%). ER-Positive status is high (75.81%), while HER2-Negative (66.13%) and Trastuzumab-Negative (66.13%) are more common. High percentages are observed for Conventional Radiation Dose (64.52%) and Negative Radiation Boost Dose (67.74%). Tamoxifen is the most common Hormone Therapy (37.10%).
  - Statistically Significant Differentiating Features ($p < 0.05$ from ANOVA/Chi-squared): Age (p=0.0172), Dissected_Nodes (p=0.0000), Chemotherapy_Session (p=0.0000), N (Nodal involvement, p=0.0000), Node_Dissection (p=0.0000), Chemotherapy_Regimen (p=0.0000), Treatment_Schedule (p=0.0082), Radiation_dose (p=0.0016), and Radiation_Boost_Dose (p=0.0069).
- **Cluster 1 (n=123):** This larger cluster, encompassing approximately 66.5% of the cohort, presents a contrasting profile with a higher mean number of dissected lymph nodes (Mean = 9.91 ± 5.44) and more intensive chemotherapy sessions (Mean = 15.02 ± 2.63). Correspondingly, these patients showed higher ordinal N-stages (Mean N = 2.13 ± 1.08) and more extensive node dissection (Mean Node_Dissection = 1.93 ± 0.25, closer to ALND). Regarding nominal features, this cluster is also predominantly Female (96.75%), with a higher proportion of Pre-menopausal patients (65.04%) compared to Post-menopausal (34.96%). Histological Type is predominantly IDC (93.50%). Total Mastectomy (65.04%) by General Surgeons (73.17%) is common. ER-Positive (65.04%) and PR-Positive (60.16%) statuses are prevalent, while HER2-Negative (71.54%) and Trastuzumab-Negative (71.54%) are common. Higher percentages are observed for Conventional Radiation Dose (86.99%) and Positive Radiation Boost Dose (54.47%). Tamoxifen is the most common Hormone Therapy (39.84%).
  - **Statistically Significant Differentiating Features ($p < 0.05$ from ANOVA/Chi-squared):** Age (p=0.0172), Dissected_Nodes (p=0.0000), Chemotherapy_Session (p=0.0000), N (Nodal involvement, p=0.0000), Node_Dissection (p=0.0000), Chemotherapy_Regimen (p=0.0000), Treatment_Schedule (p=0.0082), Radiation_dose (p=0.0016), and Radiation_Boost_Dose (p=0.0069).

This analysis reveals that K-Means clustering primarily differentiates patients based on their age, the extent of nodal involvement (Dissected_Nodes, N, Node_Dissection), and the intensity of systemic treatments (Chemotherapy_Session, Chemotherapy_Regimen, Treatment_Schedule, Radiation_dose, Radiation_Boost_Dose).

C. Hierarchical Clustering Results

Applying Agglomerative Hierarchical Clustering with Ward's linkage and extracting two clusters (k=2) from the dendrogram (Fig. 28) resulted in the following cluster sizes: Cluster 0 (n=139) and Cluster 1 (n=46).

Analyzing the mean of the numerical and re-mapped ordinal features for these hierarchical clusters, and the frequency distributions of the nominal features, provided insights into their characteristics:

- **Cluster 0 (n=139):** This larger cluster, comprising approximately 75.1% of the cohort, is characterized by a mean age of 47.05 ± 10.51 years. These patients tend to have a higher mean number of dissected nodes (9.37 ± 5.56), receive more chemotherapy sessions (mean 13.81 ± 3.90), and show higher N-stage (mean N = 2.01 ± 1.06) and Node_Dissection (mean 1.95 ± 0.22, closer to ALND). In terms of nominal features, this cluster is predominantly Female (97.12%). It has a higher proportion of Pre-menopausal patients (69.78%) compared to Post-menopausal (30.22%). Histological Type is overwhelmingly IDC (93.53%). The majority underwent Total Mastectomy (64.03%) by General Surgeons (76.26%). ER-Positive (63.31%), PR-Positive (57.55%), HER2-Negative (71.94%), and Trastuzumab-Negative (71.94%) statuses are common. Conventional Radiation Dose (84.17%) is prevalent, and Positive Radiation Boost Dose (52.52%) is slightly more common than Negative. Tamoxifen is the most common Hormone Therapy (39.57%). GnRH Analog use is largely Negative (82.01%).
  - **Statistically Significant Differentiating Features (p < 0.05):** Age (p=0.0000), Dissected_Nodes (p=0.0000), Chemotherapy_Session (p=0.0000), N (Nodal involvement, p=0.0013), Node_Dissection (p=0.0000), Chemotherapy_Regimen (p=0.0000), Menopausal_Situation (p=0.0000), ER (p=0.0112), Treatment_Schedule (p=0.0017), Radiation_dose (p=0.0047), Radiation_Boost_Dose (p=0.0151), Hormonetherapy (p=0.0055**).**
- **Cluster 1 (n=46):** This smaller cluster, approximately 24.9% of the cohort, presents a contrasting profile with patients being noticeably older on average (Mean Age = 55.54 ± 10.43 years). They tend to have a lower mean number of dissected lymph nodes (4.61 ± 5.26) and significantly fewer chemotherapy sessions (Mean = 8.09 ± 4.69). Their N-stage (mean N = 1.41 ± 1.17) and Node_Dissection (mean 1.26 ± 0.74, closer to SLND/No Dissection) also reflect less extensive nodal involvement and dissection. Regarding nominal features, this cluster is also entirely Female (100.0%). It shows a higher proportion of Post-menopausal patients (69.57%) compared to Pre-menopausal (30.43%). Histological Type is predominantly IDC (80.43%), but with a notably higher proportion of ILC (13.04%) than Cluster 0. The majority underwent Total Mastectomy (69.57%) by General Surgeons (67.39%). ER-Positive (84.78%) and PR-Positive (67.39%) statuses are highly prevalent. HER2-Negative (63.04%) is still common, but a higher proportion are HER2-Positive (36.96%) and Trastuzumab-Positive (36.96%). Conventional Radiation Dose (65.22%) is common, but a higher percentage received No RT (34.78%) compared to Cluster 0, and Negative Radiation Boost Dose (69.57%) is more common. Tamoxifen (36.96%) and Letrozole (32.61%) are common Hormone Therapies. GnRH Analog use is largely Negative (86.96%).
  - **Statistically Significant Differentiating Features (p < 0.05):** Age (p=0.0000), Dissected_Nodes (p=0.0000), Chemotherapy_Session (p=0.0000), N (Nodal involvement, p=0.0013), Node_Dissection (p=0.0000), Chemotherapy_Regimen (p=0.0000), Menopausal_Situation (p=0.0000), ER (p=0.0112), Treatment_Schedule (p=0.0017), Radiation_dose (p=0.0047), Radiation_Boost_Dose (p=0.0151), Hormonetherapy (p=0.0055).

