Multimodal learning to predict breast cancer prognosis

Ndèye Maguette Mbaye

PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à Mines Paris-PSL

Apprentissage à partir de données multimodales pour prédire le pronostic du cancer du sein

Multimodal learning to predict breast cancer prognosis

Soutenue par
Ndèye Maguette MBAYE
Le 05 septembre 2024

Ecole doctorale n° 621
Ingénierie des Systèmes, Matériaux, Mécaniques, Energétiques

Spécialité
Bio-informatique

Composition du jury :

Marc LELARGE Directeur de recherche INRIA & ENS, France	*Président du jury*
Martin BØGSTED Professeur Aalborg University, Danemark	*Rapporteur*
Alice OTHMANI Maîtresse de conférences Université Paris-Est, France	*Rapporteuse*
Michal ROSEN-ZVI Docteur IBM Research & The Hebrew University of Jerusalem, Israel	*Examinatrice*
Antoine RECANATI Ingénieur de recherche SANCARE, France	*Examinateur*
Chloé-Agathe AZENCOTT Professeure Mines Paris-PSL, France	*Directrice de thèse*

MINES PARIS | PSL

Acknowledgement

Tout d'abord, je remercie mes rapporteurs Alice OTHMANI et Martin BØGSTED, d'avoir pris le temps pour relire mon manuscrit. Je remercie également mes examinateurs Michal ROSEN-ZVI, Marc LELARGE et Antoine RECANATI et à Anne-Sophie HAMY-PETIT d'avoir accepté de juger ce travail et pour leur remarques pertinentes qui ont grandement contribué à améliorer cette thèse. Je suis honorée d'avoir pu discuter de mes travaux de thèse avec d'éminents médecins et chercheurs comme vous.

Cette thèse marque l'aboutissement de plusieurs années de travail et de sacrifices. Elle n'aurait pu voir le jour sans le soutien, l'encouragement et l'aide de nombreuses personnes, auxquelles j'exprime ici toute ma gratitude.

Tout d'abord, je tiens à adresser ma profonde reconnaissance à ma directrice de thèse, Chloé. Merci d'abord de m'avoir fait confiance et de m'avoir donné l'opportunité de mener ce projet. Merci pour ton expertise, la richesse de tes enseignements et tes conseils avisés tout au long de ce parcours. Tu as su m'aider à perserverer et me guider méthodologiquement quand les obstacles se faisaient nombreux. Merci aussi pour ta très grande disponibilité et surtout ta positivité et ta patience. Ca a été une très belle expérience avec toi comme encadrante et j'espère que nos chemins se recroiseront un jour dans le futur.

I would also like to thank Michal and Michael for their guidance and their invaluable advices. Thank you for you availability, as well as for the stimulating and constructive discussions that enriched this research.
Thanks also to each member of the MLFPM ITN network. Walking this journey with you has been incredibly inspiring. I am truly grateful to be surrounded by such driven and brilliant individuals. A special thanks to Katharina for all the help through this journey.

Un grand merci à la famille du CBIO: Adeline, Asma, Arthur, Aurélie, Daniel, Florian, Giann, Gwenaëlle, Gwenn, Héctor, Julian, Julie, Katia, Lotfi, Matthieu, Mélanie, ML, Mounir, Paul, Philippe, Romain, Simon, Thomas B., Thomas W., Tom, Tristan, Véronique, Victor, Vincent et Vivien. Merci pour les moments de convivialité qui ont ponctué ces années de travail. Je suis reconnaissante pour la sympathie de chacun de vous, pour les discussions lors des pauses cafés, et surtout, pour votre patience à la cantine. Je dois dire qu'une ambiance comme celle au CBIO aide à rendre ce défi qu'est la thèse plus léger à porter.

Je tiens aussi à remercier les membres du RT2Lab, spécialement Elise, Anne-Sophie, Aullène et Beatriz. De même que Johan et Julien du département des données à l'Institut Curie, pour leur disponibilité et leur aide, toujours précieuse, à chaque fois que j'en ai besoin. Merci également à Antoine, Eric, et à Charles pour votre accompagnement et vos idées lors de nos discussions.

Je tiens également à exprimer ma gratitude à l'ensemble du personnel administratif des Mines et de l'Institut Curie, à savoir Katy, Caroline, Christine et Catherine pour leur aide précieuse et leur disponibilité.

À ma famille, je souhaite exprimer toute ma reconnaissance pour leur soutien, leur patience et leur soutien indéfectible. Merci à mes parents, Diewo et Djibril, pour leurs prières, pour leurs encouragements, pour leur confiance en moi et pour avoir toujours cru en mes capacités. À mes frères et soeurs: Mbathio, Serigne, Sali, Marième et Pape Diamé, merci pour votre encouragements teintés de chariades, et de m'avoir aidé à garder le cap et à ne perdre de vue mon objectif. Finalement, je termine cette salve de reconnaissance par ma famille: mon cher époux Yerim et mon fils. Merci à mon mari pour sa compréhension et son soutien inconditionnel tout au long de cette aventure et à mon fils d'être ma source de motivation au quotidien.

Enfin, je remercie toutes les personnes, nommées ou non, qui ont contribué de près ou de loin à l'aboutissement de ce travail. Votre aide et votre soutien ont été inestimables.

Contents

List of Figures

List of Tables

List of abbreviations

ACR American College of Radiology

AI Artificial Intelligence

ANN Artificial Neural Network

AOPC Area Over Prediction Curve

APS Average Precision Score

BC Breast Cancer

BCE Binary Cross Entropy

BERT Bidirectional Encoder Representations from Transformers

BI-RADS Breast Imaging Reporting And Data System

CNIL Commission National de l'Informatique et des Libertés

DL Deep Learning

DFS Disease Free Survival

EFS Event Free Survival

EHR Electronic Health Records

EPG Excellent Prognosis Group

ER Estrogen Receptor

FFNN Feed Forward Neural Network

FN False Negative

FP False Positive

GALE Global Aggregation of Local Explanations

GDPR General Data Protection Regulation

GPT Generative Pre-trained Transformers

HER2 Human Epidermal Growth Factor Receptor 2

HR Hormone receptor

ICU Intensive Care Unit

IG Integrated Gradients

LIME Local Interpretable Model-agnostic Explanations

LR Logistic Regression

LSTM Long Short-Term Memory

MAE Mean Absolute Error

MDA Mean Decrease in Accuracy

MDI Mean Decrease Impurity

ML Machine Learning

MLM Masked Language Modelling

MLP Multi-layers Perceptron

MSE Mean Square Error

MPG Moderate Prognosis Group

NER Name Entity Recognition

NLP Natural Language Processing

NPI Nottingham Prognosis Index

NSP Next Sentence Prediction

ORI Out of Range Index

OS Overall Survival

pCR Pathological Complete Response

PgR Progesterone Receptor

PPG Poor Prognosis Group

RF Random Forests

RFS Relapse Free Survival

RNN Recurrent Neural Network

NER Name Entity Recognition

SHAP SHapley Additive exPlanations

SVM Support Vector Machine

TN True Negative

TNBC Triple Negative Breast Cancer

TP True Positive

t-SNE t-distributed stochastic neighbor embedding

VPPG Very Poor Prognosis Group

WHO World Health Organization

Chapter 1

Introduction

Breast cancer is one of the most prevalent forms of cancer worldwide. According to the World Health Organisation [WHO (2022a)], it accounts for 12,5% of all new annual cancer cases globally and in 2022, approximatively 2.3 million women were diagnosed with breast cancer and it resulted in more than 666,000 deaths worldwide. These statistics underscore the critical need for ongoing research to improve care strategies. Over the past few decades, research has made major progress in understanding the biology, risk factors and the disease's progression, which have lead to significant improvements in diagnosis, treatment and patient outcomes. And this has been achieved with the use of a comprehensive set of information from different domains and different modalities.

Electronic Health records have revolutionized the landscape of healthcare delivery and clinical research and its integration in clinical practice has offered multiple advantages that contribute to the improvement of breast cancer survival rates. In clinical research, it provides a robust infrastructure for data collection and analysis and real-world data from EHR provide valuable insights into everyday clinical practice. Many studies have made use of this information for breast cancer research. One of the primary advantages of using EHR is their multimodal nature. EHRs integrate various types of data, including patient history, lab results, imaging, and treatment strategies, all in one accessible platform. For Breast cancer studies, most of the papers in the literature has used a combination of structured data and imaging as these are the most accessible information. For instance, many robust BC diagnosis tools have been developed using clinical descriptors or family background and/or breast imaging (mammogram, ultrasound or MRI). Those tools have make clinicians' workflow more efficient and allow them to identify BC that would otherwise been undetectable in its early stages.

However, up to my knowledge, there are few BC studies that use the combination of structured information from EHR and free-text medical reports. Yet, it exists a huge amount of meaningful information in biological measurements, clinical information and mostly in patients reports, in a patient journey that remain unexploited. The main challenge in using unstructured information such as text that it can contain sensitive patient information, therefore, it needs to ensure compliance with data protection regulation. Moreover, the unstructured patient data might be complex to analyze (med-

ical jargon, abbreviations, variability in terminology etc.). These factors may result in a greater reliance on structured EHR data and imaging for BC studies.

In my thesis, we propose to develop and apply machine learning techniques to predict BC outcome (such as recurrence or survival), using a multimodal breast cancer patient data that include medical notes in natural languages and in french, the outcome of various lab analysis and clinical descriptors from a vast cohort from the Institut Curie. *First*, we built and trained multiple classical machine learning models for different modalities integration to predict survival endpoint and give first insight about models' predictive factors. *Then*, we developed deep learning models that used a sequential data representation for tabular information in the EHR to predict BC survival endpoints. *After that*, we performed the same task with the sequences of free-text reports throughout the patient trajectory. *Finally*, we combined deep learning models built on the different modalities into one. *Altogether*, we found features within these different modalities that can be used to improve breast cancer outcomes.

1.1 Organization and contributions of the thesis

In **Chapter** 2, I present the background knowledge about breast cancer, which includes its epidemiology, its intrinsic characteristics, the possible treatment strategies and its related survival endpoints. I also define important notions of Electronic Health Records (EHR), its challenges and its importance and limitations in clinical research.

The **Chapter** 3 is about the mathematical and ML concepts that underlies this thesis. I provide the general intuition behind methods and techniques that have been used during this Ph.D. All these concepts are specific to medical multi-modal patient data.

In **Chapter** 4, I present a challenge that I worked on, prior to the Ph.D project, called PhysioNet Computing in Cardiology Challenge (2012). I show in this chapter methods that I develop for mortality rates prediction in Intensive Care Unit (ICU). The focus on this challenge is due to the similarity between the challenge's data and my project's data.

In **Chapter** 5, I describe the multimodal cohort used during this thesis. It combined structured information from the SEIN database in Institut Curie and the free-text reports from the ELIOS database for the same patients. Then, I apply multiple ML models that use different integration methods to the dataset to predict survival outcome. We compare results for the different integration methods and we propose a set of important features using interpretation methods. This chapter was presented during the Machine Learning Frontier for Precision Medicine (MLFPM) conference in Munich in 2022.

In **Chapters** 6, 7 and 8, I propose a sequential representation of the multimodal EHR. I apply transformers-based deep learning models to predict survival endpoints. I first apply it to tabular patient trajectories, then to text reports trajectories and I finally combine both models into a multimodal transformers-based model for EHR we called M-BEHRT. I present the results and discuss about most predictive features from these robust DL methods. These chapters were presented during the Winter School in Computer Science and Engineering in Jerusalem in 2023 and during the Personalized Health Conference in Basel in 2024. They will be subject of a future publication.

Chapter 2

General context

Abstract:

Breast cancer treatment remains a significant global health challenge, necessitating continual advancements in diagnostic, treatment, and management strategies. This chapter explores the use of Electronic Health Records (EHR) to uncover previously unidentified prognostic factors in breast cancer. Leveraging the wealth of patient data within EHR, my thesis employs advanced ML algorithms to analyze complex relationships and patterns, ultimately aiming to enhance the accuracy and precision of breast cancer prognosis. In this chapter, I will highlight the general context of my project. It includes the scientific context, such as several important notions in Breast Cancer (BC) and Electronical Health records (EHRs), but also the technical environment in which the project took place.

Résumé:

Le traitement du cancer du sein reste un défi sanitaire mondial important, qui nécessite des progrès constants dans les stratégies de diagnostic, de traitement et de gestion. Ce chapitre explore l'utilisation des dossiers médicaux électroniques (DME) pour découvrir des facteurs pronostiques précédemment non identifiés dans le cancer du sein. En s'appuyant sur la richesse des données des patients contenues dans les DME, cette étude utilise des algorithmes ML avancés pour analyser des relations et des modèles complexes, dans le but ultime d'améliorer l'exactitude et la précision du pronostic du cancer du sein. Dans ce chapitre, je mettrai en évidence le contexte général de mon projet. Il comprend le contexte scientifique tel que plusieurs notions importantes dans le domaine du cancer du sein (CB) et des dossiers médicaux électroniques (DME), mais aussi l'environnement technique dans lequel le projet se déroulera.

Contents

2.1 Introduction to Breast Cancer

In this section, I will present important notions of breast cancer (BC), including its statistics and crucial definitions.

2.1.1 Cancer epidemiology

Cancer refers to a large number of diseases characterized by any type of rapidly growing of cells that spreads throughout the body. It develops when the body's normal control mechanisms of cell division stop working. Hence, it is manifested by the development of abnormal cells that divide uncontrollably and have the ability to infiltrate and destroy normal body tissue. Old cells do not die and instead grow out of control, forming new, abnormal cells. These extra cells may form a mass of tissue, called a tumor. Cancer can occur anywhere in the body. There are more than 200 types of cancers with various treatments and various prognosis.

Cancer is one of the most common diseases in the world. According to the GLOBOCAN estimations in 2022 [Bray *et al.* (2024)], produced by the International Agency for Research Cancer, there were nearly 20 millions new cases of cancer in 2022. Per their findings, approximately one in five men or women develop cancer in a lifetime. Lung cancer was the most commonly diagnosed cancer globally, accounting for nearly 2.5 million new cases, i.e, 1 in 8 of all cancers worldwide (12.4%), followed closely by breast cancer (11.6%), colorectal cancer (9.6%), prostate cancer (7.3%), and stomach cancer (4.9%). Cancer incidence varies between countries, as shown in Figure 2.1. It is the second leading cause of death globally, accounting for an estimated 9.7 million deaths in 2022 [WHO (2022b)]. Lung cancer retained its status as the primary cause of cancer-related deaths, claiming approximately 1.8 million lives (18.7%), trailed by colorectal cancer (9.3%), liver cancer (7.8%), breast cancer (6.9%), and stomach cancer (6.8%). Lung, prostate, stomach, liver and colorectal cancer are the most lethal cancer types among men, while breast, cervical and lung cancer are the most lethal ones among women (2.2).

Additionally, studying variations in cancer incidence among different demographic groups and geographic regions helps identify disparities and target interventions to reduce cancer burden and improve outcomes. Regional disparities were evident, with incidence rates varying considerably, as depicted in Figure 2.3. Also, nearly half of all cancer cases (49.2%) and the majority of cancer-related deaths (56.1%) occurred in Asia. Moreover, in Asia and in Africa, the disparity between the cancer incidence and mortality is particularly pronounced. Indeed, cancer incidence and mortality are correlated with the Human Development Index [Bray *et al.* (2024)]. Incidence and mortality rates are higher in higher HDI countries.

These figures underscore the pervasive impact of cancer within society.

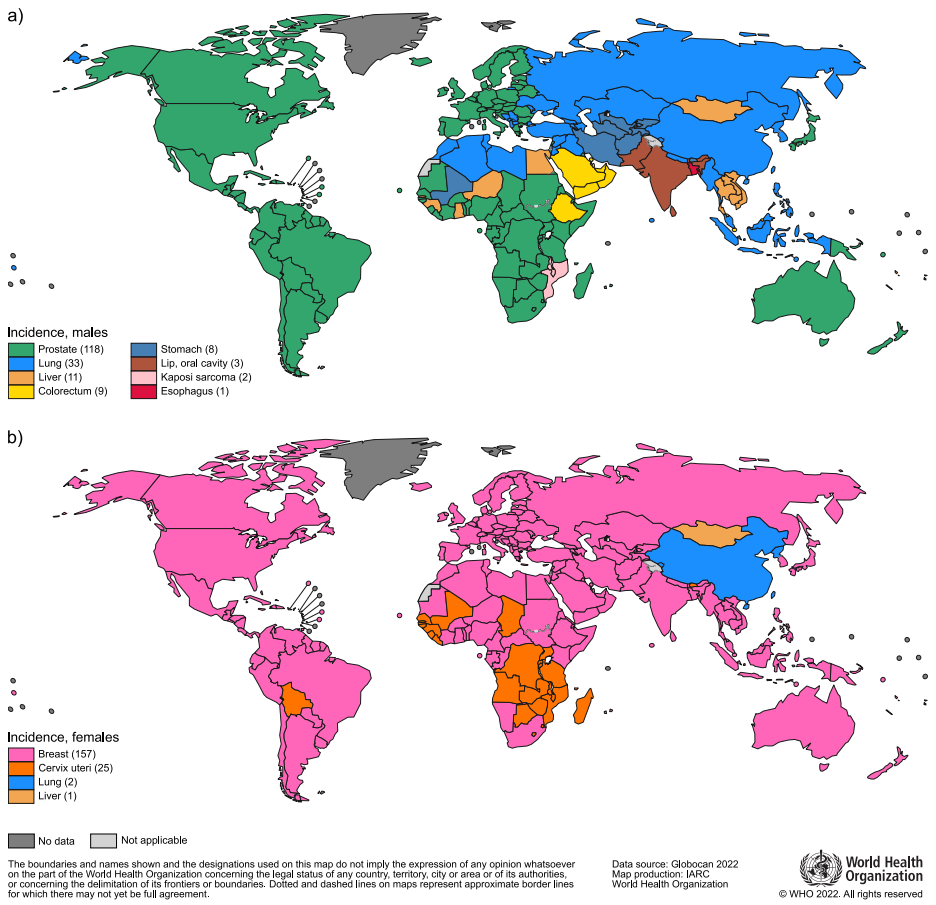
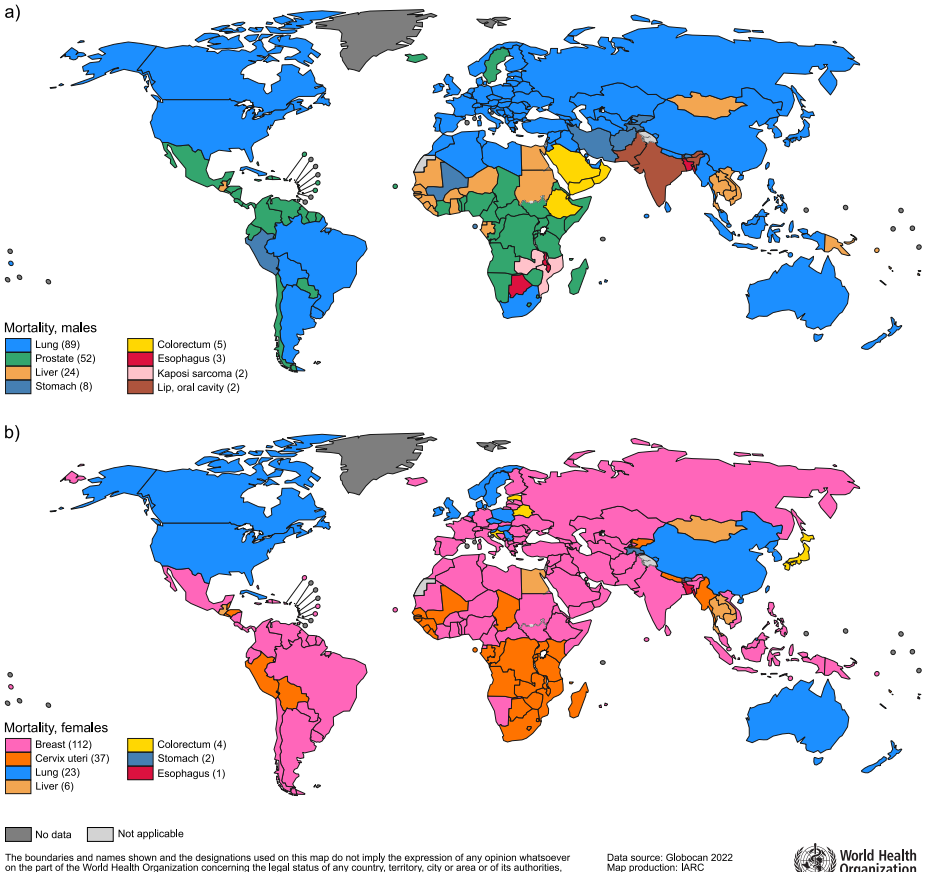


Figure 2.1: Most common cancer types by country in 2022. a. Males, b. Females. Source: GLOBOCAN 2022.



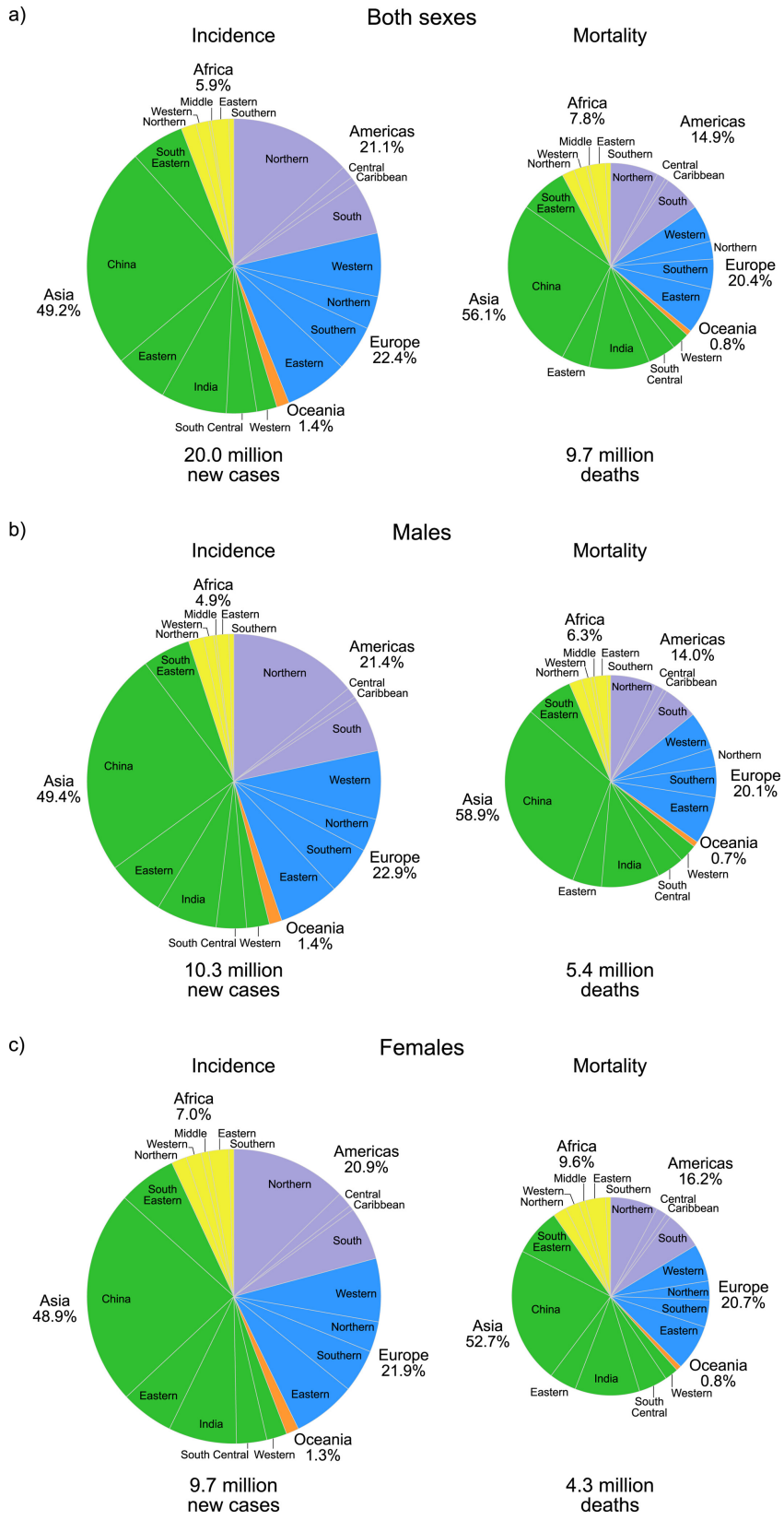


Figure 2.3: Pie chart with distribution of the incidence and the mortality rate by country in 2022 among A. males and among B. females. Female. Source: GLOBOCAN 2022.

The cancer burden continues to grow globally, exerting tremendous physical, emotional and financial strain on individuals, families, communities and health systems. Many health systems in low- and middle-income countries are least prepared to manage this burden, and large numbers of cancer patients globally do not have access to timely quality diagnosis and treatment. In countries where health systems are strong, survival rates of many types of cancers are improving thanks to accessible early detection, quality treatment and survivorship care. Among these, breast cancer stands out as a prime example of progress in combating this disease. (not sure about this phrase)

Breast cancer is the most common cancer among women. It affects around 2 million women each year. In 2022, 2.3 million women received a diagnosis of breast cancer, resulting in 670,000 deaths [WHO (2022a)]. Breast cancer can affect women in every country worldwide, striking at any age post-puberty, though its incidence tends to rise notably in later stages of life. Global estimates highlight significant disparities in the burden of breast cancer based on levels of HDI. For example, in nations with very high Human Development Index (HDI) scores, approximately 1 in 12 women will receive a breast cancer diagnosis during their lifetime, and 1 in 71 will succumb to the disease. Conversely, in countries with low HDI scores, the likelihood of a woman being diagnosed with breast cancer in her lifetime decreases to 1 in 27, but the mortality rate remains notable, with 1 in 48 women dying from the disease [WHO (2022a)].

2.1.2 Definition of BC

Breast cancer is the type of cancer that starts in the breast. It is one of the most common cancers affecting women worldwide and can also occur in men, albeit less frequently. Breast cancer is composed of multiple subtypes with distinct morphologies and clinical implications. It differs greatly among different patients (inter- tumoral heterogeneity) and within an individual tumor (intra-tumoral heterogeneity).

Traditional histopathological classification aims to classify tumors into subgroups to facilitate clinical decisions. Nowadays, recently developed high-throughput microscopic analyses can provides a more precise understanding of cancer heterogeneity. This heterogeneity arises from many different factors, such as the tumor origin, the tumor invasiveness, molecular alterations etc. Typically, BC forms in either the lobules or the ducts of the breast: we distinguish ductal BC from lobular BC (Figure 2.4). A breast lobule is the gland that produces milk whereas a breast duct is the tube that bring the milk from the breast lobule to the nipple. Moreover, according to the spreading state, BC can be broken into two other main categories: 'invasive' and 'noninvasive' (or in situ). In invasive breast cancer, the tumor has spread outside (metastasized) the breast duct to surrounding normal tissue.

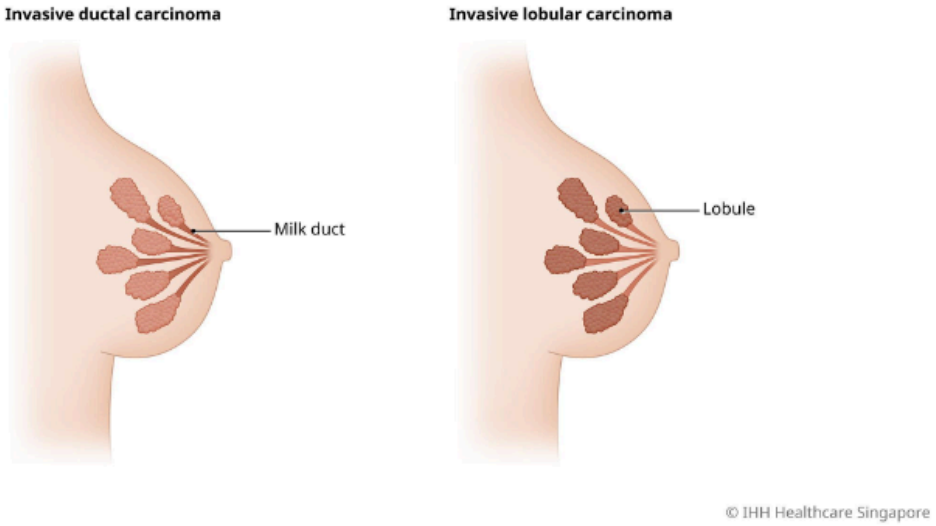


Figure 2.4: Ductal (left) and lobular (right) invasive carcinoma.

It can include the blood cells or the lymph. Noninvasive breast cancer stays with the ducts or lobules. Whether the cancer is noninvasive or invasive, ductal or lobular, will determine your treatment choices and the treatments response. Breast cancer can also be divided into stages based on how large the tumor is and its invasiveness. Cancers that are large and/or have invaded nearby tissues or organs are at a higher stage than cancers that are small and/or still contained in the breast. BC has five (5) main stages (0-5).

In addition to this classification, several studies ([Perou *et al.* (2000a)], [Sørlie *et al.* (2001)]), have shown, with gene expression profiling, that BC could also have molecular heterogeneity within these classes. Research efforts are now focusing on how these different molecular subtypes could provide information on the prognosis and thus, the treatment response, for each BC type. The general aim being to develop more specific new therapies: it is the precision medicine advent.

2.1.3 Molecular characteristics of BC

Breast cancer cells have receptors (proteins) that bind with certain hormones such as progesterone, estrogen and the human epidermal growth factor receptor 2 (HER2). Those hormones are generally involved in the breast cells growth. However, in certain breast cancers, those receptors can be more numerous than in a normal cell. BC cells may have one, both or none of these receptors. As shown in the figure 2.5, according to the presence of certain hormonal receptors, several tumor types can also be distinguished:

- The **hormone receptor tumor (HR+)** is characterized by the pres-

Hormonal receptors	ER	PgR	HER 2	Molecular subtypes
HR+	+	+/-	-	Luminal A
	+/-	+	-	
HER2+	+	+/-	+	Luminal B HER2 +
HER2+	-	-	+	HER2+ non luminal
TNBC	-	-	-	TNBC / Basal like

Table 2.1: Summary of the different molecular breast cancer types

ence of hormone receptors, estrogen and/or progesterone (ER+ and/or PgR+). The tumor development is influenced by the presence of one or of both of these receptors in the cancer cell [You *et al.* (2018)]. This subgroup is also characterized by the absence of overexpression of the growth factor HER-2. Also known as Luminal A, the hormone receptor positive breast cancer has good prognosis.

- The **HER-2 positive (HER2+)** breast cancer is characterized by an overexpression for the HER2 receptor. This hormone, also referred to as HER2/neu proteins is encoded by the gene HER2/neu. Normally, they help control a healthy breast cells growth and repair. But HER2+ breast cancer, this receptor is overexpressed and promotes the uncontrolled growth of cancer cells. When it is combined with the presence of estrogen receptors (ER+) and/or progesterone receptors (PgR+), the subtype is called Luminal B HER2 positive, otherwise, when there is an absence of those receptors, it is called HER2 positive non luminal. The targeted HER2 therapy has helped improving the prognosis of cancer overexpressing HER2 receptors.

- The **triple-negative breast cancers (TNBC)** constitute a heterogeneous group characterized by the lack of estrogen receptors (ER+) and progesterone receptors (PgR+) and the absence of overexpression of the growth factor HER-2. This breast cancer type is associated with a more unfavorable clinical profile, with a high risk of early metastatic relapse because of the aggressive nature of these tumors their partial response to chemotherapy and the absence of a clear therapeutic target, allowing to propose a specific treatment.

Some breast cancers are both hormone receptor-positive and HER2-positive, meaning that estrogen and progesterone can stimulate cell growth [You *et al.* (2018)]. These cancers are often referred to as triple-positive breast cancers. Breast cancers that are estrogen receptor-positive (ER+) and/or progesterone receptor-positive (PR+) are "fueled" by hormones. They are different from breast cancers that are HER2-positive [Callahan & Hurvitz (2011)], in which tumor growth is driven by growth factors that bind to HER2 receptors on the cancer cells [Akshata Desai (2012)]. The molecular

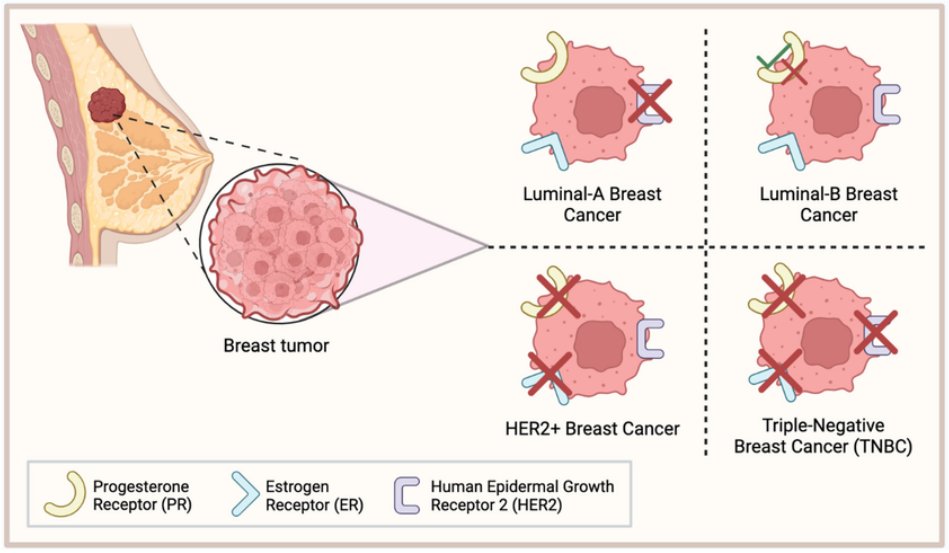


Figure 2.5: Molecular Breast Cancer types [Kirkby *et al.* (2023)].

profile of each BC types is useful in planning treatment and developing new therapies, in conjunction with the stage and the type of the BC.

2.1.4 BC treatment strategy

Fighting cancer usually requires more than one treatment. In most of the cases, clinicians adopt a multidisciplinary approach. The type of cancer, its aggressiveness and the presence or absence of several receptors allow to define the type of treatment:

- **Surgery**: During this process the tumor is removed from the patient's body by a surgeon. It is often the initial treatment for breast cancer and may involve either breast-conserving surgery (lumpectomy) or removal of the entire breast (mastectomy). During breast cancer surgery, the surgeon may also remove one or a few lymph nodes (sentinel nodes) from the underarm area to check for the presence of cancer cells. If cancer cells are found in the sentinel nodes, additional lymph nodes may need to be removed. Moreover, in some cases, particularly if cancer has spread to the lymph nodes, a more extensive lymph node dissection may be performed to remove a greater number of lymph nodes from the underarm area. Surgery may be used alone or in combination with other treatments such as radiation therapy, chemotherapy, hormone therapy, targeted therapy, or immunotherapy, depending on the characteristics of the cancer and the patient's overall health.
- **Chemotherapy**: This cancer treatment works by stopping or slowing the growth of cancer cells. Chemotherapy drugs can be given orally (in

pill form) or intravenously (through a vein). The treatment schedule and duration depend on the specific chemotherapy regimen prescribed by the oncologist. There are several types of chemotherapy drugs used to treat breast cancer, including

  - Anthracyclines: these drugs intercalate with the DNA of cancer cells, interfering with DNA metabolism and the production of RNA, preventing them from multiplying.
  - Taxanes: Examples include paclitaxel (Taxol) and docetaxel (Taxotere). Taxanes disrupt cell division by stabilizing microtubules, structures essential for cell replication.
  - Platinum-Based Drugs: Examples include cisplatin and carboplatin. Platinum-based drugs bind to DNA, causing damage and cell death.
  - Antimetabolites: Examples include 5-fluorouracil (5-FU), capecitabine, and gemcitabine. Antimetabolites interfere with DNA synthesis, preventing cancer cells from reproducing.

  The choice of chemotherapy regimen depends on factors such as the stage and subtype of breast cancer, the patient's overall health, and their individual treatment goals. Chemotherapy is often used in combination with other treatments such as surgery, radiation therapy, hormone therapy, targeted therapy, or immunotherapy to provide comprehensive care for breast cancer patients. It is more generally combined with a surgery that can be done before or after the chemotherapy:

  - Neoadjuvant chemotherapy: the chemotherapy is done in order to reduce the tumor size before the surgery. This treatment is widely used for TNBC.
  - Adjuvant chemotherapy: the chemotherapy is done after the surgery to prevent the risk of relapse by destroying remained cancer cells.

- **Hormone therapy**: It has proven to be an effective and well-tolerated treatment option for HR+ breast cancer, helping to improve outcomes and reduce the risk of recurrence. Hormone therapy works by blocking the effects of estrogen and/or progesterone or reducing their production in the body, thereby slowing or stopping the growth of hormone receptor-positive breast cancer cells. Drugs involved in this therapy are usually taken orally in pill form, once daily or as prescribed by the oncologist. The duration of hormone therapy treatment may vary depending on factors such as the stage and subtype of breast cancer, the patient's menopausal status, and individual treatment goals. Among drugs used for this treatment, we have different drugs family: Selective Estrogen Receptor Modulators (SERMs) such as tamoxifen

or toremifene, Aromatase Inhibitors and the Selective Estrogen Receptors Degraders.

- **Targeted HER2 treatment**: Targeted therapy drugs specifically target proteins or genes that contribute to the growth and spread of cancer cells. They are often used to treat HER2-positive breast cancers, which overexpress the HER2 protein. The HER2 is blocked in order to stop its uncontrolled proliferation. Targeted HER2 therapy drugs are typically administered intravenously through infusion or injection, although some oral formulations are available. As the other treatments, schedules and durations may vary depending on the specific drugs used, the stage and subtype of breast cancer. We count three (3) main groups of antibodies for targeted HER2 treatment:

  - monoclonal antibodies such as trastuzumab (Herceptin), pertuzumab (Perjeta), and trastuzumab emtansine (T-DM1 or Kadcyla). Trastuzumab and pertuzumab may be used in combination to enhance their effectiveness.
  - Tyrosine Kinase Inhibitor (TKIs) such as lapatinib (Tykerb) and neratinib (Nerlynx).
  - Antibody-Drug Conjugates: such as Trastuzumab emtansine (T-DM1 or Kadcyla).

- **Radiotherapy**: Radiation therapy uses high-energy beams to target and destroy cancer cells. It is commonly used after surgery to eliminate any remaining cancer cells in the breast or nearby lymph nodes, or before surgery to shrink the tumor. It is usually administered in a hospital or specialized radiation oncology center. During treatment, the patient lies on a treatment table while a machine delivers the radiation beams to the targeted area.

- **Immunotherapy**: Immunotherapy is a type of cancer treatment that works by harnessing the body's immune system to recognize and attack cancer cells. Drugs used for this treatment are typically administered intravenously through infusion, although some may be given as subcutaneous injections. Imunotherpay is particularly promising for breast cancer treatment because it offers the potential for long-lasting responses and fewer side effects compared to traditional treatments like chemotherapy.

These therapies are often used in combination depending on the characteristics of the breast cancer and the patient's overall health. Indeed, this multidisciplary approach can help improve outcomes and reduce the risk of a cancer relapse.

2.1.5 BC studies endpoints

Several notions have been introduced to help evaluate the efficacy of a treatment. Among them we can distinguish:

- **Overall survival (OS)** is the length of time from either the date of diagnosis or the start of treatment, that patients diagnosed with the disease are still alive. In a clinical trial, measuring the overall survival is one of the most meaningful enpoints in oncology because it reflects the ultimate goal of cancer treatment. It provide a comprehensive measure on how well a new treatment works. Moreover, OS helps to identify patient subgroups with better or worse prognosis, thus help guiding treatments. However it requires long follow-up periods to obtain valuable survival data. The OS is assessed by measuring the following patients over time and recording the occcurrence and timing to deaths.

- **Disease free survival (DFS)** is length of time after primary treatment for a cancer that the patient survives without any signs or symptoms of that cancer and is still alive. In a clinical trial, measuring the DFS is also one way to see how well a new treatment works. It evaluates the duration of time patients that remain free of disease recurrence or progression following primary treatment. It is assessed by measuring following patients over time for any signs or symptoms of a cancer relapse (distant, local, regional, contralateral, or death).

- **Relapse free survival (RFS)** and **Distant Relapse Free Survival (DRFS)** represent the length of time after the primary treatment for a cancer ends that the patient survives without any signs or symptoms of, respectively, local, regional, or contralateral relapse or metastasis. It is assessed the same way as the DFS but without aknowledging the death event. The occurrence and timing of relapse events are recorded to calculate DFS (or DRFS) rates.

- **Event-Free Survival (EFS)** is similar to DFS but includes additional events such as a progression of the cancer, the development of a second primary cancer or discontinuation of treatment due to toxicity. EFS is assessed by recording the occurrence and timing of the corresponding events (local, regional, metastasis, progression) by following patients over time.

- **Pathological complete response (pCR)** is defined as disappearance of all invasive cancer in the breast after completion of a neoadjuvant setting, typically chemotherapy. Pathologists examine the tissue samples for residual cancer cells in the breast and lymph nodes using microscopy techniques. The absence of any residual invasive cancer

cells confirms the achievement of pCR. Other investigators have defined pCR as a complete response in the breast, irrespective of axillary nodal involvement as well [Buzdar *et al.* (2005)] [Bear *et al.* (2006)] [noa (2001)] [Al-Hilli & Boughey (2016)]. It has been used as an endpoint for several trials of neoadjuvant therapy for breast cancer.

These endpoints play a major role in guiding treatment decisions and clinical practice guidelines in BC management. The selection of appropriate endpoints will depend on the study aims, patient population, and treatment setting. Existing studies show that Luminal A and B subgroups have better OS and DFS when it decreased in TNBC and HER2+ subtypes. When patients with TNBC reach a pCR, the OS and DFS get significantly improved [Abrial *et al.* (2012)].

2.1.6 Risk factors of BC

A risk factor refers to anything that increase the likelihood of getting a disease. Various cancers are associated with distinct risk factors. Breast cancer risk factors include a range of genetic, hormonal, lifestyle and environmental factors. In fact, having certain genetic mutations, such as BRCA1 and BRCA2 genes, are widely known to elevate significantly the risk of deceloping breast cancer at a certain stage of a life, especially if there is a family history of breast or ovarian cancer. Hormonal factors, such as early onset of menstruations, late onset of menopause, or hormone replacement therapy can have an impact on developing BC. The other risks comprise chemical or radiation exposure, alcohol and tobacco consumption, obesity. While age and genetics can be beyond one's control, regular screening can play an important role in reducing the risk an improving the outcomes. It is important to note that possessing one or more of these risk factors does not guarantee the development of breast cancer. Many individuals with these factors never develop the disease, while others without any known risk factors may still be diagnosed with it.

2.2 Electronic Health Records

In this section, I will delve into the realm of Electronic Health Records (EHRs), a tool revolutionizing healthcare delivery.

2.2.1 EHR overview

Electronic Health Records (EHRs) are an electronic version of a traditional patient's paper chart. They includes an ensemble of patients health information throughout their medical journey. EHRs represent an digitalized approach to medical record-keeping and allows to collect, store, manage

and easily share patient information to authorized users. Unlike traditional paper charts, EHR systems are designed to consolidate a patient's medical history, including diagnoses, treatments, medications, lab results, etc. into a single digital platform. This centralization facilitates collaborations among healthcare providers and thus, reduces errors and enhances the quality of care. EHRs may also benefit to patients with a greater access to their health information, fostering a more informed relationship between patients and clinicians. Moreover, handwritten paper medical records may lack legibility in addition to a healthcare workflow inefficiency, i.e, managing paper records requires significant manual effort from filing and storing to retrieving and transporting. Consequently, as a result of these good characteristics, many developing countries have already adopted this approach to enhance their healthcare systems. However, the implementation of Electronic Health Records (EHRs) comes with its own set of challenges, particularly concerning privacy and data security. As mentioned in [Sahney & Sharma (2018)], it requires a coordinated effort from all the involved parties and given the potential risks associated with information technology, security should be a priority. Therefore, regulations and incentives have driven the adoption of EHRs to improve healthcare efficiency and interoperability.

2.2.2 EHR System components

An EHR system represents the integrated system digital platform that include various different components to create a comprehensive health record for patients. It comprises multiple key components that work together to form an useful technology for medical organizations to ramp up their facilities' productivity and efficiency not just clinically but economically as well. The major components of an EHR system comprises :

- **Patient Demographics**: It refers to the information that identifies and describes a patient. It includes name, address, date of birth, gender, marital status and all the contact details for a defined patient. This section is a critical component of EHR, as it serves as the basis for a patient identification. Furthermore, this information is essential to maintain accurate up-to-date records, coordinating care and ensuring that the healthcare services are delivered to the correct individual.

- **Medical History**: It includes the patient's medical information over time. It encompasses a thorough record of the past and previous health-related information, such as diagnosis, surgeries, allergies, and family medical history. This section is important for physicians as it offers insights into a patient's medical background, which facilitates accurate diagnosis, treatment planning, and coordinated care by considering all relevant factors in a patient's medical history.

- **Clinical Notes**: This section includes physicians' notes, and other clinical documentation from the healthcare providers during patient visits or hospital stays. Clinical notes in an EHR are crucial for care continuity and communication among healthcare providers. They allow different professional to stay informed about a patient's condition and treatment. Moreover, they provide a record for quality assurance, legal purposes, and billing, making accurate and exhaustive clinical notes an essential component of any EHR system.

- **Medication Management**: It tracks medication prescription, administration instructions, monitoring, and related information like side-effect reactions or drugs interactions, to ensure safe and effective use of prescriped medications. This component plays a important role in maintaining patient safety, optimizing therapeutic outcomes, and preventing medication errors.

- **Lab and Diagnostic Results**: It stores the patient's laboratory test results, radiology images and other diagnostic findings. This component help clinicians to determine the type and severity of a disease. It can help also to monitor conditions and treatments.

- **Decision Support Tools**: Theses tools may be part of an EHR system. It is a feature or system that helps healthcare professionals to make informed clinical decisions by offering relevant and evidence-based information, alerts and guidance. They are designed to improve the quality of care of care and support the overall efficiency of healthcare processes.

- **Patient Portals**: It is an online platform or website that allows patients to access various healthcare services and their own health information. They can schedule appointments, and communicate with healthcare providers, as well. This component is a key component of many EHR systems, that helps enhance the overall patient experience and healthcare outcomes by improving patient engagement, communication, and involvement in their own care.

- **Interoperability and Data Exchange**: This component refers to the capability of different healthcare systems, and organizations to communicate, share and use health information. Data exchange can occur in different ways, such as electronic health information exchanges (HIEs), direct messaging between systems, or through application programming interfaces (APIs).

- **Security and Privacy**: This last component is a fundamental component of healthcare, particularly when dealing with EHRs. They ensure that patient information is protected from unauthorized access,

theft or loss. In parallel, it needs to respect patients' rights to control their own personal health data. This encompasses the measures and protocols to protect patient data, ensuring compliance with privacy regulations and data security standards. It exists keys aspect of privacy in EHR, such as patient consent, information confidentiality, compliance with regulations (laws like the General Data Protection Regulation (GDPR) in the European Union set standards for protecting patient privacy and outline patients' rights regarding their health information), and patient rights to access.

One important actor in EHR systems is data managers. They play a crucial role to ensure that all parties have access to the records whenever need while protecting privacy. There are 2 main types of EHR systems, each one providing different levels of functionality depending on the capabilities required by a specific facility.

- **Physician hosted**: This is the EHR type where the data is hosted on servers within the physician's facility. The main advantage in this type of EHR system is the ability for physicians to directly access to records which can save time and increase efficiency by eliminating the need to communicate with other staff members. Moreover, they have the possibility to choose the optimal combination of hardware ans software to maximize functionality and interoperability. On the other hand, in this category, the medical institution requires additional IT support because it is responsible for the software maintenance and it have to make significant efforts to prevent data loss. However, it balances out with the fact that data storage fees won't have to be paid to external vendors.

- **Remotely-hosted**: Under these EHR systems' types, the server is outside the physician facility, i.e, they are not responsible for storing and managing the patients' data. The EHR stytem is stored on third-party servers owned by EHR vendors. The benefit of this EHR type is that the medical practitioner only focuses on collecting information and not about managing the IT system. This type of system allows a broader range of tools that can enhance clinicians' efficiency and thus provide a higher quality care. Unlike physician-hosted EHR systems, remoted-hosted ones require a smaller IT team due to a reduce number of technical error within the system, and additionally, the server owner handles all updates and security issues. Advantages include being more cost-effective and knowing where your data will be stored in the future. However, the downside is that this type of system may offer a limited flexibility for patients, as it has been designed by a data manager who controls information that can be seen. Remote EHR systems are generally hosted in the cloud.

2.2.3 Implementation and Adoption of EHR systems in modern healthcare

Over the past few decades, Electronic Health Records (EHR) have become increasingly popular in modern healthcare all around the world, due to their ability to facilitate patient information management, improve care coordination, and enhance data accessibility ([Oza *et al.* (2017)], [Heart *et al.* (2017)], [Liang *et al.* (2018)], [Oumer *et al.* (2021)], [Fraser *et al.* (2022)]). As depicted in [Woldemariam & Jimma (2023)], EHR is at the forefront of implementation in healthcare institutions to improve the quality of given care in high-income and low-income countries alike. This shift from paper-based records to digital systems has been driven by the need for more efficient healthcare processes and better patient outcomes. EHR systems allow healthcare providers to have instant access to patient medical information throughout their medical journey and allow them to consider the whole medical timeline to take informed decisions quickly, which can be crucial for emergency cases. Some studies have shown an improved efficiency following EHR implementation, in workflow ([Nguyen *et al.* (2014)], [Jha *et al.* (2009)], [McAlearney *et al.* (2010)]), through a time gaining in medical information retrieval ([Kossman (2006)], [Kossman & Scheidenhelm (2008)], [Zhang *et al.* (2012)], [Howard *et al.* (2013)], [Noblin *et al.* (2013a)]) and reduction in documentation time ([Poissant (2005)], [Skinner *et al.* (2011a)], [Chao (2016)]). Moreover, EHR systems have shown to enhance not only communication between providers, but physician-patient communication as well as the patient's medical information is centered using EHR systems ([Archer *et al.* (2011)], [Skinner *et al.* (2011a)], [Goldberg *et al.* (2012)], [Zhang *et al.* (2012)]). However, implementing an EHR system requires to consider several critical steps to ensure its effectiveness. It can be a tricky process, because it can directly have an impact on an institution medical practices' workflow and can possibly affect the clinicians' performance, which is not insignificant in the medical field. The implementation strategy will consist of:

- Define the specific needs and preferences of the institution for every department to be able to choose, or design the most suitable EHR system that will align with those needs and preferences.

- Select an EHR vendor if the organization wants a remoted-hoted system. Different EHR vendors can be evaluated based on cost, functionality, interoperability, security and other factors more or less important depending on the institution. Contracts are also an important aspect of the EHR implementation, because they should cover key aspects like data ownership, support, and compliance with regulations. If the institution prefers a local (physician-hosted) system, they defined specifications according to their requirements and select the competent

person to develop the EHR system in local.

- Customize and configure the workflow to fit the institution's needs and workflow. This step includes setting up user roles, access controls, data fields and other system parameters.
- Migrate data to the system from existing systems (paper records for example). This step needs to be done carefully to ensure accuracy and avoid data loss, but most importantly to maintain the privacy of the existing EHR in the healthcare institution.
- Train each user to be effective in using the system and to make the transition smoother [McAlearney *et al.* (2012)].
- Test the new system for a varying duration before the full-time use, to check good functionality of each unit of the system and their synergy, to identify and resolve potential technical issues, or security problems [Aguirre *et al.* (2019)]. This step allows to ensure the reliability of the new system in terms of EHR system quality standards.
- Use of the new EHR system in the healthcare setting and close monitoring by either the EHR vendors (in Remoted-hosted ssytems) or the IT team (in physician-hosted systems).

The transition to the adoption of EHR systems involves many challenges. In fact, implementing EHR systems require significant financial investments and technical resources, but most importantly, it needs to prioritize data security and privacy. Medical institutions invest in robust measures to protect patient data and to stay compliant with regulations. But still, there are clinicians and patients that expressed concerns about potential "information leakage" ([Jha *et al.* (2009)], [Archer *et al.* (2011)], [Yau *et al.* (2011)], [Priestman *et al.* (2018)]). On another hand, it exists reviews in the litterature that shows mixed observations on EHR quality, adoption and satisfaction. In [Lo *et al.* (2007)] the adoption of an EHR system did not significantly improved their average time spent in treating patients across all specialties. Also, when the training is not done properly, EHR implementation may result in clinicians spending more time to retrieve the correct information, and thus impact on their efficiency ([Skinner *et al.* (2011b)], [Sockolow *et al.* (2012)], [Noblin *et al.* (2013b)]). Moreover, system inherent problems (technical issues or slowness of the system) may be challenging barriers for a efficient EHR system utilization, and impact negatively on the workflow as well, as mentioned in ([Alsohime *et al.* (2019)], [Al-Rawajfah & Tubaishat (2019)], [Bruns *et al.* (2018)]). Other important aspects in EHR system evaluation, such as resources constraints (system updates/maintenance, limited access, limited network), or related to the lack of administrative/IT support, etc., can be potential barriers for EHR systems adoption ([Tsai

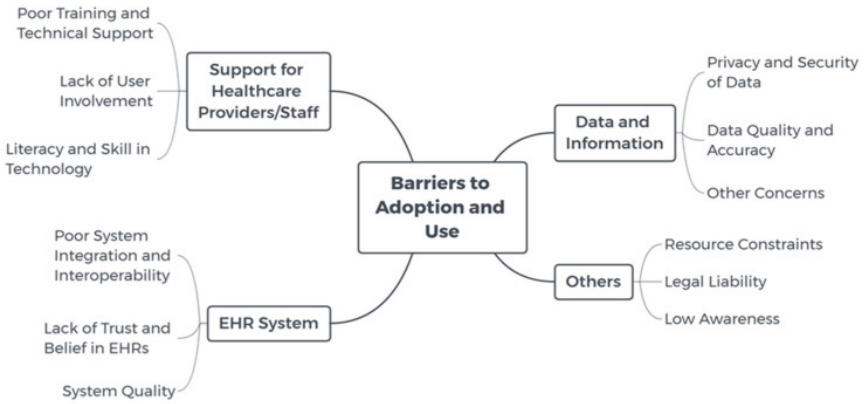


Figure 2.6: Mind map of barriers in the use of EHR systems. Source: [Tsai *et al.* (2020)]

et al. (2020)]), as shown in Figure 2.6. But ultimately, by addressing these challenges and following best practices for EHR implementation, healthcare facilities can exploit the full benefit of EHR.

2.3 Subject definition

In this section, I will describe the state-of-the-art in multimodal EHR within medical studies and how their use offers novels perspectives for comprehensive research and improved patient healthcare.

2.3.1 The use of multimodal EHR in medical studies

Clinical research involves developing knowledge that helps in identifying the best treatments and practices for a wide range of diseases and for different conditions. It has undergone significant evolution over the past years, from simple observations to modern large-scale healthcare studies involving large participant cohorts. Through the years, healthcare scientists continue to innovate, discovering news methods for screening, preventing, diagnosing, and treating diseases, as well as refining the manner of healthcare delivery. This process entails extensive research and methods conducted over years of dedicated effort, that can be considered as a series of decisive questions leading to the discovery of facts or information that have improved our understanding of health and human diseases' mechanisms. The term "clinical research" encompasses a broad spectrum of research questions and methodologies. From clinical trials by the pharmaceutical industry to develop and approve new drugs, to the monitoring of healthcare system by facilities to monitor the ongoing health conditions and the effectiveness of interventions over time and provide data for public health decisions, all the clinical re-

search areas relies on data from medical records which often derive from multiple modalities.

