POLITECNICO DI TORINO

FACOLTÀ DI INGEGNERIA

CORSO DI LAUREA IN
INGEGNERIA BIOMEDICA

Tesi di Laurea Magistrale

Multi-Modal Deep Learning for Time-to-Event analysis of Head and Neck Squamous Cell Carcinoma patients

Relatore:

Prof. Filippo Molinari

Co-relatori:

M.Sc. Adam Hilbert

Dr. med. Julian Weingärtner

Dr. med. Sebastian Zschaeck

Candidato:

Beatrice Cattaneo

Matr. 289775

Dicembre 2023

Summary

The objective of the following Thesis is to conduct an explorative study on the implementation of multi-modal Deep Learning techniques for the Time to Event analysis in patients with Head and Neck Squamous Cell Carcinoma.
While the first introductory chapter depicts the clinical framework within which this research is situated, the second chapter describes the technical aspects of Time to Event analysis and its methods of implementation.
Together, these first two chapters aim to provide the clinical and technical context necessary for a deep understanding of the role that this methodology could play in clinical practice.
After the third chapter, where our materials and methods are extensively described, the fourth chapter presents and discusses our results. The different approaches implemented in this study are compared and analysed to establish a meaningful interpretation of the achievements of this method.
Lastly, in the final chapter our conclusions are presented.
The present research was conducted at Berlin Charitè Lab of Artificial Intelligence in Medicine in collaboration with M.Sc. Adam Hilbert, Dr. med. Julian Weingärtner and Dr. med. Sebastian Zschaeck.

Index

INTRODUCTION

Head and Neck Squamous Cell Carcinoma (HNSCC) constitutes the seventh most common cancer diagnosis worldwide, and the long-term survival of the affected patients is highly influenced by the possible development of distant metastasis and tumor relapse. Hence, a prognostic model able to predict such occurrences would significantly benefit these patients and could be employed for treatment recommendation in order to optimize the handling of the individual subject.

For such purpose, Machine Learning could find a good fit. This technology is a subset of Artificial Intelligence that focuses on the development of algorithms able to learn from input data to generate decisions or predictions on new, unseen samples. Deep Learning, in turn, is a Machine Learning approach that employes Artificial Neural Networks and has the potential to identify and interpret even more complex relationships within the data.

In the depicted context, this Thesis aims at describing the development of a multi-modal Deep Learning model for Time to Event analysis in HNSCC patients; our approach involves the employment of clinical and imaging data for the prediction of distant metastasis, loco-regional failure, and overall survival of the individual patient.

In the context of automated models for clinical outcome prediction, our work involves the innovation given by the use of a combination of clinical and imaging (i.e., CT and 18FDG-PET) data. While the latter are universally known to hold a significant amount of information in the investigation of tumors, in this study clinical data are thought to provide further knowledge to the conditions of the patient, and their prognostic power is therefore assessed and employed for the predictions. Therefore, a substantial weight is given to pre-treatment clinical

variables and the investigation of their influence on the future evolution of the disease.

Moreover, this study aims at investigating the predictive power of CT and PET volumes without primary and lymph node Gross Tumor Volume segmentation; indeed, our goal is the development of an automated model that allows to disengage from a manual segmentation while granting an equally satisfactory performance.

The ultimate goal is to achieve a model able to accurately predict the timing of HNSCC-related events for a single patient, in order to contribute to the goal of personalized medicine by allowing, with such a technology, a step closer to the development of individualized therapies.

The findings presented in this Thesis provide valuable information concerning the prognostic power of clinical and imaging data in the progression of HNSCC, together with notable insights resulting from the comparison of different DL-based prediction models.

CHAPTER 1
CLINICAL FRAMEWORK

1.1. Introduction to Head and Neck Squamous Cell Carcinoma

1.1.1. Tumor properties and oncogenesis

Head and Neck Squamous Cell Carcinoma (HNSCC) is a class of tumors that constitutes the seventh most common cancer diagnosis worldwide [1]. It refers to malignancies affecting the oral and nasal cavity, larynx, oropharynx, and hypopharynx.

HNSCC arises from the squamous cells lining the tissue – the mucosal epithelium - of different head and neck regions. From the histological point of view, the progression to HNSCC follows a specific series of steps:

- epithelial cell hyperplasia: the epithelial tissue enlarges due to the higher reproduction rate of its cells; while the number of cells increased significantly, their structure and organization have not changed.
- dysplasia: abnormal development of tissue structure; cells present irregular shape, size, and organization.
- carcinoma in situ: anomalous (i.e., cancer) cells are confined to their origin site; this is an early stage of the development of the tumor.

- invasive carcinoma: cancer cells penetrate through the surrounding tissue, reaching healthy regions outside the origin site.

As HNSCC is a category that contains tumors affecting different anatomical locations (i.e., oral cavity, larynx, pharynx, etc.), the cell type of origin strongly depends on the affected site as well as the cause of the cancer. However, the cell of origin is usually found in normal adult stem cells; subsequently, the cell will transform, through the described oncogenic steps, into a cancer stem cell (CSC). Such CSCs are characterized by self-renewal and pluripotency, which respectively refer to the ability to give rise to more stem cells of the same kind and further develop to different types of cells.

HNSCC was found to be genetically unstable: this property was extensively researched in order to be exploited to analyze and predict the progression of the cancer. Noticeably, these findings proved that several genetic alterations are inevitable during the transition from in situ to invasive HNSCC; this could serve to detect pre-malignant HNSCC lesions, and consequently cure them before reaching a more advanced, and therefore harder to treat, phase of the cancer. This cancer cells, in fact, present frequent gain or loss of chromosomal regions (e.g., 9p21, 3p21, 17p13, etc.) [2], and research allowed to associate certain chromosomal anomalies to specific stages of development of HNSCC [3], [4].

The tumor microenvironment in HNSCC is made of tumor, stromal and immune cells and cancer-associated fibroblasts (CAFs). Tumor cells and CAFs secrete specific growth factors (i.e., VEGF) that secure the conditions necessary to life and support of the cancer: they trigger neovascularization for the provision of nutrients and oxygen, while endothelial cells provide factors that ensure survival and reproduction of CSCs [5].

The anatomical location and etiological agent also significantly affect to what extent HNSCC is infiltrated by immune cells, and from what kind of cells; a significant difference was in fact found between smoking- and HPV-derived HNSCC [6]. Moreover, research showed that this tumor originates different immune responses, and consequently different patterns of markers can be

detected and used to observe the development of the pathology and predict the possible reaction of the specific subject to different kinds of treatments [7].

1.1.2. Incidence

HNSCC constitutes 4.5% of cancer diagnoses and deaths: 890000 new cases and 450000 deaths annually according to GLOBOCAN statistics [8].
As it is highly correlated with alcohol and tobacco consumption, HNSCC's incidence increases in areas in which such habits are more common (e.g., lowering in developed nations, increasing in the developing world). A third crucial factor that was more recently found to be associated with HNSCC, is high-risk human papilloma virus (HPV); specifically, HPV has been assessed as an important pathogenic factor for such cancer, although its influence strongly depends on the kind of HPV and on the specific head and neck region (this aspect will be described more in detail in paragraph [1.2]).
HNSCC is more diffused to men than women and in people older than 50 years of age [8]. Generally, differences in lifestyle and access to healthcare also have a significant impact on the incidence and the survival of HNSCC among geographic regions and socio-economic statuses.
While the highest incidence was detected in India, probably caused by the extensive consumption of tobacco, the overall incidence is increasing globally especially in younger people due to currently common lifestyle (i.e., increased alcohol and tobacco consumption).

1.1.3. Overview of HNSCC types

As mentioned, the term Head and Neck Squamous Cell Carcinoma is used to refer to a category of tumors that are located in head and neck regions, but they can affect different sites and can, therefore, be further classified according to their primary region of origin. In the following, an overview of the types of HNSCC is presented; such a categorization is important as etiological agent and prognosis

are strongly related to the specific site in head and neck where the cancer originates. In Figure 1 [9] a visual representation of the anatomy of the throat is presented in order to allow, together with the brief overview in the following, a deeper understanding of the difference between the different anatomical sites and, consequently, the corresponding cancers.

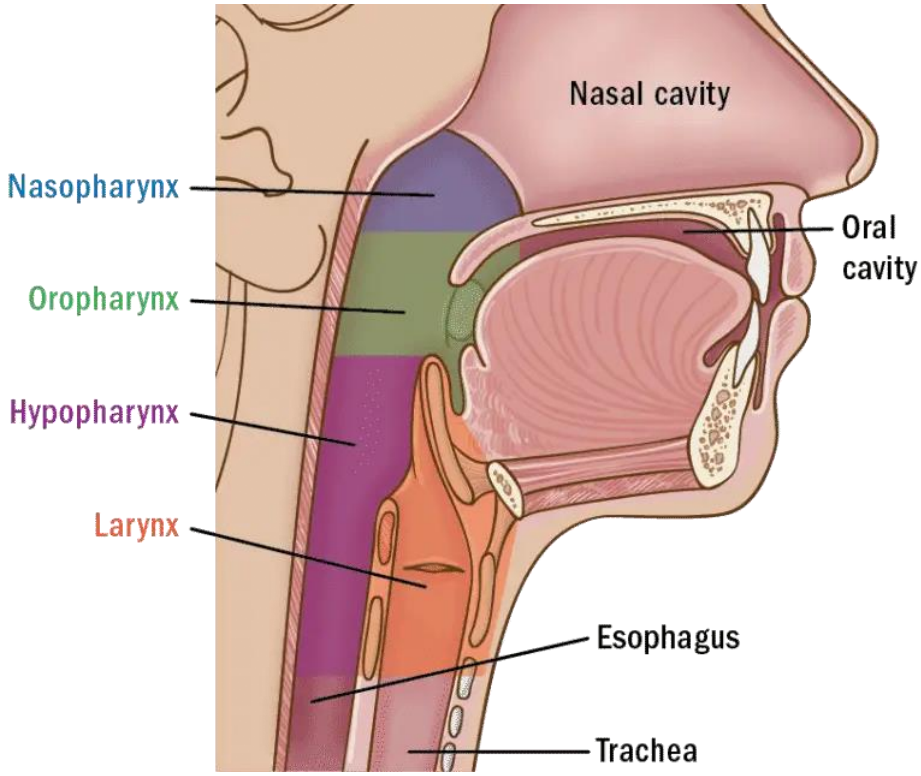

Figure 1 Anatomy of the throat

Oral cavity HNSCC

This type of cancer can affect the tongue, the gums, the lips, floor and internal lateral side of the mouth. In such HNSCC, the oral microbiota is known to play a significant role in tumor microenvironment; its composition is highly affected by oral health which is therefore, together with tobacco (smokeless or combustible), alcohol consumption and HPV status, significantly related to the risk of developing this tumor.

Nasopharynx HNSCC

The nasopharynx is the upper part of the throat: it constitutes a canal for air coming from the nose and directing to the trachea; it is in fact located above the palate and behind the nasal cavity.
This type of HNSCC is related to Epstein-Barr virus (EBV) infection.

Oropharynx HNSCC

The oropharynx is the middle part of the throat and participates in breathing, digestion and speaking processes by allowing swallowing and vocalization.
This type of HNSCC affects, besides the walls of the pharynx, the base of the tongue, the tonsils and the soft palate. It is highly associated with human papilloma virus and is in fact typically referred to as HPV-positive HNSCC; as such, it determines a better outcome in respect to the other ones with a 70-80% 5-year survival rate, granted by its better response to radio- and chemotherapy [10].

Hypopharynx HNSCC

The hypopharynx is the lower part of the throat; it plays a role in the digestion process by constituting a canal to the esophagus for food and liquid.
This type of HNSCC is strongly associated with smoked tobacco and alcohol consumption, and it's usually associated with a worse outcome when compared to the others.

Larynx and glottis HNSCC

The larynx is located below the hypopharynx, and glottis is the part that holds the vocal cords; they allow vocalization by serving as passage for the air that makes vocal cords vibrate to produce a sound.
Laryngeal and glottic HNSC, can consequently affect the patient's breathing and ability to speak; significant risk factors are combustible tobacco and alcohol consumption.

1.2. Staging

HNSCC is diagnosed through biopsy; after the histopathological evidence of its occurrence, the staging process involves the following steps:

- head, neck and oral cavity examination
- cross-sectional Computed Tomography (CT), Magnetic Resonance Imaging (MRI) or Positron Emission Tomography (PET) of head and neck to determine the dimensions of the locoregional tumor
- chest CT to assess the presence of distant metastasis

This procedure allows the staging of the cancer; traditionally, the major prognostic factor in head and neck cancers (HNC) clinical outcome prediction is the tumor node metastases (TNM) staging system [11]; this standard classifies the malignant tumor basing on:

- T: size of the primary tumor and whether it has spread to the adjacent tissue.
- N: affection on regional lymph nodes (i.e., components of the immune system that filter anomalous cells from the lymphatic fluid); this occurrence involves the beginning of the spreading of the cancer in the body and, therefore, the delocalization from the origin site. In general, lymph node involvement indicates a more advanced cancer and has significant implications in treatment planning since it requires more aggressive therapies.
- M: presence of distant metastasis (i.e., secondary tumor originated from the diffusion of cancer cells from the primary tumor to other locations).

Such staging system has a crucial role in the definition of the treatment, since it allows the discrimination of cancers basing of the properties that make them more or less suitable for different approaches.

This value, however, while providing a first indication of the future development of the patient's conditions, does not take into account the heterogeneity of the tumor within each stage, consequently leaving room for further improvement by integrating such valuable information [12].

1.3. Prognostic indicators

In dealing with HNSCC patients, as well as patients affected by any pathology, investigating prognostic indicators is pivotal. Such factors provide determinant information regarding the future progression of the disease and, consequently, valuable insights for treatment planning. In the definition of the prognosis, clinical and imaging data are considered together with etiological agents due to their crucial influence on the course of the pathology.

The prognosis of a patient affected by HNSCC, and any cancer in general, is strictly related to the stage of the tumor and the eventual presence of metastasis in particular; the patent's outcome is therefore highly affected by any factor that indicates the development of such occurrence.

Indeed, a determinant phenomenon in cancer development is the epithelial-mesenchymal transition (EMT), in which a tumor cell turns into a mesenchymal phenotype: a state in which the cell presents an anomalous shape and, more importantly, an increased mobility, that allows it to move and penetrate thorough the tissues. Understandably, this event is decisive to tumor cell invasion and metastasis [13] which are, in turn, key factors to a patient's prognosis since they refer to the phenomenon with which tumor cells diffuse to other locations of the body. In HNSCC cases a specific class of substances, matrix metalloproteinases (MMPs), plays a crucial role in the development of metastasis; MMPs contribute to the degradation of ECM and, consequently, tumor cell invasion and metastasis and, accordingly, poor prognosis [14].

A further crucial aspect to examine is tumor's hypoxia, which triggers the secretion of factors that induce the angiogenesis and the degradation of ECM. A high level of hypoxia foretells discouraging prognosis and ineffectiveness of radiation therapy [15], [16].

Noticeably, research showed that HPV status has a tremendous influence on HNSCC prognosis and effectiveness of different treatment strategies; indeed, disparities in optimal treatment and resulting survival are such that HPV-positive

HNSCC was established as a pathology with different properties, etiology and prognosis than HPV-negative one [17].
Such discriminant property will be examined in the following, where the differences between HPV-positive and -negative will be depicted.

1.3.1. HPV-positive HNSCC

Human Papilloma Virus (HPV) is among the most common sexually transmittable infections in the world [18] and, besides being highly correlated to risk of cervical cancer, it can provoke malignant lesions to the upper aerodigestive tract (e.g., pharynx, larynx, etc.); thus, it is considered a crucial risk factor to the development of HNSCC.
HPV is a DNA tumor virus that can infect the stratified squamous epithelia, the mucosal and cutaneous tissues causing epithelial proliferation and, eventually, carcinogenic transformation [19]. Among the over 100 strains of this virus, HPV16 is the one related to HNSCC and, specifically, oropharyngeal cancer.
HNSCC prognosis is strongly affected by HPV status, and therefore testing is mandatory in case of detected head or neck tumor. Specifically, HPV-positive HNSCC is significantly more favorable than its HPV-negative corresponding [10]; the latter is, in fact, found to determine a much higher probability of relapse. Such different outcomes could be determined by the wider infiltration of immune B-cells into the tumor microenvironment and the fewer genetic mutations in the tumor's cells [20]. These factors contribute to the better reaction to radio- and immunotherapy which, in turn, allows a significantly lower risk of death when compared to HPV-negative HNSCC [21].
As a result, HPV vaccination is acknowledged as the most effective primary prevention to contain HPV-positive HNSCC.
Furthermore, as the key role of HPV status on the staging of the HNSCC was acknowledged, in 2017 the Union for International Cancer Control (UICC) and the American Joint Commission on Cancer (AJCC) introduced the new Cancer Staging Manual [22] which, among other innovations (e.g., integration of depth

of invasion in oral cavity in the staging process), considers the HPV-positive status when staging the HNSCC. Such new policy proved more effective in prognostic discrimination compared to the previous one [23].

1.3.2. HPV-negative HNSCC

HPV-negative HNSCC is highly associated with tobacco and alcohol consumption, and therefore with the lifestyle conducted by the patient.
In general HPV-negative HNSCC patients are more commonly characterized by inadequate oral hygiene and lower socioeconomic status, and they are likely provided with unfavorable prognosis [10].
Indeed, tobacco is widely recognized as one of the leading risk factors for the development of tumors in the aerodigestive regions, and HNSCC is no exception to this association. This correlation is due to the several carcinogenic substances contained, such as aromatic amines, polycyclic hydrocarbons and nitrosamines that can severely affect the human body, especially when associated to high-temperature combustion [1]. Such habit damages cells of the oropharynx, consequently increasing 5 to 25 times the risk of HNSCC [24].
On the other hand, among the several unfavorable effects alcohol consumption is known to have on health, an increased risk of HNSCC development is universally recognized. Specifically, the impact of its consumption on the risk of HNSCC occurrence is proportional to the dosage [25]. Alcohol, in fact, makes mucosal tissues susceptible to carcinogens such as smoke or substances in food; interestingly, in fact, consumption of tobacco and alcohol, which are both individually recognized as important risk factors to HNSCC development, seem to increase noticeably (i.e., 40-fold) when such products are combined [26].
Tobacco and alcohol consumption also highly affect other factors that, in turn, can have a significant impact on the overall outcome for the patient; they are, in fact, found to cause chronic inflammation in the exposed tissues. This originates the twofold unfavorable effect of promoting local production of cytokines and growth factors that encourage carcinogenesis, while also causing the patient to be

less responsive to chemo- and radiotherapy treatments; therefore, such habits can indirectly lower survival chances of the subject [27].

1.3.3. Prognostic value of imaging data

Medical images such as CT and PET allow to visualize the tumor and hold, understandably, a crucial amount of information for diagnosis, prognosis, and treatment planning purposes.

Specifically, CT employs ionizing radiations to provide cross-sectional images of the body allowing clinicians to assess tumor location, size and characteristics; such properties are determinant in the staging process and, therefore, in prognosis definition.

During and after the treatment of the tumor, CT is also employed to monitor the response of the cancer and, in case of effective cure, to assess possible relapses during the follow-up.

On the other hand, PET achieves similar tasks through a different process. A radioactive substance that is metabolically active for the tumor is introduced in the patient: it accumulates in cancerous tissues and emits positrons as it decays; such particles can be detected to reveal the position and extent of the cancer. As a result, PET scans allow to monitor the metabolic activity of the tumor and, consequently, the effectiveness of the treatment.

Recently, [18F]fluorodeoxyglucose PET (FDG-PET) was identified as the most informative imaging technique for HNC [28], owing to its ability to present physiological manifestations due to tumor metabolism; such technique is therefore usually employed for HNSCC cases.

The information held in such images can be effectively processed to obtain values that have a crucial role in the prognosis and treatment planning of HNSCC. In fact, besides allowing the visualization and the determinant assessment of lymph node involvement, imaging scans provide the reconstruction of the volume of the tumor that can be described by Gross Tumor Volume (GTV) parameter; this value indicates the extent of the cancer with all its visible parts. Monitoring the GTV,

doctors can determine the staging of the tumor and, later, assess whether the treatment is succeeding in shrinking it.
It is worth highlighting that despite CT and PET's valuable role in oncology, these procedures involve dangerous exposure to radiations and their employment must, therefore, be carefully considered in order to maximize their benefits while containing the implicated risks.

1.4. Treatment

The employed treatment strategy strongly depends on the single patient's stage and properties of the disease, as well as the subject's characteristics and wish. It is important to consider whether the HNSCC is HPV- or tobacco-related and the age and general health of the patient, since these factors have a significant impact on the choice of the most appropriate treatment approach (e.g., intensive therapy can be inappropriate for elderly people).
The main modalities to cure HNSCC are resection, radiation and systemic therapy. The treatment is chosen in order to maximize the curative effects, while minimizing damages to functionalities of the patient.
In the following, a brief overview of the different treatment approaches for HNSCC is provided:

- Resection: it refers to the surgical removal of the tumor; it can be successfully executed when the cancer is localized.
- Radiotherapy: this approach is based on beams of radiation that reduce the tumor by destroying cancer cells. The dosage and properties to apply this technique are set up in order to irradiate the Planning Target Volum (PTV); such parameter is obtained through the processing of GTV: the latter is in fact employed to extract the Clinical Target Volume, that includes the visible tumor volume (i.e., GTV) expanded with a margin that accounts for microscopic cancerous extensions that are likely to be present. In turn, CTV is further increased to obtain the Planning Target Volume (PTV), that considers additional factors that require a slightly wider coverage of the

delivery of radiations (e.g., small deviations in patient positioning when acquiring imaging scans or when delivering the treatment). In conclusion, the PTV accounts for all the determinant aspects and is therefore the definite area that will be completely irradiated during radiotherapy.

- Systemic therapy: this treatment employes substances that can be introduced in the body to reach cancer cells and is typically used when the tumor has diffused to other regions in the body. This approach includes chemotherapy, which involves the use of certain drugs (e.g., cisplatin) to hinder the growth and survival of cancer cells. It is administered orally or through intravenous infusion and it is usually used in combination with other treatments in case of advanced or metastatic cancer.
- Chemoradiotherapy (CRT): combination and simultaneous administration of radiotherapy and chemotherapy. It enhances the effectiveness of radiations.

When considering the primary treatment, surgery is the preferred choice; however, it is only appropriate for certain cases: the tumor must be at an initial stage, located in a surgically accessible site, and its resection must not compromise the patient's vital functions. For the cases that do not do not verify such conditions, radiotherapy is typically chosen as primary treatment. On the other hand, chemotherapy is not usually employed as primary approach for this specific cancer, but it can still be suitable for patients who are not eligible for both surgery and radiotherapy (e.g., fragile subjects, advanced tumors, etc.).

In the unfortunate case of primary surgery or radiation treatment failure to the complete removal of the tumor, the employment of the alternative approach provides encouraging probability of success [29].

On the other hand, even in cases where the primary treatment proves effective, a secondary approach can be employed; this practice is adopted in order to minimize the risk of tumor relapse. Such precaution is chosen since even though the primary treatment is successful, some cancer cells could still be present in the original site – arising the risk of developing a new tumor. The most common

secondary treatment employed in HNSCC cases is postoperative adjuvant radiotherapy.

Currently, the different kinds of HNSCC are usually treated in the same way: the approach is mainly chosen according to the tumor stage, which in turn is affected by HPV status (i.e., whether it is HPV-positive or HPV-negative HNSCC), rather than basing on the origin site.

Considering the key role played by the severity of the disease, its influence on treatment planning is noteworthy. Optimal cases are small, restrained tumors with no involvement on lymph nodes (i.e., T1/2, N0): in this instance a single modality treatment, based on surgery or radiation, can be enough to effectively treat the pathology [30]. Radiation is typically involved in laryngeal and pharyngeal cancers, while surgery is usually employed for oral cavity cancers treatment.

In less fortunate cases of more advanced tumor or nodal stages (i.e., T1/2, N+), adjuvant radiotherapy after the surgery is required as it improves survival probabilities by significantly lowering the risk of recurrence. In these cases, the PTV to execute radiotherapy is obtained by increasing by 3mm the CTV.

In highest risk groups (i.e., T1/2/3/4, N+), a big tumor is associated with lymph nodes extracapsular growth (i.e., spread of cancer cells beyond the boundary – capsule – of the lymph node); notably, the latter is considered a sign of particularly aggressive tumor. This very complex condition requires surgery, radio- and chemotherapy (i.e., tri-modal therapy) [31]. When implementing this approach, PTV is obtained by increasing GTV by 1,5-2 cm, and the contour of the lymph nodes presenting extracapsular growth is provided by augmenting the lymph nodes size by 1-1,5 cm.

In conclusion, CRT is usually adopted for advanced tumors, regardless of HPV status, and it is usually implemented starting 6 weeks after surgery.