This analysis reveals that Hierarchical Clustering's differentiation primarily revolves around Age, Menopausal_Situation, nodal involvement (Dissected_Nodes, N, Node_Dissection), and the intensity/type of systemic treatments (Chemotherapy_Session, Chemotherapy_Regimen, Treatment_Schedule, Radiation_dose, Radiation_Boost_Dose, Hormonetherapy, ER).

D. Comparison of K-Means and Hierarchical Clustering Results

Comparing the outcomes of the K-Means and Hierarchical Clustering algorithms reveals both areas of substantial agreement and notable differences in the identified patient subgroups. Both methods, when constrained to produce two clusters, succeeded in partitioning the patient cohort based on distinct clinical characteristics.

To provide a detailed quantitative assessment of the patient archetypes identified by both methods, Table 5 presents a side-by-side analysis of key demographic, pathological, and treatment features for the two clusters derived from each algorithm.

This quantitative comparison, presented in Table 5, reveals that while the overall patient assignment agreement between the two distinct clustering solutions was moderate (as quantified by an Adjusted Rand Index (ARI) of 0.4697), both algorithms consistently identified groups sharing fundamental clinical characteristics.

For instance, K-Means Cluster 0 (n=62) is

characterized by lower mean dissected lymph nodes, fewer chemotherapy sessions, and lower N-stages. Conversely, K-Means Cluster 1 (n=123) shows a contrasting profile with a higher mean number of dissected lymph nodes, more intensive chemotherapy sessions, and higher N-stages. Notably, both K-Means clusters are predominantly female, with varying proportions of menopausal status and histological types, but share high rates of ER-positive status.

Similarly, the Hierarchical Clustering algorithm also partitioned the cohort into two clusters: Cluster 0 (n=139) and Cluster 1 (n=46). Hierarchical Cluster 0 largely aligns with the higher intensity/severity profile, showing higher mean dissected nodes and chemotherapy sessions. Hierarchical Cluster 1 presents a profile with a lower mean number of dissected lymph nodes and fewer chemotherapy sessions.

The statistical tests confirm several key differentiating features. For K-Means, Age, Dissected_Nodes, Chemotherapy_Session, N, Node_Dissection, Chemotherapy_Regimen, Treatment_Schedule, Radiation_dose, and Radiation_Boost_Dose were all found to be statistically significant differentiators ($P < 0.05$). For Hierarchical Clustering, in addition to all the features significant for K-Means (except for Age, which showed a stronger P-value, and N, which showed a slightly weaker P-value, but still significant), Menopausal_Situation, ER Status, and Hormone Therapy also emerged as statistically significant differentiating features ($P < 0.05$).

Table 5: Quantitative comparison of identified patient subgroups by K-Means and hierarchical clustering

