UNIVERSITÀ
DI PARMA

UNIVERSITA' DEGLI STUDI DI PARMA

DOTTORATO DI RICERCA

in Medicina Molecolare

Ciclo XXXVII

OCCURS: Oral Cancer Clustering Using Risk Stratification

Coordinatore:

Chiar.mo Prof. Prisco Mirandola

Tutore:

Chiar.mo Prof. Tito Poli

Dottoranda:

Michela Bergonzani

Anni Accademici 2021/2022 - 2023/2024

Table of Contents

Abstract

Head and Neck Squamous Cell Carcinoma (HNSCC), and in particular Oral Squamous Cell Carcinoma (OSCC), is characterized by marked biological heterogeneity and limited prognostic accuracy when relying solely on conventional staging systems such as TNM. This limitation often results in suboptimal risk stratification, potentially leading to overtreatment in low-risk patients and insufficient therapeutic intensity in high-risk cases.

This thesis investigates the application of machine learning (ML) approaches for risk stratification and clustering in HNSCC, with a focus on oral cavity tumors. The primary objective was to evaluate whether the integration of multi-omic data—including clinical, genomic, and radiomic variables—could improve prognostic prediction and better capture tumor heterogeneity compared to traditional methods.

A multi-center dataset from the European BD2Decide project was analyzed using both supervised and unsupervised ML techniques. In the supervised learning framework, Artificial Neural Networks (ANNs) and Random Forest (RF) models were developed to predict 5-year survival outcomes. RF consistently outperformed ANN, achieving higher accuracy and greater robustness, particularly in the context of limited sample size and high-dimensional data. Model performance improved with increasing dataset size, reaching approximately 70% accuracy when larger cohorts were used. Feature importance analysis revealed that, beyond traditional staging variables, functional and systemic parameters—such as swallowing function and hematological markers—contribute significantly to prognostic prediction, supporting the relevance of host-related factors alongside tumor characteristics. Additionally,

the identification of genes involved in key oncogenic pathways supports the biological plausibility of the models and suggests their potential role in uncovering mechanisms underlying tumor progression. Unsupervised learning techniques, including UMAP and hierarchical clustering, enabled the identification of distinct molecular subgroups within HNSCC. Tumor anatomical site emerged as a major driver of transcriptomic variation, with perfect clustering concordance (ARI = 1). Within the OSCC cohort, clustering revealed biologically relevant subgroups that strongly aligned with an established 13-gene prognostic signature, confirming the clinical relevance of the identified patterns. Furthermore, differential expression and functional enrichment analyses identified key biological processes—such as protein translation and acetylation—suggesting a potential role of epigenetic regulation in shaping tumor behavior and patient outcomes.

Despite moderate predictive performance, likely due to limited sample size and data complexity, this study demonstrates that ML-based integration of multi-dimensional data provides meaningful insights into tumor biology and prognosis. These findings support the development of more refined, data-driven stratification models and highlight the potential of ML as both a predictive and hypothesis-generating tool in precision oncology.

Future research should focus on expanding cohort size, improving data harmonization, and validating models across diverse populations. The integration of these approaches into clinical decision support systems may ultimately contribute to more personalized treatment strategies, reduced overtreatment, and improved patient outcomes.

Riassunto

Il carcinoma a cellule squamose del distretto testa-collo (HNSCC), e in particolare il carcinoma orale (OSCC), è caratterizzato da una marcata eterogeneità biologica e da una limitata accuratezza prognostica quando si utilizzano esclusivamente i sistemi di stadiazione convenzionali, come il TNM. Questa limitazione può condurre a una stratificazione del rischio non ottimale, con conseguente rischio di sovratrattamento nei pazienti a basso rischio e trattamento insufficiente nei pazienti ad alto rischio.

La presente tesi esplora l'applicazione di approcci di Machine Learning (ML) per la stratificazione del rischio e l'identificazione di sottogruppi tumorali nell'HNSCC, con particolare attenzione ai tumori del cavo orale. L'obiettivo principale è stato valutare se l'integrazione di dati multi-omici — inclusi dati clinici, genomici e radiomici — possa migliorare la previsione prognostica e descrivere più accuratamente l'eterogeneità tumorale rispetto ai metodi tradizionali.

È stato analizzato un dataset multicentrico proveniente dal progetto europeo BD2Decide mediante tecniche di apprendimento supervisionato e non supervisionato. Nell'ambito dell'apprendimento supervisionato, sono stati sviluppati modelli basati su Reti Neurali Artificiali (ANN) e Random Forest (RF) per la previsione della sopravvivenza a 5 anni. I modelli RF hanno mostrato prestazioni superiori rispetto alle ANN, con maggiore accuratezza e robustezza, in particolare in presenza di dataset di dimensioni limitate e ad alta dimensionalità. Le prestazioni dei modelli sono risultate migliorare con l'aumento della numerosità del campione, raggiungendo circa il 70% di accuratezza nei dataset più ampi.

L’analisi dell’importanza delle variabili ha evidenziato come, oltre ai parametri di stadiazione tradizionali, fattori funzionali e sistemici — quali la funzione deglutitoria e parametri ematologici — contribuiscano in modo significativo alla previsione prognostica, sottolineando il ruolo dei fattori legati all’ospite accanto alle caratteristiche tumorali. Inoltre, l’identificazione di geni coinvolti in pathway oncogenici noti supporta la plausibilità biologica dei modelli e suggerisce un potenziale contributo nella comprensione dei meccanismi alla base della progressione tumorale. Le tecniche di apprendimento non supervisionato, tra cui UMAP e clustering gerarchico, hanno permesso di identificare sottogruppi molecolari distinti all’interno dell’HNSCC. Il sito anatomico del tumore è emerso come un determinante principale della variabilità trascrittomica globale, con una perfetta concordanza nel clustering (ARI = 1). All’interno della coorte OSCC, il clustering ha evidenziato sottogruppi biologicamente rilevanti in forte accordo con una firma prognostica a 13 geni precedentemente descritta, confermando la rilevanza clinica dei pattern identificati. L’analisi dell’espressione differenziale e l’arricchimento funzionale hanno inoltre evidenziato processi biologici chiave, tra cui la traduzione proteica e l’acetilazione, suggerendo un possibile ruolo dei meccanismi epigenetici nel determinare il comportamento tumorale e l’outcome clinico.

Nonostante le prestazioni predittive moderate, verosimilmente dovute alla limitata dimensione del campione e alla complessità dei dati, questo studio dimostra come l’integrazione di dati multi-dimensionali mediante approcci di ML possa fornire informazioni rilevanti sia dal punto di vista prognostico sia biologico. Questi risultati supportano lo sviluppo di modelli di stratificazione più raffinati e sottolineano il potenziale del Machine Learning come strumento sia predittivo sia esplorativo nell’ambito dell’oncologia di precisione.

Studi futuri dovranno concentrarsi sull'ampliamento delle coorti, sulla validazione esterna dei modelli e sull'integrazione in sistemi di supporto decisionale clinico, al fine di favorire strategie terapeutiche più personalizzate e migliorare gli outcome dei pazienti.

1 INTRODUCTION

1.1 Squamous Cell Carcinoma of the Oral Cavity (OSCC)

Oral Squamous Cell Carcinoma (OSCC) represents the predominant histological subtype in its anatomical site, accounting for over 90% of all malignant neoplasms affecting the oral cavity and approximately 8% of the global oncological burden.

In the European context, the epidemiological impact is significant, with approximately 140,000 new cases diagnosed annually and a mortality rate of roughly 70,000 deaths.[1] While the average age at diagnosis is 64 years and 55% of patients initially present with early-stage disease (Stages I and II)[2], a clinical paradox persists. Despite the direct visibility of lesions and the clinical accessibility of the oral cavity, a substantial proportion of patients are still diagnosed at an advanced stage, a factor that critically undermines the overall prognosis.[3]

The etiology of OSCC is inherently multifactorial, arising from a complex interplay of behavioral, infectious, genetic, and environmental determinants:

- Behavioral Factors: Tobacco use remains the primary risk factor, particularly when combined with alcohol consumption, which synergistically amplifies the carcinogenic risk.

- Infectious and Genetic Drivers: While HPV infection is a major driver of oropharyngeal cancers, its contribution to the oral cavity proper is comparatively minimal. Genetic mutations, however, play a pivotal role, particularly in explaining the incidence of OSCC in younger populations or those without traditional environmental exposures.

- Emerging Hypotheses: Notably, up to 20% of OSCC cases occur in the absence of classic risk factors. This suggests the involvement of less-understood mechanisms, such as epigenetic alterations and shifts in the oral microbiome.

Patients with a history of head and neck malignancies or those presenting with potentially malignant oral lesions (PMOLs) remain at high risk and necessitate rigorous long-term surveillance. Consequently, current research efforts are increasingly focused on the identification of molecular biomarkers to enhance early diagnosis, refine prognostic stratification, and facilitate the personalization of treatment protocols.[4]

1.1.1 **TNM staging system and Clinical Implications**

Cancer staging serves as a standardized, anatomically-based framework for describing a patient's overall tumor burden. According to the American Joint Committee on Cancer (AJCC) and the Union for International Cancer Control (UICC), staging is defined by three primary components: the primary tumor (T), regional nodal involvement (N), and distant metastasis (M).

Initially, the clinical stage (cTNM) is established by clinicians through the integration of data derived from physical examination, radiological imaging, biopsy findings, and other diagnostic investigations. Following surgical intervention and tumor resection, the clinical staging is replaced by a pathologic stage (pTNM), which incorporates definitive information from the final histopathological examination of the specimen.[5]

In the management of oral cancer, a notable discrepancy often exists between clinical and pathological staging. The prognostic value of cTNM is generally considered inferior to that of

pTNM, largely because clinical accuracy is heavily dependent on the sensitivity of radiological methods and the clinician's diagnostic expertise. While pTNM is recognized as a highly significant predictor of patient survival, the independent prognostic quality of clinical staging remains less clear.

This distinction carries critical therapeutic weight:

- Primary Treatment Selection: Decisions regarding the initial surgical approach or primary modality are dictated by the cTNM.
- Adjuvant Therapy: The requirement for postoperative radiotherapy (RT) or chemoradiotherapy is determined by the pTNM, based on the microscopic analysis of the resected tumor and neck dissection specimens.

This highlights the imperative for high-precision clinical staging. Current protocols for accurate staging are rigorous and challenging, typically necessitating MRI or CT of the head and neck to evaluate the primary lesion and cervical metastases, complemented by thoracic and abdominal CT to exclude distant metastatic spread.[6]

The ultimate goal of any staging system is to provide an accurate representation of a patient's prognosis. Ideally, each prognostic group (Stages I through IV) should demonstrate high predictive power. According to the Groome criteria, an ideal staging system must exhibit four fundamental characteristics:

- Accurate Outcome Prediction: Precise forecasting of patient results within each stage.
- Hazard Consistency: Similar survival rates among all patients within the same prognostic group.

- Hazard Discrimination: Distinct and significantly different survival rates between different stage groupings.
- Balanced Distribution: An even spread of patients across the various stage categories.

The integration of these features ensures that a staging system remains a robust tool for both clinical decision-making and oncological research.[5]

The 8th Edition of the AJCC Cancer Staging Manual, released in October 2016 (Table 1), represents a comprehensive synthesis of contemporary oncological knowledge across all clinically significant anatomic sites. This edition builds upon a legacy of international synergy and robust classification principles, yet marks a definitive shift in the philosophy of cancer staging.

T-category	AJCC 8
Tx	Primary tumour cannot be assessed
Tis	Carcinoma in situ
T1	≤ 2 cm and ≤ 5 mm DOI
T2	≤ 2 cm and > 5 mm DOI **or** > 2 cm but ≤ 4 cm and ≤ 10 mm DOI
T3	> 2 cm but ≤ 4 cm and > 10 mm DOI **or** > 4 cm and ≤ 10 mm DOI
T4a	Moderately advanced local disease > 4 cm and > 10 mm DOI **or** tumour invades adjacent structures only (e.g., through cortical bon of the mandible or maxilla or involves the maxillary sinus or skin of the face) *Note*: Superficial erosion of bone/tooth socket (alone) by a gingival primary is not sufficient to classify a tumour as T4
T4b	Very advanced local disease Tumour invades masticatory space, pterygoid plates, or skull base and/or encases the internal carotid artery

Table 1. Tumor staging according to the 8th edition AJCC staging of OSCC.[5]

While the anatomic extent of the disease remains the fundamental pillar, the 8th Edition makes a tangible transition from a strictly population-based approach toward a more personalized framework. By integrating biological and molecular markers, the system has evolved into a tool that is not only valid for large-scale epidemiological analyses but also highly effective for determining individualized bedside therapy.[7]

To further refine risk stratification, the 8th Edition emphasizes that the molecular profiling of tumor specimens—including those obtained from pretreatment biopsies—can provide objective insights that enhance clinical staging and optimize prognostic accuracy.[8,9]

Compared to the previous 7th Edition (2010), the 8th Edition introduced several transformative updates specifically for the staging of OSCC and related head and neck cancers (Table 2):

- Introduction of Depth of Invasion (DOI): For OSCC, the DOI is now a primary factor in determining the T classification, recognizing its critical role in predicting nodal metastasis and overall survival.
- Refined Nodal Staging via Extranodal Extension (ENE): In HPV-negative cases, the presence of ENE is now explicitly included in the N classification, as it serves as a powerful indicator of aggressive tumor behavior.
- Differentiation of Oropharyngeal Squamous Cell Carcinoma (OPSCC): The manual now distinguishes OPSCC based on p16 expression status, acknowledging that HPV-mediated tumors represent a distinct biological and clinical entity.

- Revised HPV-Positive Nodal Staging: A specific N classification was developed for HPV-positive tumors to reflect their typically more favorable prognosis despite nodal involvement.

These updates reflect a modern "personalized approach" to oncology, ensuring that staging criteria align more closely with the biological reality of the individual patient's disease (Table 3,

Table 4 and

Figure 1).

Table 2 Summary comparing 7th and 8th ed. AJCC staging of OCSCC and OPSCC.[10]

Change	7th Ed. (2010)	8th Ed. (2017)		
		Oral Cavity	Oropharynx (p16-negative)	Oropharynx (p16-positive)
T-Classification	**TX:** primary tumor cannot be assessed **T0:** no primary **Tis:** carcinoma *in situ* **T1:** size ≤2 cm **T2:** size 2-≤4 cm **T3:** size >4 cm or extension to lingual surface of epiglottis **T4:** ◦ **T4a:** moderately advanced (extrinsic tongue muscle involvement constituted T4a) ◦ **T4b:** very advanced	**T1:** size ≤2 cm and **DOI ≤5 mm** **T2:** size ≤2 cm and **DOI 5-≤10 mm** or size 2-≤4 cm and **DOI ≤10 mm** **T3:** size >4 cm or **any tumor > 10 mm DOI**		**T0** if provenp16+ disease without evidence of primary tumor All locally advanced combined to **T4**
N-Classification	**NX:** regional node involvement cannot be assessed **N0:** no LN involved **N1:** single ipsi LN ≤3 cm in size **N2:** **N2a:** single ipsi LN, 3-≤6 cm in size **N2b:** multiple ipsi LNs, all ≤6 cm in size **N2c:** any bi or ctr LNs, all ≤6 cm in size **N3:** any LN >6 cm in size	**Clinical** **N0—N2:** same as previous and **ENE(-)** **N3** now with subcategories: **N3a:** previous N3 (size >6 cm) and **ENE(-)** **N3b:** any **ENE(+)**, either clinical or radiographic **Pathologic** **Criteria for pathologic ENE(+):** Oral cavity: only macroscopic (>2 mm) Oropharynx: micro- (≤2 mm) or macroscopic (>2 mm) **N1—N2:** same as previous and **ENE(-)** with exception: **N2a** includes lymph node **≤3 cm, ENE(+) LN** **N3** now with subcategories: **N3a** is previous N3 (size >6 cm) and **ENE(-)** **N3b: ≥3 cm and ENE(+) LN** or **> 1 ENE(+) LNs**		• Previous N1, N2a, N2b combined to **N1 (≤6 cm with or without ENE)** • Previous N2c isN2 (≤6 cm with or without ENE) • **ENE status not incorporated** •**N1:** ≤4 LNs involved •**N2:** >4 LNs involved
TNM Stage Grouping	Clinical or pathological TNM used for same grouping system	Same as previous		**Separate clinical and pathological TNM groupings**

DOI: depth of invasion; LN: lymph node; ENE(+): extranodal extension present; ENE(-): extranodal extension absent; ipsi: ipsilateral; bi: bilateral; ctr: contralateral.

Table 2. Summary comparing 7th and 8th ed. AJCC staging of OSCC and OPSCC.[10]

Table 1: Critical Changes in AJCC8 for Head and Neck Cancer Staging

Tumor Type	T Designation	N Designation
Oropharyngeal SCC		
HPV-related oropharyngeal SCC	New staging system for T, N, and prognostic groupings and used when oropharyngeal primary tumor is p16 or HPV positive	New clinical nodal table as follows: cN1: ipsilateral nodes cN2: bilateral or contralateral nodes; and cN3: nodes > 6 cm
	T4 no longer divided into T4a andT4b	New pathologic nodal table if neck dissection performed
Non-HPV oropharyngeal SCC	No changes	Addition of clinical ENE which determines N3b status
		New pathologic nodal table which also incorporates pathologic ENE
Nasopharyngeal carcinoma	Pterygoid muscle involvement now T2, previously included as part of *masticator space* and designated as T4	Level IV and Vb nodes now designated N3 disease with removal of term *supraclavicular fossa*
	Invasion to prevertebral muscles clarified as T2	Nodal masses > 6 cm also N3, removed N3a/N3b
	Invasion to parotid now T4	
Oral cavity	Lip SCC now staged under cutaneous carcinomas of HN	Uses the non-HPV, non-EBV nodal table
	Extrinsic muscle involvement no longer determines T4	
	DOI is new pathologic criterion for determining T status	

Note.—DOI = depth of invasion, EBV = Epstein-Barr virus, ENE = extranodal extension, HN = head and neck, HPV = human papilloma virus, SCC = squamous cell carcinoma.

Table 3. Critical changes in AjCC8 for Head and Neck Cancer staging.[11]

Table 2: Additional AJCC8 Changes to Head and Neck Cancer Staging

Tumor Type	Staging Changes
HPV-mediated oropharyngeal SCC	Marked change to prognostic groupings: N1 = stage I, N2 = stage II, N3 = stage III. Stage IV only if M1 disease
Unknown primary tumors	Part of new chapter on lymph nodes with new direction as to how to stage these tumors
	Directs pathologic evaluation of nodes to first determine whether p16 positive favoring HPV-mediated OPSCC
	No T category is assigned if node is HPV and EBV negative unless a skin primary site is clinically strongly favored
Cutaneous carcinoma of HN	New chapter and T category table for all skin lesions including lesions of the external dry lip (vermillion), but excluding eyelid tumors
Thyroid carcinoma	Thyroid-differentiated carcinoma chapter has been moved out of HN section to the Endocrine section of AJCC8
	For prognostic grouping to distinguish between stage I and stage II, the age has been changed from 45 years to 55 years
Soft-tissue sarcomas	New chapter in AJCC8 and is included in the Soft-Tissue Sarcoma section of AJCC8, not the HN section

Note.—AJCC = American Joint Committee on Cancer, EBV = Epstein-Barr virus, HN = head and neck, HPV = human papilloma virus, OPSCC = oropharyngeal squamous cell carcinoma, SCC = squamous cell carcinoma.