On the other hand, the use of tri-modality treatment involves increased effects of the toxicity of radiations, resulting in higher risk of non-cancer-related mortality [32]; thus, it is crucial to accurately analyze the pathology before the beginning of

the treatment, in order to avoid the addiction of CRT upon other approaches originally chosen as single modalities.

1.5. Impact of Machine Learning on Oncology:

In the following, the increasing relevance of Machine Learning methods on oncology is presented; specifically, the properties that make such technology highly effective in this specific application and the possible progress and benefits that it could introduce in healthcare are depicted.

1.5.1. Classification, regression and Time to Event analysis

Over the last few years, Machine Learning (ML) has had an increasing impact on radiation oncology, and in healthcare in general. The combination of complex 3D imaging and numerous clinical data required to develop the estimates and procedures involved in this field, makes this specific application highly suitable for ML techniques. This technology, in fact, has proven great potential in the handling of vast amounts of data due to its ability to identify patterns and relationships between variegated information.

The most widely performed tasks by ML algorithms in oncology are classification, regression and Time to Event analysis. While classification allows the categorization of input data into predefined classes based on specific characteristics, regression quantifies the cause-effect relationships in data by exploring the influence of independent variables on the dependent one. More recent ML algorithms integrate Time to Event analysis, which is a statistical method that provides predictions of whether and when an event of interest will occur; as such technique constitutes the main focus of this thesis, it will be extensively described in Chapter 2.

The employment of ML-based systems for prognosis and treatment recommendation could significantly improve patient care and comfort through more efficient handling.

Moreover, this technology has proven great potential in executing accurate predictions, that would be of determinant advantage in medicine. More specifically, the crucial benefit that this technology could introduce to oncology, becomes clear when considering the critical impact of early diagnosis, and consequently early intervention, on survival rates of people affected by tumors.
In effect, ML-based algorithms can efficiently process data in order to extract the information of clinical interest and, eventually, identify subtle patterns that indicate early-stage cancer. Such achievement would allow milder yet more effective treatments and, as a result, significantly improved survival rates and conditions.
Specifically, Head and Neck Squamous Cell Carcinoma is no exception to this reasoning: the accurate prediction of a ML model could signal the risk of tumor relapse, development of metastasis, or death in a much shorter time – giving the clinicians the time and opportunity of treating the patients earlier and therefore with much higher probability of success.

1.5.2. Impact on treatment strategy

Machine Learning could also be exploited to compare the consequences to different treatment approaches in order to support clinicians in making informed decisions. This application holds a twofold advantage; the first aspect to consider is the difficulty involved in the choice of the most effective treatment strategy in the single case, which could highly benefit the insights provided by the data analysis conducted by the ML-system. Such approach could in fact account for all the input clinical variables and therefore infer that a certain therapy is the most effective for the single case – thus potentially leading to a personalized treatment selection.
As mentioned, HNSCC treatment is currently planned basing on tumor stage and HPV status; the origin site is not appreciably taken into consideration: this provides an example of case in which the contribute of a ML model could introduce a significant benefit by extracting some valuable information related to

the region of the primary tumor that influences patient's prognosis or response to different treatment approaches.

The second benefit introduced by ML is related to the tradeoff between therapy effectiveness and damage to functionalities of the patient; it is in fact widely recognized that more aggressive therapies can be more effective in terms of removal of a tumor but, in some cases, at the cost of devastating effects on the subject. Therefore, an analysis of the data that allows the accurate prediction of the patient response consequently to different treatment strategies could prevent the rising of a new, maybe even worse, problem even in the case of successful cure of the tumor (e.g., death due to the toxicity of the treatment). Each therapy, in fact, implicates serious adverse reactions, therefore a careful balance based on the specific patient is required in order to optimize the efficacy of the treatment while containing the side effects.

Indeed, considering the crucial role of head and neck area in human functions, the implications of HNSCC and its treatment crucially affect health-related quality of life that, in turn, is closely related to survival [33]. On average, 1 and 2 years after treatment, head and neck patients' quality of life is found to be worsened compared to pretreatment [34]. Each of the many possible combinations of therapies implicates certain physical, functional and psychological consequences and deeply affect patients' life in general.

Among the different ML tasks, regression and Time to Event analysis allow the exploration of the development of certain statuses over time, and therefore the accurate quantification of the passing of time before an adverse event occurs. This additional information could introduce further benefit in supporting clinical decision making – by the clinician or the patients themselves. In particularly unfavorable cases, in fact, such knowledge could contribute to different clinical decisions as a result of the balance between expected survival time and treatment implications (i.e., damaged functionalities, pain, time obligation etc.).

CHAPTER 2
METHODS FOR TIME TO EVENT ANALYSIS IN MEDICINE

In healthcare, the possibility of predicting clinical conditions is crucial to medical decision making and treatment definition. In light of this, many efforts were made over the years on the development of computational and statistical methods to predict clinical outcomes. These techniques are suitable for any kind of application in which a certain outcome is to be foreseen basing on specific data; for instance, some typical applications of interest in the medical field are the estimation of the risk of developing a disease, or the probability of benefitting from a certain treatment.

The pivotal effect of accurate clinical outcome predictions is widely recognized, and therefore a further step has been taken in the development of technical methods for this kind of application: Time to Event analysis allows not only to predict the condition - or "event" - of interest, but also to locate it in time, as opposed to general outcome prediction methods, which only refer to a certain follow-up time point.

In this chapter, the technical aspects of Time to Event analysis will be extensively examined. Firstly, a detailed overview of this technique it's provided in order to gain a deep understanding of its functioning and its main critical issues. Subsequently, a review of the traditional methods employed to carry out this procedure is presented, completed with their corresponding limitations. Lastly,

more advanced techniques are described: a detailed review of operating principles and experimental applications of Machine Learning approaches for Time to Event analysis is conducted.

2.1. Time to Event analysis

Time to Event analysis, or survival analysis, is a set of statistical approaches aimed at investigating the expected duration of time before an event of interest takes place. The event in question can be of many different kinds, and therefore survival analysis is popularly employed to conduct studies in different fields (e.g., duration of unemployment after job loss, time to failure of industrial machine parts, etc.). Nevertheless, this technique is understandably widely used in medical research: some typical examples of application can be the quantification of the probability of survival of a patient at a certain time point, the investigation on the impact of a clinical data to a patient's survival, the comparison of survival curves in different groups of patients. The event isn't necessarily the death of the patient or the emerging of a disease, but can be any condition that occurs in a specific time point, provided that it is unambiguous and well defined.

Besides death, in fact, other kinds of events are often of interest; the analyzed occurrence is intrinsically related to the investigated condition or pathology: while in many potentially fatal diseases death is evidently the event to be predicted, in other clinical conditions there are some aspects to which the elapse of time plays a key role, and therefore survival analysis allows to gain an even deeper insight to the development and the handling of the illness. In cancer related studies, for example, survival analysis is widely used to observe the effectiveness of treatment by quantifying time between its administration and potential relapse or plausible formation of distant metastasis.

The definition of the event comes with its difficulties: in case the event is the death of a patient, it is crucial to distinguish an all-cause from a specific-cause mortality. In some other cases, the event is the emerging of a disease or any condition that

does not actually occur instantaneously and is therefore harder to locate at a specific time point.

Time to event analysis can overcome the limitations of the employment of logistic regression analysis. The occurring of an event can in fact be treated as a binary data and therefore analyzed with logistic regression that allows, for example, to investigate the relationship between a predictor variable (also referred to as covariate) and the probability of experiencing a certain event. This analysis, however, is only related to a specific time point and therefore does not allow to investigate when the event occurs throughout the whole length of a period. Time to event analysis, in contrast, allows both to estimate whether the event will occur and when in the observation window as well as permitting - like logistic regression - to investigate whether and how much survival times are related to specific covariates.

In order to conduct Time to Event analysis, survival function and hazard rate are the crucial concepts. The survival function describes the probability of not experiencing the event of interest, or of surviving, by a specific time point. In contrast, the hazard rate is the rate of occurrence of the event during a certain time interval [40], and the hazard function describes the instantaneous rate of occurrence over time, which represents the hazard rate in correspondence of an infinitesimally small time interval. The survival and hazard functions are related in the way that, the higher is the survival rate, the lower is the hazard and vice versa. It can also be useful to consider the hazard ratio, which indicates the ratio of hazard rates between 2 different groups; this value allows the comparison of different subsets of subjects.

2.1.1. Critical issues

The main obstacle met when dealing with survival data is the phenomenon of censoring: typically, not all subjects experience the event before the end of the observation period, and therefore for these *censored* observations the actual survival times are unknown; it is crucial to handle these cases properly in order

to obtain valid inferences, and for this reason specific statistical methods are required. Censoring can occur if a subject experiences some different event that prevents further follow-up, or simply stops showing up at a certain time of the study period. A visual representation is shown in Figure 2 [35].

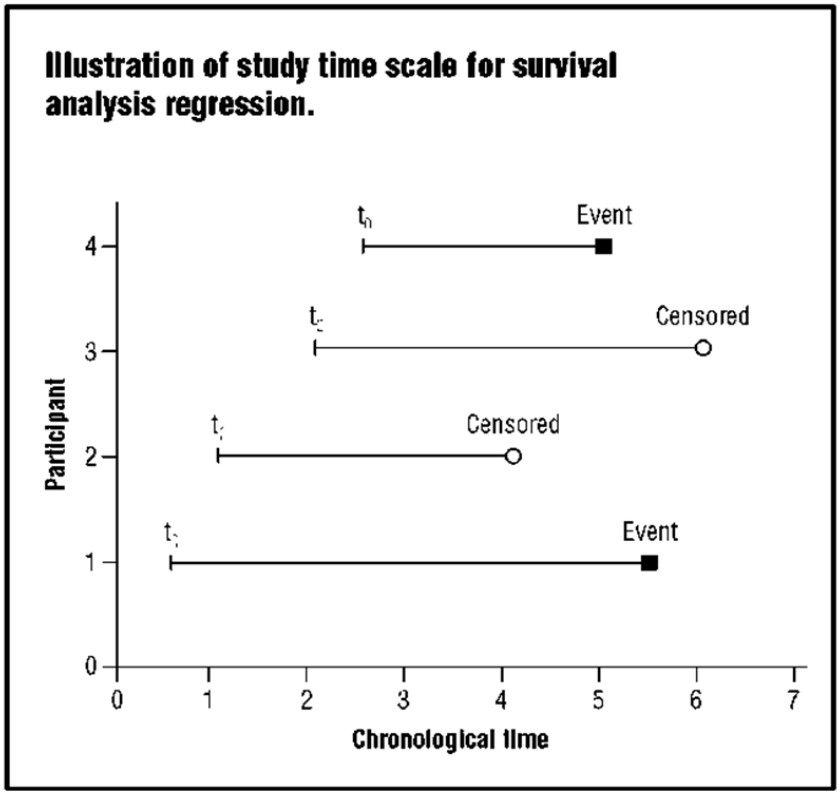

Figure 2: Example of visual representation of censoring

In the aforementioned cases, it is only known that the survival time is longer than the observation time; ignoring such cases would bias the analysis, because of the underestimation of survival time in censored subjects. This phenomenon is called *right censoring* as opposed to *left censoring*, which refers to cases in which a subject is known to have experienced an event before the beginning of the observation, but it is not known the exact time when it happened. Lastly, *interval censoring* takes place when the event is only known to have occurred between two considered time points, and not exactly when; this is however a less common issue, especially when death is the event of interest.

Although there is no ideal solution for censoring, survival analysis addresses it, but it is still important to minimize it as much as possible by promoting complete follow-ups.

In order to have unbiased inferences, the time of censoring must not be related to the event time, meaning that the censoring must be noninformative [36]. An example of informative censoring is when patients cease attending follow-up appointments due to a physical condition related to the risk of the event, preventing them from participating physically in the clinical visit.

It is also possible to store and analyze data related to circumstances and reasons that led to censoring, in order to subsequently evaluate a potential bias.

Another issue with survival data is called truncation, which is related to the subjects' selection; there's often going to be subjects who have experienced the event before the study and the subject identification took place, and therefore cannot be identified and considered in the analysis. These patients suffered from occurrence of interest but they were not known to exist, and therefore this is a different case than left censoring. This phenomenon is called left truncation and implies a selection of subjects who haven't experienced the event to take part in the analysis.

On the other hand, right truncation takes place when subjects who undergo the event are selectively considered in the study. The bias introduced by truncation can be partly handled during survival analysis, but it is still important to prevent it as much as possible during the definition of the target population.

As mentioned, a very sensitive factor for time to event analysis is the choice of the data; censoring constitutes a significant issue and has to be carefully handled since it can lead, due to the unknown times of event, to the generation of too few or too many at-risk subjects.

It is also worth mentioning the issue related to the correlation between medical data and a single hospital or even a single operator, especially in those cases where data from multiple centers are collected to achieve the amount of data required by survival analysis. This can result in the formation of a non-perfectly homogenous dataset and therefore to a lower performance in the predictions.

Furthermore, the width of the interval in survival analysis can play a significant role in the outcome prediction [37], and must therefore be carefully picked, especially in those cases where the employed model does not grant robustness to this factor.
In general, the lack of data is due, among other factors, to costs and privacy constraints [38], that limit the possibility to use and share medical information. It is crucial, when creating datasets for this application, to carefully consider possible under-represented groups in order to avoid biases, and to choose a proper dataset size to ensure the robustness of the prediction model.
These issues and the best way to address them are still being investigated in several studies, that proposed innovative solutions. A noteworthy example are generative models (survivalGANs) [39], that allow to overcome some of the mentioned limits by generating synthetic medical datasets that mime the statistical properties of the original data. This approach is particularly helpful when dealing with costs, privacy, uneven under-representation, and lack of data in general.

2.1.2. Study design

As mentioned, for survival analysis to be effective, the design of the study is crucial as it is the main way to prevent the many obstacles related to this method (e.g., censoring, truncation, under-representation of the event, etc.). Thus, several choices (i.e., length of follow-up, time of origin, selection of target population etc.) are to be made carefully and in accordance with a well-defined strategy.
It is important, in fact, to carefully choose the length of follow-up and follow-up intervals, in order to have a proper number of observed events. It is worth highlighting that longer observation times involve the risk that time could change other factors that influence survival times, causing outcomes that are not completely attributed to the covariates of interest.

The time of origin also needs to be well specified, and it should be chosen so that all the subjects are starting from the same conditions and are as comparable as possible. For instance, in a study that investigates the impact of some therapeutic treatment in patients' survival, it is chosen as the time at which it is administered, while in an epidemiologic study the origin usually corresponds to the moment the disease is diagnosed.

It is also worth considering that, for the survival analysis method to work properly, the sample size is not as important as the number of observed events; in fact, the first step is often the estimation of the required number of events to detect a clinical effect size. This can be done, for example with the Shoenfeld method for log-rank tests or proportional hazards methods. It will then be possible to proceed by estimating the number of subjects who are likely to experience the event.

2.2. Traditional methods for survival analysis

Although other kinds of research are possible (i.e., competing risks analysis, recurrent events analysis), in many cases the considered event only occurs once and puts an end to the observation of the subject who experiences it (e.g., death). For this kind of survival data analysis, 3 classes of methods are mainly used and distinguished:

- Nonparametric methods, that don't make assumptions related to the distribution of survival times nor assume a specific relationship between covariates and survival times.
- Semiparametric methods, that also don't impose assumptions on the distribution of survival times but assume a specific relationship between covariates and survival times.
- Parametric methods, which assume both a distribution of the survival times and a functional form of the covariates.

In the following, a review of the most widely used methods for Time to event analysis is provided.

2.2.1. Kaplan Meier estimator

The Kaplan-Meier estimator is a nonparametric method that estimates the probability of surviving beyond a certain time point [41]. Given a patient at risk at the beginning of an interval, the probability of surviving until the end of that interval is calculated. The survival probability at time t_i, in fact, can be calculated as follows:

$$S(t_i) = S(t_{i-1})(1 - \frac{d_i}{n_i})$$

Where:

- $S(t_{i-1})$ is the probability of survival at timepoint t_{i-1}
- n_i is the number of subjects alive right before t_i
- d_i is the number of events at t_i

Furthermore, it is known that:

$$at\ t_0 = 0, S(0) = 1$$

A Kaplan-Meier curve is the plot of estimated survival probability against time (i.e., survival function). The curve is a step-function where each vertical drop represents an occurrence of the event, therefore the estimated survival probability is constant between two events. It also reports right censoring as marks at the corresponding censoring times, and it is possible to plot confidence bands around the survival function. It is common to plot more Kaplan-Meier curves related to different subsets of subjects in one graph, in order to visually compare survival probabilities. It is also common to report the median survival time, which corresponds to the time (x-axis) value at which the event has occurred to 50% of the subjects.

This method is based on some assumptions, namely non-informative censoring and that there is no correlation between survival probability and the time in which patients took part in the study. As other nonparametric methods employed for survival analysis, Kaplan Meier estimator holds the advantage of being suitable for time-varying covariates and for cases that do not meet the proportional hazards assumption (that will be described in paragraph 2.2.3). On

the other hand, these methods do not allow an estimation of the effects of multiple covariates on the outcome prediction; semi- or fully parametric methods overcome this limit, providing estimates of how much each predictor variable affects the survival function.

2.2.2. Log-rank test

Log-rank test, also a nonparametric method, allows to assess whether there is a statistical difference in the probability of an event happening at a certain time point between different subsets. It can be very useful, in fact, to divide a target population into 2 or more groups according to a certain factor (e.g., age, gender, presence of a medical condition) and compare the survival curves with hypothesis testing in order to evaluate the possible incidence of such a factor on the investigated event. This is done computing a test statistic that quantifies the observed differences is survival between the different subsets. This process is based on the comparison of the number of observed events in respect to the expected number of events that would take place under the null hypothesis (i.e. condition in which there is no difference in survival between the groups). This comparison involves the generation of a p-value: if it is less than a predetermined level (a commonly used one is 0.05), than the null hypothesis is rejected and the statistically significant difference in survival between the two groups of subjects is assessed.

2.2.3. Cox proportional hazard model

Cox Proportional Hazard is a widely used semiparametric method for survival analysis that models the hazard function, rather than the survival function [42], and allows to assess the effect of covariates on survival probabilities. This technique estimates the relationship between the probability of the event occurrence and the considered covariates. It is based on the assumption that all subjects have a common baseline hazard function that depends on time, and each

patient's hazard function is a multiple of the baseline function; each subject's multiplicative factor is a constant that depends on a time-independent function of their covariate values. A crucial consequence of this, is that the effect of a single covariate is the same at all time points and therefore the hazard ratio (HR) between different subjects is constant over time. This implies that, if at the beginning of the study a certain subject has a higher risk of experiencing the event than another patient, at all following time points the latter will present a higher survival probability: the hazard curves will be proportional and will never cross. It is a rather strict assumption that needs to be tested, as it is not always met. This phenomenon is referred to as *proportional hazard* assumption, and is the parametric component of this approach. Consequently, the employment of this method also implies the assumption of a linear relationship between the covariates and the log-hazard and that the covariates are time-independent.
As mentioned, the Cox Proportional Hazard model is a semiparametric method, and as such it does not make assumptions about the shape of the hazard function; it allows instead to use the estimated parameters to extract the survival function. The baseline hazard is non-parametrically estimated, therefore its shape is arbitrary and it is not necessary to specify it; this is a significant advantage, since it can sometimes be hard to determine and will lead to inaccurate inferences if wrongly identified.
Despite the mentioned assumptions, which are not always easy to meet, the semiparametric Cox model is widely used in survival data analysis because it is known to provide safe inferences while not having the need to specify a data distribution, which is an important advantage. It is then possible to plot the survival probability of different groups, in order to compare them comfortably.

2.2.4. Parametric methods

It is crucial to consider that the proportional hazard assumption is not always met, and therefore in these circumstances Cox model cannot be used, while parametric methods can be more suitable. These techniques start from the

assumption of a specific distribution of the survival times and often carry the advantage of high efficiency [43], which can be particularly useful when dealing with small sample sizes. For parametric methods to work properly, however, it is pivotal that the data distribution is identified as precisely as possible (e.g., Weibull distribution, Log-Normal distribution, Gamma distribution etc.), and its choice strongly depends on the specific clinical event to be analyzed; although this can be very challenging, it is fundamental since if not done correctly, the model will lead to misleading inferences. Therefore, a crucial difference respect to parametric and semiparametric methods, is that parametric methods require an appropriate prior knowledge about the data but, on the other hand, they are less flexible as the chosen distribution must align as much as possible with the observed data.

2.3. Machine Learning approaches for Time to Event analysis

2.3.1. Artificial Intelligence in medicine

Over the last few years, Machine Learning (ML) techniques have become increasingly popular in clinical literature, and Artificial Intelligence applications are positively affecting medicine and clinical practice [44] [45]. The strength of this approach lies in its ability to recognize patterns held in big amounts of medical data, that can be employed for purposes such as disease prediction [46] and, more in general, clinical outcome prediction [47]. By enabling a more in-depth analysis of the relationships between clinical features and outcomes, Machine Learning is allowing the approach to the goal of personalized medicine and customized treatments, which would change the landscape of medical practice.

The progress introduced by AI-based algorithms was also enabled by the advances in technology of the last past decades, that led to the development of

countless different kinds of sensors and, consequently, of a great amount of data; this huge availability of information found a good fit in Machine Learning techniques, that not only allow to deal with big datasets, but even require them in order to provide better performances.

Nowadays only few particular medical applications exploit Artificial Intelligence: some examples are detection of epilepsy seizures [48], atrial fibrillation [49] and examination of bioimages or histopathology samples for diagnostic purposes [50]. In fact, although some AI-based algorithms have already been approved by the European Medicines Agency (EMA) in Europe and Food and Drug Administration (FDA) in United States [51], most AI-powered medical technologies are still at a research stage, as they are not ready to be employed for clinical practice. Given the early stage of development of such a technology in fact, these methods are currently only applied on specific tasks and in controlled environments, in order to contain the risk of uncertainty around possible errors. In conclusion, while this approach has shown great potential and given proof of being effective on some level, it still needs to be carefully investigated and improved, in order to be allowed to be used in health systems.

It is also worth mentioning the difficulties regarding legal framework and ethical implications; it is in fact important to consider that these methods will have to be validated like every other medical device or technique, and this requires the outline of suitable trials. An appropriate regulatory mechanism, in fact, is critical to avoid the delivery of unregulated healthcare services or by unregulated providers. WHO has worked with a group of experts to redact the principles to ensure a safe and ethic use of such systems [52]; in order to exploit the potential of this technology, in fact, it is crucial to grant that in every situation, the conclusions drawn by the AI are fully understandable and dictated by public interest. While moving the first steps in the employment of this technology, focusing on well-defined and specific tasks allows, besides the mentioned containment of the risks, easier definition of regulation mechanisms and validation processes.

Despite the current limitations of such a technology, due to the need of further investigation and refinement in the specific intended use, the advantages it can introduce when improved and ready to be applied, are several and clear: it can allow earlier diagnosis, less invasive options, more optimized and personalized treatments, all while reducing the burden on health staff and the length of hospitalization; therefore, besides the advantages in terms of comfort and efficacy of treatments for the patients, this techniques would also be of great benefit when facing health staff shortage, which is both a common and serious problem [54].

This technology has been simultaneously received with enthusiasm and resistance by healthcare professionals and in general experts of the field. It was found, in fact, that some healthcare workers fear AIs replacing clinicians, but it is worth highlighting that the goal is rather to provide them new tools and optimize their interactions with algorithms, in order to combine their strengths and maximize both performances and comfort for patients and healthcare staff.

Among Machine Learning techniques, Deep Learning (DL) methods have proven even further potential in applications on the medical field. DL-based algorithms, based on complex Neural Networks, are characterized by a high number of layers capable of modelling highly non-linear associations. This allows the progressive extraction of hierarchical, compounding features from the input data, thus achieving the detection of complex patterns. This approach implies a significant computational burden, especially for the training phase; this is one of the factors that prevented the spread of the research in this field until recent years, with the rise of latest computing resources.

As outlined above, Neural Networks can model complex relations between input and output but, comparing to other Machine Learning techniques, they require an even higher amount of data, and in particular labeled samples. However, in some cases this obstacle can now be overcome by relying on datasets that are widely available through Electronic Health Records, data-collection platforms or large-scale studies (e.g., Image Data Resources [55], UK Biobank [58]).

Specifically, Convolutional Neural Networks (CNNs) are acknowledged as one of the most powerful methods for image analysis [59] [60]. Such technique is based

on the convolution, the mathematical operation that allows to effectively exploit the spatial structure of images. A CNN consists of a sequence of convolutional layers, each of which computes a set of feature maps – collections of detected image patterns, using the previous layer's output and propagating its results forward. This approach allows an effective extraction and processing of spatial relations through high-level features, thus justifying the efficacy when applied to images.

CNNs are consequently considered crucial to the satisfying outcome of Deep Learning applied to medical image analysis, and their worth becomes clear when considering the importance of image data in the medical practice (i.e., diagnosis purposes). Most recently, new approaches such as Vision Transformers have emerged, that are being investigated and getting close, or sometimes exceeding, CNNs' performances.