Feature	K-Means Cluster 0 (n=62)	K-Means Cluster 1 (n=123)	Hierarchical Cluster 0 (n=139)	Hierarchical Cluster 1 (n=46)
Mean Age	51.89	47.79	47.05	55.54
Mean Dissected Nodes	4.77 (Lower)	9.91 (Higher)	9.37 (Higher)	4.61 (Lower)
Mean N Stage	1.34 (Lower)	2.13 (Higher)	2.01 (Higher)	1.41 (Lower)
Mean Node Dissection	1.47 (Less Extensive)	1.93 (More Extensive, closer to ALND)	1.95 (More Extensive, closer to ALND)	1.26 (Less Extensive, closer to SLND)
Mean Chemotherapy Sessions	7.16 (Fewer)	15.02 (More Intensive)	13.81 (More Intensive)	8.09 (Fewer)
Mean Chemotherapy Regimen	1.65 (Lower Ordinal)	2.90 (Higher Ordinal)	2.73 (Higher Ordinal)	1.74 (Lower Ordinal)
Treatment Schedule	40.32% Sx>ChT>RT>HoT; varied schedules	51.22% Sx>ChT>RT>HoT; varied schedules	48.20% Sx>ChT>RT>HoT; varied schedules	45.65% Sx>ChT>RT>HoT; varied schedules
Radiation Dose	64.52% Conventional RT	86.99% Conventional RT	84.17% Conventional RT	34.78% No RT / 65.22% Conventional RT
Radiation Boost Dose	67.74% Negative	54.47% Positive	52.52% Positive	69.57% Negative
Post-Menopausal	50.00%	34.96%	30.22%	69.57%
ER-Positive	75.81%	65.04%	63.31%	84.78%
HER2-Positive	33.87%	28.46%	28.06%	36.96%
Ki67-Negative	43.55%	34.96%	38.13%	36.96%
Hormone Therapy	37.10% Tamoxifen (Most Common)	39.84% Tamoxifen (Most Common)	39.57% Tamoxifen (Most Common)	36.96% Tamoxifen / 32.61% Letrozole

However, the quantitative analysis also highlights notable discrepancies in specific feature distributions between the conceptually similar clusters, underscoring the influence of the algorithmic approach on patient stratification in complex data. For example, while K-Means differentiation for nominal features primarily revolved around Treatment_Schedule, Radiation_dose, and Radiation_Boost_Dose, Hierarchical Clustering showed a broader differentiation across Menopausal_Situation, ER, and Hormonetherapy as well. This suggests that Hierarchical Clustering might be more sensitive to demographic and biomarker-related nuances in defining its groups.

Furthermore, despite similar overall profiles, the specific compositions and sizes of the most comparable clusters can differ. For instance, the older age and predominantly post-menopausal status of Hierarchical Cluster 1 (Mean Age = 55.54 years, 69.57% Post-menopausal) create a more distinct demographic profile than what is primarily driven by age in K-Means. This further illustrates the differing sensitivities of the two algorithms to various feature combinations, leading to unique patient groupings not perfectly mirrored across methods.

As discussed in the 'Cluster Evaluation (Internal Validity)' subsection, the internal validity of the two-cluster solutions was assessed using key metrics including the Silhouette Score and the Calinski-Harabasz Index. The relatively low values obtained for both methods (e.g., K-Means Silhouette Score of 0.1475 and Hierarchical Silhouette Score of 0.1364) are indicative of inherent overlap and non-distinct separation within these patient groupings. This underscores that while both algorithms identified potential groupings, the identified patterns are not perfectly distinct, and interpretations should be cautious. The hierarchical structure revealed by the dendrogram offers a different perspective compared to the discrete clusters produced by K-Means. While we chose to cut the dendrogram at a level yielding two clusters for comparison, the visual representation suggests that other numbers of clusters might also be meaningful and could capture different aspects of the data's underlying structure.

Discussion

This study successfully employed K-Means and Agglomerative Hierarchical Clustering algorithms to identify potential patient subgroups within the breast cancer treatment dataset from Iranmehr Hospital. Based on rigorous internal validation metrics (Elbow Method, Silhouette Analysis, Calinski-Harabasz Index, and Largest Jump Analysis for Hierarchical Clustering), the optimal number of clusters was consistently determined to be two (k=2) for both methods, providing a robust and clinically interpretable stratification of the patient cohort.

The K-Means algorithm partitioned the patient population into two distinct clusters: Cluster 0 (n=62, 33.5% of cohort) and Cluster 1 (n=123, 66.5% of cohort). Cluster 0 is predominantly characterized by patients with a lower mean number of dissected lymph nodes (Mean = 4.77 ± 5.11), fewer chemotherapy sessions (Mean = 7.16 ± 3.70), and generally lower ordinal N-stages (Mean N = 1.34 ± 0.99), suggesting a less aggressive disease profile or less intensive treatment approach. Conversely, Cluster 1 presented a contrasting profile, indicative of more advanced disease or intensive treatment, with a higher mean number of dissected lymph nodes (Mean = 9.91 ± 5.44), more intensive chemotherapy sessions (Mean = 15.02 ± 2.63), and higher ordinal N-stages (Mean N = 2.13 ± 1.08). Both K-Means clusters showed high rates of ER-positive status, with HER2-negative and Trastuzumab-negative statuses being more common. Statistically significant differentiators for K-Means clusters ($P < 0.05$) included Age, Dissected_Nodes, Chemotherapy_Session, N, Node_Dissection, Chemotherapy_Regimen, Treatment_Schedule, Radiation_dose, and Radiation_Boost_Dose.