In recent years, clinical research has undergone a significant shift in its paradigm with the advent of personalized medicine. In clinical research, personalized medicine is an approach that tailors medical treatments based on unique characteristics of each patient. These individualities comprise genetic variations and biomarkers, lifestyle factors and environmental influences. They are used to identify predispositions to certain diseases, predict treatment responses and stratify patients into subgroups with different prognosis. Therefore, personalized medicine is expected to aid to further guide decisions-making in the prevention, diagnosis and treatment of diseases, in addition to other medical information. Electronic Health Records (EHR) have facilitated the development of personalized medicine to revolutionize clinical research. They provide comprehensive and real-time access to a patient's medical history, genetic information and treatment response. This helps enhance the precision of clinical trials, enables the identification of patient-specific treatment strategies, and improves the efficiency of data collection and analysis. Consequently, the convergence of personalized medicine and EHRs is paving the way for more effective and individualized healthcare solutions, ultimately transforming patient outcomes and advancing medical research.

Electronic Health Records represent a rich repository of patient data that includes various types of information, making them inherently *multimodal*. The multimodality refers to the presence of diverse data types or data formats (image, text, video, audio, genetics for instance), that can coexist for a same patient. These modalities can be divided into two main types of data: structured and unstructured. *Structured data* refers to information that is organized and easily searchable, such as patient descriptors and demographics, vital signs, laboratory results, and medication lists. This data is typically entered into predefined fields and can be efficiently used for statistical analysis and reporting. Much clinical research has shown that many features from structured information in EHRs are a major cause of preventable death around the world. Among them, the use of tobacco and alcohol has long been known to cause numerous diseases and complications, including cancer, heart disease, stroke, infections and pregnancy complications [Ezzati *et al.* (2002), Doll *et al.* (1994b), Saracci (1995), Doll *et al.* (1994a)]. Also, overweight has been associated to hormones levels alterations, and may induce cancer, stroke, cardiovascular disease and type II diabetes among others [Ezzati *et al.* (2002), Pati *et al.* (2023), Ma *et al.* (2021), Klein *et al.* (2022), Powell-Wiley *et al.* (2021)]. According to published studies, the incidence rate of breast cancer significantly varies based on race, ethnicity, and geographic location [Bray *et al.* (2024), Youlden *et al.* (2014)]. Moreover, patient and cancer information from structured EHR has allowed to develop widely used prognostic tools for breast cancer such

as the Notthigham Prognostic Index [Haybittle *et al.* (1982b)] or Predict-Breast [Wishart *et al.* (2010)].

On the other hand, *unstructured data* includes free-text notes and images. While this data type may seem richer in information and provide more context, it is more challenging to analyze due to its lack of predefined structure. Hence the wealth of unstructured data in EHR is often underutilized. Advances in Natural Language Processing (NLP) and computer vision are now shifting the trend, enabling the extraction and analysis of this type of data. Moreover, several studies have demonstrated the value of unstructured data in clinical research. The first clinical applications for medical reports or images include assisting healthcare professionals with retrospective studies and clinical decision making [Cheng *et al.* (2010), Do *et al.* (2013), Raciti *et al.* (2020), Eloy *et al.* (2023), Lin *et al.* (2013)]. Other papers have analyzed clinical texts in various languages and images to predict survival outcomes in cancer [Mazo *et al.* (2022), Harnoune *et al.* (2021)].

In clinical research, structured and unstructured data can be integrated for medical investigations. These multimodal EHRs enable a more complex view of a patient health and deeper insights into disease processed, treatment effectiveness and patient outcomes. Researchers can leverage multimodal EHRs to conduct various types of studies, including observational research, comparative effectiveness studies, and translational research [Zhang *et al.* (2020)]. For instance, in observational studies, researchers can analyze general patient descriptors alongside imaging data or gene expression data to track disease progression or treatment response over time [Yao *et al.* (2022a), Rabinovici-Cohen *et al.* (2020), Rabinovici-Cohen *et al.* (2022a)]. In comparative effectiveness studies, multimodal EHRs allow researchers to compare the outcomes of different treatment approaches based on a combination of clinical data and patient imaging data [Peeken *et al.* (2019)].

In conclusion, significant discoveries have been made in clinical research using EHRs. However, the use of multimodal EHRs has been limited until now, although it can offer substantial advantages by providing comprehensive health information and enhancing the accuracy and depth of medical studies. In breast cancer research, there are only few studies [Wang *et al.* (2020), Sun *et al.* (2019), Zeng *et al.* (2019a)] that have utilized multimodal EHRs to identify prognostic factors, especially those involving text data.

2.3.2 Challenges in multimodal EHR use in medical studies

While the litterature extensively portrays the potential of Electronic Health Records (EHR) to revolutionize various aspects of healthcare, it also highlights the challenges accompanying it. First of all, the use of multimodal EHR in medical studies raises significant privacy and security concerns. Patient health data is highly sensitive, necessitating regulations that govern its use for research purposes. To protect patient confidentiality and main-

tain data integrity, researchers must consider these security and privacy rules. Researchers must adhere to stringent privacy regulations, such as the General Data Protection Regulation (GDPR) in the European Union for data protection legislation. However, France has its own data protection law, the French Data Protection Act (the Act) , which govern the use and disclosure of protected health information in France. Robust security measures, including encryption, access controls, and audit trails, are implemented to safeguard EHR data from unauthorized access, breaches, or tampering [Fernández-Alemán *et al.* (2013)]. Additionally, researchers must obtain informed consent from participants, clearly outlining how their data will be used, shared, and protected. Collaborative research efforts involving multiple institutions necessitate careful data sharing agreements and protocols to maintain compliance with privacy regulations while facilitating data exchange. In practice, multiple papers agree on privacy concerns related to EHR systems, highlighting both technical and ethical limitations [Gariépy-Saper & Decarie (2021)]. Additionally, inconsistent implementation across facilities and countries poses significant obstacles to external collaborations.

Using multimodal EHR for medical studies can also be challenging due to the varying formats of data, which are usually structured differently and may have different semantics. The integration of diverse data types, such as text, images or tabular information such as lab results, can complicate data management and analysis [Zitnik *et al.* (2019a), Gligorijević & Pržulj (2015a)]. For instance, gene expression data can be structured as matrices, where each entry represents the expression level of a gene in a sample. Conversely, free-text reports are unstructured, and each word within these reports must be analyzed in the context of its surrounding words to gain deeper understanding of language patterns. Some data modalities present high dimensionality challenges as well, such as gene expression profiles or a history of medical reports. All these different challenges make difficult to define a strong integration method for a multimodal learning study.

Inconsistencies in data entry, arising from variations in terminology, measurement units, and documentation practices, can lead to fragmented or incomplete datasets. These issues hinder the ability to standardize and harmonize data, which is crucial for accurate analysis and meaningful comparisons in research. Additionally, the variability in data quality and completeness across different healthcare facilities and systems further complicates the use of multimodal EHRs, making it difficult to draw reliable conclusions and collaborate effectively on a global scale [Häyrinen *et al.* (2008), Jawhari *et al.* (2016)].

Chapter 3

Methodology

Abstract:

The use of multimodal learning in clinical studies involving multimodal electronic health records (EHRs) represents a notable advancement in scientific research. In my thesis, I use ML techniques to analyze such multimodal data in the context of breast cancer. The plurality of formats inherent to multimodal EHR requires using specific machine learning methodologies.

Despite the challenges associated with managing and integrating these heterogeneous data, these method offer considerable potential for improving predictive accuracy and personalizing therapeutic approaches. In this chapter, I give an overview of different machine learning methods applicable to both tabular and sequential EHR data. Subsequently, I will present the different families of approaches for integrating these diverse modalities, as well as techniques for interpreting multimodal learning models, which will be useful for better understanding model outputs.

Résumé:

L'utilisation de l'apprentissage multimodal dans les études cliniques impliquant des dossiers médicaux électroniques (DME) multimodaux représente une avancée notable dans la recherche scientifique. Dans le cadre de ma thèse, j'utilise des techniques d'apprentissage automatique pour analyser ces diverses données sur le cancer du sein. Les différents formats inhérents aux DME multimodaux permet de discerner multiples méthodologies d'apprentissage automatique applicables à l'analyse des DME. Malgré les complexités associées à la gestion et à l'intégration de données aussi diverses, cette approche offre un grand potentiel pour améliorer la précision des prédictions et adapter les stratégies thérapeutiques. Dans ce chapitre, je donne une vue d'ensemble des méthodes d'apprentissage automatique applicables aux données tabulaires et séquentielles des DME. Ensuite, je présente les différentes méthodes d'intégration de ces diverses modalités, ainsi que des techniques d'interprétation des modèles d'apprentissage multimodaux, qui seront utiles pour mieux comprendre les résultats des modèles.

Contents

3.1 Foundations of Machine Learning

Machine Learning is a field of Artificial Intelligence (AI) that aims to teach machines to learn from data and improve with experience. It can be summarized as learning a function f that applies to input variables X. The form of the function f is unknown and machine learning algorithms allow to approximate the underlying function. Different algorithms make different assumptions or biases about the form of the function and how it can be learned. In particular, we distinguish parametric models from non-parametric models.

A learning model that summarizes data with a set of parameters of fixed size (independent of the number of training instances) is called a parametric model. Regardless of the amount of data provided, a parametric model will not change the number of parameters it requires [Stuart (2015)]. Parametric algorithms are most appropriate for problems where the complexity of the model is controlled a priori.

Algorithms that do not make strong assumptions about the form of the mapping function are called non-parametric machine learning algorithms. By not making assumptions, they are free to learn any functional form from the training data. Consequently, they can accommodate a wide variety of functional forms. Most ML methods are parametric. In this thesis, I use both parametric and non-parametric machine learning models.

The two main families of problems addressed by ML are supervised and unsupervised learning. They differ in the nature of the learning process itself, and the condition of the training data that is required. Each approach has different strengths, so the task or problem faced by a supervised vs unsupervised learning model will usually be different. Unsupervised learning is a type of ML that learns from unlabeled data. Its goal is to analyze itself the underlying structure of the input data and to discover patterns and relationships without any explicit guidance. Given a set of N unlabeled examples ($\{x_i\}$) where $i = 1, ..., N$ and $x_i \in \mathbb{R}^n$, the unsupervised machine learning model will approximate a function f that will describe the best the inputs set X. Among the well-known unsupervised ML tasks, clustering, dimensionality reduction and anomaly detection are illustrated in 3.1

Supervised learning involves training model on labeled data to make predictions, adjusting itself to minimize error These datasets are labeled for context, providing the desired output values to enable a model to give a "correct" answer. Given a training set of N labeled samples ($\{x_i, y_i\}$), where $i = 1, ..., N$, and $x_i \in \mathbb{R}^n$ is an input vector and $y_i \in$ Y is the corresponding output, Y being the set of all possible outputs, the goal of a supervised learning algorithm is to approximate a function f that best maps the inputs to the outputs: $y_i \approx f(x_i)$. For a regression task, in which the model aims to predict a continuous value, the output space Y is $Y = \mathbb{R}$. For a classification task, where the goal of the model is to predict a label in a finite list of K categories, $Y = \{1, 2, ..., K\}$.

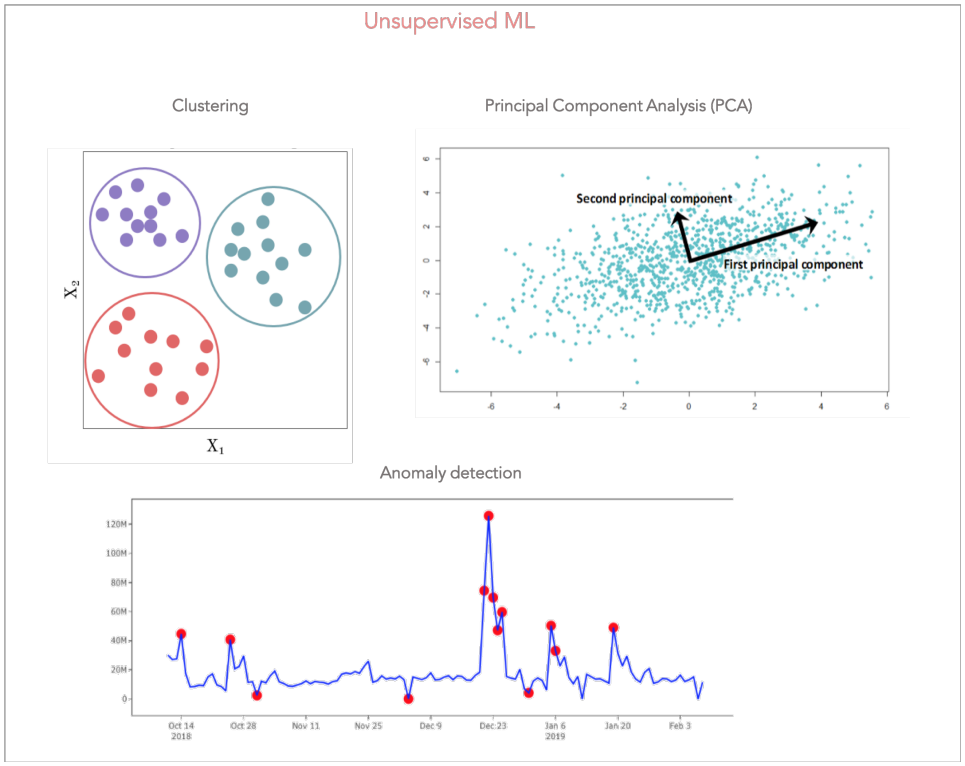


Figure 3.1: Unsupervised Machine Learning algorithms examples

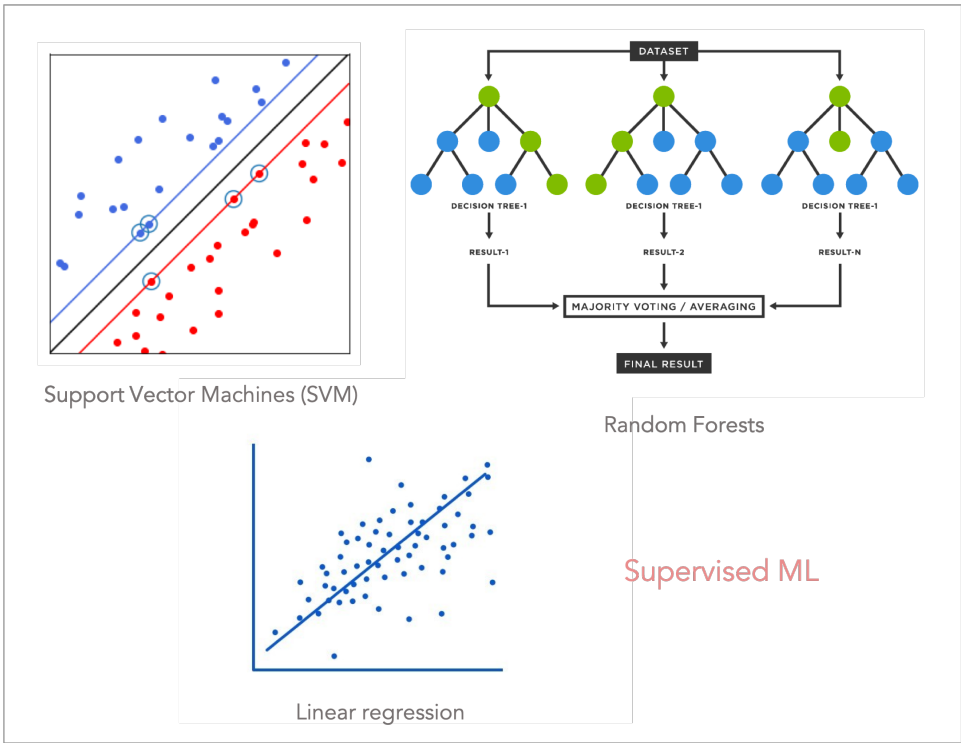


Figure 3.2: Supervised Machine Learning algorithms examples

Some common supervised learning algorithms include linear or logistic regression, random forests (regression or classification), support vector machines (whether for regression or classification) or Naive Bayes, represented in 3.2

In this thesis all developed models are supervised learning machine learning models. Indeed, the main task addressed by this thesis is the prediction of disease free survival (DFS) status using multiple samples of multimodal EHRs from the Institut Curie, framing it as a classification task. Models used for this task will be detailed in the next sections.

3.2 Classical Machine Learning models for tabular data

This section will delve into the main ML methods used for tabular datasets in this thesis. We will introduce three (3) supervised machine learning algorithms that can be used both for classification and for regression, but for this discussion we will only focus on classification task, as this is how we formulated the problem of DFS prediction. For the prediction of DFS status using multimodal electronic health records, these models will serve as the

first steps to have a glimpse of the data distribution and the complexity of the task, before to delve into more sophisticated solutions.

3.2.1 Random Forest Classifier

Random forests are an ensemble learning method widely used in machine learning tasks. RF models were first proposed by Salzberg and Heath in 1993 [Heath (1993)], then developed further by Ho in 1995 [Ho (1995)]. The current version of random forests was introduced by Breiman in 2001 [Breiman (2001)]. They are defined as a combination of tree predictors such as each tree depends on the values of a random vector sampled independently from the data and with the same distribution for all trees in the forests [Breiman (2001)]. Decision trees are trained in parallel and each tree casts a unit vote for the most likely class for input X. The random forests output is the mode of the classes among all the individual trees composed by the random forest classifier. To gain a better understanding of random forests, it is essential to first comprehend its basic unit: the decision tree.

Decision Trees Decision trees are simple and work by partitioning the input space into cuboid regions, whose edges are aligned with the axes, and then assigning a simple model (for example, a constant) to each region [Bishop (2006)]. There are, as trees, composed of nodes that represent features and branches that represent the answer to a question "asked" on nodes. Each leaf node represent an outcome, as depicted in figure 3.3. In this example, the root node is the starting point of a decision tree and represents the entire input space. It is then divided into regions according to the first given condition: whether $X_1 > 0.4$ or $X_1 \leq 0.4$. The input space is then divided into two regions depending on the condition and each of these regions will be subdivided according to splitting rules. This process is repeated on each derived sub-regions in a recursive manner. The recursion is completed when the leaf nodes, after which, no further splitting helps for a better performance in the classification task, is reached.

How does a random forest classifier work? During training, random forests employs random feature selection to ensure that each tree in the forests brings a unique perspective of the data and that the trees operate indepedently from each other. A random subset of features is selected to split the first node. Given a dataset D with N samples, the model with first generate M bootstrap samples $\{D_1, D_2, ..., D_M\}$ by sampling with replacement. For each random bootstrap sample D_i, a decision tree is built by randomly selecting a subset of features at each split and splitting the node based on the chosen feature subset either to optimize a criterion. Common choices of criterion are *entropy* or *Gini impurity*, which must be minimized, and *information gain*, which must be maximized.

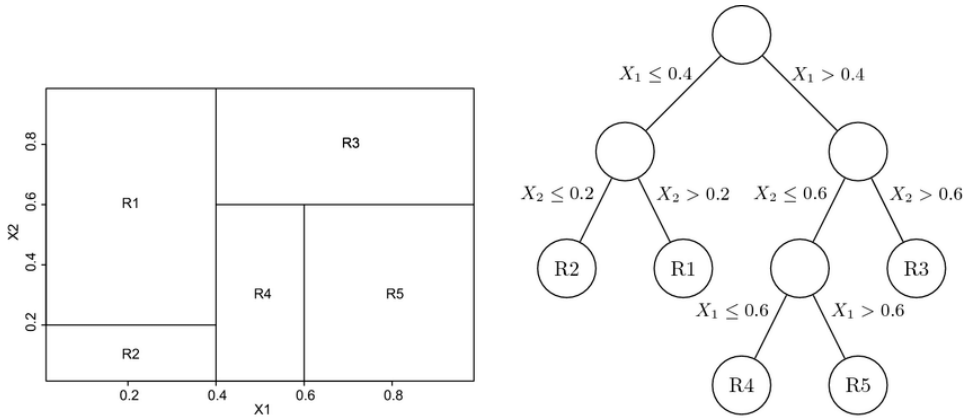


Figure 3.3: Example of decision tree partition of the predictor space (left) and the corresponding decision tree (right)

Entropy The entropy is a measure of information that indicates the disorder or randomness of the features with the target. In order to determine the right splitting criteria for each node, the entropy is computed for each feature and its potential splits. Its value varies between 0 and 1 and the optimum split is given by the lower entropy. Leaf nodes which have all instances belonging one class would have a entropy equal to 0. The entropy is calculated using the following formula:

$$\text{Entropy} = -\sum_{j} p_j . log_2(p_j) \tag{3.1}$$

where the p_j represents the probability that a randomly selected example belongs to class j. In the context of the example shown in figure 3.3, the node entropy is given by the probability of belonging to either of two classes. Let us assume that in a fictive dataset of 10 samples, with 7 samples belonging to class 0 and 3 samples belonging to class 1, then the entropy at the root node will be calculated using $p_0 = 0.7$ and $p_1 = 0.3$.

Information gain Information gain is the change in entropy from children nodes to the parent node. It measures the reduction in entropy throughout nodes and provides the amount of information a feature has in relation to the output. The feature that has a minimum impurity will be considered as the root node. It is further used to decide which feature to split at each step in building the tree. The more homogeneous the child node is, the more the variance will be decreased after each split. Thus Information Gain is the variance reduction and can be calculated as by how much the variance decreases after each split. Information gain of a parent node can be calculated as the entropy of the parent node subtracted entropy of the weighted average of the child node.

$$\text{Information Gain} = E_{parent} - Avg\,(E_{children}) \tag{3.2}$$

Gini impurity The Gini impurity, or Gini index is another way of splitting a decision tree. The Gini index measures the probability for a random instance at the node to be misclassifed when randomly labeled according to the distribution of labels at the node. The lower the Gini index, the lower the likelihood of getting a misclassification. Its values vary from 0 (highest level of purity) to 0.5 (random classes assignment). In practice, the Gini index favors larges partitions and performs only binary splits.

$$\text{Gini} = 1 - \sum_j p_j^2 \tag{3.3}$$

Pruning After completion of decision trees' building, the model may tend to overfit data due to noise or outliers present in datasets. The pruning process is used to remove redundant or unnecessary nodes in trees. During that step, a whole sub-tree is replaced by a leaf node, when it is established that the corresponding decision rule leads to a greater error rate than in a single leaf. Pruning can be done prior to the completion of the full tree (pre-pruning), or after the tree is finished (post-pruning).

The entropy and the Gini index are better measures than the misclassification rate for growing the tree because they are more sensitive to the node probabilities. Also, unlike misclassification rate, they are differentiable and hence better suited to gradient based optimization methods. For subsequent pruning of the tree, the misclassification rate is generally used.

Advantages and disadvantages By combining multiple predictions from single decision tree, random forests tend to have high performances [Grinsztajn *et al.* (2024)], and mitigate the risk of overfitting, which is common with individual classifiers. In fact, with a sufficient number of decision trees in a random forest, the averaging of uncorrelated trees lowers the overall variance and prediction error. Thus, the ensemble model learns less noise and leads to a model with better generalization ability. In addition, by training each tree on a different subset of data and features, random forest classifiers capture a more comprehensive view of the data distribution and allow less biased prediction. They are also more efficient at handling high-dimensional data, because only a subset of features is considered during the construction of each decision tree.

In addition, random forests are intuitive models. For "experts" and "non experts", they provide a measure of feature importance, which indicates the contribution of each feature in the prediction. This can be calculated with the decrease in node impurity (Gini impurity for example) attributed to

Algorithm 1 Random Forest classifier algorithm

Training Phase:

Given

- D: training set with N instances $\{(x_i, y_i)\}_{i=1}^{N}$, where x_i is a feature vector of p features and y_i is the label
- K: number of classes in target variable
- M: number of decision trees classifier in the RF classifier

Procedure:

For each tree $m = 1, 2, .., M$
1. Generate bootstrapped samples D_M from the training set D.
2. Grow a tree T_m using a random feature subset from bootstrapped samples D_M. For a given node in the tree:
(i) Randomly select $m \approx \sqrt{p}$ features from the total p features.
(ii) Find the best split features and cutpoints using the random feature subset based on a splitting criterion (Gini impurity or Entropy).
(iii) Split the node into two child nodes based on the best selected features and cutpoints.
Repeat (i) - (iii) until stopping rules are met.
3. Construct trained classifiers C_b of M decision trees $\{T_m\}_{m=1}^{M}$ by repeating steps 1. and 2.

Testing Phase:

Aggregate the M trained classifiers using a majority vote. For a test instance x, the predicted label from classifiers C_M is given as:

$$C_M(x) = \text{mode}\{T_m(x) : m = 1, 2, ..., M\} \tag{3.4}$$

each feature across all trees in the forest (mean decrease in impurity, MDI method) or using the permutation importance measure also known as the mean decrease in accuracy (MDA) method, which identifies the average decrease in acccuracy by randomly permutating the feature values in samples [Breiman *et al.* (2017)]. This interpretation method is widely used in other machine learning models and this will be discussed in detail in Section 3.7. This "white box" particularity is valuable to understand the data, the predictions and for features selection for further models.

Random forests can also handle missing data by using the similarity of data points to fill in missing values. For instance, missing points can be imputed by averaging the values of the k-nearest neighbors in the feature space. Additionally, they can maintain high performances even with missing values.

However, the ensemble nature of random forest model results in a high computational cost and memory-intensive usage when building multiple trees. As shown in Figure 1, the training process involves repeated splitting and calculation of splitting criteria, which is time-consuming and resource-intensive. This can also lead to slower predictions, i.e, a prediction using a large number of trees requires passing the input through all trees in the model and then make the predictions slower than in a single model. In real-world applications, that might require a high number of trees to acquire higher accuracy, run-time performance may be favored and other approaches would be preferred.

Another important aspect in random forests being the feature importance, we must be cautious with datasets that contain features with more levels or higher cardinality, and those with many unique values. These can lead to bias in feature importance scores generation and to misleading interpretations [Louppe (2014)].

In conclusion, random forests offer a powerful and flexible approach to classification tasks with tabular datasets.

3.2.2 Logistic regression

Logistic regression (LR) is a statistical method used to model the probability of a binary outcome. It was first developed by Pierre François Verhulst between 1838 and 1847 [Verhulst (1845)], initially as a model of *population growth* and named "logistic". In machine learning, it is used to predict the probability of a categorical dependent variable, which is a binary variable that contains data typically coded as 0 ar 1. In other terms, the LR predict $\mathbb{P}(Y = 1)$ as a function of X and parameters θ used to estimate the logistic model.

$$h_\theta(x) = \mathbb{P}(Y = 1|x; \theta) \tag{3.5}$$

Sigmoid function Logistic regression is an extension of linear regression for classification tasks. Linear regression models the relationship between continuous dependent variables (inputs) and one or many independent labels (predictor). It aims to find the best fitting linear function between input X and the label Y as: $Y = \theta_0 + \theta_1 X + \varepsilon$, with θ being coefficients or parameters and ε the error term that represents the deviation of the true value from the predicted values. While linear regression predict continuous outcomes with a linear function, logistic regression use the logistic (sigmoid) function to transform a linear combination of input features X into a probability value ranging between 0 and 1. The sigmoid function, represented by an S-shaped curve (Figure 3.4), is the essential mechanism of logistic regression model and effectively maps any values from the input vector to a probability within the 0 to 1 interval. This model is commonly used in real-world binary classification problems. For instance, LR is suitable for DFS status classification using tabular EHR data. The logistic function is given by:

$$h_\theta(x) = \mathbb{P}\left(Y = 1 | x; \theta\right) = f\left(\theta^\top x\right), \text{with:} \tag{3.6}$$

$$f\left(\theta^\top x\right) = \frac{1}{1 + e^{-\theta^\top x}} \tag{3.7}$$

θ being the parameters vector that is determined by minimizing a cost function. The cost function estimates the likelihood of observing the given outcomes in the dataset:

$$\text{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) \text{ if y=1} \\ -\log(1 - h_\theta(x)) \text{ if y=0} \end{cases} \tag{3.8}$$

In Equation 3.8, when the actual target is 1, the model's prediction $h_\theta(x)$ should be close to 1. In that case, the cost function should increase the penalty as $h_\theta(x)$ goes farther away from 1 and towards 0 and the opposite as the prediction is close to 1. This function is given by $-\log(h_\theta(x))$. Similarly, when the label is 0, $h_\theta(x)$ have to be as close as possible to 0. Therefore, the cost function should lower the penalty for values closer to 0 and higher penalty for values farther from 0 and towards 1. The appropriate function is given by $-\log(1 - h_\theta(x))$. These functions are described in figure 3.5.

During training, the model estimates the parameters that best fit the data by optimizing the cost function. This is typically done using iterative optimization algorithms such as gradient descent.

Gradient Descent In machine learning, gradient descent is used to optimize algorithm during the training phase. The idea is to find the values of a function's parameters (the sigmoid function in LR), that minimize a cost function as much as possible. Gradient descent is an iterative algorithm,

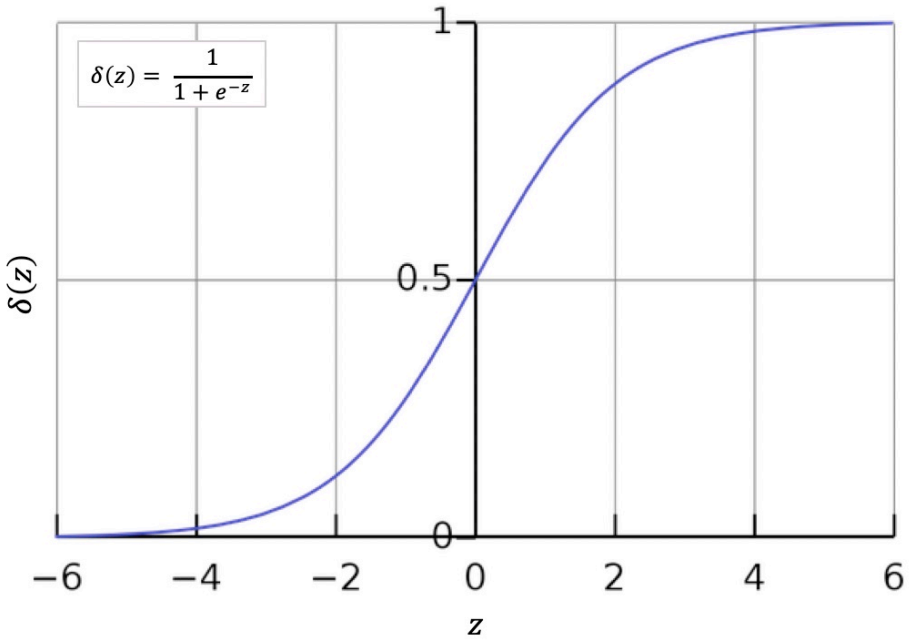


Figure 3.4: Sigmoid function represented here as $\sigma(z)$, where $z = \theta^\top x$. It maps any real-values number into a value between 0 and 1.

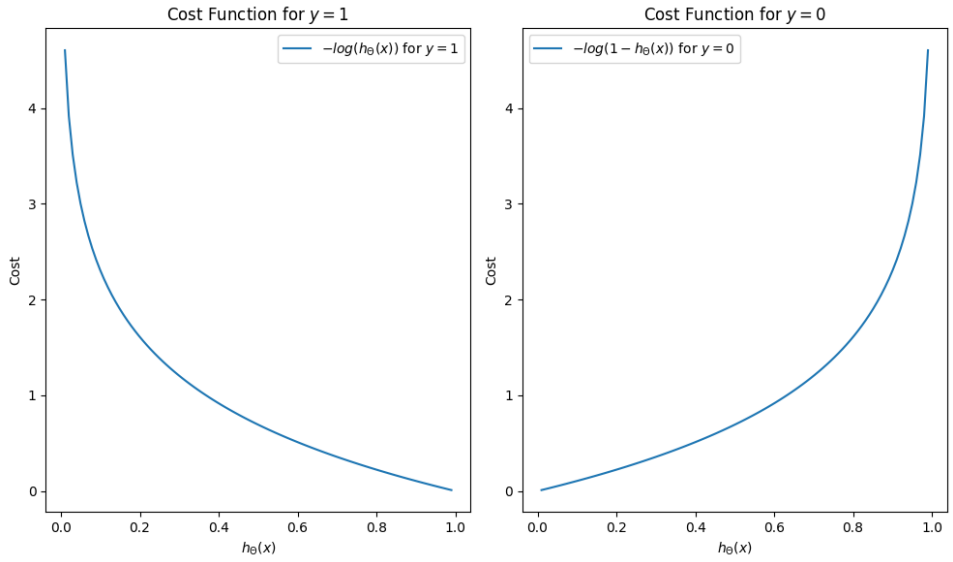


Figure 3.5: Cost function for logistic regression.

that test different values of parameters and update them to reach the optimal ones. I refer the reader to Section 3.3.3 of this chapter that detailed the gradient descent algorithm.

Algorithm 2 Logistic regression algorithm

Training Phase:

Given

- D: training set with N instances $\{(x_i, y_i)\}_{i=1}^{N}$, where x_i is a feature vector of p features and y_i is the label
- α: Learning rate
- max_iter: Number of iterations

Procedure:
Define the cost function (example of the binary cross entropy loss): $J(\mathbf{w}, b) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log(\hat{y}^{(i)}) + (1-y^{(i)})\log(1-\hat{y}^{(i)})\right]$,
where $\hat{y}^{(i)} = \sigma(\mathbf{w}\cdot\mathbf{x}^{(i)} + b)$ is the predicted probability.
repeat
  for all training samples $\{(x_i, y_i)\}$ **do**
    Compute prediction: $\hat{y}^{(i)} = \sigma(\mathbf{w}\cdot\mathbf{x}^{(i)} + b)$
    Update the weights $\mathbf{w}$ as:
    $\mathbf{w} := \mathbf{w} - \alpha\frac{1}{m}\sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})\mathbf{x}^{(i)}$
    Update the bias b as:
    $b := b - \alpha\frac{1}{m}\sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})$
  end for
until convergence or the maximum number of iterations is reached.

Testing Phase:

For each test sample $x_{\text{test}}^{(i)}$:

- Compute the predicted probability: $\hat{y}_{\text{test}}^{(i)} = \sigma(\mathbf{w}\cdot\mathbf{x}_{\text{test}}^{(i)} + b)$
- Classify $x_{\text{test}}^{(i)}$ based on a threshold (0.5 for instance):
- $\hat{y}_{\text{test}}^{(i)} = \begin{cases} 1 & \text{if } \hat{y}_{\text{test}}^{(i)} \geq 0.5 \\ 0 & \text{if } \hat{y}_{\text{test}}^{(i)} < 0.5 \end{cases}$

Regularization techniques Overfitting is a common problem in all machine learning models. In order to avoid overfitting, we use additionnal techniques that include regularization techniques. Those strategies consist in adding a penalty term to the loss function to control the model complexity

during the fitting process. Common methods include the L1 regularization or Least Absolute Shrinkage and Selection Operator (LASSO) regression, the L2 regularization or Ridge regression and the Elastic Net regularization.

- The **Lasso** regularization adds the sum of the absolute values of magnitudes of the coefficients as a penalty term to the loss function.

$$J(\theta) + \lambda \sum_i |\theta_i| \tag{3.9}$$

  where the $J(\theta)$ is the original loss function and λ is the regularization parameter that controls the strength of the penalty.

- The **Ridge** regularization adds the squared magnitude of the coefficients as a penalty term to the loss function.

$$J(\theta) + \lambda \sum_i \theta_i^2 \tag{3.10}$$

- The **Elastic Net** linearly combines both L1 and L2 penalties.

$$J(\theta) + \lambda_1 \sum_i |\theta_i| + \lambda_2 \sum_i \theta_i^2 \tag{3.11}$$

The main difference between the LASSO and the Ridge regularization methods is that the LASSO shrinks the less important features' coefficients to zero, and thus remove them from the prediction. LASSO can be used in feature selection for some tasks in case we deal with high dimensional data. The Ridge regularization creates models with smaller coefficients for those features and this results to a model that considers all features but with reduced importance. The Elastic Net regularization provide a balance between the sparsity of the L1 and the smoothness of the L2 regularization.

Advantages and disadvantages LR is an easy algorithm to implement that works well for linearly separable data, and is computationally efficient. Compared to ensemble models such as random forests described in Section 3.2.1, LR requires less processing and memory. It makes it more appropriate for large scale datasets and real-worlds applications. LR is interpretable as well, as its predicted coefficients give inference about the importance of each feature and make it simple to understand the relationship between predictor variables and the probability of the outcome. LR is less prone to overfitting compared to random forests classifiers, especially when dealing with high-dimensional data. However, to help prevent overfitting, previously presented regularization techniques can be used during the training process. Starting from the linearity assumption of LR, this algorithm is not suitable for data

that are not linearly separable. When LR assumes a linear relationship between input variables and label, this may not always be true, specially in real world applications. Therefore, this limits its performance when data truly follows more complex relationships. In those cases, we may need to manually transform or combine features to better fit the linearity assumption. This feature engineering requirement can be laborious, necessitate domain knowledge and add an additional step to using LR. Moreover, unlike RF, LR does not handle missing values inherently, preprocessing steps that are involved in handling missing data are necessary.

In this part, we presented the logistic regression model as a simple and efficient algorithm specifically designed for binary classification problems.

3.2.3 Support Vector Machine

Support Vector Machines (SVMs) are one of the most commonly used machine learning algorithms for both linear and non-linear classification problems. They have been first developed at the AT&T Bell Laboratories by Vladimir Vapnik and colleagues between 1992 and 1996 [Boser *et al.* (1992), Cortes & Vapnik (1995), Vapnik (1997)]. Their theory is based on learning non-linear classifiers by using a method called the *kernel trick*. This kernel method aims to find the optimal hyperplane that best separates data points of different classes in a high-dimensional space.

Kernel trick So far, we have discussed on methods used for in ML that mostly treat classification problems as linear problems. In SVM algorithm, the key concept consists in using linear classifiers to solve non-linear problems using a method called the kernel trick. The kernel trick is a powerful technique that enables SVMs to find a linear decision boundary in a transformed feature space without explicitly computing the transformation. In fact, many real-world problems are not linearly separable in their original feature space. To make them separable, we can transform the data into a higher-dimensional space where a linear separation is possible. This transformation is achieved through a function known as a kernel function.

For all x_i and x_j in a original input space χ, we can expressed a function $k(x_i, x_j)$ in another space ν. The function $k : \chi \times \chi \to \mathbb{R}$ is a kernel function that calculates the dot product of the transformed data points in a higher-dimensional space. The kernel trick compute the inner product of data points without explicitly computing their coordinates in that space, which can be computationally expensive. Depending on the problem, there will be different kernel functions that will be best suitable for the case. Among them, we can count:

- Linear Kernel used when data is already linearly separable:

$$k(x_i, x_j) = x_i \cdot x_j$$

- Polynomial Kernel for non-linear problems:

$$k(x_i, x_j) = (x_i \cdot x_j + c)^d$$

- Radial Basis Function (RBF) or Gaussian Kernel used for more complex and non-linear relationships by considering the distance between data points:

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

- Sigmoid Kernel, similar to the sigmoid function described in Section 3.2.2:

$$k(x_i, x_j) = \tanh(\alpha x_i \cdot x_j + c)$$

Once data points are separated into a higher dimensional space, it is possible to have a linear decision boundary that separates different classes in the feature space: the *hyperplane*. As shown in Figure 3.6, the hyperplane is a separation line in higher dimension. Support vectors are the data points closest to the hyperplane. These points are important as they determine the position and orientation of the hyperplane. SVM relies on these support vectors to maximize the *margin* between classes (cf Figure 3.6)

Advantages and Disandvantages SVMs are efficient for classification problems especially in high dimensional spaces, i.e, when the number of dimension exceeds the number of samples. The multiple available choices of kernel functions make it easier to tailor the model to different data types and problems complexities, including non-linear relationships in data, unlike certain other models, such as logistic regression for instance. But, on the other hand, this ability to use non-linear kernels make them less interpretable than simpler models like logisitic regression. Regarding memory utilization, SVMs reduce computational usage by avoiding explicit transformation of data points into higher dimensional spaces. Moreover, it is memory efficient for predictions, particularly with large datasets and high-dimensional spaces, as it uses a subset of training points (the support vectors) in the decision function. Another important aspect of SVM algorithm is that it requires careful paramaters tuning to make it effective.
In summary, SVM is a robust and versatile algorithm for classification tasks, capable of handling both linear and non-linear boundaries through kernel tricks. This method make it a powerful tool for complex data analysis.

3.3 Deep learning for tabular data

The Feed-Forward Neural Network (FFNN) are fundamental models in ML and provide a basis for more complex architecture. It is the basic type of

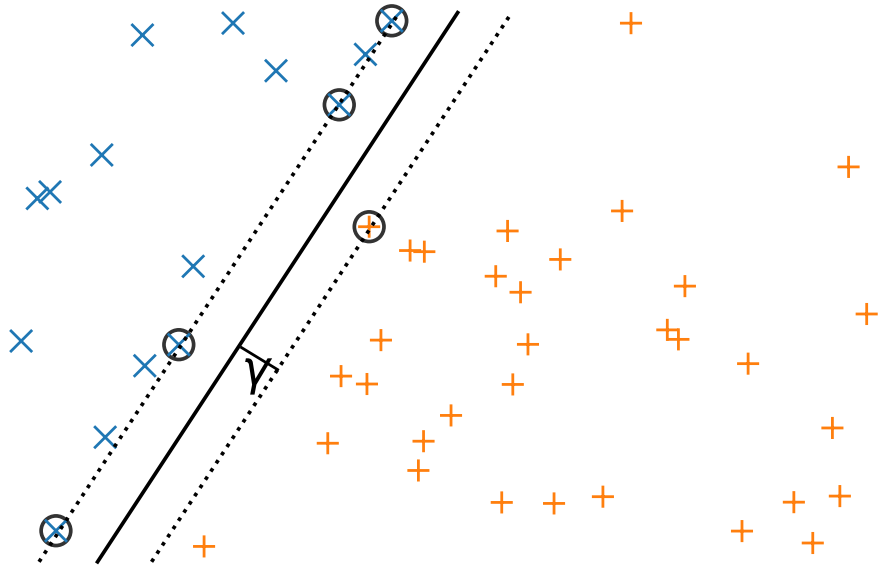


Figure 3.6: The separating hyperplane (in bold line) separates the positive and the negative samples. The margin γ is the distance betxeen the hyperplane and the nearest data points of each class. During training, the algorithm identifies the optimal hyperplane that separate the best the different classes by maximizing the margin γ. The model learn a linear decision boundary in the transformed space, which corresponds to a non-linear boundary in the original space. The circled points are the support vectors.

Algorithm 3 Support Vector Machine algorithm

Training Phase:

Given

- D: training set with N instances $\{(x_i, y_i)\}_{i=1}^{N}$, where x_i is a feature vector of p features and y_i is the label

Procedure:
Define the kernel function $k(x_i, x_j)$
Define the decision function $f(x) = \sum_i \alpha_i y_i k(x_i, x) + b$ with α being the Lagrange multiplier.
repeat
  for all $\{(x_i, y_i)\}$, $\{(x_j, y_j)\}$ **do**
    Set the number of changed α to zero.
    Compute the Error E_i and E_j with $E_{i,j} = f(x_{i,j}) - y_{i,j}$
    Optimize* values for α_i and α_j
  end for
until the number of changed α is zero or the maximum number of iterations is reached

Testing Phase:

For each test sample $x_{\text{test}}^{(i)}$:

- Compute the decision function : $f(x_{\text{test}}^{(i)}) = \sum_i \alpha_i y_i k(x_i, x_{\text{test}}^{(i)}) + b$
- Classify x based on the sign of $f(x_{\text{test}}^{(i)})$:
  - If $f(x_{\text{test}}^{(i)}) \geq x$, classify $x_{\text{test}}^{(i)}$ as a positive class
  - If $f(x_{\text{test}}^{(i)}) < x$, classify $x_{\text{test}}^{(i)}$ as a negative class

artificial neural networks (ANN) and it is characterized by the direction of the information flow throughout the layers. The first such network had a single layer and was proposed by Frank Rosenblatt in 1958 [Rosenblatt (1958)].

3.3.1 Perceptron

Description

The perceptron [Rosenblatt (1958)] is the simplest architecture for Feed-Forward Neural Networks (FFNNs).

A perceptron is initially a mathematical representation of a biological neuron. By analogy, the signal received by the dendrites of an actual neuron is represented by the input x_j. The electrical signal received by the synapses and modulated in different amounts is modeled in the perceptron by the weighted sum of the input features plus a bias term as shown in Equation 3.12. Its architecture is illustrated in Figure 3.3.1.

The perceptron is a binary classifier that maps input features $x = (x_1, ..., x_n)$ to a binary output $y \in \{0, 1\}$ using:

$$y(x) = g\left(\omega^\top x + b\right) \tag{3.12}$$

where ω represent the weights associated with each input x_j, that indicate its importance towards making the prediction and b is the bias, an additional parameter that helps the model g to adjust the output independently of the input features. The output of a perceptron is determined by applying an activation function g. The original perceptron [Rosenblatt (1958)] used the step function as its activation function. It is defined as:

$$\text{activation}(z) = g(z) = \begin{cases} 1 & \text{if } z > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.13}$$

where $z = \omega^\top x + b$ is the weighted sum of inputs plus the bias. This function produces a binary output. Since then, other activation functions have emerged in more complex neural networks. They will be discussed in detail in Section 3.3.2.

In what follows, we denote by θ the model parameters: $\theta = (b, \omega)$. During training, the perceptron adjusts θ to minimize the errors in predictions.

Given its formulation, a perceptron can only accurately model linearly separable problems with binary outcomes. But most of the times, the pattern observed in real world data points cannot be separated by a line (or, in higher dimensions, hyperplane). This limitation has laid the foundations for the development of more sophisticated models, such as multi-layers perceptrons (MLP) and deep neural networks.

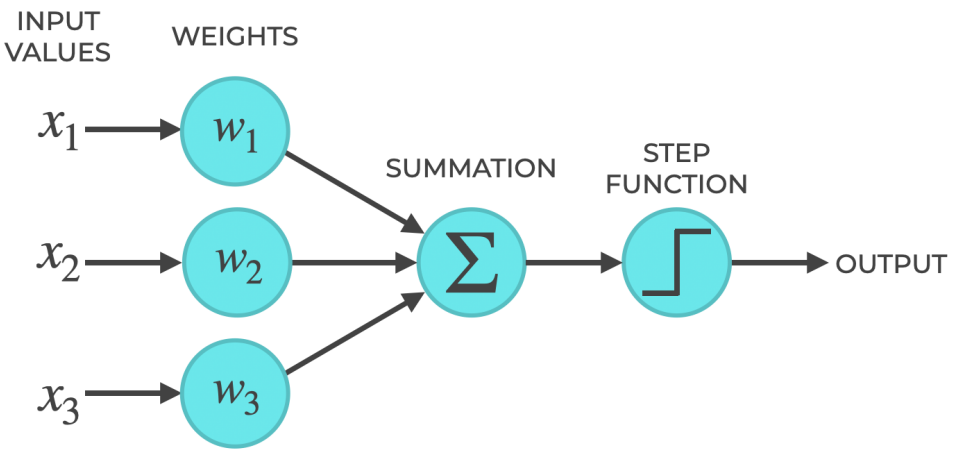


Figure 3.7: The structure of a perceptron. It is a single layer network with four parameters: input values, weights and bias, summation and activation function.

Algorithm 4 Rosenblatt perceptron learning algorithm

Training Phase:

Given

- D: training set with N instances $\{(x_i, y_i)\}_{i=1}^{N}$, where x_i is a feature vector and y_i is the label
- $\omega \leftarrow 0$
- $b \leftarrow 0$

Procedure:

while not converged **do**
  Compare the true label and the prediction:
  $\text{error}_i = y_i - g(\omega^\top x_i + b)$
  if $\text{error}_i \neq 0$ **then**
    Update the weights and the bias as:
    $\omega \leftarrow \omega + \text{error}_i \times x_i$
    $b \leftarrow b + \text{error}_i$
  end if
end while

3.3.2 Multi-layers Perceptrons

Multi-layers percerptrons (MLPs) are an extension of Rosenblatt single layer perception. By stacking multiple layers of perceptron and by using more advanced activation functions than the step function, MLPs can solve non-linear and more complex real world problems (see Figure 3.9 for the MLP architecture). A few commonly used activation functions are described below.

The **Sigmoid Function**, as given by Equation 3.7, is the same function used in logistic regression and detailed in Section 3.2.2. In logistic regression, the function maps the linear combination of input features to a value between 0 and 1. But within the framework of multiple layers perceptrons, it functions as a non-linear activation function that produces a smooth output between 0 and 1 (see Figure 3.8). Unlike the step activation, it is differentiable, which enables the use of gradient-based optimization (see Section 3.3.3. However, because the derivatives of the sigmoid function are small for a very high or a very low values of z, its use can cause what's called gradient vanishing, which is discussed in Section 3.3.3.

The **Tanh function** or **hyperbolic tangent** is also a non-linear function defined by:

$$\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{3.14}$$

Unlike the sigmoid function, the output of the tanh function ranges from -1 to 1, i.e, it is zero-centered (see Figure 3.8). This property helps in centering the data and makes optimization easier, yielding gradients that are more balanced compared to the sigmoid function.

The **Rectified Linear Unit (ReLU)** function [Fukushima (1969)] has become the default activation function for many neural networks because of its simplicity and its properties. It is defined as:

$$\text{ReLU}(z) = \max(0,\, z) \tag{3.15}$$

It ranges from 0 to $+\infty$ for positive inputs and is exactly 0 for negative inputs (see Figure 3.8). ReLU promotes sparsity in the network by setting all negative inputs to 0. This can lead to a more efficient representation of the data, reduce the likelihood of overfitting and enable computational efficiency. Nevertheless, one of its main drawback is linked to this sparsity: when neurons become inactive, that is to say, output zero for all inputs; this phenomenon is called "dying ReLU". It happens when weights are updated in a way to produce only negative inputs for a ReLU unit, which leads to a general gradient of zero, which prevents further learning. To address this limitations, several variants of ReLU have been developped: Leaky ReLU [Maas (2013)], Parametric ReLU (PReLU) [He *et al.* (2015)] or Exponential Linear Unit (ELU) [Clevert *et al.* (2015)].

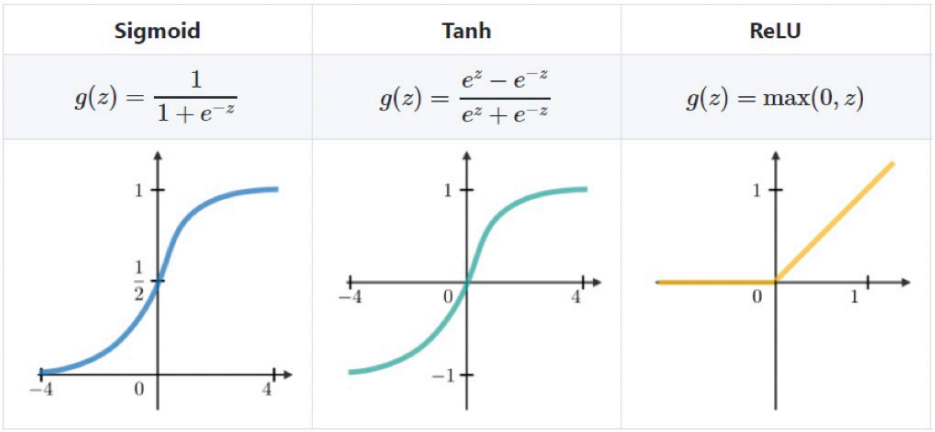


Figure 3.8: Some of the most common activation functions: the logistic (sigmoid) function, the hyperbolic tangent and the Rectified Linear Unit (ReLU)

3.3.3 Feed-Forward Neural Network

As we described above, the MLP extends the basic perceptron by adding one or more hidden layers between the input and the output layers, and by using more complex activation functions. MLPs are specific types of feed-forward neural networks (FFNNs). FFNNs are merely multi-layer neural network in which information flows forward, from the input nodes to the output nodes, while MLPs are *fully connected*, meaning that each neuron of one layer is connected to all neurons of the layer before it and of the layer after it. The first MLP model was released in 1967 [Ivakhnenko & Lapa (1967)].

How do they work?

Training We recall that the aim of training a neural network is to find the optimal set of parameters $\theta = (b, \omega)$ that minimizes an error. This process involves finding the best approximation of a function f^* into the parametric space $f_\theta|\theta$, by adjusting the parameters θ using a training dataset (x_i, y_i) and where $y_i \simeq f^*(x_i)$. The training of a feed-forward neural network involves two phases: the feed-forward phase and the back-propagation phase. The calculation done from the input to obtain the output of a FFNN is known as the feed-forward phase. During this phase, the training dataset is fed into the network and is propagated forward through the network. At each hidden layer, the weighted sum of the inputs is calculated and passed through an activation function. This process continues until the output layer is reached, and a prediction is made. Once a first prediction is made, the error between the predicted output and the actual output is calculated, using a loss function $\mathcal{L}$. Different loss functions are used, depending on the

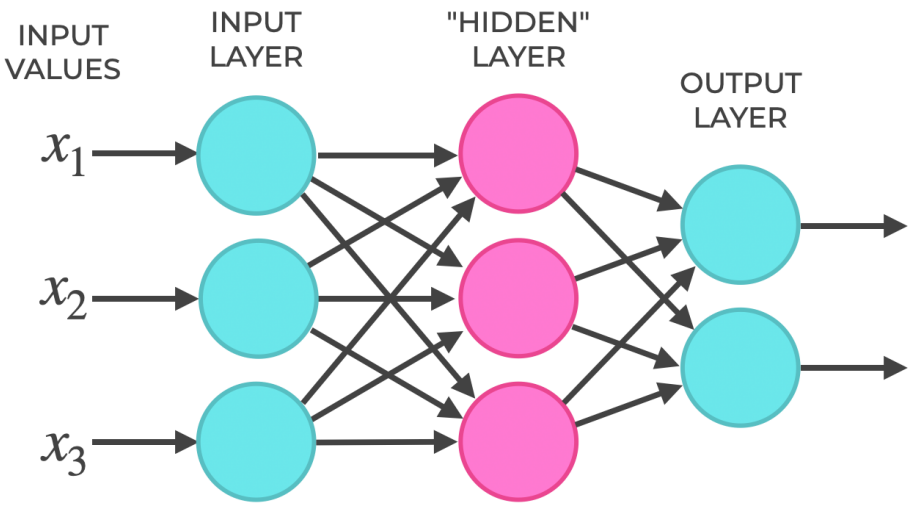


Figure 3.9: The structure of a Multi-layer perceptron (FFNN). Each layer is made up with units known as neurons, the layers interconnected by weights w, the inputs layer consists of neurons that receive inputs and pass them on to the next layer. The number of neurons in the input layer is determined by the dimensions of the input data. The network can have zero or more hidden layers and they are not exposed to the input or output and can be considered as the computational engine of the neural network. Each hidden layer's neurons take the weighted sum of the outputs from the previous layer, apply an activation function, and pass the result to the next layer. The final layer that produces the output for the given inputs. The number of neurons in the output layer depends on the number of possible outputs the network is designed to produce. Each neuron in one layer is connected to every neuron in the next layer, making this a fully connected network. The strength of the connection between neurons is represented by weights, and learning in a neural network involves updating these weights based on the error of the output.

nature of the task at hand. The most common ones are listed below, where we denote by y_i the actual target, $\hat{y}_i$ the predicted value (or $f_\theta|\theta$), N the total number of training samples:

- The **Mean Square Error (MSE)**:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

- The **Mean Absolute Error (MAE)**:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

- The **BCE**:

$$\text{Binary Cross-Entropy} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

The binary cross entropy loss is used for binary classification tasks. In the deep learning models developed in this thesis, we used the Binary Cross Entropy loss as we are dealing with binary classification tasks.