A further advantage of neural networks lies in the fact that it is possible to easily merge different neural networks, and they can also be of different kinds (e.g., RNN, CNN, etc.). This enables the processing of different kinds of data while using the most suitable network for each kind (e.g., CNN for images) and additionally learn dependencies between data modalities during training. This results in an improved accuracy since each type of data has been handled in the most appropriate way.

2.3.2. Machine Learning applications for Time to Event analysis

It is widely recognized how important early diagnosis in many medical conditions and pathologies is, and Machine and Deep Learning techniques have shown great potential in bringing improvements to this practice due to their remarkable propensity for clinical prognostication [62]. Indeed, when dealing with survival analysis, Cox Proportional Hazard (PH) regression model is currently one of the most used among the traditional methods, although it is sometimes thought to work based on oversimplified assumptions. Furthermore, the recent increase in

data availability has led to a higher computational burden for this approach, and regression-based survival analysis methods in general. These two factors affected the increasing interest towards the application of Machine Learning techniques on this kind of purpose. The AI-based approach proved particularly appropriate to this purpose thanks to its ability to recognize complex patterns among big amount of data and can therefore often outperform other methods. Thus, the extensive interest to this kind of research becomes clear when considering how a technology that can effectively conduct clinical output prediction would have a huge impact on the handling of almost any kind of disease or pathology. Additional examination shows the even further propensity of ML-based techniques for quantitative image analysis, consequently allowing encouraging results for risk stratification.

Consequently, several recent papers have described the successful use of neural networks or other Machine Learning techniques to predict future clinical outcomes.

In accordance with the specific goal, the first step of this methodology consists in defining the most suitable dataset to employ for the training, validating, and testing of the model; this will allow it to establish the patterns contained in the data and use it to predict the labels.

In general, in order to design a model for any kind of prediction, it is crucial to define a proper loss function, or cost function, which quantifies the discrepancy between the true label and the one predicted by the model. Many kinds of loss functions can be used (e.g., mean absolute error, mean squared error, binary or multi-class cross-entropy, etc.), and it is important to choose the one that is most suitable for the specific application.

It is also necessary to employ a convenient optimization algorithm (e.g., gradient descent, ADAM, etc.), that will handle the minimization of the selected loss function. The minimization of the loss function will guide the evolution of the model as its configuration (i.e., weights) will be accordingly updated.

Different types of material can be used as inputs to output prediction models, such as clinical or claims data [63], images or even speech. Data extracted from

EHRs are increasingly employed for this purpose: they hold diverse information related to each patient, including vital signs, laboratory data, demographics and medically relevant events that occurred to the individual.

Once the data to use as input are chosen, it is time to gather meaningful features through feature extraction techniques; like in any other field of application of ML approaches, it is crucial to carefully design and carry out this stage in order to obtain satisfying performances of prediction. As mentioned, in fact, Machine Learning allows the employment of raw data as inputs, as they will be automatically processed by the algorithm in order to achieve the desired output; however, an appropriate selection of features or a strategic manipulation of the input data, can result in a more performant model. The extracted features can then be used for the training of the prediction model.

To gain a comprehensive understanding of the state-of-the-art research in the application of Machine Learning approach to perform Time to Event analysis, the findings of some noteworthy studies are synthetized in the following.

In the 1990s Artificial Neural Networks (ANN) started to be employed for this aim, and allowed more flexible modeling of the effect of covariates on the survival function [67] [68]. In the 2000s, the use of Random Forest (RSF) and Support Vector Machine (SVM) for survival analysis provided advanced performances in the recognition of significant covariates [69] [70] [71].

Consequently, in recent years the application of Machine Learning on survival analysis was extensively researched and led to numerous studies in which different ML-based algorithms are compared with traditional methods and with each other. In their work, Gong et al. [72] compared the performances of ML-based methods with the Cox regression model; the models processed simulated time to event data, and the performances were evaluated through concordance index (C-index), an evaluation metric commonly used in this application since it addresses the possible presence of censored data . The comparison included two ML algorithms, RSF and ANN, and six sets of synthetic survival data that presented different relations between predictor variables and hazard function (i.e., linear, non linear, dependent, independent predictors). In this study, the

ML-based methods did not only outperformed the Cox model, but also proved more robust to data size and censoring rates.

Among Machine Learning techniques, a further step was taken when Deep Learning showed an even higher potential for applications on survival analysis. The employment of this approach, in fact, highly benefits from the flexibility of the models, that leads to higher performance.

Hence, several authors have proposed different solutions for handling time to event data with DL models. A popular example is DeepSurv, an extension of non-linear Cox proportional hazards with deep neural network presented by Katzman et al. [73]. In their study, the authors compared the performances of the model with those of Cox Proportional Hazards (CPH) and of a ML-based algorithm, the RSF, carrying out experiments on simulated and real survival data, and simulated and real treatment data. This research, did not only prove the outperformance of DeepSurv in respect to CPH and RSF, but also its ability to provide personalized treatment recommendations by modeling the interactions between each patient's covariates and treatment effectiveness; the proposed individual treatments are then shown to improve the survival rate of the patients.

In their work, Lee et al. [74] propose DeepHit, an approach that uses a Deep Neural Network to learn the distribution of survival times. This method allows to address possible changes between covariates and risk over time by not making assumptions on the form of the stochastic process, with the consequent advantage of not having to know the underlying disease process in order to lead the survival analysis. In this research, real and synthetic data have been used to compare DeepHit with conventional survival regression models (i.e., Cox Proportional Hazards, Threshold Regression, Random Survival Forest) and ML and DL-based methods (i.e., Random Forest, Logistic Regression, AdaBoost, DeepSurv); comparing to these models, the proposed network, DeepHit, provided statistically relevant performance improvements respect to the other methods, possibly by not restricting to a proportional assumption.

Similarly to Katzman et al., Kvamme et al. [75], presented a combination of CPH model with Neural Networks characterized by the removal of the proportionality

constraint of the Cox model. This Cox-Time model parameterizes the relative risk function of the Cox model through neural networks: this allowed to benefit from the flexibility of the neural network combined with the ability to model event times continuously. In this research, the proposed approach provides the best overall performance in comparison with classical Cox regression, RSF, DeepHit and DeepSurv.

A different approach, was carried out by Bennis et al. [76] who developed a neural network that, assuming that the survival times distribution can be appropriately modeled through a mixture of Weibull distributions, was able to accurately estimate the parameters of such distribution. The proposed network, called DeepWeiSurv, was able to model a continuous survival function; the method was evaluated on two real-world datasets (METABRIC and SSER), and resulted to outperform the semi-parametric CPH and the aforementioned DeepHit.

In their work, Gensheimer et al. [77], describe Nnet-survival, a discrete-time survival model able to handle non-proportional hazards and large datasets thanks to mini-batch gradient-descent. The SUPPORT study dataset was used to test and compare the Nnet-survival with the standard CPH, Cox-nnet and Deepsurv. Through this analysis, Nnet-survival was found to provide good discrimination and calibration performance both with simulated and real data. The standard Cox model resulted in worse performance because several predictor variables violated the proportional hazard assumption; furthermore, both Cox model and Cox-nnet seemed to under-predict survival probability for the best-prognosis subjects. On the other hand, Deepsurv proved a lower performance when assessed using HCI and Brier Score compared to Nnet-survival. In conclusion Nnet-survival provided satisfying discrimination performance and the best calibration performance.

A very interesting study regarding the time to event analysis was carried out by Wang et al. [78] who were able to develop a Deep Learning-based approach for the prognosis of events related to head and neck cancer; this method, in fact, allowed the authors to obtain time to event models to predict overall survival (OS) and distant metastasis (DM) with the aim of producing recommendations for personalized radiation therapy (RT). In this study different kinds of inputs, based

on combinations of PET and CT images, were compared basing on Harrel's Concordance Index (HCI) and Kaplan-Meier curves. The main focus of this research, besides the comparison of single and multi-modality imaging inputs, was the assessment of Time to Event analysis without the employment of segmentation masks for the tumor; these models resulted in satisfying predictive accuracy and the ability to provide recommendations for individualized RT, with PET single-modality achieving the best overall performance for a segmentation-free time to event analysis.

In conclusion, these noteworthy studies reflect the progressing steps that established the evolution of Time to Event practice in the clinical field. While these are only few of the numerous Machine Learning approaches that were employed for survival analysis in medicine, they constitute a meaningful expression of the potential and effectiveness of this approach.

Indeed, in the present chapter the depiction of Time to Event analysis and its traditional methods, together with the review of some DL-based approaches, aimed at providing a deep understanding of the key role that such methodology can play in this clinical application. The suitability of this approach for the processing of clinical and imaging data, enabled by its ability to extract meaningful insights and use them to generate predictions or estimations, is currently deeper and more effective than other methods. The employment of Deep Learning techniques for Time to Event analysis, therefore, constitutes a significant progress comparing to traditional approaches and could enable new and unexpected achievements.

In closing, the proven ability of Machine and Deep Learning techniques to successfully analyze clinical and imaging data has elicited the belief that they could significantly contribute to the generation of prognoses in the context of HNSCC. Thus, the objective of the present Thesis is to explore such possibility through the implementation of various experiments and approaches, aiming to provide meaningful insights and reflections on the practical execution of this ambitious yet remarkable task.

CHAPTER 3 RESEARCH PROJECT AND EXPERIMENTAL MATERIALS

3.1. Objectives of the study

Head and neck cancers are currently estimated to affect approximately 21.8 per 100000 people in Europe with mortality rates of 15.6 per 100000 [79], and are usually treated with surgery, radiotherapy and chemotherapy or combined modalities. Long term survival of affected patients is highly influenced by possible development of Distant Metastasis (DM) and Loco-Regional Failure (LRF). Hence, a prognostic model that provides survival analysis and prediction of these events would significantly benefit these patients and could be employed for treatment recommendation in order to optimize the handling of the single subject.

In this context, the goal of this project is the development of a Deep Learning-based prediction model for DM, Loco-Regional Control (LRC, i.e., as opposed to LRF) and Overall Survival (OS) Time to Event analysis in head and neck squamocellular carcinoma (HNSCC) patients. In this study various clinical data, positron emission tomography (PET) and computed tomography (CT) volumes are employed as inputs to the model.

In the context of automated models for clinical outcome prediction, this approach involves the innovation given by the use of a combination of clinical and imaging data for the generation of the output: while the latter are universally known to hold a significant amount of information in the analysis of tumors, in this study

the clinical information is thought to hold further knowledge to the condition of the patient, and its prognostic power is therefore assessed and employed.
Besides the investigation of the clinical data, this project also aims at the evaluation of the predictive power of PET and CT imaging techniques without primary and lymph node Gross Tumor Volume (GTV) segmentation; the goal, in fact, is the development of an automated model that allows to disengage from a manual segmentation while granting an equally satisfying performance.
Our approach is based on Deep Learning techniques and involves the comparison of single- and multi-modality inputs to investigate the predictive performance of clinical data in different combinations with imaging data. To this purpose, original clinical data in combination with PET and CT volumes were employed to train three-dimensional Convolutional Neural Networks (CNN).
The networks were trained on the retrospective dataset, and used to predict individual clinical outcomes; the resulting model will then be validated during the prospective validation.
The ultimate goal is to achieve a model able to recommend a specific treatment for a single patient. The hope is to contribute to the goal of personalized medicine by allowing, with such a technology, the development of individualized therapies: conceived for the specific patient and, consequently, more efficient.

3.2. Data and patients' population

As this thesis will focus on the retrospective data analysis stage of this research, a vast dataset was made up collecting data from public repositories (i.e., cancerimagingarchive [80]) and from the collaboration with seven international universities. A large retrospective cohort was established, providing clinical and imaging data of patients affected by HNSCC from several places in the world. On this basis, several considerations were made in order to decide what patient were to be kept with the aim of obtaining the most appropriate dataset. Such considerations and the definite dataset they led to are described in the following.

3.2.1. Patients' selection and collection from centers

In this research, all the patients were united in a single, vast dataset to create a retrospective cohort of subjects that would provide clinical and imaging data.
The dataset involved 1100 patients treated with primary Chemoradiation (CRT) and 200 who underwent primary surgery at Charité Hospital in Berlin. 500 additional patients were provided by collaborating university hospitals; most of them had been treated with primary surgery. Furthermore, a dataset of 300 surgical patients was publicly available in Cancer Imaging Archive [80].
These subjects had pre-treatment PET/CT images and presented a follow-up of at least two years. All of them presented HNSCC, but the ones affected by metastasis at the time of diagnosis were excluded.
This research allows to take a closer look at the impact of different risk factors on different classes of subjects (e.g., males and females, different age groups, etc.); with this purpose, a particular attention was paid to include a sufficient number of patients belonging to underrepresented cases, in order to achieve a robust analysis.
One among the different versions of feature sets that were created and compared (as will be depicted in paragraph [3.3.3]) included radiomic features, that were extracted and evaluated through criteria of the Radiomics Quality Score (RQS); these variables are handcrafted features extracted from medical images. They were investigated for comparison in order to establish whether it is possible to exclude them and conduct a proper clinical outcome prediction basing only on clinical and raw imaging data.
Inevitably, despite the efforts to create a large dataset, the risk that it is still too small to properly train the models is present and must be investigated and accounted for. Additionally, some clinical parameters are not available for all the patients, resulting in different feature sets presenting a different number of subjects; the numerosity of the feature sets, however, was found to be similar in all the cases, and therefore assumed not to affect the comparable efficacy of the training.

3.2.2. Clinical data

The obtained dataset integrated several clinical parameters, some of which were demographics, some other referred to pretreatment variables, and some others were treatment indicators (i.e., treatment modality, radiation dosage, etc.). Considering that the aim of this research is the development of a model able to predict the clinical outcome of new, unseen patients and before their treatment, such treatment indicators were not employed in the training of the model, while a deeper attention was spent on clinical variables held as possible prognostic indicators for HNSCC.

An other aspect worth considering is that this project also aims at developing a model able to make predictions basing on raw imaging data, and therefore radiomic data were also excluded from the training of the model; in this case, however, some features sets including such radiomic data were also created in order to compare the effectiveness of the model in both cases – namely to assess whether satisfying predictions are guaranteed even without including data provided by a previous processing of imaging data.

All the considered clinical variables, were then grouped in different combinations into feature sets (paragraph [3.3.3]) in order to compare their combined predictive power.

The considered dataset was fed into the Deep Learning framework code in pickle and CSV formats. The complete CSV file, comprehensive of all the patients, is then processed by a custom function that separates it into individual CSV files for each subject; this allows the establishment of the batches to be fed into the neural networks resulting from the desired number and combination of patients.

3.2.3. Imaging data

For this study, PET and CT volumes were used as inputs to the model; consequently, patients lacking CT images were excluded from the dataset. Hence,

all patients in the dataset were provided with pre-treatment CT and PET imaging[1].

Specifically, 18F-fluorodeoxyglucose PET (FDG-PET) was employed as it proved its superiority in predictive power for staging and treatment planning of HNSCC [28].

Primary tumors affecting the patients in the dataset were semi-automatically segmented to compute radiomics [81] [82], and such parameters, when calculated basing on FDG-PET, resulted suitable to identify patients at higher risk of LRF [83].

Specifically, some of the radiomic parameters of interest are based on the concept of Standardized Uptake Value (SUV), which is a parameter that quantifies the metabolic activity of a specific volume and is extracted from PET images.

The radiomics contained in the dataset are:

- Maximum SUV: it represents the highest SUV in the Region of Interest (ROI) and therefore it refers to a single voxel. It is useful to locate the most active point inside the tumor.
- Mean SUV: it is obtained from the averaging of the SUV in the ROI. It allows the assessment of the overall metabolic activity of the considered ROI.
- Peak SUV: it provides the highest SUV computed considering a small area around the voxel that manifests the maximum SUV. It provides insights about the activity of the surroundings of the hottest point of the tumor.
- Metabolic Tumor Volume (MTV): it quantifies the extent and metabolic activity of the tumor.
- Aspericitiy: it describes the shape of the tumor by quantifying the resemblance to a sphere; the higher is such value, the higher is the degree to which the cancer deviates from a spherical shape.

[1] The devices employed for the extraction of CT and PET scans are: Gemini TF 16 (Philips Medical Systems), Discovery STE (General Electric Medical Systems), Biograph 16 PET/CT scanner (Siemens Medical Solutions Inc.), Gemini TF PET-CT (Philips Medical Systems), Biograph mCT (Siemens Healthineers), Discovery IQ (General Electric Medical Systems), Gemini TF TOF 16 (Philips Healthcare Inc.) and Discovery (General Electric Medical Systems)/ Biograph (Siemens Medical Solutions Inc.).

3.3. Feature Selection and Image Preprocessing

As elucidated in chapter 1, the traditionally major prognostic factor in HNSCC clinical outcome prediction is the tumor node metastases (TNM) staging system [11].

This value, however, while providing an important indication of the future progression of the patient's conditions, does not consider the heterogeneity of the tumor within each stage, possibly related to other prognostic indicators that are to be identified and that, therefore, have the potential to improve clinical outcome predictions.

Furthermore, more recently a significant correlation between HNSCC survival and Human Papillomavirus (HPV) has been found [6], leading researchers to investigate an approach that allows the employment of HPV status as a prognostic factor for HNC survival analysis. As explained in Chapter 1, HPV-positive subjects are found to be more responsive to chemotherapy and show a longer survival, especially when affected by oropharyngeal cancer [84]. While it seems to be related to oropharyngeal cancer more than other head and neck tumors, HPV status has predictive value for HNSCC in general, although it is limited by the high variability of survival rate within the same status [85].

Deep Learning techniques have been widely investigated for survival analysis purposes, due to the shown potential in clinical outcome prediction. These models are typically used with imaging data combined with primary and lymph node GTV masking. The GTV masking, while introducing improvements in the predictions, inevitably binds the analysis to the limits of manual segmentation (i.e., time consumption, interobserver variability).

Notably, DL-based survival analysis research was mainly related to single modality input (i.e., PET or CT); a comparative analysis of the predictive power of different modalities was still not extensively investigated, and therefore the ideal input for HNC prognosis purposes is currently not clear.

On this basis, this study aims at evaluating a DL-based approach for HNSCC Time to Event analysis without the use of GTV masking, and at identifying the best input modality for DM, LRC, OS or overall Event Free Survival (EFS) prediction.

In order to achieve this, it is crucial to conduct a careful analysis of the predictive power of each possible clinical and radiomic feature, with the hope that their combination allows a stronger predictive power than mere HPV-status and stage. Indeed, even though this research aims at developing a model able to accurately predict clinical outcomes basing on raw clinical and imaging data, some prior manipulation is still necessary. As mentioned, it is in fact crucial to identify, among the pool of variables, the clinical parameters that have the highest informative power; once this selection is done, the most appropriate ones will be fed into the prediction model. This investigation was conducted through the analysis of Kaplan Meier curves, and will be described in the following paragraph. For what concerns the imaging data, as stated, the model will receive un-masked volumes; no manual intervention will be needed as they will be automatically segmented, however some minor preprocessing proved advantageous to the objective. Such preprocessing will be described in paragraph [3.3.2].

3.3.1. Kaplan Meier curves analysis

Kaplan Meier (KM) estimator allows the visual investigation of the influence of a clinical parameter on a specific condition. Therefore, its employment guaranteed the selection of the most informative features to introduce in the model with the aim of OS, LRC and DM prediction.
Specifically, the main focus was directed to the analysis of the role of gender, age, UICC stage, HPV status and tumor site in the development of these events. This choice resulted from the goal of developing a prediction model that can be applied to unseen and untreated patients, and therefore should only be fed with data that don't require treatment information or imaging scans processing.
In order to conduct this analysis, for each clinical feature the Kaplan Meier curves corresponding to different subgroups were drawn and subsequently compared through a log-rank test in order to assess their statistical correlation. This procedure was separately executed for each event of interest (i.e., EFS, LRC, DM)

in order to investigate the possibility of features having different impact on each of them.

When observing the Kaplan Meier curves, the critical aspect to consider is their superimposure: the farther the subgroups' curves are, the more different are their survival times; therefore, distanced KM curves indicate that the considered subgroups were established according to a discriminating factor and, therefore, a prognostic indicator. On the other hand, overlapped KM curves indicate a similar trend in survival of different subgroups and, therefore, the considered stratification cannot guarantee to benefit risk prediction.

On the superimposed KM curves median survival times are also indicated as dashed, vertical lines; for each subgroup, these values represent the timepoints corresponding to a 50% probability of developing the event. Hence, these values effectively provide a further factor to quickly and comfortably compare different subgroups. It is worth highlighting that such dashed lines will only be present on KM curves corresponding to subgroups in which 50% of the patients have experienced the event within the follow-up time.

The subsequent log-rank test allows the quantification of such considerations through the comparison with a 0.05 p-value threshold to assess the statistical correlation of the stratification of interest and the risk of EFS, DM or LRC.

Gender

The impact of the patient's gender on Event Free Survival (EFS) was investigated through the drawing of separate KM curves for males and females (as shown in Figure 3:) and their comparison through log-rank test.

Through a quantitative visual investigation of the superimposed curves, it is possible to assume that gender does not seem to affect appreciably the probability of developing any kind of event (i.e., EFS). Such assumption is in fact confirmed by the log-rank test, that results in a 0.59 p-value; as such value is significantly higher than the 0.05 established threshold, this test confirms that EFS and gender, basing on this dataset, are not statistically correlated.

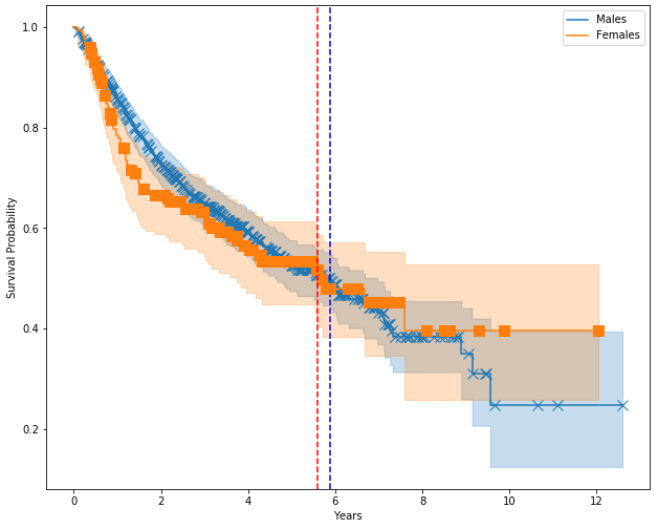

Figure 3: superimposed Kaplan Meier curves according to gender stratification for EFS probability estimation; the vertical lines indicate the median survival time of each subgroup.

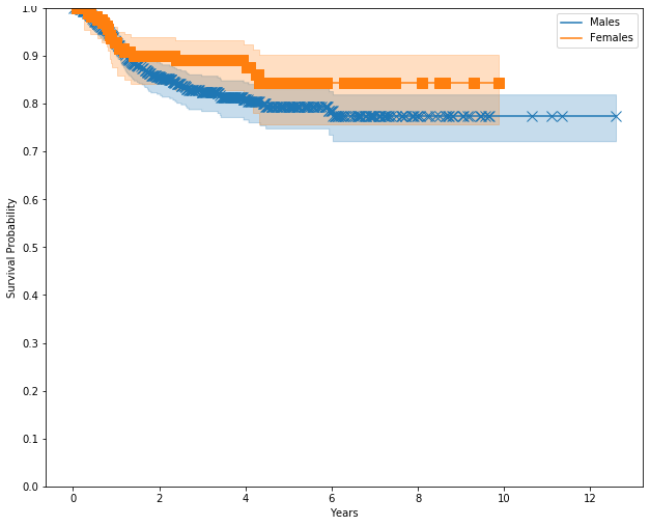

Figure 4: superimposed Kaplan Meier curves according to gender stratification for DM probability estimation.

The same analysis was conducted to investigate DM and LRC development and, unsurprisingly - as EFS integrates both DM and LRC - provided the same results: KM estimator (presented in Figure 4 and Figure 5) and log-rank test confirmed that such occurrences are not affected by the gender of the subject.

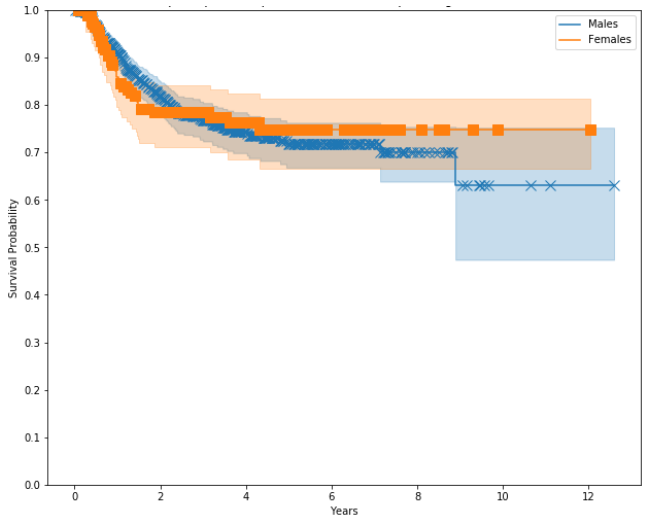

Figure 5: superimposed Kaplan Meier curves according to gender stratification for LRC probability estimation.