Similarly, the Hierarchical Clustering algorithm also partitioned the cohort into two clusters: Cluster 0 (n=139, 75.1% of cohort) and Cluster 1 (n=46, 24.9% of cohort). Hierarchical Cluster 0 largely aligned with the K-Means "higher intensity/severity" profile, exhibiting higher mean dissected nodes (9.37 ± 5.56) and more chemotherapy sessions (mean 13.81 ± 3.90), along with higher N-stages. In contrast, Hierarchical Cluster 1 presented a distinct profile, characterized by patients who were noticeably older on average (Mean Age = 55.54 ± 10.43 years) and had a lower mean number of dissected lymph nodes (4.61 ± 5.26) and significantly fewer chemotherapy sessions (Mean = 8.09 ± 4.69). This cluster also showed a higher proportion of Post-menopausal patients (69.57%), a higher percentage of ER-Positive (84.78%) and HER2-Positive (36.96%) statuses compared to Hierarchical Cluster 0, and a notable percentage receiving no radiation therapy. Statistically significant differentiators for Hierarchical clusters ($P < 0.05$) encompassed all those for K-Means, with the crucial addition of Menopausal_Situation, ER Status, and Hormonetherapy, highlighting a broader set of discriminating factors.

A quantitative comparison of the two distinct clustering solutions revealed a moderate overall patient assignment agreement, as quantified by an Adjusted Rand Index (ARI) of 0.4697. This score, significantly above random chance, indicates that while the algorithms identified a shared fundamental partitioning of the patient population, they also exhibited differences in specific data point assignments or cluster boundary definitions. Both algorithms consistently highlighted the importance of nodal involvement (Dissected_Nodes, N, Node_Dissection) and the intensity/type of systemic treatments (Chemotherapy_Session, Chemotherapy_Regimen, Treatment_Schedule, Radiation_dose, Radiation_Boost_Dose) as key differentiating factors. However, the Hierarchical Clustering method demonstrated a more pronounced ability to differentiate based on demographic and biomarker features such as Menopausal_Situation, ER Status, and Hormone Therapy, which were less statistically significant in the K-Means solution. This suggests that Hierarchical Clustering may be more sensitive to these nuanced patient characteristics, leading to a cluster (Hierarchical Cluster 1) that is more distinctly defined by age and menopausal status. The differing cluster sizes between the two methods further reflect these algorithmic sensitivities.

As discussed in the 'Cluster Evaluation (Internal Validity)' subsection, the internal validity scores for the two-cluster solutions (K-Means Silhouette Score = 0.1475; Hierarchical Silhouette Score = 0.1364) were relatively low. These values, while confirming the identification of potential groupings, suggest a degree of

inherent overlap and non-distinct separation within the clusters. This is a common and often expected characteristic when clustering complex, high-dimensional real-world clinical datasets, where patient heterogeneity and the continuous nature of clinical variables rarely result in perfectly isolated subgroups. Despite this, the observed robust and statistically significant differences in feature profiles across the identified clusters, validated by ANOVA and Chi-squared tests, combined with insights from collaborating oncology specialists, suggest the clinical relevance and practical utility of these groupings for knowledge discovery.

A. Clinical Relevance

The identification of two distinct patient subgroups within the breast cancer cohort at Iranmehr Hospital holds significant implications for understanding disease heterogeneity and advancing personalized treatment strategies. The characterization of these clusters, grounded in statistically significant differentiating features as detailed in Table 5, provides valuable insights for generating clinically relevant hypotheses.

Cluster 0 (K-Means and Hierarchical alignment in general concept): This cluster, broadly defined by higher nodal involvement and intensive systemic treatments, likely represents patients with a more aggressive disease presentation or those requiring more comprehensive therapeutic approaches. For these patients, the clusters suggest a need for vigilant follow-up, potentially more aggressive adjuvant therapies, or consideration for novel treatment regimens to mitigate the risk of recurrence.

Cluster 1 (K-Means and Hierarchical alignment in general concept, with Hierarchical showing a unique subset): This cluster, generally characterized by lower nodal involvement and less intensive systemic treatments, may represent patients with a more favorable prognosis or those for whom de-escalation of therapy could be considered, thereby minimizing unnecessary exposure to treatment-related toxicities and improving quality of life.

Crucially, Hierarchical Cluster 1 (the older, predominantly post-menopausal subgroup with lower disease intensity but distinct biomarker profiles) highlights a particularly relevant patient archetype. Understanding the specific factors influencing treatment decisions and outcomes in this older cohort, and the implications of their hormonal and HER2 statuses, could inform more tailored management guidelines for geriatric breast cancer patients. This subgroup's unique profile suggests a potential for distinct therapeutic considerations that go beyond general age-based guidelines.

The identified stratifications provide a data-driven foundation for generating specific clinical hypotheses. For example, correlating these clusters with long-term patient outcomes (e.g., disease-free survival, overall survival) in future prospective studies is essential. Such validation could establish the prognostic or predictive value of these subgroups, ultimately guiding personalized treatment decisions and patient counseling based on a more granular understanding of their clinical and biological profiles.

B. Strengths and Limitations

A key strength of this study lies in the development of a dedicated dataset for breast cancer patients at Iranmehr Hospital, collected with the direct collaboration and expertise of two oncology specialists. This collaboration ensures the clinical relevance and accuracy of the included features, reflecting real-world data and treatment practices within this specific medical center. The subsequent application of K-Means and Hierarchical Clustering to this local dataset allowed for the identification of patient subgroups specific to this population, potentially capturing nuances missed in broader, more heterogeneous datasets. Furthermore, the statistical validation of feature differences across the identified clusters using ANOVA and Chi-squared tests adds rigorous scientific support to the interpretation of these subgroups. The consistent identification of two optimal clusters across multiple internal validation metrics (Elbow method, Silhouette analysis, Calinski-Harabasz index, and Largest Jump analysis) and the moderate Adjusted Rand Index (ARI) of 0.4697 between the two distinct clustering solutions further strengthens the robustness and interpretability of these findings, indicating a stable underlying data structure.