One the loss has been computed, it is then propagated back through the network, and the weights are adjusted to minimize the loss: this is the back-propagation phase. The process of adjusting weights is typically done using a *gradient descent* optimization algorithm. The process of iteratively passing the dataset through the network multiple times, and each time updating the weights to reduce the loss is a form of gradient descent. This process continues until the network performs well on the training data. But what is *gradient descent* and how does it work?

Gradient descent Gradient descent is a method in convex optimization, which corresponds to using the first order derivatives of the loss function $\mathcal{L}$ to adjust the parameters to be determined, so as to find a local minimum. The gradient $\nabla_\theta \mathcal{L}$ (partial derivatives) of the loss function are computed with respect to each parameter of θ. The resulting gradient indicates the direction and rate of the steepest increase of the loss function. From here, the model updates the parameters θ in the direction opposite to that of the gradient, so as to minimize the loss. The size of this step is controlled by the *learning rate* $\gamma \in \mathbb{R}^+$:

$$\theta_{t+1} = \theta_t - \gamma \nabla_\theta \mathcal{L} \tag{3.16}$$

The term $\gamma \nabla_\theta \mathcal{L}$ is substracted from the current parameters θ_t because we want to take the direction against the gradient, against the local minimum

of the loss $\mathcal{L}$. This process of computing gradients is repeated and allows to update the parameters θ until the loss function converges to a minimum value, or after a predefined number of iterations (epochs) is reached. There are multiple variants of gradient descent, each of them with defined use cases and characteristics. They include Batch Gradient descent, Stochastic Gradient Descent or Mini-Batch Gradient Descent. For a more detailed description, we refer the reader to the book by Christopher Bishop [Bishop (2006)].

Challenges of using feed-forward neural networks

Working with neural networks come with various challenges that can affect the performance or the deployment of the model. We will discuss in this section few of them that may arise prior, during, or after the training phase.

Data quality and quantity: In general, feed-forward neural networks require a large amount of data of good quality . Insufficient data can lead to overfitting, where the model performs well on training data but not on unseen data. In such cases, alternative machine learning methods should be considered for tasks and data types that allow it; in addition, one can explore data augmentation strategies, which consists in generating synthetic data, if possible.

In addition, poor-quality data may cause poor model performance. This characteristic is actually common to all machine learning models. Therefore, we must ensure of the high-quality data before the training phase.

Another challenge related to the dataset is the imbalancedness , when many more samples belong to one class (the majority class) than the other (the minority class). In classification problems, the model tends to become biased towards the majority class and then lead to poor performance for the minority class. This problem exists in many real-world datasets, including EHR datasets. Several approaches have been proposed to address this issue: We can use resampling techniques such as undersampling or oversampling [Chawla *et al.* (2002)]; class-weights [Fernando & Tsokos (2022),Rezaei-Dastjerdehei *et al.* (2020)] which consist of penalizing the loss of the majority class more than that of the minority class; or create balanced batches during training, which allow to sample the same amount of positive than negative samples during training, to address this problem.

Model Optimization: The optimization phase is a crucial one: the model parameters are expected to truly minimize its loss. However, the neural network loss functions are typically non-convex, with many local minima and saddle points. Optimizers can get stuck in those points, preventing them from finding the global minimum. Another major problem may occur when using certain activation functions such as sigmoid functions or hyperbolic tangent: the vanishing gradient and exploding gradient phenomena.

In FFNNs, gradients can become very small (vanishing) or very large (exploding) during backpropagation. Vanishing gradients make it difficult to train lower layers, because the parameters are not updating enough anymore, while exploding can cause numerical instability, with high range of gradients updates, which then makes the process unstable. Therefore, the choice of the learning rate is an important step when training a FFNN. A learning rate that is too high can cause the model to converge too quickly to a suboptimal solution, while a too low learning rate value lead to a slow optimization process. Adaptative learning rate methods such as Adam [Kingma & Ba (2014)] can be used to avoid this problem.

Model Generalization: The risk of overfitting (when the model learns the noise in the training data) is important in FFNNs, due to the possibly very large number of parameters to learn. We can use regularizations techniques such as L1/L2 regularization as detailed in section 3.2.2, dropout, and data augmentation to avoid overfitting. Moreover, monitoring the model's performance on a validation set and stopping training when performance starts to degrade is a efficient way to prevent overfitting. This method is called Early-Stopping. Many of these techniques aim at reducing the complexity of model architectures so as to improve generalization. We will not give further details about these techniques in this thesis, but we refer the reader to the book of Christopher Bishop [Bishop (2006)] for more details.

3.4 ML models development and evaluation

Overall, we have seen that developing machine learning models involves multiples phases, each of them crucial for a reliable and effective model. Moreover, there are many phenomena associated to those phases that need to be considered during the model development. Developing a machine learning involves a systematic approach, starting from problem definition and data collection to model deployment and maintenance. In this section, I will depict the whole pipeline to build a supervised model that can be deployed and used for real-world tasks. I will divide the pipeline into 3 main phases: (i): Prior-to-training phases, (ii) Training phases and (iii) Post-training phases.

Prior-to-training phases: These phases include significant steps in the process of building a machine learning model. First, we need to clearly **define the problem**, namely the main objectives, and the scope and the constraints linked to the project. It is essential to understand the task, determine available IT resources, time and data before getting into the model development. Secondly, we will focus on data-related steps: **data collection, data preprocessing** and **feature engineering**. This is probably the most important step in this process. After gathering relevant anno-

tated data, we will evaluate the quality of the data to ensure it is accurate and complete. Next, data preprocessing can include cleaning steps, such as handling missing values, outliers or noise, normalize or standardize the data, encoding techniques for certain features (binarization, one-hot encoding etc.). Moreover, we need to choose a suitable data format for the task. After that, a feature engineering steps comprises feature selection, which aims at identifying and selecting the most relevant features for the model, and feature creation, which consists in generating new features from existing features. Finally, we will split the cleaned data into three different sets: the training set, that will be used to train the model, which corresponds to the highest proportion of the data (around 70% in general), the validation set, that will be used to evaluate the model during the training and prevent overfitting (around 10% of the data), and the test set that will be kept unused until the end of the training and that will help determine the performance of the model on until now unseen data (it represents around 20% of the whole dataset).

By the end of these phases, we must ensure that the data is accurate and adapted for the task and for the model.

Training phases: As presented in previous sections, the aim of these phases is to learn patterns within the data, i.e, to find the best approximation of a function f^* by adjusting the parameters $\theta = (b, \omega)$ using and a training dataset (x_i, yi) such as $y_i \simeq f^*(x_i)$. We first will need to **select a model** by choosing the suitable machine learning model or the appropriate architecture for deep learning models, based on the problem. We only focus on some machine learning models in this discussion, but it also applies to other variants of supervised machine learning models (Gradient Boosting, Convolutional Neural Networks, Recurrent Neural Networks etc.).

After that, the training phase as described in previous sections is computed with the appropriate IT resources. During that phase, the parameters θ of the model are optimized, and in parallel we must **tune the set of hyperparameters** of the model.

What are hyperparameters? Until now, we only mentioned the parameters θ that are used to best fit the model to the data. The hyperparameters, by contrast, are all the parameters whose values can control the learning process and determine the model parameters. They include the number of trees, the maximum depth of trees or the criterion for the split, in RF classifiers, the regularization penalty or the regularization coefficient in LR or SVM, and the learning rate, the topology of the network, the optimizer or the regularization for deep learning models. The difference between parameters and hyperparameters is that the hyperparameters cannot be inferred while fitting the model to the data, unlike the parameters. However, most performance variations can be attributed the the hyperparameters. So, it is an important step to tune the hyperparameters in order to tweak model

performance for optimal results.

Several methods have been developed to find the best set of hyperparameters for a given model. Among them, we can use Grid Search [Hsu *et al.* (2003)], which is an approach where a predefined set of hyperparameters and their values are specified, and the model will be trained and evaluated for all possible combinations. This method is simple to implement, but computationally expensive because all the combinations will be considered. Another method is the Random Search [Bergstra & Bengio (2012)]. In this case, the hyperparameters values are sampled randomly from predefined distributions and the model is trained and evaluated on random combinations. This method is more efficient in high dimensional hyperparameter spaces, however, we can miss optimal combinations, since the search is random. Finally, the Bayesian optimization method [Osborne *et al.* (2009)]uses probabilistic models to model the performances of hyperparameters configurations and to select the next set of hyperparameters to evaluate based on expected improvement. This method is efficient because it can find optimal hyperparameters with fewer combinations tested, but it is more complex to implement and computationally intensive. For this thesis, all the hyperparameters have been tuned using this last method.

These training phases are completed once the best model is chosen, i.e, we select the model with the optimal parameters and hyperparameters sets trained on the training set and with the best results with the validation sets.

Post-training phases: In these phases, we **evaluate** the model on the unseen set: the test set, in order to determine its ability to generalize to new data. In Machine Learning, there are multiple metrics that are commonly used to monitor the performance of the model. In the context of classifications tasks there are terms used to describe model's predictions: a true positive (TP) occurs when the model correctly labels an instance of the positive class, a true negative (TN) is when the model correctly labels an instance of the negative class, a false positive (FP) is when the model incorrectly labels as positive an instance that is actually negative, and a false negative (FN) occurs when the classifier predicts the negative class while the instance was actually positive.

The threshold is another important concept in model evaluation. In deep learning, models predict probabilities of an instance belonging to the positive class. These probabilities are converted into class using a threshold. If the predicted probabily is for instance above a given threshold (0.5 by default), the instance is classified as positive and negative otherwise.

Let us note that the metrics presented here are also used in the hyperparameter tuning step described in the previous section, but also in all machine learning models presented in this thesis. I will present few of them used for classification task, which are the ones mainly used in this thesis.

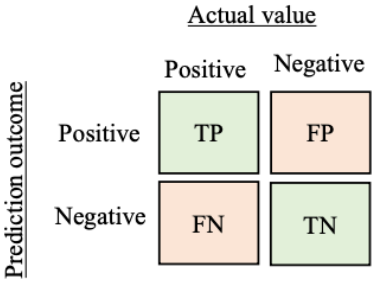


Figure 3.10: Confusion matrix

- **Precision** and **Recall**: The Precision is defined as the proportion of true positives among all predicted positives, while recall or sensitivity is the proportion of true positive among all actual positives. The Precision score demonstrates the ability of a classifier to correctly predict positive samples and the recall is intuitively the ability of the classifier to find all positive samples. The best score for Precision and Recall is 1 and the worst is 0.

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{3.17}$$

$$\text{Recall(Sensitivity)} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{3.18}$$

- **F1-score**: This metric is the harmonic mean of precision and recall, and provides a single score that balances both precision and recall. This metric is suitable for imbalanced datasets.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{(\text{Precision} + \text{Recall})} \tag{3.19}$$

- **Confusion matrix**: It is not by definition a metric, but it is a table that allows to display the TP, TN, FP and FN proportions. It is generally in the following format:

- **APS and Precision-Recall curve**: The Average Precision Score (APS) is calculated by integrating the precision values over all possible recall levels. This metric shows the ability of a classifier to correctly predict accross all possible thresholds. The Precision-Recall curve (PR-curve) is the graphical representation of the trade-off between precision and recall for different thresholds. APS is a scalar value that summarizes the PR-curve. Its range is from 0 (bad performance) to 1 (good performance). APS is suitable for imbalanced datasets where positive class is rare.

- **AUC-ROC and ROC curve**: The Receiver Operating Characteristic (ROC curve) plots the trade-off between the TP rate and the FP rate. It provides a visual summary of the model's ability to discriminate accross different thresholds. The AUC-ROC score (Area Under the ROC curve) is calculated by integrating the area under the ROC curve. As previous metrics, it ranges from 0 to 1. A random classifier has an AUC-ROC of 0.5.

3.5 Deep Learning for sequential data

I now describe in depth deep learning algorithms that can learn from sequential data. Although they have typically been developed in the context of Natural Language Processing for sequences of letters or words, they can be extended to learn from any sequence, including sequences of DNA base-pairs or amino acids in bioinformatics, and sequences of medical events in the context of EHRs.

3.5.1 Key concepts of natural language preprocessing

Before we dive into models for sequential data, we will present some of the concepts that form the foundation of NLP and that are essential for building applications that handle sequences effectively.

Tokenization is a the first step in many NLP pipeline. It is the process of breaking down sequences into smaller units called **tokens**. These tokens can be words, subwords, characters, or sentences, depending on the level of tokenization. Each token will be turned into a vector representation that are suitable for computational analysis.

Name Entity Recognition (NER) is an NLP method that aims to identify and classify *named entities* (e.g people, organizations, locations) in an unstructured text. It is particularly useful for quickly extract information from a large amount of unstructured data.

The **Stemming and Lemmatization** steps aim to reduce the dimensionality of the text data and deal with word variations. The Stemming method reduces each word to its base or root form by removing suffixes. The Lemmatization is an alternative, that will use the base form of the word based on the vocabulary or the morphological analysis.

Stopwords represent most common words that carry little semantic values ("the" or "and": for instance). One main step in NLP pipeline is the stopwords removal. This step reduces noise and focuses on more meaningful words. Stopwords lists are readily available in public package and in many languages.

Token embedding or word embedding is a crucial step in sequential models. It allows to define a meaningful representation for tokens. In the

context of deep learning models for sequential data, defining a word embedding method enables a representation of word as dense vectors in continuous space, which vector captures the semantic meanings and relationships between items in the sequence.

3.5.2 Transformers and Attention mechanisms

Until recently, models such as Recurrent Neural Networks (RNN) and Long Short-Term Memory neural networks (LSTM) have been the state-of-the-art models for Natural Language Processing (NLP) tasks. These deep learning methods are popular types of neural network architectures for sequential data processing. As shown in Figure 3.11 their architecture have directed circular connections that allow them to maintain the memory of the latest inputs.

As presented in Section 3.3, training neural networks with gradient based learning and backpropagation [Hochreiter (1998)] can be challenging for long sequences due the vanishing gradient phenomenon. In RNNs, this problem make it difficult for the network to learn long-range dependencies. LSTMs adress the vanishing gradient problem by leveraging gating mechanisms to control the flow of information and grandients. As a result, LSTM have become widely used in different NLP applications. More recently, the use of encoder-decoder LSTMs methods has started to emerge for certain applications such as machine translation [Sutskever *et al.* (2014), Bahdanau *et al.* (2014)].

However, the recent advent of more modern deep learning architectures has revolutionized the NLP field. The "Attention is All You Need" paper [Vaswani *et al.* (2017a)] has introduced a novel deep learning architecture in 2017 for NLP tasks: *Transformers*. Transformers have had a major impact on NLP and have found applications in various other domains. The model relies on attention mechanisms to capture the dependencies between words in a sentence, bypassing the need for recurrent neural networks (RNNs or LSTMs).

Attention mechanisms The core idea behind Transformers is attention mechanisms. They allow models to focus on different parts of the input sequence when making predictions, rather than treating all inputs equally. In encoder-decoder based models (RNNs or LSTMs), the model's final prediction is made based on the final hidden state of the encoder, which may not greatly capture long-range dependencies in the input sequence. With the attention mechanism, the model selectively pays *attention* to certain parts of the input sequence when generating output. The first applications that made use of the attention score were natural language transduction tasks, or sequence-to-sequence (Seq2Seq) tasks. Seq2Seq tasks consist in transforming an input sequence into another output sequence (machine translation

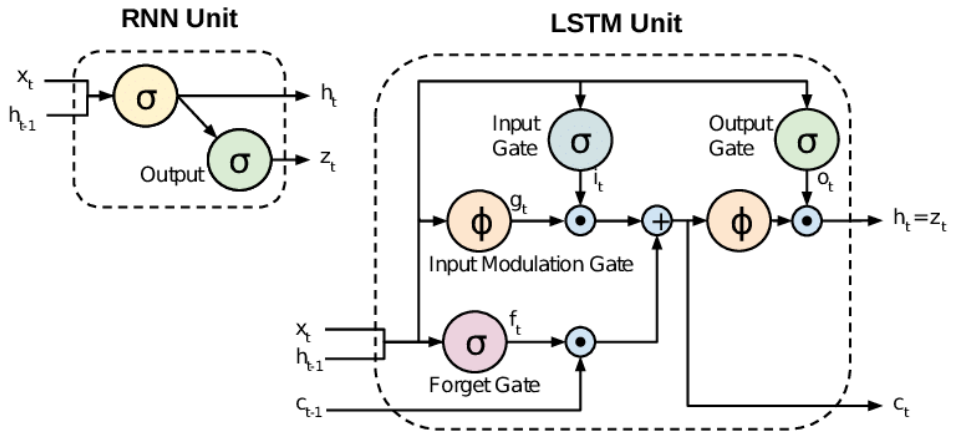


Figure 3.11: RNN unit (left) combines the input x_t with a hidden state h_{t-1} that captures information about previous inputs at $t-1$ in the sequence through an activation function σ. The LSTM units (right) have a more complex structure with gates that control the flow of the information. The key feature in LSTM is the ability to maintain a cell state c_t which allows information to flow unchanged across many time steps. Moreover, we count many gates, including the input gate i_t, the forget gate f_t and the output gate o_t, that allow to selectively learn information to retain and to discard over time. We denote $\odot$ as an element-wise multiplication. The input x_t is combined with the hidden state h_{t-1} from the previous input through an input gate i_t which controls how much new information from the current input should be added to the cell state. the same information is given to a forget gate f_t that controls the amount of information retained. The output is given by the output gate o_t and combines the actual input x_t and the information from the cell state c_t to generate a activation vector as the output of the unit. The output represent the hidden state at t.

for instance). To do so, the most suitable method was the use of an encoder-decoder architecture to first process and encode the input sequence to fixed-length context vector and use the decoder to generate the output sequence from the encoded information. In modern Seq2Seq models, the attention mechanism is incorporated to allow the decoder to focus on the important parts of the input sequence at each step of the decoding process, instead of relying on a single fixed-length context vector. The important parts are determined by the *attention score*, which is a numerical value that informs about the relevance of each specific element in the input sequence. These methods have significantly improved the performance achieved for Seq2Seq tasks, especially for longer sequences. In the "Attention is All You Need" paper [Vaswani *et al.* (2017a)], the attention mechanism has been reformulated into a general form that can be applied to many NLP task. Its concept relies on a function that maps a query and a set of key-value pairs to an output.

Query, key and value and Scaled-dot product *Queries* Q, *keys* K and *values* V are the three main components used to compute the attention score. For each item in the input, the model generates these three vectors Q, K and V.

Queries, *keys* and *values* are so named by analogy to retrieval systems. For instance, when we want to make a scientific search on a search engine, this latter will map a **query** (the text put in the search bar) against a set of **keys** (article titles, abstract) associated with the most reliable results, and finally shows the most suitable articles for your research: the **values**. The query, key and value vectors for a given input item (typically, a token) x are obtained as projections of this input: the *query* vector of dimension d_k is obtained as $q = w_q \cdot x$, the *key* vector of identical dimension d_k is obtained as $k = w_k \cdot x$ and the *value* vector of dimension d_v (often chosen to be equal to d_k, but not necessarily) is obtained as $v = w_v \cdot x$. The *attention weight* between a query q_i and a key k_j, $\alpha(q_i, k_j)$, scores the compatibility between the query and the key.

Let us assume that we want to perform machine translation from French language to English language of a given sentence. During training, the output sequence is fed to the model to allow the model to learn the translation between the two languages. At a particular word x_i in the English sentence (represented by the query q_i), the keys k_i are the representation of all words x_j in French sentence and the values v_j are the words to be translated . At a particular word in the sentence, the attention score computed between the query q_i and the key k_j demonstrates the importance of the key to predict the query in the output sentence. Then, the attentions scores are used to weigh the values v_j, which represents words to be translated. The higher the score is, the more that word is important for the output.

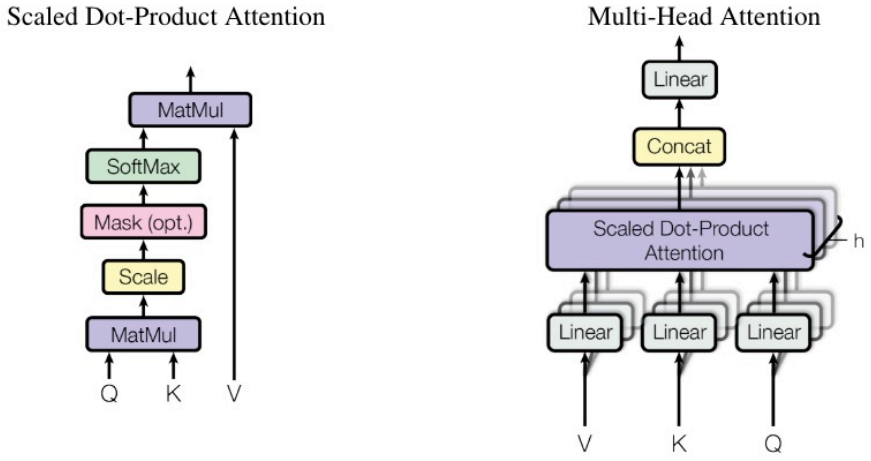


Figure 3.12: Scaled dot-product (left) and Multi-Head Attention module (right) [Zhou *et al.* (2019)]. Attention scores are computed using the scaled dot-product. Multi-head attention involves applying the scaled dot-product attention mechanism in parallel across multiple attention heads The outputs of each head are concatenated and linearly transformed to produce the final output.

By definition, the attention score represents a weighted sum of the attention values [Bahdanau *et al.* (2014)].

$$\text{Attention}(q, k, v) = \sum_i \alpha(q, k_i) V_i \tag{3.20}$$

In the [Vaswani *et al.* (2017a)] paper, the weights α_i are determined by a compatibility function between the query and the corresponding key, and can all be computed simultaneously in matrix form: 3.21.

$$\text{Attention} = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V \tag{3.21}$$

This function calculates attention weights between all pairs of input tokens in parallel. Each element in an input sentence would be attributed its own query, key, and value vectors, generated by multiplying the encoder's representation of the specific element under consideration with three different weight matrices W^Q, W^K and W^V that would have been learned during training.

Is it interpretable? Attention scores can be visualized (see Figure 3.13). They can be further inspected to provide insights into the model's decision-making process. By examining which part of the input sequence the model

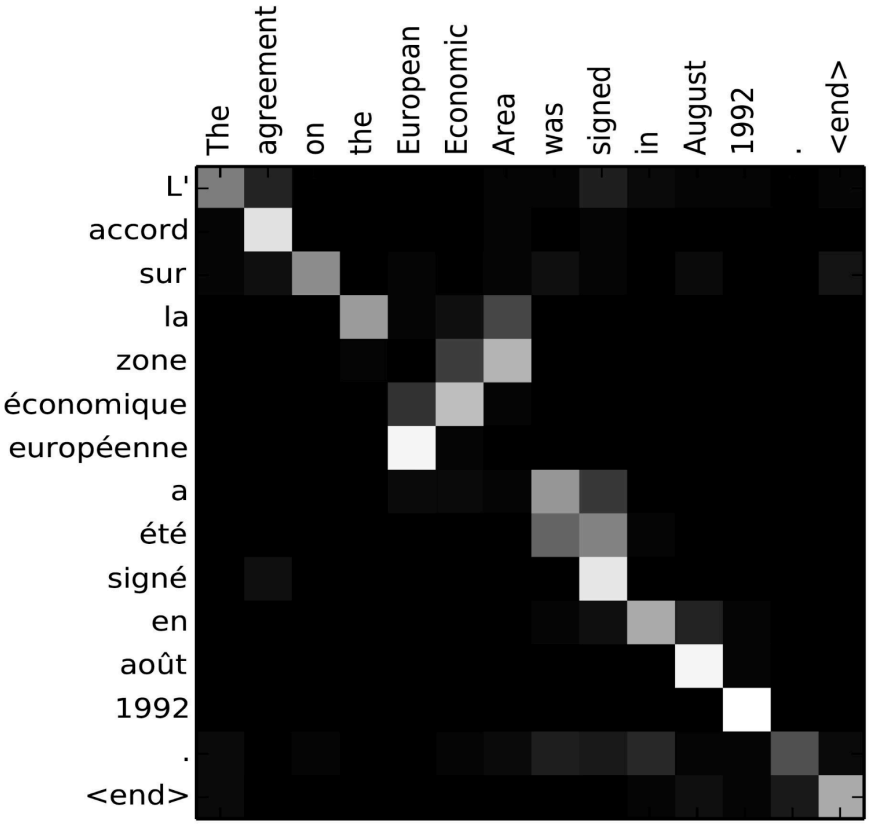


Figure 3.13: Attention scores between tokens of two sentences in a translation task. The brighter the square is the higher the score is between the two corresponding tokens [Bahdanau *et al.* (2014)].

has focused to output the prediction, we can infer important parts within the input for a given task (text classification, sentiment analysis, etc.). In this example of machine translation 3.13, we can tell which words in the source sentence the model is attending to for each word generated in the predicted translated sentence.

In addition, attentions scores can be used to analyze model errors by looking at the model's focuses with incorrect predictions. It can give insights on the model behavior and its potential biases. Thus, it can help researchers to better refine and improve the model. Attention scores provide some degree of transparency in model outputs, which can be crucial for certain applications, such as healthcare research. However, according to certain researchers, attention is not necessarily associated to importance, and attention scores can even be inconsistent and noisy predictors [Jain & Wallace (2019), Wiegreffe & Pinter (2019), Serrano & Smith (2019), Pandey *et al.* (2022)]. They often do not correlate with other measures of feature im-

portance, such as gradient-based methods [Jain & Wallace (2019),Serrano & Smith (2019)]. In these studies, the authors did experiments that show that removing features considered as being important by the attention scores lead to less decision flip in the model than the features described as important by gradient-based models [Jain & Wallace (2019),Serrano & Smith (2019)].Additionally, in certain cases, various attention distributions will still have the same output predictions [Jain & Wallace (2019)]. In another study [Bai *et al.* (2020)], the authors demonstrates that, in NLP tasks, attention mechanism can focus on uninteresting tokens because of an effect they called "combinatorial effect". Nonetheless, attention scores can offer a form of explanation for the model's output in certain NLP taks. Attention mechanisms usually provide faithful explanations in syntax-related tasks, such as Part-of-speech task or syntactic annotation [Clark *et al.* (2019),Vig & Belinkov (2019)]. On the other hand, Zhang and al. [Zhang *et al.* (2018)] agree to the ability of attention scores to capture the importance of abstract features when dealing with images. In all cases, attention scores should be combined with other interpretability methods to ensure a comprehensive understanding.

Finally, we have to admit that interpretability remain a pressing concern for many NLP models (details on 3.7), especially as modern deep learning models become increasingly complex, despite their high performances.

Transfomers' architecture Transformers are composed of encoder and decoder stacks, where each layer consists of multi-head self-attention and feed-forward neural networks, followed by layer normalization and residual connections, as depicted in Figure 3.14

The multi-head attention mechanism extend to attention mechanism (heads) by running multiple heads in parallel (Figure 3.12). Each head operates independently. The final model's output is the concatenation followed by a linear transformation for each head's output, given by Equation 3.22 and 3.12. It allows the model to focus on various parts of the sentence simultaneously, improving its ability to understand context.

$$\begin{aligned} \text{MultiHead}(Q,K,V) &= \text{Concat}(\text{head}_1,\ldots,\text{head}_h)W^0, \text{where} \\ \text{head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{3.22}$$

The *cross-attention* module extend the capabilities of traditional transformer architectures to handle multiple modalities of data, in a unified framework. In a cross-attention module, the attention mechanism computes weights between the elements of the query sequence and the key-value pairs of another sequence. The formula for cross-attention is similar to that of self-attention but applied across different sequences.

Training transformers Training steps are the same as training a basic deep learning model (FFNN) as described in Section 3.3.

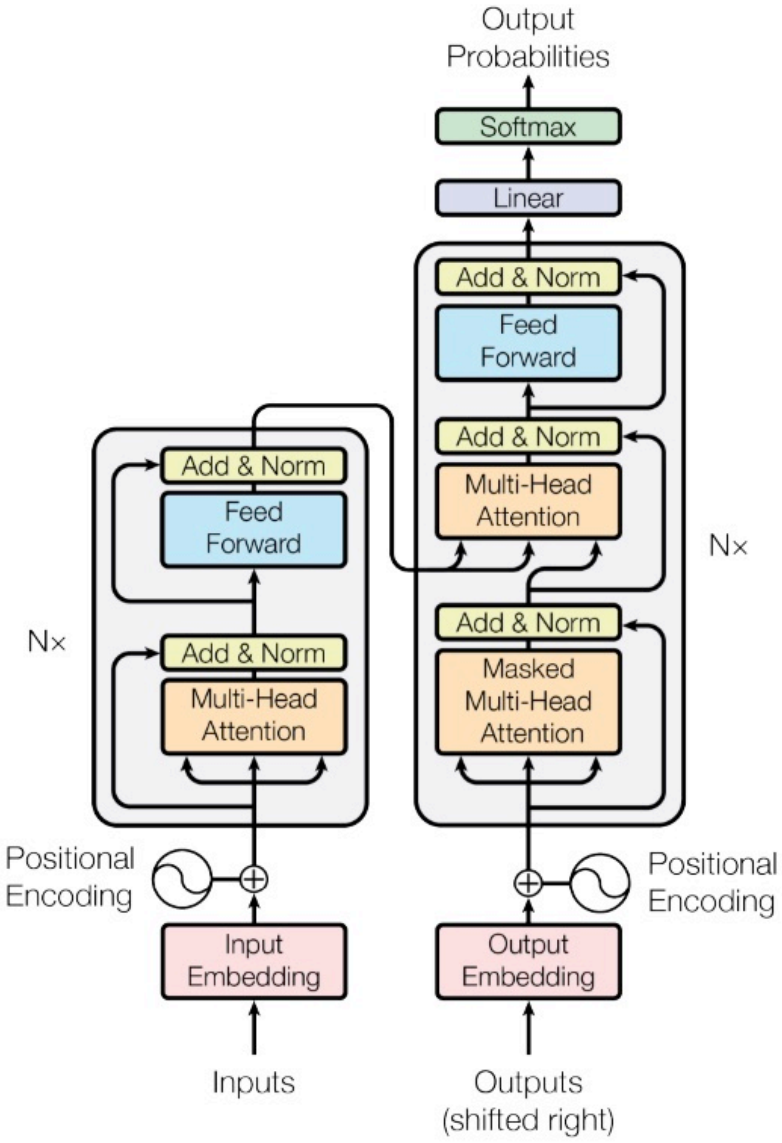


Figure 3.14: Transformers architecture as proposed in [Vaswani *et al.* (2017a)]. This model consists of an encoder and a decoder. The encoder layers (left) includes multi-head self attention mechanism and a FFNN, while the decoder layer (right) includes a additional multi-head attention mechanism over the encoder's output. Let us recall that an encoder-decoder architecture is used for Seq2Seq tasks. For classification task, the encoder part is only considered.

Advantages and Disadvantages Transormers have outperformed many languages modeling approaches, mostly in applications such as Seq2Seq models and classification tasks. Their architecture offers a significant advantage for the NLP field. The scalability and parallelism of transformers make them highly efficient for training on large datasets. In fact, they can process mutliple parts of a sequence at the same time. Since its introduction, the transformer architecture has become the basis for many state-of-the-art models in natural language processing, including Google's BERT (Bidirectional Encoder Representations from Transformers) [Devlin *et al.* (2018)] and OpenAI's GPT (Generative Pre-trained Transformer) series [Radford *et al.* (2018), Radford *et al.* (2019), Brown *et al.* (2020)]. But despite their impact on the NLP field, transformers still struggle with many aspects. First, transformers require large amonts of data to train effective models. For instance, the well known chatbot ChatGPT was trained on a massive corpus of text data, around 570GB of datasets, including web pages, books, and other sources. Additionnally, training and running transformers models need a great amont of computational requirements that are often inaccessible for most of the facilities and make it difficult to democratize modern AI.

3.5.3 Pretrained Models

All machine learning models that have been previously trained on a large dataset and can be used as a starting point to a downstream task are called pretrained models. Pretrained models are common in certain fields such as NLP, computer vision or speech recognition. For this thesis, we will focus only on the NLP field. Pretrained models facilitate transfer learning by allowing users to leverage the knowledge caputred during the pretraining task. Pretrained models demonstrate enhanced **transfer learning** capabilities, that make them more adaptable to new tasks.

How does it work? In the pretraining process, a model is trained typically on a large dataset to learn general features of the data. The aim is to expose the model to a wide variety of context and the language patterns. The idea is to leverage a large amount of unlabeled data to learn meaningful representations that can be further adapted to other tasks further. The model is trained in an unsupervised or self-supervised manner, meaning that it has to learn from the data itself without any other information (labels for instance). Outputs are dense vector representations of words, phrases, or sentence. Those embeddings capture semantic information from the input, learned from the pretraining. Common techniques that include Masked Language Modeling (MLM) or the Next Sentence Prediction (NSP) will be discussed in detail in Sections 3.5.3 and 3.5.3.

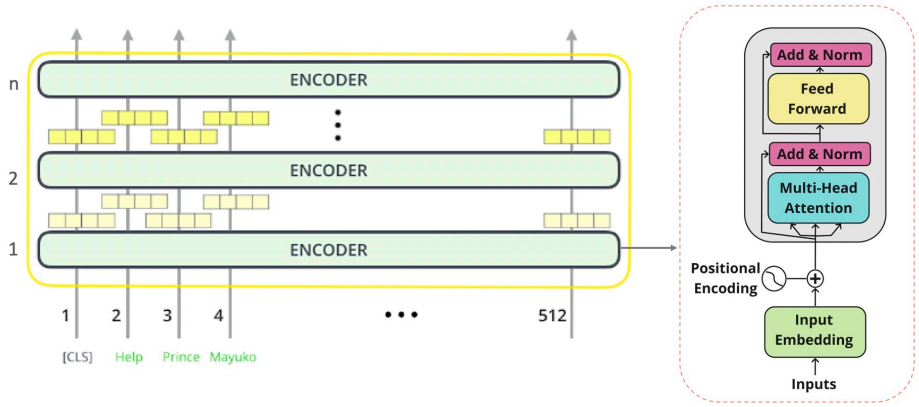


Figure 3.15: BERT architecture consists of a n stacked encoders blocks. n = 12 for Bert_base and n = 24 for Bert_Large.

Why do we pretrain? Pretraned models provide a meaningful foundation for many NLP tasks. First, by starting the fine-tuned task with already adjusted weights learned during the pretraining, rather than from scratch, it allow a gain of training time and performance, and a computational resources savings. Pretraining allows downstream models to be optimized quickly and it tend to facilitate generalization to new tasks. Moreover, the downstream task require less labeled data to converge and to make it efficient in performance.

Bidirectional Encoder Representations from Transformers (BERT) The BERT (Bidirectional Encoder Representations from Transformers) have been introduced in 2018 by researchers in Google AI [Devlin *et al.* (2018)]. BERT is a transformers based deep learning model that was designed to pretrain deep bidirectional representations of words by jointly conditioning both left and right context. It has achieved state-of-the-art results on a variety NLP applications and represent in recent language method one of the most powerful approach. The BERT model has been pretrained on unsupervised tasks (MLM and NSP) on a large corpus of 800M words called BooksCorpus [Zhu *et al.* (2015)] and the English Wikipedia (2500M words). BERT is an open source model and its architecture is shown in 3.15

Bidirectionality Most of the NLP models that preceded BERT were unidirectional, meaning that the model processes sequences in a single direction (right to left or left to right) . The BERT model considers both preceding and following words to form the context of a word. This leads to a broader view of input context and a deeper understanding of the language

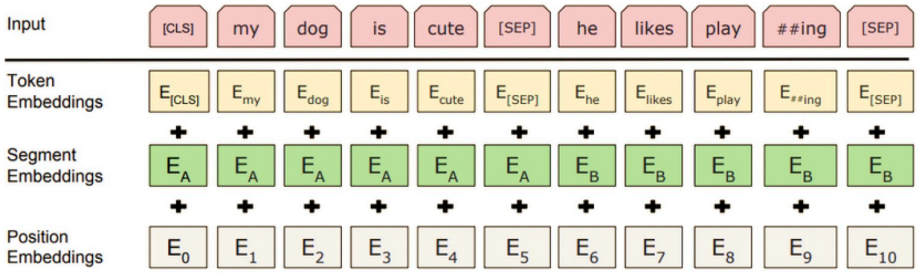


Figure 3.16: BERT input representation. The input embeddings are the sum of the token embeddings, the segment embeddings and the position embeddings.

patterns.

BERT input representation The input representation for BERT is obtained using many embedding layers. A visualization of the input is shown in 3.16

The first layer is the input sequence. It refers to the input token sequence that can either contains one sequence or two sequences packed together in Seq2Seq models. The [CLS] (stands for classification) token always begin a sentence. The [CLS] token is an important token, as its final hidden state is used to represent the aggregated sequence for classification tasks [Devlin *et al.* (2018)]. [SEP] is another BERT special token that separates successive sentences. The second embedding layer is called segment embedding. It reflects the alternation of representations throughout sentences. Also called *token_type_ids embeddings*, this layer contain a vector : `[0, 0, ..., 0, 1, 1, ..., 1, 0, 0, .., 0]` where the first sequence of tokens with the *token type ID* $= 0$ corresponds to sentence A, and the second with the *token type ID* $= 1$ corresponds to the sentence B. These two token type ID distinguish two following sentences. The position embeddings as its name shows the position of the sentence within the whole sequence. For several reasons related to computational efficicency and model architecture, that will be detailed in the chapter 6, BERT is designed to handle a maximum input sequence of 512 tokens. Therefore the position embedding is a defined by a sequence that starts at 0 up to 511.

Masked Language Modeling MLM is the first pretraining phase used in BERT. This concept was used in earlier litterature with the name Cloze procedure, a tool that measures the effectiveness of communication [Taylor (1953)]. In this pretraining task, the model is trained to predict masked words using the context provided by the remaining words around

the masked words. The final hidden state output vector that contain masked tokens is fed to a softmax over the entire corpus vocabulary to retrieve the masked words. As a result, we obtain a probability distribution over the vocabulary. In BERT, 15% of the words are randomly masked in each input sequence. The selected tokens are replaced with the BERT special token [MASK] with a probability of 80%, other tokens are swapped with another random token with a probability of 10% and the remaining tokens are kept unchanged.

Next Sentence Prediction The second pretraining task aims to learn latent representations of sentences in the context of the corpus. It takes as input a pair of sentences of the corpus. Some of these sentences are consecutive in a text and are labeled "IsNext", others are random pairs from the corpus and are labeled "NotNext". The model is trained to predict whether the second sentence in the pair follows the first in the original text. It is formulated as a binary classification task. During training, for each input sample, 50% of the time the second sentence is the actual next sentence in the text and 50% it is a random sentence from the text.

Fine-tuning BERT After pretraining phases, BERT can be fine-tuned on specific tasks. During the fine-tuning phase, the model is trained end-to-end, updating the pretrained weights based on the downstream task. BERT's transformers-based architecture allows for easy adaptation to many downstream tasks, including classifcation, named entity recognition, question answering, and more. In tasks such as sentiment analysis, topic classification or Q&A tasks on datasets such as SQuAD (Stanford Question Answering Dataset) , BERT has new achieved state-of-the-art results [Devlin *et al.* (2018)] , as it benefits from the pretraining's contextual understanding, unlike the other methods.

Advantages and disadvantages While BERT offers multiple advancements in NLP tasks, including its ability to capture bidirectional context, and its pretrained language model approach, which allowed improved performance and efficiency for several tasks thanks to fine-tuning to specific downstream tasks, it also come with multiple challenges. First, BERT requires significant computational resources. The model's complexity and size lead to high memory usage and longer training times, which limits its deployment on devices with not enough resources. We also count among limitations the token limitation, as mentioned in Section 3.5.3, which can affect performance on tasks that require full-document understanding. Another limiting aspect is that BERT requires domain adaptation. We need to fine-tune the pretrained model on specialized applications in order to have optimal results. For instance, to solve NLP tasks in specialized domains

such as breast cancer reports related corpora, we will need to fine-tune the model to the new corpora to have better performance. This additional step can be time-consuming and require the availability of enough computational resources. However, researchers continue to explore improvements and alternatives to address these limitations for a broader deployment in various applications.

BERT derivatives Since its release in 2018, many models inspired by the BERT model have emerged in the NLP community. Each new variant and extension of the model aims to specialize in certain other aspects of language modelling. First, we have BERT extensions that have been pretrained again with more data and longer sequences, to add more contextual understanding to the model, such as RoBERTa (Robustly Optimized BERT Pretrained Approach [Liu *et al.* (2019a)]). We also have BERT models that aim to reduce the number of parameters, or to make the model smaller and lighter, which is more suitable for many facilities with not enough computational resources. They include for example ALBERT (A Lite BERT [Lan *et al.* (2019)]), DistiBERT [Sanh *et al.* (2019)], or TinyBERT [Jiao *et al.* (2019)]. Moreover, the original BERT has been pretrained on a corpus written in English. Therefore, BERT was only adapted to tasks on English written texts. Therefore, models pretrained on other languages corpus have also been developed, for instance, MultiLingual BERT (pretrained on a large mutlilingual corpus of 104 languages), Chinese BERT [Devlin *et al.* (2019a), Devlin *et al.* (2019b)] (developed by Google among other language variants BERTs, alongside the original English BERT), or FlauBERT [Le *et al.* (2019)] and CamemBERT [Martin *et al.* (2019)] exclusively pretrained on French corpora.

Other BERT-like models have been developed for specific domains and using specific corpora instead of general text data. They include ClinicalBERT [Huang *et al.* (2019)], pretrained on the MIMIC III clinical database [Johnson *et al.* (2016)], which aims to improve performance on healthcare related NLP tasks such as clinical text classification or Q&A in the clinical domain. SciBERT [Beltagy *et al.* (2019)] is another specific BERT-like model pretrained on scientific publications and allows to enhance performance on paper classification task or relation extraction within scientific literature for instance.

Another notable example derived from BERT is the BEHRT model [Li *et al.* (2020a)].Rather than sequences of words, it considers sequences of medical events and uses longitudinal patients electronic health records for pretraining and aims to predict future medical events. BEHRT has been a pionner among BERT models in using longitudinal patients records to perform medical-specific tasks, paving the way for subsequent models in this domain: Med-BERT [Rasmy *et al.* (2021)], CEHR-BERT [Pang *et al.*

(2021a)], ExBEHRT [Rupp *et al.* (2023)] etc. In this thesis, I developed a model called M-BEHRT, which is an adaptation of BEHRT, tailored to the specific data and task of my research as fully described in Chapter 5 5.

3.6 Integration methods for different data modalities

Multimodal machine learning is an approach that offers a powerful way to extract deeper insights from data by combining different data types. In clinical research, it enables more accurate results by integrating the different information from medical images, patient tabular records and clinical notes, for example. These different modalities are expected to be complementary. For instance, in oncology, a patient state is characterized by a whole spectrum of information from different modalities, ranging from radiology or genomics to clinical reports. For a given patient, multimodal EHR can be characterized by the patient's demographic data (name, age, gender, etc.) that will give general description of the patient, questionnaires that provide baseline information about the patient health, laboratory results, MRI results and clinical notes during the patient's stay in the hospital, which indicate their ongoing health state. These different information allow to form a comprehensive view of the patient status when doing clinical research. However, the heterogeneity associated with multimodality remains a challenge when developing integrative models. First of all, the data have different formats and require different types of processing. For instance, the demographic data or biological tests are structured as matrices, whereas clinical notes are unstructured, in a free text format. In addition, even modalities that have similar formats (for instance, different types of medical images, or tabular data encoding different type of information) are not straightforward to combine. Therefore, integrating the different data modalities is for multimodal machine learning is not straightforward. Multiple integration methods have been proposed to be able to combine these different modalities [Gligorijević & Pržulj (2015b),Zitnik *et al.* (2019b)]. In this section, we will present three main categories of methods that are currently used for mulitmodal learning: the *early, intermediate* and *late* fusions (see Figure 3.17).

3.6.1 Early integration

In *early integration*, or *feature-level integration*, the fusion between the different modalities is done at an early stage 3.17.a. They are combined at the input level before being fed into the multimodal learning model. This combination usually takes the form of a concatenation or merging into one single input representation.

Formally, let us denote the input in each modality i as the-$d^{(i)}$ dimen-

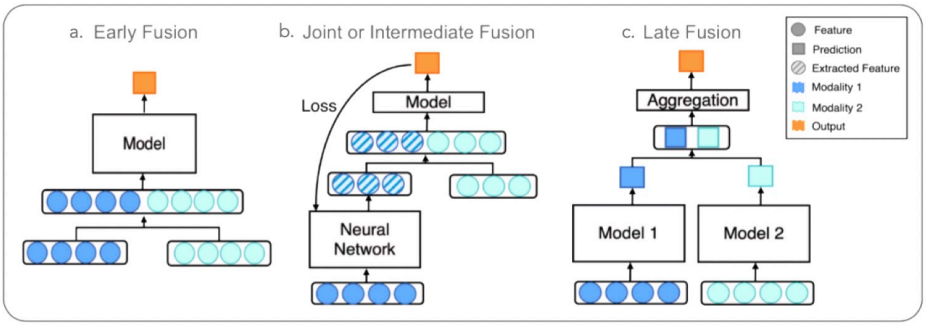


Figure 3.17: Illustration of early (a), intermediate (b) and late (c) integration methods.

sional input vector $x^{(i)}$ where $i = 1, ..., m$ and m is the number of modalities. Early integration is the most straightforward approach, where the input is a single input vector : $x = (x^{(1)}, x^{(2)}, ..., x^{(m)})$, which is a $\sum_{i=1}^{m} d^{(i)}$ dimensional vector. The model is trained with this input vector as input [Alpaydin (2018)]. Applications of early integration for multimodal learning model include integration of similar modalities such as multiview ultrasound images, or MRI data and PET scans, for cancer detection [Qian *et al.* (2021),Le *et al.* (2017)], treatment planning [Lipkova *et al.* (2019)] or survival prediction [Nie *et al.* (2019), Captier *et al.* (2023)]. The main advantage of this method is the simplicity of its implementation. In addition, it offers the possibility for the model to leverage all the mixed information from the beginning and to learn a combination of patterns. However, concatenating the data may lead to a high-dimensional feature vector as input. The classifier becomes more complex with a high number of features, and contains more parameters to fit. For that reason, it may need a high number of samples as well, if we do not want the model to overfit.This is not always possible, patient data being difficult to acquire, and will also require more computational resources.

Early integration also requires all the modalities to be present at the same time, whereas data can often be missing in one or more modalities for certain patients. We must therefore consider effective imputation strategies tailored for each data modalities. Additionally, it will be important to address the difference in units and scales of the different modalities within the input. In fact, the joint space defined by the concatenation of the different modalities may be difficult to compute for a model.Overall, we can think of using an early integration approach in scenarios where the concatenation of the different modalities does not result in an excessively high-dimensional representation, or where capturing feature-level interactions between modalities is crucial for the task In other cases, other integration methods may be more suitable.

3.6.2 Late integration

Late or *decision-level integration* merges predictions from separate learners trained with the corresponding representation of each input for each modality. In other words, each modality is processed independently, and their outputs are combined at a latter stage 3.17.c. The aggregation can be achieved using the average of all individual predictions, a majority vote, or Bayes-based rules [Ramanathan *et al.* (2022)].

In *late integration*, each model i makes its predictions for each modality i: $\hat{y}^{(i)} = f_{(i)}(x^{(i)})|\theta_{(i)}$, independently and in parallel. The final prediction among all the m modalities is given by the choosen aggregation method F, where, $\hat{y} = F(f_{(1)}(x^{(1)})|\theta_{(1)}, f_{(2)}(x^{(2)})|\theta_{(2)}, ..., f_{(m)}(x^{(m)})|\theta_{(m)})$.

There are examples of late fusion in the litterature, used for cancer prediction, with the fusion of MRI images and PSA-blood tests [Reda *et al.* (2018)], for survival prediction, with the combination of genomics and histology profiles [Chen *et al.* (2022)], or for the response to chemotherapy treatment, with the fusion of MRI or CT scans and EHR [Joo *et al.* (2021)].

The main advantages of using late integration are that it allows separate optimization of each modality, it is more flexible with missing data, and it offers a relatively more suitable dimensionality in input vectors. However, late integration might lose some correlations between modalities that could be captured if they were fused earlier. In addition, the final prediction may be heavily influenced by the modality that is the most dominant modality. In general, *late integration* approach can be used when there is no major interdependencies between modalities.

3.6.3 Intermediate integration

Middle or *intermediate integration* combines aspects of both early and late integration 3.17.c. It merges features at an intermediate level of the model, embedding the different modalities in a common feature space. In deep learning models, each modality is first processed separately, and their individual representations/embeddings are combined into a joint representation in further layers of the networks. Multiple multimodal learning models that use an intermediate integration have been proposed, such as multiple kernel learning (MKL) [Lanckriet *et al.* (2003)] or multimodal cross-attention mechanisms [Aiello *et al.* (2023)], among others. In oncology, intermediate integration has been used for example to combine different imaging modalities for cancer detection [Sedghi *et al.* (2020),Kumar *et al.* (2020)] or diverse multi-omics data for cancer subtyping [Liang *et al.* (2015)] or survival prediction [Lai *et al.* (2020)].

Intermediate integration balances the benefits of both early and late integration, by capturing interactions between modalities while allowing some independent processing. However, it is more complex to design and to im-

plement.

3.7 Interpretation of machine learning models

Interpreting machine learning models means extracting insights from the model's behavior so as to gain understanding of the underlying relationships in the data and the features that drive the model's predictions. Interpretation is important for multiple reasons: (i) it helps with machine learning model transparency and makes them more understandable to stakeholders; (ii) it can uncover potential biases in the model, and therefore, allows developers to debug more easily, (iii) it outlines features that contribute the most to the model's prediction, thus providing valuable insights into the underlying mechanisms of the studied outcome, (iv) it allows domain experts to validate the model's prediction against their domain knowledge.

Some models are *intrinsically interpretable*, such as logistic regression 3.2.2 for instance. There are known as *white box* models. They are designed to be already transparent and provide visibility into the decision making process. Applications that require high accountability and trust will most favor *white box* models. However, *black box* models can model more complex relationships, which can make them preferable. The particularity of these models is their lack of transparency, i.e, they output results based on the data, but do not clarify how the predictions are made. The main example include all deep-learning models and boosting models, among others. For deep learning models, we cannot easily leverage knowing the the network parameters to understand relationships between features, and their impact on the prediction, as they are too numerous. Interpretation methods discussed in this section will only involve *black box* models. There are various techniques for interpreting machine learning models, which can be categorized into model-agnostic interpretations methods and specific interpretation methods. We review these two categories in detail below.

3.7.1 Model-agnostic interpretation methods

Model-agnostic methods consist in using a separate model to provide explanations. Model-agnostic methods can be applied to all type of models, regardless of their architecture or complexity. These are post-hoc techniques that explain predictions without inspecting the internal model parameters. They often create interpretable approximations or surrogate models that capture the behavior of the original model to explain, in a more transparent way. We distinguish *global* from *local* model-agnostic methods. Global methods provide explanations for the average behavior of the model, while local methods provide explanations for a specific prediction. Local explanations can also be aggregated into global methods, as detailed in Section 3.7.3.

Global model-agnostic methods

Global methods are able to describe the average behavior of the machine learning model to explain. They can output which features were important in the model construction. Examples of global model-agnostic interpretation techniques include:

- **Partial dependence plots** (PDP) [Friedman (2001)], which are a features effect plot that show the dependence between the expected prediction and a set of inputs when all other features are marginalized out.
- **Accumulated local effect plots** [Apley & Zhu (2020)] are similar to partial dependence plots but can be used when featuress are correlated.
- **Permutation feature importance** [Altmann *et al.* (2010)] measures the contribution of each feature as the increase in model error when the feature is perturbed by permutations.
- **Global surrogate models** approximate the predictions of a black box model with a simpler but more interpretable model.

Local model-agnostic methods

In contrast to global interpretation techniques, *local* model-agnostic interpretation techniques provide explanations for a specific prediction of the model. The general intuition for local model-agnostic interpretation methods is that the ML predictions in a neighborhood of a given instance can be approximated by a *white box* interpretable model. This local model must mimic the behavior of the original model within a small region around the instance of interest. The main local model-agnostic methods used for interpretation are LIME and SHAP, which we detail below.

LIME LIME (Local Interpretable Model-agnostic Explanations) was introduced in 2016 [Ribeiro *et al.* (2016)]. This method can explain individual predictions of any model by approximating an intrinsically white-box local surrogate model. The aim of LIME is to understand how the machine learning model changes when we perturb data samples. It works by generating a new dataset, consisting of perturbed samples and their corresponding predictions from the original black box model. It then trains the interpretable model on this new data samples and weights each instance by its proximity to the sampled instances. This surrogate model approximates the behavior of the black-box model locally. The surrogate interpretable model can be a sparse linear regression model, such as LASSO [Tibshirani (1996)], or a decision tree. It should be a good approximation of the black box model predictions locally.

Mathematically, local surrogate models with interpretability constraint can be expressed as follows:

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{3.23}$$

Where the explanation $\xi(x)$ for an instance x is the model that minimizes the loss $\mathcal{L}(f, g, \pi_x)$ called *locality-aware loss* in the original paper [Ribeiro *et al.* (2016)], which is a measure of how unfaithful the function g is to approximate f in the neighborhood π_x, while keeping the the surrogate model complexity $\Omega(g)$ low. In other terms, it measures how far the explanation can be compared to the prediction for the instance x of the original model fIn practice, LIME only optimizes the loss part. The user has to determine the complexity, by selecting the maximum number of features that the linear regression model may use for instance, or the depth of the tree if the explanation model is a decision tree.

SHAP Like LIME, the SHAP (SHapley Additive exPlanations) [Lundberg & Lee (2017)] is a method for explaining individual predictions. Different from LIME coefficients, SHAP for feature contributions do not directly come from a local regression model. Instead, they explain the prediction of an instance by computing the contribution of each feature to the prediction. SHAP is based on Shapley values. As shown on Figure TODO, Shapley values are a concept in cooperative game theory. They provide a way to fairly distribute the total gains among players based on their individual contributions to the overall outcome. Shapley values are useful in situations where the outcome is the result of a collaboration between multiple players. In theory, it measures the value of the contribution of each player in a coalition, and the sum of the individual Shapley values equals the total payoff for the whole coalition. In the context of model interpretation, players correspond to features, games to making predictions, and payoffs to predictions (or, more accurately, to differences between a prediction and the average prediction).