Table 4. Additional AJCC8 changes to Head and Neck Cancer staging.[11]

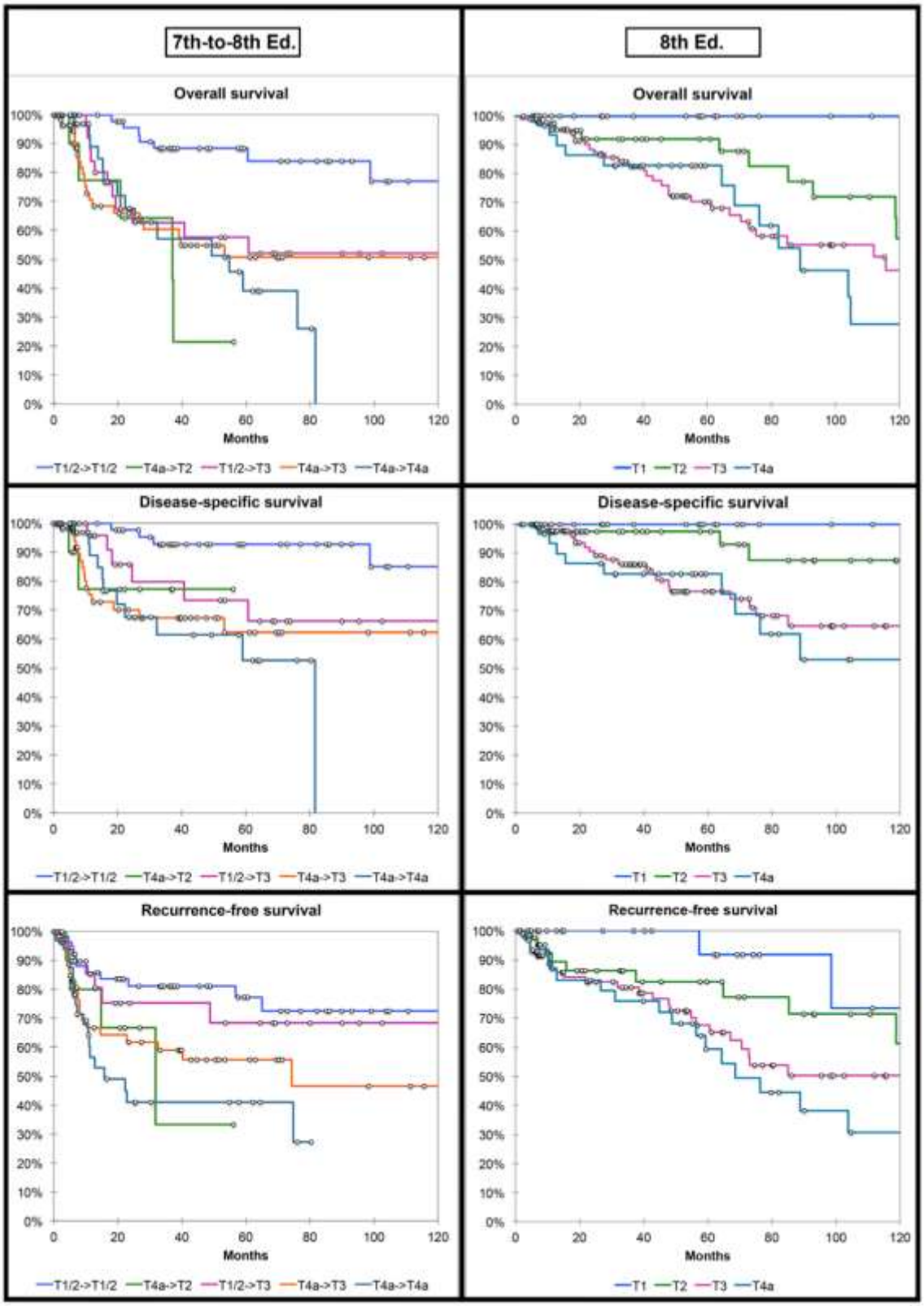

Figure 1. Kaplan-Meier plots illustrating overall, disease-specific, and recurrence-free survival according to T category modification passing from the 7th to the 8th TNM edition (left column) and according to T category of the 8th TNM edition (right column)12

1.1.2 **TNM Staging System Limitations**

Despite its global adoption as the cornerstone of clinical oncology, the TNM staging system exhibits several inherent limitations in the management of OSCC, particularly concerning its sensitivity in prognosticating survival and guiding individualized therapeutic decisions.

While the 8th Edition of the AJCC/UICC classification marked a significant advancement through the integration of Depth of Invasion (DOI) and Extranodal Extension (ENE), the system remains primarily focused on anatomical burden. Consequently, it fails to formally incorporate a spectrum of critical histopathological and biological determinants, such as:

- Perineural Invasion (PNI) and Lymphovascular Invasion (LVI).
- Tumor Grade (differentiation).
- Surgical Margin Status.

These parameters are well-documented drivers of local recurrence and unfavorable outcomes, yet they are not currently utilized to upstage or downstage a patient within the TNM framework.[13]

Furthermore, the system's reliance on morphology introduces significant challenges in clinical practice:

- Staging Discrepancies: Research indicates a persistent gap between clinical (cTNM) and pathological (pTNM) staging, with discrepancies reported in up to 40% of cases, potentially leading to suboptimal or inconsistent treatment planning.[14]

- Demographic Omissions: The current system does not account for age as a prognostic variable. This is a notable omission given emerging evidence that patients over the age of 65 exhibit reduced overall survival rates, even when diagnosed with early-stage disease.
- Nodal Complexity: The N category in the 8th Edition does not fully encapsulate the nuances of regional disease. Factors such as the absolute number of metastatic lymph nodes, their specific dimensions, and the precise volumetric extent of ENE are not adequately represented, despite their proven independent prognostic value.

In summary, these shortcomings highlight a critical need to move beyond purely anatomical staging. There is an increasing imperative for an integrated and personalized risk stratification model—one that synthesizes molecular, histological, and demographic data to refine prognostic accuracy and optimize therapeutic strategies for patients with OSCC.[15]

1.1.3 **Treatment and Prognosis**

Surgical resection remains the cornerstone of treatment for Oral Squamous Cell Carcinoma (OSCC), particularly in early-stage disease (Stages I and II), where it provides the highest probability of local control and long-term survival. Small, accessible lesions are typically managed via a transoral approach, whereas tumors located in the posterior oral cavity or the maxilla may necessitate more invasive surgical access to ensure negative margins.

A critical component of surgical planning is the management of the neck, which is dictated by the nodal status:

- cN0 (Node-negative): Elective neck dissection (END) is generally recommended for patients at significant risk of occult metastasis.
- cN+ (Node-positive): More extensive therapeutic neck dissections are required.

Interestingly, despite the major updates to T and N classifications in the AJCC 8th Edition, the overarching treatment guidelines by stage have remained largely unchanged. This has sparked a significant debate regarding overtreatment in early-stage disease, leading to increased research into treatment de-intensification strategies. These protocols aim to minimize surgical and adjuvant morbidity without compromising oncological efficacy.[13]

For advanced-stage OSCC (Stages III and IV), a multimodal approach is standard, typically combining radical surgery with postoperative radiotherapy, with or without concurrent chemotherapy. In these cases, surgical planning must achieve a delicate balance between oncological radicality and the preservation of essential functions, such as speech and deglutition, making sophisticated reconstructive strategies a vital element of the patient's care path.

The prognosis of OSCC reflects a complex interplay between disease stage, etiology, and molecular profile. While early-stage disease carries a 5-year survival rate of approximately 80%, this figure declines sharply to nearly 20% in advanced-stage cases. Despite continuous refinements in diagnostic techniques and multimodal therapeutic protocols, OSCC continues to exhibit relatively low five-year survival rates, averaging 64.2%.[12] In contrast, HPV-positive cases, frequently localized in the oropharynx, demonstrate significantly better long-term survival outcomes.

The overall poor prognosis of conventional OSCC can be attributed not only to late diagnosis but also to the frequent development of multiple primary tumors. Notably, OSCC has the highest annual incidence of second primary tumors (3–7%) among all cancers.[16] Furthermore, recurrence remains a major concern, with 27% to 50% of patients experiencing relapse within the first two years post-treatment, a stage at which subsequent salvage interventions are often ineffective.

Beyond clinical staging, demographic and molecular factors modulate outcomes. Data from the National Cancer Data Base indicate that younger patients (under 36 years) experience more favorable survival rates (63.7%) compared to those over 65 (47.6%).[17] At the molecular level, the overexpression of biomarkers such as survivin and EGFR correlates with increased lymph node metastasis, while PD-L1 expression shows a strong correlation with advanced TNM staging.

1.1.4 **Limit of current approach**

Despite the global standardization of the TNM staging system and its synergy with postoperative histopathological evaluation, both frameworks exhibit intrinsic limitations when applied to the paradigm of personalized oncology. The TNM classification is based on anatomical extent and does not capture the biological heterogeneity of tumors, limiting its applicability in tailoring therapy to individual patients. Similarly, while histological analysis of tumor specimens plays a crucial role in postoperative assessment, it offers limited value in pre-treatment planning or prognostic evaluation.[18]

One of the most complex clinical decision-making processes in OSCC management involves the clinically node-negative (cN0) neck. Clinicians are currently forced to weigh the benefits and risks of elective neck dissection—which is associated with non-negligible morbidity—against those of active surveillance, which carries the possibility of missing occult lymph node metastases that may manifest later. In this context, molecular profiling emerges as a promising transformative tool. Molecular analysis can be performed on pretreatment biopsy samples, offering objective data to refine clinical staging and enhance the precision of prognostic assessments before tumor resection.[9,19]

The consequences of inadequate risk stratification are not merely statistical but deeply functional. Overtreatment—resulting from aggressive surgical protocols in patients who might have benefited from de-escalation—can lead to irreversible impairment of fundamental physiological functions. The loss of integrity in speech, deglutition, and mastication significantly diminishes the patient's quality of life, underscoring the urgent necessity for a more nuanced, biomarker-driven approach to head and neck oncology.

1.1.5 **Prognostic Models in OSCC**

Prognostic models are designed to synthesize multiple independent variables to estimate clinical outcomes on an individual basis.[20] In the oncological landscape, these tools serve a dual purpose: they refine research methodologies and act as critical adjuncts in clinical decision-making, facilitating tailored treatment planning and providing a standardized framework for patient and family counseling regarding long-term expectations.[21]

However, a significant translational gap remains. Although various prognostic models have been developed for Oral Squamous Cell Carcinoma (OSCC), none have yet achieved integration into routine clinical practice. This lack of clinical adoption is attributed to several systemic challenges:

- Methodological Barriers: Insufficient clinical validation and a lack of reproducibility across different molecular profiling platforms.
- Biological Variability: Inconsistencies related to primary tumor sites and the diverse HPV status of study populations.
- Technical Constraints: High costs associated with high-throughput transcriptomic analyses and the limited applicability of these models to formalin-fixed paraffin-embedded (FFPE) tissue samples—the gold standard in surgical pathology.[19]

A fundamental limitation of existing predictive models is their failure to account for the profound biological heterogeneity of OSCC and the role of population-specific factors in disease progression and therapeutic response. For instance, earlier models—along with the AJCC 7th Edition—omitted the etiological impact of HPV. While it is now established that HPV-mediated carcinogenesis results in a distinct clinical entity with a superior prognosis due to unique molecular drivers, these insights were not adequately reflected in older prognostic frameworks.[22]

The development of clinically reliable models is heavily dependent on rigorous protocols and robust statistical analysis plans. Many existing OSCC models rely on traditional statistical methodologies that may struggle to process the high dimensionality and volume of data

required for precision medicine. These traditional approaches often fail to capture complex non-linear interactions within datasets, thereby diminishing the model's overall predictive power.

Furthermore, the lack of external validation in diverse, real-world cohorts remains a major hurdle. Because OSCC is a relatively rare malignancy, conducting large-scale randomized trials is inherently difficult. As a result, many models are validated within highly selective clinical trial populations that do not accurately reflect the broader and more heterogeneous Head and Neck Squamous Cell Carcinoma (HNSCC) patient population.[23]

In summary, current prognostic models frequently fall short of effectively stratifying patient risk or providing actionable guidance for treatment de-escalation or intensification. These limitations underscore the urgent academic and clinical need for validated, population-representative prognostic tools that can finally bridge the gap from research to the bedside.

1.2 The BD2Decide Project: Advancing Precision Oncology in HNC

The BD2Decide project represents a major advancement in the development of high-precision prognostic systems for Head and Neck Cancers (HNC), with particular focus on Oral Squamous Cell Carcinoma (OSCC). As previously discussed, traditional staging systems such as TNM rely primarily on anatomical criteria and often fail to provide accurate individualized risk stratification, especially in advanced-stage disease. This limitation underscores the need for

integrative approaches capable of improving therapeutic decision-making and reducing overtreatment.

To address these challenges, BD2Decide was designed as a multidimensional prognostic platform integrating clinical, biological, and population-level data. The system combines HPV status, radiological imaging, radiomics, genomic data, and epidemiological variables to stratify patients into distinct prognostic subgroups.

The core objectives of the project include:

- Precision prognostication, improving outcome prediction beyond TNM through population-adjusted models
- Identification of prognostic signatures, both tumor-specific and patient-specific
- Therapeutic prioritization, aligning treatment strategies with evidence-based indicators
- Clinical standardization, reducing subjectivity in multidisciplinary decision-making
- Shared decision-making, supported by visual and interactive tools[24]

Funded under the Horizon 2020 program, BD2Decide addresses the societal challenge of improving disease diagnosis and treatment through digital health innovation. The project is coordinated by the Azienda Ospedaliero-Universitaria di Parma (Italy) and involves a consortium of major cancer centers, including institutions in Italy, the Netherlands (VU University Medical Center and MAASTRO Clinic), and Germany (University of Düsseldorf), with additional collaborations in Austria and Portugal[22].

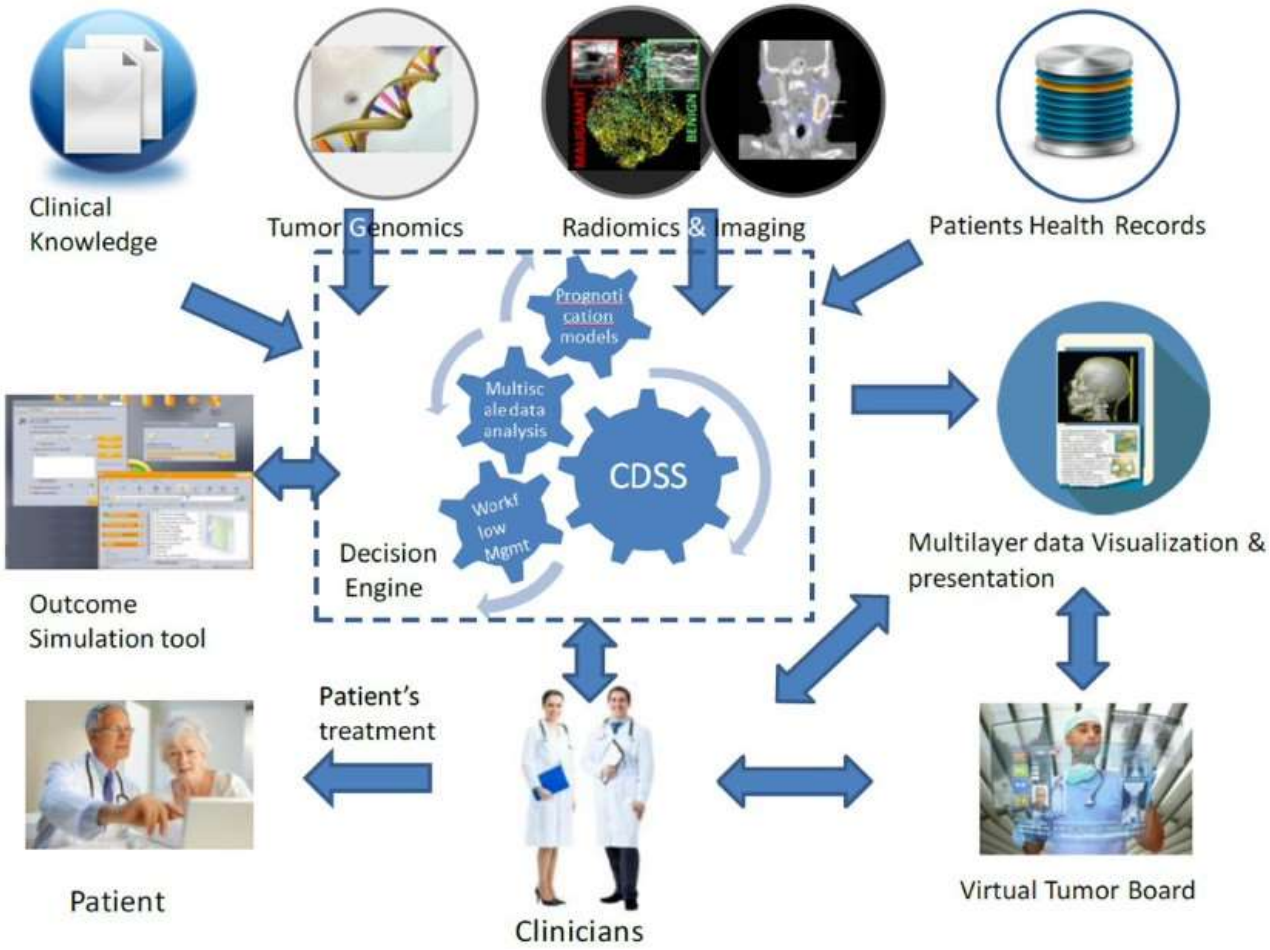

Figure 2. BD2Decide project: clinical knowledge, tumor's genomics data, imaging and radiomics features and data related to patient are integrated with previous prognostic models and analyzed with new statistical methods to create the Decision Support System in order to guide clinical practice.[23]

At the core of the project is the development and validation of an Integrated Decision Support System (DSS), which combines multiscale patient data with advanced visualization tools (Figure 2). Central to this system is the Virtual Patient paradigm, which creates a dynamic digital representation of the individual, enabling clinicians to contextualize patient-specific data within population-based risk models and support personalized treatment decisions[24].

1.2.1 **Objectives, Data and Sources**

The BD2Decide project aims to bridge the gap between Big Data technologies and clinical practice by developing validated prognostic tools for advanced-stage HNSCC. The platform is structured to achieve several key goals:

- improving predictive accuracy beyond TNM staging
- validating models across heterogeneous populations
- identifying novel biomarkers and prognostic patterns
- enhancing quality of life through personalized treatment strategies
- strengthening multidisciplinary collaboration through improved usability

A dedicated pilot study further explored the translational potential of the system, focusing on:

- multidisciplinary validation of imaging, genomic, and histopathological signatures
- integration of clinical, radiomic, and genomic data into unified prognostic models
- evaluation of individual prognostic factor contributions
- assessment of radiotherapy resistance using radiomic features
- scalability of the OraMod model to additional HNC subtypes[25]

The project is grounded in Big Data methodologies, integrating large-scale, heterogeneous datasets that include molecular, imaging, environmental, and socioeconomic variables. This multidimensional approach enables a more comprehensive characterization of tumor biology and patient-specific factors.

The dataset is centrally managed at the University of Parma and represents one of the largest European repositories for advanced-stage HNSCC. Data collection and integration are supported by the Open Clinic platform, which enables standardized acquisition, secure storage, and real-time visualization of patient data[19,26].

Resource to be reused	Source	Functionality / Usage in BD2Decide
Prognostic models	See *Table 2*	Prediction of prognosis for HNC (various subtypes) and extraction of related prognostic biomarkers
INT-172 signature	INT	172-genes prognostic signature for HNC
CosMo model	NeoMark + OraMod (Fraunhofer IGD)	Segmentation of 40 different features in the head and Neck (CT/MRI)
Image Analysis Tools	NeoMark + OraMod (Fraunhofer IGD)	Extraction of prognostic risk features from CT and MRI scans (tumor, lymph-nodes)
Big Data infrastructure	All-in-Image	Big data management, Big Data analysis
NK gene signature	NeoMark	Genomic signature for OSCC recurrence (blood and tumor tissue)
OSCC retrospective data and biologic specimens	>500 NeoMark, OraMod, VU/VUMC retrospective datasets (AOP, UDUS, VU/VUMC	Models refinement, retrospective study
Other HNC retrospective data	>1000 INT, VU/VUMC, MAASTRO, UDUS previous trials and projects	Models refinement, retrospective study
HNC Prospective data	~ 300 DIETINT project (INT) ~ 300 MAASTRO ongoing trials	Clinical assessment of the models and of the **BD2Decide** system (see clinical study outcomes measurement)
Epidemiologic data from cancer registries	RARECARENet study (INT)	• Calibration of prognostic models (including the identification of essential prognostic variables) • Assess the applicability of ***BD2Decide*** in the routine clinical practice
Comedication data Smoking behaviours Alcohol consumption Environmental exposure	Ist. Superiore di Sanità (AOP subcontractor) and other open data sources	• Training of models and of Big Data Analysis tools • System validation • Assess the applicability of ***BD2Decide*** in the routine clinical practice

Table 5. Resources and Uses in the BD2Decide Project. INT = Fondazione IRCCS Istituto dei Tumori Milano; Fraunhofer IGD = Fraunhofer-Institut für Graphische Datenverarbeitung; NK = NeoMark; AOP = Azienda Ospedaliero- Universitaria di Parma; UDUS = University of Düsseldorf; VU/VUMC = VU University Medical Centre Amsterdam; MAASTRO = Maastro Clinic (the Netherlands).[26]

The Decision Support System was developed through a two-phase validation strategy:

- Phase I (retrospective): training on over 1,000 patients from multicenter studies such as De-ESCALaTE, RTOG1016, FP7 NeoMark, PredictCancer, and OraMod[26],[27]
- Phase II (prospective): validation on more than 400 patients with longitudinal follow-up, including 300 cases from the DIETINT Transcan study (spanning Italy, Germany, Portugal, Slovenia, Poland, and Austria) and 100 patients from the MAASTRO Clinic.