Age

The effects of age on the probability of EFS, DM and LRC were also evaluated through the same procedure. In order to achieve such evaluation, patients were divided by age into 5 subgroups:

- younger than 40 years old;
- between 40 and 50 years old;
- between 50 and 60 years old;
- between 60 and 70 years old;
- older than 70 years old;

as reported in Figure 6, for each subgroup the corresponding KM curve was drawn, and all their combinations were compared through log-rank test.

It is noteworthy that the group containing patients younger than 40 years old is significantly less numerous than the others, and this could highly affect and distort the statistical analysis; this is consistent with the blue shaded area, indicating the confidence interval of survival estimations of this subgroup, that appears significantly large and overlaps the majority of the other curves.

The subsequent lor-rank tests confirmed such consideration: no statistical correlation was, in fact, found between patients younger than 40 and patients of

any other age; as statistical discrimination was found between other age groups (i.e., even in patients closer in age), it is reasonable to ascribe this to the numerosity of the subgroup.

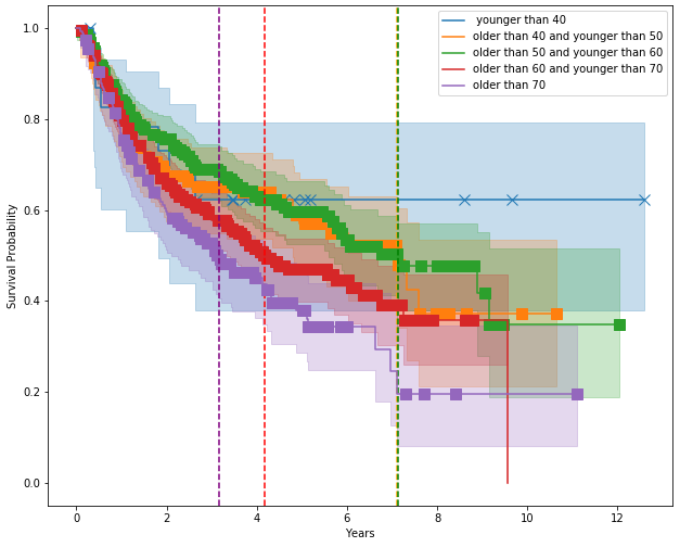

Figure 6: superimposed Kaplan Meier curves according to age stratification for EFS probability estimation; the vertical lines indicate median survival times of the corresponding subgroup.

Indeed, the test proved that, for patients younger than 50 years, a gap of at least 20 years is associated with a higher risk of developing an event; in the other hand, for patients older than 50 years, even a 10-year gap is enough to increase the risk of occurrence.

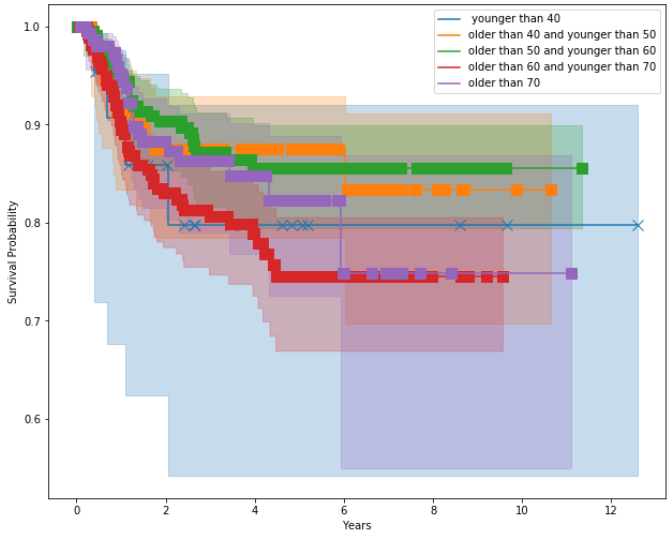

Figure 7: superimposed Kaplan Meier curves according to age stratification for DM probability estimation.

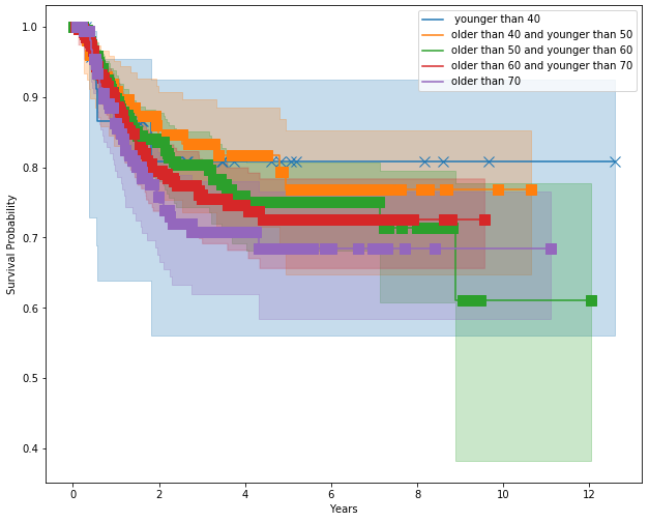

Figure 8: superimposed Kaplan Meier curves according to age stratification for LRC probability estimation.

For what concerns DM and LRC probability, reported in Figure 7 and Figure 8 respectively, the log-rank test only proved a statistical correlation between patients aged in ranges [50,60] and [60,70] years, with a higher risk of occurrence for older patients.

UICC stage

KM curves for EFS probability, reported in Figure 9, appear highly overlapped for different UICC stages. Indeed, the investigation of the impact of UICC on EFS assessed that there is no statistical correlation between them (i.e., p-value>0.05).

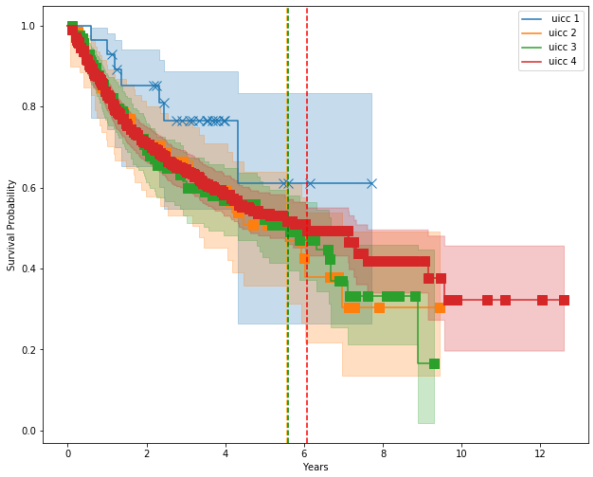

Figure 9: superimposed Kaplan Meier curves according to UICC staging stratification for EFS probability estimation; the vertical lines indicate median survival times of the corresponding subgroup.

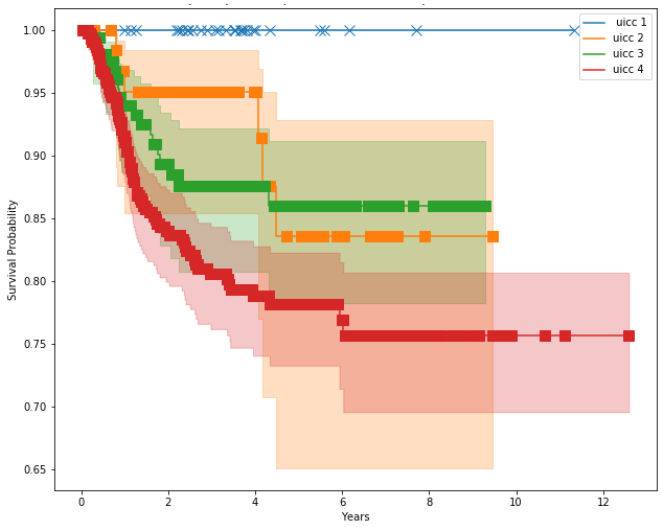

Figure 10: superimposed Kaplan Meier curves according to UICC staging stratification for DM probability estimation.

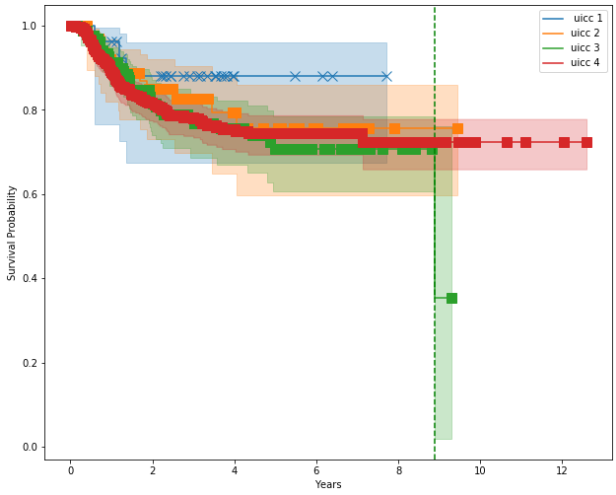

Figure 11: superimposed Kaplan Meier curves according to UICC staging stratification for LRC probability estimation; the vertical lines indicate median survival times of the corresponding subgroup.

However, as shown in Figure 10, this feature plays a role in the probability of developing DM; specifically, UICC stage 4 is statistically correlated with a significantly higher probability of developing DM when compared to UICC 1 and 3. On the other hand, such influence was not detected when considering KM curves for LRC probability, presented in Figure 11.

In conclusion, while UICC staging does not seem to significantly affect LRC probability, a higher stage proved to increase the risk of DM development.

HPV status

As shown in Figure 12, HPV status seems to significantly affect EFS probability in HNSCC patients. A visual investigation of the KM curves, in fact, highlights the clear detachment between the HPV-positive and -negative curves; indeed, in accordance with the clinical evidence, HPV-positive HNSCC seems to provoke much more favorable prognosis. Therefore, as expected, such distant curves imply that HPV status can be successfully employed to predict the risk of developing an event.

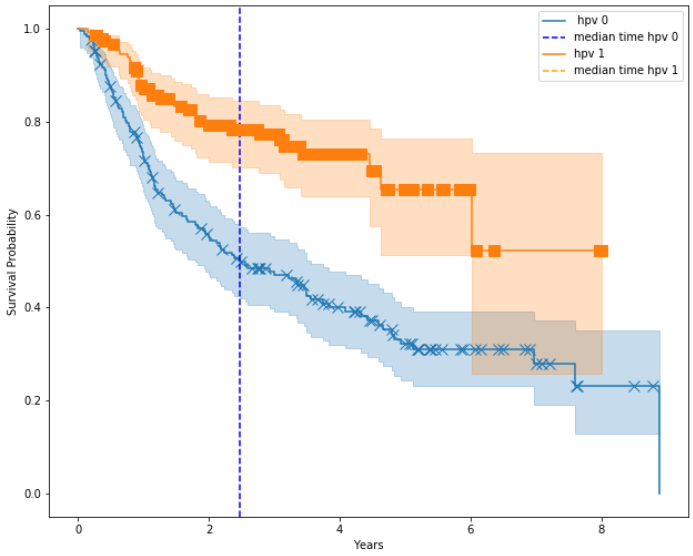

Figure 12: superimposed Kaplan Meier curves according to HPV status stratification for EFS probability estimation; the vertical lines indicate median survival times of the corresponding subgroup.

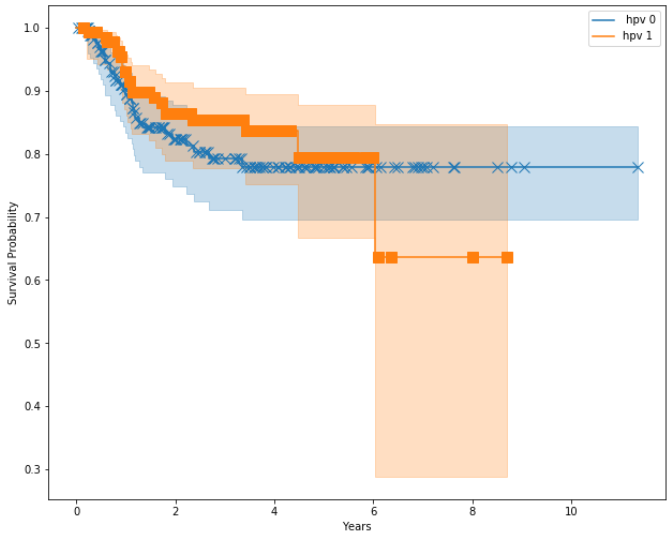

Figure 13: superimposed Kaplan Meier curves according to HPV status stratification for DM probability estimation.

The same analysis was conducted to investigate HPV's role in DM and LRC risk. More specifically, a statistical correlation was found with the probability of LRC (shown in Figure 14) but not DM (Figure 13). Interestingly, while it does not affect DM development risk, HPV-negative HNSCC appears to determine a significantly higher risk of tumor relapse.

Hence, it is reasonable to assume that the strong influence of this clinical variable on EFS probability was due to its relationship with LRC risk.

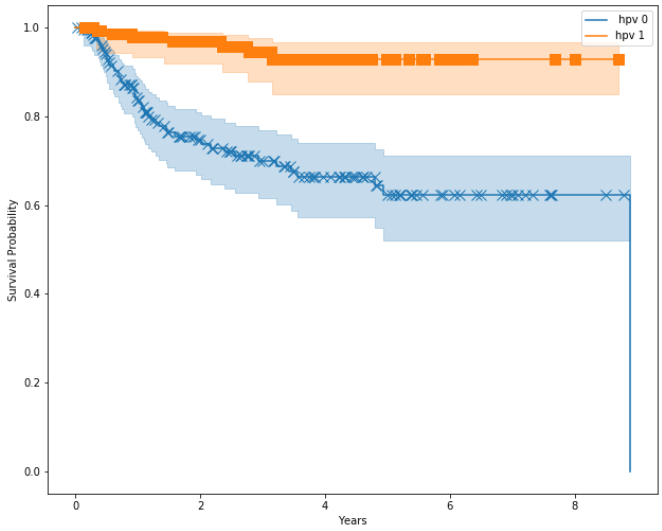

Figure 14: superimposed Kaplan Meier curves according to HPV status stratification for LRC probability estimation.

Tumor site

The influence of the anatomical location of the tumor on EFS, DM and LRC was also investigated. The considered subgroups are the following:

- 1 = Hypopharinx HNSCC;
- 2 = Larynx HNSCC;
- 3 = Nasopharynx HNSCC;
- 4 = Oropharynx HNSCC;
- 5 = Oral Cavity HNSCC;
- 6 = Paranasal Sinuses Carcinoma
- 7 = HNC of Unknown Primary;
- 0 = other HNSCC;

As shown in Figure 15, the tumor site appears to highly impact EFS, but it is important to consider the strong underrepresentation of sites 0, 6 and 7. However, the conclusion regarding the other sites can be considered reliable. Specifically, the log-rank tests confirmed a statistical correlation between the anatomical location of the tumor and the EFS probability that could be summarized through a ranking of the tumors in an increasing order of event risk development: nasopharynx, oropharynx, larynx, hypopharynx, oral cavity.

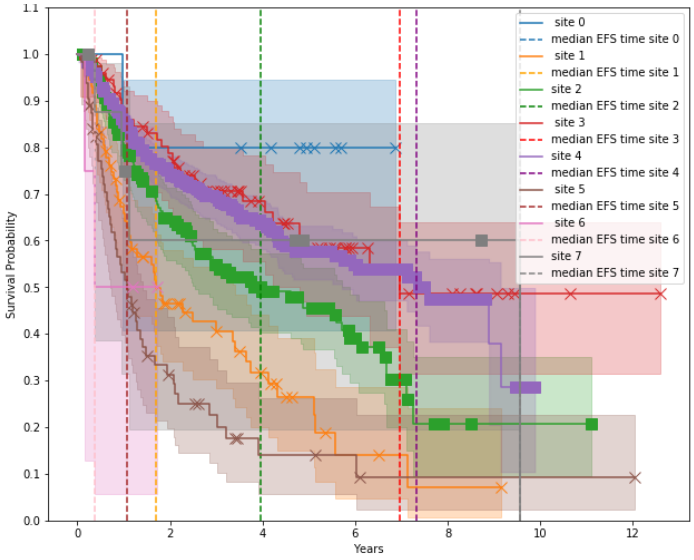

Figure 15: superimposed Kaplan Meier curves according to tumor site stratification for EFS probability estimation; the vertical lines indicate median survival times of the corresponding subgroup.

This analysis also highlights that hypopharynx, larynx, oropharynx and oral cavity affect differently the probability of DM development (Figure 16); specifically, while this event appears less influenced by the tumor sites, oral cavity HNSCC presents a statistically significant correlation with such occurrence as it is associated with a higher risk.

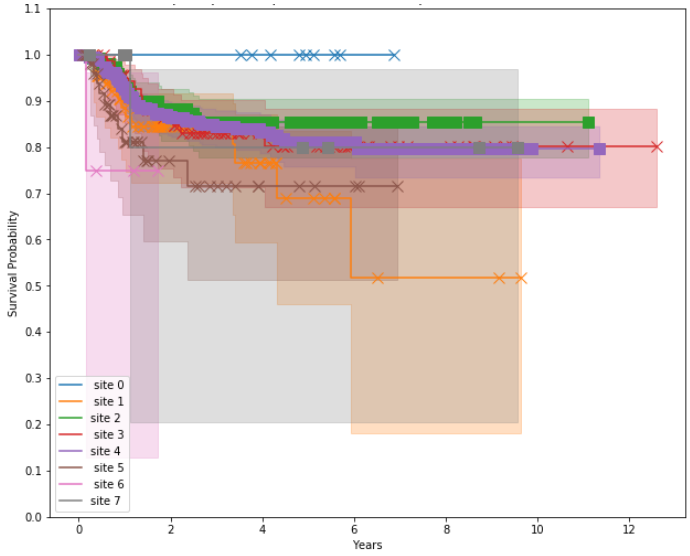

Figure 16: superimposed Kaplan Meier curves according to tumor site stratification for DM probability estimation.

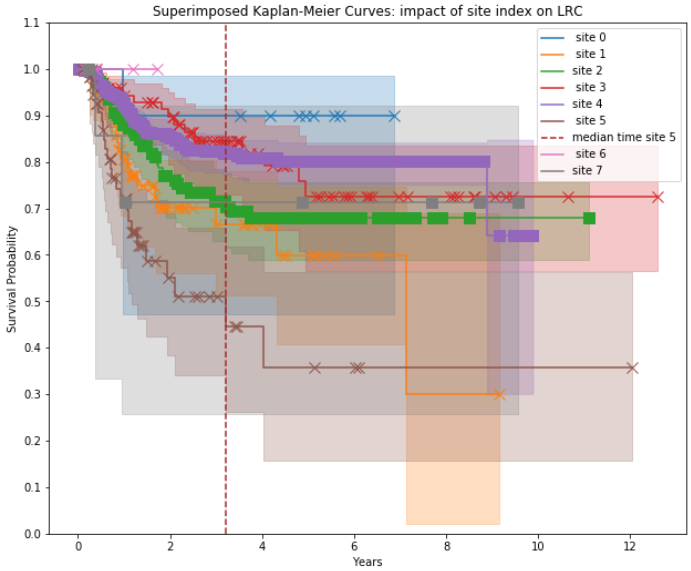

Figure 17: superimposed Kaplan Meier curves according to tumor site stratification for LRC probability estimation; the vertical lines indicate median survival times of the corresponding subgroup.

As shown in Figure 17, instead, when considering LRC probability a clear correlation with the tumor anatomical location is visible; LRC KM curves present the same definite trends as the EFS ones. Hence, the same risk stratification as EFS probability can be made for LRC as the latter determines the former; specifically, oral cavity HNSCC involves a significantly higher risk of tumor relapse.

In Table 1 a summary of the conclusions provided through the previous analysis is reported.

	EFS	***DM***	***LRC***
GENDER	No predictive value	No predictive value	No predictive value
AGE	Some predictive value	Slight predictive value	Slight predictive value
UICC	No predictive value	Some predictive value	No predictive value
HPV STATUS	Strong predictive value	No predictive value	Strong predictive value
TUMOR SITE	Strong predictive value	Some predictive value	Some predictive value

Table 1: predictive value of clinical features estimated through Kaplan Meier curves and log-rank test.

It is worth highlighting that, while this analysis was useful as it provided valuable insights on the influence of each of the considered clinical features on each of the events of interest, in this investigation the variables were considered individually; indeed, this procedure is to be considered as a preliminary step as it did not provide any consideration regarding the informative value of their combination. Subsequently, even the features that proved less informative in respect to the considered events, were still employed as inputs for the predictions as DL-based models are widely recognized to be effective in finding – if present - patterns among data. Therefore, experiments carried out with different feature sets will assess whether the clinical features provide a higher predictive value when combined.

3.3.2. Feature sets

In this study, different kinds of datasets were investigated and compared. Starting from the initial dataset, seven different feature sets of interest were built by excluding different combinations of features. The feature sets are shown in Table 2, and each feature is briefly explained in the following index:

- age of the patient;
- gender of the patient;
- UICC stage;
- site: anatomical location of the tumor;
- HPV status;
- MTV: Metabolic Tumor Volume of the tumor;
- MaxSUV: maximum SUV;
- asphericity of the tumor;
- Tstage: T value according to TNM staging;
- Nstage: N value according to TNM staging;

In order to employ HPV status, a minor preprocessing was necessary. Several patients (i.e., 570) did not present such clinical variable and therefore, in order

to avoid the exclusion of so many subjects, such missing feature was filled with a -1 value, as opposed to 0 (i.e., HPV negative) and 1 (i.e., HPV positive). This manipulation allowed the integration of all the patients while introducing a significant difference in value when compared to patients associated to a known HPV status.

V1	V2	V3	V4	V5	V6
Age	Age	Age	Age	Age	Age
Gender	Gender	Gender	Gender	Gender	Gender
	UICC	Site	UICC	UICC	UICC
			Site	Site	Site
				HPV	MTV
					MaxSUV
					Asphericity
					Tstage
					Nstage

Table 2: features sets employed to experiment the prediction model.

3.3.3. Image preprocessing

When considering imaging data, the adoption of a windowing technique was tested: basing on the hypothesis that such manipulation could increase the performance of the model, some experiments to assess this aspect were carried out. This technique involves the processing of the images in order to enhance the visibility of the object of interest, in this case the tumor. As each pixel that composes the image presents a value, it is changed in order to improve the overall brightness and contrast levels. To achieve this, two parameters are used: window level and width; they respectively determine the new midpoint and width of the consented range of values for the pixels. Such values refer to Hounsfield Units (HU), which provide the radiodensity of tissues in the body; specifically, denser tissues are associated with positive and high values HU and vice versa.

Hence, research in literature was conducted in order to identify the best range of HU to employ for this specific application of windowing. Such investigation suggested, firstly, to only employ this technique on CT and not on PET volumes, and secondly, it indicated [50,160] as the range of HU values manifested by HNSCC [86]. Consequently, several experiments were conducted and resulted in a [0,190] range windowing to provide a higher performance of the model compared to the employment of other ranges and absence of windowing. Keeping some margin around the values of interest guarantees the preservation of all the possibly useful information allowing some small deviance from the expected values.

3.4. Implementation of Time to Event Analysis for Head and Neck Squamous Cell Carcinoma

In the context of HNSCC, Time to Event analysis has the potential to effectively predict different events of interest: Overall Survival (OS), Loco-Regional Control (LRC) and Distant Metastasis (DM) development. Such achievement is the aim of extensive research due to its crucial impact on the overall conditions of the patients; local recurrence and distant metastasis are, in fact, major threats to the survival of the affected subjects.

Specifically, such phenomena weight directly and indirectly on the future conditions of the patient: while indicating the severity of the tumor and therefore directly affecting their health, they also highly influence the employed treatment approach that, in turn, have important physical effects on the patients. Indeed, cancer treatment always involves an amount of harm to the patient's body in the attempt of targeting the tumor. While such side effects are inevitable, their entity can significantly range, depending on several factors, from minor to devastating – involving damages to certain functionalities or even death.

As a result, a method that allows the accurate prediction and location in time of OS, LRC and DM could significantly and positively influence the handling of HNSCC patients and, consequently, highly improve their prospects.

As the relevance of such achievement is widely recognized, the efforts applied on the related research become clear and, among the several studies dedicated to this goal, the present thesis project can be found.
In order to address such a complex challenge as the Time to Event analysis of HNSCC patients, a versatile Python-based Deep Learning framework was developed. This system was designed with the aim of providing the flexibility necessary to the accommodation of different tasks (i.e., classification, regression and Time to Event) and modalities. Much effort was made in order to allow an efficient yet comfortable configuration of the experiments by allowing the choice of several parameters depending on the specific use; indeed, while several aspects of the prediction system remain unchanged throughout the different tasks, some more specific characteristics are to be chosen according to the specific intended use.
Hence, in the following paragraphs a detailed description of our implementation of Time to Event analysis is presented: starting point, development and details of implementation of the model are extensively described; a comprehensive depiction of how the model was tailored to the specific task will be provided together with a discussion of the available options provided to suit different types of experiments.