The Shapley value is defined via a value function v of players in S, S being a coalition of players, corresponding for model interpretation to a subset of the features used in the model. For a game with n players, the Shapley value for player i is the contribution payout, weighted and summed over all possible feature value combinations:

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(S)] \tag{3.24}$$

where N is the set of all players, S is a subset of N not including player i, and $|S|$ is the number of players in subset S. This contribution can be interpreted as:

Probability	Order of arrival	A's marginal contribution	B's marginal contribution	C's marginal contribution
$\frac{1}{6}$	first A then B then C : ABC	v({A}) = 40	v({A,B}) - v({A}) = 30	v({A,B,C}) - v({A,B}) = 30
$\frac{1}{6}$	first A then C then B: ACB	v({A}) = 40	v({A,B,C}) - v({A,B}) = 30	v({A,C}) - v({A}) = 20
$\frac{1}{6}$	first B then A then C: BAC	v({A,B}) - v({B}) = 40	v({B}) = 30	v({A,B,C}) - v({A,B}) = 30
$\frac{1}{6}$	first B then C then A: BCA	v({A,B,C}) - v({B,C}) = 50	v({B}) = 30	v({B,C}) - v({B}) = 20
$\frac{1}{6}$	first C then A then B: CAB	v({A,C}) - v({C}) = 40	v({A,B,C}) - v({A,B}) = 30	v({C}) = 20
$\frac{1}{6}$	first C then B then A: CBA	v({A,B,C}) - v({B,C}) = 50	v({B,C}) - v({C}) = 30	v({C}) = 20

$$\phi_i(v) = \frac{1}{\text{number of players}} \sum_S \frac{\text{marginal contribution of } i \text{ to S}}{\text{Number of coalitions excluding } i \text{ of this size}} \tag{3.25}$$

Let us consider an example of a simple game with three players A, B, and C where the value of the coalitions is given by $v(\{A,B,C\}) = 100$, $v(\{A,B\}) = 70$, $v(\{A,C\}) = 60$, $v(\{B,C\}) = 50$, $v(\{A\}) = 40$, $v(\{B\}) = 30$, $v(\{C\}) = 20$, and $v(\emptyset) = 0$.

To find the Shapley value for player A, we calculate the marginal contributions for all possible coalitions without A. Next, we average the marginal contributions, weighted by the number of permutations of the coalition sizes. We perform similar calculations for players B and C, we get their respective Shapley values. This ensures that the total value (100 in this case) is distributed fairly among the three players.

With 3 players, the Shapley value is calculated by considering all the possible orders of arrival of players and give a marginal contribution:

$$\begin{aligned}
\phi_A(v) &= \frac{1}{6}(40 + 40 + 40 + 50 + 40 + 50) \\
\phi_B(v) &= \frac{1}{6}(30 + 30 + 30 + 30 + 30 + 30) \\
\phi_C(v) &= \frac{1}{6}(30 + 20 + 30 + 20 + 20 + 20)
\end{aligned} \tag{3.26}$$

LIME or SHAP? LIME and SHAP are both popular techniques for interpretring models. The choice between them depends on many aspects including the desired robustness of explanations. Indeed, despite its multiple advantages such as its simplicity or its flexibility, its explanations have been shown to be unstable, i.e, explanations greatly differ in a small neighborhood [Alvarez-Melis & Jaakkola (2018)]. LIME works by perturbing the

input data and observing the changes in the model's output. Those perturbations can lead to variations in the local approximation model. Different perturbation samples for neighboring points result in different surrogate models, which can produce different explanations. SHAP ensure more consistency and fairness in explanations, but is still not very robust for non-linear model [Lakkaraju *et al.* (2020)].

3.7.2 Model-specific interpretation methods

In contrast to agnostic methods, *model-specific interpretation* tools are limited to specific types of model. By definition, the interpretation of intrinsically interpretable models is a model-specific interpretation method; the analysis of regression weights is specific to linear model, for instance. Many tools have been designed to improve neural networks explicability, such as DeConvNets (for convolutional neural networks) [Zeiler *et al.* (2011)], Guided back-propagation [Springenberg *et al.* (2014)], Deeplift [Shrikumar *et al.* (2017)] or Integrated Gradients (IG) [Sundararajan *et al.* (2017)]. In this discussion we will only focus on Integrated Gradients methods, which we further use in this thesis for interpretation.

The IG method is designed to attribute the prediction of a neural network to its input features. They are based on computing the gradients of the output with respect to the input, integrated over a path from a baseline input to the actual input. We denote by $f : \mathbb{R}^n \to [0, 1]$ the function that represents a neural network for a binary classification problem. Here $x \in \mathbb{R}^n$ is an input data point to the neural network, and $x' \in \mathbb{R}^n$ is a *baseline* input. The baseline x' usually represents the absence of features, or an input that is expected to have no predictive power for the model to interpret. A zero-vector is commonly used for the baseline input. We consider a segment that links x to x', which is represented by the set of all interpolated inputs points along the straight line path from x to x'. The idea is now to calculate gradients of the model f for each interpolated input with respect to the input x. These gradients indicate how changes in the input would change the output of the model:

$$\mathrm{IG}_j(x) = (x_i - x'_j) \times \int_{\alpha=0}^{1} \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i} d\alpha, \tag{3.27}$$

where j represents the index of the input features, α controls the position of the interpolated point between x and x', and $\frac{\partial f(x)}{\partial x_j}$ is the gradient of the model output with respect to feature j. In practice, this integral is approximated by computing m interpolated inputs $x_{\alpha k} = x' + \alpha_k(x - x')$.

To summarize, IG follow 4 main steps:

- Choose a baseline input x'.
- Generate m interpolated inputs $x_{\alpha k} = x' + \alpha_k(x - x')$

- Compute the gradient for each $x_{\alpha k}$
- Approximate the integral by summing all the interpolated inputs' gradients for each $x_{\alpha k}$

The final value $IG_j(x)$ represents the attribution of feature x_j to the model output y.

The IG method is widely used in neural network interpretation. It stands out from other interpretation methods for neural networks for its sensitivity, meaning that a variation in the input leads to a proportional variation in the output. It also provides consistent attributions values regardless of the neural network architecture and implementation details. All it all, it is a robust method and helps to make the black box neural networks more transparent [Sundararajan *et al.* (2017)].

3.7.3 Aggregation of local interpretations

In this thesis, we will use three main interpretation methods: LIME in the LIME library [Ribeiro *et al.* (2016)], SHAP in the SHAP library [Lundberg & Lee (2017)] , and the Integrated Gradients tool implemented in the CAPTUM library [Kokhlikyan *et al.* (2020)]. All these methods are local interpretation methods: they provide explanations for a specific prediction.

In order to gain more general insights into the models developed during this thesis, we also used methods that are able to aggregate all individual explanations into a comprehensive global explanation. Those methods are known as GALE for Global Aggregation of Local Explanations (GALE) [van der Linden *et al.* (2019)]. They involve combining the feature importances from multiple local explanations.

We count two main approaches for GALE, decribed in the original paper [van der Linden *et al.* (2019)] and used in this thesis: **Global LIME importance** and **Global Average importance**.

Global LIME Importance: In order to provide a global explanations of ML models, the authors of LIME have developed a tool that aims to select the main features that contribute to the global behavior of the model [Ribeiro *et al.* (2016)]. They propose a global feature importance denoted I_j^{LIME} and defined as follows:

$$I_j^{\text{LIME}} = \sqrt{\sum_{i=1}^{N} |W_{ij}|}, \tag{3.28}$$

where W_{ij} is the attribution value of the feature j for instance i. With this method, features with high attributions are expected to have a bigger global impact on the model's predicitions than features with low attributions values.

Moreover, features that occur more often are expected to have a higher effect on the global attribution. This is particularly problematic when features are sparse, which is the case for text inputs where features correspond to words. Therefore, we expect the global LIME attribution to be biased towards most common features in the text. To overcome this situation, the authors of GALE [van der Linden *et al.* (2019)] also propose a variant called Global Average Importance.

Global Average Importance: Global Average Importance allows features to have similar effect in all of their occurrences. The average importance denoted as I_j^{AVG} is defined as follows:

$$I_j^{\text{AVG}} = \frac{\sum_{i=1}^{N} |W_{ij}|}{\sum_{i:W_{ij} \neq 0} 1} \tag{3.29}$$

The global LIME importance is averaged over the features occurrences in the dataset. Local explanations from the cited model-agnostic interpretation methods (SHAP, LIME and IG) can be aggregated using these methods to provide global insights into the model's behavior.

3.7.4 Interpretation of transformers-based models

The interpretation of transformers-based models such as BERT has become a crucial area of research given their widespread use. Different methods can be used to this end:

- Attention mechanisms: These are the core item of transformers-based models. As described in Section 3.5.2, to this date, there are debates about the ability of attentions mechanisms to provide reliable model explanations. However, for many NLP tasks, they can help by giving insights into which part of the input sequence the model has focused for the defined task.

- Gradient-based methods: As mentioned in Section 3.7.2, they can be applied to any neural network model, including transformers based models such as BERT.

- Feature Attribution methods: Methods such as LIME and SHAP (cf Section 3.7.1) can be adapted to interpret transformer models. Indeed, they are mode-agnostic. For instance, the lime package (`https://lime-ml.readthedocs.io`) has a submodule called LimeTextExplainer which can be used for any text inputs, including BERT; TransSHAP [Kokalj *et al.* (2021)] is a variant of SHAP developed specifically for text.

3.8 Conclusion

This chapter presents different machine learning models that can be used in particular to perform binary classification of disease-free survival (DFS) status. Among them, I included more "classical" machine learning methods, such as random forests, support vector machines or logistic regression, as well as more recent models such as transformers-based models.

In this thesis, I use the "classical" models as baselines, as detailed in Chapter 5 . Furthermore, I have proposed and implemented BERT-based models adapted for the task of DFS status prediction on EHRs from Institut Curie. More specifically, I have provided a method for learning from multimodal events in the patient journey, suitable for multimodal tabular data as presented in Chapter 6, for free-text reports in the EHR as presented in Chapter 7, and their combination as a final multimodal EHR model, presented in Chapter 6.

I also use some of the methodology depicted in this chapter for a challenge I reproduced during the first part of my ph.D: The PhysioNet CinC challenge.

Chapter 4

PhysioNet challenge

Abstract:

PhysioNet is an online platform established in 1999 and managed by members of the MIT laboratory for Computational Physiology. Its mission is to help improving biomedical research and education, by offering free access to large collections of physiological and clinical data. Furthermore, PhysioNet crowdsources solutions on unsolved clinical problems by providing an annual series of challenges, in collaboration with the annual Computing in Cardiology conference. The PhysioNet CinC challenge 2012 [25] was about Intensive Care Unit (ICU) patient's mortality prediction with time series data, and it took place in August 2012.

Résumé:

PhysioNet est une plateforme en ligne créée en 1999 et gérée par des membres du laboratoire de physiologie computationnelle du MIT. Sa mission est de contribuer à l'amélioration de la recherche et de l'enseignement dans le domaine biomédical en offrant un accès gratuit à de vastes collections de données physiologiques et cliniques. En outre, PhysioNet propose des solutions à des problèmes cliniques non résolus en organisant chaque année une série de défis, en collaboration avec la conférence annuelle "Computing in Cardiology". Le défi PhysioNet CinC 2012 (REF), qui s'est déroulé en août 2012, portait sur la prédiction de la mortalité des patients des unités de soins intensifs (USI) à l'aide de données chronologiques.

Contents

4.1 Introduction

This chapter delves into the exploration of a parallel challenge that shares similarities in data structure and characteristics with the main project in this thesis. The decision to participate in this challenge was driven by the aim to not only benchmark the proposed methodologies but also to leverage insights gained from a broader context. The comparative challenge serves as a valuable opportunity to validate the robustness and generalizability of the developed models, as well as to unearth potential nuances that may influence performance.

PhysioNet is an online platform established in 1999 and managed by members of the MIT laboratory for Computational Physiology. Its mission is to help improving biomedical research and education, by offering free access to large collections of physiological and clinical data. Furthermore, PhysioNet crowdsources solutions on unsolved clinical problems by providing an annual series of challenges, in collaboration with the annual Computing in Cardiology conference. The PhysioNet CinC challenge 2012 [PhysioNet (2012)] was about Intensive Care Unit (ICU) patient's mortality prediction with time series data, and it took place in August 2012.

4.2 Challenge characteristics

Several ICU scoring systems that are widely used as clinical decision systems. Acute Physiology and Chronic Health Evaluation (APACHE II), Simplified Acute Physiology Score (SAPS II) and Sequential Organ Failure Assessment (SOFA) were designed to provide a score that will indicate an ICU patient status through the time. Each score is punctual and related to a mortality rate. However, these scores are more appropriate to account for populations differences in studies aiming to compare how medications, care guidelines, surgery, and other interventions impact mortality in ICU patients. The aim of the PhysioNet Computing in Cardiology challenge was to predict the in-hospital mortality rate in an ICU population, in a more specific way. Features used for that purpose are not only the parameters used to compute the acuity scores listed above, but also other observations including time series physiological measurements. The particularity will be to take into account the dynamic of these features, which is not done for to the current acuity scores.

4.3 Datasets presentation

4.3.1 Data characteristics

The challenge dataset is extracted from the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) II clinical database [Saeed *et al.* (2011)]. It

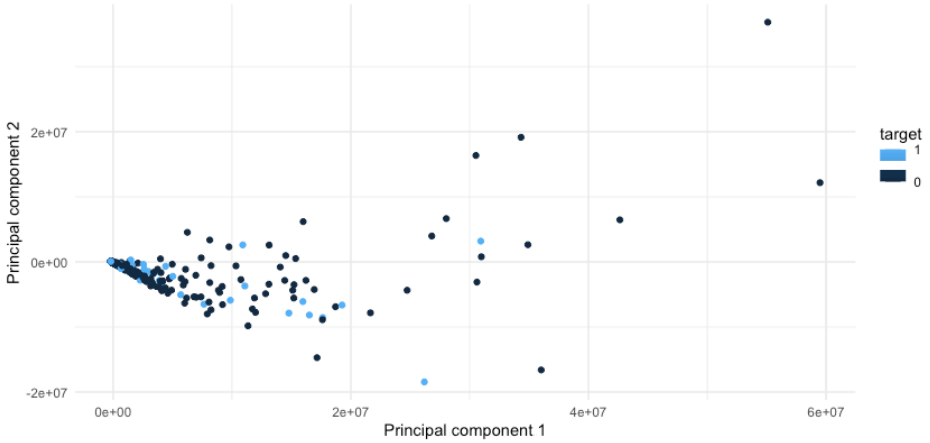


Figure 4.1: Features distribution according to the outcome.

contains 12,000 records from ICU stays. All the patients were adults (>16yo) who were admitted in different specialized ICU (cardiac, medical, trauma and surgical) for at least 48hours. Patients with DNR (do not resuscitate) or CMO (comfort measures only) order were included to the dataset [Silva *et al.* (2012)]. The dataset has been splitted by the organizers into 3 sets (training set a, open test set b and hidden test set c that will be used to evaluate challengers' models and rank participants). It describes all the physiological states during the first two days of each patient stay.

4.3.2 Available features

Up to 42 features have been collected at least once for the patients. Among them, six are vitals measurements and 37 are time series physiological measurements. Each of these physiological measurements has an associated timestamp which represents the elapsed time between the measurement and the patient ICU arrival. A timestamp of 24:20 for a measure means that the measurement was collected 24 hours and 20 minutes after the ICU admission.

General descriptors

Time-series data

Outcome

The task is a binary classification aimed at predicting in-hospital death. The outcome is skewed to the negative class (survivor), which is common when working with clinical data. All the models will be trained taking into account this imbalance.

```
RecordID: Patient ID Record
Age (years)
Gender (0: Female, 1: Male)
Height (cm)
Weight (kg)
ICUType: (Coronary Care Unit, 2: Cardiac Surgery Recovery Unit, 3: Medical ICU, 4:
Surgical ICU)
```

Figure 4.2: General descriptors.

```
Albumin (g/dL)
ALP (IU/L)
ALT (IU/L)
AST (IU/L)
Bilirubin (mg/dL)
BUN (mg/dL)
Cholesterol (mg/dL)
Creatinine (mg/dL)
DiasABP (mmHg)
FiO2 (0-1)
GCS (3-15)
Glucose (mg/dL)
HCO3 (mmol/L)
HCT (%)
HR (bpm)
K (mEq/L)
Lactate (mmol/L)
WBC (cells/nL)
Mg (mmol/L)
MechVent (0-1)
Na (mEQ/L)
NIDiasABP (mmHg)
NIMAP (mmHg)
NISysABP (mmHg)
PaCO2 (mmHg)
PAO2 (mmHg)
pH (0-14)
Platelets (cells/nL)
Resprate (bpm)
SaO2 %
SysABP (mmHg)
Temp (°C)
TropI (ug/L)
TropT (ug/L)
Urine (mL)
```

Figure 4.3: Time series data.

```
In- hospital death (0: survivor, 1: died in-hospital)
```

Figure 4.4: Binary outcome

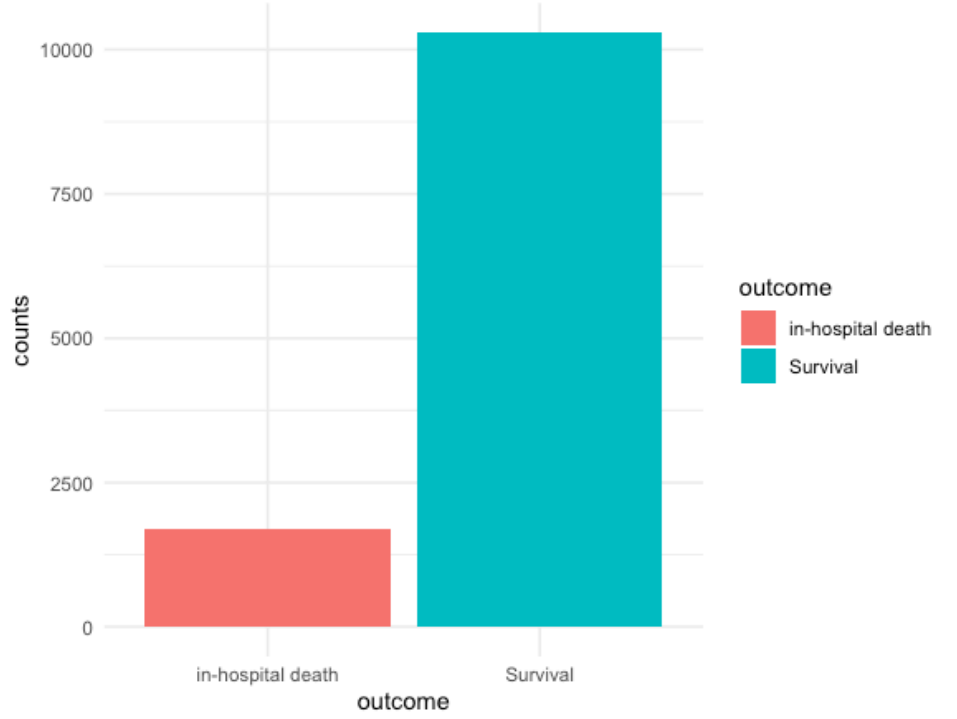


Figure 4.5: Outcome distribution

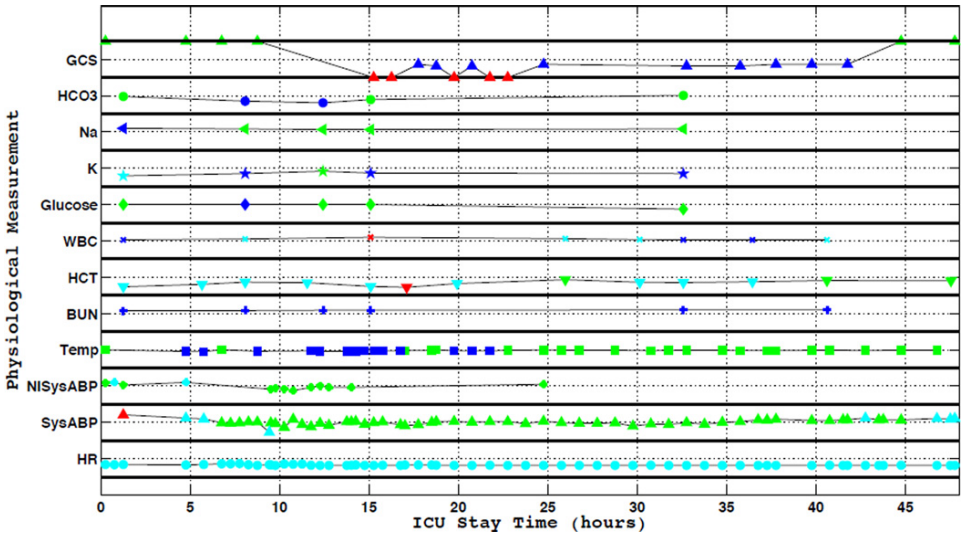


Figure 4.6: An example of ICU stay data used for the challenge [Silva *et al.* (2012)]

Model	Score	References
Bayesian Ensemble	**0.5353**	[Johnson *et al.* (2012)]
Cascaded SVM-GLM	0.5345	[Citi & Barbieri (2012)]
Logistic Regression	0.5009	[Vairavan *et al.* (2012)]
Linear Bayes	0.4928	[Macaš *et al.* (2012)]
Neural network	0.4923	[Xia *et al.* (2012)]

Table 4.1: Scores on the PhysioNet challenge

This dataset has been chosen because of its multimodality and the similarity with the data from Institut Curie that will be the focus of this thesis starting from Chapter 5. It contains vitals and biological time series data such as the Institute Curie dataset.

4.3.3 Scoring criteria

The aim of the challenge was to predict from these biological analysis results the in-hospital death. The scoring is based on 2 metrics: the sensitivity (recall) and the positive predictive (precision) which are dependent on the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). The scoring for the challenge is given by:
Score = min(Se, P+) : the minimum of the sensitivity and the positive predictivity with:
Se: the fraction of correct predictions of in-hospital deaths
P+: the fraction of in-hospital deaths that are predicted

$$Se = \frac{TP}{TP + FN} \tag{4.1}$$

$$P+ = \frac{TP}{TP + FP} \tag{4.2}$$

The challengers' ranking is set up with the set c. The highest ones (see Table 4.1) reached around 50% of good classification. This threshold is due, on the one side, to the wide variety of life- threatening conditions among patients. Similar physiological values can lead to different outcomes depending on prior or post-collect conditions. This can be observed in Fig. 10, where several points close to each other in the space still have different outputs. On the other hand, there is a high number of missing values and/ or outliers for some features that can be potentially important for the prediction.

Moreover, several patients with DNR (do not resuscitate) and CMO (comfort measures only) requirements have been kept in the dataset. In

fact, patients with DNR directives could have had a good prognosis but made the decision not to receive cardiopulmonary resuscitation (CPR) if their heart stops beating. On the opposite, patients with CMO directives may have been in a dying process, but still survive because of the "comfort care" given. This has made the prediction more complex.

Finally, the features used was collected for the 48 first hours of the ICU stay. The data are not exhaustive enough to describe the overall health trend of a specific patient. All the scores that I obtained were compared to these highest scores.

4.4 Related word

4.4.1 Features extraction

In Machine Learning, feature extraction is a process by which the initial data set is used to derive meaningful other features that can provide hidden information. It is one of the main steps of machine learning algorithms. It has many advantages such as accuracy improvements and the overfitting/underfitting risk reduction, among others. Different feature extraction techniques exist (supervised or unsupervised). They can be used differently according the dataset and the task.

For our case, I extract several statistical features, for each temporal measurement: the minimum, the maximum, the mean value, the variance, the kurtosis, the skewness, the frequency of collection, the maximum rate of change, daily trends and the entropy. Other features that can probably have a statistical meaning were also included such as the ORI (Out of range index) and the number of alarms. The ORI represents the differences of a physiological measurement's amplitude within its normal range and the time the normal range goes out of normal range (see Fig. 14). The ORI is a clinically intuitive measure as clinicians believe that the amount of time that a physiological variable is out of normal range or in a dangerous zone is as important as the number of times that it surpasses the normal limits [Sejdic (2018)]. Furthermore, it has been used for prediction with clinical time series data and the results have shown it as an excellent predictor of outcome [Jalali *et al.* (2013)]. The number of alarms will indicate the number of times that a measurement has crossed the normal range threshold.

Furthermore the Body Mass Index ($BMI = weight/height^2$) is included as a new feature and ages have been splitted into groups:
1 : < 30 ; 2 : [30-40]; 3: [40-50]; 4 : [50-60]; 5 : [60-70]; 6 : [70- 80] and 7 : > 80.
In final, twelve (12) features have been extracted from the time series data for all the records, over and above the five (5) remained general descriptors (RecordID, age_bins, gender, BMI and ICUType).

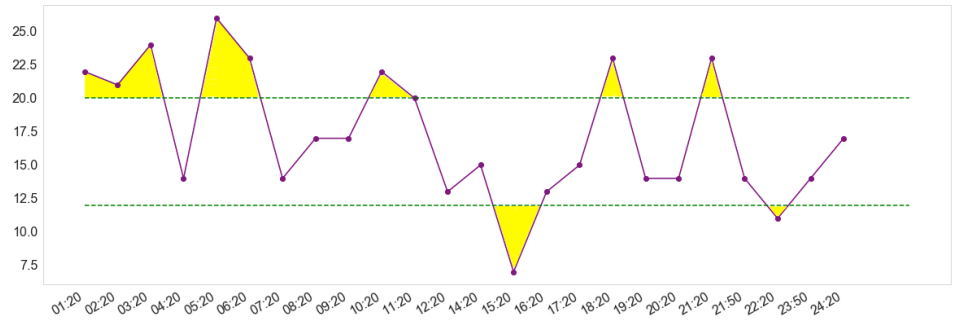


Figure 4.7: Out of range index (ORI) for the RespRate feature

4.4.2 Data preprocessing

The data processing is an important step in machine learning where raw data is transformed into a cleaned and understandable format. For this purpose, I used the z-score method to detect outliers for certain features such as the height, the weight etc. Those outliers are removed by removing the involved observation. I handle the missing data for the mean value, the minimum and the maximum by imputing them with the mean, minimum and maximum in the normal range [Sejdic (2018)] of each feature. All the other statistical features are imputed by zero (0). The categorical features (ICUType and Gender) are handled in a more specific way. I used an iterative imputer [Little (2002)] which will impute a value according to the values of the other features. The categorical features will be imputed depending the pattern of the remained features. I performed the BMI imputation with the mean value by gender.

4.4.3 Models

I have set up several models for the prediction of in-hospital death. They will be used as baselines for my own data in order to predict cancer treatment response 5. The models used for this challenge, namely logistic regression and random forests are detailed in chapter 3. Another model have been built based on random forest classifier.

Random forests classifier with physiological knowledge

In this algorithm, I used an algorithm tested on ICU data in [PhysioNet (2012)]. Clinical physiological knowledge has been integrated by the authors, in order to provide a more accurate decision support. In this case, the model apprehends the mortality prediction by grouping the different features into their belonging organ. The authors have used a deep learning algorithm: neural network and the different organ as first layer of the network.

Organ	Features
Heart	MAP, DiasABP, SysABP, NISysABP, NIDiasABP, NIMAP, HR, K, cholesterol
Neuro	GCS, glucose, MAP, SaO2
Lung	FiO2, RespRate, SaO2, PaO2, PaCO2, pH, MechVent, HCO3
Liver	Bilirubin, Albumin
Kidney	Creatinine, BUN, K, Lactate, Urine
Infection	WBC, Temp

Table 4.2: Organ classification for ICU mortality prediction

Based on expert clinical physiological knowledge, the authors have divided the features as shown on Table 4.2.

The table shows the organ and their respective associated features defined by clinicians [Sejdic (2018)]. Grouping the features into organs allows us to describe the state of each organ according to the value of the features. For my model, I set up a classifier for each organ instead of a neural network. Each organ has a classifier that will return the probability of having a positive outcome (in-hospital death). The final outcome will be a consensus vote between the different classifiers. The authors have added the infection case because it is common in ICUs. As the basic random forest, the hyperparameters that gave the best F1-score have been chosen. I used the balanced class_weight which put more emphasis on the minority class and class_weight tuned manually.

4.4.4 Results

For any binary classifier with imbalanced data, it is preferable to have high prediction score for the minority class, while maintaining a good accuracy for the majority class. To evaluate all the models developed, I will compare the precision, the recall, the f1-score and the AUC score of the minority class 4.3.

I also compared the built models scores with the highest scores in the challenge.

4.4.5 Conclusion

I presented five different models for the in-hospital death prediction with ICU time series data. All the models put more weights on the minority class, either automatically with a balanced class weight and/or by choosing the best weight for the minority class manually. Thus, they penalize more heavily on misclassifying the minority class (in- hospital death). Logistic regression model, which is the only linear model, has shown better recall for the minority class, but the lowest precision. Random forests (with physiological

Model	Precision	Recall	f1-score	AUC score
Logistic Regression	0.29	**0.67**	0.41	**0.77**
Random Forests class_weight : balanced	**0.35**	0.61	0.41	**0.77**
Random Forests class_weight : {0:1, 1:7}	0.29	0.66	0.41	**0.77**
RF + Physiological knowledge class_weight : balanced	0.34	0.51	0.41	0.74
RF + Physiological knowledge class_weight : {0:1, 1:7}	0.31	0.54	0.40	0.73

Table 4.3: Performance comparison across models

Models	min(Se, P^+)
Logistic regression	0.29
Random Forests class_weight : balanced	**0.35**
Random Forests class_weight : {0:1, 1:7}	0.29
RF + Physiological knowledge class_weight : balanced	0.34
RF + Physiological knowledge class_weight : {0:1, 1:7}	0.31

knowledge or not) have shown better precision with reasonable recalls. Compared to the highest performance in the challenge (0.5353), all the models did not reach the expected scores yet 4.3. However, the increasing precision score while adding physiological knowledge is encouraging. Following this challenge, I explore machine learning models using the Institut Curie breast cancer dataset to classify DFS status.

Chapter 5

Multi-modal machine learning models to predict breast cancer endpoints

Abstract:

Breast cancer is a complex disease that affects millions of people and is the leading cause of cancer death worldwide. There is therefore still a need to develop new tools to improve treatment outcomes for breast cancer patients. Moreover, it exists a huge amount of meaningful information in medical reports, biological measurements and clinical information in a patient journey that remain mostly unexploited. In that context, I propose to develop in my thesis, several machine learning models that use the multi-modal EHR to predict prognosis endpoints. In this chapter, I will first present the cohort that have been used for this thesis, then I will present ML models developed with different integration methods and finally I will provide results about the model interpretation.

Résumé:

Le cancer du sein est une maladie complexe qui touche des millions de personnes et constitue la principale cause de décès liés au cancer dans le monde. Il est donc toujours nécessaire de développer de nouveaux outils pour améliorer les résultats du traitement des patientes atteintes d'un cancer du sein. De plus, il existe une énorme quantité d'informations significatives dans les rapports médicaux, les analyses biologiques et les informations cliniques dans le parcours des patientes qui restent pour la plupart inexploitées. Dans ce contexte, je propose de développer dans ma thèse plusieurs modèles d'apprentissage automatique qui utilisent le DME multimodal pour prédire les caractéristiques du pronostic. Dans ce chapitre, je présenterai d'abord la cohorte qui a été utilisée pour cette thèse, puis je présenterai les modèles d'apprentissage automatique développés avec différentes méthodes d'intégration et enfin je fournirai des résultats sur l'interprétation des modèles.

Contents

In this chapter, I present different machine learning models applied to multimodal EHR using different integration methods. I also give insights about models' behavior and which features was found to be the most predictive.

5.1 Introduction

As detailed in chapter 2, breast cancer is the most commonly diagnosed cancer among women (almost 2.3 million cases worldwide in 2022) and the leading cause of cancer death [fer (2024)]. In order to enhance prognosis, its treatment strategies should be tailored to a patient's specific diagnosis and needs. Among the various treatment options, adjuvant chemotherapy is proposed after first-line surgery to lower the chance that the cancer will return. However, recurrence or death are still possible. Accurately identifying the patients most likely to relapse is therefore important to inform both treatment selection and future research to propose better therapeutic options. In healthcare delivery, predicting breast cancer relapse using machine learning techniques is a critical area of research that can be very useful for clinicians in order to better manage and treat breast cancer patients. Moreover, Electronical Health Records (EHRs) serve as a valuable source of data of patient data. They contain a wealth of meaningful information, from pathological reports to biological measurements, that remains unexploited. The more recent improvements in machine learning models allow us to come up with innovative and efficient methods to use this information to improve patient care.

In fact, through the years, multiple machine learning tools have been developed to improve breast cancer patients' treatments outcome. Among those, models that use one or multiple modalities within electronic health information. Some works focused on building machine learning models that are capable of detecting almost all true positive regarding breast cancer relapse. By using patient's clinical information and ensemble methods (AdaBoost and cost-sensitive learning) [Yang *et al.* (2021)] achieved a high sensitivity rate at 94.7% with a cohort of 1061 breast cancer patients from Shin Kong Wu Ho-Su Memorial Hospital between 2011 and 2016. For [Alzu'bi *et al.* (2021)], the use of a bagging classifier allows to reach a sensitivy of 92.3%. These models can serve as a supportive aid during follow-up visits for both early-intervention and advanced treatments, contributing to the reduction of cancer mortality rates.

Over the years, other modalities have been included into these studies, with the intention of combining different information for a better prediction performance. [Yao *et al.* (2022b)] integrated in a multi-modal deep learning prediction model histopathological image, clinical information and gene expression data for 196 breast cancer istances from the Cancer Genome Atlas.

Their method acheived a AUC score of 75% and was capable of capturing the multimodal aspect of EHR.

However, these modalities are not always available for all patients treated. For this reason, other authors have taken advantage of the considerable information present in medical reports that constitute the Electronic Health Records of patients [Zeng *et al.* (2019b), González-Castro *et al.* (2023a)]. [Zeng *et al.* (2019b)] developed a support vector machine to identify breast cancer local recurrences using concepts extracted from text reports by MetaMap, and the number of pathological reports recorded for each patient. Indeed, there is a need of using a multimodal approach to predict breast cancer relapse, combining clinical, pathological and molecular information. Their model achieved a high AUC of 93% in cross-validation. For [González-Castro *et al.* (2023b)], medical concepts are also extracted from reports to constitute features that will be combined to clinical information. In the 5-year cancer recurrence their best model (eXtreme Gradient Boosting) reached a great AUC of 80.7%.

Ongoing research continues to develop machine learning models to enhance our understanding of breast cancer mechanisms and improve prognostic models, and ultimately aiming to provide more effective care for patients. Data scientists use multimodality into their studies more and more, as it makes it possible to have a broader view of the disease mechanism. However, few studies combine structured information from EHR and medical records in a free-text format. In that context, I propose, in this thesis the integration of those information (clinical data, biological measurements and free-text reports) in multimodal machine learning models to predict prognosis. In this chapter, I present the data from Institut Curie that I have been working with, as well as multimodal models based on classical machine learning approaches. The following chapters will present deep learning models. I also propose predictive factors from these models interpretation that can be considered as potential multimodal prognostic factors after more investigations.

5.2 Construction of a data set of breast cancer patients for breast cancer disease free survival prediction

5.2.1 Data sources

In this work, we used databases extracted from the Electronical Health Records (EHR) system from Institut Curie in Paris (France). All data collected were pseudonymized. Additionally, individuals under 18 years of age, with a history of previous cancer, under guardianship, or unable to provide consent were excluded from this cohort. Every patient included in the study

has completed and signed a research informed consent form. The study was approved by the Breast Cancer Study Group of Institut Curie and was conducted according to institutional and ethical rules concerning research on tissue specimens and patients.

The first database (SEIN database) contains patient-level biological and clinical longitudinal information, for patients treated with adjuvant chemotherapy for breast cancer from 2005 to 2012. It contains 15 150 unique patients, male and female. More specifically, the SEIN database includes, for each date at which they were measured, tumor markers used to monitor treatment, namely levels of cancer antigens (CA15-3 and CA19), prostate specific antigen (PSA), cytokeratine fragment (CYFRA), angiotensin-converting enzyme (ACE) and neuroson-specific enolase (NSE). It also contains other immune markers, such as counts of lymphocytes (LYMP), monocytes (MONO), leukocytes (LEUK), neutrophils (PN), basophils (PB) and eosinophils (PE),

Moreover, it contains general descriptors of patients (such as age, sexe, or weight), medical background, as well as diagnosis and treatment information. Finally, the patients are annotated with survival and recurrence information. The clinical information includes 162 features in total.

In this work, we focus on disease-free survival (DFS) as a binary endpoint. At a given time after surgery, DFS is defined as the absence of either death, loco-regional recurrence or distant recurrence. 90.7% of the patients in the SEIN database have a positive DFS status, making it a highly imbalanced data set to work with.

In addition, free-text visits notes for all admissions in Institut Curie are stored in a EHR system. This data refers to unstructured narrative descriptions or notes entered by healthcare professionals. Unlike the structured data, which is organized into predefined fields, free text allows healthcare providers to input progress reports and relevant patient information recorded during patient journey, in a more natural manner. Free text reports from cytopathology or radiology also capture key information from medical images, as captured by experts. Those medical reports comprise free-text clinical notes for consultations, as well as free-text reports of cytopathology, radiology, surgery, and blood tests. All reports are written in French. For this study, we selected medical reports from the patients of the SEIN database, hence creating a retrospective cohort containing both structured information and free text for each patient.

5.2.2 Ethics

This study was conducted in accordance with institutional and ethical rules, and French regulations, regarding the use of patients data for scientific research purposes. The study was reviewed and approved by the Institutional Review Board of Institut Curie (Paris, France). The patients/participants provided their written informed consent to participate in this study.

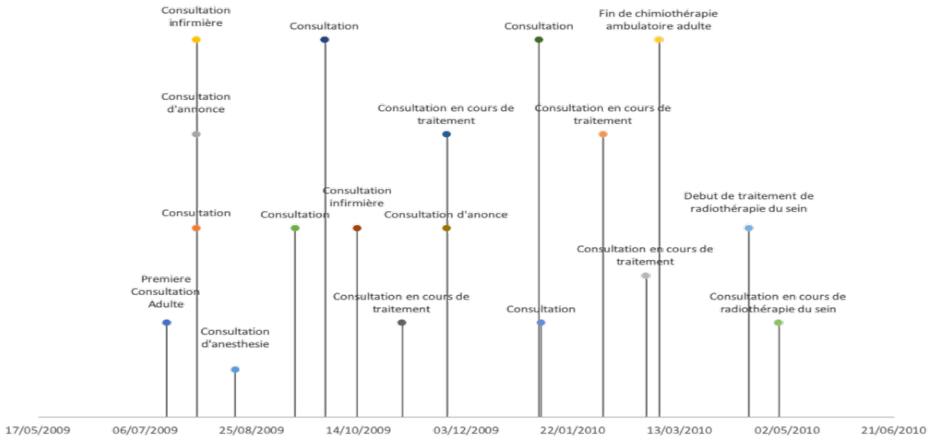


Figure 5.1: Example of procedure timeline for a patient

As already mentioned, all the used data are pseudonymised under the responsability of The Institut Curie's Data department. The patient/identifier correspondence table is built and stored by the same department. Moreover, the Data department provides a secure space, which can be upgraded according to the needs of the study, in terms of environment, power and tools made available.

Finally, the study was authorized by the French data protection agency (CNIL, under declaration number 1023665)

5.2.3 Data preprocessing and Data engineering

Clinical features

Clinical features are already in a tabular format, a suitable format for classical machine learning models. One of the main challenge in medical data remain the missingness in values for certain features. In this study, within the 162 features, we only kept the ones with less than 30% of missing values, which corresponds to 36 features.

The preprocessing steps applied to the clinical features from the SEIN database include the removal of redundant information and duplicate rows, the creation of new features (BMI for instance), the imputation of missing values for the remaining features, and the correction of outliers for continuous features (height for instance: 1766cm to 176cm). First the duplicate rows, the redundant features (according to prior knowledge and to a correlation matrix) and dates are dropped from the data. Then, we perform feature categorization for multiple features (tumor size into clinical and pathological T stages, number of lymph nodes into clinical and pathological N stages, the Body Mass Index (BMI) into 5 classes: 1, $\leq$18.5|2, 18.5-24.9|3, 25-29.9|4,

30-34.9|5, ≥35). The following step is to handle the remained missing values. We built different imputation strategies depending on the features type. For categorical features, we impute the missing values with an aberrant value 999. In fact, in medical data, missingness can be random, when for example a information is missing due to a technical error. It can be caused by an unobserved data itself, when the patient change facility in the middle of the care journey. Sometimes, the missingness also provide an information, when for instance PSA (prostate specific antigen) is only measured for men. Therefore, imputing with an aberrant value give an information about the feature and the sample. The continuous features are imputed with the mean value (age at menarche, clinical and histological tumor sizes). The missing BMI are filled with the BMI mode class according to the gender. Regarding the menopausal age, I replace missing values by the mode value (50) if the menopausal status is positive and 999 otherwise. The final input for machine learning models will be a table of 25 cleaned features and 15150 samples.

Biological features

As clinical data, biological features are in a structured format. We also kept features that have at most 30% of missing values, which corresponds to 3 features: CA15-3, LEUK, MONO as shown in Figure 5.2. Biological features are time-stamped measurements, that is to say, they are sequential features. In order to integrate them in a machine learning model, we computed statistical features from the remaining features: the mean value, the maximum value, the minimum value, the variance, the number of measurements, the number of alerts, the ORI feature (I refer the reader to Chapter 4 for more details about the Out of Range Index feature), the entropy, the skewness, the kurtosis and the maximum delta (dmax). The number of alerts refers to the number of times the feature value is out of the normal range (defined as in Table 5.1). The entropy provides a measure of uncertainty of a distribution. The kurtosis describes the shape of the distribution's tails and the skewness measures the asymmetry of a distribution.

The preprocessing pipeline applied to biological features starts by duplicates removal. I fill missing values of the mean, min and max features with respectively the mean, min and max values from the normal range interval of each feature. We assume that the absence of values for biological features is related to the lack of necessity for measurement. Therefore, we fill the missing values with their supposed normal values. For the remained features (the number of measurements, the number of alerts, the ORI feature, the entropy, the kurtosis, the skewness and the maximum delta), we impute with 0.

The biological input for the models will be a table of 45 features for 8998 samples.

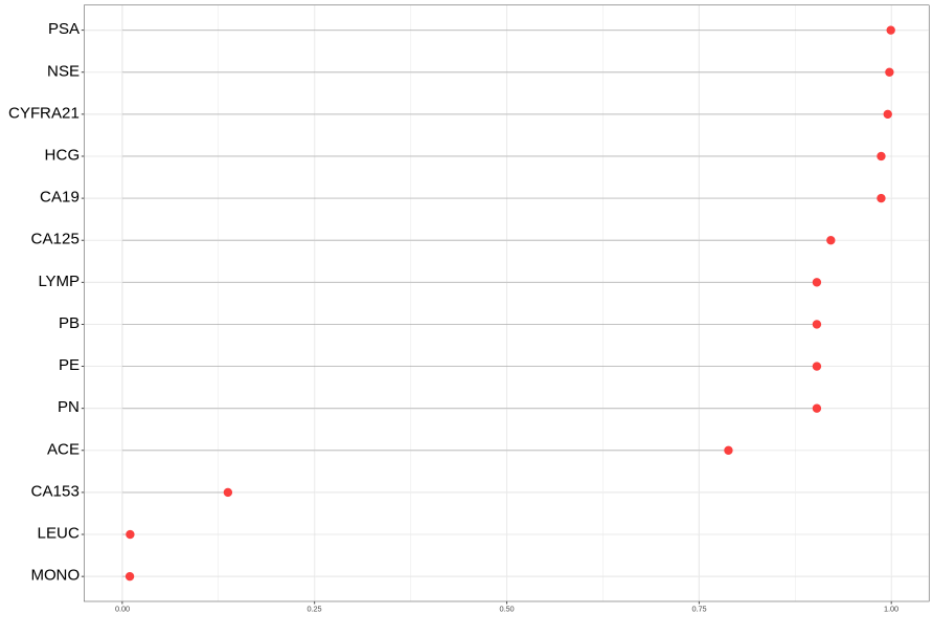


Figure 5.2: Percentages of collection for each feature

Feature	Normal range	Mean value ± std	missing
CA15-3 (U/mL)	$N < 30$	19.73 ± 27.84	1 236
LEUK (g/L)	$4 < N < 10$	6.81 ± 2.94	92
LYMP (g/L)	$1.4 < N < 4$	1835.19 ± 644.91	8 126
MONO (g/L)	$0.2 < N < 1$	42.81 ± 144.48	88
ACE (μg/L)	$N < 3.4$	2.99 ± 16.75	7 092
PN (g/L)	$1.7 < N < 7$	$4\,391.71 \pm 2\,161.66$	8 126
PE (g/L)	$N < 0.5$	136.84 ± 123.81	8 126
PB (g/L)	$< N < 0.2$	29.22 ± 18.65	8 126
CA125 (U/mL)	$N < 35$	22.78 ± 49.62	8 290
CA19 (U/mL)	$N < 35$	75.79 ± 681.22	8 879
HCG (ng/mL)	$N < 10$	$255.30 \pm 2\,836.84$	8 879
CYFRA21 (ng/mL)	$N < 3$	1.39 ± 1.40	8 956
NSE (ng/mL)	$N < 15$	13.67 ± 3.88	8 976
PSA (ng/mL)	$N < 4$	3.07 ± 3.62	8 992

Table 5.1: Normal ranges for the biological features

Free-text reports

Free-text reports represent unstructured textual descriptions of medical information recorded by medical experts. They can be clinical notes, that is to say, information recorded during patient encounters with clinicians, or reports made by specialists (laboratory biologists, radiologists, histopathologists) to interpret the results of medical exams. Unlike tabular data, that is recorded in a standardized way at least within a hospital, medical reports are highly variable, as they allow each healthcare provider to be distinctive in format, style, or terminology.

The semantic related to the medical field is complex, using abbreviations, acronyms, and medical jargon [Grossman Liu *et al.* (2021)]. Therefore, in addition to common NLP preprocessing steps (normalization, removal of noisy entities, adverbs, stopwords and text delimiters), the Text BEHRT preprocessing pipeline includes steps that are specific to medical reports, such as removing proper nouns or correct abbreviations among others.

Removing proper nouns is one of the key step of the preprocessing pipeline. This is important as specific doctor names may serve as proxy for the DFS classification, for example, when a doctor mostly handles severe cases. Patient names are already excluded from the reports, which had been anonymized before we accessed them. The first stage of this process consists in using part-of-speech tagging to remove proper nouns tags that follow titles such as *Dr*, *M.* ("Mr" in English), *Mme* ("Mrs" in English). However, proper nouns may appear without a title. We thus further constructed a list of proper nouns to remove from the text. We first built a list of names of Institut Curie's health practicioners, obtained through the public directory of practicioners [Cur (1 30)] as retrieved in 2023, and therefore only partially matching practicioners that were involved in the care of patients in the 2005–2012 period covered by our cohort). We additionally considered surnames given at least 30 times in France from 1891 to 2000 (n=218 912) and first names given at least 20 times from 1946 to 2022 in France (n=36 964), as provided by Institut National de La Statistique et des Etudes Econonomiques (INSEE) ([Ins (1 30)], [INS (1 30)]). We then removed from this list the proper names that correspond to disease names, such as Paget.

One other main difficulty that occur with free-text reports is the high number of typos. To address this issue, we used the pyspellchecker spell checking algorithm [Barus (2023)] which identifies, for each word of the corpus that is not found in a given dictionary, the most likely correct replacement for this presumably misspelled word. More specifically, the spell checker generates, for each word of the corpus that is not found in a given dictionary, a list of candidate words based on the Levenshtein Distance [Levenshtein (1966)] (based on single-character edits calculations: insertions, deletions, replacements, or transpositions) which are potential corrections

for the unknown word. Finally, the spellchecker selects the candidate word that is both within an acceptable Levenshtein Distance and has a higher frequency of occurrence in the language, thus increasing the likelihood of providing the correct replacement for the misspelled word. For effective spellchecking, it is crucial to have a rich dictionary that contains medical jargon. Therefore, we augmented the French vocabulary from OpenSubtitles [Lison & Tiedemann (2016)] (implemented by default in pyspellchecker) with the contents of the French open dictionary Usito [ush (1 30)], as well as the 3 184 words from a French online medical dictionary [Thomsen (1 30)], the CAS corpus of French clinical cases [Grabar *et al.* (2018)] which contains over 397 000 word occurences, a list of drug names in French [vid (1 30)], and two lists of French medical abbreviations specific to oncology [moz (2020), Poletto (2023)]. If, following this step, any words from the dictionary remain unidentified, we replaced them with the most likely correct spelling suggestion from Wikipedia [wik (1 30)].

The full text preprocessing pipeline is described on Figure 5.3.

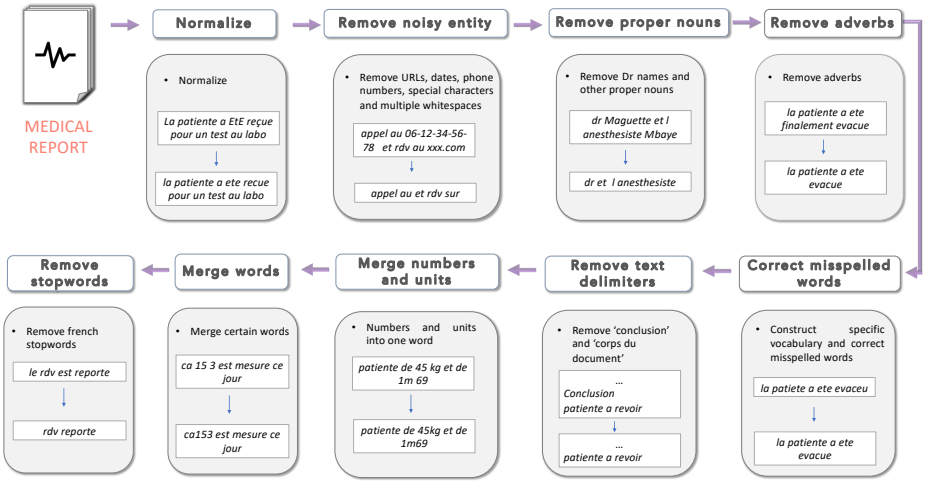


Figure 5.3: Medical text preprocessing pipeline

For the DFS prediction task, using multimodal data, we derive vectors from the free texts using the TF-IDF (Term Frequency-Inverse Document Frequency) of bi-grams and tri-grams. Bi-grams and tri-gras are respectively groups of 2 and 3 consecutive words taken from the text. If we consider the sentence : "The cat sat on the mat", the bi-grams are: ("The", "cat"), ("cat", "sat"), ("sat", "on"), ("on", "the"), ("the", "mat") and the tri-grams are: ("The", "cat", "sat"), ("cat", "sat", "on"), ("sat", "on", "the"), ("on", "the", "mat"),. We used bi-grams and tri-grams instead of single words (unigrams) because it can capture more context and meaning from the text. For instance, it allows to capture the meaning "not good" that the individual

words as independent entities "not" and "good" cannot capture. The next phase of the free-text vectorization is the use of the TF-IDF method. This method is used to evaluate the importance of a word in a document relative to a collection of documents (a corpus). In other terms, by using the TF-IDF for the free-text reports, we assign for each sample, a score to each word of each report relative the whole free-text information of the sample throughout the patient journey. TF-IDF is defined by two components: the Term Frequency (TF) which measures how frequently a term appears in a document and the Inverse Document Frequency (IDF) defined as the measure of the importance of a word within the entire corpus.

$$TF(t,d) = \frac{\text{Number of times term } t \text{ appears in document } d}{\text{Total number of terms in document } d} \tag{5.1}$$

$$IDF(t,D) = \log\left(\frac{\text{Total number of documents in corpus } D}{\text{Number of documents containing term } t}\right) \tag{5.2}$$

$$TF\text{-}IDF(t,d,D) = TF(t,d) \times IDF(t,D) \tag{5.3}$$

For the text modality, we thus obtain as final input a table of 1756 features representing the total number of unique bi-grams and tri-grams in the whole corpus, for 15 150 samples.

Moreover, we derived from the free-text reports a fourth modality that we called "Frequency_of_events", describing, for each possible medical procedure, how often it appears in the patient's history. The procedure's name is available in every medical reports' headers. This new dataset counts for each sample the occurrences of the unique procedure. We have a table of 123 procedures as events for the 15 150 patients for the machine learning models.

We assess two binary classification tasks: disease free survival (DFS) 3 years after surgery (called T1) and 5 years after surgery (called T2), using patient history up to one year after first surgery and starting from 6 months before the breast cancer diagnosis is made as shown in figure 5.4. This choice of one year after the first surgery as an index data ensures that we use as much of the patient's history as possible, without capturing an actual relapse. We removed patients who relapsed before the index date, as well as patients censored before 3 (resp. 5) years after the first surgery. This test set contains 520 patients, with a proportion of positive samples similar to that of the whole dataset: 6.1% and 11.9% samples with negative DFS status, respectively. We used a cross validation to assess the performance of models, typically the k-Fold Cross Validation method with k = 5. The best cross-validation APS allows to choose the best model during the Random Search in the hyperparameter tuning.

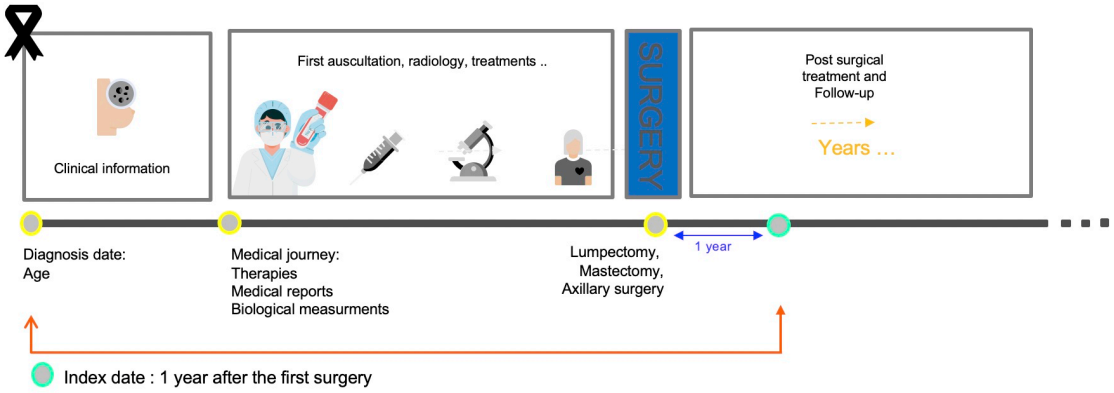


Figure 5.4: Index date definition

5.3 Machine learning methods

5.3.1 Models

Models that are used in this chapter are machine learning models described in 3. These are random forests (RF)and FFNN.

To address class imbalance, we used sampling methods in combination with Random Forests, such as the Synthetic Minority Over-Sampling Technique (SMOTE) [Chawla *et al.* (2002)] which is a popular method used for generating synthetic samples from the minority class to balance the class distribution, and Balanced Random Forests methods [Chen & Breiman (2004)], which is an adaptation of the Random Forest algorithm that aims to handle imbalancedness by balancing the classes in each bootstrap sample before training the tree. Moreover, we used two different integrations methods for these multimodal data 3.6: early integration, where all the modalities are concatenated and used as one input for the 3 models, late integration, where the modalities are modeled separately and a weighted majority vote is used to set the final prediction. We choose to use early integration to evaluate how the model capture relationships between different modalities, and late integration to allow the different models to be tailored to each modality's unique characteristics. We can compare these two strategies to see which method fits the most. Importantly, these two different approaches allow to find a balance between model complexity and model efficiency.

5.3.2 Interpretation methods

We defined a set of predictive features using the feature importance scores from the Random Forests models. These methods offer a global interpretation of models. We also used SHAP values which provides the impact measure of each feature of individual predictions.

The top features identified using these different techniques are provided, and their impact is evaluated through the Area Over the Perturbation (AOPC) score. This score measures how well a feature contributed to the model's predictions by removing them and observing the impact on the model's performance. It is defined as follows: Let us note the original performance of the model on the test set as Score($\mathbf{X}$), where $\mathbf{X}$ is the original dataset. We identify the top-k features based on their importance scores. For each j in the range from 1 to k:

- We perturb the top-j important feature to create a new dataset $\mathbf{X}_j$. We can perturb it by changing its value to 0 or to an aberrant value.
- We measure the new model's performance Score($\mathbf{X}_j$) on the perturbed dataset.
- We calculate the performance drop for perturbing the top-j features is $\Delta\text{Score}_j = \text{Score}(\mathbf{X}) - \text{Score}(\mathbf{X}_j)$.