In addition to primary patient data, BD2Decide represents a "meta-integration" of existing oncological knowledge, integrating external knowledge sources and validated prognostic models (Table 5), ensuring a strong evidence-based foundation[26].

1.2.2 **Big Data and Machine Learning**

In the medical context, Big Data refers to the integration of large, complex, and heterogeneous datasets, including electronic health records, imaging data, genomic profiles, and epidemiological information. These datasets are characterized by the “Five Vs” framework:

- **Volume**: large-scale data generation
- **Velocity**: rapid data acquisition and processing
- **Variety**: diversity of data types and formats
- **Veracity**: reliability and quality of data
- **Value**: clinical usefulness of extracted information[29]

Due to this complexity, traditional analytical methods are insufficient, and advanced computational infrastructures are required. BD2Decide leverages a Big Data platform

developed by All-in-Image, enabling secure integration and real-time processing of heterogeneous data sources[28].

Machine Learning (ML) represents a central component of this framework. ML algorithms identify patterns within complex datasets and generate predictive models through data-driven learning processes. Within BD2Decide, ML approaches include:

- Supervised Learning: This approach involves training algorithms on labeled datasets containing known input-output pairs. This allows the system to establish a mapping function to predict outcomes for future, unseen cases. It remains the most prevalent methodology in clinical oncological research.
- Unsupervised Learning: In this modality, the algorithm identifies intrinsic patterns or clusters within data without pre-existing labels. It is particularly valuable for data exploration and patient segmentation, having been utilized in BD2Decide to group similar patient profiles.
- Semi-supervised Learning: A hybrid framework that integrates both labeled and unlabeled data, providing a strategic balance between precision and exploratory breadth.
- Reinforcement Learning: An iterative approach where the model learns by interacting with its environment, receiving feedback through rewards or penalties to optimize a cumulative performance metric.[31]

In particular, clustering techniques have been used to identify patient subgroups with similar prognostic characteristics, while Deep Learning approaches enhance pattern recognition in

imaging and genomic data[30]. The integration of ML enables the development of adaptive, personalized predictive models that significantly extend beyond traditional statistical approaches[26].

1.2.3 **Prognostic Models and DDS Architecture**

The BD2Decide project systematically curated, validated, and refined several existing prognostic models specifically for the stratification of patients with Oral Squamous Cell Carcinoma (OSCC). By consolidating these models into a unified Decision Support System (DSS), the project seeks to enhance transparency and clinician trust, thereby facilitating their practical application in both therapeutic decision-making and patient counseling.

The foundational phase of the project centered on the validation and recalibration of these models using the extensive Big Data repository compiled during the study. Recalibration is a critical optimization process that aligns prognostic models with population-specific epidemiological data and risk factors. To maintain individual relevance, a Model Synthesis Tool was engineered to aggregate data from multiple prognostic sources (Table 6) and tailor the output to specific patient profiles. Additionally, the architecture is designed for longitudinal application; a dedicated Model Updating Tool allows for the continuous integration of emerging data.

To enhance clinical applicability, a stepwise approach assigns weights to individual prognostic factors, prioritizing the most relevant variables. This method improves interpretability and cost-effectiveness by reducing reliance on expensive or difficult-to-obtain data. A cost-utility analysis module further evaluates the benefit of additional diagnostic inputs[32].

Type of HNC	Input	Output	Target group
Oral Cancer: The ORAMOD Cancer Evolution Model			
Oral	Clinical Data; TNM; Tumor deph; Risk factors; Biopsy genomic data; Imaging data; Laboratory data.	Prediction of loco-regional reoccurrence, lymph-nodes metastasis and overall survival probabilities.	Patients with oral cancer.
Oral Cancer UK model [25]			
Oral	Age; metastatic disease; discohesive pattern of invasion; smooth muscle actin (SMA) expression.	Prediction of loco-regional reoccurrence, lymph-nodes metastasis and overall survival probabilities.	Patients with oral cancer treated with primary surgery with or without Radiotherapy and not receiving chemotherapy.
TNM staging system			
All HNC	TNM	Overall survival for all cancers	All HNC
Head and Neck Cancer Outcome Calculator [1]			
Oral Cavity **Oropharynx** **Hypopharynx** **Nasopharynx** **Larynx**	Patients' characteristics: current age, sex, race, years since diagnosis. Tumor characteristics: site, diameter, extension, N stage, # of possible nodes, extracapsular spread, histological type.	Cancer Mortality and Life Expectancy.	All HNC patients

Table 6. OSCC prognostic models used within the BD2Decide project.[26]

1.2.4 **Key Results and Achievements**

The BD2Decide project has produced several important outcomes, including:

- establishment of a large, high-quality multicenter dataset
- identification of novel genomic and radiomic biomarkers
- development and validation of integrated predictive models
- creation of advanced "supermodels" combining multi-omic data
- implementation and clinical validation of the DSS

A key finding was the identification of a 48-gene signature capable of outperforming traditional TNM staging in OSCC. Additionally, the integration of radiomic and genomic data significantly improved prognostic accuracy.

Machine Learning analyses confirmed the importance of large datasets for reliable prediction, while the integration of predictive models into clinical workflows demonstrated their practical utility in improving patient management.

1.2.5 **Dissemination and Relevance to the Present Study**

Although the BD2Decide project formally concluded on September 30th, 2019, its scientific and clinical impact extends well beyond its official duration. This work is situated within the ongoing dissemination and exploitation framework mandated by Horizon 2020, in which dissemination supports the advancement of scientific knowledge, while exploitation facilitates the translation of research findings into clinical innovation and public health improvement[1,33].

The continued analysis of the BD2Decide dataset has enabled further exploration of its extensive clinical, genomic, and radiomic information, fostering new research directions and the development of increasingly refined prognostic models. Addressing the intrinsic biological heterogeneity of Head and Neck Squamous Cell Carcinoma (HNSCC) requires precisely this type of large-scale, structured, and multidisciplinary effort, supported by international collaboration and robust data integration strategies[34].

In this context, the BD2Decide dataset represents a unique opportunity to develop predictive models that extend beyond traditional prognostic indicators by incorporating underutilized clinical and biological variables.

Within this framework, the present thesis builds upon the BD2Decide infrastructure to investigate the application of Machine Learning techniques for prognostic stratification and

tumor clustering in HNSCC, with a particular focus on OSCC. By leveraging both supervised and unsupervised approaches, this work aims to identify predictive features, characterize tumor heterogeneity, and research more accurate prognostic tools.

Ultimately, the integration of Big Data and Machine Learning approaches represents a critical step toward the development of personalized and clinically actionable prognostic tools. In this perspective, the present study contributes to the broader BD2Decide initiative by supporting the transition from data-driven research to precision oncology, with the potential to improve patient stratification, guide therapeutic decisions, and optimize clinical outcomes.

2 MATERIALS AND METHODS

This study utilizes the database collected during the BD2Decide project, extending previous clinical analyses through advanced Machine Learning (ML) methodologies. During the Bd2Decide a large, multicenter collection of heterogeneous data was carried out across several European clinical centers. Sources included:

1. Biological Samples: Formalin-fixed tumor specimens for genomic sequencing.
2. Imaging Data: Diagnostic images processed via standardized radiomic pipelines.
3. Clinical & Behavioral Data: Electronic Health Records (EHRs) containing demographic and epidemiological variables, alongside patient-reported Quality of Life (QoL) data gathered through validated questionnaires.[22]

The integration of these diverse data types, complemented by expert clinical knowledge, was achieved through Machine Learning and advanced statistical tools. This phase was conducted in collaboration with the Department of Engineering and Architecture at the University of Parma. The primary objective was the development and calibration of predictive prognostic models designed for routine clinical integration, thereby facilitating increasingly personalized therapeutic decision-making.[35]

2.1 Materials

2.1.1 Patient Selection

The study cohort comprises patients diagnosed with locoregionally advanced head and neck squamous cell carcinoma (HNSCC) (Stages III–IVA/B, according to the 7th edition of the TNM

classification) who were enrolled in the BD2Decide project. While initial staging at diagnosis followed the TNM 7th edition, all patients were subsequently re-staged during the project according to the 8th edition guidelines.[1]

The inclusion criteria were defined as:

- Histological confirmation of HNSCC at primary subsites, including the oral cavity (OSCC), oropharynx (OPSCC), larynx (LCSCC), and hypopharynx (HPCSCC);
- Age ≥ 18 years;
- Clinical stage III, IVA, or IVB (TNM 7th edition);
- Treatment including any combination of surgery, radiotherapy, and chemotherapy;
- Availability of a pre-treatment tumor specimen and contrast-enhanced CT or MRI scans of the head and neck, performed according to standardized protocols;
- Mandatory p16/HPV status assessment for all OPSCC cases.

Exclusion criteria were defined as:

- Diagnosis of metastatic disease (Clinical Stage IVC);
- Prior malignancies within 5 years of the current HNSCC diagnosis;
- Previous head and neck malignancies treated with surgery and/or radiotherapy.[34]

Overall Survival (OS) was defined as the duration from initial diagnosis to death from any cause or the date of last follow-up, with a minimum required follow-up of 5 years. Patients alive with less than 5 years of follow-up were classified as "lost to follow-up" and excluded from the final survival analysis. Survival probabilities were estimated using the Kaplan-Meier method.[1]

The final study population reached a total of 1,537 patients diagnosed and treated between 2008 and 2017, with follow-up data concluded in September 2019.[36]

The study population was divided into two distinct cohorts for model development and verification:

- Training (Retrospective) Cohort: Approximately 1,000 patients treated between 2009 and 2013.
- Validation (Prospective) Cohort: Approximately 400 patients treated more recently.

Individual protocols were developed for both the retrospective and prospective arms, both of which received formal review and approval from the Ethics Committees of all participating clinical centers. All subjects provided informed consent prior to enrollment. The demographic range included both male and female patients, ages ranging from 20 to 93 years (

Patient Characteristics	Retrospective Study		Prospective Study*	
	No.	%	No.	%
Number	**1086**		**389**	
Male/Female	786 /300	72/28	267/121	69/31
Age [median age (range)]	62 (20-93)		60 (20-93)	
Site	n= 1086		n=389*	
Oral cavity	273	25%	126	32%
Oropharynx	443	41%	175	45%
Hypopharynx	130	12%	36	10%
Larynx	240	22%	52	13%
Stage				
III	297	27%	82	21%
IVa	678	63%	277	71%
IVb	111	10%	30	8%
Median follow-up [months]	43		20	

Table 7).[34]

Patient Characteristics	Retrospective Study		Prospective Study*	
	No.	%	No.	%
Number	**1086**		**389**	
Male/Female	786 /300	72/28	267/121	69/31
Age [median age (range)]	62 (20-93)		60 (20-93)	
Site	n= 1086		n=389*	
Oral cavity	273	25%	126	32%
Oropharynx	443	41%	175	45%
Hypopharynx	130	12%	36	10%
Larynx	240	22%	52	13%
Stage				
III	297	27%	82	21%
IVa	678	63%	277	71%
IVb	111	10%	30	8%
Median follow-up [months]	43		20	

Table 7. BD2Decide study population

2.1.2 **Data Collection**

Patient data were systematically collected across several domains, including:

- *Clinical data*
  - Demographics: Patient sex and age at diagnosis.
  - Performance Status: Functional assessment via the ECOG (Eastern Cooperative Oncology Group) scale.
  - Hematological Profile: Laboratory values for hemoglobin, leukocytes (WBC count), and platelets.
  - Clinical Staging: at diagnosis according to the AJCC/UICC TNM 8th Edition.
- *Histopathological and Anatomical Parameters*
  - Morphological Features: Presence of basaloid features, Depth of Invasion (DOI), and maximum tumor diameter.

- Aggressiveness Indicators: Presence of Extranodal Extension (ENE) (formerly referred to as extracapsular extension), Lymphovascular Invasion (LVI), and the Pattern of Invasion (POI).
- Surgical Factors: Status of the surgical margins.
- Pathological Staging: T, N, and overall pathological stage according to the TNM 8th Edition.
- Anatomic Mapping: Specific primary tumor site classified according to the International Classification of Diseases (ICD-10) codes.[37]

- *Risk Factors and Comorbidities*
  - Presence of precancerous lesions.
  - Lifestyle Factors: Smoking habits and alcohol consumption patterns at the time of diagnosis.

- *Radiomic Data*

  Quantitative analysis of pre-treatment radiological imaging (CT and MRI) to extract objective tumor characteristics. Radiomics involves the high-throughput extraction of numerical features using specialized digital computational tools, identifying patterns—such as shape, volume, and voxel-intensity distributions—that are beyond the limits of human visual perception. These features provide a granular definition of the tumor microenvironment and heterogeneity. Radiomic data were avilable for 1,239 patients (80% of the total cohort).

- *Genomic and Transcriptomic Data*

  Genomic profiling focused on the analysis of gene expression within pre-treatment tumor specimens. RNA was extracted from formalin-fixed paraffin-embedded (FFPE)

tumor tissues, followed by transcriptomic sequencing to identify molecular signatures associated with treatment response and prognosis. This molecular characterization was available for 1,284 patients (83% of the cohort).

- *Quality of Life (QoL) Information*

  To assess the functional impact of the disease and its treatment, Patient-Reported Outcome Measures (PROMs) were collected. These evaluations focused on critical domains such as swallowing, speech, and general health status. Data were recorded at three distinct time points: at baseline (pre-treatment) and during follow-up consultations at 6 and 18 months post-therapy.

All multi-modal data were uploaded to the OpenClinica interoperable digital platform. This centralized repository enabled the seamless integration of clinical, radiomic, and genomic variables, facilitating the development and validation of prognostic models.[37]

In the present study, the initial analysis was conducted using only the data from patients recruited at the Maxillofacial Surgery Unit of Parma. This preliminary approach was adopted to ensure the highest possible consistency and reliability in data collection procedures, as all patients were enrolled and evaluated within the same clinical setting following uniform protocols. In the second phase of the study, the analysis was extended to the complete dataset, allowing comparison of predictive accuracy across different cohorts and enabling further exploration of the data using additional machine learning models.

2.2 **Data Analysis**

The data analysis was performed using Machine Learning (ML) models, computational approaches designed to automatically learn patterns and relationships from data. Unlike traditional rule-based programming, ML algorithms are capable of identifying complex and often non-linear interactions within high-dimensional datasets. In the context of this study, these models were applied to analyze the heterogeneous clinical, genomic, and radiomic data associated with Head and Neck Squamous Cell Carcinoma (HNSCC), with the aim of extracting predictive information and improving prognostic assessment.

In this study, both supervised and unsupervised learning paradigms were utilized, each serving a distinct analytical purpose.

Supervised learning was employed primarily for the development of prognostic models. In this approach, the algorithm is trained on a "labeled" dataset, where each patient record is associated with a known clinical outcome (e.g., survival status). The model learns to map input variables—such as clinical, radiomic or genomic features—to these specific outcomes. The main advantage of supervised learning is its typically high predictive accuracy when sufficient labeled data are available. The primary goal is to minimize the error between the model's prediction and the actual clinical "ground truth."

Conversely, unsupervised learning was used for exploratory data analysis and patient phenotyping. This method operates on "unlabeled" data, meaning the algorithm is not informed of the clinical outcome beforehand. Instead, it identifies inherent structures, clusters, or hidden correlations within the data based solely on the similarity of the features.

2.2.1 **Supervised learning**

In the first part of the study, two supervised learning algorithms were compared: Artificial Neural Networks (ANNs) and Random Forest (RF).

2.2.1.1 *Artificial Neural Network*

Artificial neural networks are computational models composed of multiple interconnected layers that process input data to generate predictions. Inspired by the structure of biological neural systems, they consist of simple processing units (neurons) that are densely connected and organized into sequential layers. Deep learning represents a mathematical framework for learning hierarchical data representations through multiple layers of nonlinear transformations. Each layer progressively extracts more abstract features from the input data.

Figure 3 illustrates how a multilayer neural network processes the image of a handwritten digit in order to classify the number represented.

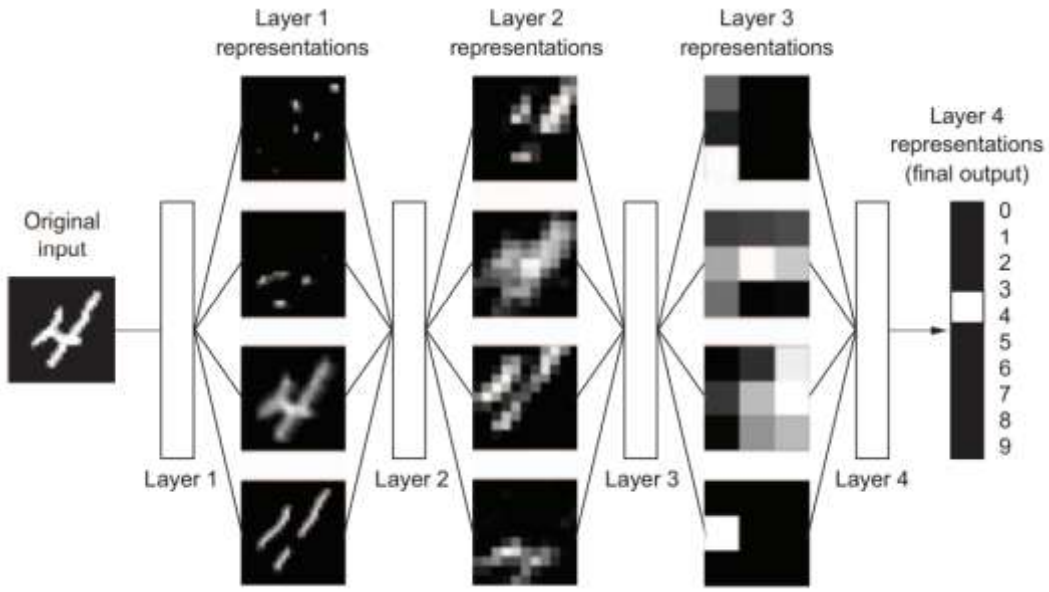

Figure 3. Representation of a multi-layered network which learns numbers classification

During the learning process, the network transforms the input data into internal representations that progressively move closer to the final prediction while becoming less similar to the original input. In this sense, a neural network can be interpreted as a feature extraction system that progressively distills the input information to retain only the components relevant to the predictive task.

The operations performed by each layer are determined by a set of numerical parameters known as weights, which define how input signals are transformed as they propagate through the network (Figure 4).

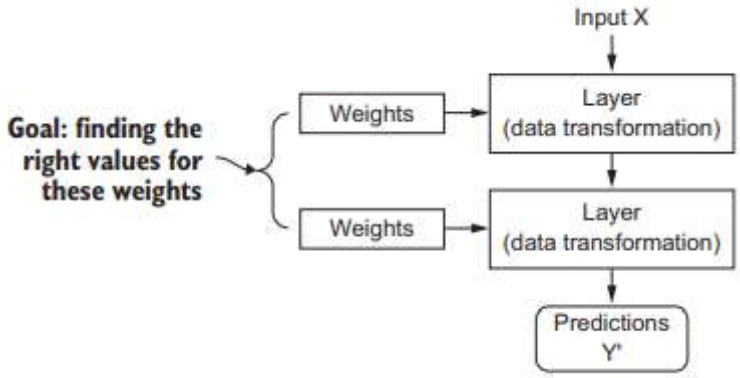

Figure 4. A neural network adjusted by its weights

The learning process consists of identifying the set of weights that allows the network to correctly map input variables to their corresponding labels. In practice, neural networks may contain thousands or even millions of parameters, making the optimization process computationally complex.

To evaluate model performance during training, loss functions are used. These functions measure the discrepancy between the predicted output and the true label. The resulting value provides a feedback signal that is used to iteratively update the network weights through optimization algorithms, thereby reducing prediction error.

At the beginning of training, the network weights are randomly initialized, resulting in a high loss value. As the training process progresses, the model iteratively adjusts the weights using the training data, gradually minimizing the loss.

The supervised learning workflow consists of two main phases:

- Training phase: During training, the neural network receives input data together with the corresponding correct output labels. This allows the model to learn the mapping between input features and target outcomes.
- Testing phase: After training, the model is evaluated using previously unseen data. The predictive performance is assessed by comparing the predicted outputs with the true labels and calculating performance metrics such as classification accuracy.