3.4.1. Benchmark paper

The first step in the execution of this study was a literature review, which was outlined in paragraph [2.3.4]. Specifically, among the several examined papers, one provided our starting point and benchmark; such paper depicts Wang et al. work [78]. As described in the literature review, the authors developed a ML model for the Time to Event analysis of OS and DM in HNSCC patients; this study successfully assessed the possibility of conducting such predictions basing on CT and PET volumes without the employment of segmentation masks for the tumor. As this study provided promising results but also some limitations, it was used as benchmark for the present work. In particular, while Wang et al.'s work provided

noteworthy conclusions, a further step can be taken: the aim of the present research, is to improve such practice by integrating in the inputs pool some clinical data.
As mentioned, in fact, such prior work only based the predictions on CT and PET images; and while the resulting predictions are satisfactory, the strong prognostic value of some clinical variables (e.g., HPV status, tumor staging) led to the hypothesis that by integrating such information the achievement of an even better performance is possible.

3.4.2. Data preprocessing

As previously described, the employed input data were provided by gathering clinical variables of the retrospective cohort. Specifically, such dataset was composed by clinical features of the patients and the data collected during their follow-up of the duration of at least two years.
Consequently, the data related to the clinical outcomes provided the occurrences of LRC, DM and OS that were registered at each visit during the follow-up. Such medical appointments did not occur at regular intervals, hence some prior standardization was required. Indeed, during all the phases of the establishment of the DL model (i.e., training, validation, testing) the labels related to each patient are required for each timepoint: they represent the true values that the model is supposed to predict. In the case of a Time to Event analysis of LRC, DM and OS for HNSCC patients, such labels must be configured as binary values that indicate the occurrence of the event at each timepoint. In order to extract them, the original clinical data were manipulated.
Specifically, as the focus of the research was the execution of a Time to Event analysis with 6 months interspersed timepoints, the times of the registered occurrences were propagated accordingly in order to regularly provide an update of the conditions of the patients.
By applying such manipulation, the registered outcomes were forced to the desired interval width.

In case of patients who developed an occurrence before the end of the fourth year, the related information (i.e., presence and time of the event) was propagated until the end of the minimum follow-up length (i.e., 4 years).

An example is provided in Figure 18 for clarity.

timeEFS	statEFS	timeEFS12	statEFS12	timeEFS24	statEFS24	timeEFS36	statEFS36	timeEFS48	statEFS48
16,5	1	12	0	16,5	1	16,5	1	16,5	1
52,9	0	12	0	24	0	36	0	48	0
6,6	1	6,6	1	6,6	1	6,6	1	6,6	1
19,3	1	12	0	19,3	1	19,3	1	19,3	1
36,1	1	12	0	24	0	36	0	36,1	1

Figure 18: table extracted by the clinical dataset where each row regards one patient. Variables "timeEFS" and "statEFS" indicate, respectively, when and whether the event occurred (i.e., statEFS=1 states that either DM, LRF or death occurred). Subsequently, the following features describe whether and when the event had occurred at the yearly timepoints reported in months (i.e., 12, 24, 36, 48 months).

This precaution allowed the obtainment of the required data for the feeding of the DL model, as the information for each patient and at each timepoint is necessary. Furthermore, a significant peculiarity of Time to Event analysis is the obligation of handling censoring cases. Such phenomenon, that due to its important influence was extensively described in Chapter 2, determined the lack of a portion of the original clinical data. A part of the patients, in fact, is expected to stop taking part in the follow-up (i.e., censor); consequently, such patients will not provide clinical information after a certain timepoint. While handling inappropriately these cases would mislead the model and therefore determine significant errors in its future predictions, their complete exclusion from the dataset would imply a severe loss of information. A strategic compromise is, therefore, necessary.

Specifically, as previously stated, Time to Event analysis allows to address such problem, as this technique was specifically designed to also account for this factor. As a result, one last touch to the clinical data was necessary to gather all the required inputs: the labels had to indicate, somehow, the event of censoring; this way, the model could recognize this occurrence and consequently handle it. This was implemented through the creation of further binary labels that, for each timepoint, indicated censoring. Specifically, such variable called “Cens” reported, every 12 months, whether the patient was present (i.e., Cens=1) at the clinical

appointment or censored (i.e., Cens=0). As this analysis only focuses on the occurrence of the first event, the successive information is not taken into account and therefore the censoring feature is null after the timepoint of the first occurrence.

This procedure provided the definite dataset; an example is shown in Figure 19.

	Cens12	Cens24	Cens36	Cens48	statEFS12	statEFS24	statEFS36	statEFS48
A	1	1	1	0	0	0	0	0
B	0	0	0	0	1	0	0	0
C	1	0	0	0	0	1	0	0
D	1	1	1	1	0	0	0	0

Figure 19: example of censoring and statistics labels provided in months. Patient A: censored between 36th and 48th month with no event development (i.e., statEFS=0) until time of censoring; patient B: development of an event before 12th month; patient C: development of an event between 12th and 24th month; patient D: event-free survival throughout the entire 4 years follow-up (i.e., did not censor nor develop any event).

In conclusion, the integration of the clinical outcome labels with the censoring labels allowed the model to learn whether the absence of detected event was real or due to the patient not participating in the clinical visit.

3.4.3. Data standardization

The employed framework provided the possibility of clinical data standardization. Specifically, a standardized scaling was implemented in order to obtain a new data distribution characterized by a null mean and unitary standard deviation.

This method can enhance the performance of the algorithm in case of features presenting variegated scales. Different ranges of values, in fact, could lead the model to the assumption that they have a different weight on the establishment of the output. Consequently, this technique simplifies the interpretation of the data for the model by making the different kinds of features more comparable; as a result, it allows the model to investigate the actual relative significance of the variables.

This approach also allows to address the outliers reducing their influence on the training of the model, and it permits a more informative computation of distance metrics.

In the implementation of our model, standardized scaling was considered for the processing of the following numerical clinical data: age, tumor volume (i.e., MTV), maximum SUV, asphericity.

3.4.4. Network architectures

As previously mentioned, the possibility of effectively combining different neural networks was exploited in this study. Specifically, clinical data were fed into an ad hoc structured Clinical Neural Network, while imaging data were input into a Convolutional Neural Network. This approach allows the handling of each kind of input data with the most appropriate network, resulting in a more effective training and, therefore, a performant model.
As will be explained in the following paragraph [3.4.5], these networks will then be combined according to the desired modality.

Clinical Neural Network

The Clinical Neural Network is the portion of the algorithm that will only handle and process the clinical data, and it is therefore created basing on the desired clinical inputs.
In the structuring of this network, the input layer is established in accordance with the data (i.e., input shape, batch size). The subsequent hidden layers are created basing on the chosen configuration, which specifies the number of layers and neurons. These values were chosen basing on the optimization technique that will be described in paragraph [3.6] in order to select the most effective structure for the Clinical Neural Network.
Each hidden layer is a fully connected layer, and the activation function applied to the layer's output is a ReLU[2]. Kernel and bias regularization terms are applied to the layer's weights and biases. A dropout layer is introduced after each dense layer: this regularization technique randomly ignores 10% of the neurons during

[2] ReLU activation function is described by the following equation: $f(x) = \max(0, x)$.
This implies that all the positive values provided by the previous layer remain unchanged, while all the negative ones are substituted by a 0.

training in order to prevent overfitting. Depending on the task and modality, a certain final layer is added to the network; in order to implement Time to Event analysis, the number of neurons in such last layer is equal to the number of predictions for each patient (i.e., one every 6 months for the desired number of years). A sigmoid activation function is employed to process the obtained values and the same kernel regularizer as the hidden layers is used to prevent overfitting.

Convolutional Neural Network

When building the CNN, the input layer is defined basing on the imaging data shape, which is (128,128,140).

The first convolutional layer has a 5-sized kernel, while the subsequent ones present size 3. In each layer the number of filters is doubled, and the initial number was set to 16. After each convolutional layer, the same dropout layer as the Clinical NN was added to prevent overfitting. Lastly, as the final layer only depends on the task, the same one as the Clinical NN is appended to the network.

3.4.5. Modalities

The term “modality” refers to the method used to combine the different neural networks. This factor is correlated with the number and kinds of desired inputs for the model and affects the way the different data will be joint together. Since this aspect affects the amount of data that are simultaneously fed into the model, it can influence its development and performance.

The modalities implemented in the framework code are “single”, “multiple” and “merged”, and will be described in the following.

Single modality

In single modality, the data employed as inputs are of a single kind (i.e., clinical, PET or CT). In this case, one type of input data will be fed into the corresponding network (i.e., Clinical NN or Convolutional NN) and therefore the model only presents one channel.

Consequently, this configuration is only appropriate for experiments in which either clinical or one kind of imaging data are employed. Such setting is therefore mainly considered for investigative analysis purposes (e.g., assessment of the informative power of each input type separately).

Merged modality

Unlike the previous case, with merged modality different kinds of imaging input (i.e., CT and PET), are combined before being fed into one single convolutional network. Such merging was implemented by concatenating the 3D arrays representing the imaging volumes along the fourth dimension. The resulting merged 4D arrays were then saved as NIfTI files and consisted in 3D imaging volumes, in which each element contained two values representing CT and PET data; therefore, in this case the model presents two channels.

The employment of a merged modality CNN allows to analyze the informative power of all imaging data at once.

Multiple modality

In multiple modality different kinds of data are fed to separate networks that are merged at the final layer. This powerful approach allows to process each kind of data with a different network, namely the most suitable for the specific data (i.e., Clinical NN for clinical data, Convolutional NN for imaging data) and subsequently combine the resulting information.

Each network processes its data and provides its output features that are then concatenated along the last dimension into a single features vector that, in turn, will be fed into a single final layer. An example is shown in Figure 20 [61].

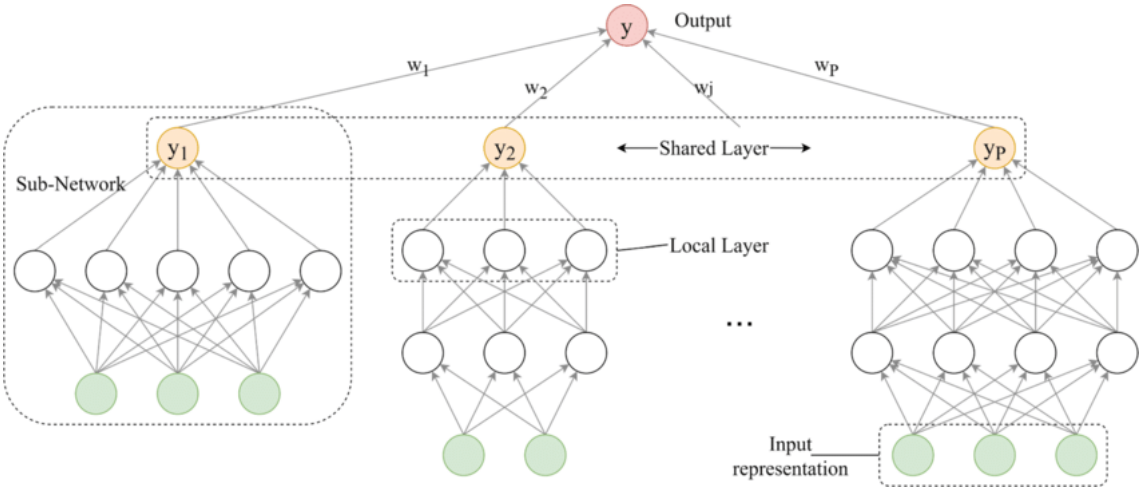

Figure 20: example of multiple modality network.

This modality allows the combination, for instance, of one Clinical NN with one CNN or of two different CNNs that process, respectively, CT and PET data; these kinds of experiments are likely to provide enhanced predictions when compared to single modality due to the integration of the informative power of both clinical and imaging data.

While such approach is powerful, this modality has the potential of taking an even further step by defining a multiple-modality NN that combines a Clinical NN with a Merged imaging CNN; hence, the informative value of clinical, CT and PET data would be combined and simultaneously fed into one single predictive model.

Indeed, while using different modalities to analyze the influence and informative power of some kinds of data provides fundamental insights, this last method is the strongest and is expected to provide the best performance as it exploits all the available data, and is therefore the ultimate modality chosen for the validation of this research.

3.4.6. Loss function and metrics

The loss function chosen for Time to Event analysis of HNSCC patients, is a custom Keras function provided by Michael Gensheimer and available on Github named surv_likelihood() [86]. Such function quantifies the errors made by the model by comparing predicted survival times with ground truths. Specifically, the

model predicts the probability with which the subject is estimated not to experience the event at each timepoint.

This loss calculates the negative log-likelihood, namely the negative logarithm of the predicted probabilities. Such approach allows the quantification of how accurately the model predicts survival probabilities while considering both censored and uncensored patients. By minimizing such loss function, the optimization algorithm can improve the model parameters and therefore its performance.

Furthermore, Harrel's Concordance Index (HCI) was computed and employed to assess the performance of the ultimate model (i.e., resulting from training, validation and testing phase). This parameter is a statistic commonly used to assess the predictive accuracy of a survival analysis model. For each pair of patients, it compares their probability of survival within a specific time interval. A pair of subjects is considered "concordant" if the first who experienced the occurrence also presents a higher predicted risk; conversely, the pair is referred to as "discordant". It is subsequently possible to compute the HCI as follows:

$$\text{HCI} = \frac{\text{Number of concordant pairs}}{\text{Total number of pairs}}$$

HCI ranges from 0 to 1, and the higher is the value, the higher is the predictive accuracy.

This index, while being effective in the evaluation of the model accuracy, mainly owns its popularity in Time to Event analysis due to its ability to account for censored cases; this critical aspect, in fact, is what prevents the use of several other performance indexes that would not allow the consideration of the incomplete information introduced by censoring patients.

When dealing with censoring, in fact, HCI allows to integrate the predictions of the corresponding cases, and therefore avoid the loss of potential information, while simultaneously guaranteeing not to introduce biases. In order to achieve this, an initial categorization of the pairs is established: the concordant pairs are only assessed among "comparable pairs", which refer to cases in which one of the individuals experienced the event before the other censored; this allows to

consider the valuable information introduced by the censored patient (i.e., until the moment of censoring, they are known not to have experienced the occurrence) – this individual can therefore be compared to another non-censored individual without the risk of biasing the predictions.
In conclusion, this parameter allows the evaluation of the model's predictive accuracy by assessing its ability to rank the subjects according to their survival times.

3.4.7. Data postprocessing

As previously described, HCI provides valuable information about the model's predictive accuracy, and was therefore employed as the main parameter for the evaluation and comparison of the several implemented models.
Nonetheless, there are further considerations worth making.
While the HCI does effectively indicate the model's ability to conduct accurate predictions, it is based on the evaluation of its propensity to correctly rank the subjects according to their survival times; namely, a high HCI model will estimate longer survival times for lower-risk patients and vice versa. This approach is powerful and effective in the assessment of predictive accuracy, however it can also be considered as the basis to conduct some further analysis.
Indeed, provided that the model outputs the estimated probability with which each patient will experience the event at each timepoint, such outputs will require some further processing and, therefore, an ulterior accuracy metric to evaluate the new – postprocessed – predictions.
In order to develop the most effective procedure, it is firstly necessary to define the most appropriate postprocessing and gain, therefore, adequate knowledge for the choice of the metric.
As previously mentioned, the model provides as outputs the probability of developing the occurrence at each timepoint; as the goal of Time to Event analysis is the provision of a binary variable that indicates, at each moment of interest, the occurrence of the event, such probabilities must be binarized.

In order to achieve this, a threshold has to be defined to discriminate the probabilities and associate the lower values to 0 and the higher values to 1. Such threshold strongly depends on the model and the characteristics of its predictions; the more balanced the model, the closer the threshold is to 0.5, while if it tends to overestimate the patients' risk (and therefore to provide short estimations of survival times) it will be lower and vice versa.
Such threshold will be chosen in order to maximize the number of true positives and negatives.
This procedure allows the obtainment of a binary variable representing the occurrence of the event at each timepoint. This leads to the possibility of considering this analysis like a multiple classification (i.e., in classes 0 and 1), executed once for each timepoint. On this basis, it is possible to evaluate such new outputs like those of a classification and, therefore, a Time-Dependent Area Under the Receiver Operating Curve (AUC) was chosen for this purpose. Such parameter represents the ability of the model to distinguish between subjects who experience the event and those who don't. As the ROC shows the proportion between True Positive (TP) and False Positive (FP) rates in function of the decision threshold of the model, its area (i.e., AUC) reflects the overall predictive accuracy of the model; the better is the ratio of TP and FP, in fact, the higher is the AUC. In conclusion, computing the AUC at each timepoint, provides insights of the model's discriminant power and therefore its ability to correctly classifying a patient as at-risk.
A twofold analysis was conducted: both Time-Dependent and Cumulative AUC were computed. The former separately considers the comparison between ground truth and prediction at each timepoint, while the latter accounts for propagated predictions and labels; this means that even if an event was predicted in advance in respect to reality, the classification will be considered correct in the timepoints equal to and succeeding to the correct one.
In conclusion, while Time-Dependent AUC reflects the ability of the model to predict the event and locate it at the right time, the Cumulative AUC additionally considers acceptable the advanced predictions. This implies a more conservative

use that, in the context of a clinical application, could prove advantageous; it would involve, in fact, the acknowledgement that even though technically inaccurate, a slightly shorter, rather than longer, estimated survival time could still highly benefit the handling of the patient.

3.5. Alternative approaches

With the aim of conducting a comparative analysis of the effectiveness of different ML-based models for clinical outcome prediction for HNSCC patients, other approaches were implemented during the course of this study. The new attempted methods were analyzed basing on the assumption that they are a simplification of the task comparing to the Time to Event analysis.
Specifically, the new implemented tasks are classification and regression of clinical outcomes in HNSCC, and they were developed basing on the same dataset as the Time to Event analysis.
The mentioned simplification lies in the fact that, while Time to Event analysis provides one clinical prediction at each timepoint, both these new techniques only aim at providing one output, namely the recurrence of the event in case of classification, and the estimated time before its occurrence in case of regression – while Time to Event task attempted to predict both.
While implementing classification and regression, a new feature set was also established with the aim of further enhancing the predictive power of the model; such feature set was named v7 and is reported in Table 3. Such new feature set results from the integration of v6 with the "ChemoIndex"; this feature indicates the kind, if any, of chemotherapy that was chosen for the specific patient. As we originally aimed at assessing the possibility of a Time to Event analysis model to be used on pretreatment data of the patient, we did not initially integrate such feature. This approach was, therefore, subsequently chosen in order to evaluate whether some information regarding the therapy would introduce an improvement in the performance of the model.

V1	V2	V3	V4	V5	V6	V7
Age	Age	Age	Age	Age	Age	Age
Gender	Gender	Gender	Gender	Gender	Gender	Gender
	UICC	Site	UICC	UICC	UICC	UICC
			Site	Site	Site	Site
				HPV	MTV	MTV
					MaxSUV	MaxSUV
					Asphericity	Asphericity
					Tstage	Tstage
					Nstage	Nstage
						ChemoIndex

Table 3: new feature set V7 shown together with the other ones for comparison.

Such new approaches required some adaptation of the data and considerations in order to be executed; this procedure, together with the implementation details, will be described in the following.

3.5.1 Classification of Head and Neck Squamous Cell Carcinoma

Classification is a procedure that involves the categorization of input data into predefined classes basing on specific features. While this is a fundamental concept in several field of application, it found its major fit in the context of Machine Learning; this technology, in fact, proved highly effective in executing such procedure. Among the numerous applications where this technique can introduce significant advantages, medicine stands out. Specifically, classification can be employed for clinical outcome prediction and can therefore be considered an appropriate tool in our study.

Indeed, the propensity in analyzing a vast amount of data in order to find and employ the relationships that bind them, makes such technology highly effective for outcome prediction tasks; in the context of HNSCC, this technique was in fact extensively and, in many cases, successfully researched.

Data preprocessing

A crucial difference between classification and Time to Event analysis tasks is the output that will be provided by the model; namely, while the latter predicts whether the event has occurred at each timepoint – consequently providing information on when it happened – classification only reports the eventual occurrence. Consequently, classification provides one single output, and it does not supply any information regarding the timing. A noteworthy consideration, is that such lower informative value of the output allows to significantly simplify the problem, and is therefore the reason why this approach was tested for this specific application; indeed, while Time to Event analysis provides more information regarding the clinical output of the patient, since the classification model is requested to consider less aspects (i.e., it does not need to learn time implications), the mere prediction of whether the subject will experience the condition of interest (OS, LRC or DM) could result in being more accurate.

In conclusion, such approach was tested basing on the considerations regarding the advantages that would be introduced by a potentially less informative, yet more reliable, method.

This fundamental difference between these tasks implies that the dataset needs to be handled differently in order to provide the appropriate labels.

Indeed, in the implementation of classification task only the occurrences at 24 or 36 months were employed; they were simply extracted by the dataset, while the values regarding the time of occurrence were excluded.

A critical aspect regarding the employment of classification for this purpose is the handling of censoring cases. In effect, classification task does not anticipate this phenomenon; as a matter of fact, when employing a supervised learning algorithm in general, missing labels are not permitted at all. This obligation arises from the fact that it is not possible for the model to learn how to classify input data without their corresponding ground truth.

Consequently, as censoring cases are not allowed for this task, they had to be excluded from the dataset. It is worth noting that, especially in cases of many censoring patients, such discrimination involves the loss of significant

information (i.e., the patient is known to not have experienced the event until the moment of censoring); for this exact reason, Time to Event analysis is technically a more effective – and less wasteful – approach.
In conclusion, Time to Event analysis was the initial and definite goal of this research due to its higher informative power, but this alternative approach was considered as, despite its lower predictive potential, the previous considerations highlighted a possible advantage implied by its higher simplicity.

Networks architectures

When considering the architectures of the neural networks, they are the same as the ones in Time to Event analysis task (described in paragraph [3.4.4]). The only part that changes according to the selected task, is the last layer to be appended to the clinical network; in case of classification, the dense output layer is made of a single neuron and employes a sigmoid activation function.

Loss function and metrics

A noteworthy difference when compared to Time to Event analysis task, are the loss function and metrics employed to achieve classification. As its role is determinant to the development and therefore the performance of the model, the loss function was chosen basing on its effectiveness for this specific purpose; hence, Mean Absolute Error was selected to guide the training of the networks. This metric presents the average absolute difference between predicted and real values and is computed as follows:

$$MAE = \frac{1}{n}\sum |y - \hat{y}|$$

where n is the number of predictions, y is the value of the label and $\hat{y}$ is the predicted value.
It is worth noting that this is an uncommon metric to use as a loss function in a classification task (as in such case labels are binary values), however, it was chosen as resulted in improved performance when compared to other loss functions.

As the dataset held the possibility of some imbalance in the classes (i.e., 42% incidence of positive samples) a class weighting algorithm was tested in order to assess whether it would enhance the performance of the classification model.
This technique allows to assign different weights to the classes in order to improve the ability of the model of recognizing the minority class. Specifically, a higher weight is assigned to the minority class and multiplied to the loss function computed at each epoch; this way, the misclassifications of such class will have a higher impact on the learning process of the model.
This approach was tested and compared to the respective model, in order to evaluate a possible improvement of the performance and, subsequently, the impact of class imbalance in the dataset.
Furthermore, other metrics were employed to quantify and assess the performance of the model; for this purpose, AUC (paragraph [3.4.6]), Binary Accuracy and Binary Crossentropy (BCE) were chosen. While these functions were not used to guide the training of the networks, they were still imposed as outputs to be reported by the model as they provide further and useful insights regarding the predictive power of the networks. Specifically, Binary Accuracy is the ratio between number of correct predictions and total number of predictions, while Binary Crossentropy penalizes uncertainty and inaccuracy of the model according to the following equation:

$$BCE = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log\,(1 - \hat{y}_i)$$

The classification task provides the probability of the input belonging to classes 0 and 1, indicated by float numbers in the range [0,1]. Hence, a first analysis of the model's predictive effectiveness was conducted using its ROC.
Subsequently, in order to assess the accuracy of the model, such probabilities had to be turned into binary variables and to achieve this, the Youden Index was employed. This statistic provides the sum of sensitivity and specificity of the model for each decision threshold, and it is maximized to optimize both. Hence, the obtained optimal threshold, that was computed on the training dataset, was applied to the testing set to evaluate the classification performance of the model.

In order to achieve this in a comfortable and immediate way, Confusion Matrix were also established.
This procedure constitutes the postprocessing of the data that allows the definite assessment and comparison of classification task performance in respect to the other implemented tasks.