The AOPC score is then the average performance drop over all perturbation steps:

$$\text{AOPC} = \frac{1}{k}\sum_{j=1}^{k}\Delta\text{Score}_j$$

5.4 Results

5.4.1 Model performance

Early integration

The first presented plot, Figure 5.5, illustrates the comparative performance of a random forests classifier, SMOTE random forests, Balanced Random Forests and the FFNN.

We compared the random forests classifiers with the different sampling methods with the best FFNN in the following table 5.2. Among the different sampling strategies and the random forest classifier, the random forest classifier shows the highest performance in terms of f1-score. The synthetic minority class samples used for the SMOTE RF did not provide more valuable information and diversity within the minority class. Moreover, the FFNN showed lower scores than the remained models, which demonstrates the robustness of random forests classifiers over neural network for data and this given task.

The AUC scores are similar across all the random forests classifiers. However, as it is reflecting with the other metrics, the Random forests classifiers outperform the neural network. For further experiments, we will use the balanced random forest as the best early integration for T1 and T2, as it achieves the highest validation APS.

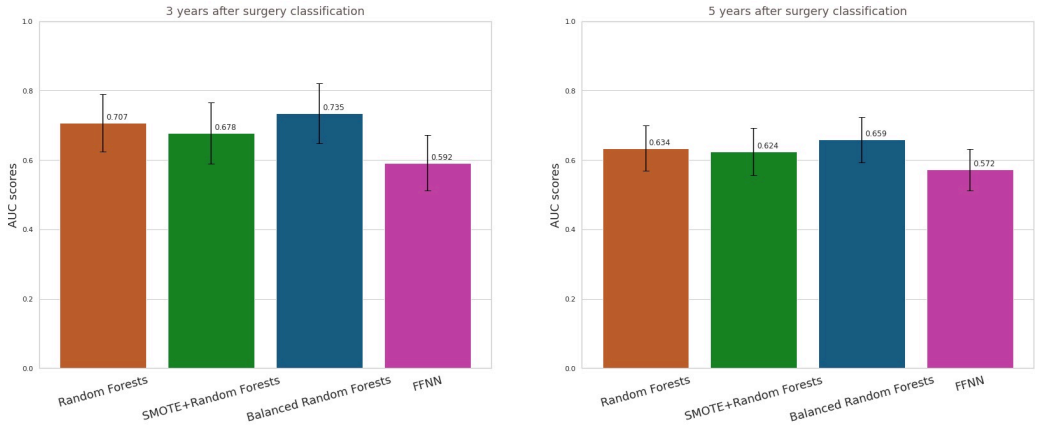


Figure 5.5: ROC-AUC scores for random forests models (with different sampling methods) using early integration method, compared with the best FFNN.

Late integration

Regarding the Late Integration method, we built the previously described models for each of the modality. We aggregated the best model predictions from each modality. The model with the higher validation APS is used to perform late integration. For task T1, those models are the Balanced Random Forests for biological and text features, random forests classifiers used on over-sampled training set (SMOTE) for the clinical data and the random forests classifier for the frequency of procedures modality. Regarding the task T2, they include random forests for biological and text features, balanced random forests for the frequency of events data and random forests for over-sampled (SMOTE) clinical data. The higher the validation APS is for a model, the higher its prediction will be weighted for the majority vote to find the final prediction (see table 5.3). Scores for each modality are shown in figure 5.6 and their aggregation are shown in figure 5.7.

Overall, clinical and text related features (medical reports and frequency of procedures) lead to the best performance compared to biological data. The late integration scores are shown in Table 5.4; ROC curves are plotted on Figure **??** in the appendix. When combining the different modalities, the score have slightly improved when compared to the each separate modality's model. By allowing each model to contribute with its corresponding weight, the late integration leverages the strenghts of each modality more effectively.

We compare the performance of the different integration methods. In terms of AUC scores, both of the methods achieve the same performance 5.8, while in terms f1 score, the late integration method performs better than the early integration method 5.5. Globally, the comparable AUC scores suggest that the core predictive information is being leveraged similarly in

	Scores for T1		
Models	Precision	Recall	f1 score
Baseline = RF	0.5707	0.579	**0.574**
SMOTE RF	0.538	**0.642**	0.548
Balanced RF	**0.703**	0.573	0.559
FFNN	**0.617**	0.536	0.463
	Scores for T2		
Models	Precision	Recall	f1 score
Baseline = RF	0.591	**0.616**	**0.600**
SMOTE RF	0.570	0.604	0.579
Balanced RF	**0.617**	0.577	0.577
FFNN	**0.551**	0.530	0.433

Table 5.2: Scores comparison for early integration for T1 and T2

	Modalities	**Biological features**	**Clinical features**	**Frequency of events**	**Text data**
T1	Best model	Balanced Random Forests	SMOTE + RF	Random Forests	Balanced Random Forests
	Validation APS	0.133	0.227	0.222	0.218
	Weights	0.17	0.28	0.28	0.27
T2	Best model	Random Forests	SMOTE + RF	Balanced Random Forests	Random Forests
	Validation APS	0.173	0.273	0.315	0.290
	Weights	0.16	0.26	0.3	0.28

Table 5.3: Late aggregation weights for each modality for T1 and T2.

	Scores for T1		
Models	Precision	Recall	f1 score
Biological data	**0.538**	0.509	0.412
Clinical data	0.534	**0.603**	0.543
Text data	**0.647**	0.550	0.517
Procedures occurrences	**0.591**	0.562	0.571
	Scores for T2		
Models	Precision	Recall	f1 score
Biological data	0.528	0.532	**0.534**
Clinical data	0.530	**0.563**	0.529
Text data	0.584	**0.593**	0.588
Procedures occurrences	**0.618**	0.573	0.565

Table 5.4: Scores comparison for the best model of each individual modality for both tasks T1 and T2.

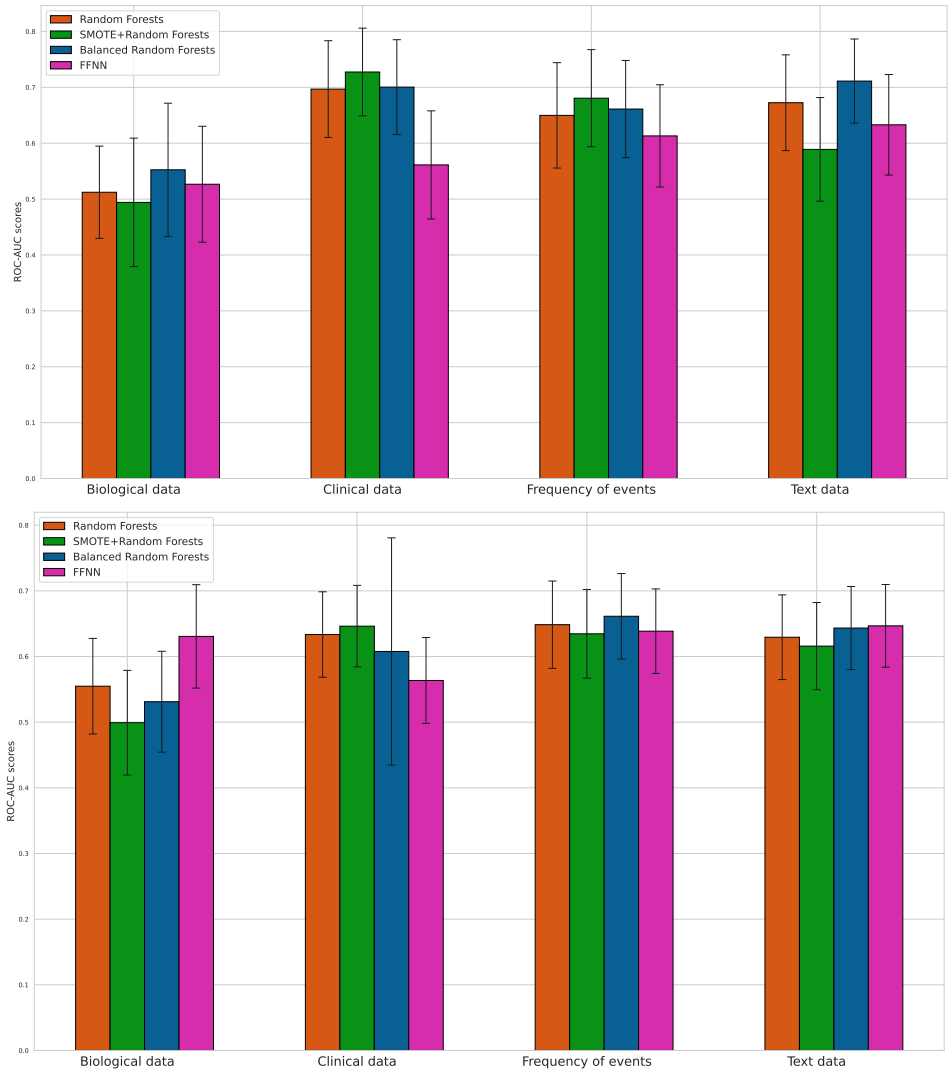


Figure 5.6: ROC-AUC scores for each modality and for each model for T1 and T2 (top and down respectively)

both methods. Moreover, multi-modal models' performance are better than all the individual models' performance for each modality taken separately, which indicates that integrative models take advantage of the complementarity of the different modalities. When combining multiple modalities, we can capture more comprehensive information about the DFS status prediction. This uses various aspects of the patient's EHR, and helps the model to discern patterns that might not be evident when using single modalities.

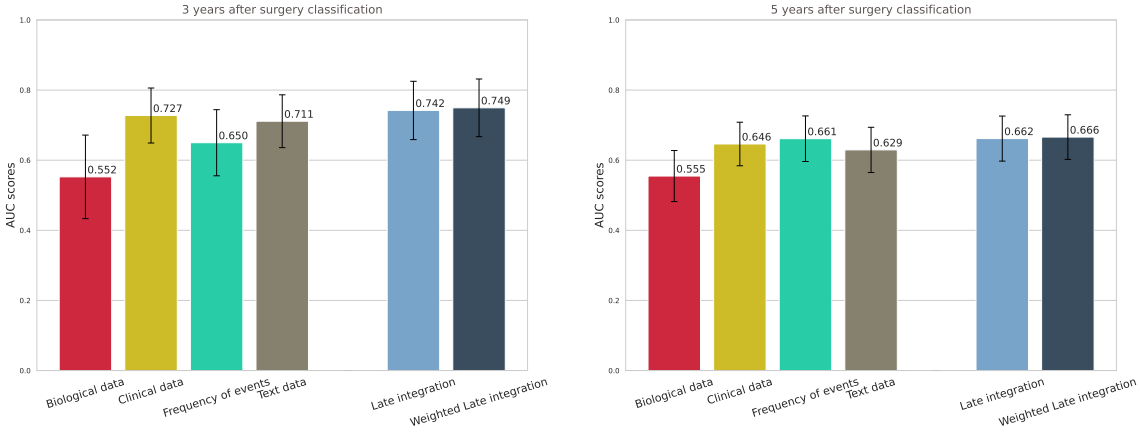


Figure 5.7: AUC scores per modality, for their late integration and their late integration weighted by their validation APS

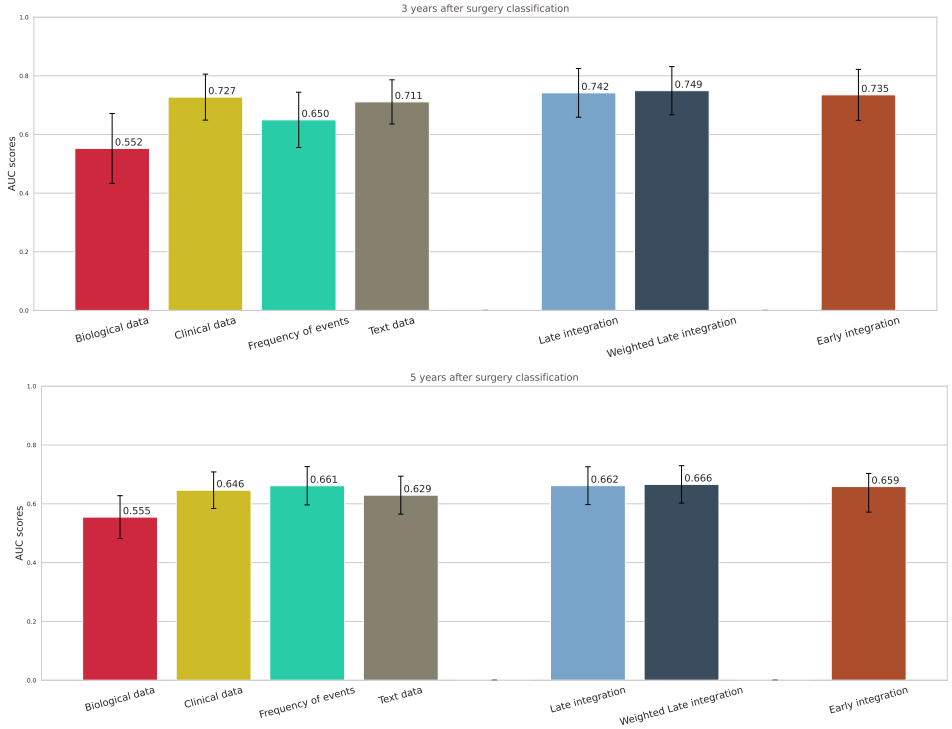


Figure 5.8: AUC scores for individual modalities and early vs late integration, for T1 and T2. Each modality model is mentioned in 5.4.1

	Scores for T1			
Models	**Precision**	**Recall**	**f1 score**	**AUC**
Late integration	0.616	0.660	0.633	0.742
Weighted Late integration	0.620	0.689	**0.645**	**0.749**
Early integration	0.703	0.573	0.559	0.735
	Scores for T2			
Models	**Precision**	**Recall**	**f1 score**	**AUC**
Late integration	0.588	0.607	0.595	0.661
Weighted Late integration	0.5917	0.615	**0.600**	**0.665**
Early integration	0.617	0.577	0.577	0.659

Table 5.5: Performance of early integration and late integrations for T1 and T2. Models used for the different integration methods are detailed in 5.4.1 and 5.4.1.

5.4.2 Interpretation

Early integration

We display the random forest's most predictive features using early integration method for both tasks T1 and T2 in Figure 5.9. It highlights key insights derived from the feature importance analysis. First, by comparing the feature importance across modalities, we see that most of the features in the most predictive ones are from the clinical data modality. This is in line with the individual modality performance, where models for clinical data have outperformed the model for the other modalities. Specifically, we find that the features that have the highest importance scores among the clinical features are related to well documented clinical prognosis factors, namely *'nbggpos'* which correspond to the number of affected lymph nodes, *'grade_3cl'* which is the tumor grade, and *'histo_size' and 'tclin'*, which are features related to the tumor size.

On the other hand, we find bi-grams and tri-grams such as *'antecedents carcinologiques' (carcinological history)*, *'scintigraphie osseuse' (bone scan)* or *'suites operatoires simples' (simple operating sequences)* from the textual features that give insights on the patient clinical and pathological condition. For instance, the bone scintigraphy is part of the standard work-up following diagnosis of a breast lesion but may also be requested for surveillance purposes, or in response to suspicious pain or elevated markers. Familial cancer history is also been shown to be a significant prognosis factor for breast cancer outcomes [Song *et al.* (2017), Lafourcade *et al.* (2018)].

Regarding the frequency of events and the biological data, no feature have been found to be predictive for the random forests model.

For the next interpretation results, I use SHAP to compute global explanations for all the samples of the test set. Figure 5.10 shows the 20

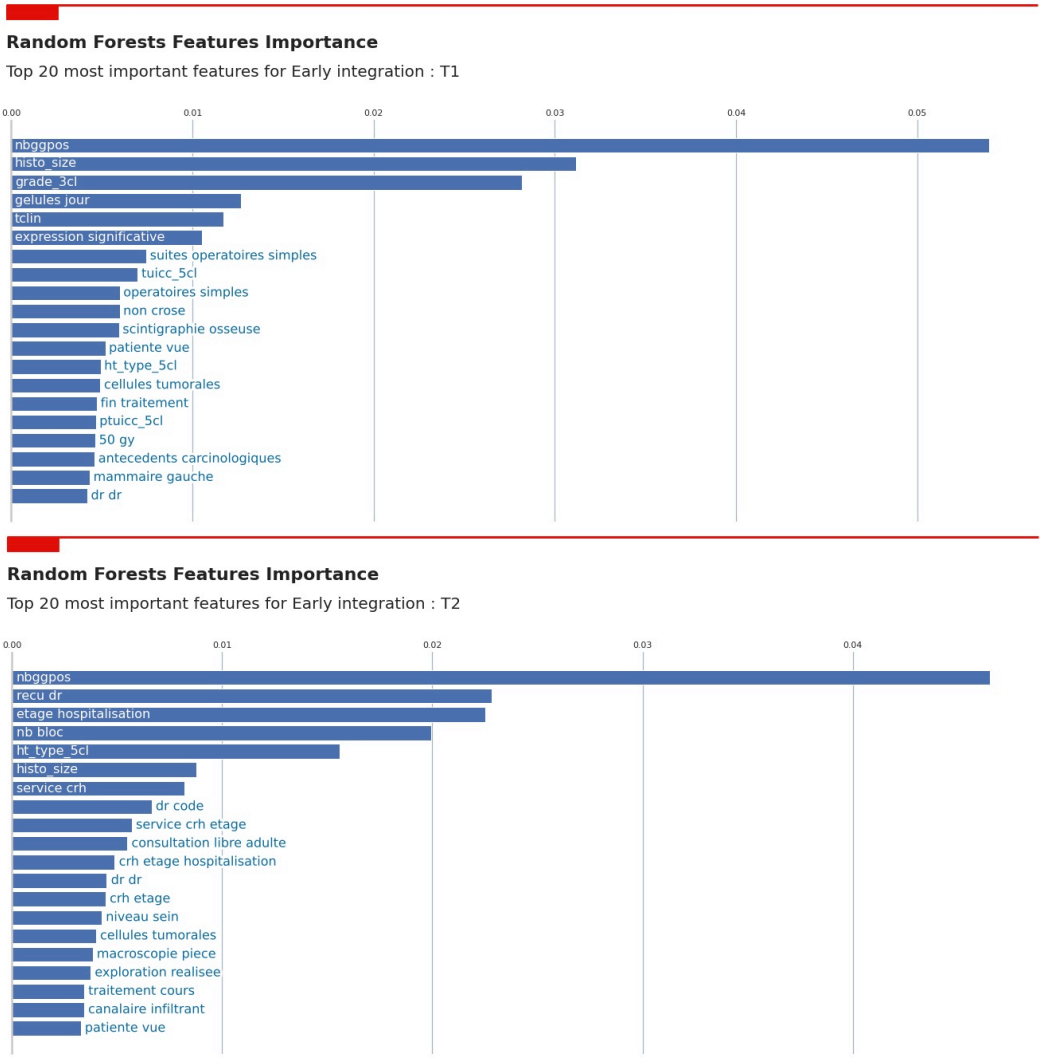


Figure 5.9: Predictive Features according to the Random Forests' Feature Importance method for Early integration method (T1 and T2)

most important features given by the SHAP interpretation model. Overall, we find common features that are outputted by the Random forest feature importance method: *'grade_3cl'* , *'nbggpos'*, *'histo_size'*, *'antecedents carcinologiques'*, *'suites operatoires simples'* or *'scintigraphie osseuse'*.. Moreover, we have features such as *'mitotique faible' (low mitotic)* for T1 or the frequency of *'consultation libre adulte' (adult consultation)* for T2, that highlight respectively a slower growth rate and a favorable prognostic factors in various types of cancers and a potential correlation between the number of consultations with the severeness of the breast cancer.

To give global insights on the most important features given by the early

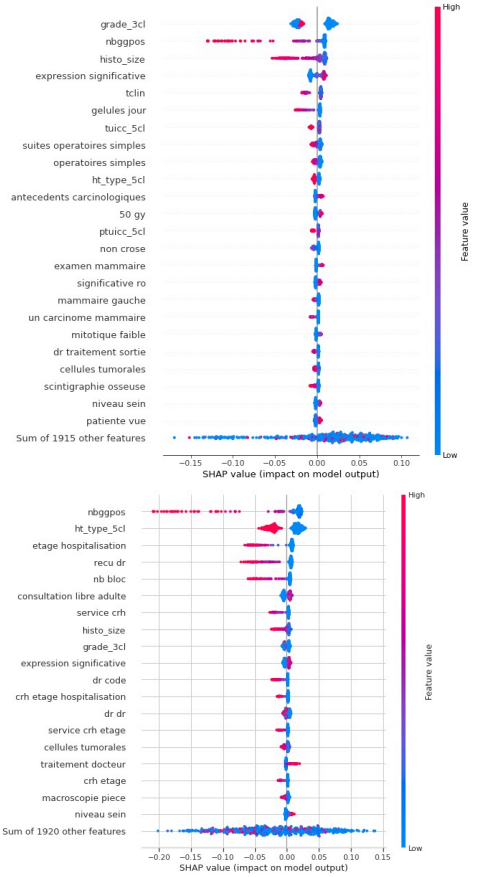


Figure 5.10: Top 20 most important features from SHAP for both tasks T1 (top) and T2 (down) for early integration.

integration method, I plot common important features from both interpretation method in Figure 5.11.

Late integration

We output the most important features for each modality's model used to perform late integration. Those features from the Random Forest feature importance method are shown in the section A.2 in the appendix. We compute the most important features across the separated results using their attribution values for the different interpretation methods. The top 20 are shown in figure 5.12 for the Random forest feature importance method and in figure 5.13 for the SHAP method.

Unlike the early integration most important features, the most importances features given by the different interpretation methods include biological features *(MONO, LEUC and CA15-3)*. Abnormal values factors can

Figure 5.11: Early integration: Top features across the different interpretation methods. From top to down: features and attributions from the random forest feature importance method, features and attribution from SHAP and their mean.

be assimilated to a post-surgical complications or infections. We also find factors that are already mentioned in the early integration method (nbggpos, histo_size, tclin or grade_3cl).

To give global insights on the most important features given by the late integration method, I also plot common important features across both interpretation methods in Figure 5.14.

Early and Late integration

We gather the top features given by all the interpretation method for both early and late integrations. We find respectively six (6) and four (4) features that have been found predictive using both methods and for both integration methods for T1 and T2 (see Figure 5.15). For T1, clinical features play an important role in predicting early relapse, while for T2, models used features related to post-treatment conditions such as the number of consultations from the frequency of events modality or *recu dr* which also refers to an occurrence of consultations.

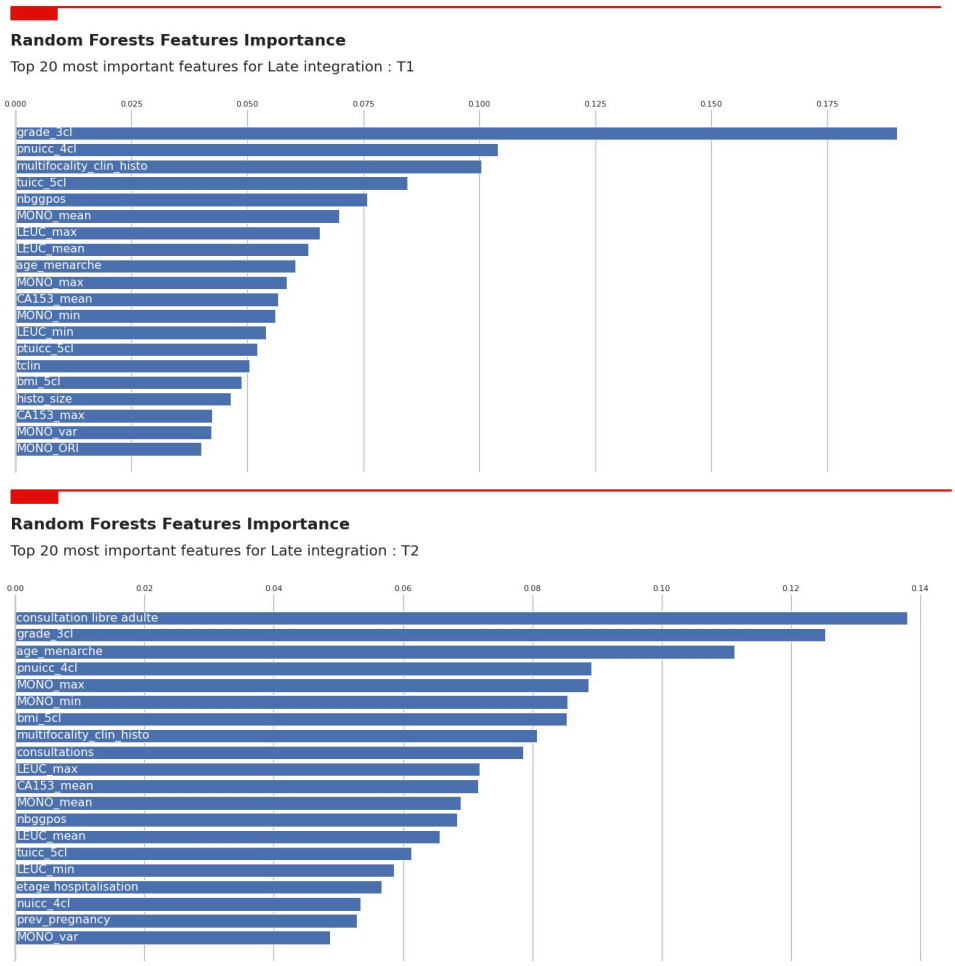


Figure 5.12: Predictive Features according to the Random Forests' Feature Importance method for Late integration method (T1 and T2)

Area Over the Pertubation Curve - AOPC

AOPC is measured using the top features for T1 and T2. We compared the results with baselines that removed random features to perturb the data. Removing the top features predicted by SHAP and RF feature importance method degrades performance. This means that these features do indeed have a non-negligible impact on the model. Therefore, these approaches seem to indeed find relevant features that explain the model.

5.5 Conclusion

We used different integration methods using four modalities of EHR to predict DFS status. In general, the different fusion methods perform simi-

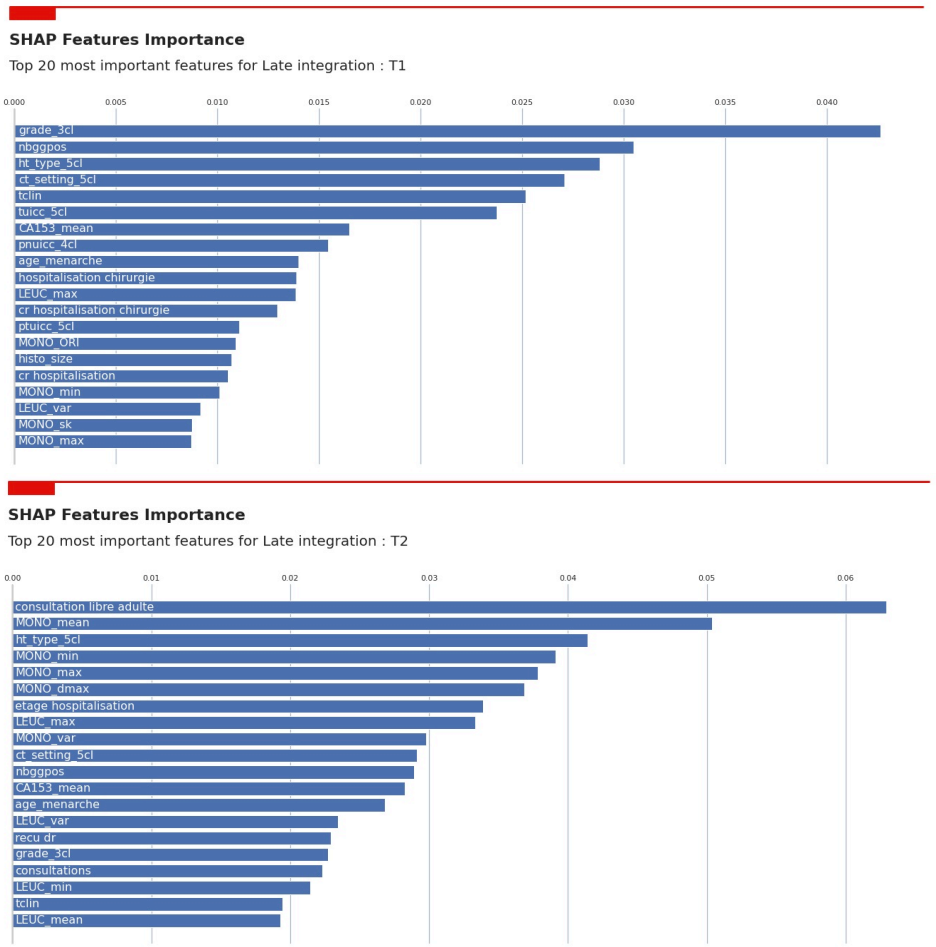


Figure 5.13: Predictive features according to SHAP for the late integration method (T1 and T2)

larly. They both manage to combine valuable information from the different modalities at their different level to achiveve the prediciton task. Moreover, the top features are reliable from a medical point of view and will require further investigations or more complex data representation to have a broader view of the explanations.

In the next chapter, I will introduce another way to represent multimodal patient data that better reflects the EHR nature: a sequential representation. This data will be used with deep learning models that are efficient with sequential represented data: transformers.

Figure 5.14: Late integration: Top features across the different interpretation methods. From top to down: features and attributions from the random forest feature importance method, features and attribution from SHAP and their mean.

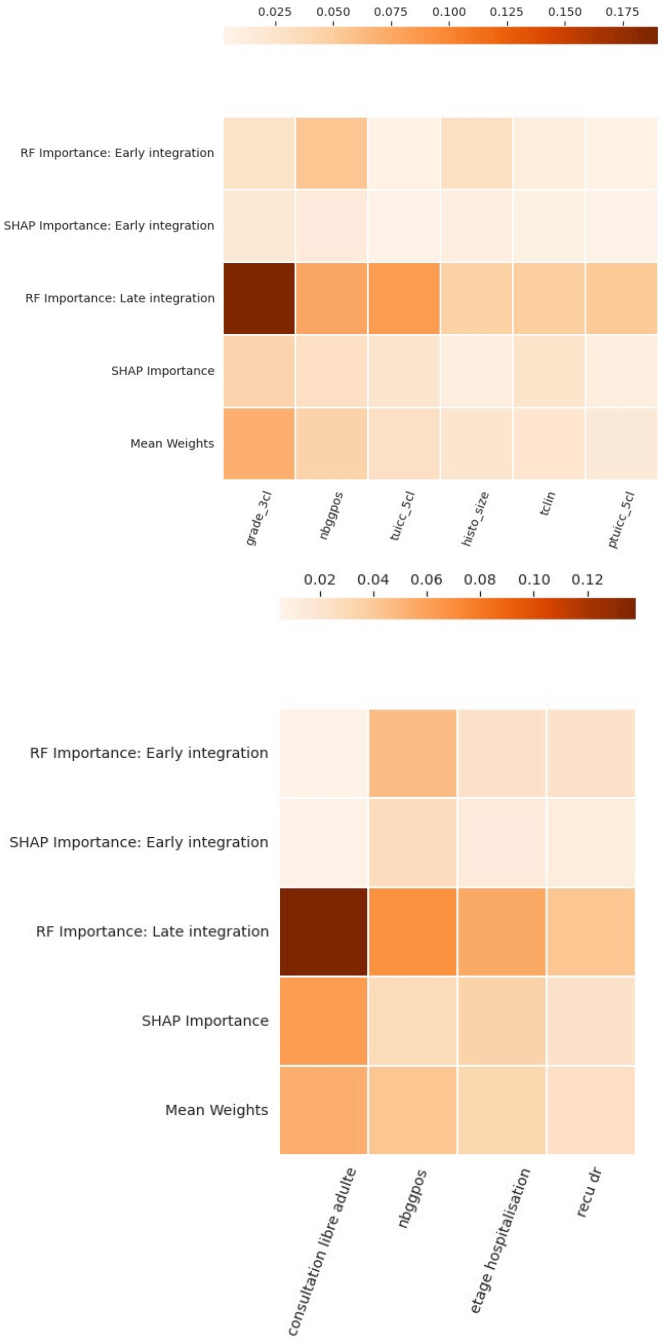


Figure 5.15: Top features from all methods and using Early and Late integration methods for T1 (top) and T2 (down).

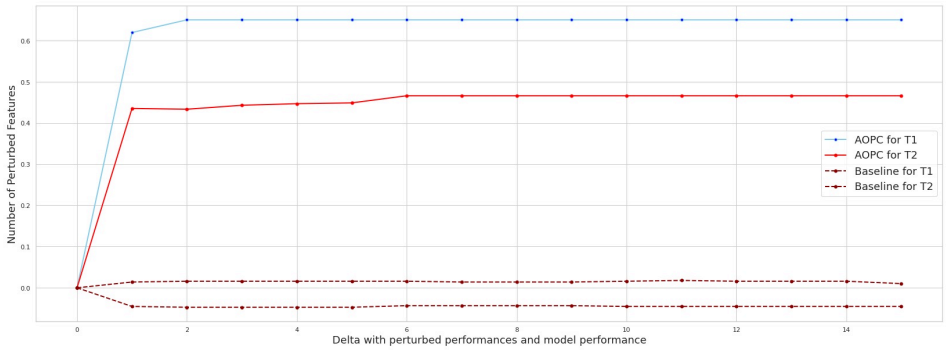


Figure 5.16: Evaluation of the interpretation methods used with the early integration models for T1 and T2. We used the 15 most important features that are common for both interpretation methods.

Chapter 6

Tabular BEHRT: Pre-trained transformers models for tabular EHR

Abstract:

Classical machine learning techniques have shown their ability to predict cancer outcome such as disease free survival status (DFS status) for multimodal breast cancer patient data. During my thesis, I also explore more complex machine learning models: transformers based models. I developed Tabular BEHRT which is a readaption of a deep neural sequence transduction for electronical health records called BEHRT using as input data the sequential information through the patient journey. The BEHRT model is inspired by one the most powerful transformer-based model in Natural Language Processing: BERT. In this chapter, I will present the readaptation of medical events into a sequential format. Then I will present whole pipeline of Tabular BEHRT. And I will finally give insights on the model interpretation.

Résumé:

Les techniques classiques d'apprentissage automatique ont montré leur capacité à prédire les résultats d'un cancer tels que le statut de survie sans maladie pour les données multimodales des patientes atteintes d'un cancer du sein. Au cours de ma thèse, j'ai également exploré des modèles d'apprentissage automatique plus complexes : les modèles basés sur les transformers. J'ai developpé Tabular BEHRT, qui est une adaptation d'un modèles basé sur l'architecture des transformers et utilisé pour des dossiers medical electroniques (DME) appelé BEHRT. Ce modèle s'inspire de des modèles les plus puissants dans le domaine du traitement du langage naturel: BERT. Dans ce chapitre, je présente la réadaptation des dossiers patients en séquence d'évènements. Ensuite je présente le pipeline de Tabular BEHRT et enfin je montrerais des résultats d'interpretation du modèle

Contents

In this chapter, I present Tabular BEHRT, a BEHRT-based model applied to tabular information form Electronic Health Records. Contrary to the previous chapter, I explore representations of the tabular patient data in a *sequential* format that depicts the patient journey over time. More specifically, I show how to adapt specific features as sequences of events, and how to train transformers-based models using such data.

6.1 Introduction

In breast cancer research, accurately identifying the patients that are most likely to relapse is important to inform both treatment selection and future research to propose better therapeutic options. One of the most commonly used prognostic tool for breast cancer is the Nottingham Prognosis Index (NPI), which uses a combination of three clinical features (tumor size, tumor grade, and number of lymph nodes) and was proposed in 1982 [Haybittle *et al.* (1982a)]. Since then, many authors have used statistical and machine learning algorithms to build breast cancer relapse predictors from clinical features; however NPI still seems to be the most robust criterion [Phung *et al.* (2019)], despite its limitations.

In addition to the NPI or the classical machine learning models such as those presented in the previous chapter, recent methodological developments in deep learning have opened the way to developing new tools to use EHR data to improve patient care. Indeed, deep learning techniques have proven useful to model complex patient trajectories based on multimodal EHR data [Amirahmadi *et al.* (2023)]. In these models, information about different time points in the patient trajectory are flattened together. By contrast, a growing body of literature is taking advantage of the sequential nature of EHRs, using deep learning architectures such as long short-term memory (LSTM) networks to capture patient trajectories as a sequence of ordered time-stamped events [Liu *et al.* (2018), Amirahmadi *et al.* (2023)].

Among those, transformer-based models inspired from BERT (Bidirectional Encoder Representations from Transformer) [Devlin *et al.* (2019c)], an architecture that has significantly outperformed previous methods on a large variety of natural language processing tasks ands continue to drive advancements in the field (see Chapter 3), have recently gathered a lot of interest. Their superiority is explained by the use of self-supervised pretraining tasks, such as masked language modeling and next sentence prediction, which allows them to learn better representations of the data. These architectures have been successfully transposed to patient trajectories by seeing them as sequences of medical events rather than of words [Li *et al.* (2020b), Pang *et al.* (2021b), Rasmy *et al.* (2021), Rao *et al.* (2022), Li *et al.* (2023b)]. To the best of our knowledge, however, none of these have considered cancer-related clinical outcomes, possibly because they are typically applied to very

Feature	Normal range	Mean value ± std	missing
CA15-3 (U/ml)	$N < 30$	63.39 ± 484.44	6 390
LEUK (g/l)	$4 < N < 10$	6.99 ± 6.82	2 525
PN (g/l)	$1.7 < N < 7$	$718.85 \pm 1\,789.66$	9 419
LYMP (g/l)	$1.4 < N < 4$	289.63 ± 714.26	9 448
MONO (g/l)	$0.2 < N < 1$	33.29 ± 123.59	3 675

Table 6.1: Normal ranges for the biological features

large cohorts of millions of patients.

In this chapter, we present a new transformer architecture for binary classification from multimodal EHR data, which combines biological measurements, therapies, and medical reports into a sequence of medical events describing a patient's trajectory. We evaluate our proposed method on two classification tasks: the prediction of relapse after 3 and 5 years, respectively. We pretrain the models on the equivalent of a masked language model.

6.2 Materials and Methods

6.2.1 Data description

The data used in this work is from the cohort described in the previous chapter; I refer the reader to the Section data in chapter 5. However, features used for this model are described in the following table.

6.2.2 Data preprocessing

From biological measurements, we only kept features that have less than 30% of missing values: MONO, LEUK, LYMP, PN and CA 15-3. Transformers require categorical inputs; hence all numerical values have to be discretized. We binarized biological measurements into two values: 1 if the value is outside the normal range for the biological measurement, and 2 otherwise. Figure 6.1 shows the distribution of biological measurements; the medical normal range of these biological features can be found in Table 6.1.

In addition, we also computed the differences $\Delta_t = v_t - v_{t-1}$ between the current visit's biological value v_t and the previous visit's value v_{t-1}. We then discretized the Δ values by dividing them by ten and rounding. This captures more subtle variations in biological measurements evolution than the mere abnormal/normal values.

From the clinical information, we included both longitudinal and non-longitudinal features: age, undergone therapies, and tumor size on the one hand, tumor grade and number of nodes involved at diagnostic as

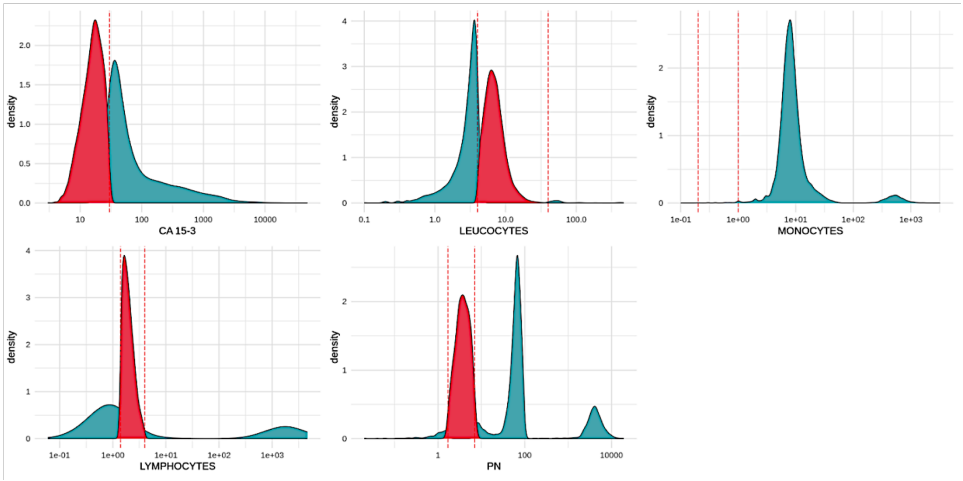


Figure 6.1: Binarization of biological features into 1 and 2. For each of the 5 biological features, the dashed red lines delineate the normal range, highlighted in red, and mapped to 2, from the abnormal range, highlighted in green, and mapped to 1

well as breast cancer molecular subtypes (Luminal, TNBC, HER2+/RH-, HER2+/RH+) on the other. Age is computed at each visit and discretized by rounding to the nearest integer. Descriptive statistics of the age, breast cancer subtype, grades, number of lymph nodes involved, tumor size and biological measurements are given in Table S1 in the appendix.

We combined tumor size, tumor grade and the number of lymph nodes involved into the Nottingham Prognosis Index (NPI) [Haybittle *et al.* (1982a)], a commonly used, clinically relevant and robust prognostic tool [Phung *et al.* (2019)]. The NPI is computed as NPI $= 0.2 \times$ tumor_size (cm) $+$ tumor_grade $+$ lymph_nodes_stage, where the lymph nodes stage is computed as 1 (0 nodes), 2 (1 to 3 nodes) or 3 ($>$ 3 nodes). The lower the score, the higher the chance of survival 5 years after surgery. The tumor size is measured at various points in the cancer journey. We kept for this study the clinical tumor size (clinical_ts) assessed at diagnosis when the tumor is palpable, and the pathological tumor size (pathological_ts) which is the histological size of the tumor extracted at the surgery. The NPI is recalculated with each new tumor size measurement, hence termed as the dynamic NPI (dNPI). For patients with at least one available feature among the three required for calculating the dNPI, we imputed missing tumor sizes using the mode value among samples of the same clinical or pathological tumor stage (TNM) status. The number of involved lymph nodes is the sum of the number of affected sentinel nodes and axillary nodes. We imputed missing number of nodes to zero and missing tumor grade to G2 (grade 2), based on the most frequent values in our data. The higher the dNPI, the lower the

Therapies	Sub-therapies
Surgery	Lumpectomy
	Mastectomy
	Axillary node dissection
	Sentinel node biopsy
Radiotherapy	Axillary irradiation
	Internal mammary chain irradiation
	Mammary gland/chest wall irradiation
	Supra/sub-clavicular irradiation
Hormone therapy	Tamoxifen
	Aromatase
	LHRH agonist
Anti-HER2 therapy	Trastuzumab
	Pertuzumab
	Lapatinib

Table 6.2: List of possible therapies and sub-therapies in our data.

chance of survival. Following Blamey et al. (2007) [Blamey *et al.* (2007)], we categorized dNPI into six groups: Excellent Prognostic group (EPG) (NPI $\leq$ 2.4), Good (GPG) (2.4 $<$ NPI $\leq$ 3.4); Moderate I (MPG I) (3.4 $<$ NPI $\leq$ 4.4), Moderate II (MPG II) (4.4 $<$ NPI $\leq$ 5.4), Poor (PPG) (5.4 $<$ NPI $\leq$ 6.4) and very poor (VPG) (NPI $>$ 6.4). We display the survival curves according to the value of clinical and pathological NPI in the figures 6.2

Therapies are inferred by considering the occurrence date for the surgery, the start and end dates for hormone-therapy, chemotherapy and anti-HER2 treatment, and the number of doses administered for the radiotherapy. This inference incoporates the therapeutic protocol of Institut Curie (see Figure 6.3). Subtherapies, also inferred from this protocol, provide additional information about the specific molecules given in the case of chemotherapy or anti-HER2 therapy, radiation types in the case of radiotherapy, and specific surgical procedures including both breast and axillary surgeries. A list of all possible values for the "therapies" and "subtherapies" field is given in Table 6.2.

Because Tabular BEHRT can handle missing values (see Section 6.2.3), we did not impute missing values for longitudinal features. However, for the baselines, we opted to impute the tumor size, number of nodes, grades and cancer subtype by an aberrant value of 999. Using an aberrant value allows the model to explicitly indentify and differentiate imputed values from the actual data, by analogy with not locating a token within a sentence when using M-BEHRT.

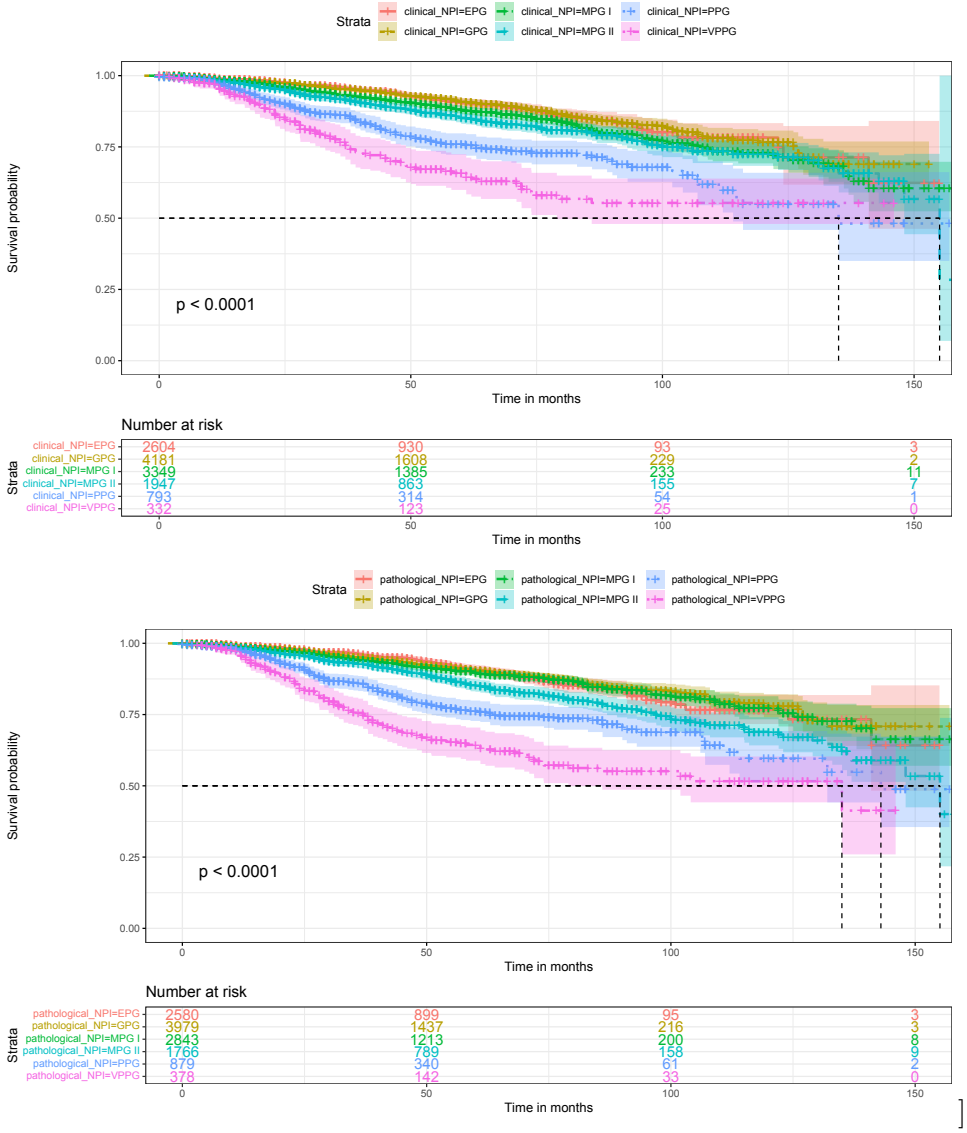


Figure 6.2: Survival curves showing the number of surviving patients at successive time points following breast cancer diagnosis for the different NPI groups (clinical NPI on the left and pathological NPI on the right), in the SEIN cohort (N=15150). The y-axis represents the probabilty of survival, ranging from 0 to 1, while the x-axis represents time. The worst NPI prognosis group reflects the curve that drops more quickly (VPPG group for both), which indicate a higher rate of the event: the Disease Free Survival here

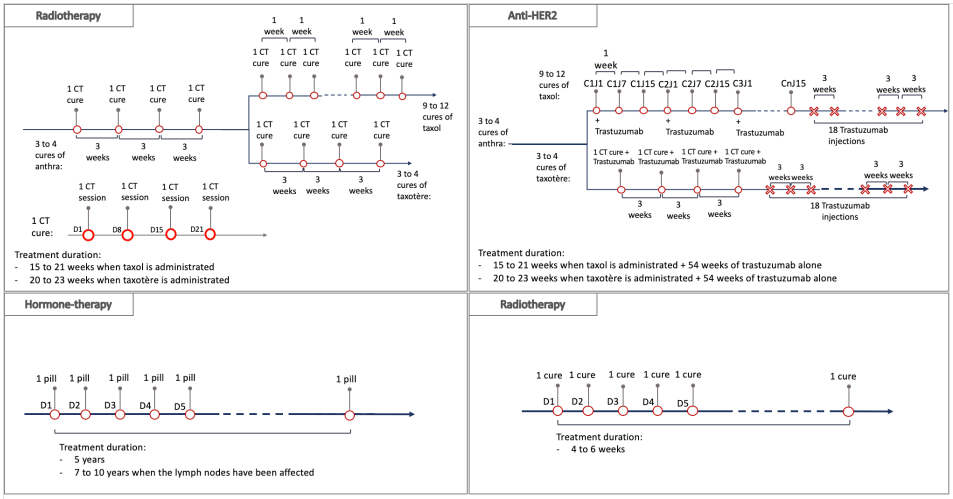


Figure 6.3: Institut Curie Therapeutic Protocol

Finally, medical visit department and procedure names are available within the headers of free-text reports. They represent the hospital's department from which the report is written, such as *consultations*, *consultation génétique* or *imagerie* (consultations, genetic consultations and imaging in English), and the procedure stands for the type of event that the report is about, for example *réunion de concertation pluridisciplinaire*, *information* or *échographie* (multidisciplinary consultation meeting, medical information and ultrasound in English). Thus, this information is used to describe the events that occur during each visit, and is extracted from each report for every patient. We normalized department and procedure names by removing accents, punctuation and special characters. We merged synonyms into a single word: for example, *anapath*, *anatomopathologie* and *anatomo-cyto-pathologie* are merged into *anatomo-cyto-pathologie* (anatomical cytology in English). To do so, we sifted through the corpus vocabulary, identifying and unifying synonyms and/or differently written terms to enhance coherence of the medical history. We also removed words that appear fewer than 100 times in the whole corpus.

We defined two classification tasks: the prediction of DFS 3 years (T1) and 5 years (T2) after surgery, as depicted in Figure 6.4, which are the same tasks as the previous chapter 5. We kept patients that had at least 3 visits in their medical history. For pre-training tasks (see Section 3.5.3), we used all patients and their full history.

This results in 8 089 patients for T1 and 5 192 for T2, with respectively 6.2% and 17.1% of negative DFS status.

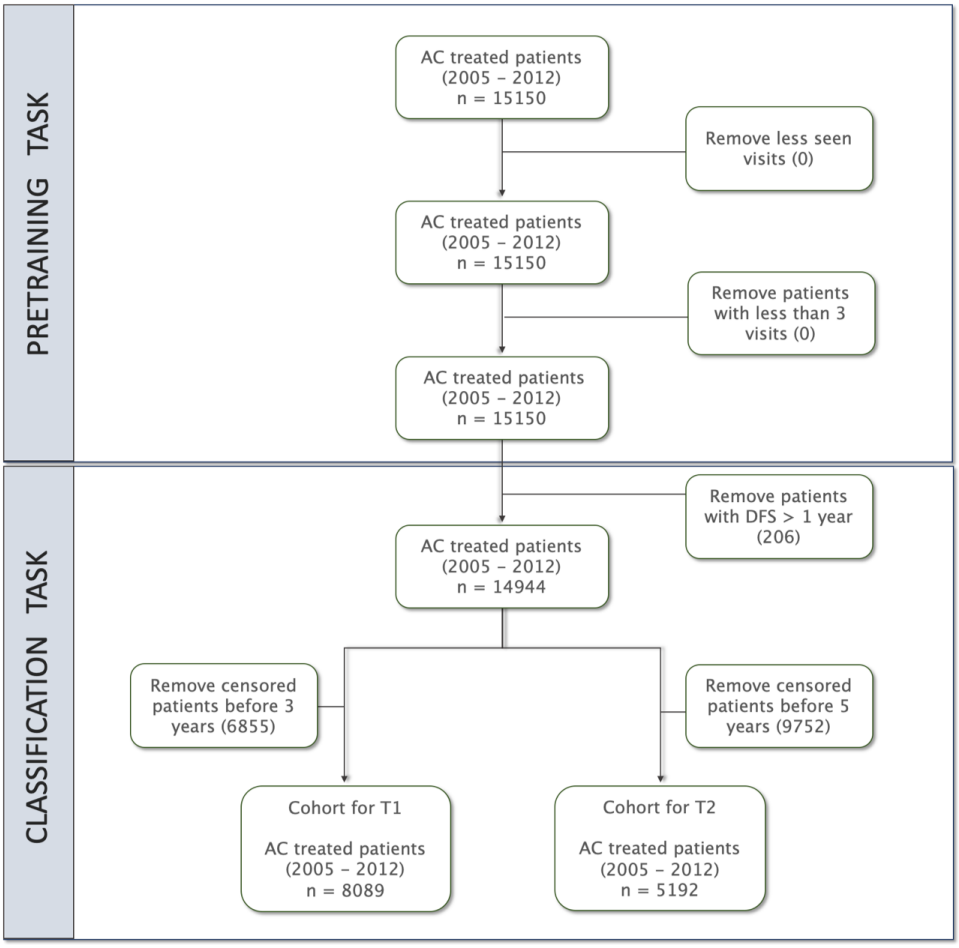


Figure 6.4: Flowchart of study inclusion and exclusion.

6.2.3 Tabular BEHRT

Information retrieved from EHR are generally time stamped events. As in Natural Language Processing, EHR can be transformed into sequences of tokens, where each token represents a unit of information from the EHR rather than a linguistic unit. These sequences can then be fed into language models such as transformers [Vaswani *et al.* (2017b)]. This was first proposed by [Li *et al.* (2020b)], who proposed BEHRT (BERT for EHR), an architecture based on that of BERT (Bidirectional Encoder Representations from Transformers) [Devlin *et al.* (2019c)] to predict future conditions from a sequence of diagnoses. Here we propose Tabular BEHRT, which is a transformer-based deep learning model whose architecture is inspired by BEHRT's. Tabular BEHRT considers that each medical visit is described using structured data: the department in which it took place, the corresponding procedure, as well as clinical and biological measurements available at this time. Like BERT and BEHRT, Tabular BEHRT combines a pre-training task (Masked Language Model) with a downstream task (the classification task), but applies it to a multimodal tabular EHR dataset.

Patient trajectory representation from structured EHR

By analogy with Natural Language Processing data, a patient's history can be seen as a document, where visits serve as sentences, and the events within the visits act as tokens. In our final data, the medical sequence consists of a sequence of visits that are chronogically ordered.

We used dates from the medical reports to construct medical chronological sequences. Each visit is described by the specific department and procedure from which the report originates, which contextualizes additional features, which are incorporated as available.