Despite these strengths, several limitations must be acknowledged. Firstly, the study is based on a retrospective dataset from a single institution, which inherently limits the generalizability of the findings to other populations or healthcare settings with different treatment protocols and patient demographics. While the local specificity can be a strength for understanding patterns within Iranmehr Hospital, it necessitates caution when extrapolating these findings.

Secondly, while k=2 was consistently identified as the optimal number of clusters based on internal validity metrics, the choice of 'k' inherently involves a degree of subjectivity in unsupervised learning. Exploring alternative 'k' values or employing different clustering algorithms might reveal alternative or more clinically relevant patient segmentations, although for this study, the two-cluster solution offers a parsimonious and interpretable stratification.

Thirdly, the relatively low Silhouette Scores for both K-Means (0.1475) and Hierarchical Clustering (0.1364) suggest a degree of overlap and heterogeneity within the identified clusters, indicating that the boundaries

between these subgroups may not be sharply defined. This is often attributable to the inherent complexity and continuous nature of clinical variables in real-world patient data, where subgroups rarely form perfectly discrete or geometrically isolated clusters.

Fourthly, the observed imbalance in the distribution of certain treatment combinations reflects the real-world clinical practices at this institution. While this is an inherent characteristic of the data, future analyses could explore the impact of this imbalance on the clustering results and consider using techniques specifically designed for imbalanced datasets, if deemed necessary. Fifthly, consistent with the methodology, no explicit feature selection or dimensionality reduction techniques were applied at this stage to specifically optimize cluster separability. While this approach provided an unbiased view of patient characteristics, it is possible that such techniques could enhance cluster compactness and separation.

Finally, while we have discussed potential clinical relevance, this study is primarily descriptive. Further research correlating these clusters with long-term clinical outcomes (e.g., survival, recurrence) is needed to validate their prognostic or predictive value and to establish their utility in guiding treatment decisions. The cross-sectional nature of the data also limits our ability to infer temporal relationships or the evolution of treatment strategies over time.

C. Future Work

Several promising avenues for future research emerge from this study, building upon the identified two-cluster patient stratifications. Firstly, to address the limitation of single-center data, it would be invaluable to validate the identified patient subgroups in larger, potentially multi-center datasets. To facilitate this while respecting data privacy, future research could explore the application of Federated Learning techniques. This approach would allow for collaborative analysis across institutions without the need to centralize sensitive patient information.

Secondly, future work should explore the application of a wider range of clustering algorithms beyond K-Means and Hierarchical Clustering, including density-based (e.g., DBSCAN), distribution-based (e.g., Gaussian Mixture Models), and other partitioning methods. Comparing the results of these algorithms and evaluating their performance using appropriate internal and external validation metrics could lead to a more robust and clinically meaningful patient segmentation.

Thirdly, given the potential clinical relevance suggested by the characteristics of the identified two clusters, a critical next step is to correlate these clusters with long-term clinical outcomes such as disease-free survival, overall survival, and 5-year survival rates. This would provide strong evidence for the prognostic value of these subgroups and their potential utility in guiding personalized treatment decisions and patient counseling.

Fourthly, future studies could investigate the integration of other relevant data sources, such as detailed molecular and genomic information, imaging data, and patient-reported outcomes, to further refine the identified clusters and gain a more comprehensive understanding of the underlying biological and clinical characteristics of these patient subgroups.

Fifthly, to address the challenge of feature dimensionality and potentially enhance cluster separability, future analyses will rigorously investigate various feature selection and dimensionality reduction techniques. Methods such as Principal Component Analysis (PCA) for linear dimensionality reduction, Recursive Feature Elimination (RFE), or filter methods based on statistical tests (e.g., correlation-based feature selection) will be explored to identify the most informative and discriminative features for patient stratification.

This systematic approach is anticipated to reduce noise, improve computational efficiency, and potentially yield more compact and well-separated clusters, thereby further refining the interpretability of identified patient subgroups.

Finally, future research could explore the implications of the inherent treatment imbalance in the dataset on the identified clusters and investigate whether alternative clustering approaches or specific techniques for imbalanced data analysis could provide further insights. Sensitivity analyses on the clustering parameters would also be beneficial to assess the robustness of the identified clusters.

Conclusion

This study effectively utilized K-Means and Agglomerative Hierarchical Clustering to identify two distinct potential patient subgroups of breast cancer patients within the Iranmehr Hospital dataset, revealing significant stratifications based on their clinical and treatment characteristics.

Through rigorous internal validation, k=2 was consistently identified as the optimal number of clusters for both methodologies. The identified clusters exhibited statistically significant differences across key features such as age, chemotherapy intensity, nodal involvement, menopausal status, and ER expression, suggesting underlying heterogeneity in the patient population and treatment approaches. The moderate agreement between the two clustering methods, quantified by an Adjusted Rand Index of 0.4697, indicates a shared foundational partitioning while also highlighting areas of

distinct algorithmic sensitivity.

While the findings offer valuable initial insights into patient stratification within this specific clinical context, the study's limitations, including its single-center, retrospective nature, and the inherent complexity of clustering real-world clinical data, necessitate further investigation.