3.5.2 Regression for Head and Neck Squamous Cell Carcinoma

Regression is a technique that allows the analysis, and consequent quantification, of the relationship between independent and dependent variables. In medical applications, this method allows the investigation of the influence of clinical data on the outcome of interest; indeed, an analysis of the prognostic indicators and of how they affect the patient's condition is possible. As one would expect, such procedure can be applied to investigate HNSCC.
Hence, regression's proven potential for clinical outcome prediction was investigated in this research, with the aim of assessing its ability to quantify the relationship between input clinical variables and the duration of time before the event (OS, LRC or DM) is expected to occur.

Data preprocessing

In the employment of this approach, some considerations are similar to the ones involved in the execution of classification, as opposed to Time to Event analysis. Namely, like classification, also regression provides one single output; in this case, however, such value indicates the time supposed to pass before the occurrence of the event (instead of *whether* the event will occur). Such crucial property of this approach highlights the intrinsically higher informative value of the output supplied by the regression model.
This significant difference, however, arises a similarity in the required handling of the dataset, in the sense that in order to implement regression each patient will be associated with one single label. Again, this has a twofold implication: firstly,

the clinical data will be processed with the aim of excluding every aspect except for the time at which the first event occurred. Secondly, similarly to the classification case and for the same reasons, missing ground truths are not allowed; subsequently, all the censoring patients were to be excluded from the dataset.

This procedure caused the same – negative – implications of data loss as the classification approach.

Despite the similarities, however, a significant difference stands out: comparing to classification, regression involves the advantage of providing some information regarding the elapsing of time in the development of the event of interest; this factor is a further progress comparing to the mere prediction of whether such event will occur. However, despite the advantages when compared to classification, regression is still far from supplying the comprehensive information provided by Time to Event analysis.

A critical consequence of the outlined aspects is that this approach does not consider the possibility of not developing the event; an effective model for this kind of application is, therefore, expected to provide a higher amount of time estimated before the occurrence for lower risk groups (i.e., people that are actually never going to experience the event). Hence, the most effective approach could involve a further postprocessing of the estimated times in order to find a threshold of time to discriminate people who are not considered at risk of developing the event.

Nonetheless, it is worth highlighting the tradeoff between informative power and complexity of the task: the higher predictive power of regression task when compared to classification, comes with the cost of a harder prediction requested to the model and, consequently a higher risk of unsatisfying performance.

In order to implement regression, a further processing of the clinical data was necessary; specifically, the propagation in time that was executed on clinical data with the aim of implementing Time to Event analysis was not suitable for this purpose. Hence, the labels identifying the time at which a certain event would occur were extracted from the original clinical follow-up data and left unvaried;

for this task, in fact, the standardized detection of the occurrences at specific timepoints (i.e., at multiples of 6 months) would be disadvantageous as it would lead to biased inferences.
After executing this procedure, that allowed the gathering of the inputs and labels required for the implementation of the regression, it is possible to proceed with the definition and set up of the model.

Networks architectures

When executing a regression task, architectures are the same as those for Time to Event analysis (depicted in 3.4.4), with the only exception being the output layer of the clinical network. In order to execute a regression, the final layer is represented by a dense, single neuron layer without an activation function.

Loss function and metrics

For this task, the most appropriate loss function and metrics were chosen. Like in classification task, Mean Absolute Error was employed as loss function.
Other metrics were implemented to observe the effectiveness of the model in the prediction of event by 2 years: precision, recall and f1 score. Such parameters are actually classification metrics and were employed after a binarization of the prediction. Specifically, the model was asked whether the patient would experience the event by 2 years, and therefore the estimated times were associated to a 1 when lower than 2 years, and 0 vice versa. Subsequently, the metrics could be computed. Precision indicates the portion of True Positive predictions among all the positive predictions, while Recall measures the ratio between True Positive predictions and all the positive labels. Lastly, F1 score combines such parameters through their harmonic mean[3].

Data postprocessing

[3] F1 score is computed through the following formula:

$$F1\ Score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

The prediction model output the estimated time of occurrence of the event that was to be compared with the real one in order to assess its effectiveness. For this purpose, some statistics describing such estimations were computed; specifically, the percentage of correct, delayed and advanced predictions were calculated:

- predictions correct within ±1 year from the real occurrence;
- delay or advance in a ±[1,2] years range;
- delay or advance in a ±[2,3] years range;
- delay or advance of more than ±3 years;

This approach provided a quick and effective report of the model's tendency in providing correct, delayed or advanced predictions.

Furthermore, the estimated times were converted into binarized variables that established, at each year, whether the event was estimated to be occurred; such variables were then compared to the corresponding ground truths and used to draw Confusion Matrix. This allowed the establishment, for each year, of the binary accuracy of the regression model. This analysis could provide interesting insights regarding the robustness of the model in the estimation of events that occur at different timepoints.

3.6. Experimental setup

Different approaches were adopted to train the neural networks: the models were trained with different combinations of clinical, segmentation-derived and imaging data, and to provide different kinds of outputs.

The performance of the models was evaluated when predicting Event Free Survival (EFS), Overall Survival (OS), Distant Metastases (DM) and Loco-regional Failure (LRF).

As usual procedure, in order to develop the DL model, the dataset is to be divided into training, validation and testing set; specifically, a portion (i.e., one quarter) of the dataset provides the testing set, and the remaining data constitute validation and training sets with a 0.25 ratio.

Furthermore, in order to validate such model 4-fold cross-validation was applied; this technique consists in the division of the dataset into 4 subsets (i.e., folds) that are combined to develop four different models and calculate their average performance. Specifically, the model is trained and tested 4 times, with each time the testing set being constituted by a different portion of the original dataset; a visual representation is provided in Figure 21 [86]. The resulting performances are then averaged in order to assess the model's ability to generalize to new, unseen data.

Figure 21: 4-fold cross-validation; each portion of the dataset is used as testing set once, while the rest constitutes the training set.

With the aim of enhancing the performance of the model, hyperparameter tuning was carried out through RandomSearch optimization technique[4]. This approach randomly selects a set of hyperparameters from ranges specified by the user to train the model. This procedure is repeated multiple times with different combinations of values; each time, the resulting performance is assessed on the testing set and stored so that the hyperparameters that led to the best predictions are then selected as optimal for the specific model.

Such hyperparameters highly affect the performance of the system as they specify the model and how the training will be executed. Therefore, their influence on the predictive power of the model was analyzed in order to choose the most appropriate values.

[4] The parameters optimized through SGD are: learning rate, momentum, Nesterov's momentum, kernel regularizer, dropout rate, initial number of filters and spatial reduction for the CNN, clinical depths of the Clinical NN.

Stochastic Gradient Descent (SGD) was selected as optimization algorithm; it handles the minimization of the loss function by adjusting the network parameters (i.e., weights and biases) basing on the gradient of the loss function computed with respect to a random mini-batch extracted from the dataset.

Lastly, the obtained models were compared basing on the evaluation metrics chosen for final assessments.

Since, as previously described, the different tasks were evaluated through different metrics, they cannot be directly compared; therefore, the comparison of the different tasks requires some careful considerations to gain a deep understanding of the implications of each evaluation metric.

CHAPTER 4 RESULTS AND DISCUSSION

4.1 Results

In the following, the results achieved through Time to Event analysis, classification and regression will be reported and described. As each of the implemented tasks was achieved through a different methodology and evaluated with different metrics, they will be separately presented.

4.1.1 Results of Time to Event Analysis

The parameters chosen to execute Time to Event analysis resulting from the hyperparameters tuning are reported in Table 4.

Parameter name	**Optimized value**
Num epochs	50
Batch size	32
Learning rate	0.01
Momentum	0.90
Kernel regularizer	1e-4
Dropout rate	0.1
Initial filters	32
Clinical depths	[64,128]

Table 4: Model parameters resulting from the optimization.

The experiments proved that the selection of such parameters brought a 10% improvement of the model's performance.

Once such optimized parameters were determined, they were employed to develop the definite Time to Event analysis model, which was successively evaluated through different metrics and graphs that were established as depicted in paragraph 3.4.7.

Basing on the HCI, the best configuration and models were chosen and subsequently compared to each other and to those related to other tasks. Firstly, the different feature sets were compared through experiments that only employed clinical data as inputs; the results are shown in Table 5, in which for each experiment the HCI's mean and standard deviation indicates the performance obtained through each features set.

Feature set	**HCI** (mean ± std)
v2	0,512±0,019
v3	0.489±0.0156
v4	0.512±0.023
v5	0,521±0.085
v6	0.569±0.041

Table 5: results of clinical experiments with different feature sets.

While the performance achieved through the first 4 feature sets do not differ significantly, the data reported in Table 5 prove the higher informative power of feature sets v5 and v6 that were, therefore, chosen to perform the experiments including the imaging data.

The experiments conducted through the integration of clinical and imaging data are shown in Table 6: different inputs and modalities were compared.

A visual inspection of Table 6 allows a proper understanding of the performances of the model according to the employed features set and modality. The union of feature set v6 with merged CT and PET volumes achieves the best performance.

In general, v6 seems to provide, in any configuration, better results when compared to v5.

Feature set	**HCI** (mean ± std)
v5	0,520±0.037
v5 + CT	0,523±0,016
v5 + PET	0,513±0,028
v5 + merged	0,508±0,020
v6	0,563±0,033
v6 + CT	0,582±0,037
v6 + PET	0,573±0,028
v6 + merged	0,596±0,030

Table 6: mean and standard deviation of HCIs for different features sets and modalities.

While the HCI provided valuable information regarding the predictive capacity of the model, a further postprocessing of the predictions and a computation of different metrics, as described in paragraph 3.4.7, allow an even deeper understanding of the achievements of the models.

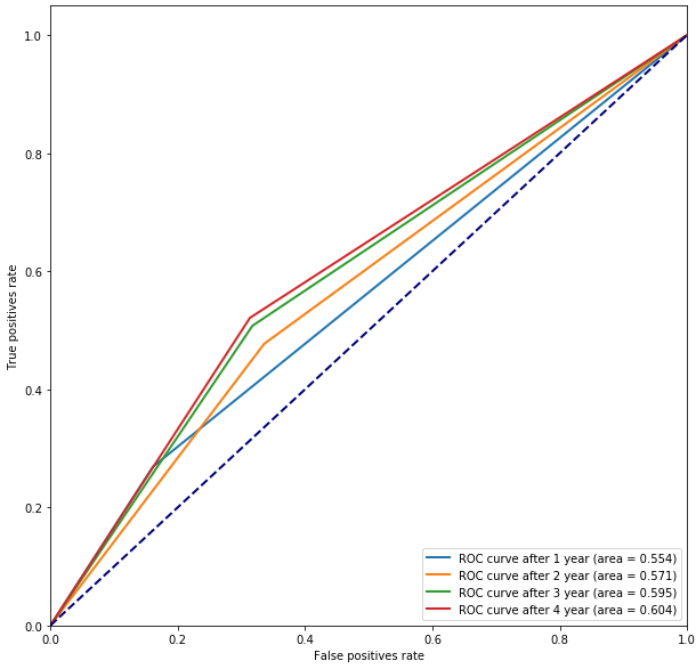

Figure 22: Superimposed ROCs for the detection of events at each year.

For the ultimate models a 0.8 threshold was selected to transform the float values indicating the probability of belonging to each class, to the binary labels; such high value reflects the tendency of the model to provide very high probability of survival. Once this binarization of the labels was executed, ROCs for each year

were drawn, as shown in Figure 22. For each curve, the corresponding AUC is reported in figure.

The AUCs were then reported on a new graph to allow a quick and informative representation of how well the model predicts an event at each year. A visual investigation of Figure 23 allows to assess that the model predicts with a slightly higher accuracy events occurred between the second and third year; however, the accuracy does not deviate significantly through the years, and the model can therefore be considered robust to this factor. Moreover, as shown in Figure 24, a cumulative ROC was drawn; conversely to the previous graph, in this case the positive labels were propagated before their comparison with the ground truths. This approach will, subsequently, provide higher AUCs as years go by and the difference between subsequent AUCs reflects the number of advanced predictions.

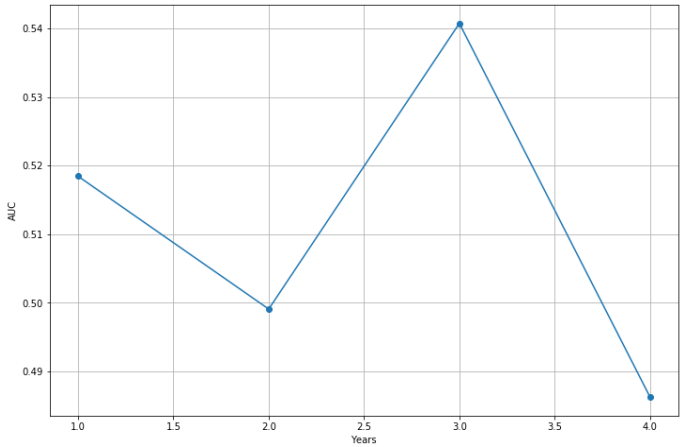

Figure 23: Time-dependent AUC value.

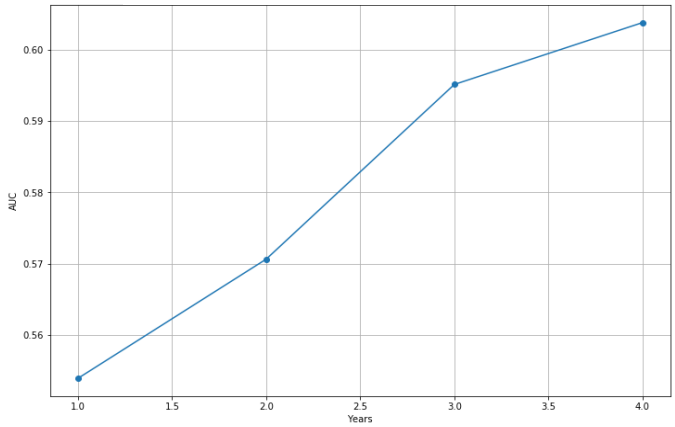

Figure 24: Cumulative AUC; positive predictions are propagated to consider correct the early ones.

4.1.2 Results of Classification

As extensively described in 3.5.1, in order to implement the classification task, patients who censored before 3rd year were excluded from the dataset; this allowed the model to train on correct statistics of the occurrences.

A class weighting algorithm was tested in order to assess whether it would enhance the performance of the classification model. As it introduced an almost 10% improvement in balanced accuracy, it was chosen to implement this technique in the ultimate model.

The best performance resulted from the application of the model on feature set v7 for the prediction of patients who would develop an occurrence before the 3rd year from the diagnosis. For the evaluation of the classification model, the balanced accuracy was employed together with the AUC.

As reported in Table 7, it emerges that, surprisingly, the imaging data do not appreciably achieve higher performances in the model. In order to gain a deeper understanding of the performance of the model and of the distribution of the accuracy among classes 0 and 1, Confusion Matrices were drawn and reported in Figure 25 and Figure 26.

Metric	Balanced Accuracy	AUC
v7	62.09%	0.663
v7 + PET	62.00%	0.683
v7 + CT	60.89%	0.648
v7 + merged	62.89%	0.677

Table 7: results of the best classification model for different inputs and configurations.

As emerges from a comparison of the ROCs in Figure 28 and Figure 27, the overall predictive power of the model does not deviate notably depending on the selected inputs.

In conclusion, in accordance with the results shown in Table 7, the performances of these two configurations are similar, even in the values of sensitivity and specificity.

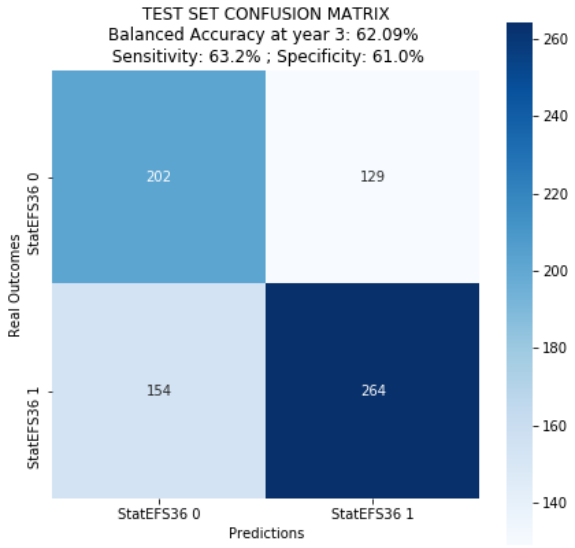

Figure 25: Confusion Matrix of the classification model trained only on clinical data.

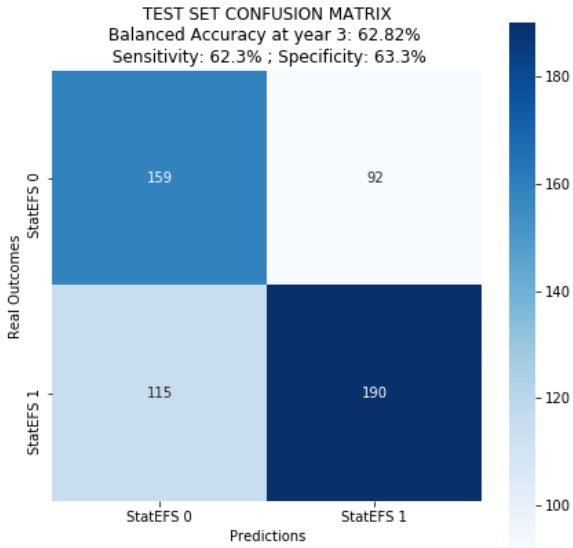

Figure 26: Confusion Matrix of the classification model trained on merged clinical and imaging data.

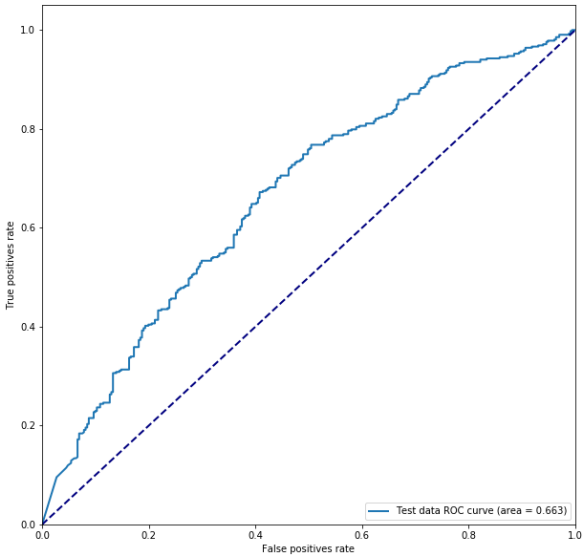

Figure 27: ROC of the classification model trained only on clinical data.

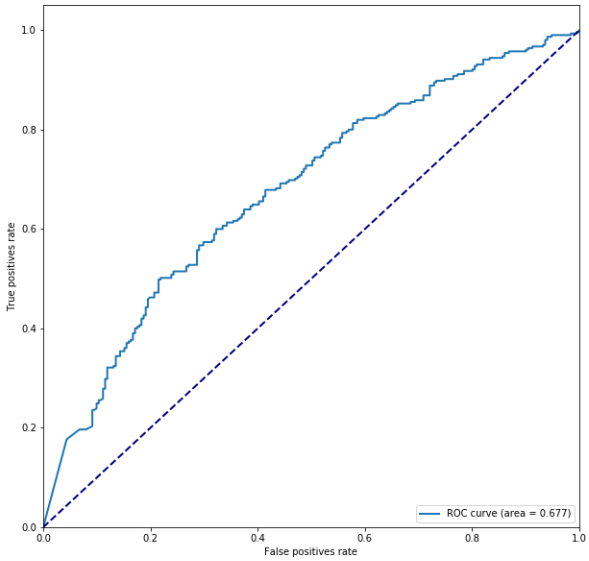

Figure 28: ROC of the classification model trained on merged clinical and imaging data.

4.1.3 Results of Regression

Similarly to classification task, the most successful model was the one trained on feature set v7. The balanced accuracies resulted in the detection of an event within the 3rd year from the diagnosis, are reported in Table 8. Like in the case of classification, the balanced accuracy is consistently around 60% across all

configurations, and imaging data does not seem to improve the performance of the model.

Feature set	Balanced Accuracy
v7	60.51%
v7 + PET	57.17%
v7 + CT	59.07%
v7 + merged	57.46%

Table 8: results of the best regression model for different inputs and configurations.

In order to provide a quantitative idea of how the predictions are distributed, a scatter plot is shown in Figure 29. While such graph does not allow an assessment of the accuracy of the model as the correspondence between predictions and labels is not clear, it can still provide useful information regarding the range of the predictions. Indeed, a visual inspection of the plot allows to assess that the model appears to effectively space throughout the whole range of [0;75] months; specifically, the majority of the patients experienced an event within 60 months, and therefore a model able to locate an occurrence in that range can be considered satisfactory.

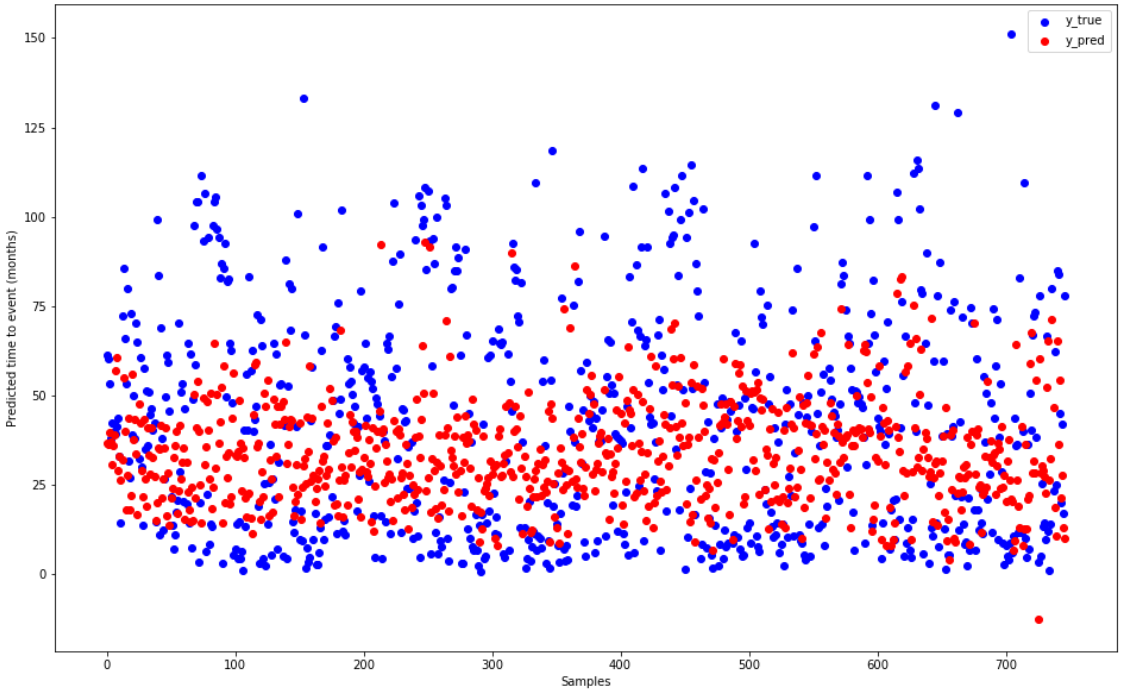

Figure 29: Scatter plot of time labels and predictions provided by the regression model.

Such consideration, however, does not account for an estimate of the accuracy of the predictions, which must be conducted through a different approach. For this purpose, the percentage of correct, late and advanced predictions were computed in order to complete this evaluation with a quantitative understanding of the accuracy of the model.

The obtained percentages, that refer to the timing of the predictions, are listed in the following index:

- correct within ±1 year from the real occurrence: 28.38%;
- late in a +[1,2] years range: 15.80%;
- advanced in a -[1,2] years range: 11.11%;
- late in a +[2,3] years range: 8.84%;
- advanced in a -[2,3] years range: 11.24%;
- more than 3 years late: 5.22%;
- more than 3 years in advance: 19.41%;

Such percentages seem in accordance with the not completely satisfactory Balanced Accuracy of the model. Indeed, while the time of survival of a higher percentage of subject was correctly predicted with a margin of a ±1 year, the number of mispredictions is still too high. On the other hand, these values also highlight the slight overall tendency of the model to underestimate the time of survival of the patients: such behavior, when not too prominent, could benefit the application of a clinical outcome prediction. It would in fact be desirable the obtainment of a model that slightly underestimates the times of survival, rather than overestimating them.

In order to allow a deeper analysis of the results, a time threshold was selected in order to discriminate the patients who are predicted to develop an event before such threshold, and the patient who did not and can therefore possibly never experience it (i.e., low-risk patients). This proves convenient when considering that, as mentioned, the majority of HNSCC patients is estimated to either experience an event before 2 or 3 years, or to fully recover from the cancer. On this basis, a 3-year threshold was employed.