By analogy with Natural Language Processing data, a patient's history can be seen as a document, where visits serve as sentences, and the events within the visits act as tokens. In our final data, the medical sequence consists of a sequence of visits that are chronogically ordered.

As illustrated on Panel C of Figure 6.5, each visit is therefore described by at most 12 features: 5 biological measurements, the medical department where the visit took place, the type of procedure the visit corresponded to, the therapy and sub-therapy administered, the patient's age, the dNPI and the breast cancer subtype (which is static but repeated at each visit).

A separate modality layer indicates what kind of feature each measurement corresponds to. Generally speaking, this could be set to simply indicating the modality (biological, clinical, visit), but here we chose to be specific and encode the feature name. This allows us in particular to deal with missing values, which can simply be skipped as the modality layers provides the information of what feature is at each position. The modality

layer allows the algorithm to treat each modality differently.

As in BERT and BEHRT, a sequence of visits starts with the special token CLS, and visits are separated with the special token SEP.

Whereas BEHRT captures temporal information by including the age of the patient in a separate layer, we kept age as other clinical descriptors in the main input layer, but added another special embedding layer that represents the delay between the next visit and the previous. We discretized delays, as in Pang et al. 2021 [Pang *et al.* (2021b)], into W0-3 (under 1, 2, 3, or 4 weeks) for delays shorter than 4 weeks, M1-12 (under 1 month up to under 12 months) for delays shorter than a year and LT (long term) for delays longer than a year.

One of the notable constraints in BERT-like models is token capacity: they process tokens in fixed-size sequences of at most 512 tokens. While this size is arbitrary and varies depending on the exact BERT architecture and implementation, it cannot take much larger values, as it is linked to the memory usage of the self-attention mechanism of BERT, which grows quadratically with the number of tokens (each token being attentive to every other token). There is therefore a tradeoff between the number of features/tokens used to describe each visit, and the number of visits that can be considered by Tabular BEHRT. This is alleviated by the exclusion of both missing values and biological delta values equal to zero (corresponding to an absence of change in measurement), which is possible as the modality layer informs the architecture as to the kind of feature each token corresponds to. In practice, if the patient trajectory still exceeds 512 tokens, we only consider the first 512 tokens, which represent the initial interactions of the patient with the healthcare system, and inform about initial diagnostic visits and treatment decisions. Figure S4 in appendix shows how much information is excluded from patient trajectories due to restricting data to the 512 first tokens.

Panel C of Figure 6.5 illustrates the representation of each patient's sequence of visits that will be fed to Tabular BEHRT.

Model

Tabular BEHRT uses the transformers architecture to model temporal dependencies in the built sequence. Its architecture can be broken down into three (3) key components.

The *Input layer* as depicted on panel C of Figure 6.5 represents the sum of multiple embedding layers. Let E be the embedding matrix for medical events, i.e, the token embedding or input embedding have each unique event e_i, converted into a dense vector. The modalities embedding M represents a list of learnable dense vectors for each modality token m_i. The same is seen for each delay token d_i for the delay embedding, denoted D. Each visit or encounter is assigned an unique embedding to distinguish different visits in the Segment Visits embedding S and the position of each token within

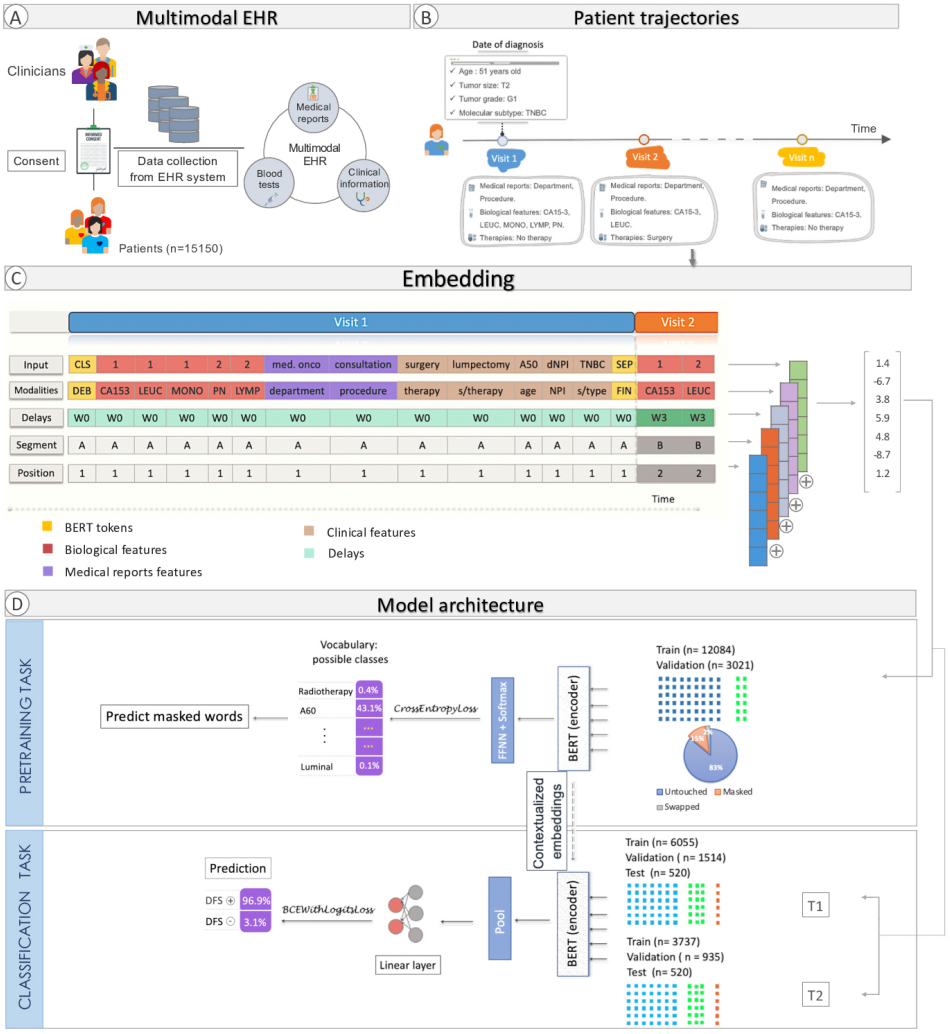


Figure 6.5: Tabular BEHRT architecture. Tabular BEHRT considers as input patient trajectories extracted from multimodal EHR (panel A) and represented as sequences of medical events where each event is characterized by tabular (or structured) data (panel B). Panel C shows an example of patient trajectory embedding. Panel D shows the architecture of Tabular BEHRT.

the visit sequence is encoded by P, to capture temporal order.

For a patient trajectory $\{e_1, e_2, ..., e_n\}$ across different visits, the input representation x_i for each token is given by:

$$x_i = E(e_i) + M(m_i) + D(d_i) + S(v_i) + P(p_i) \tag{6.1}$$

where m_i and d_i denote respectively the modality and the delay at that index, v_i denotes the visits' index and p_i denotes the position index within the visit.

In the original BERT [Devlin *et al.* (2019c)] and BEHRT [Li *et al.* (2020b)] papers, the position embedding layer allows to capture the location of an entity in the input embedding layer so that each position is assigned a unique representation. Transformers use a smart positional encoding scheme where each index is mapped to a vector where for a given position p_i. Therefore, the output of the positional encoding layer is a matrix where each row represents an encoded object of the sequence summed with its positional infomation. Suppose we have an input sequence of length L, for the i^th object within the sequence, the position embedding P is computed as follows:

$$P(k, 2i) = \sin\left(\frac{k}{m^{2i/d}}\right) \tag{6.2}$$

$$P(k, 2i+1) = \cos\left(\frac{k}{m^{2i/d}}\right), \tag{6.3}$$

where k is the position of an object in the input sequence ($0 \leq k < L/2$), d is the dimension of the output embedding space, $P(k, j)$ is the embedding mapping position k in the input sequence to index (k, j) of the positional matrix, m is a scalar defined by the user ($m = 10\,000$ by default in the original paper [Vaswani *et al.* (2017b)]) and i is used for mapping to column indices $0 \leq i < d/2$. The even positions are mapped using a sine function and the odd positions using a cosine functions. Sine and cosine functions are periodic, which allows the model to generalize the relative positions to tokens and to capture both absolute and relative positional information. The denominator $m^{2i/d}$ scales the position p differently for each dimension allowing the model to learn to attend to different positions.

Each *transformer encoder layer* processes the input representation through self-attention, as in Chapter 3 and a feed-forward network. Various aspects of the model architecture of the transformers and the training process are optimized through a hyperparameters tuning process using Bayesian optimization. The hyperparameters that have been chosen for the final model are the most that achieve the best validation performance in terms of average precision score (APS). Details are shown in the Appendix.

The final hidden states from the last transformer layer are used for prediction; this is the *output layer*. For a binary classification task such as DFS

prediction for instance, the sigmoid function is applied to the final state corresponding to the CLS token.

$$\hat{y} = \sigma(h_{[\text{CLS}]}.W + b) \tag{6.4}$$

Pretraining task To improve the embeddings of patient trajectories built from structured data, we follow the example of BEHRT and pre-train a Masked Language Model (MLM) on the representations described in Section 6.2.3. As in Natural Language Processing, the MLM is designed to predict missing or masked tokens within a patient's history, using the bidirectionaly context provided by the surrounding tokens. Its goal is to learn contextual representation of the medical events in the patient's history. For this purpose, in this pre-training phase Tabular-BEHRT uses the whole cohort of 15 150 patients and the entire sequence of events for each patient, from the date of diagnosis to the date of death or censorship, with a length average of 506($\pm$466) tokens (Figure xx shown in appendix). We randomly replaced 15% of the tokens with a special MASK token. We swapped another 2% with another token at random; this adds a limited amount of noise, encouraging the model to learn a more robust and generalizable representation of patient trajectories. As shown in panel D of Figure 6.5, the MLM part of Tabular BEHRT is a transformer-based architecture that generates probabilities for each token in the vocabulary, computed using softmax over the model's output logits, as a multilabel learning task.

We first split the dataset into a training (90%) and a validation set (10%) in order to prevent overfitting. Then, all the embeddings from the training set are randomly initialized and fed to the MLM. We use Bayesian optimization to find the best set of hyperparameters, with precision as a criterion.

For robustness, we run the model five times with five different random seeds for the sequence masking, and use as final token embeddings for the downstream classification tasks the mean values of standardized embeddings from these five runs.

Implementation details for the pretraining task In this thesis, I implemented BEHRT as a custom BERT model, following the BERTOnlyMLMHead implementation from HuggingFace [Wolf *et al.* (2019)]. The BERT model is as already described in chapter 3, a stacked encoders layer that take as inputs the embeddings vectors from Equation 6.1. The encoded information from the BERT model serve as an input the the MLM module. It consists of a linear layer to map the hidden states to a vocabulary-sized vector, followed by a normalization layer to normalize the hidden state and a softmax activation to produce a probability distribution over the vocabulary. The MLM is trained by minimizing the Cross Entropy Loss. Formally, the MLM can be written using the following ingredients:

1. Input Sequence: Let $X = (x_1, x_2, ..., x_n)$ be the input token sequence.
2. Masked Sequence: Let $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ be the masked sequence, where some tokens x_i are replaced by MASK.
3. Hidden layers intermediate output computation :
$$\mathbf{z}_i = W_h \mathbf{h}_i + b_h$$
4. Hidden layer output computation via Layer Normalization [Ba *et al.* (2016)]:
$$\mathbf{z}'_i = \text{LayerNorm}(\mathbf{z}_i)$$
5. Softmax:
$$P(y_i|\tilde{X}) = \text{softmax}(W_h \mathbf{h}_i + b_h)$$
where $\mathbf{h}_i$ is the hidden state corresponding to the i-th token position, and W_h and b_h are the weights and bias of the output layer.
6. Loss Function:
$$\mathcal{L}_{\text{MLM}} = -\sum_{i \in M} \log P(y_i = x_i|\tilde{X})$$
where M is the set of masked token positions.

We used Bayesian optimisation to choose the best set of hyperparameters for the model. The evaluation of the MLM is done using the precision of the model, i.e, the percentage of correctly predicted masked tokens.

DFS prediction In this study, we assessed two binary classification tasks: disease-free survival (DFS) 3 years after surgery (prediction task T1) and 5 years after surgery (prediction task T2). We used the same test set than the previous machine learning models that remained untouched throughout the whole model development process with number of samples $N = 520$. We randomly split the remaining data into a training (90%) and a validation set (10%). We used Bayesian optimization to find the optimal set of hyperparameters during the training phase, using Average Precision Score (APS) on the validation set, as a performance criterion.

Implementation details for DFS prediction For the classification task, I used the same BERT architecture as for the MLM task. As shown on Figure 6.5, only the last layer is different between pre-training and fine-tuning: here the patient history embeddings are fed to a single feed-forward layer. This layer maps the encoded information to a 2D vector appropriate for binary classification. The training starts by loading the pretrained

weights from the MLM. We obtain logits from the classification layer. The loss used to compute the loss between the predicted logits and the true label is the a Binary Cross Entropy With Logits Loss (BCEWithLogitsLoss).

More formally, let $X = (x_1, x_2, ..., x_n)$ be the input token sequence. It is fed to the BERT-like model which returns as output $\mathbf{h}_{[\text{CLS}]}$, which is the hidden state corresponding to the CLS token. The model then applies a logit classification layer to $\mathbf{h}_{[\text{CLS}]}$:

$$\text{logits} = W \cdot \mathbf{h}_{[\text{CLS}]} + b \tag{6.5}$$

where W is the weight matrix and b is the bias vector. The best parameters are found by minimizing the BCEWithLogitsLoss:

$$\text{BCEWithLogitsLoss(logits, y)} = \text{mean}(\ell(\sigma(\text{logits}), y)) \tag{6.6}$$

with:

$$\ell = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(P(y=1|\text{input}_i)) + (1-y_i)\log(1-P(y=1|\text{input}_i))]$$

where N is the number of samples, y_i is the true label (0 or 1), and input_i is the input sequence.

Because the labeled data is typically imbalanced (see Section 3.3.3), we also implemented a stratified batches strategy to ensure that the the model learns to recognize all classes more effectively and does not become biased towards the majority class. This technique consists in loading the same proportion of positive and negative samples for each batch, with replacement for the positive instances (the minority class), during training.

I also implemented other techniques to address class imbalance during this thesis, such as the "class weights" technique, which consists in adjusting the loss function to weigh samples from the minority more heavily. By penalizing the majority class, the model is ensured to have enhanced performance on minority classes. However, this technique did not show as good validation performances as balanced batches in practice.

I also experimented with various techniques to represent delays. They included encoding the *delay embedding layer* using the Time2Vec [Kazemi *et al.* (2019)] encoding method. This method, similar to the positional encoding described above, is designed for temporal sequential information. It has been particularly effective in multiple applications that involve temporal data [Kazemi *et al.* (2019), Ozair *et al.* (2020), López-Andreu *et al.* (2023)], and for various neural network architectures (LSTM, GRU or Tranformers). It is defined as follows:

Given a time point t, the Time2Vec representation $\mathbf{t2v}(t)$ is computed using a combination of a linear term and periodic terms:

$$\mathbf{t2v}(t) = [w_0 t + b_0, \sin(w_1 t + b_1), \cos(w_2 t + b_2), \ldots, \sin(w_{d-1} t + b_{d-1})]$$

where w_i and b_i are learnable parameters, the first term $w_0 t + b_0$ is a linear component and the remaining terms are periodic components (sine and cosine functions) that capture cyclical patterns. However, this technique has failed to improve the classification task.

Comparison baselines

To evaluate Tabular-BEHRT's effectiveness, we developed baselines that served as benchmarks for gauging M-BEHRT's performance in a DFS status classification task. Those are standard machine learning methods: random forests classifiers (RF), logistic regression (LR), and support vector machines (SVM) (see chapter 3). Moreover, we used the NPI measured at the date of diagnosis, which is a tool that has helped clinicians to determine prognosis, as another benchmark. Machine learning models (RF, LR and SVM) used the same features used for Tabular-BEHRT presented in a structured representation (Table B.2 in appendix). For patient trajectories, sequence of events are transformed into occurrences of events for certain features (procedure and department names, and biological measurements in or outside of the normal range). Clinical features (age, therapies, tumor size, tumor grade, breast cancer molecular sub-type and number of nodes) are kept as static clinical features in the table.

6.3 Results

6.3.1 Events' embedding

The optimal hyperparameters we identified for the MLM are 5 hidden layers with 12 attention heads, a hidden size of 144, an intermediate layer size of 133, a training duration of 120 epochs, using Adam optimizer with a learning rate set to 1e-3 and a batch size of 64.

To assess the performance of the MLM, we ran the model five times with five different random seeds for the sequence masking. We also compute a baseline by running the MLM on a data set in which tokens have been randomly reordered within each sequence. This approach disrupts the inherent sequential structure of the data, and creates a scenario where the model should not be able to rely on contextual relationships between tokens. Hence, comparing the MLM's performance on shuffled sequences against its performance on original sequences offers a benchmark for assessing the impact of contextual information on the model's predictive capabilities. The precision of these models (proportion of correctly predicted masked tokens) on the held-out validation set is shown on Figure 6.6.

The MLM is able to predict masked tokens with a precision of 72% on the validation set, a performance that is not significantly different from the one on the training set, highlighting the absence of overfitting. In addition,

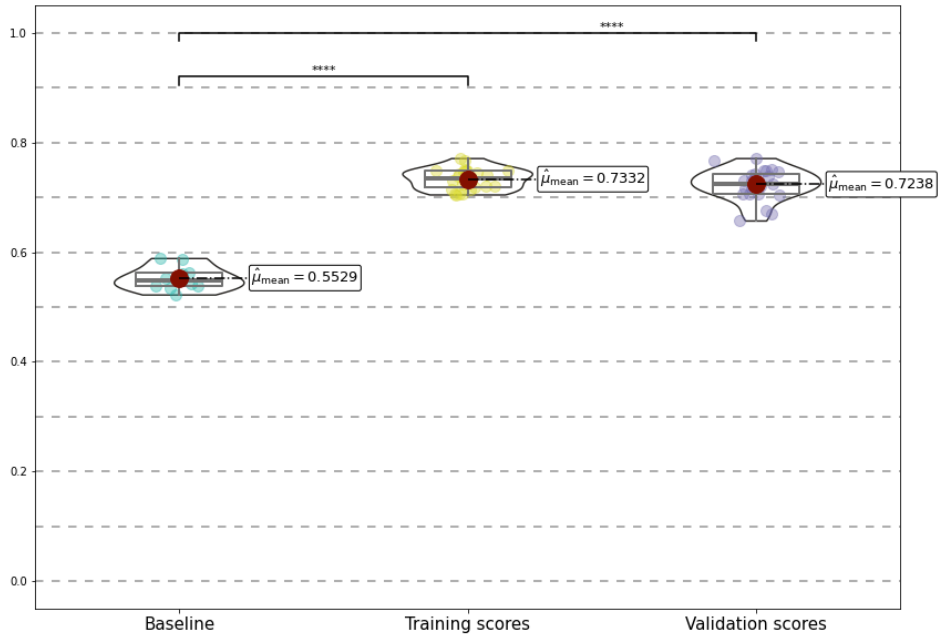


Figure 6.6: Precision scores for the Masked Language Model. The baseline scores are obtained from the MLM ran on shuffled sequences.

this precision is significantly higher than the precision of 55% obtained when shuffling the sequences, which shows that the MLM does indeed capture contextual information. We also note that the precision of the MLM of BEHRT reported by [Li *et al.* (2020b)] on sequences of diagnoses is of 66%. While it is difficult to compare this performance to ours due to the different nature of the tasks, it indicates that the MLM provides embeddings of sufficient quality to perform supervised learning in a second stage.

We further evaluate embeddings generated by the MLM by visualizing token embeddings through two-dimensional plotting along the first two components of a t-distributed Stochastic Neighbor Embedding (t-SNE) as shown on Figure 6.7. This figure shows how the MLM capture semantic relationships between tokens and contextual information. Tokens belong to the same modality (therapies, variation in biological features, breast cancer subtypes) tend to cluster together, with the exception of procedures and departments, which tend to be mixed together. This is however unsurprising, as some procedures and departments are tightly linked; for example, panel F shows that the embedding of the "nuclear medicine" service is quite close to the embeddings of "radiology", "scanner" and "MRI" procedures, while panel D shows that the embedding of the "radiotherapy" service is quite close to the embeddings of several procedures all relating to the proposal, prescription, initiation, unfolding and ending of treatment by radiotherapy.

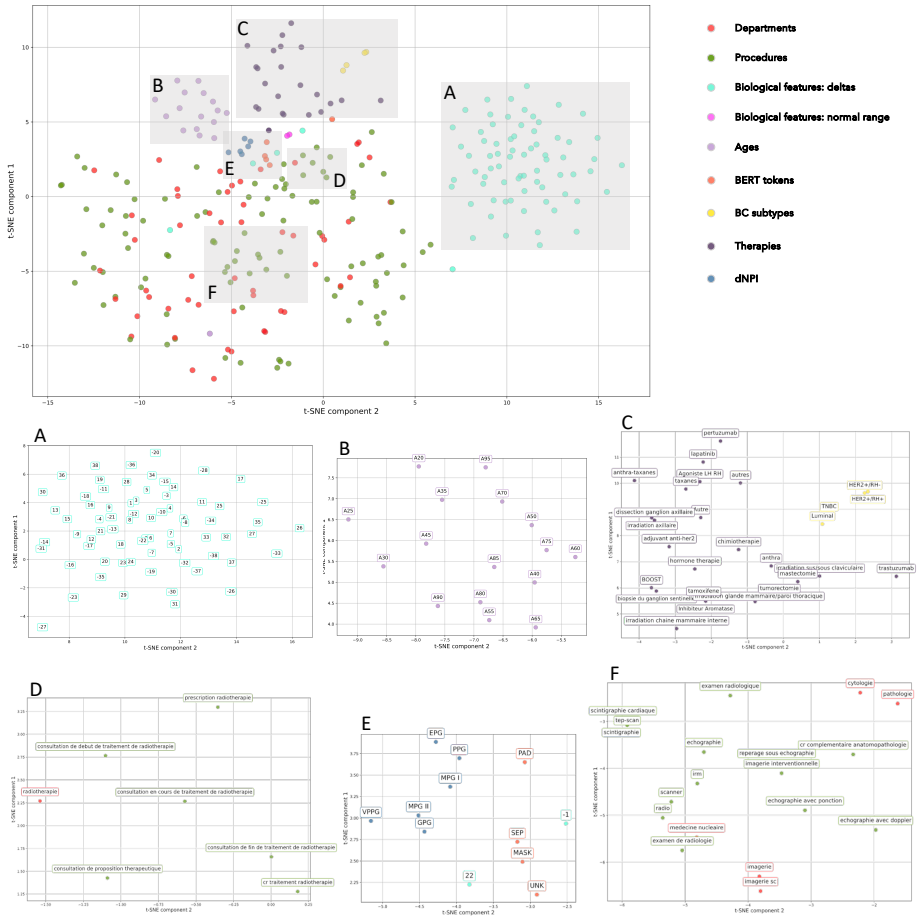


Figure 6.7: t-SNE of Tabular BEHRT tokens embeddings as learned by the Masked Language Model. Panels A through F zoom in on specific section of the plot. Panel A corresponds to a cluster of deltas in biological measurements. Panel B shows that age tokens cluster together. Panel C shows that therapy token, on the one hand, and breast cancer subtypes, on the other, cluster together. Panel D and F show two different clusters of procedures and departments. Panel E show that dNPI tokens cluster together, as well as BERT special tokens.

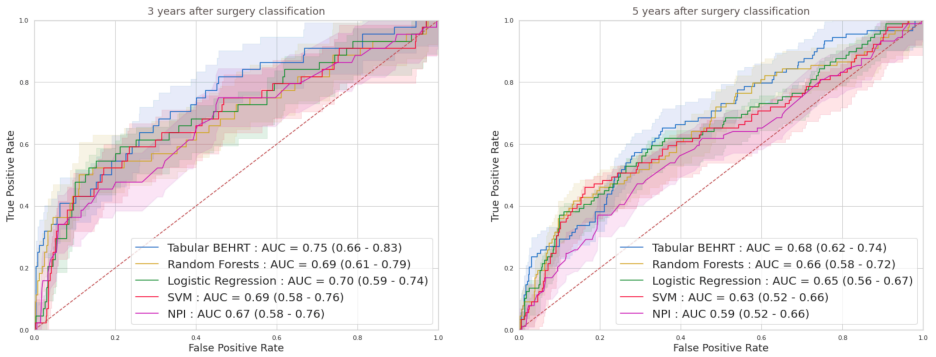


Figure 6.8: ROC curves for baselines and Tabular BEHRT, for predicting disease-free survival 3 years (T1, left) or 5 years (T2, right) after surgery.

6.3.2 Classification task results

Comparison to state-of-the-art predictive algorithms

We determined the best set of hyperparameters on the validation sets and evaluated the best-performing model on the held-out test set. For T1, we trained Tabular BEHRT for 5 epochs, using Adam optimizer with a learning rate of 10^{-4} and a batch size of 16. For T2, we trained Tabular BEHRT for 20 epochs, using Adam optimizer with a learning rate of 3.10^{-4} and a batch size of 32. We then evaluated the performance of Tabular BEHRT on the until now untouched test set, using Receiver Operating Characteristic (ROC) curves. APS performances can be found in the appendix. For direct numerical comparison, we also report the Area Under the ROC Curve (AUC-ROC).Figure 6.8 compares the performance of Tabular BEHRT with the baselines presented in Section 6.2.3.

Tabular BEHRT outperforms all comparison partners. All methods perform significantly better than random classifier. However predicting DFS after 5 years (T2) is more difficult than after 3 years.

Tabular BEHRT with small size datasets

While BEHRT and its variants have been trained on millions on samples, focusing on a specific disease (here breast cancer) for a single moderately-sized hospital (here Institut Curie) drastically reduces the number of samples available for training. To put to the test the ability of Tabular BEHRT to learn from small training sets, we experimented with reducing the size of the training set even further. To this end, we created smaller training sets by randomly selecting subsets of the training data, starting from 10 samples, and compared on the test set the performance of Tabular BEHRT and classical machine learning algorithms trained on these small training sets. Figure 6.9

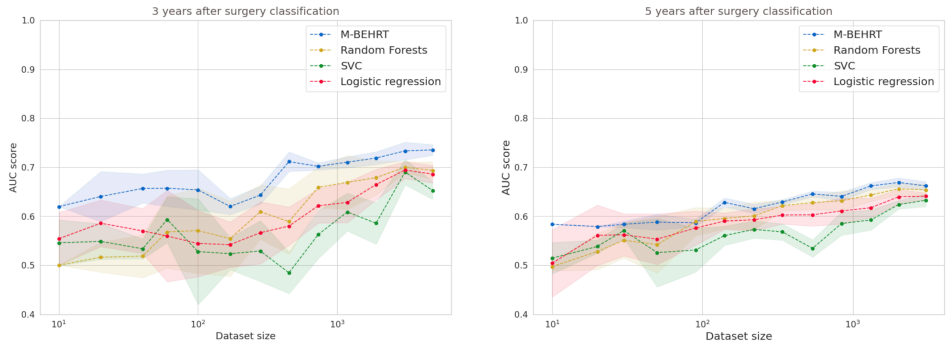


Figure 6.9: APS (top) and AUC-ROC (bottom) on the test set for M-BEHRT, random forests, support vector classifier, and logistic regression trained on dataset of increasing sizes (x-axis).

shows that Tabular BEHRT clearly outperforms the classical machine learning algorithms, especially random forests, in the few-shot learning setting (when training set sizes are very small), achieving better-than-random performance with as little as 10 training samples and outperforming NPI with a few hundred training samples. We attribute this performance to the ability of the pretraining phase to learn meaningful representations of patient trajectories. While Tabular-BEHRT remains above the others, the gap between methods narrows as sample sizes increase.

Ablation study

In order to better understand the contribution of each modality to the performance of M-BEHRT, we performed an ablation study, in which we evaluated how the model performs when removing some of the modalities. As shown on Figure 6.10, dNPI contributes the most to the performance. However, the addition of the other features, in particular the remaining clinical features (including age and more notably therapies), increases performance significantly. This is in line with observations from other studies, in which clinical features are the one providing the most information towards the prediction of breast cancer relapse [Perou *et al.* (2000b), Dent *et al.* (2007)]. Biological features contribute the least to performance, although they still contain information, as they allow for better-than-random prediction. However, it seems that this information is redundant with that captured by the other features. However, early experimentations with including a quantized version of the variation between biological measurements did not improve performance (results not shown). Moreover, therapies administrerded for breast cancer ofrten serve as a proxy for a wealth of information about the tumour characteristics, and those characteristics are mostly correlated

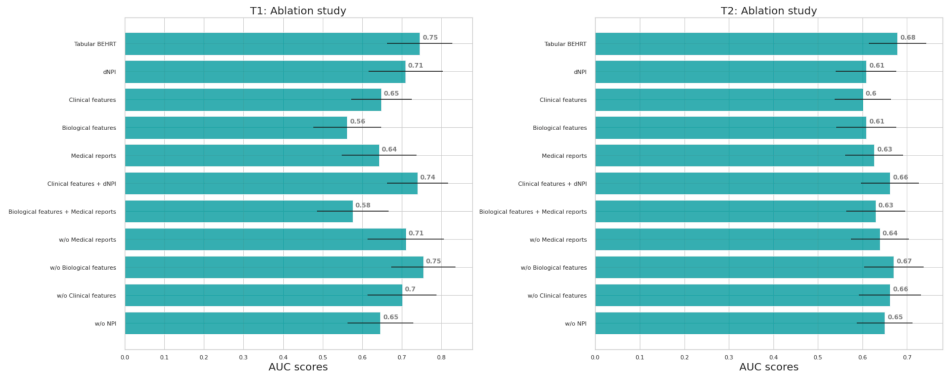


Figure 6.10: Ablation studies AUC-ROC on the test set for Tabular BEHRT. We present results for the full model (Tabular BEHRT), then using only one of the 4 modalities (dNPI, clinical features, biological features, medical visits), two modalities (dNPI+clinical or biological+visits), then removing one of the 4 modalities. Here "medical records" stands for features extracted extracted from the medical record headers, that is to say, visit department and procedure. Performance scores are presented on the test set.

xwith the patient's prognosis. There are therapies choices that are made in more advanced or aggressive tumors, such as adjuvant or neo-adjuvant chemotherapies, which generally indicate a poorer prognosis. On another hand, treatments such as hormone-therapy are mainly used for early stage or less aggressive tumors, which may foreshadow a better prognosis. Performance also drops substantially if information about the nature of the medical visit (department and procedure) is omitted. These observations are consistant across both tasks.

Tabular BEHRT performance per cancer subtype

Figure 6.11 presents the AUC-ROC of Tabular BEHRT on the test set, stratified by patient age, tumor grade, molecular subtype, or node status. Tabular-BEHRT is better at predicting DFS at three years on older patients, with at least one affected lymph node. Stratification of results by NPI range is available in the appendix.

Model interpretation

I perform interpretation for Tabular BEHRT using the IG method implemented in the CAPTUM python library [Kokhlikyan *et al.* (2020)] (see section 3.7). I show, in following figures (6.12, 6.13), interpreted sequences for 3 different samples that have different prognosis groups in both tasks T1 and T2. In these sequences, the 'white' tokens do not have an impact on

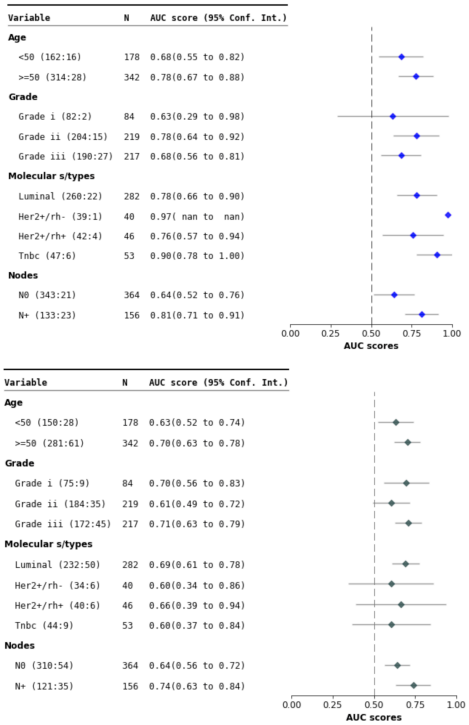


Figure 6.11: AUC-ROC stratified by patient age, cancer grade, molecular subtype and node status, for tasks T1 (prediction of DFS 3 years after surgery, top) and T2 (prediction of DFS 5 years after surgery, bottom).

the model prediction, the tokens in 'green' have an impact on the positive endpoint (influence on the relapse), and the tokens in 'red' have an impact on the negative impact (influence on the non-relapse status). First, it becomes evident that the model heavily relies on the well-defined prognostic groups for making its decisions. The bad NPI groups (VPPG and PPG) have high attributions towards the positive group (relapse) while the good NPI group (GPG) have high attributions towards the negative group (non-relapse). These examples are only showing TP samples. Therefore, NPI tokens are mainly used by Tabular BEHRT for the poor prognosis groups (VPPG and PPG). Regarding the good and moderate prognosis groups, tokens that provide critical insights into the aggressiveness and progression of the disease, have been used by Tabular BEHRT to accurately predict relapse. They includes a high number of 'RCP' *(multidisplinary consultation meetings)*, a high number of 'consultations' *(auscultations)*, a second surgical procedure or abnormal value for the CA15-3 and the LYMP biological markers. Moreover, Tabular BEHRT uses well-documented factors in the litterature to predict a positive DFS status such as age. Delays sequences have also been explored in the interpretation step, however I did not find a reliable explanations for the model predictions.

Additionally, we gather the most frequent events with higher attribution across all the samples. We output features that have been predictive for at least 10% across samples. The result include 'CLS' token, 'consultations' tokens and 'rcp' token. The CLS token's primary function is to capture a summary representation of the entire input sequence for the purpose of classification tasks, which explain its high influence to prediction. Consultations tokens cover 'consultation during treatment', 'consultations', 'adult consultation' and 'announcement consultations'. From a medical point of view, these events are not specifically linked to cancer prognosis. The 'rcp' token can however give insights on the cancer severeness. We plot survival plots to visualize time to relapse for different patient groups with varying 'rcp' numbers post-surgery 6.14. The number of 'rcp' is not necessarily linked to prognosis, but it indicates the difficulty level of treating the cancer.

6.4 Conclusion

In this chapter, we proposed Tabular BEHRT, a deep learning architecture which considers structured data to describe each medical event. Our work is motivated by applications to oncology, and applied to the prediction of disease-free survival for breast cancer patients. In Tabular BEHRT, we chose to perform early integration of the different modalities in the tabular data describing medical events by concatenating the corresponding features in a single input layer. However, an additional modality embedding layer keeps track of which modality each feature comes from and modulates the

Figure 6.12: Interpretation examples for true positive samples in T1, from a bad prognostic group (VPPG), to a good prognostic group (GPG) (top to bottom).

Figure 6.13: Interpretation examples for true positive samples in T1, from a bad prognostic group (PPG), to a good prognostic group (GPG) (top to bottom).

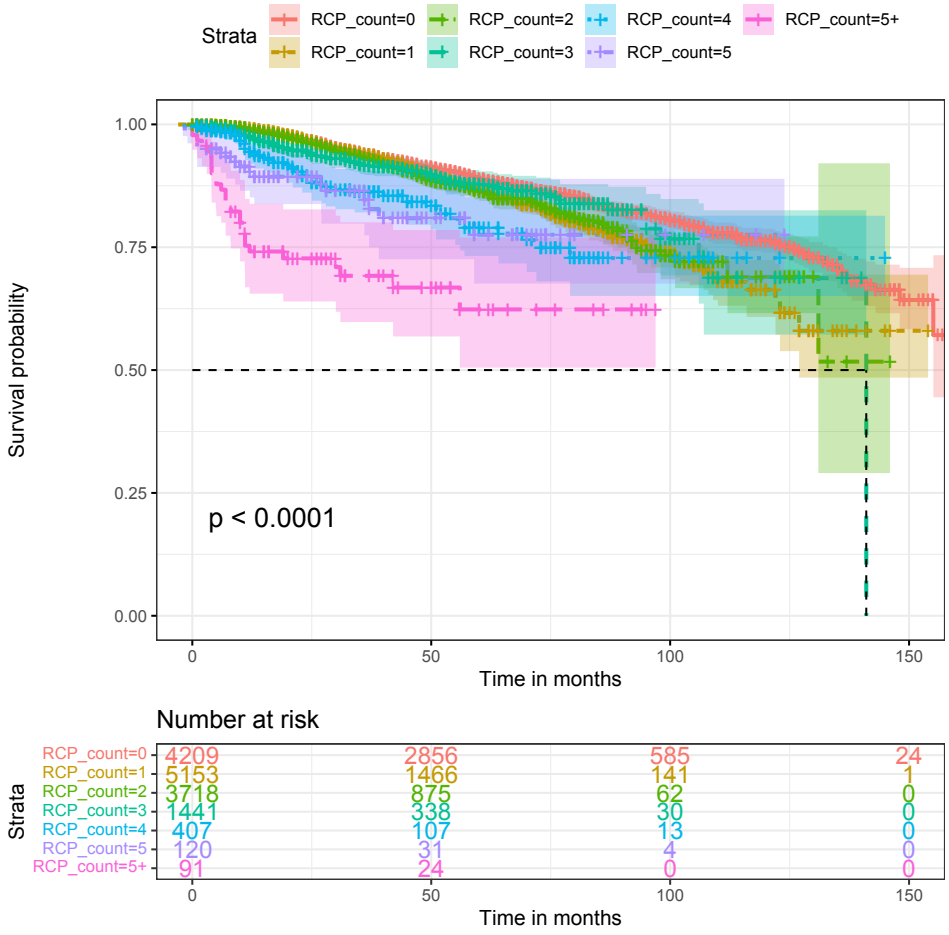


Figure 6.14: Survival plots for samples with varying "RCP" counts post-surgery.

contribution of each modality to the final model, allowing the model to treat each modality differently.

In this approach, there is a tradeoff between the number of visits that can be considered and the amount of information that can be used to describe each visit, because the underlying BERT architecture is limited to processing 512 tokens. This number is arbitrary, but constrained by the memory usage of the self-attention mechanism. We have found this number to be sufficient for the DFS prediction tasks at hand and the available features and modalities. However, this might be too small for other applications, in which case one might want to use approaches that approximate the self-attention matrices so as to reduce their memory footprint, such as Big Bird [Zaheer *et al.* (2020b)] or Nyströmformer [Xiong *et al.* (2021)].

To the best of our knowledge, this is the first study predicting breast cancer endpoints from sequences of EHR data, whether considering solely multimodal dynamic tabular data, solely the contents of free-text reports, or combining both. Our results underscore the usefulness of such data for future research on prognosis modeling, and outline the importance of integrating medical information collected over time to gain previously unknown insights into the understanding of breast cancer evolution.

Tabular BEHRT achieve AUCs on a held-out data set of 0.75 for the prediction of DFS 3 years after surgery and 0.68 for the prediction of DFS 5 years after surgery. Predicting DFS 3 years after surgery seems much easier than 5 years after surgery (AUC of 0.77 vs 0.69). This is in line with previous observations that earlier events are easier to predict than long-term ones [Witteveen *et al.* (2015)]. Their time-dependent prognostic tool, designed to estimate the yearly risk of locoregional recurrence in early breast cancer patients showed an AUC of 0.84, 0.77, 0.70 0.73 and 0.62, respectively, for each successive year after the primary treatment. This decline in performance over time might stem from the influence of by more subtle and/or complex effects on long-term DFS. Moreover, there may be factors than have not been captured in the input sequence up to 1 year after surgery that could have more impact on long-term DFS.

Intuitively, deep-learning methods, specially transformers-like models, are better to fit data if the training size is sufficient (in the range of 10K at least). In comparable transformer models that analyze EHR, [Li *et al.* (2020b)] introduced a model that was trained and evaluated on nearly 1.6 million samples. Indeed, many other studies use high number of samples: Med-BERT [Rasmy *et al.* (2021)] has been pre-trained using more than 28 millions patients EHR and fine-tuned on different prediction tasks using from 29,405 to 672,647 samples. In a similar approach, [Pang *et al.* (2021b)] fine-tuned CEHR-BEHRT with high numbers of samples, which range from 97 758 to 590 578 for their different downstream tasks. Unlike these studies, our model is constrained by a smaller sample size: about 15 000 patients for pretraining, and 5 000 to 8 000 patients for fine-tuning. And despite its

smaller scale, this dataset serves as a valuable resource for DFS prediction in breast cancer. The possibility to apply such methods to much smaller data sets is very encouraging for future research, as many studies, especially on very specific diseases and endpoints, only have access to a limited number of patients. Keeping the same pretrained model, we experimented with further reducing the number of patients used for training the classifier. While performance was of course degraded, the models learned with Tabular BEHRT using only 100 samples had much better AUC than their classical machine learning counterparts, which performed close to random. This highlights the importance of pretraining in this context.

The primary goal in this study is to predict DFS using the longitudinal information from the different modalities throughout the patient medical trajectory. Tabular BEHRT performances underscore the capability of transformer architectures to capture event dependencies across the entire sequence, a practice actually comparable to clinicians' considerations when predicting potential breast cancer evolution for individual patients. Our results also highlight the limitations of relying solely on clinician predictions at specific time points, NPI for example, as these may lead to miss relevant information. Additionally, NPI's utilization of only three clinical features, albeit important, appears insufficient to capture the entirety of predictive factors. Notably, [Kim *et al.* (2012)] and [Wu *et al.* (2017)] report a lower performance for NPI (AUC of 0.70 and 0.751, respectively) compared to their prognosis models' performances (respectively, an AUC of 0.85 with a SVM and an AUC of 0.807 with a Cox Regression analysis) for recurrence prediction at 5 years after breast cancer surgery. These models incorporated other clinical features in their process in addition to those already used by NPI, such as lymphovascular invasion (LVI), ER status, and metastatic lymph node.

We stratified the data based on features that are expected to define patients with similar prognoses (age, grade, number of lymph nodes involved, molecular subtype). We found that the prediction ability of Tabular BEHRT varies depending on subgroups and that the model works better on older patients with more aggressive disease (at least one lymph node involved). In addition, Tabular BEHRT is better at predicting relapse after 5 years than after 3 years for luminal tumors, suggesting that it correctly identifies predictive factors with long term influence for these tumors that tend to recur later than others [Ignatov *et al.* (2018)].

In this chapter, we have only used tabular information to describe each medical visit. In the next chapter, we will show a similar architecture to learn from free text describing the visits instead.

Chapter 7

Text BEHRT: Pre-trained transformers models for Free-text reports

---

Abstract:

Prognosis prediction in breast cancer research is particularly crucial for patient management. And many work have been done to leverage machine learning tolls to enhance de reliability and the accuracy of those predictions and EHR have been a major part of that process. In this thesis, I explore the integration of free text medical reports with advanced deep learning methods such as transformers for DFS status prediction. I developed Text BEHRT, a transformer-based models that use the sequence of free text reports through the patient journey, to assess its efficacity in DFS status prediction. As the previous chapter, Text-BEHRT is inspired by the BERT model. In this chapter, I will present the whole pipeline that: (i) processes french free-text medical reports into valuable information, (ii) readapt them into sequences of free text trajectories and that (iii) models this information through Text-BEHRT.

Résumé:

La prédiction du pronostic dans la recherche sur le cancer du sein est particulièrement cruciale pour la gestion des patients. De nombreux travaux ont été réalisés pour exploiter les outils d'apprentissage automatique afin d'améliorer la fiabilité et la précision de ces prédictions, et les dossiers médicaux électroniques ont joué un rôle majeur dans ce processus. Dans cette thèse, j'explore l'intégration de rapports médicaux en texte libre avec des méthodes avancées d'apprentissage profond telles que les transformers pour la prédiction de la DFS. J'ai développé Text BEHRT, un modèle basé sur un transformers qui utilise la séquence de rapports en texte libre tout au long du parcours du patient, afin d'évaluer son efficacité dans la prédiction de la DFS. Comme dans le chapitre précédent, Text-BEHRT s'inspire du modèle BERT. Dans ce chapitre, je présenterai l'ensemble du pipeline qui : (i) traite les rapports médicaux en texte libre en informations utiles, (ii) les réadapte en séquences de trajectoires en texte libre et qui (iii) modélise ces informations par le biais de Text-BEHRT.

Contents

In this chapter, I also present a BEHRT-based model but applied to the text data present in medical reports. The medical reports are treated in a chronological order through a free-text patients trajectories and I learn a transformers based models using that data.

7.1 Introduction

The previous chapter showed a new transformer-based architecture for learning from EHRs where each visit is represented by tabular data. However, as discussed in Chapter 5, we also have free-text reports available for each of these visits.

Many other works had taken advantage of the considerable information present in text reports. [Zeng *et al.* (2019b)] developed a support vector machine to identify breast cancer local recurrences using concepts extracted from text reports by MetaMap, and the number of pathological reports recorded for each patient. there is a need of using a multimodal approach to predict breast cancer relapse, combining clinical, pathological and molecular information. Their model acheived a high AUC of 93% in cross-validation. For [González-Castro *et al.* (2023a)], medical concepts are also extracted from reports to constitute features that will be combined to clinical information. In the 5-year cancer recurrence their best model (eXtreme Gradient Boosting) reached a great AUC of 80.7%.

In this chapter, we present a transformer architecture for DFS status prediction from free-text reports, applied to the SEIN database. As Tabular BEHRT, Text BEHRT is evaluated on T1 and T2, namely, 3 and 5 years after surgery relapse, respectively.

7.2 Materials and Methods

Data description

The data used for this study is the same than in the chapter 5. Therefore, I refer the reader to the Data section in that chapter for the free-text reports data description.

Data preprocessing

The preprocessing steps that have been applied to the free-text data is detailed in the section 5.2.3 in the chapter 5.

7.2.1 Text-BEHRT

Patient trajectory with free-text report

In addition, we assume that important information is contained within the text itself of the free-text reports. We therefore build a sequence of free-text reports, ordered chronologically from the date of the diagnosis until the index date (one year after the first surgery). As shown in Table B.1 in the chapter 5, the number of reports per patient and the length of each report are such that these create very long documents (on average 34 reports, averaging 159 words each, for a total of more than 5 000 words per patient history, See appendix for words histogram). However, while BERT has proven to be highly effective in capturing contextual relationships and semantic nuances in text, it can only process sequences of at most 512 tokens, due to the memory footprint of the self-attention mechanism.

This constraint again poses challenges when dealing with lengthy documents such as a sequence of medical reports [Gao *et al.* (2021)]. Using transformers to classify long documents is still a topic of open research [Park *et al.* (2022)]. The most straightforward approach consists in truncating inputs to fit within the allowed number of tokens, typically by using the first, last or middle tokens. However, limiting patient history to 512 tokens may result in major information loss and hence produce incomplete representation of medical reports. Other approaches such as Big Bird [Zaheer *et al.* (2020b)] or Nyströmformer [Xiong *et al.* (2021)] use sparse or low-rank approximations of the self-attention matrices. However, existing pretrained models typically do not handle more than 4 096 tokens, which is still too short for some of the patients in our data set. In addition, they have only been trained on English corpora whereas our medical notes are in French. Nevertheless, our corpus is much too small to train a transformer model from scratch. Finally, many approaches consist in dividing long text into chunks smaller than 512 tokens and combining their embeddings, whether through an additional layer of self-attention in a hierarchical model [Pappagari *et al.* (2019)] or by pooling [Li *et al.* (2023a)].

In the absence of a clear consensus on which of these strategies is likely to perform best [Park *et al.* (2022), Li *et al.* (2023a)], we chose to test two methods: CLS pooling, by using token embedding that starts every report, as the representation for the whole medical report, as it has been defined to contain the most information of the sequence report, and a simple aggregation strategy. However, by comparing results in downstream tasks, the simple aggregation pooling have shown to be more efficient for the task, therefore we will only discuss on that method for the rest of the discussion.

More specifically, we construct the embedding of every report by summing the embeddings of all tokens it contains, and construct sequences not of token embeddings, but of reports embeddings. We compute token em-

beddings from the three following models: word2vec CBOW [Mikolov *et al.* (2013)], word2vec skipgram [Mikolov *et al.* (2013)], and DrBERT [Labrak *et al.* (2023)].

- CBOW or Continuous Bag Of Words, a word embedding technique as part of the Word2Vec framework [Mikolov *et al.* (2013)], designed to learn words representations by predicting a target word based on its context words within a given window size (see Figure 7.1)
- Skip-Gram, another word embedding technique from the Word2Vec framework, which predicts surrounding context words from a target word to learn words embeddings.
- DrBERT [Labrak *et al.* (2023)], a state-of-the-art pre-trained transformer model, based on the RoBERTa architecture [Liu *et al.* (2019b)] and trained on a French biomedical corpus which contains 7GB of clinical data from multiple sources.

We can then train a BERT model on the sequences of reports embeddings.

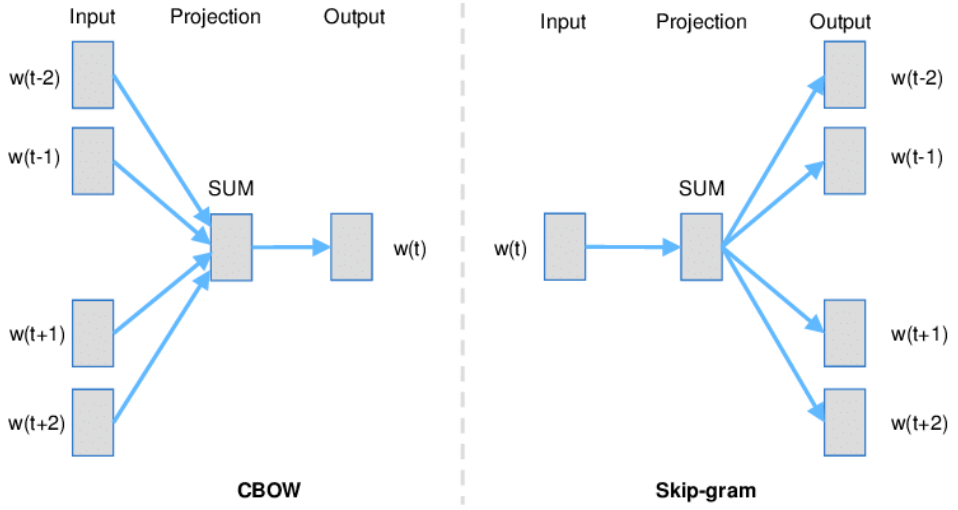


Figure 7.1: CBOW and Skipgram training model illustration adapted from [Mikolov *et al.* (2013)]. The task is iterated over the whole corpus, word by word)

To account for temporality, we add an embedding layer of delays between reports. Finally, we use BERT special tokens: CLS for the start of a medical history and SEP to separate reports from different visits. This representation is illustrated on panel B of Figure 7.2. In Text BEHRT, we didn't run a MLM task, as running MLM on the whole corpus would require more computational resources than available.

Figure 7.2: Text BERT architecture

The architecture of Text BEHRT is illustrated on Panel D of Figure 7.2. It is again a transformer-based model, which uses report embeddings obtained through the aggregation of DrBERT embeddings as described in Section 7.2.1. The same sampling strategy as the one depicted in the previous section is used for this task.

Binary classification

Text BEHRT is again a transformer-based model, which uses report embeddings obtained through the aggregation of the best word embedding model as described in Section 7.2.1. The tasks are the same than in Tabular BEHRT: disease-free survival (DFS) 3 years after surgery (prediction task T1) and 5 years after surgery (prediction task T2). The same test set will be used for evaluation and the same sampling strategy as the one depicted in the previous section is used for this task. We also used Bayesian optimization to find optimal set of hyperparameters using the best validation Average Precision Score (APS).

Implementation details The implementation of Text BEHRT is similar to that of Text BEHRT. During training, the optimization is done by minimizing a Binary Cross Entropy With Logits Loss. We also used stratified batches to load the same number of negative than positive samples during training.

Comparison baselines Machine learning models (RF, LR and SVM) and the NPI are used as benchmarks to evaluate Text BEHRT's performance. As inputs, the ML model used flattened reports embeddings, from a vector of 768 dimensions to a table of 768 features for each report. We imputed missing values with zero (0) for missing values in the inputs.

7.2.2 Results

Words embeddings

Embeddings generating by CBOW, Skipgram and DrBERT are evaluated through the measure of the quality of the embedding based on downstream tasks (T1). We perform T1 and T2 using the different embeddings methods. The Figure B.6 in the appendix shows the different ROC-curves for both tasks. The DrBERT method has shown better results and will be used for Text BEHRT.

Medical reports embeddings

We first evaluate the quality of the medical reports embeddings obtained by pooling tokens embeddings extracted from DrBERT by visualizing them after their projection into a 2D space using t-SNE. The proximity of reports within this space corresponds to their semantic similarity. As shown in Figure 7.3, this visualization provides a comprehensive overview of the clustering patterns, demonstrating the potential of DrBERT embeddings in representing French medical text data.

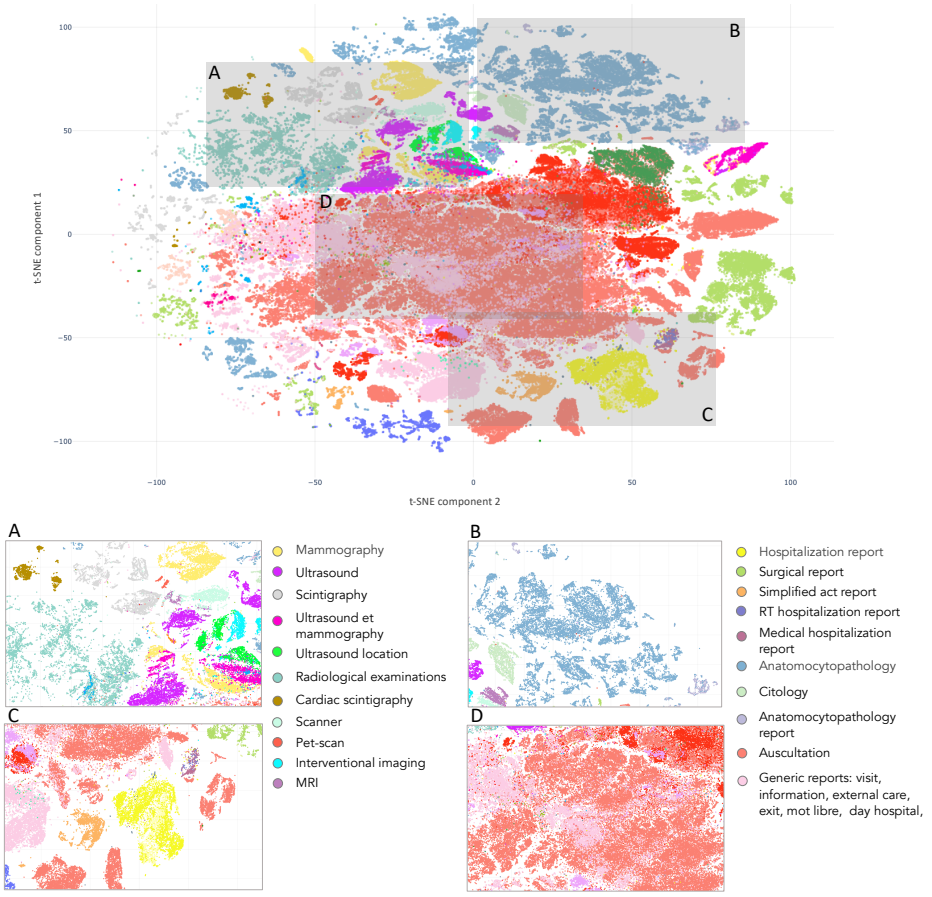


Figure 7.3: t-SNE of Text BEHRT medical reports embeddings. Each panel correspond to a different departments' reports with similar information, cluster together.