Nevertheless, the identified two clusters provide a data-driven foundation for future research aimed at understanding their clinical relevance, particularly in terms of long-term treatment outcomes and potential for personalized medicine strategies.

Future work should focus on validating these findings in larger, potentially multi-center cohorts, and exploring the utility of alternative clustering algorithms and feature selection techniques, including dimensionality reduction methods. To facilitate analysis across multiple institutions while preserving data privacy, future research could also explore the application of Federated Learning techniques. Importantly, future studies should correlate the identified patient subgroups with crucial clinical endpoints such as recurrence rates and survival outcomes. Integrating multi-omics data could further refine our understanding of these patient stratifications and pave the way for more tailored and effective breast cancer management.

Author Contributions

In the current study, the roles of each individual were as follows:

N. Mehrshad, Supervision, Conceptualization, Methodology, Investigation, Reviewing, and Editing.

R. Bakhshali: Advisor, Collecting the dataset, Investigation, Reviewing, and Editing.

A. Sebzari: Advisor, Collecting the dataset, Investigation, Reviewing, and Editing.

O. Dehghantanha: Collecting the dataset, Conceptualization, Methodology, Investigation, Writing, results.

All authors discussed the results.

Acknowledgment

We sincerely thank the respected referees for their accurate review of this paper.

Funding

This research received no external funding.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

AI	Artificial Intelligence
ALND	Axillary Lymph Node Dissection
ARI	Adjusted Rand Index
BC	Breast Cancer
BCS	Breast-Conserving Surgery
CAD	Computer-Aided Diagnosis
ChT	Chemotherapy
CNN	Convolutional Neural Networks
DL	Deep Learning
EGFR	Epidermal Growth Factor Receptor
ER	Estrogen Receptor
GnRH-Ana.	Gonadotropin-Releasing Hormone Analog
Gy	Gray
HER2	Human Epidermal Growth Factor Receptor 2
HoT	Hormone Therapy
IDC	Invasive Ductal Carcinoma
ILC	Invasive Lobular Carcinoma
KI67	Ki67 (nuclear protein marker of cellular proliferation)
KNN	K-Nearest Neighbors
ML	Machine Learning
PCA	Principal Component Analysis
PR	Progesterone Receptor
RFE	Recursive Feature Elimination
RT	Radiation
Sx	Surgery
SLND	Sentinel Lymph Node Biopsy
SVM	Support Vector Machines

TNBC	Triple-Negative Breast Cancer
TNM	TNM staging system
WCSS	Within-Cluster Sum of Squares

References

[1] F. Bray, M. Laversanne, E. Weiderpass, I. Soerjomataram, "The ever-increasing importance of cancer as a leading cause of premature death worldwide," Cancer, 127(16): 3029–3030, 2021.

[2] NCD Countdown 2030 Collaborators, "NCD Countdown 2030: Pathways to achieving Sustainable Development Goal target 3.4," Lancet, 396(10255): 918, 2020.

[3] H. Sung et al., "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA Cancer J. Clin., 71(3): 209–249, 2021.

[4] A. G. Renehan, M. Tyson, M. Egger, R. F. Heller, M. Zwahlen, "Body-mass index and incidence of cancer: A systematic review and meta-analysis of prospective observational studies," Lancet, 371(9612): 569–578, 2008.

[5] A. McTiernan et al., "Recreational physical activity and the risk of breast cancer in postmenopausal women: The Women's health initiative cohort study," JAMA, 290(10): 1331–1336, 2003.

[6] M. E. Levine et al., "Low protein intake is associated with a major reduction in IGF-1, cancer, and overall mortality in the 65 and younger but not older population," Cell Metab., 19(3): 407–417, 2014.

[7] N. Hamajima et al., "Collaborative reanalysis of individual data from 53 epidemiological studies, including 58,515 women with breast cancer and 95,067 women without the disease," Br. J. Cancer, 87(11): 1234–1245, 2002.

[8] U.S. Department of Health and Human Services, The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General. Atlanta, GA: U.S. Department of Health and Human Services, 2014.

[9] D. J. Hunter et al., "Oral contraceptive use and breast cancer: A prospective study of young women," Cancer Epidemiol. Biomarkers Prev., 19(10): 2496–2502, 2010.

[10] A. K. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognit. Lett., 31(8): 651–666, 2010.

[11] A. Ahmad, L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data," Data Knowl. Eng., 63(2): 503–527, 2007.

[12] Z. Huang, "Clustering large data sets with mixed numeric and categorical values," in Proc. The First Pacific-Asia Conference on Knowledge Discovery and Data Mining: 21–34, 1997.

[13] D. Xu, Y. Tian, "A comprehensive survey of clustering algorithms," Ann. Data Sci., 2(2): 165–193, 2015.

[14] F. Murtagh, P. Legendre, "Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion?," J. Classif., 31(3): 274–295, 2014.

[15] G. Pison, A. Struyf, P. J. Rousseeuw, "Displaying a clustering with CLUSPLOT," Comput. Stat. Data Anal., 30(4): 381–392, 1999.

[16] A. K. Dubey et al., "Analysis of k-means clustering approach on the breast cancer Wisconsin dataset," Int. J. Comput. Assist. Radiol. Surg., 11(11): 2033–2047, 2016.