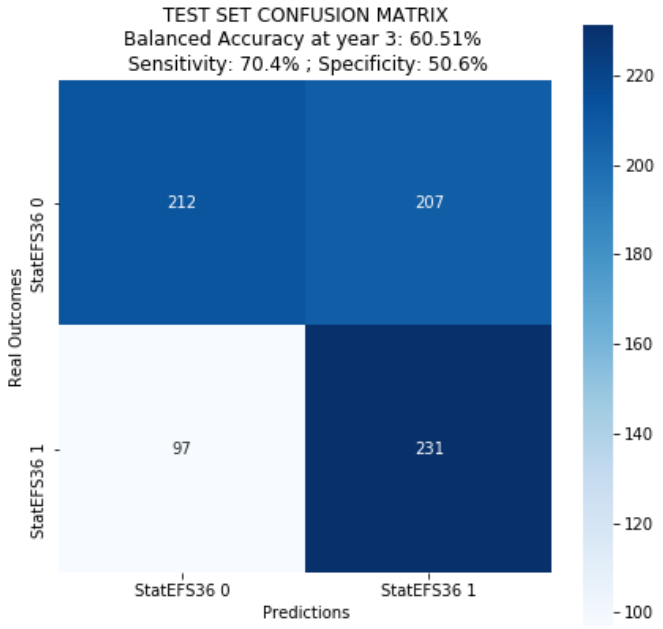

Figure 30: Confusion Matrix of the regression model for the prediction of events by the 3rd year from the diagnosis.

The resulting Confusion Matrix, reported in Figure 30, highlights the very high number of False Positive samples. This is in accordance with the reported sensitivity and specificity values: while the former results more satisfactory, the latter proves extremely low.

4.2 Discussion of the Results

4.2.1 Comparison of the methods

The choice of implementing further approaches originated from several factors; in general, as time to event analysis is a more complex task than classification and regression, an assessment of whether the former would lead to lower performances seemed reasonable.

Specifically, as our main goal is the prediction of certain events rather than the detailed understanding of the underlying phenomena (i.e., survival function), classification and regression models had the potential to prove more effective for a single, time-fixed prediction.

It is also crucial to account for the censoring: it is universally recognized that if the dataset is highly affected by this phenomenon, Time to Event analysis cannot be effectively achieved; this is, in fact, our case as 52% of the considered patients did not participate in the follow-ups for the complete length of 4 years.
In general, as classification and regression require less labeled data than Time to Event analysis, they could provide better performances when executed on the same dataset. Furthermore, the higher simplicity of the implemented alternative approaches also involves a stronger robustness to overfitting and a lower computational burden.
Lastly, classification and regression tasks also provide significant insights about the importance of the different features; while Time to Event analysis has the same capacity, it deals with more complex patterns and relationships among the data and can, therefore, be much harder to interpret.
The depicted methods provide different information concerning the progression of the patients' disease. Hence, it is crucial to consider what the most relevant insights for the specific application are, together with the performance of each task. Specifically, while classification does not intrinsically provide information about the timing of the events, our data processing allowed the integration of a useful insight about the possible occurrence before a desired time threshold. As the prediction of occurrences by the second or third year after the diagnosis is considered crucial in this field of application, this approach could, despite the lower informative power of the prediction technique, still provide the desired information. Therefore, if this approach were to guarantee a higher reliability, it could still be preferrable compared to Time-to-Event analysis and regression.
On the other hand, and as extensively discussed in paragraph 3.5.2, classification as well as regression cannot handle censoring cases. Therefore, censored cases were excluded from the dataset to perform these tasks, causing a lower number of patients available for the training of the model.
A further, and critical, implication is that, as regression does not account for the possibility of not experiencing the event, only the times of occurrences were predicted. Considering how the input data are defined, patients who never

experience an event still provide the time at which they censored, which refers to the timepoint at which they stopped taking part in the clinical appointments. This implies that, even patients who never developed an occurrence stopped showing up at the follow-up at a certain point, and the corresponding time is registered as time of the event to be predicted. This means that this regression model is trying to predict, in cases of patients who never experienced any occurrence, the time at which they quit the follow-up. This aspect completely biases the predictions, as such event has no relation to the inputs of the model. Therefore, a further processing of the data allowed to account for this problem; as mentioned, in fact, patients who had censored before the third year of follow-up were removed from the dataset. On the other hand, in order to handle the patients who censored after the third year, different considerations were made. As the third year was established as a threshold to determine high-risk subjects, times of events higher than 36 months were just considered as survival times longer than 3 years; for this specific use, even for patients who did not actually develop the event after more than 3 years but just stopped showing up at the follow-up, the predictions would not be biased.

Interpretation of Time to Event analysis results

The first step in the investigation of Time to Event analysis' results, was the comparison of the performances of the clinical networks (i.e., models trained with clinical data alone); the different feature sets were compared to determine their prognostic power.

As described in paragraph 3.3.2, features sets v5 and v6 are of particular interest to the purposes of this research, as they contained the most informative features. While v5 integrates HPV status, v6 contains radiomic data, and therefore the comparison of the performances obtained through these two feature sets allows an evaluation of the predictive power of such variables. Accordingly, as reported in Table 5, this procedure allowed the assessment of the superiority of v5 and v6 comparing to v2, v3 and v4.

The further comparison of these two feature sets allows to assess whether such prediction model can function properly when only relying on pretreatment and raw imaging data, as opposed to the radiomic features (contained in v6); this analysis resulted in feature set v6 achieving the best performance. This suggests that, surprisingly, the HPV status contributes significantly less to the Time to Event analysis compared to radiomic features; such information seems discordant with the acknowledged role and prognostic value of such variable.

Furthermore, another unexpected finding is that imaging data does not seem to enhance appreciably the performance of the model or, in case of feature set v5, they even worsen it.

For each year, the performances of the model were also represented through ROCs to assess whether the model struggles to predict occurrences at specific timepoints. As reported in Figure 23, while the achieved AUCs do not vary significantly across the years, they are very close to 0.5; this consideration, together with the excessively high decision threshold (i.e., 0.8) manifest the inability of the model to effectively detect class 1.

A possible explanation and an analysis of these crucial aspects will be provided in paragraphs 4.2.2 and 4.2.3.

The outputs of the model, corresponding to predictions at each year from the start of the follow-up, were considered and compared to the corresponding ground truths to establish Confusion Matrices. This allowed to carefully analyze the prediction within each timepoint, and the corresponding balanced accuracy, sensitivity and specificity were computed.

A visual investigation of Figure 31, Figure 32, Figure 33 and Figure 34 allows an immediate understanding of the model's ability to predict events occurred at each yearly timepoint. Specifically, it is possible to affirm that the model proved robust to the passing of the time, since its accuracy does not vary appreciably throughout the years. However, the model also proved ineffective, as its balanced accuracy is very close to 50% (i.e., random prediction). The Confusion Matrices show that such value derives from the extremely low number of detected positives; indeed, this analysis highlights the difficulty of the model of predicting class 1. This also

explains the optimized decision threshold, as a very high threshold allows an increase of detected positives.

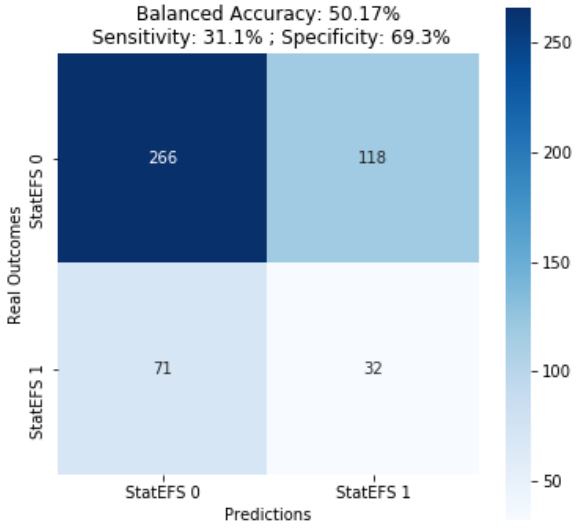

Figure 34: Confusion Matrix of subjects' classification at year 1.

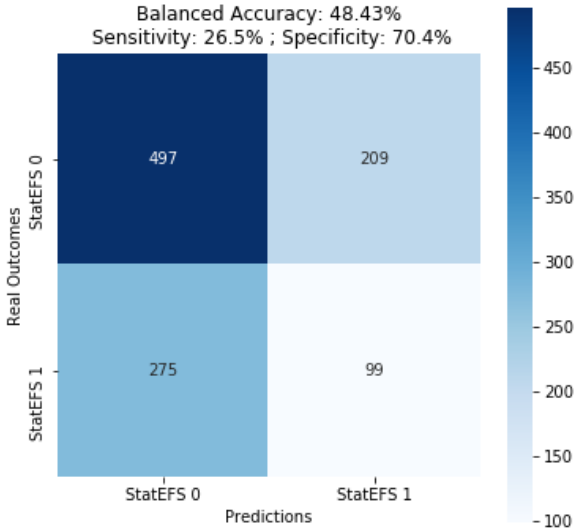

Figure 33: Confusion Matrix of subjects' classification at year 2.

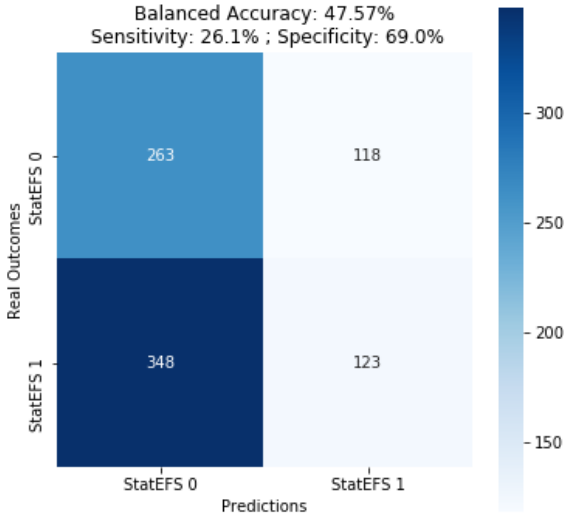

Figure 31: Confusion Matrix of subjects' classification at year 3.

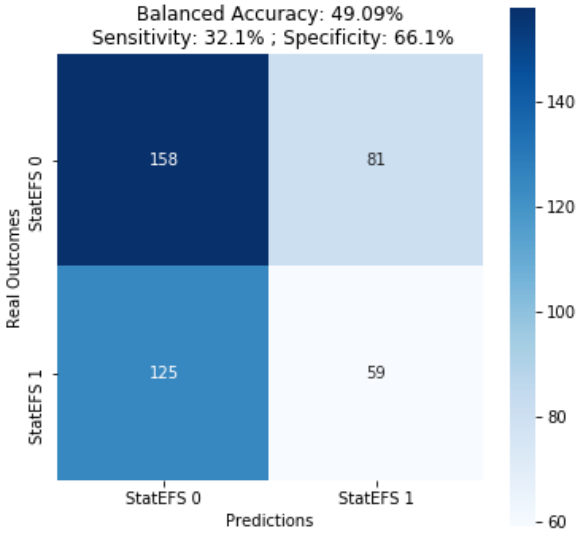

Figure 32: Confusion Matrix of subjects' classification at year 4.

For this last evaluation, balanced accuracy was employed rather than HCI for two reasons; firstly, it allows a quick assessment of the proportion of the correct predictions – while the HCI referred to its capacity of accurately assigning a time of event to each patient. Furthermore, balanced accuracy can be also employed to

evaluate classification and regression and can, therefore, be used to effectively compare these three methods.

Comparison of the Approaches

As mentioned, the postprocessing of the outputs, that allowed the predictions of all the tasks to be assessed through Balanced Accuracy, permitted an immediate and even comparison of the approaches.

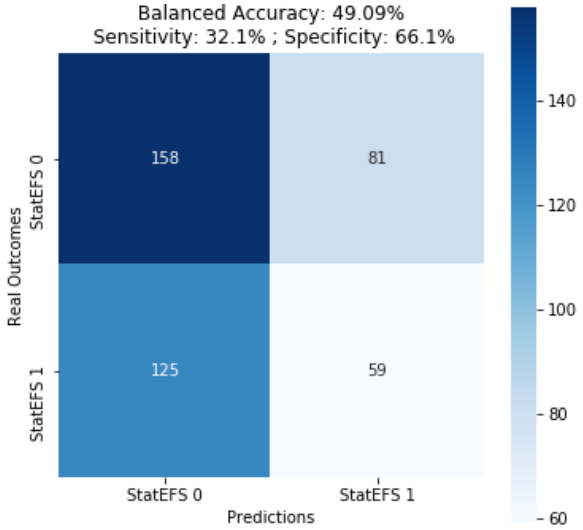

Figure 36: Confusion Matrix of Time to Event analysis for the classification of events by year 3.

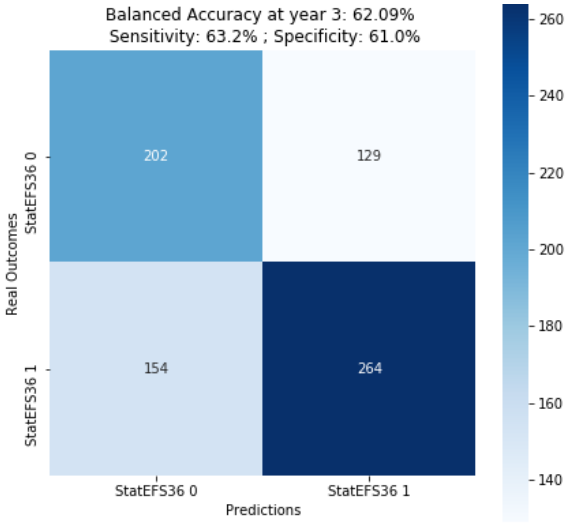

Figure 35: Confusion Matrix of classification of events by year 3.

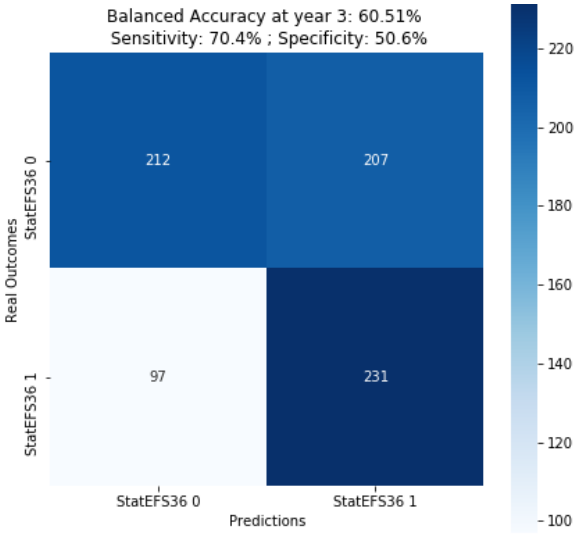

Figure 37: Confusion Matrix of regression for classification of events by year 3.

A first comparison of the Confusion Matrices, together with the balanced accuracies of the approaches, allows to quickly assess the lower performance of the Time to Event analysis. Specifically, the Confusion Matrix reports an extremely low number of predicted positives; this implies that the model struggles to predict the occurrence of the event and tends to estimate a longer survival period.

It is crucial to keep in mind that this model is to be applied as a clinical screening for patients who receive a HNSCC diagnosis; therefore, it is fundamental to discourage as much as possible the false negatives, even at the cost of a higher number of false positives. The latter would in fact be later disproved through further clinical inspection, thus involving no harm to the patients' health.

For this reason, after using the balanced accuracy to compare the different approaches, a further step was taken through the comparison of the respective sensitivities. This index, in fact, gives an immediate understanding of the capacity of the model of detecting the positive class.

In Table 9 balanced accuracy, sensitivity and specificity achieved with each task are reported. From a comparison of the balanced accuracies, an improvement of around 10% emerges with the introduction of classification and regression tasks, comparing to Time to Event analysis. Furthermore, an encouraging increase in sensitivity is also achieved through the alternative approaches: classification and regression's sensitivity are around twice the respective value provided by Time to Event analysis. This last achievement is noteworthy since, as previously mentioned, sensitivity must be prioritized in a clinical outcome prediction.

Task	Balanced Accuracy	Sensitivity	Specificity
Time to Event	49.09%	32.1%	66.1%
Classification	62.09%	63.2%	61.0%
Regression	60.51%	70.4%	50.6%

Table 9: Performance metrics of each task.

Lastly, classification implies a slightly lower specificity than Time to Event analysis; for this task however, sensitivity and specificity are very balanced.

On the other hand, while providing a more satisfactory sensitivity, regression results in a very low specificity; such factors are therefore significantly unbalanced.

4.2.2 Factors contributing to the suboptimal performances

The collected dataset is likely to have played a crucial role in the resulting performance of the model, and various aspects related to this matter can be analyzed.

First of all, it is important to consider that this study proposed to employ a Deep Learning model to assess whether patterns among the data were present and could allow the prediction of the event of interest. Thus, our results could be caused by the fact that such correlation between some of the input features and the individual outcome does not exist.

On the other hand, it is possible that a correlation between the considered clinical data and the probability of developing an occurrence does exist; in this case, the low accuracy of our model could be determined by an unsuitable dataset. Indeed, while a major effort was made in order to collect the highest and the most balanced amount of data, it might not have been enough. The number of patients could still be insufficient or, more likely, the population might not be appropriately representative. It is even possible that too much data was input in the model: such a high number of different features and values could have introduced some confusion due to their complexity.

A significant property of a dataset is class imbalance; it occurs when the distribution of classes is not equal throughout the population and one class is, consequently, over-represented when compared to the other one. This phenomenon can severely impair the predictive accuracy of a Machine Learning model. Specifically, the model might become biased and learn better to classify the majority class but poorly the other one. Moreover, if the class imbalance among the training set significantly differs from labels' representation in real

data, the generalization ability of the network could also be significantly impaired.

Our dataset, presenting 39% of occurrences, implies the risk of class imbalance. It is also crucial to highlight that even though 39% of the patients experienced an event, each of them did at a different time point: when performing Time to Event analysis, the labels of any patient who suffered from the event are therefore likely to be very imbalanced when compared to the negative subjects.

This important factor probably contributes to the extremely low sensitivity achieved by the model when performing Time to Event analysis (i.e., 32.1%). The sharp difference with the specificity (i.e., 66.1%) in fact, reflects how much better the model is in predicting the majority class, namely patients who never experience the occurrence.

In order to address this problem it is possible, for a classification task, to implement class weighting. Indeed, this method introduced a 10% improvement in the sensitivity of the classification model.

Another significant aspect to take into account is the origin of the data employed in this research; aiming at a large and appropriate dataset, data from different centers from all around the world were collected. While this allowed to increase the number of subjects to include in the study, it might have originated some heterogeneity throughout the data. The clinical variables could, in fact, have been acquired or processed in slightly different ways (e.g., different instrumentation presenting different properties and accuracies, different set-ups etc.,). When considering some of the clinical features some inter-observer variability must be accounted for. It is therefore possible that these multi-center data could have introduced some standardization problems with the data; moreover, this same assumption could be made for the labels. Indeed, the labels for Time to Event analysis, namely the times at which each occurrence did or did not occur, were not acquired at specific and homogeneous time-points; this might have introduced some difference in months in the detection of the events. This aspect, together with the timing of the diagnosis of the HNSCC itself, could suffer the influence of the prevention culture and clinical rules of the specific country:

depending on these factors, in fact, some countries could encourage earlier predictions and finer schedules for follow-up appointments. This would highly impact the origin and event times provided as labels for the survival analysis. In conclusion, the multi-center dataset could have involved a confusing heterogeneity or even biases we would not be aware of.

On the other hand, also the imaging data and their influence is noteworthy. Unexpectedly, inputting PET and CT volumes to the model did not seem to appreciably increase the performances of the models regardless of its configuration or task. As this kind of data is intrinsically highly informative, it is possible that such volumes did not add knowledge to the network due to their enormous complexity and sophistication. It is possible, therefore, that PET and CT data require a specific processing or a segmentation to provide their informative contribution to the predictions; this could imply that, while this research proposed the use of raw imaging data, this approach might not be appropriate to achieve our goal of clinical outcome prediction.

A further consideration can be made: the problem with the imaging data could also arise from the multi-modality of the inputs, rather than from their complex nature. Specifically, it is possible that imaging data were not able to contribute to the task because they needed, for example, a longer training when compared to the – much simpler – clinical data. Such different inputs, in fact, might require some training specifically designed to handle the volumes (e.g., our configurations forced the training of both clinical and imaging network to the same batch size, number of epochs, etc.).

4.2.3 Approaches to overcoming the limitations

In order to gain a deeper understanding of the unsatisfactory performance of the model, the distribution of the labels and wrong predictions was analyzed. The goal was to assess whether the model would be ineffective for a specific subclass of patients. Specifically, considering the crucial role of HPV status in HNSCC prognosis (extensively depicted in Chapter 1 and 3), the distribution of the

predictions in relationship to this feature was analyzed and reported in Figure 38.

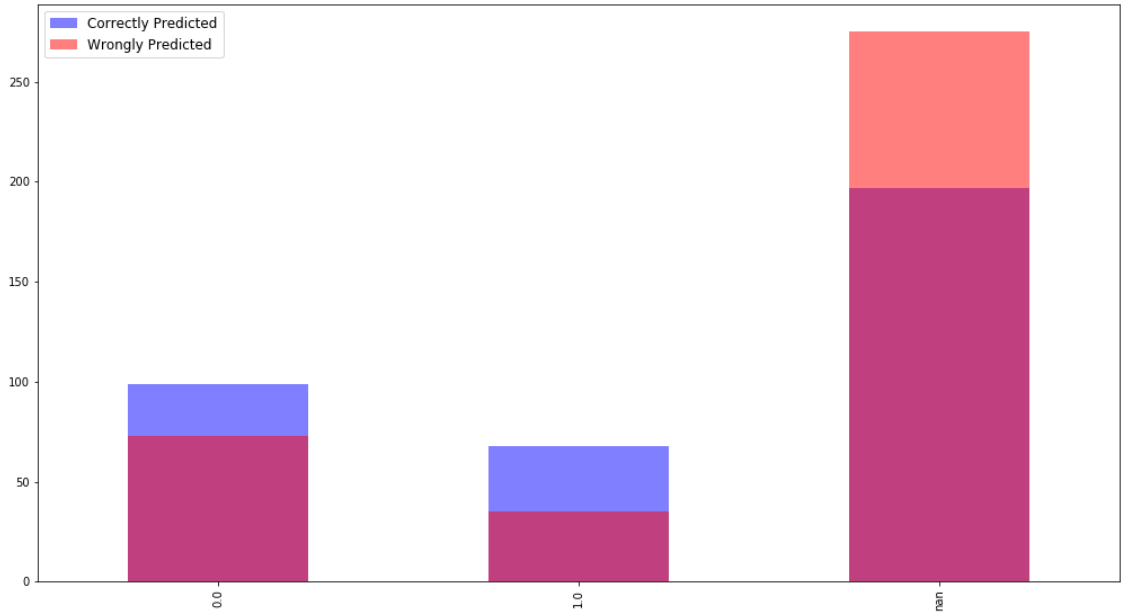

Figure 38: Distribution of correct and wrong predictions with reference to the HPV status.

Interestingly, from this graph emerges the difficulty with which the model develops a correct prediction for all the patient lacking their HPV status; such observation is in complete accordance with the clinical knowledge, that deems the HPV value a crucial prognostic factor.

As only the 35% of the patient in our retrospective cohort is provided with their HPV status, it was originally decided not to exclude all the subjects who did not present it in order to preserve a big amount of valuable data. The unsatisfactory performances, however, suggest that inputting such a high number of missing HPV data introduces confusion to the model. Additionally, the remaining – less informative – features do not compensate for such a loss of prognostic data.

As shown in Figure 39 and Figure 40, this model proved a higher performance when compared to all the previous approaches. Specifically, both Balanced Accuracy and AUC resulted in a 10% and 5% increase respectively. Moreover, sensitivity, which we aimed at increasing as much as possible, significantly enhanced without impairing the specificity, that remained unchanged in respect to the previous classification case.

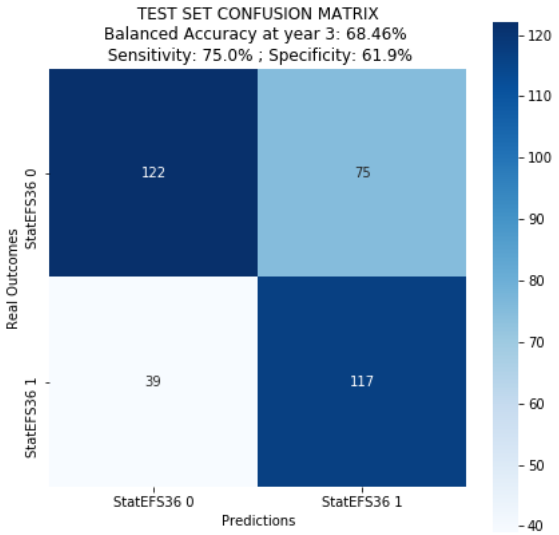

Figure 39: Confusion Matrix of the new classification model trained with patents provided with HPV status.