2.2.1.2 *Random Forest*

Random Forest is a supervised machine learning algorithm based on the aggregation of multiple decision trees (DTs).[38] The final prediction is obtained by combining the outputs of many individual trees, typically through majority voting in classification tasks.

A decision tree is a hierarchical structure composed of a sequence of decision nodes organized in a flowchart-like architecture. Each node represents a question based on a feature value, and the branches correspond to possible outcomes of that decision. The initial node is referred to as the root, while the terminal nodes are called leaves, which represent the final classification output.

The path from the root to a leaf defines a set of decision rules used to classify an observation. An example of a decision tree structure is shown in Figure 5.

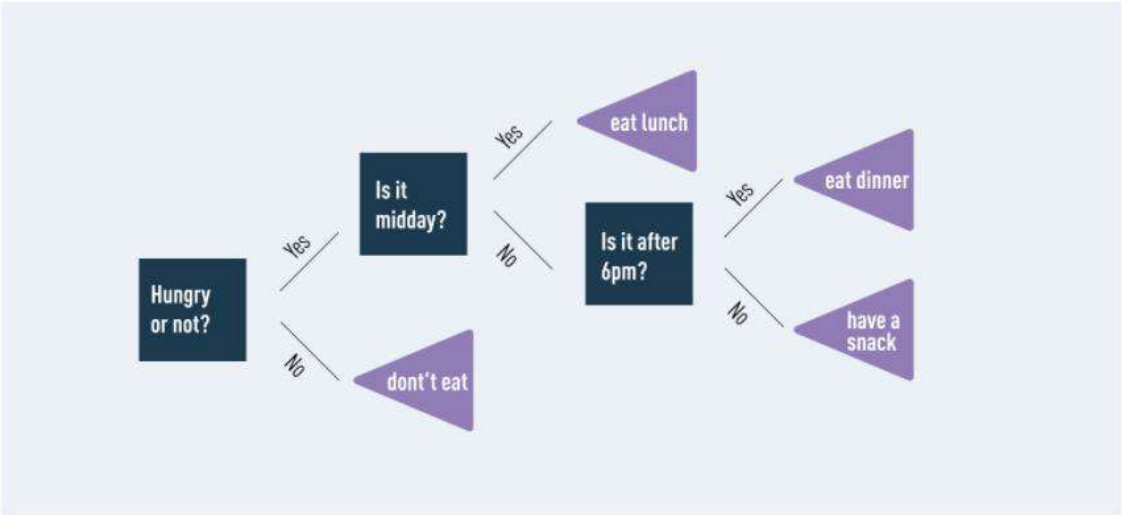

Figure 5. Decision tree example (careerfoundry.com)

When multiple decision trees are combined to form a Random Forest model, predictive performance typically improves. This occurs because each tree is trained on a slightly different subset of the data, introducing variability that reduces overfitting and increases model robustness.

In classification tasks, each decision tree produces a prediction, and the final class assigned by the Random Forest corresponds to the class receiving the highest number of votes among all trees.[38]

2.2.1.3 *Data Pre-processing*

The collected data were pre-processed to make them suitable for machine learning analysis, as most algorithms cannot directly handle textual or unstructured information.

Using Python libraries such as NumPy and Pandas, the datasets were structured, cleaned, and transformed into numerical representations suitable for computational analysis. The pre-processing pipeline included vectorization, normalization, and missing data management.

Vectorization

All input and output variables were converted into numerical tensors (multi-dimensional arrays), which represent the standard input format for neural networks. This process also included the transformation of categorical or textual variables into numerical values.

Normalization

Feature values were scaled to a homogeneous range, typically between 0 and 1. This step improves training stability and prevents features with larger numerical ranges from dominating the learning process.

Missing data handling

Missing values were addressed using several strategies depending on their relevance. Non-critical missing values were replaced with zero, while patients with excessive missing data were excluded from the analysis. In addition, columns containing only missing values were removed.

The Pandas library was used to import and manipulate the original Excel datasets, after which the processed data were converted into NumPy tensors for subsequent modeling.

For the binary classification task (“alive” vs. “dead”), the output variable did not require additional encoding. In contrast, for three-class classification problems, one-hot encoding was applied using the to_categorical function provided by Keras.

2.2.1.4 *Neural Networks and Random Forest Implementation*

Based on the available datasets, several Artificial Neural Network (ANN) and Random Forest (RF) models were developed to predict patient prognosis five years after the initial diagnosis.

The output variable was encoded as:

- 0 – death within five years from diagnosis
- 1 – alive beyond five years or still alive at last follow-up

Ground-truth labels were generated using a custom function (compute_alive_after), which extracted the dates of first diagnosis and death from a Pandas DataFrame and returned a NumPy vector containing binary survival outcomes. The dataset was subsequently divided into training and testing subsets.

The first neural network model was trained using clinical data from 92 patients, corresponding to 62 features after pre-processing. The model architecture was implemented sequentially using Keras, while the Scikit-learn library was used to perform dataset splitting (80% training and 20% testing).

To ensure robust performance estimation, the training procedure was repeated 100 times using identically structured models, and the final accuracy was calculated as the average accuracy across all runs.

Random Forest classification models were implemented using the Scikit-learn library, adopting the Gini impurity as the splitting criterion.

Since genomic and radiomic datasets were already available in numerical format, pre-processing was limited to alignment with the clinical output labels and replacement of missing values (NaN) with zero.

Given the high dimensionality of genomic and radiomic features combined with the limited sample size, a small neural network architecture composed of two hidden layers was used in order to reduce the risk of overfitting.

2.2.1.5 *Most Relevant Features Identification*

One advantage of the Random Forest algorithm is its ability to estimate the relative importance of input features in the prediction process.

To identify the most relevant variables associated with patient prognosis, feature importance scores were calculated across 100 independently trained Random Forest models with identical configurations.

Only clinical and genomic features were included in this analysis.

For each model, the importance weight associated with every feature was stored as a column in a Pandas DataFrame, with rows representing the individual variables. After running all 100 models, the mean importance value was calculated for each feature by averaging the weights across all columns.

This procedure allowed the identification of the variables that consistently contributed the most to the predictive performance of the models.

The same workflow was applied separately to both the clinical and genomic datasets.

2.2.2 **Unsupervised Learning**

In the second part of the study, unsupervised machine learning techniques were applied in order to explore latent structures within the dataset. Unlike supervised learning, unsupervised learning operates on unlabeled data, meaning that no predefined outcome variables or diagnostic classes are available.

In this context, the algorithm analyzes the intrinsic structure of the data and identifies similarities or patterns among observations without prior knowledge of clinical outcomes.

In practical terms, this means that:

- only the input variables are available (for example, gene expression profiles from multiple patients);
- no information about the correct output labels is provided (e.g., survival status or tumor subtype).

The primary goal of unsupervised learning is therefore to identify hidden structures, similarities, and patterns within the dataset.

In this study, two main analytical approaches were applied:

- Dimensionality reduction, using Uniform Manifold Approximation and Projection (UMAP);
- Clustering techniques, used to identify groups of patients with similar molecular or clinical characteristics.

2.2.2.1 *Dimensionality Reduction*

Advances in data acquisition technologies have enabled the generation of increasingly large and complex datasets. In many biomedical applications, the number of variables greatly exceeds the number of observations. For instance, genomic datasets often contain thousands of gene expression variables measured across a relatively limited number of patients.

Such high-dimensional datasets can be difficult to interpret directly because the underlying structure may be obscured by the large number of variables. Dimensionality reduction techniques address this issue by transforming the data into a lower-dimensional representation while preserving the most informative characteristics of the original dataset.

By projecting the data into a reduced space, it becomes possible to visualize patterns, relationships, and potential subgroups that would otherwise remain hidden.

One of the dimensionality reduction methods used in this study was Uniform Manifold Approximation and Projection (UMAP). UMAP is based on the assumption that high-dimensional data lie on an underlying low-dimensional manifold embedded within the original feature space.

In mathematical terms, a manifold is a space that may have a complex global structure but locally resembles a simpler Euclidean space, such as a line or a plane. This concept is particularly useful in biomedical data analysis, where complex multidimensional datasets—such as imaging or genomic profiles—can be represented in a more interpretable form.

UMAP constructs a low-dimensional representation that preserves the local neighborhood relationships between data points while maintaining the overall structure of the dataset as much as possible (Figure 6).

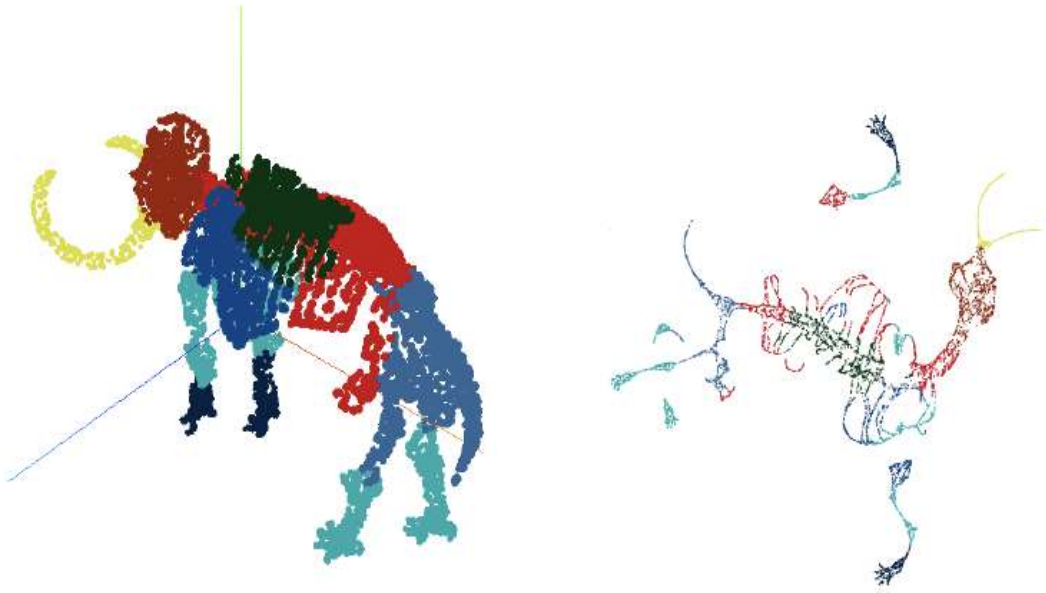

Figure 6. Comparison between the three-dimensional representation of a set of points (on the left) and its two-dimensional reduction obtained with UMAP (on the right). UMAP preserves local structures and the relative distribution of the points, allowing the topology of the underlying manifold to be maintained. The transformation highlights how, in the transition from a high-dimensional space to a two-dimensional space, some geometric details are lost, but the proximity relationships between points are still approximately preserved. This enables a compact representation of the data, useful for visualization and exploratory analysis.

2.2.2.2 *Clustering*

Clustering is an unsupervised data analysis technique used to group similar elements within a dataset. The objective is to partition the dataset into homogeneous subgroups, called clusters, such that:

- patients with similar characteristics are assigned to the same cluster;

- patients with different characteristics are separated into distinct clusters.

An effective clustering solution generally satisfies three basic conditions:

1. Completeness: every observation in the dataset must be assigned to a cluster;
2. Disjointness: each observation must belong to one and only one cluster;
3. Validity: each cluster must contain at least one observation and should not correspond to the entire dataset.

By grouping observations based on similarity, clustering can reveal previously unrecognized patterns or subtypes within the data, which may have biological or clinical relevance.

2.2.2.3 *Partitional Clustering*

In partitional clustering, the dataset is divided into a predefined number of clusters, with the aim of maximizing similarity within clusters and differences between clusters.

Each possible partition is evaluated through a cost function, which measures cluster compactness and penalizes configurations where elements within the same group are widely dispersed.

One of the most widely used algorithms for this approach is K-means clustering. This algorithm partitions the data into K clusters by minimizing the sum of the squared distances between each observation and the centroid of its assigned cluster.

The procedure follows an iterative process:

1. Initial centroids are randomly assigned.

2. Each observation is assigned to the nearest centroid.
3. Centroids are recalculated based on the mean position of all observations within each cluster.
4. Steps 2 and 3 are repeated until convergence is reached.

The distance metric most commonly used in K-means is the Euclidean distance, which is particularly suitable for continuous quantitative data such as gene expression measurements.

A key challenge in partitional clustering is determining the optimal number of clusters (K), which is generally not known in advance. One commonly used heuristic approach is the Elbow Method, which evaluates the reduction in clustering cost as the number of clusters increases. The optimal value of K is typically identified at the point where further increases in cluster number produce only marginal improvements in the objective function.

2.2.2.4 *Hierarchical Clustering*

Hierarchical clustering is an alternative clustering approach that organizes observations into a tree-like structure, known as a dendrogram, representing nested groups of data points based on their similarity.

Two main strategies can be used:

- Agglomerative approach (bottom-up): Each observation initially forms its own cluster. The algorithm then progressively merges the most similar clusters until only a single cluster remains or until a predefined stopping criterion is reached.

- Divisive approach (top-down) The algorithm starts with all observations in a single cluster and iteratively splits the dataset into smaller groups based on dissimilarities.

The similarity between observations is quantified using distance metrics. In this study, two distance measures were considered:

- Euclidean distance, representing the geometric distance between two points in multidimensional space;
- Cosine distance, which measures the angle between two vectors and is often more appropriate for high-dimensional datasets.

Before clustering begins, a distance matrix is computed containing the pairwise distances between all observations. This matrix forms the basis for the hierarchical clustering process.

Different linkage criteria can then be applied to determine how clusters are merged:

- Complete linkage, which considers the maximum distance between observations belonging to different clusters. This method tends to produce well-separated groups.
- Ward linkage, which merges clusters in a way that minimizes the increase in within-cluster variance, leading to compact and homogeneous clusters.

By analyzing the dendrogram (Figure 7), hierarchical clustering enables exploration of the dataset at different levels of granularity.

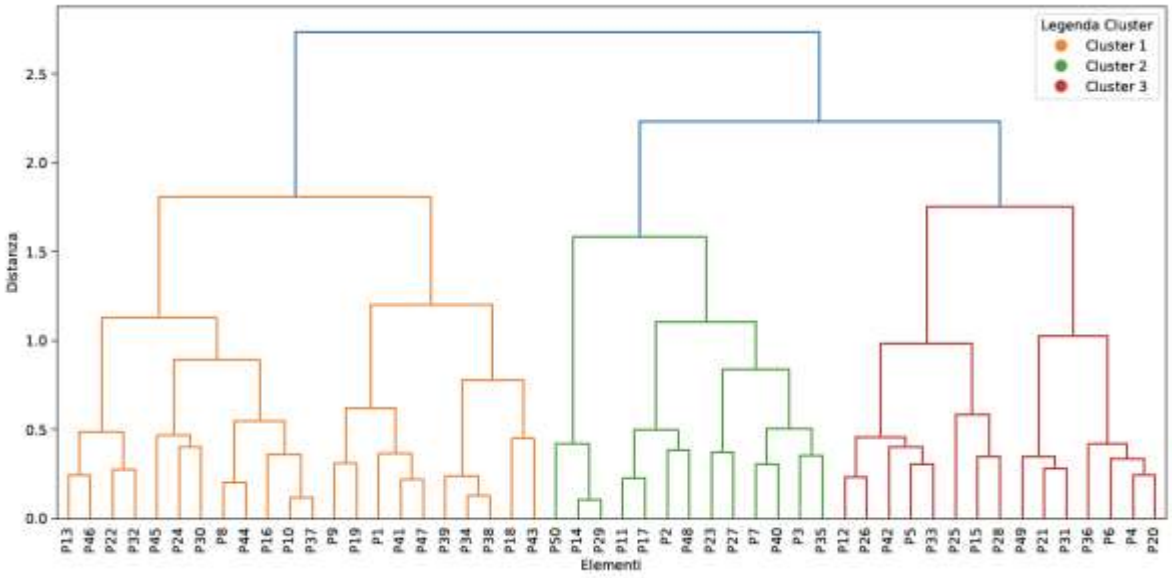

Figure 7. Dendrogram obtained with hierarchical clustering. The vertical axis represents the distance between clusters, while the horizontal axis shows the grouped elements. The colors identify the final clusters obtained.

2.2.2.5 *Cluster Validation*

After performing clustering analysis, it is necessary to evaluate the quality and validity of the identified clusters.

Several metrics can be used to assess:

- cluster compactness (similarity among elements within the same group),
- cluster separation (differences between groups),
- agreement with known clinical classifications, when available.

In this study, three evaluation methods were employed.

Adjusted Rand Index (ARI)

The Adjusted Rand Index (ARI) compares the clustering result with an existing classification, such as a clinical categorization. The metric also accounts for agreements that may occur purely by chance.

- Values close to 1 indicate strong agreement between clustering and the reference classification.
- Values close to 0 or negative indicate poor correspondence.

Confusion Matrix

A confusion matrix provides a tabular comparison between cluster assignments and known labels. It indicates how many patients were correctly or incorrectly assigned to each group, allowing identification of systematic classification errors.

Silhouette Index

The Silhouette coefficient evaluates clustering quality without relying on external labels. For each observation x:

- $a(x)$ represents the average distance between x and all other observations within the same cluster (intra-cluster distance);
- $b(x)$ represents the average distance between x and the closest neighboring cluster (inter-cluster distance).

The silhouette value ranges from −1 to +1:

- values close to +1 indicate that the observation is well matched to its cluster;
- values near 0 indicate that the observation lies near the boundary between clusters;
- negative values suggest possible misclassification.

The average silhouette score across all observations provides an overall measure of clustering consistency.

2.2.2.6 *Integration with Clinical Data*

Finally, the results of the genomic clustering analysis were integrated with the available clinical dataset.

In particular, potential associations between cluster membership and clinical characteristics—such as tumor stage, survival outcomes, and other patient variables—were investigated. This step aimed to determine whether the molecular patterns identified through unsupervised learning corresponded to clinically meaningful patient subgroups.

3 RESULTS

3.1 Supervised learning

The supervised learning phase utilized a multi-modal dataset comprising clinical, radiomic, and genomic domains. The analysis focused mainly on the cohort recruited at the University Hospital of Parma (AOP) and the broader European BD2Decide dataset for comparative validation.

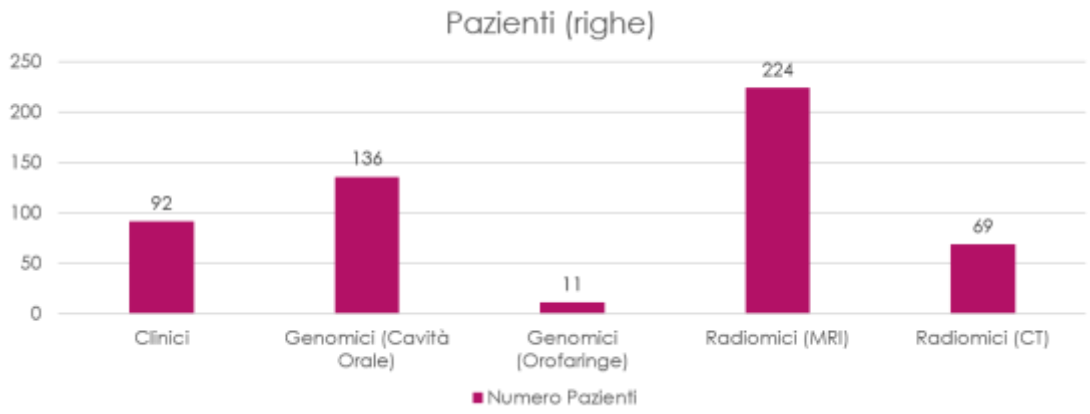

Figure 8. Number of patients for each group

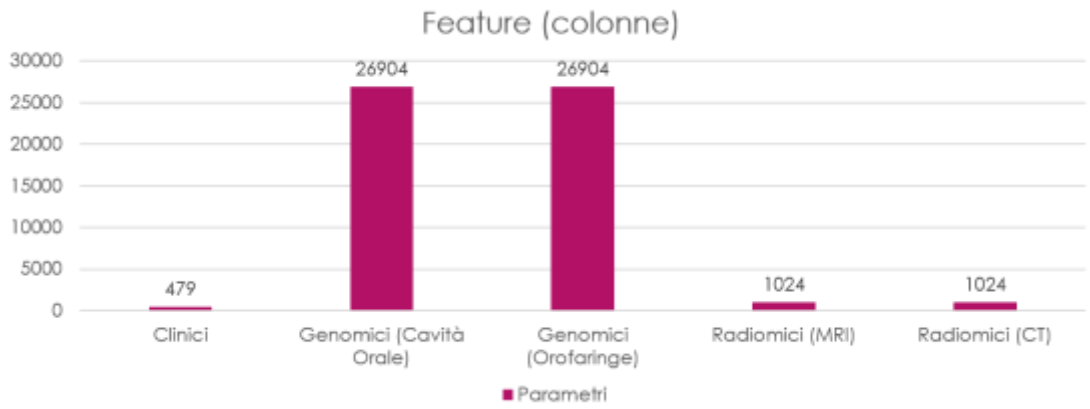

Figure 9. Number of columns for each patient

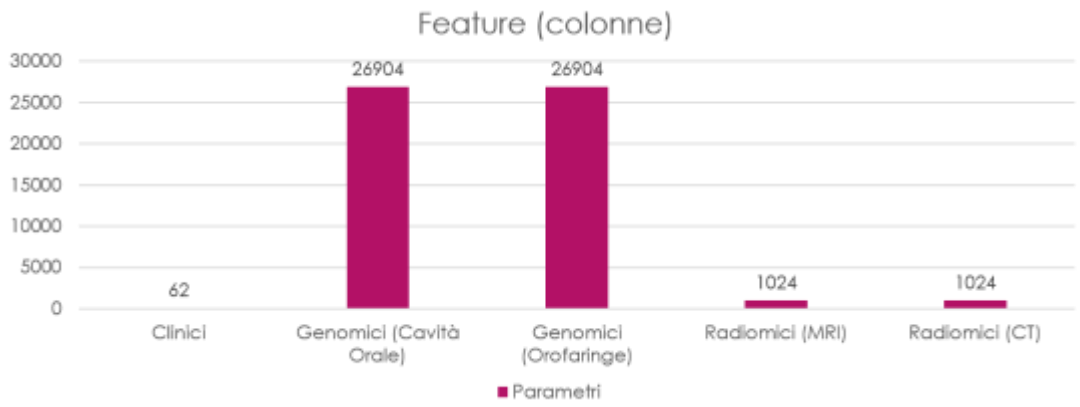

Figure 10. Number of features in the different datasets after pre-processing

As illustrated in

Figure 8 and

Figure 9, clinical data from AOUPR were available for 92 patients and included 479 clinical parameters (columns).