[17] U. Agrawal, D. Soria, C. Wagner, J. Garibaldi, I. O. Ellis, J. M. S. Bartlett, D. Cameron, E. A. Rakha, A. R. Green, "Combining clustering and classification ensembles: A novel pipeline to identify breast cancer profiles," Artif. Intell. Med., 97: 27–37, 2019.

[18] C. Wang et al., "Breast cancer patient stratification using a molecular regularized consensus clustering method," Methods (San Diego, Calif.), 67(3): 304–312, 2014.

[19] Z. Sajjadnia et al., "Preprocessing breast cancer data to improve the data quality, diagnosis procedure, and medical care services," Cancer Inform., 19: 1176935120917955, 2020.

[20] A. Ahmadi et al., "Incidence pattern and spatial analysis of breast cancer in Iranian women: Geographical information system applications," East. Mediterr. Health J., 24(4): 345–352, 2018.

[21] S. M. Hosseini , M. Parvin , P. Shokri , M. Fadaie , B. Ghaytasi , M. Khondabi , M. Olfatifar, E. Chavoshi, "Clustering of breast cancer cases among women from kurdistan province, Iran: A population-based cross-sectional study," middle east journal of cancer, 9(1): 2018.

[22] S. Dehdar et al., "Applications of different machine learning approaches in prediction of breast cancer diagnosis delay," Front. Oncol., 13: 1103369, 2023.

[23] M. Radak et al., "Machine learning and deep learning techniques for breast cancer diagnosis and classification: a comprehensive review of medical imaging studies," J. Cancer Res. Clin. Oncol., 149(12): 10473–10491, 2023.

[24] J. Xiao et al., "The application and comparison of machine learning models for the prediction of breast cancer prognosis: Retrospective cohort study," JMIR Med. Inform., 10(2): e33440, 2022.

[25] I. Guyon, A. Elisseeff, "An introduction to variable and feature selection," J. Mach. Learn. Res., 3: 1157–1182, 2003.

[26] A. Zimek, E. Schubert, H. P. Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data," Stat. Anal. Data Min.: ASA Data Sci. J., 5(5): 363–387, 2012.

[27] D. T. Dinh, V. N. Huynh, S. Sriboonchitta, "Clustering mixed numerical and categorical data with missing values," Inf. Sci., 571: 418–442, 2021.

[28] S. Boluki, S. Zamani Dadaneh, X. Qian, E. R. Dougherty, "Optimal clustering with missing values," BMC Bioinformatics, 20: 1–10, 2019.

[29] M. Sheller et al., "Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data," Sci. Rep., 10(1): 12598, 2020.

[30] Q. Yang et al., "Federated machine learning: concept and applications," ACM Trans. Intell. Syst. Technol., 10(2): 1–19, 2019.

[31] M. Ester, H. Kriegel, J. Sander, X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proc. 2nd Int. Conf. Knowledge Discovery and Data Mining (KDD-96): 226–231, 1996.

[32] D. A. Reynolds, "Gaussian mixture models," in Encyclopedia of Biometrics, Springer, pp. 827-832, 2015.

[33] G. L. Gierach et al., "Relationship between mammographic density and breast cancer death in the breast cancer surveillance consortium," J. Natl. Cancer Inst., 104(16): 1218–1227, 2012.

[34] G. C. Wishart et al., "Screen-detected vs symptomatic breast cancer: Is improved survival due to stage migration alone?" Br. J. Cancer, 98(11): 1741–1744, 2008.

[35] S. Adams et al., " Prognostic value of tumor-infiltrating lymphocytes in triple-negative breast cancers from two phase III randomized adjuvant breast cancer trials: ECOG 2197 and ECOG 1199," J. Clin. Oncol., 32(27): 2959-2966, 2014.

[36] S. Watanabe, H. Asamura, "Lymph node dissection for lung cancer: Significance, strategy, and technique," J. Thorac. Oncol., 4(5): 652–657, 2009.

[37] M. Ferrero-Poüs et al., "Comparison of enzyme immunoassay and immunohistochemical measurements of estrogen and progesterone receptors in breast cancer patients," Appl. Immunohistochem. Mol. Morphol., 9(3): 267–275, 2001.

[38] K. C. Chu et al., "Frequency distributions of breast cancer characteristics classified by estrogen receptor and progesterone receptor status for eight racial/ethnic groups," Cancer, 92(1): 37–45, 2001.

[39] A. S. Knoop et al., "Value of epidermal growth factor receptor, HER2, p53, and steroid receptors in predicting the efficacy of tamoxifen in high-risk postmenopausal breast cancer patients," J. Clin. Oncol., 19(14): 3376–3384, 2001.

[40] C. R. Wenger et al., "DNA ploidy, S-phase, and steroid receptors in more than 127,000 breast cancer patients," Breast Cancer Res. Treat., 28: 9–20, 1993.

[41] N. Falette et al., "Prognostic value of P53 gene mutations in a large series of node-negative breast cancer patients," Cancer Res., 58(7): 1451–1455, 1998.

[42] R. M. Elledge et al., "Prognostic significance of p53 gene alterations in node-negative breast cancer," Breast Cancer Res. Treat., 26: 225–235, 1993.