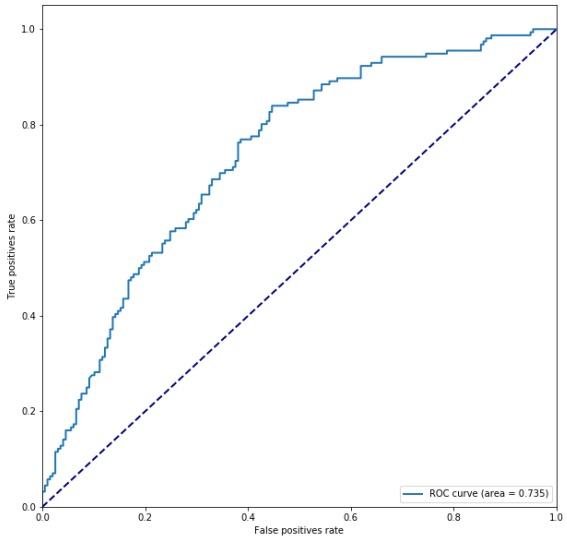

Figure 40: ROC of the new classfication model trained with patents provided with HPV status.

Task	Balanced Accuracy	Sensitivity	Specificity
Time to Event	49.09%	32.1%	66.1%
Classification	62.09%	63.2%	61.0%
Regression	60.51%	70.4%	50.6%
New classification	68.46%	75.0%	61.9%

Table 10: Performances of all the implemented methods. "New classification" refers to the experiments inputted with the new dataset resulted from the exclusion of all patients lacking HPV status.

These results reflect the strong predictive power of HPV status in the context of HNSCC clinical outcome prediction. It is in fact noteworthy how a better performing model was obtained using only 363 patients as inputs (as opposed to the 927 subjects of the whole dataset); this implies that the other clinical features, while still contributing to the predictions, do not present such a strong correlation with the risk of death, tumor relapse, or distant metastasis development.

4.3 Study Limitations and Future Developments

A significant limitation in our study is provided by the class imbalance in the dataset, especially when performing Time to Event analysis. Indeed, while this

problem was addressed for the execution of the classification task by using class weighting, such technique is not suitable to be directly applied to Time to Event analysis. Specifically, class weighting is not very commonly used for this application and would likely be quite complex to implement but, with the required modification and configuration it could be possible and provides an interesting and convenient addition to a future improvement of this approach. Specifically, for this specific task it is crucial to consider that class imbalance does not refer to the number of patients that do and do not experience the event; indeed, a class in Time to Event analysis data is composed by all the subject that experience the occurrence at a specific timepoint. Therefore, in order to compensate for class imbalance, it is necessary to gather a dataset with a comparable number of patients who experience the event at each timepoint and patients who do not experience it at all. A further aspect that could be taken into account is that it might be possible, rather than choosing a certain number of patients who suffer from the event each year, developing an algorithm in order to optimize the length of the time intervals selected for the Time to Event analysis with the aim of excluding the lowest possible number of subjects when balancing the dataset.

A different aspect that was brought up when discussing the limitations of our Time to Event analysis, is the possible heterogeneity of the multi-center data. As mentioned, the collection of data from multiple centers around the world allowed a higher number of subjects in the retrospective cohort, which is fundamental to achieve a good performance when employing a Deel Learning model. However, this approach might have introduced some unknown bias in the model as the data from different clinics might present some slight differences. In order to overcome such limitation, a specific standardization could be designed for the collection of data with the purpose of Time to Event analysis; for instance, some guidelines could describe a certain way of acquiring and processing the data, besides imposing a specific periodic schedule for clinical visits and encourage some measures that could result in a diagnosis at a similar step of the development of the HNSCC.

A crucial area that presents an encouraging margin of improvement is the handling of imaging data; our approach proposed to explore the possibility of inputting the model with raw, unprocessed imaging data, but this method resulted inefficient for this specific application. Therefore, a future work in this field of application could focus on the integration of some automatic segmentation and processing of the PET and CT volumes; this approach might significantly increase the contribution of imaging data to the clinical outcome prediction. Moreover, besides the image processing, some attention could be paid to the configuration for the combination of clinical and imaging network; the findings of our work, in fact, could suggest that clinical and imaging data are too different to be treated in the same way during the training of the model. A further improvement could therefore be some configuration that allows a slower training rate for the clinical data, in order to give time to the model to focus more deeply on the imaging data. It could also prove convenient to employ transfer learning, namely a technique that would allow the previous and separate training of an imaging data network, and the subsequent integration of such knowledge on the model that simultaneously handles clinical and imaging data.

Lastly, the fundamental role of the HPV status was assessed in this study; an issue introduced by our dataset is the significant lack of patients providing this feature. Consequently, our model proved of limited effectiveness when dealing with our original cohort, and conversely it achieved a more satisfactory performance when only the small portion of subjects provided with HPV status was selected for the development of the model. As previously depicted and discussed, the model proved rather effective when predicting the events affecting this subgroup. Therefore, while this analysis highlighted a crucial limitation in our dataset, it simultaneously emphasized the predictive strength of the HPV value. Such encouraging finding suggests that a future significant improvement could be introduced through the careful selection of the subjects provided with this information. As our model proved some predictive ability when only learning from such a low number of labeled samples, it is reasonable to believe that an

enlargement of a suitable dataset, together with an appropriate imaging data processing, could achieve an extremely effective performance.

CHAPTER 5
CONCLUSIONS

The present work aimed at exploring the feasibility of the implementation of Time to Event analysis for patients affected by Head and Neck Squamous Cell Carcinoma. Our method employes a multi-modal Deep Learning model that, when input with clinical, PET and CT data, predicts the time of occurrences of tumor relapse, distant metastasis, and death of the individual patient.

In order to implement such task, a careful analysis of the clinical variables was conducted in order to assess their prognostic value and subsequently optimize their selection and combination into different feature sets.

Once the dataset and parameters of the model were optimized, the performance of the model was evaluated; however, as Time to Event analysis did not provide the wished results, other approaches were implemented and evaluated.

Classification and regression tasks were consequently adjusted to achieve the prediction of the events of interest and to provide insights regarding their timing.

In the end, the classification model for the prediction of the patients who would develop an HNSCC occurrence within 3 years from the diagnosis, proved the most effective among our approaches.

This technique was developed through the training of the Deep Learning model on a small minority of the dataset, constituted by the patients provided with their HPV status; indeed, this achievement proved the strong predictive power of such feature in combination with the others.

Hence, our initial goal of executing a Time to Event analysis basing on clinical and raw imaging data proved unfeasible. However, while raw imaging data was not suitable to appreciably contribute to the predictions, we explored and fully exploited the informative potential of clinical variables. Indeed, the combination of HPV status with other clinical features proved to hold reliable and meaningful information.

The achieved performances are remarkable when considering that they essentially do not account for the imaging data and derive from an extremely restricted dataset. Therefore, our encouraging findings suggest that even just a larger clinical dataset could achieve more accurate predictions if the HPV-related information is ensured. Secondly, as interesting results were obtained without any processing of the imaging data, it is reasonable to assume that an appropriate manipulation and segmentation of CT and PET volumes could introduce notable contribution to the development of an accurate Deep Learning model for this application.

Bibliography

[1] J. S. A. P. R. K. S. A. B. Adam Barsouk, «Epidemiology, Risk Factors, and Prevention of Head and Neck,» *medical sciences ,* 2023.

[2] «Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas.,» *Nature,* n. 517, pp. 576-582, 2015.

[3] C. J. e. al, «Genetic progression model for head and neck cancer: implications for field cancerization.,» *Cancer Res,* vol. 56, pp. 2488-2492, 1996.

[4] S. I. &. W. W. H. Pai, «Molecular pathology of head and neck cancer: implications for diagnosis, prognosis and treatment,» *Annu. Rev.. Pathol.,* vol. 4, pp. 49-70, 2009.

[5] S. e. a. Krishnamurthy, «Endothelial cell-initiated signaling promotes the survival and self-renewal of cancer stem cells.,» *Cancer Res.,* vol. 70, p. 9969–9978, 2010.

[6] R. e. a. Mandal, «The head and neck cancer immune landscape and its immunotherapeutic implications.,» *JCI Insights,* 2016.

[7] J. M. e. a. Brooks, «Development and validation of novel microenvironment-based immune molecular subgroups of head and neck squamous cell carcinoma: implications for immunotherapy.,» *Ann. Oncol. ,* vol. 30, pp. 68-75, 2019.

[8] H. Sung, J. Ferlay, R. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal e F. Bray, «Global Cancer Statistics 2020: GLOBOCAN,» *CA Cancer J. Clin.,* 2021.

[9] «https://thancguide.org/cancer-types/throat/laryngeal/glottic/anatomy/,» [Online].

[10] S. C. Maria Elisa Sabatini, «Human papillomavirus as a driver of head and neck cancers,» *British Journal of Cancer,* vol. 122, pp. 306-314, 2020.

[11] R. D. A. S. Rosen, TNM classification, StatPearls Publishing, 2022.

[12] A. J. M. W. S. N. G. B. K. M. M. J. A. K. Garden AS, «Is concurrent chemoradiation the treatment of choice for all patients with Stage III or IV head and neck carcinoma?,» *Cancer,* 2004.

[13] M. M. e. a. Nijkamp, «Expression of E-cadherin and vimentin correlates with metastasis formation in head and neck squamous cell carcinoma patients.,» *Radiother. Oncol.,* vol. 99, 2011.

[14] B. P. S. S. V. S. S. N. S. P. M. &. P. P. S. Patel, «Clinical significance of MMP-2 and MMP-9 in patients with oral cancer.,» *Head Neck ,* vol. 29, 2007.

[15] D. M. S. G. S. P. L. R. S. R. L. &. D. M. W. Brizel, «Tumor hypoxia adversely affects the prognosis of carcinoma of the head and neck.,» *Int. J. Radiat. Oncol. Biol. Phys.,* 1997.

[16] E. L. O. C. S. P. N. B. J. &. H. E. M. Gottgens, «HPV, hypoxia and radiation response in head and neck cancer,» *Br. J. Radiol.,* 2019.

[17] K. K. e. a. Ang, «Human papillomavirus and survival of patients with oropharyngeal cancer.,» *N. Engl. J. Med.,* vol. 363, pp. 24-35, 2010.

[18] L. M. Braaten KP, «Human Papillomavirus (HPV), HPV-Related Disease, and the HPV Vaccine.,» *Rev Obstet Gynecol.,* 2008.

[19] L. C. M. A. N. C. W. D. M. W. N. Y. M. e. a. Mirabello, «The intersection of HPV epidemiology, genomics and mechanistic studies of HPV-mediated carcinogenesis.,» *Viruses,* n. 10, 2018.

[20] S. Zhang, B. Wang, F. Ma, F. Tong, B. Yan, T. Liu, H. Xie, L. Song, S. Yu e L. Wei, «Characteristics of B lymphocyte infiltration in HPV + head and neck squamous cell carcinoma.,» *Cancer Sci.,* 2021.

[21] K. Ang, J. Harris, R. Wheeler, R. Weber, D. Rosenthal, P. Nguyen-Tân, W. Westra, C. Chung, R. Jordan, C. Lu e e. al., «Human Papillomavirus and Survival of Patients with Oropharyngeal Cancer.,» *N. Engl. J. Med.,* 2010.

[22] M. B. e. a. Amin, AJCC Cancer Staging Manual, Springer, 2017.

[23] N. e. a. Wurdemann, «Prognostic impact of AJCC/UICC 8th edition new staging rules in oropharyngeal squamous cell carcinoma.,» *Front. Oncol. ,* vol. 7, 2017.

[24] S. Morita, M. Yano, T. Tsujinaka, Y. Akiyama, M. Taniguchi, K. Kaneko, H. Miki, T. Fujii, K. Yoshino, H. Kusuoka e e. al., «Genetic Polymorphisms Of Drug-Metabolizing Enzymes And Susceptibility to Head-And-Neck Squamous-Cell Carcinoma. J.,» *Cancer,* n. 80, pp. 685-688, 1999.

[25] D. Zandberg, S. Liu, O. Goloubeva, R. Ord, S. Strome, M. Suntharalingam, R. Taylor, R. Morales, J. Wolf, A. Zimrin e e. al., «Oropharyngeal cancer as a driver of racial outcome disparities in squamous cell carcinoma of the head and neck: 10-year experience at the University of Maryland Greenebaum Cancer Center.,» *Head Neck,* n. 38, pp. 564-572, 2015.

[26] H. Rumgay, N. Murphy, P. Ferrari e I. Soerjomataram, «Alcohol and Cancer: Epidemiology and Biological Mechanisms.,» *Nutrients,* n. 13, p. 3173, 2021.

[27] B. B. L. C. L. V. B. J. G. J. Johnson DE, «Head and neck squamous cell carcinoma.,» *Nat Rev Dis Primers.,* 2020.

[28] U. E. S. B. B. C. B. B. C. T. Beichel RR, «FDG PET based prediction of response in head and neck cancer treatment: assessment of new quantitative imaging features,» *PLoS One,* 2019.

[29] A. A. e. a. Forastiere, «Concurrent chemotherapy and radiotherapy for organ preservation in advanced laryngeal cancer.,» *N Engl. J. Med. ,* 2003.

[30] N. C. J. e. a. Lee, «Patterns of failure in high-metastatic node number human papillomavirus-positive oropharyngeal carcinoma.,» *Oral Oncol.,* 2018.

[31] A. A. e. a. Forastiere, «Concurrent chemotherapy and radiotherapy for organ preservation in advanced laryngeal cancer.,» *N. Engl. J. Med.,* 2003.

[32] A. A. e. a. Forastiere, «Long-term results of RTOG 91-11: a comparison of three nonsurgical treatment strategies to preserve the larynx in patients with locally advanced larynx cancer.,» *J. Clin. Oncol.,* vol. 31, 2013.

[33] E. M. e. a. Rettig, «Health-related quality of life before and after head and neck squamous cell carcinoma: analysis of the Surveillance, Epidemiology, and End Results–Medicare Health Outcomes Survey linkage.,» *Cancer,* vol. 122, pp. 1861-1870, 2016.

[34] H. M. &. M. R. P. Mehanna, «Deterioration in quality-of-life of late (10-year) survivors of head and neck cancer.,» *Clin. Otolaryngol. ,* vol. 31, 2006.

[35] «SURVIVAL ANALYSIS REGRESSION (Social Science),» [Online]. Available: http://what-when-how.com/social-sciences/survival-analysis-regression-social-science/.

[36] B. M. L. S. A. D. Clark TG, «Survival analysis part I: basic concepts and first analyses,» *Br J Cancer,* 2003.

[37] Ø. B. Håvard Kvamme, «Continuous and Discrete-Time Survival Prediction with Neural Networks,» 2019.

[38] S. F. G. L. P. S. Sabrina De Capitani di Vimercati, «Data Privacy: Definitions and Techniques,» *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* 2012.

[39] B. C. F. I. P. L. M. v. d. S. Alexander Norcliffe, «SurvivalGAN: Generating Time-to-Event Data for Survival Analysis,» 2023.

[40] W. J. F. T. Blagoev KB, «Hazard ratios in cancer clinical trials–a primer,» *Nat Rev Clin Oncol. ,* p. 178–183, 2012.

[41] L. S. M. S. Hosmer DW, «Descriptive methods for survival data. In: Applied Survival Analysis,» *NJ: John Wiley & Sons,* p. 16–66, 2008.

[42] «Cox DR. Regression models and life-tables,» *J R Stat Soc Series B.,* p. 187–220, 1972.

[43] C. T. L. S. A. D. Bradburn MJ, «Survival analysis part II: multivariate data analysis–an introduction to concepts and methods,» *Br J Cancer,* p. 431–436, 2003.

[44] A. L. B. a. I. S. K. K.-H. Yu, «Artificial intelligence in healthcare,» *Nat. Biomed. ,* pp. 719-731, 2018.

[45] G. a. L. M. O. Briganti, «Artificial Intelligence in Medicine: Today and Tomorrow,» *Frontiers in Medicine,* vol. 7, 2020.

[46] Y. H. K. H. L. a. L. M. Chen, «Disease prediction by machine learning over big data from healthcare communities,» *IEEE Access,* vol. 5, p. 8869–8879, 2017.

[47] N. T. e. al., «A clinically applicable approach to continuous prediction of future acute kidney injury,» *Nature,* vol. 572, p. 116–119, 2019.

[48] D. A. K. M. P. T. M. C. P. G. T. Kaur T, «Artificial Intelligence in Epilepsy,» *Neurol India,* 2021.

[49] O. S. M. M. J. W. P. N. David M Harmon, «Artificial Intelligence for the Detection and Treatment of Atrial Fibrillation,» *Arrhythmia & Electrophysiology Review,* 2023.

[50] S. K. R. D. V. V. M. A. Bera K, «Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology,» *Nat Rev Clin Oncol,* 2019.

[51] [Online]. Available: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices.

[52] «Ethics and governance of artificial intelligence for health: WHO guidance,» in *Geneva: World Health Organization*, 2021.

[53] K. T. N. T. S. A. C. J. D. K. Boniol M, «The global health workforce stock and distribution in 2020 and 2030: a threat to equity and 'universal' health coverage?,» *BMJ Glob Health,* 2022.

[54] C. M. S. I. K. B. Y. A. Z. G. R. M. A. I. A. M. A. A. A. A.-A. A. M. A. e. a. Annie Haakenstad, «Measuring the availability of human resources for health and its relationship to universal health coverage for 204 countries and territories from 1990 to 2019: a systematic analysis for the Global Burden of Disease Study 2019,» *The Lancet,* 2022.

[55] E. e. a. Williams, «The image data resource: a bioimage data integration and publication platform,» *Nat. Methods 14,* p. 775–781, 2017.

[56] V. S. K. L. &. C. A. E. Ljosa, «Annotated high-throughput microscopy image sets for validation,» *Nat. Methods 9,* 2012.

[57] K. C. P. &. W. M. Tomczak, «The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge,» *Contemp. Oncol. ,* pp. 68-77, 2015.

[58] C. e. a. [Sudlow, «UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age,» *PLoS Med.,* 2015.

[59] I. S. G. E. H. Alex Krizhevsky, «ImageNet Classification with Deep Convolutional Neural Networks,» *Advances in Neural Information Processing Systems 25,* 2012.

[60] E. S. T. D. Jonathan Long, «Fully Convolutional Networks for Semantic Segmentation,» 2015.

[61] F. A. A. L. F. R. Ivano Lauriola, «Learning adaptive representations for entity recognition in the biomedical domain,» *Journal of Biomedical Semantics,* 2021.

[62] Z. T. C. D. Shamout F, «Machine Learning for Clinical Outcome Prediction,» *IEEE Rev Biomed Eng,* pp. 116-126, 2021.

[63] M. G. D. M. C. a. Z. O. M. Makar, «Short-term mortality prediction for elderly patients using medicare claims data,» *Int. J. Mach. Learn. Comput.,* vol. 5, pp. 192-197, 2015.

[64] M. S. C. S. A. H. J. Hardev S. Grewal, «Prediction of the output factor using machine and deep learning approach in uniform scanning proton therapy,» *Journal of applied clinical medical physics,* 2020.

[65] X. H. G. C. T. H. S. X. S. X. G. D. Z. C. Hong L, «Prediction of low cardiac output syndrome in patients following cardiac surgery using machine learning,» *Front Med (Lausanne),* 2022.

[66] P. S. S. A. J. e. a. Mascheroni, «Improving personalized tumor growth predictions using a Bayesian combination of mechanistic modeling and machine learning,» *Commun Med 1,* 2021.

[67] E. B. P. L. &. M. E. S. M. 1. B. E. [Biganzoli, «Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach,» *Stat. Med.,* p. 1169–1186, 1998.

[68] E. B. P. &. M. E. Biganzoli, «A general framework for neural network models on censored survival data,» *Neural Networks 15,* p. 209–218 , 2002.

[69] H. K. U. B. E. &. L. M. [Ishwaran, «Random survival forests,» *Ann. Appl. Stat,* p. 841–860 , 2008.

[70] H. K. U. G. E. M. A. J. &. L. M. Ishwaran, «High-dimensional variable selection for survival data,» *J. Am. Stat. Assoc. 105,* p. 205–217, 2010.

[71] X. &. I. H. Chen, «Random forests for genomic data analysis,» *Genomics 99,* p. 323– 329, 2012.

[72] H. M. Z. L. Gong X, «Big Data Toolsets to Pharmacometrics: Application of Machine Learning for Time-to-Event Analysis,» *Clin Transl Sci.,* 2018.

[73] U. S. A. C. J. B. T. J. Y. K. Jared L. Katzman, «DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network,» *BMC Medical Research Methodology,* 2018.

[74] W. R. Z. J. Y. a. M. v. d. S. Changhee Lee, «Deephit: A deep learning approach to survival analysis with competing risks,» *AAAI Conference on Arti_cial Intelligence,* 2018.

[75] Ø. B. I. S. Håvard Kvamme, «Time-to-Event Prediction with Neural Networks and Cox Regression,» *Journal of Machine Learning Research,* 2019.

[76] S. M. M. S. Achraf Bennis, «Estimation of conditional mixture Weibull distribution with right-censored data using neural network for time-to-event analysis».

[77] B. N. Michael F. Gensheimer, «A scalable discrete-time survival model for neural networks,» 2018.

[78] L. E. A. M. Z. S. W. J. H. A. A. N. M. S. F. G. F. G. S. J. W. F. C. S. N. M. L. J. B. C. R. M. K. C. L. G. Wang Y, «Deep learning based time-to-event analysis with PET, CT and joint PET/CT for head and neck cancer prognosis,» *Comput Methods Programs Biomed,* 2022.

[79] C. D. L. L. J. G. E. S. H. S. S. D. P. L. W. G. P. M. A. V. G. C. G. C. S. C. R. L. Irma Verdonck-de Leeuw, «European Head and Neck Society recommendations for head and neck cancer survivorship care,» *Oral Oncology,* vol. 133.

[80] V. B. S. K. F. J. K. J. K. P. e. a. Clark K, «The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository,» *J Digit Imaging,* 2013.

[81] P. C. O. L. B.-B. B. S. J. K. J. e. a. Hofheinz F, «Automatic Volume Delineation in Oncological PET. Evaluation of a Dedicated Software Tool and Comparison With Manual Delineation in Clinical Data Sets.,» *Nuklearmedizin ,* 2012.

[82] L. J. P. J. B.-B. B. S. J. K. J. e. a. Hofheinz F, «An Automatic Method for Accurate Volume Delineation of Heterogeneous Tumors in Heterogeneous Tumors in PET.,» *Med Phys,* 2013.

[83] W. J. L. E. M. S. H. M. B. M. Z. D. L. Y. L. Q. A. H. T. E. v. d. H. J. B. V. K. J. F. K. K. E. K. D. G. V. H. A. A. N. N. P. e. a. Zschaeck S, «18F-Fluorodeoxyglucose Positron Emission Tomography of Head and Neck Cancer: Location and HPV Specific Parameters for Potential Treatment Individualization,» *Frontiers in Oncology,* 2022.

[84] B. B. L. Q. F. R. Cramer JD, «The changing therapeutic landscape of head and neck cancer,» *Nature reviews Clinical oncology,* 2019.

[85] H. P. Kimple RJ, « The prognostic value of HPV in head and neck cancer patients undergoing postoperative chemoradiotherapy,» *Ann Transl Med.,* 2015.

[86] M. Gensheimer, «Github,» 2019. [Online]. Available: https://github.com/MGensheimer/nnet-survival/blob/master/nnet_survival.py.

[87] «https://it.mathworks.com/discovery/cross-validation.html,» [Online].

[88] 2. L. C. B.-N.-S. 3. . IGO., «Organization, Ethics and governance of artificial intelligence for health: WHO guidance,» in *Geneva: World Health;*.

[89] A. J. K. H. S. e. a. Avati, «Improving palliative care with deep learning,» *BMC Med Inform Decis Mak 18 (Suppl 4),* 2018.

[90] M. S. C. S. A. H. J. Hardev S. Grewal, «Prediction of the output factor using machine and deep learning approach in uniform scanning proton therapy,» *Journal of applied clinical medical physics.*

[91] mnhgfds, p. hgfd.

[92] T. Y. K. T. F. M. N. K. U. Y. Torizuka T, «Prognostic value of 18F-FDG PET in patients with head and neck squamous cell cancer,» *AJR Am J Roentgenol,* 2009.

[93] X. e. a. Leon, «Second, third, and fourth head and neck tumors. A progressive decrease in survival.,» *Head Neck,* 2012.

[94] C. B. C. W. K. D. &. K. A. Gotz, «Detection of HPV infection in head and neck cancers: Promise and pitfalls in the last ten years: a meta-analysis.,» *Mol. Clin. Oncol. ,* n. 10, pp. 17-28, 2019.