Genomic data are partitioned into two distinct tables based on tumor location:

- Oropharyngeal tumors: Data from 11 patients.
- Oral cavity tumors: Data from 136 patients.

Both genomic tables consist of 26,904 features (columns), representing approximately 27,000 molecular markers per patient, as shown in

Figure 10.

Radiomic data are similarly divided into two tables based on imaging modality:

- MRI images: Data from 224 patients.
- CT scans: Data from 69 patients.

As indicated in Figure 2.2, both radiomic datasets consist of 1,024 quantitative features (columns) extracted from the medical imaging.

3.1.1 **Models' Accuracy**

Figure 11 summarizes the accuracies obtained by applying both Random Forest (RF) and Artificial Neural Networks (ANN) to the different datasets.

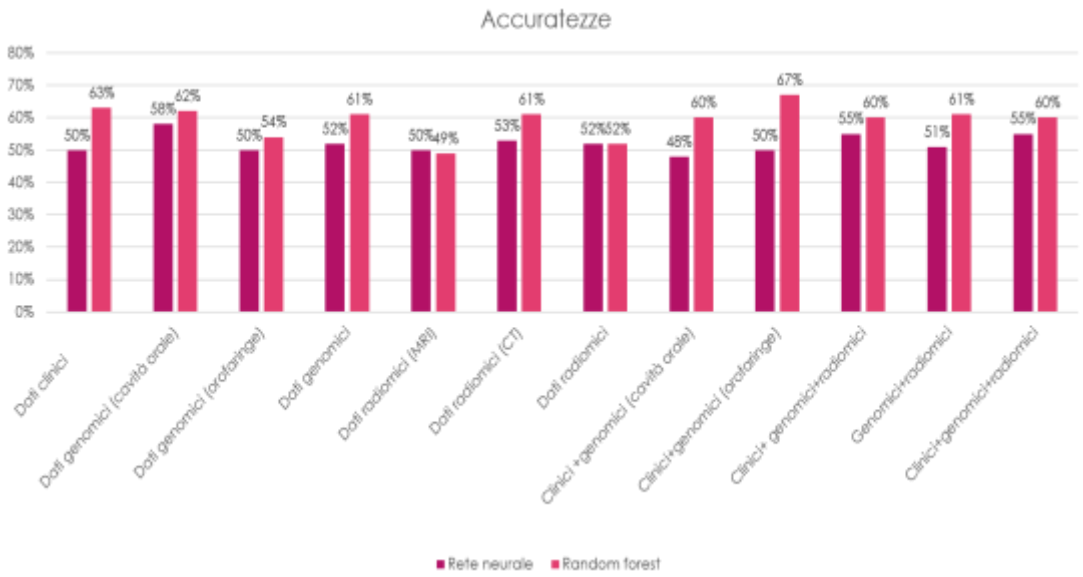

Figure 11. Comparison between all the accuracies of different datasets obtained in Data Analysis

3.1.1.1 *Clinical Data Performance*

The first neural network model was trained using clinical data from the AOP cohort (92 patients, reduced to 62 features after pre-processing, as shown in

Figure 10). The accuracy, calculated as the average of 100 identically structured runs, was 50%, likely due to the limited sample size. In contrast, the Random Forest model achieved a higher accuracy of 63% on the same data.

When extending the analysis to the full BD2Decide project clinical dataset (1,086 patients), the predictive performance significantly improved. As shown in

Figure 12, the accuracy increased to 55% for the ANN and 70% for the RF. This comparison highlights a clear trend: as the volume of available data increases, the accuracy of both machine learning architectures improves.

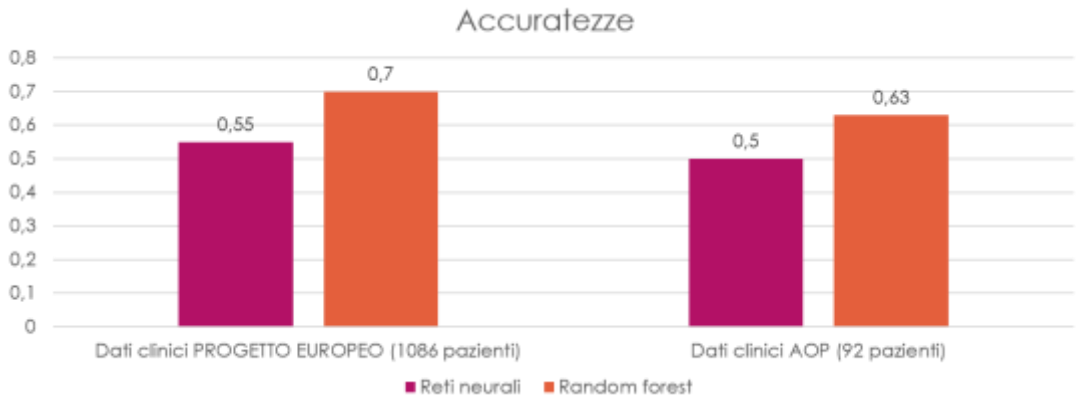

Figure 12. Comparison of model accuracies with clinical data from the European BD2Decide project and with data from the AOP

3.1.1.2 *Genomic Data Performance*

The genomic datasets—divided into oral cavity (136 patients) and oropharyngeal tumors (11 patients)—comprise 26,904 features each (

Figure 8 and

Figure 9).

- For the oral cavity group, accuracies reached 58% (ANN) and 62% (RF).
- For oropharyngeal tumors, performance was lower at 50% (ANN) and 54% (RF).

After merging these two genomic datasets, the final accuracy was recorded at 52% for the ANN and 61% for the RF (Figure 13).

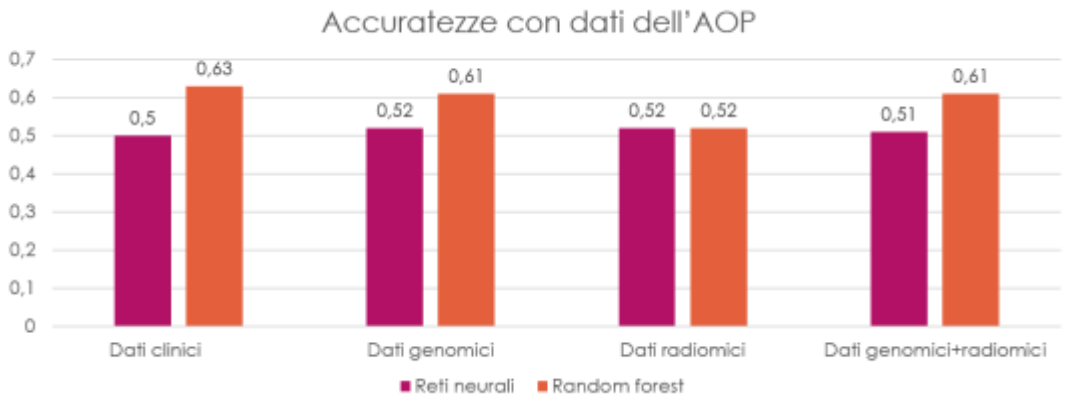

Figure 13. Neural network accuracies and random forest with the different datasets

3.1.1.3 *Radiomic Data Performance*

The radiomic data were analyzed separately for MRI (224 patients) and CT scans (69 patients), each containing 1,024 features.

- For MRI, the models achieved 50% (ANN) and 49% (RF) accuracy.
- For CT, the results were higher, with 53% (ANN) and 61% (RF).

When combining MRI and CT data into a single radiomic dataset, the accuracy settled at 52% for both models.

3.1.1.4 *Multi-modal Data Integration*

Several experiments were conducted to evaluate the impact of data fusion:

1. Genomic + Radiomic: Combining these two modalities resulted in accuracies of 51% (ANN) and 61% (RF).
2. Clinical + Genomic (AOP): To evaluate the specific contribution of clinical variables, they were integrated with the genomic data. This yielded RF accuracies of up to 67% (with oropharyngeal genes) and approximately 60% (with oral cavity genes), while the ANN performance remained stable between 48% and 50%.
3. Full Integration: Finally, integrating clinical, genomic, and radiomic data produced an accuracy of 55% for the ANN and 60% for the RF.

Overall, these results confirm that Random Forest generally demonstrated superior and more stable performance across the various experimental configurations compared to the Neural Network models.

3.1.2 Identification and Characterization of Relevant Features

To identify the variables most significantly associated with patient prognosis, we leveraged the Random Forest algorithm's inherent ability to estimate feature importance. As detailed in the methodology, importance weights were averaged across 100 independently trained models to ensure the stability and reliability of the rankings.

To focus the analysis on the most impactful predictors, only the 50 highest-ranking variables from the clinical and genomic datasets were selected for further characterization (

Table 8, 9 and Figure 14).

FEATURE	Associated Weight (%)
clinical_Age_at_Diagnosis	4,5754
clinical_Year_of_Birth	4,3778

qol_Deglutition	3,9973
clinical_HB_Level	3,6587
clinical_PLT_Level	3,5687
clinical_PLT_Level_Total	3,5669
ctn_TNM_cN_8Edition_OralCavity_Oropharynx_p16Negative_Hypopharynx_Larynx	3,4172
clinical_HB_Level_Total	3,2441
clinical_Leukocytes	3,1059
risk_Pack_Years	3,1009
clinical_Leukocytes_Total	3,0902
risk_Years_as_a_Smoker	2,8459
ctn_Anatomical_Tumor_Location	2,7911
ctn_TNM_cT_8Edition_OralCavity_Oropharynx_p16Negative_Hypopharynx_Larynx	2,7156
ctn_Stage_at_Diagnosis_8Edition_OralCavity_Oropharynx_p16Negative_Hypopharynx_Larynx	2,2074
risk_Packs_Smoked_per_Day	2,1353
risk_Whiskey	2,0028
ctn_Node_Levels_Left	1,9274
ctn_TNM_cN_7Edition	1,9197
clinical_Overall_Comorbidity_Score	1,9098
risk_History_of_Alcohol_Dependence	1,904
ctn_Node_Levels_Right	1,7915
ctn_TNM_cT_7Edition	1,6393
clinical_Lymphocytes_Total	1,5479
clinical_Lymphocytes	1,5394
patho_Degree_of_Cell_Keratinisation	1,539
risk_Oral_Hygiene	1,5203
patho_Grade_at_Diagnosis	1,3832
clinical_ECOG_Performance_Status	1,3487
ctn_TNM_cN_8Edition_Oropharynx_p16Positive	1,3265
INT_OSCC_48gene_model_risk	1,3216
clinical_ASA	1,319
risk_Alcohol_at_Time_of_Diagnosis	1,2155
ctn_Stage_at_Diagnosis_8Edition_Oropharynx_p16Positive	1,2121
risk_Family_History_of_Malignancies	1,0914
ctn_TNM_cT_8Edition_Oropharynx_p16Positive	1,0794
ctn_Tissue_Invasion_ExtraNodalExtension	1,0573
ctn_Tumor_Region	0,9678
ctn_Laterality_of_T	0,9628
INT_HNC_prognostic_LA_model_risk	0,9466
risk_Smoker_at_Time_of_Diagnosis	0,9116
qol_Respiration	0,8954
ctn_Stage_at_Diagnosis_7Edition	0,85
clinical_Hospital	0,842

patho_Basaloid_features	0,7902
risk_Std_Wine	0,7676
ctn_Extension_to_lingual_surface_of_epiglottis_is_present	0,7476
clinical_Sex	0,6752
INT_Lohavanichbutr_OSCC_HPVneg_risk	0,6508
clinical_Zubrod_Performance_Status	0,6402

Table 8. Most relevant clinical features identified

GENE	Associated Weight (%)
SEMA3A	0,1061
C9orf64	0,0915
ZNF785	0,0755
MIR5697	0,0725
LINC01176	0,0712
LOC105372069	0,0652
RAB11B	0,0647
TMEM206	0,0643
CAPG	0,0641
LINC00977	0,061
MIR7854	0,0606
IGHVIII-26-1	0,0575
GUCY2GP	0,0556
GPR26	0,0549
IGHV1-45	0,0543
RPS10P7	0,0534
EIF2S3	0,0533
MZF1	0,0532
KCTD12	0,0526
MTRNR2L2	0,0525
IMMP2L	0,0524
PAXBP1-AS1	0,0503
RHEBL1	0,049
LINC00412	0,049

KRTAP4-11	0,0486
TCHP	0,0474
DTD2	0,0468
GPR32	0,0466
SV2A	0,0463
ZGRF1	0,0435
ALDH7A1	0,0433
IGKV2D-28	0,043
PHPT1	0,0429
FCRL6	0,0418
TCEAL8	0,0418
GPR1	0,041
STOML1	0,041
SLFN11	0,041
LOC100128568	0,0407
ADAM8	0,0405
CDH12P4	0,0403
SOWAHB	0,0403
RAB29	0,0403
UBA3	0,04
KCNK10	0,0399
FANCG	0,0397
IGHVIII-5-2	0,0387
SYNC	0,038
TRIM38	0,0379
HIST3H2BA	0,0374

Table 9. Most important oral cavity genomic data

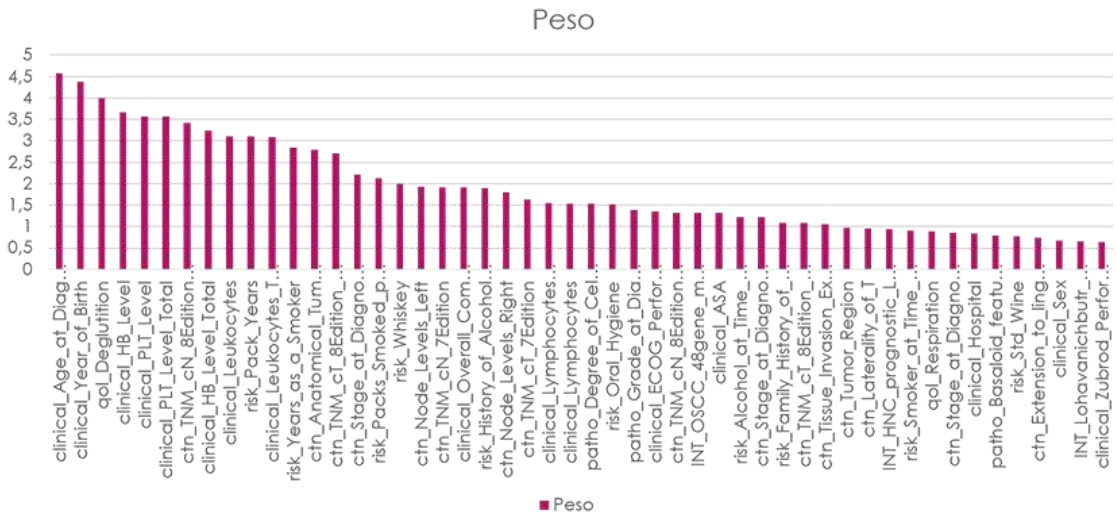

Figure 14. Clinical data features in order of importance

3.1.2.1 *Clinical feature characterization analysis*

The feature importance analysis highlights several critical domains influencing survival among in the clinical dataset:

- Demographics: Age at diagnosis and year of birth emerged as the most influential predictors, confirming the impact of biological age on HNSCC outcomes.
- Clinical and Laboratory Markers: Systemic health indicators, specifically hemoglobin (HB) and platelet (PLT) levels, showed high prognostic weight, suggesting the importance of the patient's hematological and inflammatory status.
- Disease Staging: The 8th Edition of the TNM classification (both T and N categories) consistently contributed more to the model's predictive power than older staging versions, validating its clinical use.
- Lifestyle and Quality of Life: Smoking history (Pack Years) and alcohol consumption patterns were identified as significant risk factors. Notably, the Quality of Life (QoL)

score for deglutition (swallowing) was ranked as a top-three predictor, underlining how pre-treatment functional status directly relates to long-term prognosis.

3.1.2.2 *Genomic characterization analysis*

Following the identification of the most relevant genomic features through the Random Forest weight estimation, a detailed biological characterization was performed on the top-ranked genes. To bridge the gap between statistical importance and clinical function, the OMIM (Online Mendelian Inheritance in Man) database was utilized to investigate the specific roles of the selected gene subset.

OMIM is a comprehensive, authoritative, and daily-updated compendium of human genes and genetic phenotypes. It provides full-text, referenced overviews of all known Mendelian disorders and over 16,000 genes. For this study, OMIM was used to retrieve the following data for the most significant genes identified by our model:

- Biological Function: The primary role of the gene in cellular processes.
- Chromosomal Location: The specific physical mapping of the gene.
- Expression Patterns: The tissues or conditions in which the gene is most active.
- Common Mutations: Known pathogenic variants or polymorphisms that may influence tumor progression.

The biological characterization of the most significant genes identified in our analysis is summarized in

Table 100.