[43] I. L. Andrulis et al., "neu/erbB-2 amplification identifies a poor-prognosis group of women with node-negative breast cancer," J. Clin. Oncol., 16(4): 1340–1349, 1998.

[44] A. K. Tandon et al., "HER-2/neu oncogene protein and prognosis in breast cancer," J. Clin. Oncol., 7(8): 1120–1128, 1989.

[45] M. Ferrero-Poüs et al., "Relationship between c-erb B-2 and other tumor characteristics in breast cancer prognosis," Clin. Cancer Res., 6(12): 4745–4754, 2000.

[46] M. Bolla et al., "Estimation of epidermal growth factor receptor in 177 breast cancers: Correlation with prognostic factors," Breast Cancer Res. Treat., 16: 97–102, 1990.

[47] V. Pawlowski et al., "Prognostic value of the type I growth factor receptors in a large series of human primary breast cancers quantified with a real-time reverse transcription-polymerase chain reaction assay," Clin. Cancer Res., 6(11): 4217–4225, 2000.

[48] C. A. Purdie et al., "Progesterone receptor expression is an independent prognostic variable in early breast cancer: A population-based study," Br. J. Cancer, 110(3): 565–572, 2014.

[49] J. P. Thakkar, D. G. Mehta, "A review of an unfavorable subset of breast cancer: Estrogen receptor positive progesterone receptor negative," Oncologist, 16(3): 276–285, 2011.

[50] J. Anampa, D. Makower, J. A. Sparano, "Progress in adjuvant chemotherapy for breast cancer: An overview," BMC Med., 13: 195, 2015.

[51] P. A. Francis et al., "Tailoring adjuvant endocrine therapy for premenopausal breast cancer," N. Engl. J. Med., 379(2): 122–137, 2018.

[52] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proc. Fifth Berkeley Symp. Math. Statist. Probability, Volume 1: Statistics, 5: 281-298, 1967.

[53] S. C. Johnson, "Hierarchical clustering schemes," Psychometrika, 32(3): 241-254, 1967.

[54] R. L. Thorndike, "Who belongs in the family?," Psychometrika, 18(4): 267-276, 1953.

[55] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," J. Comput. Appl. Math., 20: 53-65, 1987.

[56] T. Caliński, J. Harabasz, "A dendrite method for cluster analysis," Commun. Stat. - Theory Methods, 3(1): 1-27, 1974.

[57] A. K. Jain, M. N. Murty, P. J. Flynn, "Data clustering: A review," ACM Comput. Surv., 31(3): 264–323, 1999.

[58] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel et al., "Scikit-learn: Machine learning in python," J. Mach. Learn. Res., 12: 2825-2830, 2011.

Biographies

Oveis Dehghantanha received his B.Sc. degree in Electronics Engineering from Chamran Technical and Vocational University of Kerman, Iran, in 2010. He received his M.Sc. degree from the University of Birjand, Iran, in 2015. He is currently pursuing a Ph.D. in Electronics Engineering at the University of Birjand. His research interests include biomedical engineering, pattern recognition, machine learning, swarm intelligence, and soft computing.

- Email: O.dehghantanha@birjand.ac.ir
- ORCID: 0009-0008-6940-3189
- Web of Science Researcher ID: N/A
- Scopus Author ID: N/A
- Homepage: N/A

Nasser Mehrshad received his B.Sc. degree in Electrical Engineering from Ferdowsi University of Mashhad, Iran, in 1995. He also received his M.Sc. and Ph.D. degrees in Biomedical Engineering from Tarbiat Modarres University, Tehran, in 1998 and 2005, respectively. He is an expert in biomedical engineering, machine learning, image processing, and pattern recognition. Currently, he is a full Professor in the Department of Electrical and Computer Engineering at the University of Birjand, Birjand, Iran.

- Email: NMehrshad@Birjand.ac.ir
- ORCID: 0000-0001-8678-3402
- Web of Science Researcher ID: N/A
- Scopus Author ID: 36986754400
- Homepage: https://cv.birjand.ac.ir/mehrshad/en/

Roksana Bakhshali is an experienced radiation oncologist with extensive expertise in both clinical practice and research. She earned her Doctor of Medicine from Mazandaran University of Medical Sciences in Sari, Iran, in 2011 and completed her Radiation Oncology degree at Ahvaz Jundishapur University of Medical Sciences between 2015 and 2019. In 2020, she received board certification from the Iranian Board of Radiation Oncology. Currently, she practices as a Radiation Oncologist at Omid Cancer Center in Ahvaz, Iran.

- Email: roksanabakhshali@gmail.com
- ORCID: 0000-0002-9788-0029
- Web of Science Researcher ID: N/A
- Scopus Author: N/A
- Homepage: N/A

Ahmad Reza Sebzari is a board-certified and experienced Radiation Oncologist. He is currently working as an Assistant Professor at Birjand University of Medical Sciences in Birjand, Iran. Dr. Sebzari holds a degree in Radiation Oncology from Tehran University of Medical Sciences (2009-2012) and a Doctor of Medicine (MD) degree from Birjand University of Medical Sciences (2002-2009).

- Email: arsebzari@bums.ac.ir
- ORCID: 0000-0001-5424-0577
- Web of Science Researcher ID: N/A
- Scopus Author ID: 57195574667
- Homepage: sebzari.ir/