This figure shows clusters of reports written in the same departments. Additionally, it displays promixity between clusters that arise from similar departments. The Panel A groups all reports associated with radiology, including "mammography", "MRI", "ultrasound", or "scintigraphy". The same pattern is observed in Panel D, which contains the "generic" reports as those related to "discharge", "external care" or "information", and in Panel B, with clusters relating to cytology ("anatomocytopathology", "cytology"). Lastly, Panel C displays reports from various departments positioned closely together.

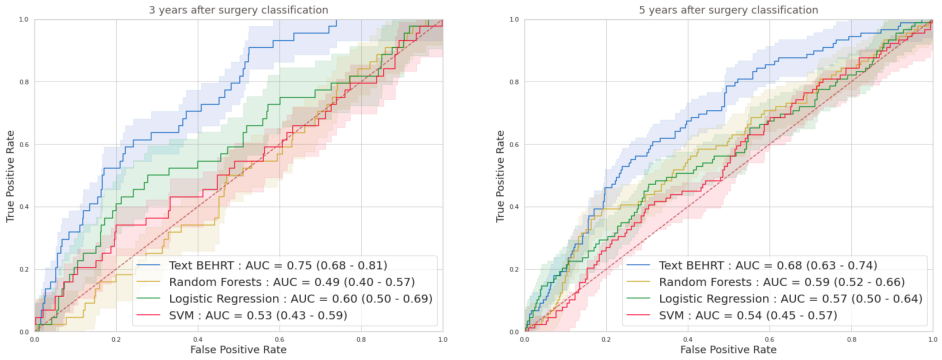


Figure 7.4: ROC curves for baselines and Text BEHRT, for predicting disease-free survival 3 years (T1, left) or 5 years (T2, right) after surgery.

DFS classification

Comparison with state-of-the art predictive algorithms We identified the optimal hyperparameters for Text BEHRT that yield to the highest performance for the validation set. For T1 and T2, Text BEHRT was trained using Adam optimizer with a batch size of 32, and a learning rate of 5.10^{-4} and 1.10^{-3} respectively. The number of epochs is respectively equal to 99 and 94.

Figure 7.4 shows how Text BEHRT performs on the test sets for these two tasks, in comparison to baseline ML models (random forests, SVM and logistic regression) trained on report embeddings as described in Section 7.2.1.

Text-BEHRT clearly outperforms all the baselines in terms of ROC curve.

Model interpretation We also attempted to interpret Text BEHRT using the IG method in the CAPTUM library [Kokhlikyan *et al.* (2020)]. The example shown in Figure 7.5 shows a list of reports, colored in green, that the model has used to predict a negative DFS status. The reports in red have been predictive for a positive DFS status. The remaining reports in white are not predictive for the model. The example belongs to a patient with a moderate prognostic group (I) (NPI=MPGI), aged between 80 and 85 years old.

Overall, the model mostly relied on the entire sequence of the reports from the diagnosis to the index date (1 year after the fisrt surgery) to make its prediction, which is represented by the CLS token. In fact, this interpretation pattern have been found in many true positive (TP) samples for both of the tasks.

Moreover, the model relies on other specific reports contents to predict the relapse, including an echography that classify the initial cancer class

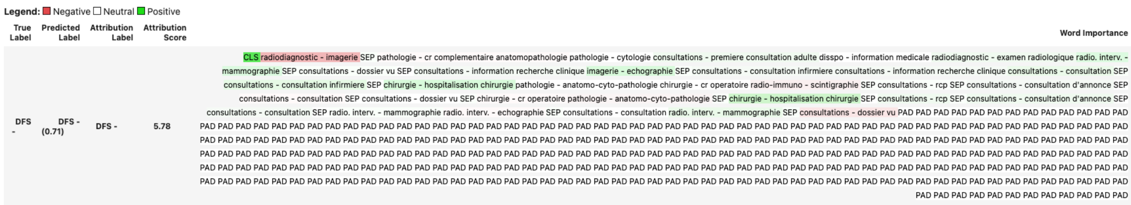


Figure 7.5: Interpretation example for a true positive (TP) sample in T1

Corps du document

SEIN DROIT : Sein en involution adipeuse partielle avec contingent glandulaire inférieur à 50% Dans le quadrant supéro-interne : on note une opacité hétérogène avec convergence fibreuse de 040 à 015 mm avec des micro-calcifications supérieures à 10 groupées irrégulières denses en foyer de 025 à 013 mm. L'évolution par rapport aux mammographies antérieures montre une progression des images.

SEIN GAUCHE : **Sein en involution adipeuse partielle avec contingent glandulaire inférieur à 50%.** Il n'existe aucune anomalie mammographique. L'aspect est stable par rapport aux clichés antérieurs.

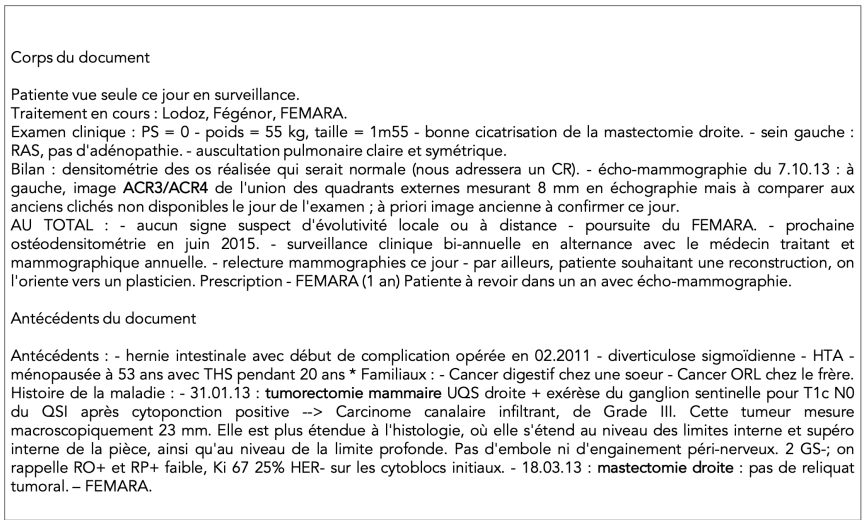
EN RESUME : Relecture de la mammographie de ville du 18/12/12 + compléments CRH. Lésion du versant interne de l'union des quadrants supérieurs droits. **ACR5D** - ACR1G. Ponction/Cytobloc prévus ce jour sous repérage clinique. Le score CRH est R5 droit correspondant à un aspect très probablement malin. Une cytoponction est conseillée. Le score CRH est R0 Gauche correspondant à une mammographie normale.

Figure 7.6: Report from the "mammography" procedure in the "radio. interv" department, index=6

based to the BI-RADS classification scheme to ACR5 (see report in Figure 7.6), that corresponds to a malignant tumor according to the medical images. In fact, the model relies on reports that shows information regarding the characterisation of a suspicious tumor, but this is not in and of itself indicative of a future relapse.

The model has also acknowledged an attribution from two of the patient's last reports (Figure 7.7), indicating another BI-RADS cancer classification post surgery to ACR3/ACR4. These correspond to a presence of a mass in the breast for which short-term monitoring is recommended, possibly suggesting that the surgery was not sufficient to remove the entire tumor. This report also depicts two surgical operations within a 3 months delay, which may potentially mean a more severe cancer. While this information, in and of itself, is not strongly indicative of a future relapse, it could, in conjunction with other reports highlighted by the model, indicate a case that is more difficult to treat.

In order to try to gain global understanding of the model, we investigated the most predictive reports for a positive DFS and for a negative DFS. We



Corps du document

Patiente vue seule ce jour en surveillance.
Traitement en cours : Lodoz, Fégénor, FEMARA.
Examen clinique : PS = 0 - poids = 55 kg, taille = 1m55 - bonne cicatrisation de la mastectomie droite. - sein gauche : RAS, pas d'adénopathie. - auscultation pulmonaire claire et symétrique.
Bilan : densitométrie des os réalisée qui serait normale (nous adressera un CR). - écho-mammographie du 7.10.13 : à gauche, image **ACR3/ACR4** de l'union des quadrants externes mesurant 8 mm en échographie mais à comparer aux anciens clichés non disponibles le jour de l'examen ; à priori image ancienne à confirmer ce jour.
AU TOTAL : - aucun signe suspect d'évolutivité locale ou à distance - poursuite du FEMARA. - prochaine ostéodensitométrie en juin 2015. - surveillance clinique bi-annuelle en alternance avec le médecin traitant et mammographique annuelle. - relecture mammographies ce jour - par ailleurs, patiente souhaitant une reconstruction, on l'oriente vers un plasticien. Prescription - FEMARA (1 an) Patiente à revoir dans un an avec écho-mammographie.

Antécédents du document

Antécédents : - hernie intestinale avec début de complication opérée en 02.2011 - diverticulose sigmoïdienne - HTA - ménopausée à 53 ans avec THS pendant 20 ans * Familiaux : - Cancer digestif chez une soeur - Cancer ORL chez le frère.
Histoire de la maladie : - 31.01.13 : **tumorectomie mammaire** UQS droite + exérèse du ganglion sentinelle pour T1c N0 du QSI après cytoponction positive --> Carcinome canalaire infiltrant, de Grade III. Cette tumeur mesure macroscopiquement 23 mm. Elle est plus étendue à l'histologie, où elle s'étend au niveau des limites interne et supéro interne de la pièce, ainsi qu'au niveau de la limite profonde. Pas d'embole ni d'engainement péri-nerveux. 2 GS-; on rappelle RO+ et RP+ faible, Ki 67 25% HER- sur les cytoblocs initiaux. - 18.03.13 : **mastectomie droite** : pas de reliquat tumoral. – FEMARA.

Figure 7.7: Report from the last "consultation" procedure in the "consultations" department.

set a threshold regarding the given attribution for each medical report. We collect all the reports with an attribution above this threshold. This yielded 921 reports that are predictive for DFS negative in the entire corpus, and 1 720 reports that are predictive for DFS positive. For each reports collection, we determine the 30 most frequent sequences (of 3 to 9 words) for both of the cohorts. We then choose to test the most frequent sequences for the DFS negative cohort that are not found in the DFS positive cohort. We ended up with the following sequences of words, some of which have been obtained with the overlapping resulting sequences:

- "sein en involution adipeuse partielle avec contingent glandulaire inferieur a 50", (*breast in partial adipose involution with less than 50% glandular contingent*)
- "Traitement anterieur par hormone de croissance extractible non facteurs de risque de transmission de la mcj", (*Previous treatment with extractable growth hormone without risk factors for mcj transmission*)
- "[avec] lymphadenectomie axillaire", (*with axillary lymphadenectomy*)
- "syndrome de masse", (*mass syndrom*)
- "[j1] solumedrol 80mg", (*solumedrol 80mg*)
- "lovenox 0 4 ml", (*lovenox 0 4 ml*)

We then plotted 7.8, 7.9, 7.10, 7.11 and 7.12, survival curves to compare the patients that have reports containing these sequences and the patients

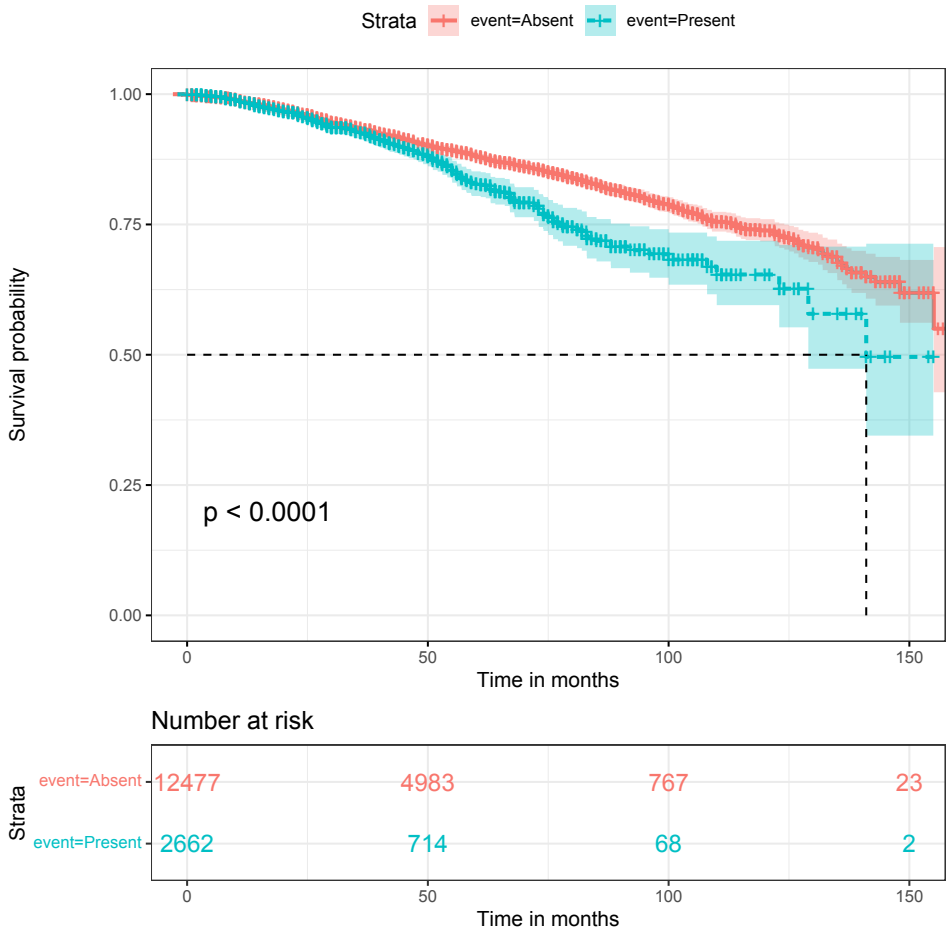


Figure 7.8: Survival plots for the sequence: "sein en involution adipeuse partielle avec contingent glandulaire inferieur a 50", Present or Absent in patients reports

that do not, using the entire cohort. DFS is the event and the log-rank test is used to compare the populations. Only significant results are displayed.

These survival plots depict the disease-free survival probabilities over time for patients that have one of the given sequences in their reports and for patients that do not. The log-rank test indicates a statistically significant difference between the two groups (< 0.05).

For the first example (Figure 7.8), the figure suggests that patients with this feature are most likely to relapse than the other population. Until now this feature has not been defined as a prognostic factors by clinicians. In fact, this feature defines a specific state of breast tissue where the glandular tissue is replaced by adipose tissue. This process naturally occurs with aging and after menopause. When it is partial, the process is not complete yet

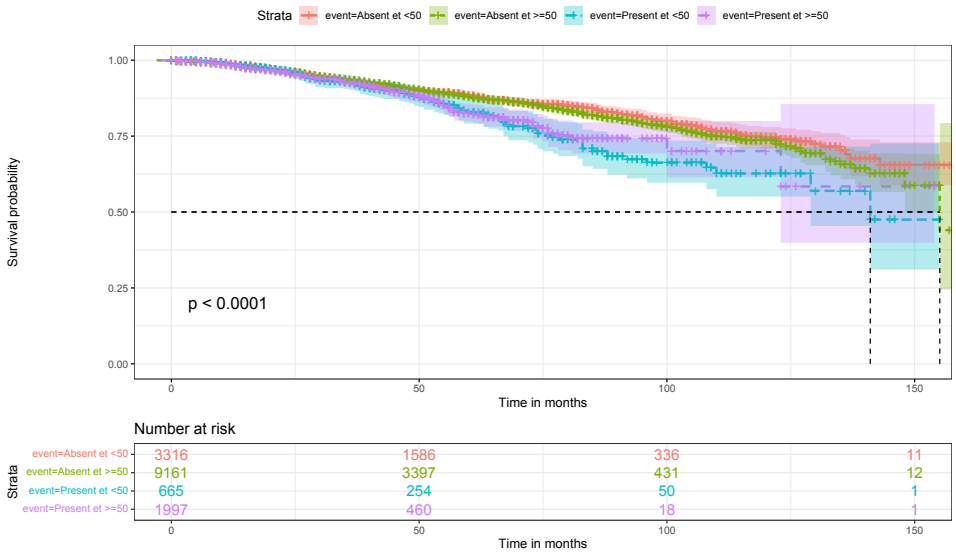


Figure 7.9: Survival plots for the sequence: "sein en involution adipeuse partielle avec contingent glandulaire inferieur a 50", Present or Absent in patients reports, associated with the feature "age"

and the glandular contingent percentage refers to the portion of the breast tissue made up with the glands and ducts. This feature can have an impact on DFS simply because it is related to the patient's age, which is already a prognostic factor. However, when compared with 2 age groups (Figure 7.9), it added more information on the survival than just $>$ 50yo and $<$ 50 yo. Young patients with this feature represent the worst prognostic groups.

The second plot (Figure 7.10) compared a population with the feature "lymphadenectomie axillaire" and a population without. This feature involves removing lymph nodes from the armpits. This information is associated with the potential affection of axillary nodes, which is found to be predictive for BC relapse.

The last survival plots concerns the features "syndrome de masse" (Figure 7.11) and "Lovenox" (Figure 7.12). The mass syndrom reflects the development of a lesion of any kind that leads to the compression of neighboring structures. The presence of a mass syndrom can be linked to a presence of a tumor or to other affections. The presence of a mass syndrom is not necessarily a cause of cancer relapse but it can indicate resistance to treatments as it could be a remaining mass that has resisted chemotherapies or radiotherapies.

Lovenox is a medication used to prevent and treat thromboembolic complications in patients that undergo surgery. It belongs to the low molecular weights heparin (LMWH) class of anticoagulant. Its survival plot depicts a better long-term prognosis for the patients for which this anticoagulant

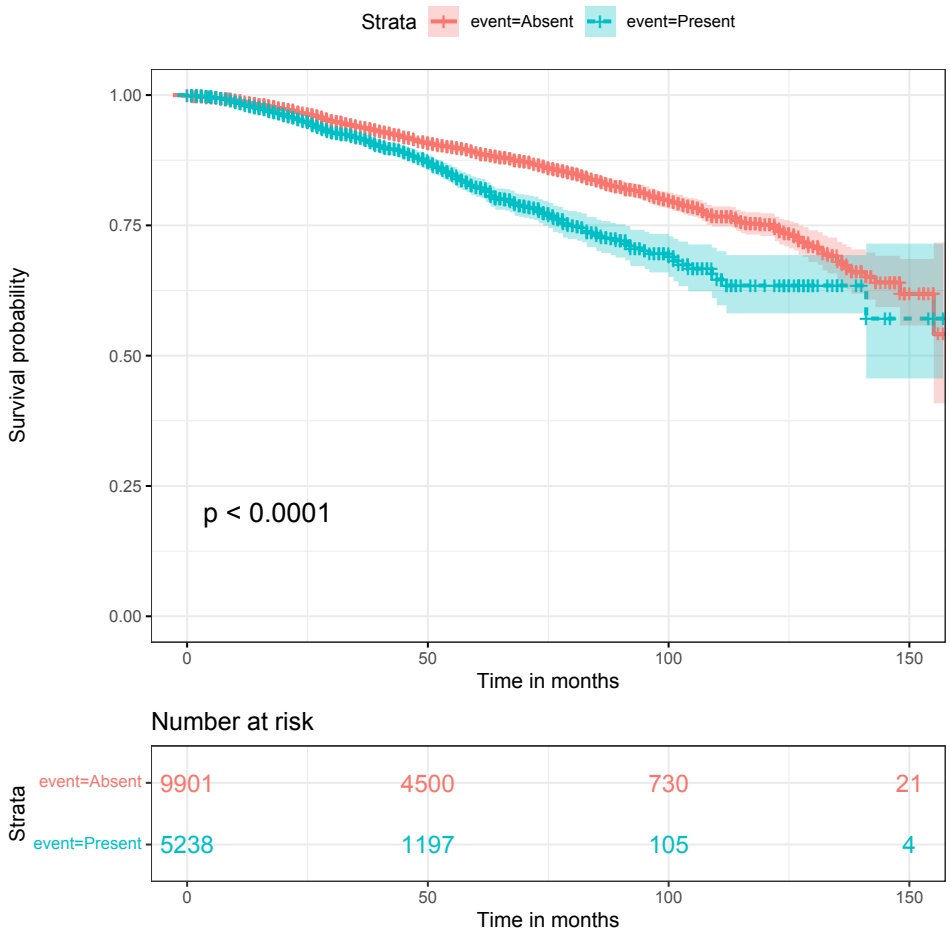


Figure 7.10: Survival plots for the sequence: “lymphadenectomie axillaire”, Present or Absent in patients reports.

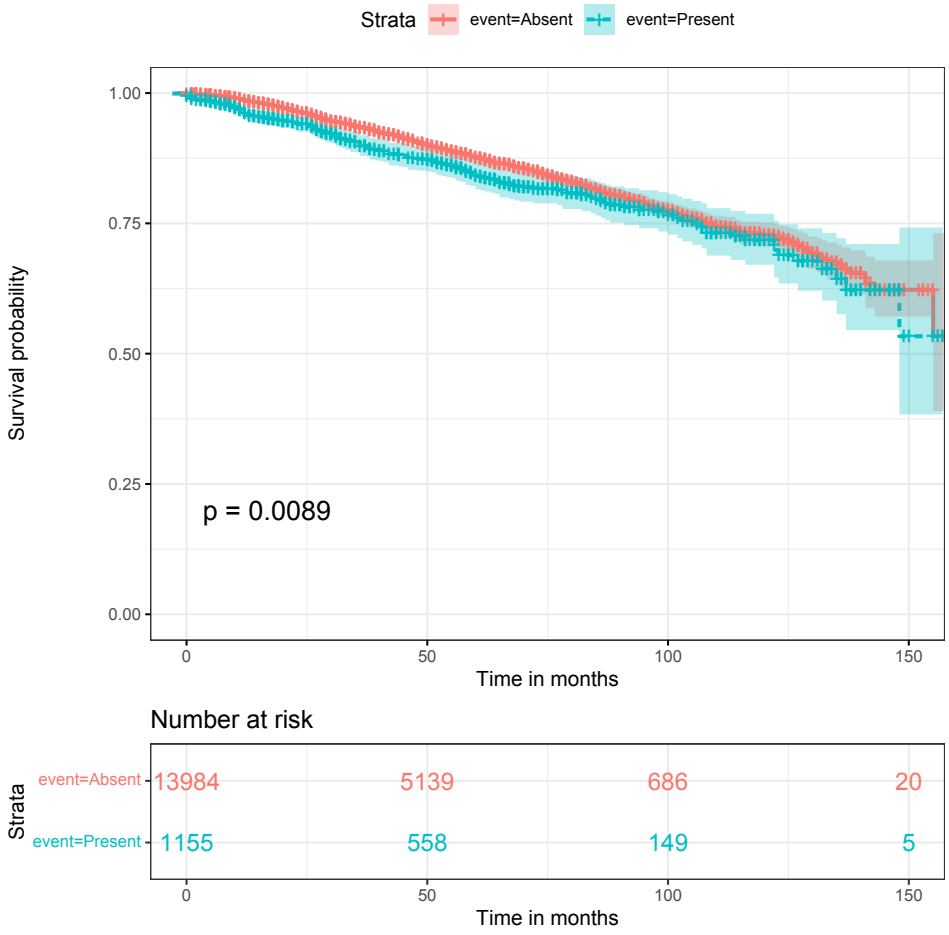


Figure 7.11: Survival plots for the sequence: “Syndrome de masse”, Present or Absent in patients reports

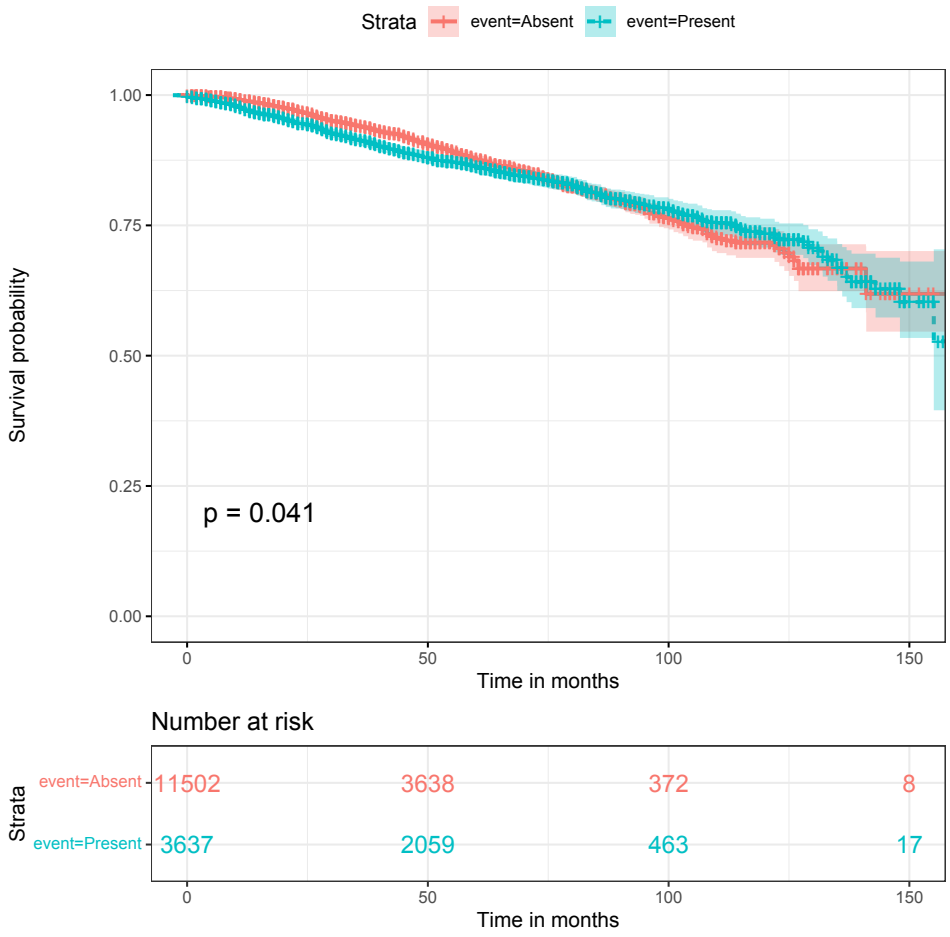


Figure 7.12: Survival plots for the sequence: "Lovenox", Present or Absent in patients reports

have been administrated. As Lovenox is used to treat thromboembolic complications, it has been shown that the occurrence of these complications can be an indicator of a more aggressive disease, which can lead to short-term relapse [Zhang *et al.* (2016)]. On the other hand, its survival plot depicts also a better long-term prognosis for the patients for which this anticoagulant have been administrated. In fact, some studies have shown improved survival rates in cancer patients receiving anticoagulants such as Lovenox [Zhang *et al.* (2016), Akl *et al.* (2008)]. All in all, we need more research to explore the underlying mechanisms by which Lovenox administration may affect tumor progression.

7.3 Conclusion

As previously described, although BERT-based models can be very powerful, one of their keys limitations is the restriction on the maximum number of tokens it can process in a single pass. Limiting the sequence length to 512 tokens helps to manage the model's computational efficiency and memory usage. This limitation makes handling sequences of medical reports really impractical, as a single report can exceed 512 tokens. We could have used models designed to handle longer sequences such as Longformer [Beltagy *et al.* (2020a)] or Big Bird [Zaheer *et al.* (2020b)]. However, the maximum number tokens for those models, albeit superior to BERT, are still not enough for situation such as sequences of reports. Despite this, Text BEHRT outperforms both NPI and classical ML baselines, which suggests it ability to capture the structure of EHR data.

To the best of our knowledge, ours is the first study to use entire free text medical reports (in a language other than English) for breast cancer prognosis. There are several limitations to our approach. First, we used token embeddings learned on French clinical text that are not specific to breast cancer; it is possible that pretraining on breast cancer clinical text could improve the performance of our model. However, this requires considerable resources, both in terms of amount of clinical records available and computing power. Second, we build medical records embedding by simply pooling all token embeddings of a record, which is likely not be optimal for capturing the information contained in a report. Several authors have proposed using convolutional neural networks (CNN) or bidirectional long-short term memory architectures (Bi-LSTM) on top of token embeddings [Gao *et al.* (2021), D'Costa *et al.* (2020), Hui *et al.* (2020)], which typically helps capturing the structure of text documents and could be an interesting future direction to explore for this research. Despite these shortcomings, our results demonstrate the ability of Text BEHRT to capture relevant information, as it performs on with Tabular BEHRT. In terms of interpretability, it is difficult to output a general behavior behind Text BEHRT. In some studied

examples (see appendix), the model relies mainly on reports that contain symptoms related information or reports from imagery. When it occurs before the first surgery, these information are normal, as we are studying a BC treated cohort. If it is not the case, these insights can be further studies in more investigations. They can, in conjunction with other reports, have a impact on the relapse. Moreover, the pooling embedding method that we have used to derive reports embeddings from the reports' content does not help with interpretability, as it does not allow to pinpoint specific parts of a medical report. On another hand, several potentially interesting text features have been found to have a impact on the DFS. Even though these results require more investigations to confirm them as prognostic factors, they seem promising, as shown in survival plots.

In the next chapter, we will discuss how to combine Tabular BEHRT and Text BEHRT into a single multi-modal model.

Chapter 8

Multimodal BEHRT: Transformers for multimodal EHR to predict BC prognosis

---

Abstract:

Electronic Health Records contains a wealth of information accross various modalities, including structured data and unstructured data. The integration of multimodal data can capture the complexity of patient health status. After Tabular BEHRT that provides a comprehensive view of structured EHR information for DFS status prediction, Text BEHRT that used the full spectrum of information avalaible in free-text medical notes, we aggregate these two models through a Multimodal transformer-based model called M-BEHRT (Multimodal-BEHRT). In this chapter, I will present M-BEHRT and its value-added for our task.

Résumé:

Les dossiers médicaux électroniques contiennent une multitude d'informations issues de différentes modalités, dont des données structurées et non structurées. L'intégration de données multimodales permet de saisir la complexité de l'état de santé du patient. Après le modèle Tabular BEHRT, qui fournit une vue complète des informations structurées des DME pour la prédiction de la DFS, et le modèle Text BEHRT, qui utilise tout le spectre des informations disponibles dans les notes médicales en texte libre, nous agrégeons ces deux modèles à travers un modèle basé sur un transformers pour les données multimodales appelé M-BEHRT (Multimodal-BEHRT). Dans ce chapitre, je présenterai le M-BEHRT et sa valeur ajoutée pour notre tâche.

Contents

In this chapter, I propose M-BEHRT, which models multimodal patient trajectories as a sequence of medical visits, which comprise a variety of information ranging from clinical features, results from biological lab tests, medical department and procedure, and the content of free-text medical reports.

8.1 Introduction

As discussed in previous chapters, many studies on cancer prognosis prediction have integrated multi-modal data, including in deep learning models. However, to the best of my knoweldge, few studies have combined clinical / biological information with free-text medical reports. This can be explained by the limited access to comprehensive multi-modal datasets, namely, much of the EHR data may not readily available for research, as well as the limited sample sizes, which can affect the performance or the generalizability of models.

In this chapter, we present a model that combines clinical, biological and free-text clinical notes in a ensemble model that will combines the predictions of Tabular BEHRT and Text BEHRT through a meta-learner to get predictions from all the different modalities. We called this model M-BEHRT. We evaluate M-BEHRT on the same tasks as Tabular BEHRT and Text BEHRT: DFS status prediction 3 and 5 years after the first surgery, respectively T1 and T2.

8.2 Relapse classification

The final stage of this thesis is to combine information derived from the two models: Tabular BEHRT and Text BEHRT, which aim to harness the complementary strengths from the the diverse modalities used, thus potentially enhancing the predictive power and robustness of our approach. After training Tabular BEHRT and the Text BEHRT , we integrate the two distinct modules and Text BEHRT using a cross-attention module [Chen *et al.* (2021)] (see section xx in chapter 3).

As show on Figure 8.1, logits from structured data trajectories and the text trajectories are computed using their respective models. The cross-attentions layer calculates attentions with one model's logits as key, and the other model's logits as value and query. We recall that words embeddings used for Text BEHRT model are provided by DrBERT pretrained model, which have embedding size set to 768. Moreover, afterthe hyperparameters tuning step, the most suitable hidden size for Tabular BEHRT input vector is 144. To compute cross-attentions scores, all models' logits must have same size, therefore, logits from Text BEHRT are fed through a single feed-forward layer to obtain an embedding of the same size as logits from Tabular BEHRT.

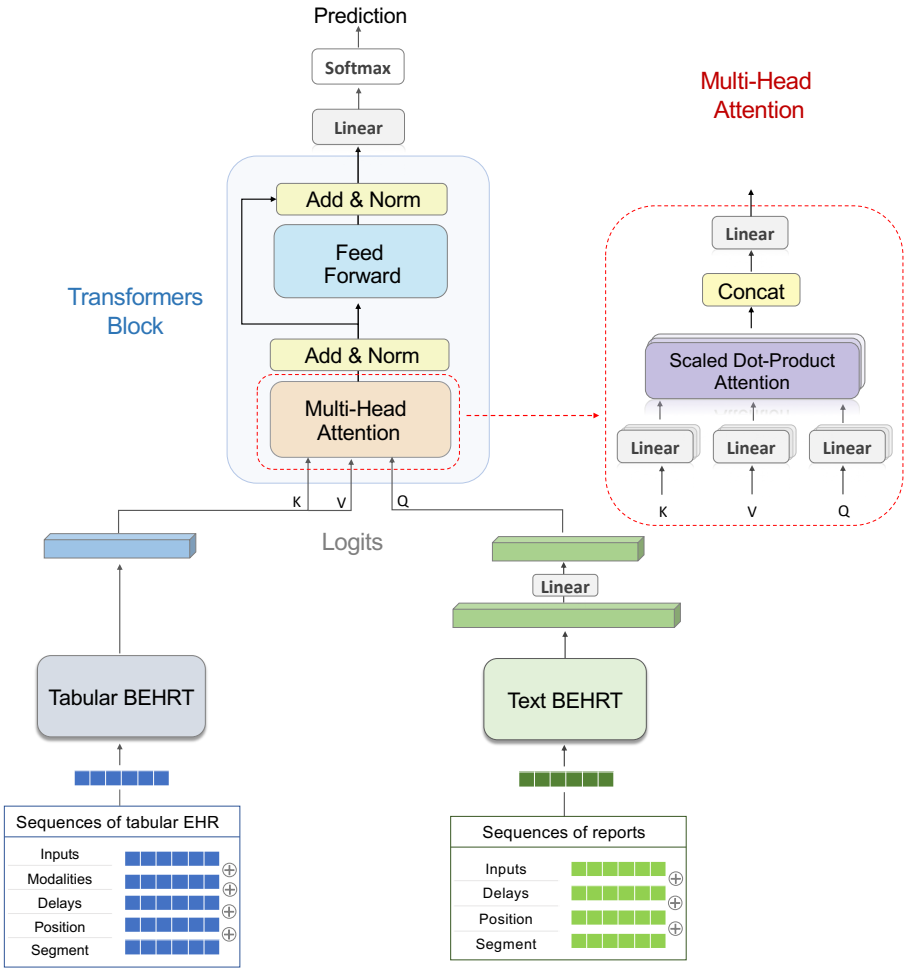


Figure 8.1: M-BEHRT architecture

8.2.1 Implementation details

For the multi-modal transformer based model M-BEHRT, the integration of the Tabular BEHRT and Text BEHRT have been done through a cross attention module, as depicted in Figure 8.1. Logits from Text BEHRT with the are projected into a Query space Q, while logits from Tabular BEHRT are porjected into a key K and value V spaces. The attention score is computed for each pair of modalities as specified in Chapter 3: $A_{ij} = \text{softmax}\left(\frac{Q_i K_j^\top}{\sqrt{d_k}}\right) V_j$ with d_k being the dimensionality of the key vectors K. The attention score are then fed to a FFNN before to be applied to a Softmax to ensure a probability distribution.

8.2.2 Comparison baselines

To assess the M-BEHRT performance, we compared its performance to machine learning models performance for the same input. In fact, we have define benchmarks models for Tabular BEHRT and Text BEHRT. These models are agregated from both modalities. Outputs from tabular data baselines and from text data baselines (specifically their logits) constitute inputs to a secondary model (meta-model) which makes the final prediction.

8.3 Results

8.3.1 Comparison with state-of-the-art predictive ML algorithms

Bayesian optimization of the hyperparameters on the validation set for the combined model M-BEHRT led us to select the M-BEHRT model trained with a learning rate of 1.10^{-3} and a batch size of 64 using Adam Optimizer, for both tasks. Specifically, for T1, the model have been trained with 6 epochs, while for T2, M-BEHRT have been trained with 18 epochs. The figure 8.2 shows that all models perform significantly better than a random classifier (AUC-ROC of 0.5). Moreover, M-BEHRT outperforms all comparison machine learning models.

8.3.2 Comparison of Tabular BEHRT, Text BEHRT and M-BEHRT

Figure 8.3 compares the ROC curves of Tabular BEHRT, Text BEHRT and their combination M-BEHRT on the test sets for the two DFS prediction tasks.

Although they use different information, Tabular BEHRT and Text BEHRT achieve similar performance on both tasks, highlighting that Text BEHRT

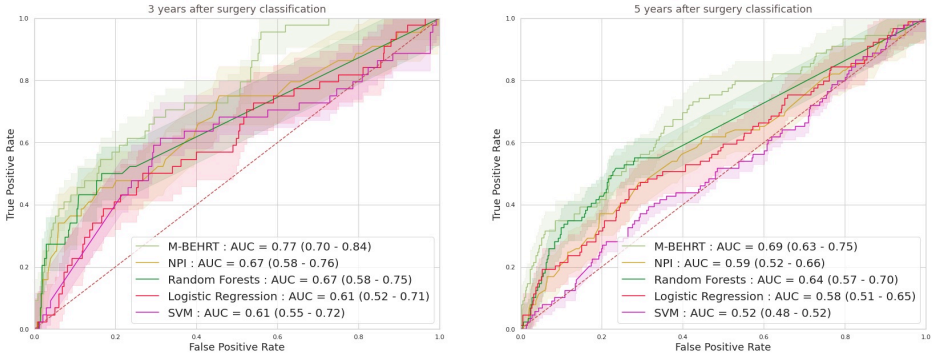


Figure 8.2: ROC Curves comparing M-BEHRT against baselines machine learning models on tasks T1 (left) and T2 (right).

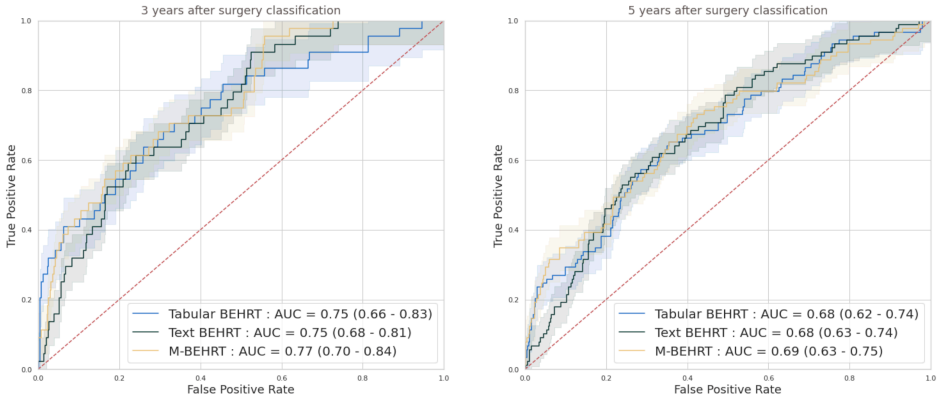


Figure 8.3: ROC Curves comparing Tabular BEHRT and Text BEHRT against their combined model M-BEHRT on tasks T1 (left) and T2 (right).

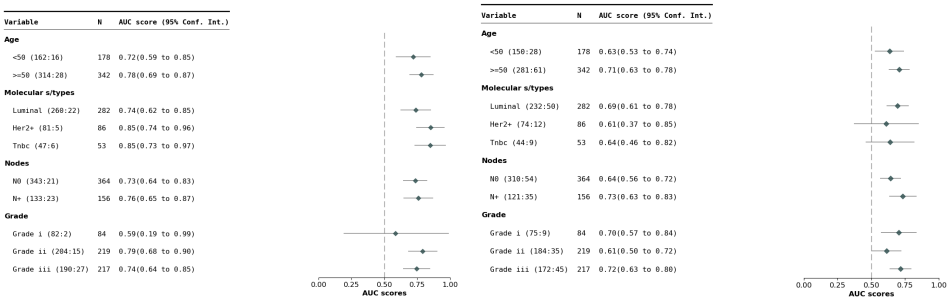


Figure 8.4: AUC-ROC of M-BEHRT stratified by patient age, cancer grade, molecular subtype and node status, for tasks T1 (left) and T2 (right).

can capture relevant information in unstructured medical reports. The combination of both models through cross-attention slightly improves their respective performance, demonstrating the synergistic effect of integrating the strengths of both Tabular and Text BEHRT into a single unified model.

8.3.3 Performance of M-BEHRT per cancer subtype

Figure 8.4 presents the AUC-ROC of M-BEHRT on the test set, stratified by patient age, tumor grade, molecular subtype, or node status. M-BEHRT is better at predicting DFS at three years on older patients, with at least one affected lymph node. Stratification of results by NPI range is available as Supplementary Figure S5.

8.4 Conclusion

In this thesis, we proposed several novel deep learning architectures inspired by BEHRT to model patient trajectories using multimodal data extracted from Electronic Health Records. As the original BEHRT model, Tabular BEHRT considers structured data to describe each medical event. In addition, it considers multiple modalities (biological lab results, clinical information, department and procedure names) simultaneously. By contrast, in Text BEHRT each visit is described via the content of free text medical reports. Finally, M-BEHRT combines both models through cross-attention. Our work is motivated by applications to oncology, and applied to the prediction of disease-free survival for breast cancer patients.

M-BEHRT uses a cross-attention module to perform the multimodal fusion between the two models. This approach allows the contextual integration of information from both transformers, i.e, that each model can attend information from the other model, and thus enable a better exploitation of the complementarity between each inputs. Although the reduction of the report embedding dimensionality from 768 (as provided by DrBERT) to

144 through a linear layer to accommodate the cross-attention module may result in a reduction of available information, the fusion of Tabular BEHRT and Text BERT slightly improves the overall performance. Compared to classical machine learning methods, M-BEHRT is therefore able to capture the sequential aspect of patient data throughout their medical journey, resulting in improved performance.

Using very different information, Tabular BEHRT and Text BEHRT achieve AUCs on a held-out data set of 0.75 for the prediction of DFS 3 years after surgery and 0.68 for the prediction of DFS 5 years after surgery. Combining them in M-BEHRT slightly increases predictive power, reaching AUCs of 0.77 and 0.69 for these same tasks. Overall, our study highlights the potential to predict DFS using solely longitudinal sequence of medical visits and evolution of clinical information and biological measurements.

Perhaps surprisingly, we do not see the same drastic increase in performance between Tabular BEHRT and M-BEHRT as others have observed in multimodal prediction of breast cancer prognosis when augmenting clinical data with imaging data [Rabinovici-Cohen *et al.* (2022b), Han *et al.* (2024a)], although Text BEHRT leverages medical reports from radiologists or cytopathologists, which are based on medical images. Although this could be due to the aforementioned limitations of Text BEHRT, this could also be because Tabular BEHRT already achieves much better performance than models based solely on static clinical data.

To date, most of the multimodal prognosis models for breast cancer use various types of medical images, as well as sometimes genetics data, combined or not with tabular information (biological measurements, clinical features). Moreover, endpoints vary between studies: DFS, but also overall survival or recurrence (sometimes separated between local, regional and distant); which can be measured 3 or 5 years after surgery as in the present work, but also at different time points. Finally, different studies use different criteria inclusions. All in all, this makes comparing our performance to other studies challenging. However, we note that M-BEHRT achieves better performance for the prediction of DFS after three years than the recent work of [Han *et al.* (2024b)], which uses ultrasound and mammography images combined with clinical, pathological and radiographic characteristics and reports an AUC of 0.739 on a held-out test set. In addition, the performance of M-BEHRT is in the same ballpark as that of [Rabinovici-Cohen *et al.* (2022b)], which predict recurrence at five years in patients who receive neo-adjuvant chemotherapy (AUC of 0.75 on a held out data set) using clinical features, immunohistochemical markers, and multiparametric magnetic resonance imaging, or [González-Castro *et al.* (2023a)], which achieve an AUC of 0.807 also for predicting recurrence at five years, but considering all cancer patients and using clinical features, immunohistochemical markers, and descriptors of clinical history such as the number and type of therapies.

As previous models, we stratified the data based age, grade, number

of lymph nodes involved, molecular subtype. We found that as Tabular BEHRT, M-BEHRT is better at predicting relapse on older patients with more aggressive disease (at least one lymph node involved). Moreover, M-BEHRT is better at predicting relapse after 5 years than after 3 years for luminal tumors, suggesting that it correctly identifies predictive factors with long term influence for these tumors that tend to recur later than others [Ignatov *et al.* (2018)].

There are some limitations to our study. In particular, our findings are restricted to a very specific cohort of patients who received adjuvant chemotherapy. We also have not been able to validate our findings on an external validation group, due to privacy concerns limiting the access to EHR of other centers; it is very possible that our models have captured idiosyncrasies of Institut Curie that do not apply to patients from other hospitals. However, our work shows that it is possible to learn from multimodal patient trajectories built from dynamic tabular data and the content of free-text reports written by practicioners at each medical visit, and paves the way for future research in understanding breast cancer prognostic factors.

Chapter 9

Conclusion and perspectives

In my thesis, I explored the application of different multimodal machine learning models to the prediction of breast cancer relapse. My main goal was to leverage the rich and diverse information available in electronic health records (EHRs) to improve the accuracy and reliability of Breast Cancer relapse prediction. This opens up opportunities to explore more advanced algorithms and address to many challenging questions.

9.1 Results of the thesis

In **Chapter 2**, I provided a general context about breast cancer, its negative impact on global health, its characteristics and its different treatments strategies. This highlighted the pressing need to find a solution for better managing the disease. I also introduced important notions in EHRs, and discussed the challenges that arise when working with them. In spite of these challenges, it remains necessary to continue to dig into these EHRs to search for prognostic factors or treatment responses that will be helpful for BC management. Most recent studies in personalized medicine, immunotherapy, early detection and so forth have made great strides in understanding the disease mechanisms, but BC is a complex disease and the EHR remains definitely unexploited. Further investigations in that direction was needed.

In **Chapter 3**, I gave an overview of the mathematical and ML knowledge that underlie this thesis. I presented the machine learning models that I have used to predict DFS status, including both the more "classical" ML models and the more "complex" deep learning models such as transformers-based models. As I have used these models to work with multimodal data, I presented the various possible methods to combine those modalities. Finally, I also described the methods used to understand the model behavior (interpretability), which is important when dealing with this kind of data.

Chapter 4 was about the PhysioNet/Computing in Cardiology challenge, which aimed to predict mortality of ICU patients given a multimodal health data. I worked on this challenge during the first year of this PhD, and it has helped me to understand EHRs characteristics, specially when sequential data is involved.

In **Chapter 5**, I built machine learning models that integrated clinical features, sequential biological information and textual reports. I performed early and late integration to integrate these different modalities. First, the

multimodal models outperform unimodal models. This demonstrates the value of integrating diverse patients information types. Second, the early integration method performs better than the late integration method, which leads us to say that the interaction between the different modalities is more efficient in our situation for this given task. Finally, I presented a set of features that have been found to be predictive for the DFS status. Those features come from the different modalities, highlighting again the usefulness of combining different modalities, which contain complementary information. Furthermore, some of them have already been shown to be prognostic in the litterature, which gives us confidence in our model; in addition, others were new, and can be used for further investigations.

In the following chapters (6, 7, 8), I used a different, sequential representation of EHRs with the aim of detecting new pattern within the patient trajectory. Moreover, transformers-based models having shown great promise in medical applications, I adapted a BERT-based model for learning from patient trajectories and applied it to the prediction of DFS status. This makes it possible to detect the events, or reports in the medical history, that have been somehow related by the model to a relapse event.

9.2 Future work and Perspectives

Overall, we were able to make use of the available EHRs data for relapse prediction. We presented different models and 2 main approach to tackle this task.

While classical multi-modal ML models perform well in integrating the clinical, biological and textual information, M-BEHRT has been able to take account the natural sequential representation of EHR. We showed the efficiency of our model compared to baselines and we were able highlighted several predictive features for DFS.

Future research could focus on the transformers-based models, namely Tabular BEHRT and Text BEHRT. *First*, the integration method used in Tabular BEHRT is an early integration method. The different tokens from the different modalities are merged together in the same input sequence, which leads to a long sequence in terms of number of tokens. Knowing the BERT number of tokens' limitation, a late integration method can be applied to the input. It will lead to using shorter sequences, and thus, the ability to take account most of the patient history while doing the prediction. In addition, this could also open the possibility to add new features such as the different components of the NPI separately to the description of each visit, which could also improve the model.

Secondly, we have found that the biological markers were not strongly related to the DFS prediction. We assume that it is due to their coarse representation, which is unsufficient to clearly define a DFS positive or DFS

negative. For certain of the features (CA15-3 or LEUK for instance), where the normal range contains patients with both positive and negative DFS status. In further investigations, these markers could be categorized depending on statistical features (quantiles, median for instance), or using key point in their distribution curve, that split the best the two classes.

Third, regarding Text BEHRT, the reports embeddings have been calculated using an available open source pretrained model: DrBERT. DrBERT has been pretrained on a large French medical corpus, which is not specialized in oncology. An important step for the model improvement will be to fine-tune the DrBERT pretrained model on a specialised corpus (the Institut Curie reports database for instance). This can make the reports embeddings more accurate and then lead to better performances.

Moreover, when dealing with sequences of reports, the biggest limitation in Text BEHRT was the number of tokens. In our task, it is important to consider the whole trajectory (or most of it) to make the prediction. On the other hand, we are aware that some sections of the reports do not contribute to the task. It will therefore be interesting to build a method to assess important sections in reports. This can be achieved through automatic text summarization methods, but the evaluation of such tasks on a corpus this size might be tricky. We could also use information extraction methods to build more specific summaries; for instance extracting medical concepts such as symptoms, medications or medical procedures, through medical specialized named entity recognition (NER) algorithms. We could then build a model that will select the neighboring words around the most important extracted concepts. This can also been done using a prior "white box" method that will use words vectors as inputs (TF-IDF for example) to perform the tasks (T1 and T2) with a subset of the corpus. Finally, the most important features (words) from this model can be used to derive reports' summary in the remained subset of the corpus. Such a strategy will allow to only acknowledge important parts in patients trajectories and thus have shorter sequences. Their length might be of more than 512 tokens, but might be small enough that we can consider using BERT-based models adapted for longer sequences such Longformer [Beltagy *et al.* (2020b)] or Big Bird [Zaheer *et al.* (2020a)]. And the main adjustment that derive from this strategy is the use of word embedding instead of reports embedding, which will make the model more efficient.

And *Finally*, we can think of validating the developed models with diverse patient populations, that is to say, with one or more other cohorts, coming from different clinics. This could still be difficult to achieve, as EHRs are characterized by many ethical considerations. We would also need to validate the features found to be predictive with further tests.

Altogether, further investigations can turn M-BEHRT into a practical clinical application, potentially improving breast cancer patients outcomes through early detection and more personalized treatment strategies.

Appendix A

Classical ML

A.1 ROC-Curves for the different integration methods

We evaluate the different machine learning models for the same test set for the different integration methods.

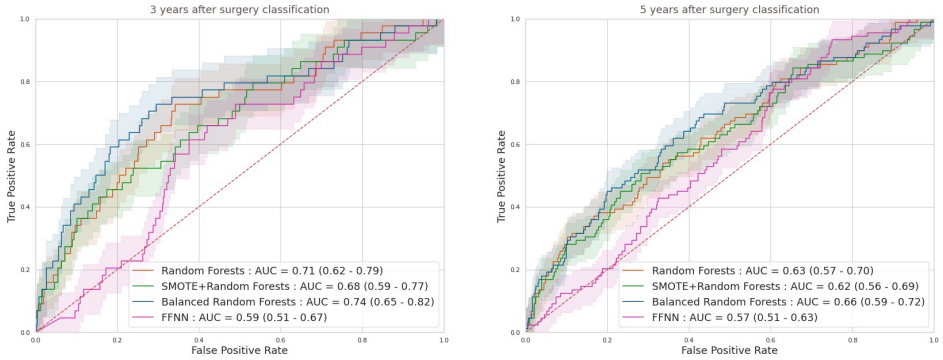


Figure A.1: ROC curves for the different models used for the early integration method for T1 (left) and T2 (right).

We observe on the ROC curves that the Balanced Random Forests outperforms the other models with an AUC of 0.74 for T1 and 0.66 for T2. In contrast, the neural network has an AUC of 0.59 and 0.57 (respectively for T1 and T2), which make it less effective for our tasks. For further experiments, we will use the Balanced Random Forests for both tasks.

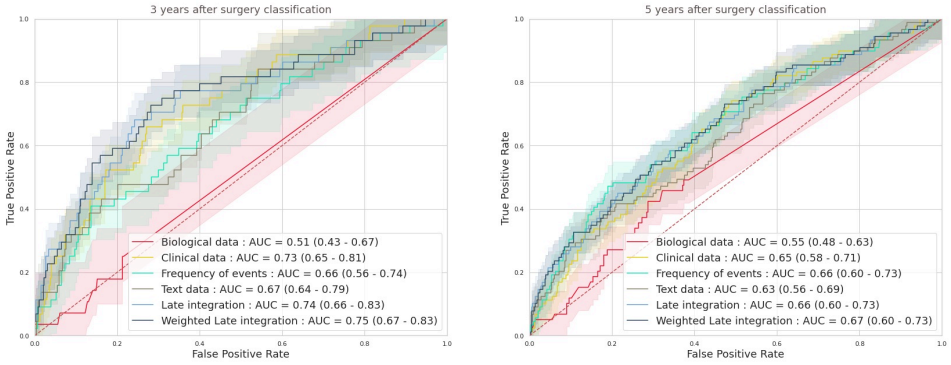


Figure A.2: ROC curves for the best model of each modality and their late integration for T1 (left) and T2 (right)

For the late integration method, the weighted integration method shows better AUC for both tasks (T1 and T2). Moreover, we failed to predict DFS status using biological features' models for all tasks, while clinical data' models outperform other modalities. Globally, this suggests that clinical

data contains more information that allow to distinguish between positive and negative DFS status.

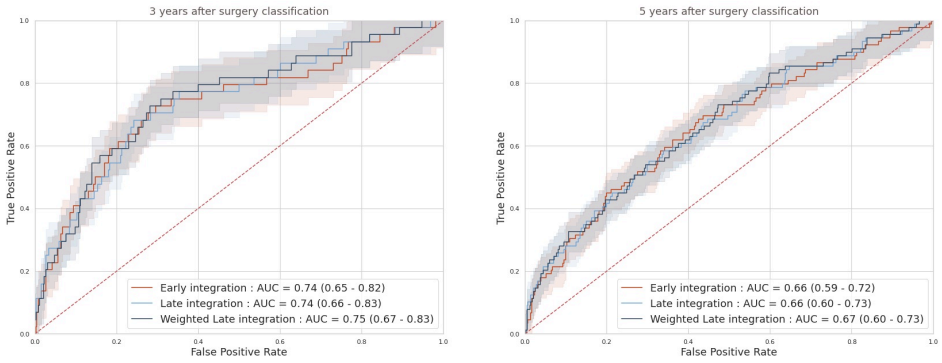


Figure A.3: ROC curves for early and late integration methods (T1 and T2).)