GENE	*OMIM*	*MAPPING*	*FUNCTION*	*EXPRESSION*
Semaphorin 3a - **SEMA3A**	***** 603961	7q21.11	• neurosignaling • osteoprotective effect in bones • signaling in taste receptor cells • Kallman Syndrome gene (Hypogonadotropic Hypogonadism 16 with or without Anosmia)	Brain Taste receptor cells Bone Thymic cells
Chromosome 9 open reading frame 64 - **C9ORF64**	* 611342	9q21.32	• Isolated in chromosome 9q deleted in acute myeloid leukemia	-
ZNF785	-	-	-	-
MIR5697	-	-	-	-
LINC01176	-	-	-	-
LOC105372069	-	-	-	-
Ras-associated protein - **RAB11B**	* 604198	19p13.2	• Part of Ras superfamily of small GTP-binding proteins with regulating exocytotic and endocytotic pathways activities • Allelic variants were isolated in some neurodevelopmental disorders	-
Proton-activated chloride channel 1 - **PACC1** or **TMEM206**	* 618427	1q32.3	• Membrane protein functioning as proton-activated chloride channel • Gene knockout partially protected HEK293 cells from acid-induced death	Brain
Capping protein, gelsolin-like - **CAPG**	* 153615	2p11.2	• Modulator of actin structures in response to external stimuli	Ubiquitary
LINC00977	-	-	-	-
MIR7854	-	-	-	-
IGHVIII-26-1	-	-	-	-
GUCY2GP	-	-	-	-
G protein-coupled receptor 26 **-** **GPR26**	* 604847	10q26.13	• Sequence homology with serotonin receptor 5-HT5A and gastrin-releasing hormone receptor BB2	-
Immunoglobulin heavy chain	* 147070	14q32.33	• N-terminal variable (V) region containing the antigen-binding	-

variable gene cluster – **IGHV1-45**			site, of immunoglobulin heavy chains • Genetic variation of these loci may be associated with genetic differences in immune response	
RPS10P7	-	-	-	-
Eukaryotic translation initiation factor 2, subunit 3 - **EIF2S3**	* 300161	Xp22.11	• Translation initiation factor • Mutations causing syndromic X-linked mental retardation (MEHMO S.)	-
Myeloid zinc finger 1 - **MZF1**	* 194550	19q13.43	• Metal-binding proteins active as transcriptional regulator in myeloid differentiation	-
Potassium channel tetramerization domain-containing protein 12 - **KCTD12**	* 610521	13q22.3	• Potassium channel: N-terminal voltage-gated potassium channel tetramerization (T1) domain	Strongly expressed in fetal tissues; barely detectable in adult tissues
MTRNR2L2	-	-	-	-
Inner mitochondrial membrane peptidase, subunit 2 - **IMMP2L**	* 605977	7q31.1	• Mitochondrial membrane protein: inner membrane peptidase	Ubiquitary
PAXBP1-AS1	-	-	-	-
Rheb-like protein 1 - **RHEBL1**	* 618956	12q13.12	• RAS family small GTPase active in control of growth, differentiation, and transformation • Involved in signaling of mTOR pathway	Ubiquitary
LINC00412	-	-	-	-
KRTAP4-11	-	-	-	-
Trichoplein – **TCHP** Keratin filament-binding Mitostatin	* 612654	12q24.11	• negative regulator of ciliogenesis, • recruitment of microtubules to centriole controller, • maintenance of mitochondrial morphology • Overexpression inhibites cell growth and enhances apoptotic response • Gene knockdown increased cell survival rate	Liver Kidney Heart Skeletal Muscle Enterocytes Reduced in bladder and breast tumours (advanced stages)
DTD2	-	-	-	-

G protein-coupled receptor 32 - **GPR32**	* 603195	19q13.33	• Chemoattractant receptors	-
Synaptic vesicle glycoprotein 2A - **SV2A**	* 185860	1q21.2	• Uptake control of neurotransmitters into vesicles • Insulin secretion regulation • Binding site of botulinum neurotoxin A	Highest levels in brain, lung, and ovary. Intermediate expression in heart, kidney, pancreas, spleen, and testis
ZGRF1	-	-	-	-
Aldehyde dehydrogenase 7 family, member a1 - **ALDH7A1**	* 107323	5q23.2	• Semialdehyde dehydrogenase in the acid pathway of lysine catabolism • Gene mutations in patients with pyridoxine-dependent epilepsy	-

Table 100. Analysis of first thirty most relevant genomic data

The objective of this analysis was to delineate the molecular characteristics of the study cohort's tumours. By identifying specific genomic profiles, this study aims to facilitate the future delivery of tailored therapeutic strategies within the framework of Precision Medicine.

Database Coverage and Expression Profiles

Of the top 32 genes identified by the Random Forest model, 16 were successfully cross-referenced within the OMIM Database. For the remaining 15 genes, data regarding their specific expression, function, or detailed characterization are currently unavailable in the literature.

Regarding tissue distribution, most selected genes exhibited ubiquitous expression across adult human tissues. However, distinct tissue-specific patterns were observed:

- Neural Expression: PACC1 (proton-activated chloride channel), SEMA3A (neurosignaling molecule), and SV2A (synaptic vesicle glycoprotein) showed predominant expression in brain tissue.
- Developmental Specificity: KCTD12 (potassium channel) demonstrated strong expression in fetal tissues but was barely detectable in adult tissues, suggesting a potential role in developmental pathways or oncogenic reactivation (recapitulation of fetal programs).

The Role of TCHP (Mitostatin) as a Tumor Suppressor

A significant highlight of this selection is TCHP (Trichoplein/Mitostatin). According to Vecchione et al. (2009), TCHP acts as a potent keratin filament-binding protein. Functional studies have demonstrated that:

- Overexpression inhibits cell growth and colony formation while enhancing apoptotic responses.
- Knockdown leads to increased cell survival and proliferation.
- Clinical Correlation: Mitostatin expression is significantly reduced in advanced stages of primary bladder and breast tumors, strongly suggesting its role as a tumor suppressor in HNSCC as well.[39]

Oncogenic Signaling: The Ras and mTOR Pathways

Several identified genes are key players in the PI3K–AKT–mTOR pathway, the most frequently altered oncogenic signaling route in HNSCC (occurring in approximately 60% of cases).[40,41,42]

- RHEBL1: A member of the RAS family of small GTPases. It functions upstream of the mTOR pathway, promoting signaling that drives growth and transformation.[43] RAB11B: A member of the Rab11 GTPase subfamily involved in vesicular trafficking and recycling endosomes. Like RHEBL1, it belongs to the Ras superfamily, reinforcing the model's focus on GTPase-mediated signaling in HNSCC progression.

Structural and Functional Diversity

Beyond signaling, the model identified genes critical to cellular architecture and homeostasis:

- Cytoskeletal Dynamics: CAPG (actin modulator) and TCHP (regulator of ciliogenesis and mitochondrial morphology) are essential for maintaining structural integrity and responding to external stimuli. Experimental studies have suggested that TCHP may function as a tumor suppressor, as its overexpression inhibits cell growth and promotes apoptosis, whereas its downregulation increases cell proliferation and survival.[39] Reduced expression of this gene has been observed in several tumor types, including bladder and breast cancers, where it has been associated with more advanced tumor stages.
- Membrane Transport and Receptors: The selection included G-protein coupled receptors (GPR26, GPR32), a mitochondrial peptidase (IMMP2L), and ion channels such as PACC1 and KCTD12. These genes are involved in a variety of cellular processes including signal transduction, mitochondrial function, and ion transport, which are fundamental mechanisms that may influence tumor cell metabolism, communication, and survival.

Gene-Phenotype Relationships and Allelic Variants

Interestingly, several genes with high "importance weights" are associated with established clinical phenotypes:

- SEMA3A: The top-ranked gene by weight, associated with Kallmann Syndrome (hypogonadotropic hypogonadism).
- RAB11B: Linked to neurodevelopmental disorders involving ataxic gait and cortical white matter reduction.
- EIF2S3: Mutations in this core subunit of eukaryotic translation initiation factor-2 (eIF2) are responsible for MEHMO Syndrome, supporting its essential role in protein synthesis.
- ALDH7A1: Mutations in this lysine catabolism enzyme are associated with pyridoxine-dependent epilepsy.

The identification of these genes—many of which are involved in fundamental cellular survival, translation, and structural regulation—suggests that the Random Forest model has captured a complex signature that extends beyond simple tumor markers to the broader biological landscape of the patient.

3.2 **Unsupervised Learning Results**

The second stage of this study involved an unsupervised analysis conducted on a comprehensive cohort of 1,246 patients diagnosed with Head and Neck Squamous Cell Carcinoma (HNSCC). The distribution of patients according to primary tumor localization is summarized below:

- Oropharynx: 491 patients
- Oral Cavity: 358 patients
- Larynx: 255 patients
- Hypopharynx: 142 patients

The analytical workflow utilized two primary data domains: genomic and clinical. The genomic datasets comprise high-dimensional gene expression profiles, which served as the primary input for dimensionality reduction and clustering algorithms. The clinical dataset was utilized as a complementary reference to support the interpretation of genomic patterns. It provides a comprehensive overview of the patients' demographic profiles, clinical histories, and therapeutic interventions, allowing for the correlation of identified molecular clusters with real-world clinical characteristics.

3.2.1 **Genomic and Clinical Datasets**

The genomic data were organized into four distinct datasets, each corresponding to a specific primary tumor site. Each dataset followed a high-dimensional matrix structure, where 26,903 genes were represented as rows and individual patients as columns.

A significant methodological advantage of this study was the uniformity and integrity of the genomic data. Due to the standardized collection and processing protocols, the datasets were characterized by a complete absence of missing values. Consequently, no data imputation or cleaning procedures were required prior to the analysis. This high level of homogeneity ensured that the subsequent exploratory data analysis and unsupervised clustering techniques were performed on a robust and unbiased substrate, maximizing the reliability of

the identified molecular patterns.

In parallel, a comprehensive clinical dataset was processed, encompassing demographic profiles, medical histories, tumor-specific characteristics, and therapeutic interventions. Initially, this dataset comprised 1,541 subjects and over 500 variables.

To ensure methodological consistency and allow for multi-omic correlation, a rigorous alignment phase was performed. Only patients with corresponding genomic data were retained, resulting in a final cohort of 1,246 subjects. Furthermore, variables with significant missingness were excluded, yielding a refined clinical matrix of 46 high-quality variables.

The finalized clinical information was categorized into the following domains:

- Demographics and Lifestyle: Age, sex, and risk factors (e.g., smoking and alcohol consumption).
- Oncological Profile: Tumor characteristics including TNM staging (7th and 8th editions) and HPV/p16 status.
- Diagnostic and Pathological Data: Radiological imaging findings and post-operative pathological parameters.
- Therapeutic Regimens: Details on surgical interventions, radiotherapy, and chemotherapy.
- Follow-up: Longitudinal survival and recurrence data.

This clinical dataset served as the interpretative framework for the unsupervised genomic analysis. By mapping these clinical variables onto the genomic clusters, it became possible to validate the biological significance of the identified molecular patterns.

3.2.2 **Analysis of results**

3.2.2.1 *UMAP*

Once the datasets were finalized, UMAP was applied to the genomic data to explore potential differences in gene expression across the various tumor sites: Oral Cavity (OC), Hypopharynx (HY), Larynx (LA), and Oropharynx (OP).

The dimensionality reduction into a two-dimensional space provided a clear and interpretable representation of the genomic landscape, as shown in Figure 15.

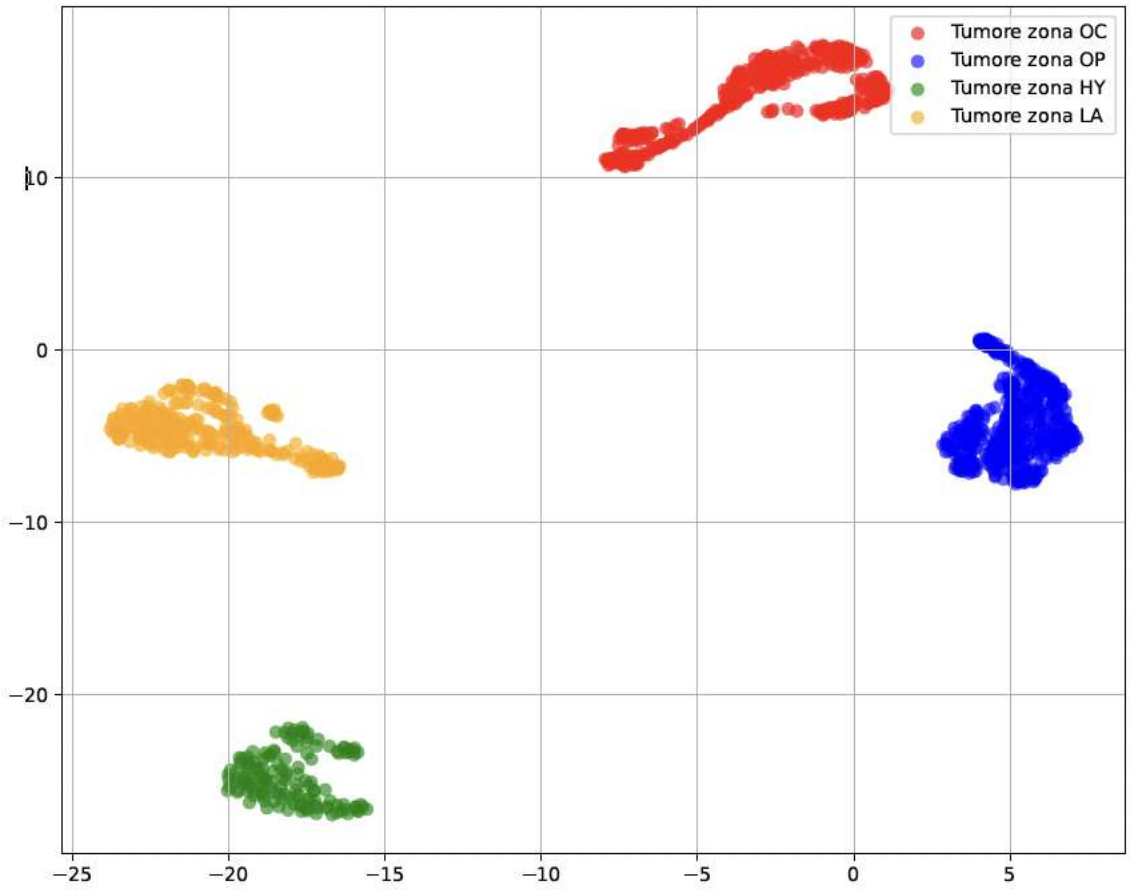

Figure 15. UMAP projection of genomic data.

Cluster Separation and Differentiation

The UMAP projection revealed the formation of four distinct and well-separated clusters, each representing a specific anatomical tumor site. Each cluster is composed exclusively of patients with tumors located in the same region, indicating a marked differentiation in genomic profiles based on anatomical origin.

To quantify the quality of this separation, we utilized two key metrics:

- Silhouette Score: A value of 0.87 was achieved, indicating high cluster density and minimal overlap between different groups.
- Adjusted Rand Index (ARI): The K-Means algorithm (with K=4) was applied to the projection, yielding an ARI of 1.

Validation and Agreement

An ARI of 1 represents perfect agreement between the true anatomical labels and the clusters predicted by the algorithm. This demonstrates that the gene expression signature is sufficiently unique for each site to allow the model to classify every patient correctly with no misclassifications.

This result was further validated by a confusion matrix, which confirmed total consistency between predicted and actual classes (Figure 16). This "perfect" clustering confirms that the anatomical site of the tumor is the primary driver of genomic variation in HNSCC, providing a robust baseline for more granular sub-phenotype analysis.

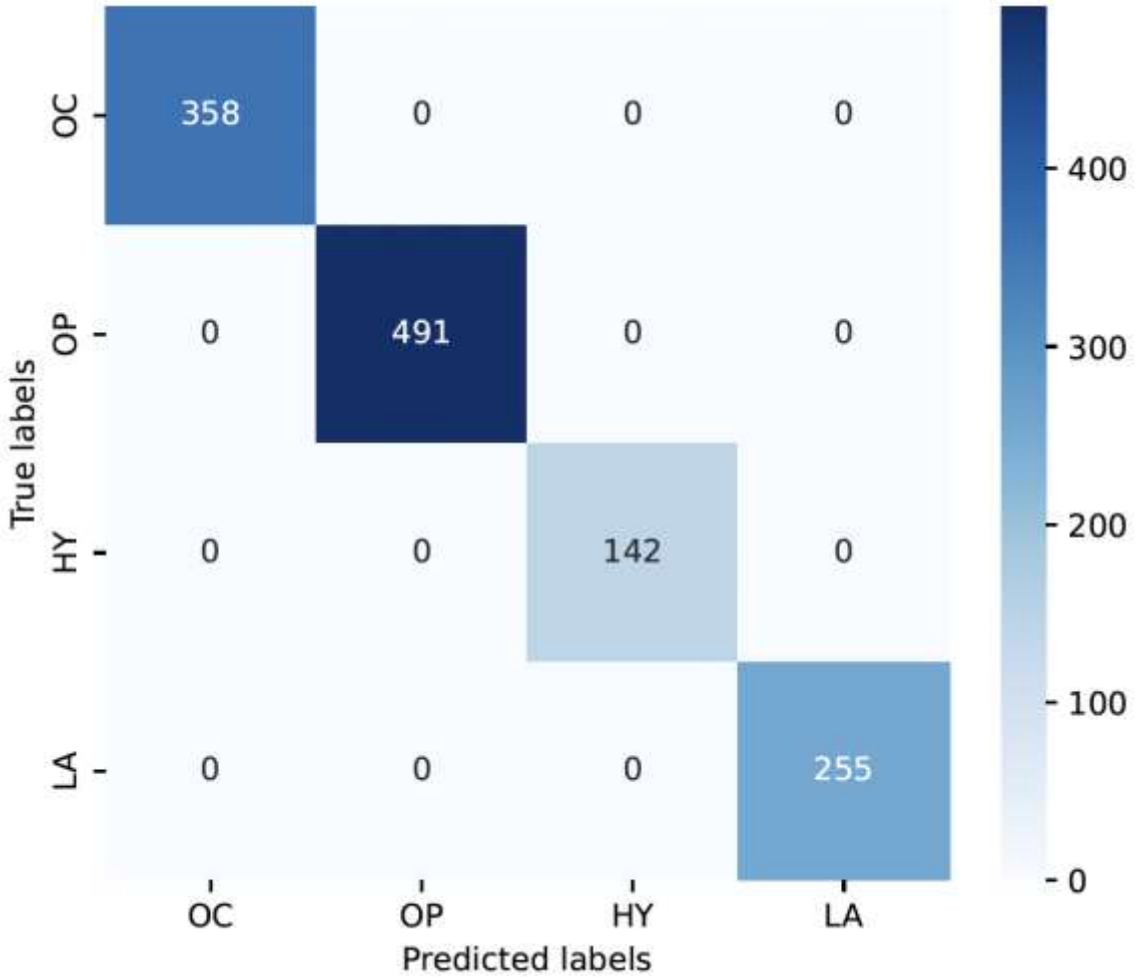

Figure 16. Confusion matrix obtained by comparing the true tumor labels with those predicted by K-Means clustering. The perfect alignment of the classes along the main diagonal confirms that each patient was correctly assigned to their respective tumor group, validating the effectiveness of the clustering.

3.2.2.2 *Clustering*

Following the successful separation of the four primary tumor sites, the next objective was to investigate whether identifiable subgroups of patients with shared genomic and clinical characteristics existed within each cluster. To achieve this, a Hierarchical Clustering approach was adopted—a method designed to reveal a nested structure of similarities among patients.

Methodology: Distance Metrics and Linkage

For each site-specific dataset, a square distance matrix was computed using Cosine Distance. This metric was selected for its efficacy in capturing patterns in high-dimensional gene expression data rather than just absolute magnitude. Subsequently, clustering was performed using the "Complete" linkage method. This is the most conservative hierarchical approach, as it determines the distance between two clusters based on the maximum distance between their most dissimilar members, ensuring compact and well-separated subgroups.

Visualization and Clinical Annotation

The results were visualized through dendrograms, where each leaf represents an individual patient. To interpret the biological and clinical relevance of the hierarchy, the dendrograms were annotated by color-coding the patient labels based on their clinical features. For each of the 46 available clinical columns:

1. Distinct categorical values (e.g., risk levels, hospital sites) were identified.
2. Unique color signatures were assigned to each value.
3. Patient labels were mapped to these colors to detect spatial "clusters" of specific clinical traits.

Key Findings: Oral Cavity Subset

While the analysis was applied across all datasets, the most significant and interpretable results emerged from the Oral Cavity cohort. Two specific variables provided the most profound insights into the genomic organization of these tumors:

- Study_Name: Representing the recruiting clinical centers, including UDUS, INT (Parma), AOP (Milan), BRESCIA, MAASTRO, and VUMC.
- INT_Lohavanichbutr_OSCC_HPVneg_risk: A specialized genomic risk signature for HPV-negative oral cavity cancers.[44]

The correlation between these variables and the genomic hierarchy allowed us to investigate whether patient grouping was driven by biological factors or potential multicenter variability (Figure 17).

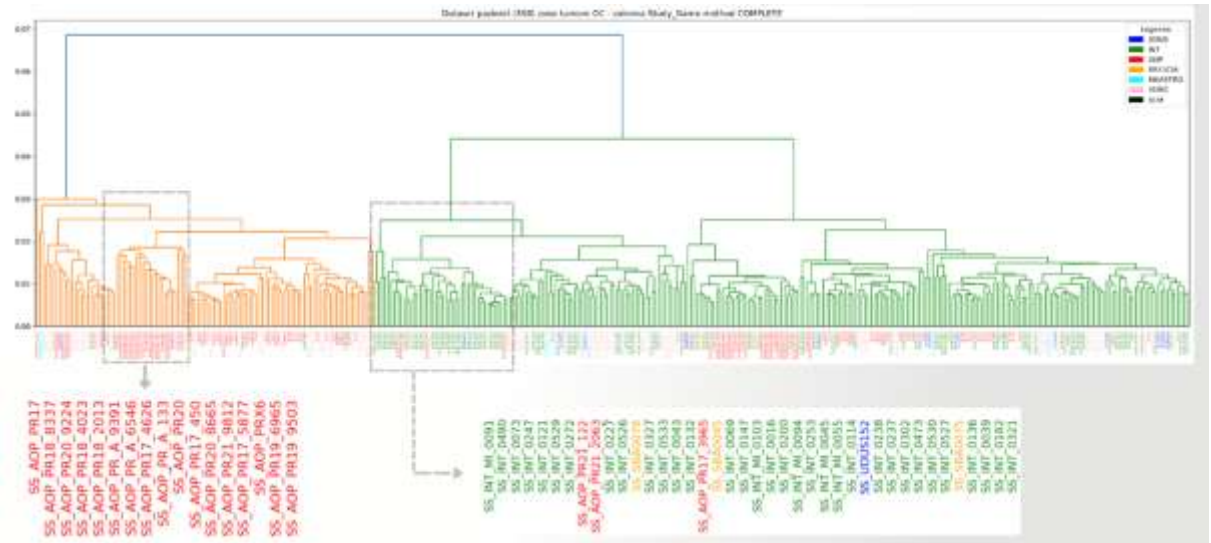

Figure 17. Dendrogram obtained through hierarchical clustering (Complete linkage method) for patients with oral cavity cancer (OC), with labels colored according to the recruiting hospital. Each color represents a different hospital (see legend). A non-random distribution of labels can be observed, with homogeneous color regions suggesting a possible bias due to environmental and social factors.

Centralization and Potential Bias Mitigation

It is essential to emphasize that while patients were recruited from various national and international facilities, all genomic sequencing and data processing were centralized at the

Milan center (INT/AOP). This centralized approach was strategically implemented to minimize experimental biases—commonly referred to as "batch effects"—which typically arise from variations in genomic extraction protocols and sequencing platforms across different institutions.

Institutional Clustering Patterns

Despite the centralized analysis, the dendrogram reveals distinct chromatic patterns within the two primary clusters (identified in orange and green). These patterns indicate a degree of genomic consistency among patients recruited from the same facility. Specifically:

- Patients from the Parma cohort (green) tend to aggregate within specific sub-clusters.
- Patients from the Milan (AOP) unit (red) and other participating centers exhibit similar grouping tendencies.

This non-random distribution suggests the presence of systematic differences among the patient cohorts. Such discrepancies may be attributed to a combination of factors:

1. Socio-Environmental Contexts: Variations in regional risk factors and lifestyles.
2. Clinical Specialization: The level of expertise and the specific multidisciplinary approach of each hospital in managing advanced HNSCC cases. Centers of excellence often treat a higher proportion of complex or late-stage cases, which is reflected in the underlying genomic landscape of their patient populations.

Integration of the Lohavanichbutr Risk Signature

To further validate the clinical relevance of the hierarchical structure, the INT-Lohavanichbutr-OSCC-HPVneg-risk signature was mapped onto the dendrogram. This specialized genomic signature, composed of 13 prognostic genes, was specifically developed to predict survival in HPV-negative Oral Cavity Squamous Cell Carcinoma (OSCC) patients.[44]

The signature categorizes patients into two distinct groups:

- High-Risk: Associated with poorer prognostic outcomes.
- Low-Risk: Associated with more favorable survival rates.

The alignment of this established 13-gene signature with our unsupervised clustering results provides a robust biological validation. It confirms that the clusters identified by the algorithm are not merely driven by institutional data fingerprints but are deeply connected to the molecular aggressiveness and prognostic profile of the tumors.

The hierarchical organization of the Oral Cavity dataset, as visualized in the dendrogram (Figure 18), revealed a clear partition of patients into two primary clusters (represented in orange and green). When the Lohavanichbutr 13-gene signature was overlaid onto these groups, a significant correlation emerged:

- High-Risk Cluster: Patients characterized by a "High" genomic risk signature demonstrated a strong tendency to aggregate within one specific primary cluster.
- Low-Risk Cluster: Conversely, patients with a "Low" risk signature were predominantly grouped within the alternative cluster.

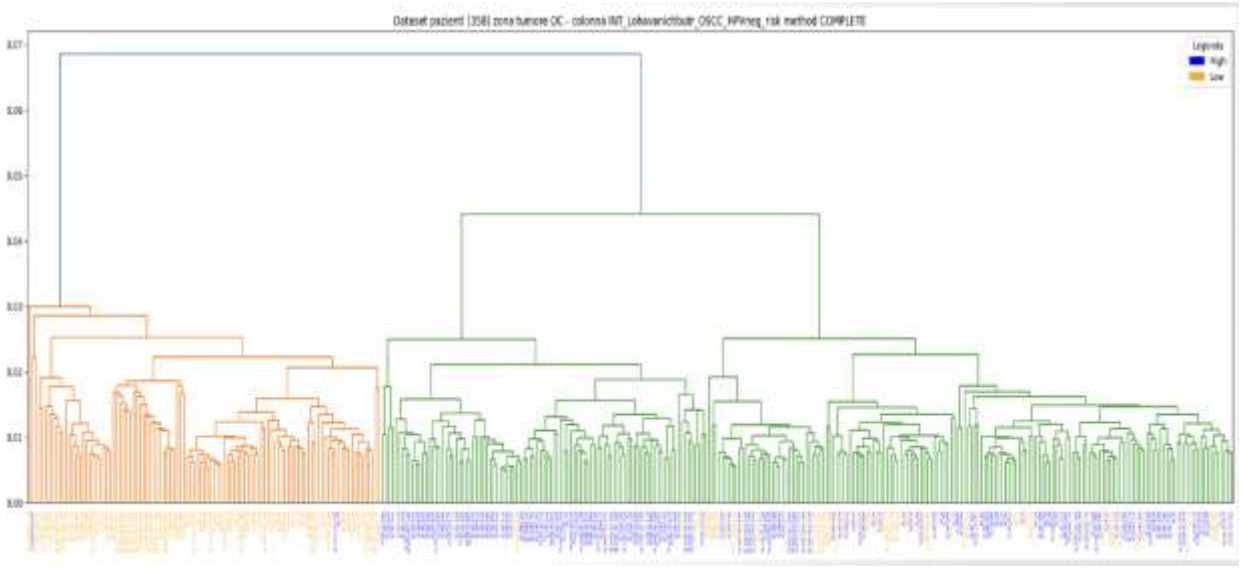

Figure 18. Hierarchical dendrogram of HPV-negative patients with oral cavity squamous cell carcinoma (OSCC). The two main clusters, highlighted in orange and green, show a clear separation based on the genomic risk signature. The orange cluster groups patients with a lower prognostic risk (Low), while the green cluster includes those with a higher risk (High).

This high degree of alignment between our unsupervised clustering and the established 13-gene prognostic model confirms that the 13 genes within the signature:

1. Effectively Discriminate Prognostic Profiles: They serve as robust biological markers that distinguish between patients with varying survival outcomes.
2. Capture Clinically Relevant Biology: The signature identifies authentic biological variations in disease progression, rather than being influenced by external data artifacts.
3. Support Patient Stratification: These findings validate the use of this genomic signature as a reliable tool for risk stratification, potentially guiding personalized treatment intensification or de-escalation strategies.

3.2.2.3 *Log Fold Change*

Following the stratification of the oral cavity carcinoma cohort into two distinct molecular clusters, a Log Fold Change (LFC) analysis was performed to identify the differentially expressed genes (DEGs) driving this separation.

For each gene, the LFC was calculated as the ratio of the mean expression levels between the two clusters, followed by a base-2 logarithmic transformation (log_2FC). This approach allows for a symmetric representation of up-regulation and down-regulation. Our analysis identified 86 genes with an absolute LFC value greater than 1 ($log_2FC > 1$), indicating a significant and substantial divergence in expression profiles between the two patient subgroups.

Biological Pathway Exploration

To translate this gene list into functional biological insights, a Functional Enrichment Analysis was conducted. This methodology allows for the systematic identification of:

- Biological Processes (BP): The broader "biological goals" the genes contribute to (e.g., cell proliferation or immune response).
- Molecular Functions (MF): The specific biochemical activities of the gene products (e.g., catalytic activity or receptor binding).
- Signaling Pathways: The coordinated molecular networks (e.g., KEGG or Reactome pathways) that are significantly associated with the 86 identified DEGs (Table 111).

This step was essential to determine if the molecular differences between the "High-Risk" and "Low-Risk" clusters correspond to specific hallmarks of cancer, such as increased metastatic potential, altered metabolism, or immune evasion.

No.	GENE	ENSEMBL_GENE_ID	No.	GENE	ENSEMBL_GENE_ID
1	ANXA1	ENSG00000135046	44	SPRR1A	ENSG00000143977
2	ATP5I	ENSG00000169020	45	SS18	ENSG00000169474
3	DST	ENSG00000151914	46	TCEB1	ENSG00000141380
4	CANX	ENSG00000127022	47	UBE2D3	ENSG00000154582
5	CAST	ENSG00000153113	48	DEK	ENSG00000109332
6	CASP4	ENSG00000196954	49	PTP4A2	ENSG00000124795
7	CDH1	ENSG00000039068	50	HIST1H4C	ENSG00000184007
8	FOXN3	ENSG00000053254	51	ATP5J2	ENSG00000197061
9	ATF2	ENSG00000115966	52	BZW1	ENSG00000101558
10	CSNK1A1	ENSG00000113712	53	N4BP2L2	ENSG00000241468
11	CSTB	ENSG00000164733	54	ATP5H	ENSG00000082153
12	DLG1	ENSG00000160213	55	ATP5L	ENSG00000244754
13	EEF1A1	ENSG00000075711	56	YWHAQ	ENSG00000167863
14	EIF2S3	ENSG00000156508	57	DSTN	ENSG00000167283
15	FABP5	ENSG00000130741	58	CBX3	ENSG00000134308
16	GJB2	ENSG00000164687	59	PRRC2C	ENSG00000125868
17	GTF3A	ENSG00000165474	60	UBR4	ENSG00000122565
18	HSPA9	ENSG00000122034	61	TRAM1	ENSG00000117523
19	IFI16	ENSG00000113013	62	SRRM2	ENSG00000127481
20	ITGB1	ENSG00000163565	63	SPATS2L	ENSG00000067167
21	KRT5	ENSG00000150093	64	RNU6-1	ENSG00000167978
22	LDHA	ENSG00000186081	65	TRAJ35	ENSG00000196141
23	COX1	ENSG00000134333	66	TMEM230	ENSG00000206625
24	NACA	ENSG00000198804	67	FERMT1	ENSG00000211854
25	NDUFB1	ENSG00000196531	68	ZDHHC7	ENSG00000089063
26	POLR2J	ENSG00000183648	69	CBWD1	ENSG00000101311
27	RMRP	ENSG00000005075	70	WLS	ENSG00000153786
28	RPL11	ENSG00000277027	71	DYNLRB1	ENSG00000172785
29	RPL27	ENSG00000142676	72	TMEM123	ENSG00000116729
30	RPL30	ENSG00000131469	73	RPL13AP20	ENSG00000125971
31	RPL34	ENSG00000156482	74	SMIM4	ENSG00000152558
32	RPL35A	ENSG00000109475	75	POLR2J3	ENSG00000234498
33	RPL37	ENSG00000182899	76	SCARNA7	ENSG00000168273
34	RPN1	ENSG00000145592	77	SNORD70	ENSG00000168255
35	RPS15A	ENSG00000163902	78	SNORD71	ENSG00000238741
36	RPS21	ENSG00000134419	79	SNORD93	ENSG00000212534
37	RPS24	ENSG00000171858	80	SNORD117	ENSG00000223224
38	RPS27	ENSG00000138326	81	SNORD105B	ENSG00000221740
39	S100A7	ENSG00000177954	82	RNU6-15P	ENSG00000201785
40	S100A8	ENSG00000143556	83	MIR548AQ	ENSG00000238531
41	S100A11	ENSG00000143546	84	RNU6-33P	ENSG00000207264
42	SMN1	ENSG00000163191	85	RNU6-39P	ENSG00000265470
43	SNRPG	ENSG00000172062	86	RNU6-8	ENSG00000207524

Table 111. List of the 86 differentially expressed genes between the two clusters (Oral cavity tumor)

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	UP_KW_PTM	Acetylation	RT		46	52,3	2,1E-15	3,6E-14
	GOTERM_CC_DIRECT	cytosolic ribosome	RT		11	12,5	2,8E-12	6,9E-10
	GOTERM_BP_DIRECT	translation	RT		13	14,8	5,2E-11	2,6E-8
	GOTERM_BP_DIRECT	cytoplasmic translation	RT		10	11,4	8,9E-11	2,6E-8
	GOTERM_MF_DIRECT	RNA binding	RT		25	28,4	3,1E-10	6,0E-8
	UP_KW_DISEASE	Diamond-Blackfan anemia	RT		6	6,8	3,6E-8	8,0E-7
	UP_KW_MOLECULAR_FUNCTION	Ribosomal protein	RT		10	11,4	6,1E-8	2,1E-6
	GOTERM_CC_DIRECT	extracellular exosome	RT		28	31,8	2,0E-8	2,5E-6
	GOTERM_MF_DIRECT	structural constituent of ribosome	RT		10	11,4	3,4E-8	3,3E-6
	UP_KW_MOLECULAR_FUNCTION	Ribonucleoprotein	RT		11	12,5	3,3E-7	5,7E-6
	KEGG_PATHWAY	Ribosome	RT		10	11,4	2,8E-7	3,3E-5
	GOTERM_CC_DIRECT	cytosolic large ribosomal subunit	RT		6	6,8	2,9E-6	2,3E-4
	KEGG_PATHWAY	Coronavirus disease - COVID-19	RT		10	11,4	4,2E-6	2,5E-4
	UP_KW_PTM	Isopeptide bond	RT		21	23,9	8,3E-5	5,5E-4
	UP_KW_PTM	Ubl conjugation	RT		26	29,5	9,6E-5	5,5E-4
	UP_KW_CELLULAR_COMPONENT	CH(0)	RT		4	4,5	2,4E-5	7,9E-4
	GOTERM_CC_DIRECT	nucleolus	RT		18	20,3	1,3E-5	8,3E-4
	GOTERM_CC_DIRECT	proton-transporting ATP synthase complex, coupling factor F(o)	RT		4	4,5	1,8E-5	9,0E-4
	UP_SEQ_FEATURE	CROSSLINK:Glycyl lysine isopeptide (Lys-Gly) (interchain with G-Cter in SUMO2); alternate	RT		10	11,4	2,3E-6	1,0E-3
	GOTERM_CC_DIRECT	cytosol	RT		39	44,3	3,4E-5	1,4E-3
	GOTERM_CC_DIRECT	proton-transporting ATP synthase complex	RT		4	4,5	4,1E-5	1,4E-3
	UP_KW_CELLULAR_COMPONENT	Cytoplasm	RT		39	44,3	8,8E-5	1,5E-3
	GOTERM_CC_DIRECT	cytoplasm	RT		40	45,5	4,8E-5	1,5E-3

Figure 19. Results of the functional enrichment analysis performed on the differentially expressed genes in the two identified clusters. The table shows the significantly enriched biological categories, including molecular processes, cellular components, and metabolic pathways, along with their respective significance values

The functional enrichment results were evaluated based on their statistical significance and the density of gene involvement. The findings are summarized in Figure 19, which displays the significance levels represented by both raw p-values and the Benjamini-Hochberg (FDR) correction.

The Benjamini correction was applied to control the False Discovery Rate (FDR), ensuring that the identified pathways are not the result of random statistical noise, a crucial step when analyzing high-dimensional genomic data. The most impactful biological terms were identified based on two primary criteria:

1. Statistical Robustness: Terms with the lowest adjusted p-values.
2. Biological Representation: Terms with the highest "gene count" (the number of differentially expressed genes belonging to that specific functional category).

These high-impact terms represent the fundamental mechanisms that differentiate the two patient clusters. Notably, one of the most prominent biological processes emerging from this analysis is acetylation.

The Significance of Acetylation

Acetylation is a vital post-translational modification that influences protein stability, localization, and function. In oncology, dysregulation of protein and histone acetylation is a well-documented driver of altered gene expression and metabolic reprogramming. The identification of acetylation as a top-tier enriched term suggests that the divergent clinical outcomes between the "High-Risk" and "Low-Risk" clusters may be deeply rooted in epigenetic modifications and proteomic signaling rather than just genomic mutations alone.

4 DISCUSSION

The results of this study highlight both the potential and the inherent challenges of integrating multi-omic data through machine learning (ML) to improve prognostic prediction in Head and Neck Squamous Cell Carcinoma (HNSCC). Overall, the findings support the value of data-driven approaches while emphasizing the critical role of dataset size, data quality, and methodological choices in determining model performance and interpretability.

4.1 Supervised Learning: Model Performance and Data Volume

A primary finding of this study is the positive relationship between dataset size and predictive performance. The improvement observed when transitioning from the AOP cohort to the larger European BD2Decide dataset reinforces the concept that data volume represents a key determinant of model accuracy in biomedical ML.

Among the evaluated models, Random Forest (RF) consistently outperformed the Artificial Neural Network (ANN), achieving higher accuracy and demonstrating greater robustness across all experimental settings. This result is consistent with existing literature, particularly in scenarios characterized by limited sample sizes and high-dimensional feature spaces (Breiman, 2001). While ANNs offer high representational capacity, their performance is strongly dependent on large datasets, and in this context, their potential was likely constrained by the limited number of available samples. In contrast, Random Forest (RF), which combines multiple decision trees built on different subsets of the data and variables, reduces overfitting and provides more stable predictions while requiring relatively few parameter adjustments.

It is also important to note that the the prediction task was defined as a binary classification problem (“low risk” corresponding to survival at 5 years and “high risk” to death within 5 years), which may oversimplify the complexity of clinical outcomes. Prognosis in oncology is inherently dynamic and multifactorial; therefore, future studies may benefit from alternative approaches such as time-to-event modeling or multi-class risk stratification to better capture disease heterogeneity and improve model performance.

Achieving highly accurate predictive models is of critical importance in oncology, as improved tumor stratification directly impacts clinical decision-making, with potential benefits ranging from increased overall survival to improved quality of life. The ultimate goal of refined stratification and staging systems is to reduce overtreatment in low-risk patients while enabling treatment intensification and closer follow-up in high-risk individuals. In this context, even modest improvements in predictive accuracy may translate into clinically meaningful differences, particularly in borderline cases where treatment decisions remain uncertain.

Supervised learning methods offer the advantage of producing clinically interpretable predictions aligned with established medical classifications; however, they require large, well-annotated datasets, and their performance is strongly dependent on the quality of the input labels. These models learn relationships between patient features (clinical, genomic, and radiomic variables) and known outcomes, enabling risk prediction for new patients.[45,46,47] Among supervised approaches, Random Forest has proven particularly effective for biomedical applications due to its ability to model complex, non-linear relationships, maintain robustness with relatively small datasets, and provide estimates of feature importance.[48]

Although the predictive accuracy achieved by the models remains moderate, for above mentioned reasons, the identification of relevant features highlights the value of ML approaches in exploring the biological basis of tumor progression and prognosis.

4.2 Clinical and Functional Predictors: Beyond TNM

The feature importance analysis derived from the RF model provides a meaningful link between computational outputs and clinical interpretation, enhancing model transparency and supporting its potential integration into clinical decision-making processes.

While the 8th Edition TNM classification emerged as a stronger predictor than the previous version, age, smoking history and alcohol consumption showed high prognostic weight. The identification of pre-treatment deglutition (swallowing) scores as a top-ranked variable is particularly noteworthy. This finding suggests that functional status and quality-of-life indicators at diagnosis may reflect underlying disease aggressiveness and patient resilience, rather than representing purely secondary outcomes. Similarly, the high importance of hematological parameters—such as hemoglobin (HB) and platelet (PLT) levels—highlights the possible role of systemic inflammation and overall physiological condition in influencing long-term survival. The identification of non-tumor-specific variables among top predictors suggests that patient-related factors play a critical role in determining outcomes and should be systematically incorporated into future prognostic models.

These results support a more comprehensive view of cancer prognosis, in which tumor characteristics and host-related factors jointly contribute to clinical outcomes.

4.3 Genomic and Biological Relevance

By highlighting genomic features that consistently contribute to predictive performance, RF models can provide valuable information on the molecular mechanisms driving tumor development and key oncogenic pathways.[48,49] The genes identified by RF were involved in a broad spectrum of biological processes, including signal transduction, cytoskeletal organization, vesicular trafficking, mitochondrial function, and cellular metabolism. The identification of genes associated with well-established oncogenic pathways, such as the PI3K–AKT–mTOR signaling axis, further supports the biological relevance of the selected features.[50] The convergence between statistically relevant features and well-established biological mechanisms suggests that the model may be capturing meaningful patterns rather than purely random associations.

At the same time, the presence of poorly characterized genes suggests the potential of ML approaches to identify novel molecular candidates for further functional investigation. These findings warrant further experimental validation to clarify their role in tumor biology and their potential clinical utility.

While supervised models identify features associated with outcome prediction, unsupervised approaches provide complementary insights into the intrinsic structure and heterogeneity of the data.

4.4 Unsupervised Learning: Global Structure of the Dataset

The results obtained from unsupervised ML approaches provided complementary insights into the structure of the genomic data and further emphasize the importance of integrating

genomic and clinical data for prognostic stratification in OSCC. Techniques such as UMAP and hierarchical clustering enabled the identification of distinct molecular subgroups, reflecting a level of biological heterogeneity not captured by traditional TNM staging systems. This observation is consistent with previous literature highlighting the limitations of TNM classification in predicting individual patient outcomes, particularly in advanced disease.[51,52,53]

The UMAP projection revealed a clear separation of tumors according to anatomical site (ARI = 1.0), strongly suggesting that tumor origin is a major driver of global transcriptomic variation in HNSCC. However, this result should be interpreted with caution, as it may also reflect dataset-specific structure or underlying biases. The UMAP analysis demonstrated also that OSCC exhibits greater heterogeneity compared to other head and neck tumor types. An unexpected finding was the absence of clear separation between HPV-positive and HPV-negative tumors. This observation contrasts with established clinical evidence and may indicate that HPV-related transcriptional differences are more subtle, context-dependent, or not fully captured by the selected feature space. Alternatively, this result may be influenced by dataset-specific factors such as sample imbalance, feature representation, or the dimensionality reduction process, and therefore requires further investigation.

4.5 Intratumoral Heterogeneity and Molecular Subgroups

Hierarchical clustering further revealed structured subgroups within the oral cavity dataset. The partial aggregation of patients according to their center of origin suggests the presence of systematic differences between cohorts. While these patterns may reflect true biological or socio-environmental variability, the potential contribution of residual batch effects—despite centralized data processing—cannot be excluded.

The strong concordance between the identified clusters and the Lohavanichbutr 13-gene prognostic signature provides external validation for the biological and clinical relevance of the clustering results.[44]

The Log Fold Change (LFC) analysis identified 86 differentially expressed genes between the two main clusters. Functional enrichment analysis of these genes revealed involvement in key biological processes, including protein translation, ribosome regulation, and acetylation, highlighting molecular mechanisms potentially associated with distinct clinical phenotypes. Notably, acetylation-related genes emerged among the most prominent features within the differentially expressed set, suggesting a potential role of epigenetic regulation in driving differences in tumor behavior and patient outcomes. This observation is consistent with growing evidence supporting the importance of epigenetic mechanisms in cancer progression.[54,55]

A potential future direction of this work involves extending clustering analyses to identify a greater number of subgroups, thereby enabling finer stratification of patient populations. This approach may allow the identification of more homogeneous molecular subsets and improve the resolution of biological interpretation.

4.6 **Limitations and Future Directions**

Overall, these findings support the concept that a multidimensional and personalized approach represents the future of OSCC management, in alignment with current trends in precision oncology.

Despite the insights provided, the predictive performance of the models remains moderate (approximately 60–70%). This performance plateau is likely attributable to the intrinsic

biological heterogeneity of HNSCC, the limited sample size, and the high dimensionality of genomic data.[56,57] However, these results should be interpreted as a proof-of-concept, demonstrating the feasibility and potential value of integrating machine learning approaches into prognostic workflows.[58]

Several limitations should be acknowledged, including the lack of external validation on independent cohorts and the potential influence of residual biases in multi-center datasets. Furthermore, the binary formulation of the prediction task may have limited the ability to fully capture the complexity of patient outcomes.[59,60]

Future research should focus on expanding cohort size, improving data harmonization, and validating the proposed models in independent and geographically diverse populations. The adoption of time-to-event modeling and the deeper integration of multi-modal data, including radiomic features, may further enhance predictive performance.

In addition, the clinical implementation of these approaches remains challenging due to the cost and complexity of genomic analyses. A key objective for future work will be the development of simplified and clinically actionable biomarker panels, as well as their integration into decision support systems such as those envisioned within the BD2Decide project.

5 CONCLUSION

This thesis investigated the application of machine learning approaches for risk stratification and clustering in head and neck squamous cell carcinoma (HNSCC), with a particular focus on oral cavity tumors. The primary objective was to explore whether the integration of multi-omic data could improve prognostic assessment beyond conventional staging systems, aiming to contribute to the growing field of precision oncology and highlighting the potential of data-driven methods to capture tumor heterogeneity and support more personalized clinical decision-making.

Overall, the findings demonstrate that integrative, data-driven approaches can provide meaningful insights into the prognostic landscape of HNSCC and OSCC. By combining heterogeneous data sources, machine learning models are able to capture complex interactions between tumor biology and host-related factors, moving beyond the limitations of traditional classification systems toward more refined and personalized prognostic frameworks.

Supervised learning models were employed due to the availability of labeled clinical outcomes and easier clinical interpretability. Among the tested approaches, Random Forest demonstrated superior performance compared to Artificial Neural networks, achieving higher predictive accuracy and greater robustness, particularly in the context of limited sample size and high-dimensional data. Nevertheless, both models showed moderate accuracy, reflecting the intrinsic complexity of the disease and the limitations of the available datasets. Despite these limitations, the identification of clinically and biologically relevant features, such as

genes associated with key oncogenic pathways, indicates that RF can provide meaningful insights into the mechanisms underlying tumor progression and patient prognosis. These findings support the potential of machine learning not only as a predictive tool but also as a means of generating biologically interpretable hypotheses and discovering potential biomarker.

The unsupervised analyses further revealed substantial biological heterogeneity within HNSCC, supporting the need for more refined and personalized stratification systems. While tumor anatomical site emerged as a major driver of global genomic variation, substantial variability was observed within individual tumor groups, particularly in OSCC. The identification of molecular subgroups and particularly the concordance between the identified clusters with a known prognostic gene signature for HNSCC reinforces the need for more granular stratification strategies that extend beyond traditional TNM staging systems.

Overall, these results underline that the integration of machine learning techniques with multidimensional clinical and genomic data represents a promising strategy for improving risk stratification in oral cancer. Future research should focus on expanding dataset size, validating models in independent cohorts, and incorporating additional data types, such as radiomic features, to enhance predictive performance.

Ultimately, the development of robust, interpretable, and clinically applicable predictive models may contribute to optimizing treatment strategies, reducing overtreatment, and improving patient outcomes, paving the way toward a more personalized and precise approach to cancer management.

In conclusion, this work supports the growing role of machine learning and multi-omic integration in precision oncology. By enabling a more comprehensive characterization of tumor behavior and patient-specific factors, these approaches have the potential to improve risk stratification, guide treatment decisions, and ultimately contribute to more personalized and effective cancer management strategies.

REFERENCES:

1. Keek SA, Wesseling FWR, Woodruff HC, et al. A prospectively validated prognostic model for patients with locally advanced squamous cell carcinoma of the head and neck based on radiomics of computed tomography images. *Cancers*. 2021;13(13):3271. doi:10.3390/cancers13133271
2. Krolls SO, Hoffman S. Squamous cell carcinoma of the oral soft tissues: a statistical analysis of 14,253 cases by age, sex and race of patients. *J Am Dent Assoc.* 1976 Mar;92(3):571-574. doi:10.14219/jada.archive.1976.0556
3. Jagadeesan D, Sathasivan K, Fuloria NK, et al. Comprehensive insights into oral squamous cell carcinoma: diagnosis, pathogenesis, and therapeutic advances. *Pathol Res Pract.* 2024;261:155489. doi:10.1016/j.prp.2024.155489. Epub 2024 Jul 26
4. Ghaderi H, Roshan-Zamir M, Jafarinia M, Kruger E. Oral squamous cell carcinoma: focus on biomarkers for screening. *J Dent (Shiraz).* 2024 Mar;25(1):1-16. doi:10.30476/dentjods.2023.96159.1924. eCollection 2024 Mar
5. Christine GM. Critical changes in the staging of head and neck cancer. *Radiol Imaging Cancer.* 2020 Jan 31;2(1):e190022. doi:10.1148/rycan.2019190022. eCollection 2020 Jan
6. Kreppel M, Nazarli P, Grandoch A, Safi AF, Zirk M, Nickenig HJ, Scheer M, Rothamel D, Hellmich M, Zöller JE. Clinical and histopathological staging in oral squamous cell carcinoma — comparison of the prognostic significance. Oral Oncol. 2016;60:68-73. doi:10.1016/j.oraloncology.2016.07.004
7. Amin MB, Greene FL, Edge SB, et al. The eighth edition AJCC cancer staging manual: continuing to build a bridge from a population-based to a more “personalized” approach

to cancer staging. *CA Cancer J Clin.* 2017 Mar;67(2):93-99. doi:10.3322/caac.21388. Epub 2017 Jan 17

8. Mes SW, Beest D te, Poli T, et al. Prognostic modeling of oral cancer by gene profiles and clinicopathological co-variables. *Oncotarget.* 2017;8(35):59312-59323. doi:10.18632/oncotarget.19576
9. Hashim D, Genden E, Posner M, Hashibe M, Boffetta P. Head and neck cancer prevention: from primary prevention to impact of clinicians on reducing burden. *Ann Oncol.* 2019;30(5):744-756. doi:10.1093/annonc/mdz084
10. Kato MG, Baek CH, Chaturvedi P, et al. Update on oral and oropharyngeal cancer staging – international perspectives. *World J Otorhinolaryngol Head Neck Surg.* 2020;6(1):66-75. doi:10.1016/j.wjorl.2019.06.001
11. Glastonbury CM. Critical changes in the staging of head and neck cancer. *Radiol Imaging Cancer.* 2020 Jan 31;2(1):e190022. doi:10.1148/rycan.2019190022
12. Mattavelli D, Ferrari M, Taboni S, et al. The 8th TNM classification for oral squamous cell carcinoma: what is gained, what is lost, and what is missing. *Oral Oncol.* 2020;111:104937. doi:10.1016/j.oraloncology.2020.104937
13. Perlangeli G, Lilloni G, Salti G, Ferri A, Ferrari S, Poli T. The ability of the eighth edition of the TNM staging system plus minor invasion criteria to predict the biological behaviour of oral cavity carcinomas. *J Oral Pathol Med.* 2023 Sep;52(8):746-750. doi:10.1111/jop.13469. Epub 2023 Aug 1
14. Tehzeeb H, Hande A, Patil S, Sonone A, Pakhale A, Chavhan A. Correlation of clinical and pathological TNM staging with histopathological grading in oral squamous cell carcinoma. *Cureus.* 2024 May 23;16(5):e60912. doi:10.7759/cureus.60912

15. Ritter A, Yosefof E, Edri N, et al. Reappraising the TNM staging system for oral cavity squamous cell carcinoma: an age-related prognosis analysis. *Eur Arch Otorhinolaryngol.* 2025 Apr 25. doi:10.1007/s00405-025-09385-x. Online ahead of print
16. Taghavi N, Yazdi I. Prognostic factors of survival rate in oral squamous cell carcinoma: clinical, histologic, genetic and molecular concepts. *Arch Iran Med.* 2015 May;18(5):314-319
17. Siriwardena BSMS, Karunathilaka HDNU, Kumarasiri PVR, Tilakaratne WM. Impact of histological and molecular parameters on prognosis of oral squamous cell carcinoma: analysis of 290 cases. *Biomed Res Int.* 2020;2020:2059240. doi:10.1155/2020/2059240. eCollection 2020
18. Human papillomavirus as a driver of head and neck cancers. *Br J Cancer.* Accessed May 24, 2024. https://www.nature.com/articles/s41416-019-0602-7
19. Du Toit DF. The tongue: structure and function relevant to disease and oral health. *SADJ.* 2003 Oct;58(9):375-376, 380-383
20. Byers RM, El-Naggar AK, Lee YY, et al. Can we detect or predict the presence of occult nodal metastases in patients with squamous carcinoma of the oral tongue? *Head Neck.* 1998;20(2):138-144. doi:10.1002/(SICI)1097-0347(199803)20:2<138::AID-HED7>3.0.CO;2-3
21. Dijkland SA, Helmrich IRAR, Steyerberg EW. Validation of prognostic models: challenges and opportunities. *J Emerg Crit Care Med.* 2018;2(0). doi:10.21037/jeccm.2018.10.10
22. López Pérez L, Hernandez L, Ottaviano M, et al. BD2Decide: big data and models for personalized head and neck cancer decision support. 2019;68. doi:10.1109/CBMS.2019.00024

23. Overview | BD2DECIDE. Accessed May 29, 2024. https://www.bd2decide.eu/content/overview
24. Resteghini C, Trama A, Borgonovi E, et al. Big data in head and neck cancer. *Curr Treat Options Oncol.* 2018;19(12):62. doi:10.1007/s11864-018-0585-2
25. OraMod Project: helping clinicians predict oral cavity cancer reoccurrence. Accessed May 28, 2024. https://www.vph-institute.org/news/oramod-project-helping-clinicians-predict-oral-cavity-cancer-reoccurrence.html
26. Big data and models for personalized head and neck cancer decision support | BD2Decide | Project | Results | H2020 | CORDIS | European Commission. Accessed May 21, 2024. https://cordis.europa.eu/project/id/689715/results/it
27. Consortium Partners | BD2DECIDE. Accessed May 29, 2024. https://www.bd2decide.eu/content/consortium-partners
28. Tsai CJ, Riaz N, Gomez SL. Big data in cancer research: real-world resources for precision oncology to improve cancer care delivery. *Semin Radiat Oncol.* 2019;29(4):306-310. doi:10.1016/j.semradonc.2019.05.002
29. Yang Y. Brief introduction of medical database and data mining technology in big data era. *J Evid Based Med.* 2020. Accessed May 31, 2024. https://onlinelibrary.wiley.com/doi/10.1111/jebm.12373
30. Dixit S, Kumar A, Srinivasan K. A current review of machine learning and deep learning models in oral cancer diagnosis: recent technologies, open challenges, and future research directions. *Diagnostics.* 2023;13(7):1353. doi:10.3390/diagnostics13071353
31. Sah S. Machine learning: a review of learning types. 2020. doi:10.20944/preprints202007.0230.v1

32. Poli T. Big data and models for personalized head and neck cancer decision support (BD2Decide). *ClinicalTrials.gov*; 2016. Accessed Jan 1, 2024. https://clinicaltrials.gov/study/NCT02832102
33. Big data and models for personalised head and neck cancer decision support (BD2Decide) – DigitalHealthEurope. Accessed Jun 3, 2024. https://digitalhealtheurope.eu/catalogue/big-data-and-models-for-personalised-head-and-neck-cancer-decision-support-bd2decide/
34. Cavalieri S, De Cecco L, Brakenhoff RH, et al. Development of a multiomics database for personalized prognostic forecasting in head and neck cancer: the Big Data to Decide EU Project. *Head Neck.* 2021;43(2):601-612. doi:10.1002/hed.26515
35. Big data and models for personalized head and neck cancer decision support – *Description of Actions Document, BD2Decide Project*
36. Cavalieri S, Serafini MS, Carenzo A, et al. Clinical validity of a prognostic gene expression cluster-based model in human papillomavirus–positive oropharyngeal carcinoma. *JCO Precis Oncol.* 2021;(5):1666-1676. doi:10.1200/PO.21.00094
37. 2024 ICD-10-CM Codes C06*: Malignant neoplasm of other and unspecified parts of mouth. Accessed Jun 4, 2024. https://www.icd10data.com/ICD10CM/Codes/C00-D49/C00-C14/C06-
38. Contrera KJ, Zafereo ME, Yaniv D, et al. Outcomes for recurrent oral cavity squamous cell carcinoma. *Oral Oncol.* 2022; [Epub ahead of print]. doi:10.1016/j.oraloncology.2022.106127

39. Vecchione A, Fassan M, Anesti V, et al. MITOSTATIN, a putative tumor suppressor on chromosome 12q24.1, is downregulated in human bladder and breast cancer. Oncogene. 2009;28:257-269. doi:10.1038/onc.2008.381

40. Marquard FE, Jücker M. PI3K/AKT/mTOR signaling as a molecular target in head and neck cancer. *Biochem Pharmacol.* 2020 Feb;172:113729. doi:10.1016/j.bcp.2019.113729. Epub 2019 Nov 27

41. Wang Z, Goto Y, Allevato MM, et al. Disruption of the HER3-PI3K-mTOR oncogenic signaling axis and PD-1 blockade as a multimodal precision immunotherapy in head and neck cancer. *Nat Commun.* 2021;12:2383. doi:10.1038/s41467-021-22619-w

42. Gougis P, Moreau Bachelard C, Kamal M, Gan HK, Borcoman E, Torossian N, Bièche I, Le Tourneau C. Clinical development of molecular targeted therapy in head and neck squamous cell carcinoma. JNCI Cancer Spectr. 2019;3(4):pkz055. doi:10.1093/jncics/pkz055

43. Tee AR, Blenis J, Proud CG. Analysis of mTOR signaling by the small G-proteins, Rheb and RhebL1. FEBS Lett. 2005 Aug 29;579(21):4763-4768. doi:10.1016/j.febslet.2005.07.054

44. Lohavanichbutr P, Méndez E, Holsinger FC, Rue TC, Zhang Y, Houck J, Upton MP, Futran N, Schwartz SM, Wang P, Chen C. A 13-gene signature prognostic of HPV-negative OSCC: discovery and external validation. Clin Cancer Res. 2013 Mar 1;19(5):1197-1203. doi:10.1158/1078-0432.CCR-12-2647. Epub 2013 Jan 14

45. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer; 2009

46. Alaa AM, van der Schaar M. AutoPrognosis: automated clinical prognostic modeling via Bayesian optimization. Proceedings of Machine Learning Research. 2018;80:139-158. https://proceedings.mlr.press/v80/alaa18b.html
47. Azuaje F. Artificial intelligence for precision oncology: beyond patient stratification. npj Precis Oncol. 2019;3:14. https://www.nature.com/articles/s41698-019-0078-1
48. Breiman L. Random forests. Machine Learning. 2001;45(1):5-32. https://link.springer.com/article/10.1023/A:1010933404324
49. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J. 2015;13:8-17. doi:10.1016/j.csbj.2014.11.005
50. Yuan TL, Cantley LC. PI3K pathway alterations in cancer: variations on a theme. Oncogene. 2008;27(41):5497-5510. doi:10.1038/onc.2008.245
51. Zhao Y, Zhang X, Li J, Wang H. Integrative multi-omics analysis and experimental validation identify molecular subtypes, prognostic signature, and CA9 as a therapeutic target in oral squamous cell carcinoma. Oral Oncol. 2025; [Epub ahead of print]. doi:10.1016/j.oraloncology.2025.105488
52. Singh MP, Rai S, Kumar V. Unsupervised machine learning-based clustering identifies unique molecular signatures of colorectal cancer with distinct clinical outcomes. Genes Dis. 2023;10(1):45-57. doi:10.1016/j.gendis.2022.12.003
53. Cao J, Li H, Chen S, Zhang Y. Unsupervised hierarchical clustering identifies immune-related gastric cancer subtypes with distinct prognoses. Front Pharmacol. 2021;12:692454. doi:10.3389/fphar.2021.692454

54. Dawson MA, Kouzarides T. Cancer epigenetics: from mechanism to therapy. Cell. 2012;150(1):12-27. doi:10.1016/j.cell.2012.06.013
55. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. Science. 2013;339(6127):1546-1558. doi:10.1126/science.1235122
56. Yeghaian M, et al. Machine learning-based treatment outcome prediction in head and neck cancer using integrated noninvasive diagnostics. Int J Comput Assist Radiol Surg. 2025; [Epub ahead of print]
57. Zhou Y, et al. Advances in cancer through machine learning models. Appl Sci. 2025; [Epub ahead of print]
58. Tam SY, Tang FH, Chan MY, et al. Prognosis prediction in head and neck squamous cell carcinoma by radiomics and clinical information. Biomedicines. 2024;12(8):1646. doi:10.3390/biomedicines12081646
59. Alabi RO, Guntinas-Lichius O, Elmusrati M, et al. Machine learning for survival outcome in head and neck squamous cell carcinoma: a multicenter validation study. Sci Rep. 2026;16:254. doi:10.1038/s41598-025-29295-6
60. Dhiman P, Ma J, Andaur Navarro CL, et al. Risk of bias of prognostic models developed using machine learning: a systematic review in oncology. Diagn Progn Res. 2022;6:13. doi:10.1186/s41512-022-00126-w