Figure A.3 shows ROC curves for the different integration methods for T1 and T2. Overall, AUC scores are similar accross the different integration methods.

A.2 Separated modalities' top features

We perform interpretation of models built for the different modalities using the random forests features importance. The top 10 features are displayed.

A.2.1 Biological data

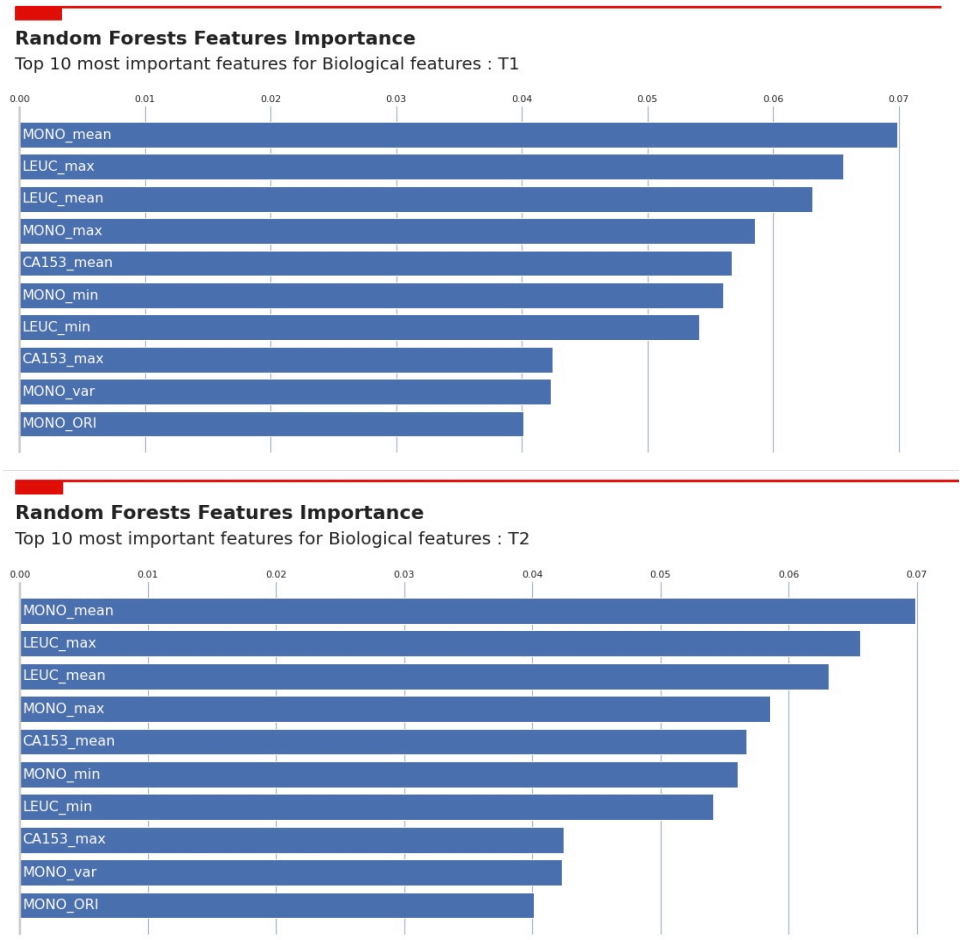


Figure A.4: Top 10 features from biological data' best model

A.2.2 Clinical data

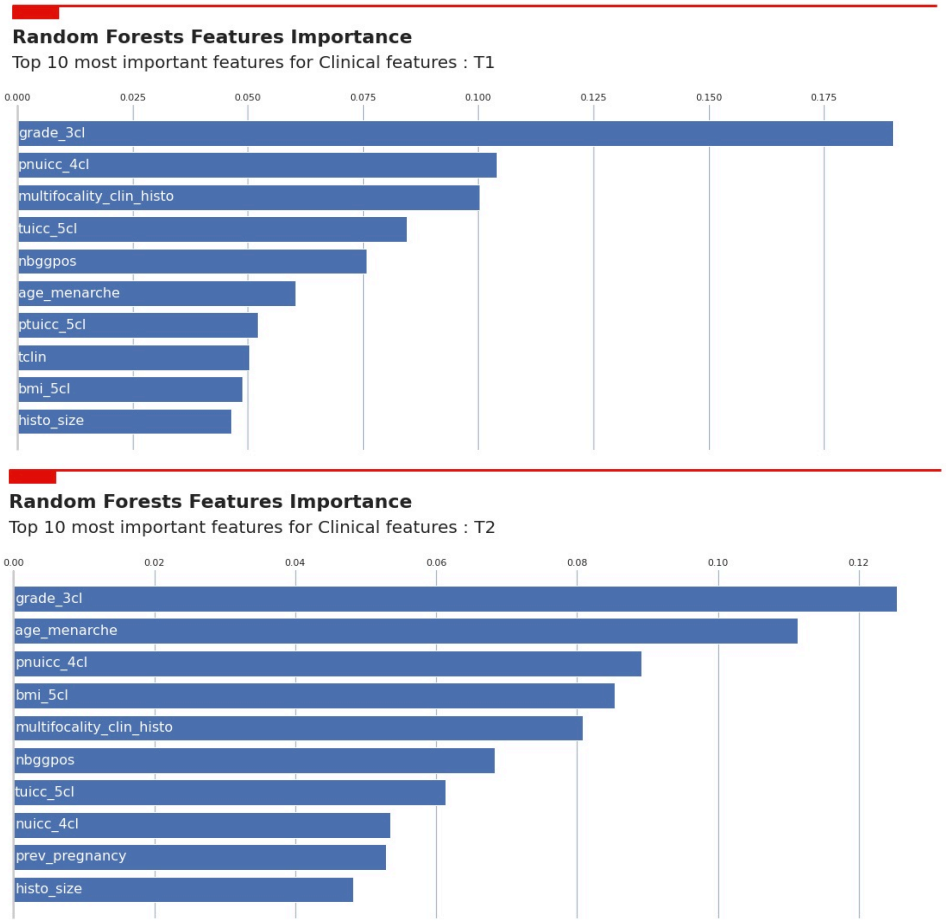


Figure A.5: Top 10 features from clinical data' best model

A.2.3 Frequency of events

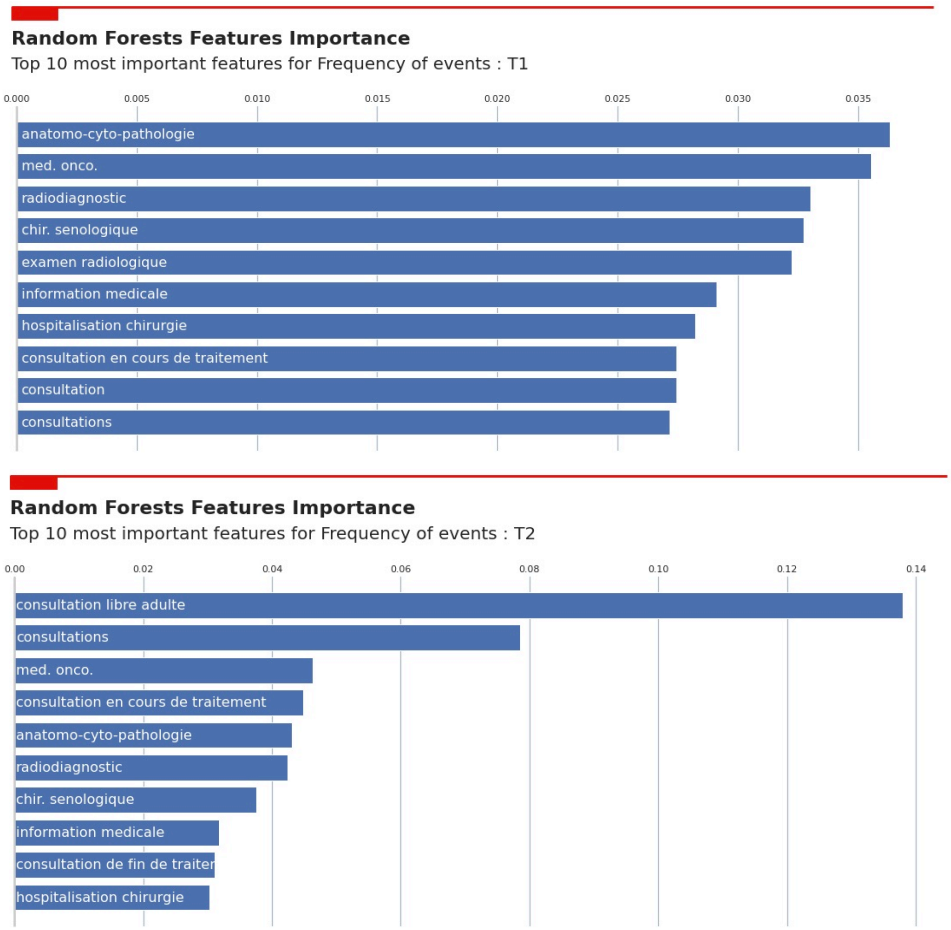


Figure A.6: Top 10 features from the frequency of events modality best model

A.2.4 Text data

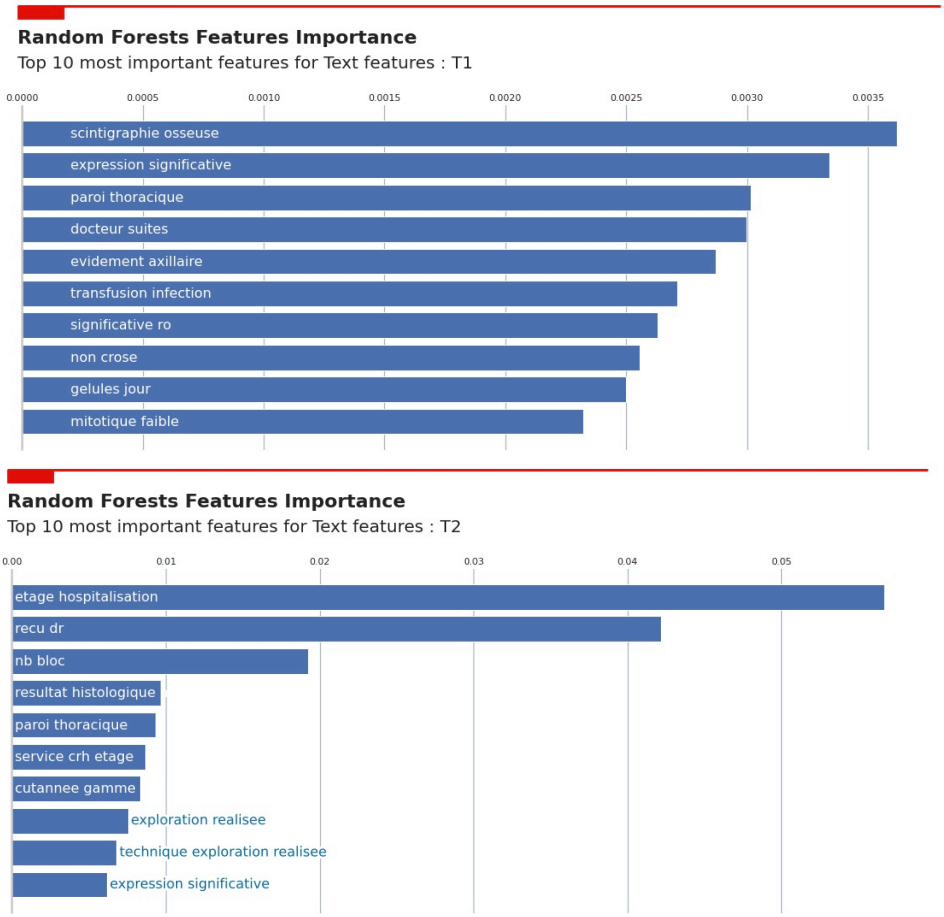


Figure A.7: Top 10 features from the text data modality best model

Appendix B

Pre-trained transformers models for multimodal EHR

Features		Entire dataset		T1 dataset		T2 dataset	
		Mean ± std	N	Mean ± std	N	Mean ± std	N
Age	< 50	58 ± 12	3 982	56 ± 12	2 493	55 ± 13	1 725
	≥ 50		11 168		5 596		3 467
BC subtype	Luminal		9 979		4 866		2 930
	TNBC		1 041		642		446
	HER2+/HR+		681		587		482
	HER2+/HR-		480		415		330
Grades	I		3 473		1 688		1 016
	II		5 911		3 057		1 941
	III		3 119		2 044		1 462
Nodes	N0	0.93 ± 2.49	9 463	1.07 ± 2.74	4 899	1.18 ± 3.01	3 132
	N+		4 045		2 405		1 597
Tumor size (mm)	Clinical	16.89 ± 12.70		17.36 ± 12.97		17.78 ± 13.18	
	Pathological	15.04 ± 12.75		15.63 ± 12.90		16.17 ± 12.94	
Biological values	CA 15-3 (U/ml)	63.39 ± 484.44	8 760	62.85 ± 535.76	3 826	75.34 ± 617.09	2 256
	LEUK (g/l)	6.99 ± 6.82	12 625	6.90 ± 7.49	6 419	6.75 ± 3.60	3 916
	PN (g/l)	718.85 ± 1789.66	5 731	976.17 ± 2007.52	2 385	1105.76 ± 2093.81	1 365
	LYMP (g/l)	289.63 ± 714.26	5 702	405.84 ± 820.08	2 373	463.92 ± 862.37	1 375
	MONO (g/l)	33.29 ± 123.59	11 475	37.54 ± 131.79	5 821	33.29 ± 123.59	3 489
Medical reports	visits	46 ± 33		25 ± 10		25 ± 10	
	reports	62 ± 50		34 ± 15		34 ± 15	
	words/report	172 ± 41		159 ± 37		159 ± 37	

Table B.1: Descriptive statistics of the data sets used in this study, for the full cohort of 15 150 patients, as well as the data set of patients uncensored after 3 years (T1) and 5 years (T2).

B.1 Tabular BEHRT

B.1.1 Datasets

In the next table B.1, we present the different statistics of features used to build the Tabular BEHRT sequences for the pretrained model (MLM) and the classification tasks (T1 and T2).

The following figure B.1 shows the distribution of tokens in tabular sequences for Tabular-BEHRT. We also displayed the number of samples that have more than 512 tokens, which represent 859 samples in T1 and 610 in T2, which correspond to around 10 visits lost for the classification tasks.

We used as datasets for baselines, occurrences of tabular sequences' tokens in each sample B.2.

B.1.2 Relapse classification

In the next figure B.3, we show the APS performance for Tabular BEHRT for T1 and T2. Regarding the APS, Tabular BEHRT still outperform the baselines models and the NPI for both tasks. The next figure B.4 shows the APS performance on the test set for Tabular BEHRT when removing some of the modalities. We present the Tabular BEHRT's APS and other modalities' APS (dNPI, clinical features, biological features and medical reports). As the AUC, this plot shows the highest contribution of clinical data. And on another hand, the biological features do not contribute much to the overall performance.

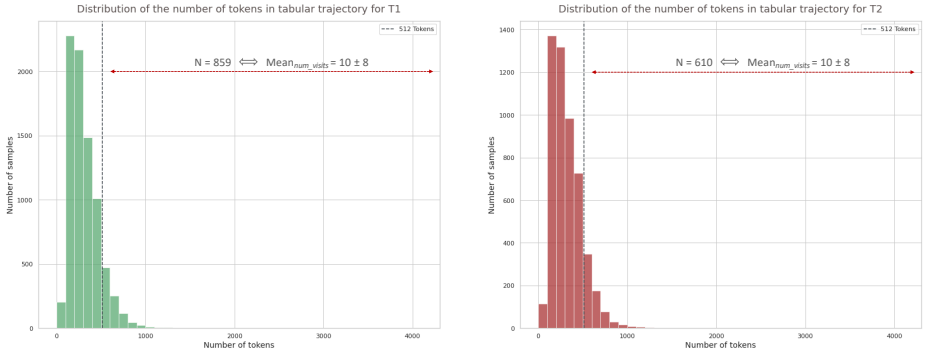


Figure B.1: Distribution of the number of tokens in tabular trajectory for T1 (left) and T2 (right).

Num_dossier	Consultation _rcp	Première consultation adulte	...	RCP	Med.Onco	...	Surgery	Therapy_2	Subtherapy_2	CA153_0	CA153_1	CA153_2	NPI	age
xxxxxxxxxxxx	4	7	...	3	2	...	1	3	4	3	4	0	2	62

Categories | Services | Therapies | Biological features | Other clinical features

Figure B.2: Baselines dataframes

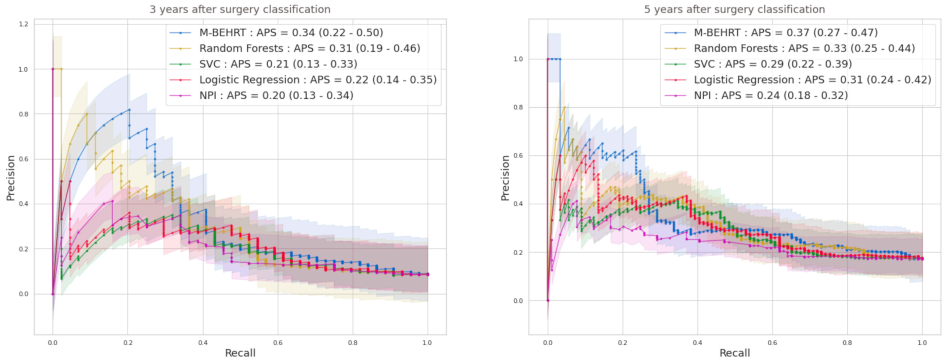


Figure B.3: APS for Binary classification

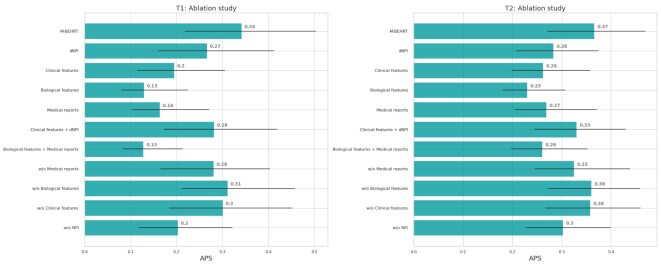


Figure B.4: APS in Tabular BEHRT ablation studies

B.2 Text BEHRT

In figure B.5, we show the distribution of reports which go from 5 to 284 reports and each report contains from 126 to 21301 tokens. This figure shows the potential high length of free-text reports' sequences. BERT-like models have a tokens limitations to 512 tokens, which lead us to use pooling embedding methods to use the entire text trajectory.

The following figure B.6 shows the ROC curves for Text BEHRT used with the three different words embedding methods for T1 and T2. These performance has lead us to the use of DrBERT embeddings for Text BEHRT.

The following figures B.7 and B.8 show statistical analysis that indicates that there is no significant difference in survival between the groups being compared. These features 'cystosteatonecrose' and 'transmissions pour vpa' are part of the most frequent sequences within the negative DFS predictive reports according to Text BEHRT. These figures allow us to say that these features has no impact regarding the DFS.

B.3 Multimodal BEHRT

In figures B.9 and B.10, we present the AUC-ROC of M-BEHRT on the test set stratified by the prognosis group. These plots show that M-BEHRT predicts well clearly defined prognosis groups (GPG for T1 and T2) as Tabular BEHRT relies a lot on the clinical information to output its predictions.

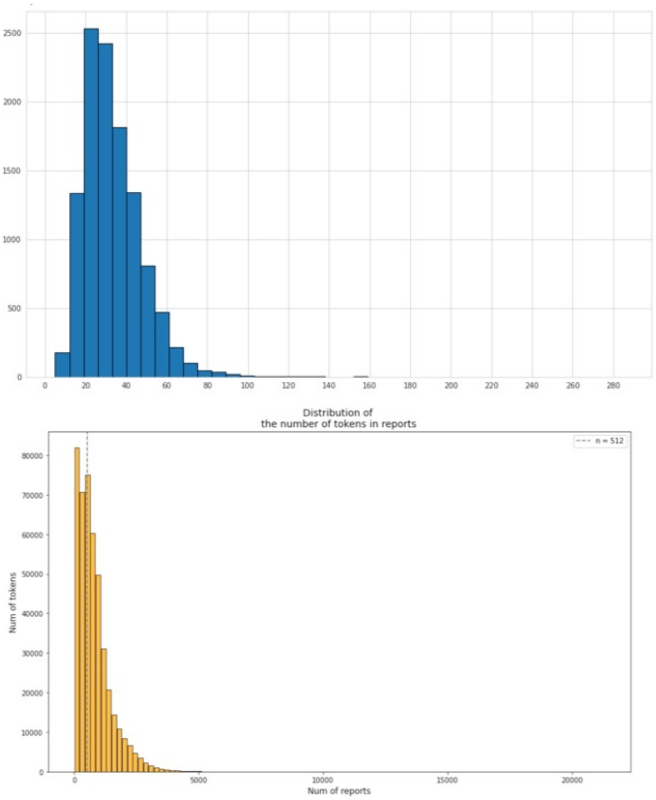


Figure B.5: Distribution of reports in the cohort (top) and the distribution of tokens per report (down).

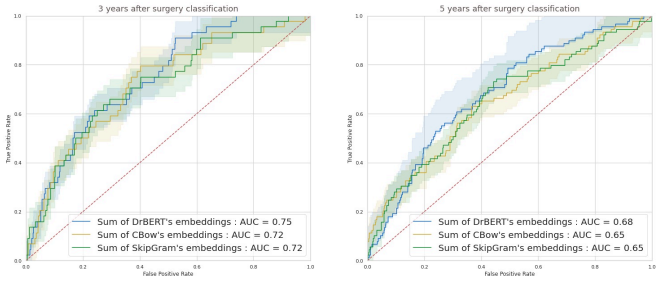


Figure B.6: ROC-curves for the three different embedding methods for T1 and T2.

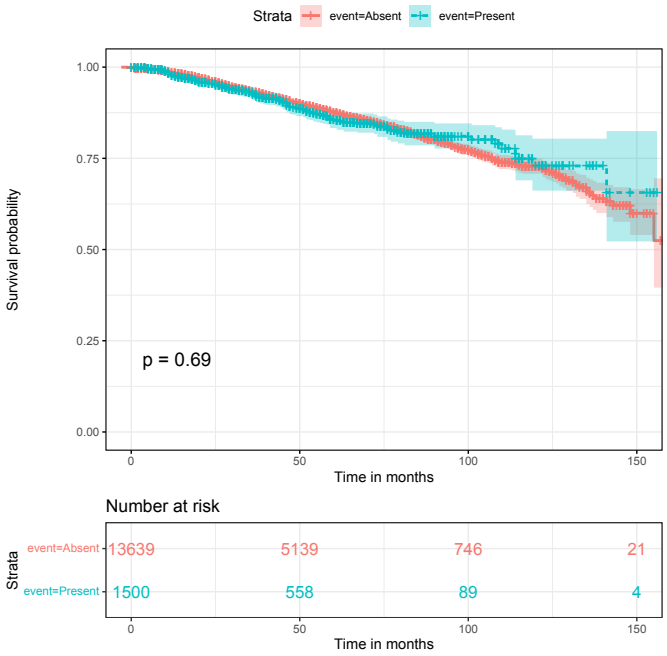


Figure B.7: Survival plots for the reports samples that contain the feature 'cystosteatonecrose' and samples that do not have the feature.

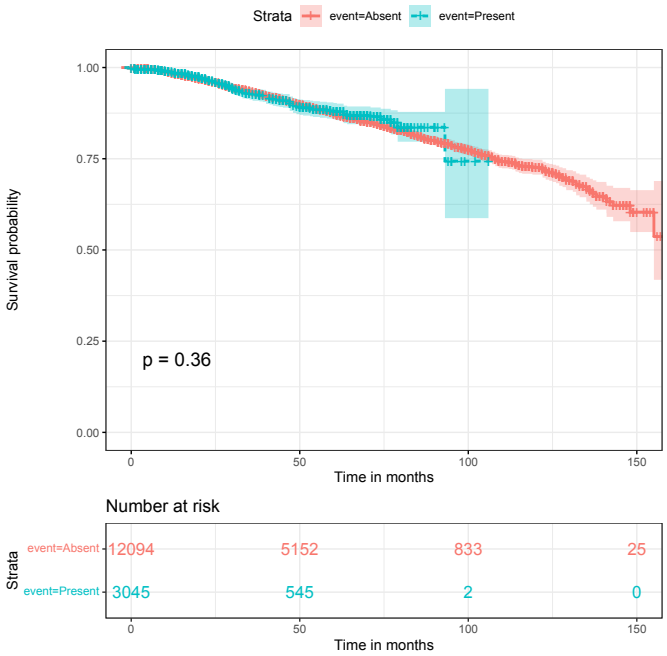


Figure B.8: Survival plots for the reports samples that contain the feature 'transmission pour vpa' and samples that do not have the feature.

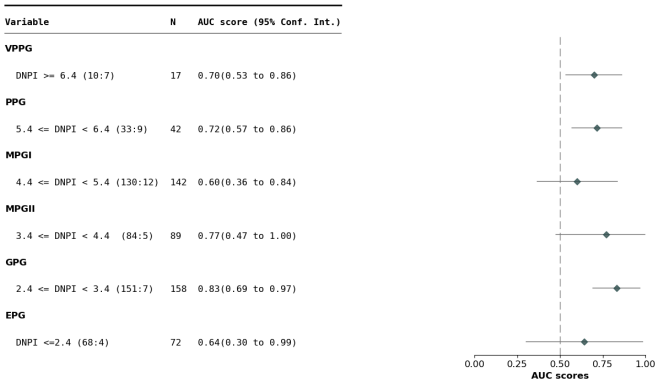


Figure B.9: M-BEHRT performance stratified by the NPI group for T1.

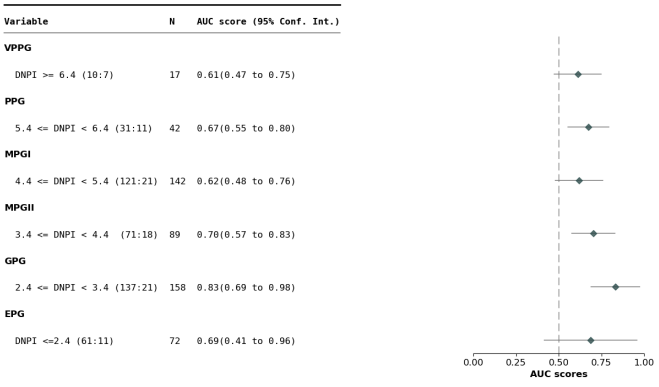


Figure B.10: M-BEHRT performance stratified by the NPI group for T2.

Bibliography

[noa (2001)] *Preoperative chemotherapy in patients with operable breast cancer: Nine- year results from national surgical adjuvant breast and bowel project B-18*Vols **30**, 96 (2001).

[moz (2020)] (2020), *Oncopod – abréviations pour l'oncologie*, https://www.mozocare.com/fr/oncopod/chemotherapy/abbreviations/ (Accessed: 2023-01-30), Mozocare.

[fer (2024)] (2024), *Global cancer observatory: Cancer today (version 1.1)*.

[Cur (1 30)] (Accessed: 2023-01-30), *Curie - annuaire 2023*, https://curie.fr/annuaire-medecins.

[Ins (1 30)] (Accessed: 2023-01-30), *Insee noms*, https://www.insee.fr/fr/statistiques/3536630.

[INS (1 30)] (Accessed: 2023-01-30), *Insee prenoms*, https://www.insee.fr/fr/statistiques/7633685?sommaire=7635552.

[vid (1 30)] (Accessed: 2023-01-30), *Le Dictionnaire VIDAL*, https://www.vidal.fr/medicaments.html.

[ush (1 30)] (Accessed: 2023-01-30), *Usito, dictionnaire général de la langue française*, université de Sherbrooke https://usito.usherbrooke.ca/.

[wik (1 30)] (Accessed: 2023-01-30), *Wikipédia – l'encyclopédie libre*, https://fr.wikipedia.org.

[Abrial *et al.* (2012)] C. Abrial, Q. Wang-Lopez, et al., *270p - long-term overall survival (os) and disease-free survival (dfs) according to pcr in breast cancers subtypes: Luminal a and b, triple negative and her2+, treated by neaodjuvant chemotherapy (nct)*, Annals of Oncology **23**, ix102, abstract Book of the 37th ESMO Congress Vienna, Austria, 28 September - 2 October 2012 (2012).

[Aguirre *et al.* (2019)] R. R. Aguirre, O. Suarez, et al., *Electronic health record implementation: A review of resources and tools*, Cureus (2019).

[Aiello *et al.* (2023)] E. Aiello, L. Yu, et al., *Jointly training large autoregressive multimodal models*, (2023).

[Akl *et al.* (2008)] E. A. Akl, M. Barba, et al., *Low-molecular-weight heparins are superior to vitamin K antagonists for the long term treatment of venous thromboembolism in patients with cancer: a cochrane systematic review*, J. Exp. Clin. Cancer Res. **27** (1), 21 (2008).

[Akshata Desai (2012)] K. A. Akshata Desai, *Triple negative breast cancer – an overview*, Hereditary Genet. (2012).

[Al-Hilli & Boughey (2016)] Z. Al-Hilli & J. C. Boughey, *The timing of breast and axillary surgery after neoadjuvant chemotherapy for breast cancer*, Chinese Clinical Oncology **5** (3) (2016).

[Al-Rawajfah & Tubaishat (2019)] O. Al-Rawajfah & A. Tubaishat, *Barriers and facilitators to using electronic healthcare records in jordanian hospitals from the nurses' perspective: A national survey*, Inform. Health Soc. Care **44** (1), 1 (2019).

[Alpaydin (2018)] E. Alpaydin (2018), *Classifying Multimodal data.*

[Alsohime *et al.* (2019)] F. Alsohime, M.-H. Temsah, et al., *Satisfaction and perceived usefulness with newly-implemented electronic health records system among pediatricians at a university hospital*, Comput. Methods Programs Biomed. **169**, 51 (2019).

[Altmann *et al.* (2010)] A. Altmann, L. Toloşi, et al., *Permutation importance: a corrected feature importance measure*, Bioinformatics **26** (10), 1340 (2010).

[Alvarez-Melis & Jaakkola (2018)] D. Alvarez-Melis & T. S. Jaakkola (2018), *On the robustness of interpretability methods*, in *WHI 2018*.

[Alzu'bi *et al.* (2021)] A. Alzu'bi, H. Najadat, et al., *Predicting the recurrence of breast cancer using machine learning algorithms*, Multimedia Tools and Applications **80** (9), 13787 (2021).

[Amirahmadi *et al.* (2023)] A. Amirahmadi, M. Ohlsson, & K. Etminani, *Deep learning prediction models based on ehr trajectories: A systematic review*, Journal of Biomedical Informatics **144**, 104430 (2023).

[Apley & Zhu (2020)] D. W. Apley & J. Zhu, *Visualizing the effects of predictor variables in black box supervised learning models*, J. R. Stat. Soc. Series B Stat. Methodol. **82** (4), 1059 (2020).

[Archer *et al.* (2011)] N. Archer, U. Fevrier-Thomas, et al., *Personal health records: a scoping review*, J. Am. Med. Inform. Assoc. **18** (4), 515 (2011).

[Ba *et al.* (2016)] J. L. Ba, J. R. Kiros, & G. E. Hinton, *Layer normalization*, (2016).

[Bahdanau *et al.* (2014)] D. Bahdanau, K. Cho, & Y. Bengio, *Neural machine translation by jointly learning to align and translate*, (2014).

[Bai *et al.* (2020)] B. Bai, J. Liang, et al., *Why attentions may not be interpretable?*, (2020).

[Barus (2023)] T. Barus (2023), *pyspellchecker – Pure python spell checker based on work by Peter Norvig*, https://pypi.org/project/pyspellchecker.

[Bear *et al.* (2006)] H. D. Bear, S. Anderson, et al., *Sequential preoperative or postoperative docetaxel added to preoperative doxorubicin plus cyclophosphamide for operable breast cancer: National surgical adjuvant breast and bowel project protocol B-27*, J. Clin. Oncol. **24** (13), 2019 (2006).

[Beltagy *et al.* (2019)] I. Beltagy, K. Lo, & A. Cohan, *SciBERT: A pretrained language model for scientific text*, (2019).

[Beltagy *et al.* (2020a)] I. Beltagy, M. E. Peters, & A. Cohan, *Longformer: The long-document transformer*, arXiv:2004.05150 (2020a).

[Beltagy *et al.* (2020b)] I. Beltagy, M. E. Peters, & A. Cohan, *Longformer: The Long-Document transformer*, (2020b).

[Bergstra & Bengio (2012)] J. Bergstra & Y. Bengio, *Random search for hyper-parameter optimization*, J. Mach. Learn. Res. **13** (null), 281–305 (2012).

[Bishop (2006)] C. Bishop (2006), *Pattern recognition and Machine Learning*.

[Blamey *et al.* (2007)] R. Blamey, I. Ellis, et al., *Survival of invasive breast cancer according to the nottingham prognostic index in cases diagnosed in 1990–1999*, European journal of cancer **43** (10), 1548 (2007).

[Boser *et al.* (1992)] B. E. Boser, I. M. Guyon, & V. N. Vapnik (1992), *A training algorithm for optimal margin classifiers*, in *Proceedings of the fifth annual workshop on Computational learning theory* (ACM, New York, NY, USA).

[Bray *et al.* (2024)] F. Bray, M. Laversanne, et al., *Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries*, CA Cancer J. Clin. (2024).

[Breiman (2001)] L. Breiman, Mach. Learn. **45** (1), 5 (2001).

[Breiman *et al.* (2017)] L. Breiman, J. H. Friedman, et al. (2017), *Classification And Regression Trees* (Routledge).

[Brown *et al.* (2020)] T. B. Brown, B. Mann, et al., *Language models are few-shot learners*, arXiv preprint arXiv:2005.14165 (2020).

[Bruns *et al.* (2018)] E. J. Bruns, A. N. Hook, et al., *Impact of a web-based electronic health record on behavioral health service delivery for children and adolescents: Randomized controlled trial*, J. Med. Internet Res. **20** (6), e10197 (2018).

[Buzdar *et al.* (2005)] A. U. Buzdar, N. K. Ibrahim, et al., *Significantly higher pathologic complete remission rate after neoadjuvant therapy with trastuzumab, paclitaxel, and epirubicin chemotherapy: Results of a randomized trial in human epidermal growth factor receptor 2–positive operable breast cancer*, J. Clin. Oncol. **23** (16), 3676 (2005).

[Callahan & Hurvitz (2011)] R. Callahan & S. Hurvitz, *Human epidermal growth factor receptor-2-positive breast cancer: current management of early, advanced, and recurrent disease*, Curr. Opin. Obstet. Gynecol. **23** (1), 37 (2011).

[Captier *et al.* (2023)] N. Captier, M. Lerousseau, et al., *Models including pathological and radiomic features vs clinical models in predicting outcome of patients with metastatic non-small cell lung cancer treated with immunotherapy*, J. Clin. Oncol. **41** (16_suppl), e21164 (2023).

[Chao (2016)] C.-A. Chao, *The impact of electronic health records on collaborative work routines: A narrative network analysis*, Int. J. Med. Inform. **94**, 100 (2016).

[Chawla *et al.* (2002)] N. V. Chawla, K. W. Bowyer, et al., *SMOTE: Synthetic minority over-sampling technique*, J. Artif. Intell. Res. **16**, 321 (2002).

[Chen & Breiman (2004)] C. Chen & L. Breiman, *Using random forest to learn imbalanced data*, University of California, Berkeley (2004).

[Chen *et al.* (2021)] C.-F. R. Chen, Q. Fan, & R. Panda (2021), *Crossvit: Cross-attention multi-scale vision transformer for image classification*, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 357–366.

[Chen *et al.* (2022)] R. J. Chen, M. Y. Lu, et al., *Pan-cancer integrative histology-genomic analysis via multimodal deep learning*, Cancer Cell **40** (8), 865 (2022).

[Cheng *et al.* (2010)] L. T. E. Cheng, J. Zheng, et al., *Discerning tumor status from unstructured MRI reports—completeness of information in existing reports and utility of automated natural language processing*, J. Digit. Imaging **23** (2), 119 (2010).

[Citi & Barbieri (2012)] L. Citi & R. Barbieri (2012), *Physionet 2012 challenge: Predicting mortality of icu patients using a cascaded svm-glm paradigm*, in *2012 Computing in Cardiology*, pp. 257–260.

[Clark *et al.* (2019)] K. Clark, U. Khandelwal, et al. (2019), *What does BERT look at? an analysis of BERT's attention*, in *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, édité par T. Linzen, G. Chrupała, Y. Belinkov, & D. Hupkes (Association for Computational Linguistics, Florence, Italy) pp. 276–286.

[Clevert *et al.* (2015)] D.-A. Clevert, T. Unterthiner, & S. Hochreiter, *Fast and accurate deep network learning by exponential linear units (ELUs)*, (2015).

[Cortes & Vapnik (1995)] C. Cortes & V. Vapnik, *Support-vector networks*, Machine Learning **20** (3), 273 (1995).

[D'Costa *et al.* (2020)] A. D'Costa, S. Denkovski, et al. (2020), *Multiple sclerosis severity classification from clinical text*, in *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pp. 7–23.

[Dent *et al.* (2007)] R. Dent, M. Trudeau, et al., *Triple-negative breast cancer: clinical features and patterns of recurrence*, Clinical cancer research **13** (15), 4429 (2007).

[Devlin *et al.* (2018)] J. Devlin, M.-W. Chang, et al., *BERT: Pre-training of deep bidirectional transformers for language understanding*, (2018).

[Devlin *et al.* (2019a)] J. Devlin, M.-W. Chang, et al. (2019a), *Bert: Pre-training of deep bidirectional transformers for language understanding*, uRL: https://huggingface.co/google-bert/bert-base-multilingual-cased, arXiv:1810.04805.

[Devlin *et al.* (2019b)] J. Devlin, M.-W. Chang, et al. (2019b), *Bert: Pre-training of deep bidirectional transformers for language understanding*, uRL: https://huggingface.co/bert-base-chinese, arXiv:1810.04805.

[Devlin *et al.* (2019c)] J. Devlin, M.-W. Chang, et al. (2019c), *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Association for Computational Linguistics, Minneapolis, Minnesota) pp. 4171–4186.

[Do *et al.* (2013)] B. H. Do, A. S. Wu, et al., *Automatic retrieval of bone fracture knowledge using natural language processing*, J. Digit. Imaging **26** (4), 709 (2013).

[Doll *et al.* (1994a)] R. Doll, R. Peto, et al., *Mortality in relation to consumption of alcohol: 13 years' observations on male british doctors*, BMJ **309** (6959), 911 (1994a).

[Doll *et al.* (1994b)] R. Doll, R. Peto, et al., *Mortality in relation to smoking: 40 years' observations on male british doctors*, BMJ **309** (6959), 901 (1994b).

[Eloy *et al.* (2023)] C. Eloy, A. Marques, et al., *Artificial intelligence–assisted cancer diagnosis improves the efficiency of pathologists in prostatic biopsies*, Virchows Arch. **482** (3), 595 (2023).

[Ezzati *et al.* (2002)] M. Ezzati, A. D. Lopez, et al., *Selected major risk factors and global and regional burden of disease*, Lancet **360** (9343), 1347 (2002).

[Fernando & Tsokos (2022)] K. R. M. Fernando & C. P. Tsokos, *Dynamically weighted balanced loss: Class imbalanced learning and confidence calibration of deep neural networks*, IEEE Transactions on Neural Networks and Learning Systems **33** (7), 2940 (2022).

[Fernández-Alemán *et al.* (2013)] J. L. Fernández-Alemán, I. C. Señor, et al., *Security and privacy in electronic health records: A systematic literature review*, Journal of Biomedical Informatics **46** (3), 541 (2013).

[Fraser *et al.* (2022)] H. S. F. Fraser, M. Mugisha, et al., *User perceptions and use of an enhanced electronic health record in rwanda with and without clinical alerts: Cross-sectional survey*, JMIR Med. Inform. **10** (5), e32305 (2022).

[Friedman (2001)] J. H. Friedman, *Greedy function approximation: A gradient boosting machine*, Ann. Stat. **29** (5), 1189 (2001).

[Fukushima (1969)] K. Fukushima, *Visual feature extraction by a multilayered network of analog threshold elements*, IEEE Trans. Syst. Sci. Cybern. **5** (4), 322 (1969).

[Gao *et al.* (2021)] S. Gao, M. Alawad, et al., *Limitations of transformers on clinical text classification*, IEEE Journal of Biomedical and Health Informatics **25** (9), 3596 (2021).

[Gariépy-Saper & Decarie (2021)] K. Gariépy-Saper & N. Decarie, *Privacy of electronic health records: a review of the literature*, J. Can. Health Libr. Assoc. **42** (1) (2021).

[Gligorijević & Pržulj (2015a)] V. Gligorijević & N. Pržulj, *Methods for biological data integration: perspectives and challenges*, J. R. Soc. Interface **12** (112), 20150571 (2015a).

[Gligorijević & Pržulj (2015b)] V. Gligorijević & N. Pržulj, *Methods for biological data integration: perspectives and challenges*, J. R. Soc. Interface **12** (112), 20150571 (2015b).

[Goldberg *et al.* (2012)] G. Goldberg, D. Kuzel, et al., *EHRs in primary care practices: benefits, challenges, and successful strategies*, Am J Manag Care **18** (2), e48 (2012).

[González-Castro *et al.* (2023a)] L. González-Castro, M. Ch vez, et al., *Machine learning algorithms to predict breast cancer recurrence using structured and unstructured sources from electronic health records*, Cancers **15** (10), 2741 (2023a).

[González-Castro *et al.* (2023b)] L. González-Castro, M. Chávez, et al., *Machine learning algorithms to predict breast cancer recurrence using structured and unstructured sources from electronic health records*, Cancers **15** (10), 10.3390/cancers15102741 (2023b).

[Grabar *et al.* (2018)] N. Grabar, V. Claveau, & C. Dalloux (2018), *CAS: French corpus with clinical cases*, in *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, édité par A. Lavelli, A.-L. Minard, & F. Rinaldi (Association for Computational Linguistics, Brussels, Belgium) pp. 122–128.

[Grinsztajn *et al.* (2024)] L. Grinsztajn, E. Oyallon, & G. Varoquaux (2024), *Why do tree-based models still outperform deep learning on typical tabular data?*, in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22 (Curran Associates Inc., Red Hook, NY, USA).

[Grossman Liu *et al.* (2021)] L. Grossman Liu, R. H. Grossman, et al., *A deep database of medical abbreviations and acronyms for natural language processing*, Scientific Data **8** (1), 10.1038/s41597-021-00929-4 (2021).

[Han *et al.* (2024a)] J. Han, H. Hua, et al., *Prediction of disease-free survival in breast cancer using deep learning with ultrasound and mammography: A multicenter study*, Clinical Breast Cancer 10.1016/j.clbc.2024.01.005 (2024a).

[Han *et al.* (2024b)] J. Han, H. Hua, et al., *Prediction of disease-free survival in breast cancer using deep learning with ultrasound and mammography: A multicenter study*, Clinical Breast Cancer 10.1016/j.clbc.2024.01.005 (2024b).

[Harnoune *et al.* (2021)] A. Harnoune, M. Rhanoui, et al., *Bert based clinical knowledge extraction for biomedical knowledge graph construction and analysis*, Computer Methods and Programs in Biomedicine Update **1**, 100042 (2021).

[Haybittle *et al.* (1982a)] J. Haybittle, R. Blamey, et al., *A prognostic index in primary breast cancer*, British journal of cancer **45** (3), 361 (1982a).

[Haybittle *et al.* (1982b)] J. L. Haybittle, R. W. Blamey, et al., *A prognostic index in primary breast cancer*, Br. J. Cancer **45** (3), 361 (1982b).

[He *et al.* (2015)] K. He, X. Zhang, et al., *Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification*, (2015).

[Heart *et al.* (2017)] T. Heart, O. Ben-Assuli, & I. Shabtai, *A review of phr, emr and ehr integration: A more personalized healthcare and public health policy*, Health Policy and Technology **6** (1), 20 (2017).

[Heath (1993)] S. S. S. Heath, D; Kasif, *k-dt: A multi-tree learning method*, , 138 (1993).

[Ho (1995)] T. Ho, *Random decision forest*, , 14 (1995).

[Hochreiter (1998)] S. Hochreiter, *The vanishing gradient problem during learning recurrent neural nets and problem solutions*, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **06** (02), 107, https://doi.org/10.1142/S0218488598000094 (1998).

[Howard *et al.* (2013)] J. Howard, E. C. Clark, et al., *Electronic health record impact on work burden in small, unaffiliated, community-based primary care practices*, J. Gen. Intern. Med. **28** (1), 107 (2013).

[Hsu *et al.* (2003)] C.-w. Hsu, C.-c. Chang, & C.-J. Lin, *A practical guide to support vector classification chih-wei hsu, chih-chung chang, and chih-jen lin*, (2003).

[Huang *et al.* (2019)] K. Huang, J. Altosaar, & R. Ranganath, *ClinicalBERT: Modeling clinical notes and predicting hospital readmission*, (2019).

[Hui *et al.* (2020)] Y. Hui, L. Du, et al. (2020), *Extraction and classification of tcm medical records based on bert and bi-lstm with attention mechanism*, in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (IEEE).

[Häyrinen *et al.* (2008)] K. Häyrinen, K. Saranto, & P. Nykänen, *Definition, structure, content, use and impacts of electronic health records: A review of the research literature*, International Journal of Medical Informatics **77** (5), 291 (2008).

[Ignatov *et al.* (2018)] A. Ignatov, H. Eggemann, et al., *Patterns of breast cancer relapse in accordance to biological subtype*, Journal of Cancer Research and Clinical Oncology **144** (7), 1347–1355 (2018).

[Ivakhnenko & Lapa (1967)] A. Ivakhnenko & V. Lapa (1967), *Cybernetics and Forecasting Techniques*, Modern analytic and computational methods in science and mathematics (American Elsevier Publishing Company).

[Jain & Wallace (2019)] S. Jain & B. C. Wallace (2019), *Attention is not Explanation*, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, édité par J. Burstein, C. Doran, & T. Solorio (Association for Computational Linguistics, Minneapolis, Minnesota) pp. 3543–3556.

[Jalali *et al.* (2013)] A. Jalali, E. Buckley, et al., *Prediction of periventricular leukomalacia occurrence in neonates after heart surgery*, IEEE journal of biomedical and health informatics **18**, 10.1109/JBHI.2013.2285011 (2013).

[Jawhari *et al.* (2016)] B. Jawhari, L. Keenan, et al., *Barriers and facilitators to electronic medical record (emr) use in an urban slum*, International Journal of Medical Informatics **94**, 246 (2016).

[Jha *et al.* (2009)] A. K. Jha, D. W. Bates, et al., *Electronic health records: Use, barriers and satisfaction among physicians who care for black and hispanic patients*, J. Eval. Clin. Pract. **15** (1), 158 (2009).

[Jiao *et al.* (2019)] X. Jiao, Y. Yin, et al., *TinyBERT: Distilling BERT for natural language understanding*, (2019).

[Johnson *et al.* (2012)] A. E. W. Johnson, N. Dunkley, et al. (2012), *Patient specific predictions in the intensive care unit using a bayesian ensemble*, in *2012 Computing in Cardiology*, pp. 249–252.

[Johnson *et al.* (2016)] A. E. W. Johnson, T. J. Pollard, et al., *MIMIC-III, a freely accessible critical care database*, Sci. Data **3** (1), 160035 (2016).

[Joo *et al.* (2021)] S. Joo, E. S. Ko, et al., *Multimodal deep learning models for the prediction of pathologic response to neoadjuvant chemotherapy in breast cancer*, Sci. Rep. **11** (1) (2021).

[Kazemi *et al.* (2019)] S. M. Kazemi, R. Goel, et al. (2019), *Time2vec: Learning a vector representation of time*, in *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS)* (NeurIPS).

[Kim *et al.* (2012)] W. Kim, K. S. Kim, et al., *Development of novel breast cancer recurrence prediction model using support vector machine*, J. Breast Cancer **15** (2), 230 (2012).

[Kingma & Ba (2014)] D. P. Kingma & J. Ba, *Adam: A method for stochastic optimization*, (2014).

[Kirkby *et al.* (2023)] M. Kirkby, A. Popatia, et al., *The potential of hormonal therapies for treatment of triple-negative breast cancer*, Cancers **15**, 4702 (2023).

[Klein *et al.* (2022)] S. Klein, A. Gastaldelli, et al., *Why does obesity cause diabetes?*, Cell Metab. **34** (1), 11 (2022).

[Kokalj *et al.* (2021)] E. Kokalj, B. Škrlj, et al. (2021), *BERT meets shapley: Extending SHAP explanations to transformer-based classifiers*, in *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, édité par H. Toivonen & M. Boggia (Association for Computational Linguistics, Online) pp. 16–21.

[Kokhlikyan *et al.* (2020)] N. Kokhlikyan, V. Miglani, et al. (2020), *Captum: A unified and generic model interpretability library for pytorch*, arXiv:2009.07896 [cs.LG].

[Kossman (2006)] S. P. Kossman, *Perceptions of impact of electronic health records on nurses' work*, Stud. Health Technol. Inform. **122**, 337 (2006).

[Kossman & Scheidenhelm (2008)] S. P. Kossman & S. L. Scheidenhelm, *Nurses' perceptions of the impact of electronic health records on work and patient outcomes*, Comput. Inform. Nurs. **26** (2), 69 (2008).

[Kumar *et al.* (2020)] A. Kumar, M. Fulham, et al., *Co-learning feature fusion maps from PET-CT images of lung cancer*, IEEE Trans. Med. Imaging **39** (1), 204 (2020).

[Labrak *et al.* (2023)] Y. Labrak, A. Bazoge, et al. (2023), *DrBERT: A Robust Pre-trained Model in French for Biomedical and Clinical domains*, in *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (ACL'23), Long Paper* (Association for Computational Linguistics, Toronto, Canada) p. 16207–16221.

[Lafourcade *et al.* (2018)] A. Lafourcade, M. His, et al., *Factors associated with breast cancer recurrences or mortality and dynamic prediction of death using history of cancer recurrences: the french E3N cohort*, BMC Cancer **18** (1), 171 (2018).

[Lai *et al.* (2020)] Y. Lai, W. Chen, et al. (2020), *Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning. sci rep 10: 4679.*

[Lakkaraju *et al.* (2020)] H. Lakkaraju, N. Arsov, & O. Bastani (2020), *Robust and stable black box explanations*, in *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 119, édité par H. D. III & A. Singh (PMLR) pp. 5628–5638.

[Lan *et al.* (2019)] Z. Lan, M. Chen, et al., *ALBERT: A lite BERT for self-supervised learning of language representations*, (2019).

[Lanckriet *et al.* (2003)] G. R. G. Lanckriet, M. Deng, et al. (2003), *Kernel-based data fusion and its application to protein function prediction in yeast*, in *Biocomputing 2004* (WORLD SCIENTIFIC).

[Le *et al.* (2019)] H. Le, L. Vial, et al., *FlauBERT: Unsupervised language model pre-training for french*, (2019).

[Le *et al.* (2017)] M. H. Le, J. Chen, et al., *Automated diagnosis of prostate cancer in multi-parametric MRI based on multimodal convolutional neural networks*, Phys. Med. Biol. **62** (16), 6497 (2017).

[Levenshtein (1966)] V. I. Levenshtein, *Binary codes capable of correcting deletions, insertions, and reversals*, Soviet Physics Doklady **10** (8), 707 (1966).

[Li *et al.* (2023a)] C. Li, A. Yates, et al., *Parade: Passage representation aggregation fordocument reranking*, ACM Transactions on Information Systems **42** (2), 1–26 (2023a).

[Li *et al.* (2023b)] Y. Li, M. Mamouei, et al., *Hi-BEHRT: Hierarchical Transformer-Based Model for Accurate Prediction of Clinical Events Using Multimodal Longitudinal Electronic Health Records*, IEEE Journal of Biomedical and Health Informatics **27** (2), 1106, conference Name: IEEE Journal of Biomedical and Health Informatics (2023b).

[Li *et al.* (2020a)] Y. Li, S. Rao, et al., *BEHRT: Transformer for electronic health records*, Sci. Rep. **10** (1), 7155 (2020a).

[Li *et al.* (2020b)] Y. Li, S. Rao, et al., *BEHRT: Transformer for Electronic Health Records*, Scientific Reports **10** (1), 7155, number: 1 Publisher: Nature Publishing Group (2020b).

[Liang *et al.* (2018)] L. Liang, M. O. Wiens, et al., *A locally developed electronic health platform in uganda: Development and implementation of stre@mline*, JMIR Form. Res. **2** (2), e20 (2018).

[Liang *et al.* (2015)] M. Liang, Z. Li, et al., *Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach*, IEEE/ACM Trans. Comput. Biol. Bioinform. **12** (4), 928 (2015).

[Lin *et al.* (2013)] C. Lin, E. W. Karlson, et al., *Automatic prediction of rheumatoid arthritis disease activity from the electronic medical records*, PLoS One **8** (8), e69932 (2013).

[van der Linden *et al.* (2019)] I. van der Linden, H. Haned, & E. Kanoulas, *Global aggregations of local explanations for black box models*, (2019).

[Lipkova *et al.* (2019)] J. Lipkova, P. Angelikopoulos, et al., *Personalized radiotherapy design for glioblastoma: Integrating mathematical tumor models, multimodal scans, and bayesian inference*, IEEE Trans. Med. Imaging **38** (8), 1875 (2019).

[Lison & Tiedemann (2016)] P. Lison & J. Tiedemann (2016), *OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles*, in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, édité par N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (European Language Resources Association (ELRA), Portorož, Slovenia) pp. 923–929.

[Little (2002)] D. B. Little, Roderick J. A.; Rubin (2002), *Statistical Analysis with Missing Data*.

[Liu *et al.* (2018)] J. Liu, Z. Zhang, & N. Razavian (2018), *Deep ehr: Chronic disease prediction using medical notes*, in *Proceedings of the 3rd Machine Learning for Healthcare Conference*, Proceedings of Machine Learning Research, Vol. 85, édité par F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, & J. Wiens (PMLR) pp. 440–464.

[Liu *et al.* (2019a)] Y. Liu, M. Ott, et al., *RoBERTa: A robustly optimized BERT pretraining approach*, (2019a).

[Liu *et al.* (2019b)] Y. Liu, M. Ott, et al., *Roberta: A robustly optimized bert pretraining approach*, arXiv preprint arXiv:1907.11692 (2019b).

[Lo *et al.* (2007)] H. G. Lo, L. P. Newmark, et al., *Electronic health records in specialty care: a time-motion study*, J. Am. Med. Inform. Assoc. **14** (5), 609 (2007).

[Louppe (2014)] G. Louppe (October 2014), *Understanding Random Forests: From Theory to Practice*, Thèse de Doctorat (ULiège - Université de Liège).

[Lundberg & Lee (2017)] S. M. Lundberg & S.-I. Lee (2017), in *Advances in Neural Information Processing Systems 30*, édité par I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Curran Associates, Inc.) pp. 4765–4774.

[López-Andreu *et al.* (2023)] F. J. López-Andreu, J. A. López-Morales, et al., *Deep learning-based time series forecasting models evaluation for the forecast of chlorophyll a and dissolved oxygen in the mar menor*, Journal of Marine Science and Engineering **11** (7), 1473 (2023).

[Ma *et al.* (2021)] L.-Z. Ma, F.-R. Sun, et al., *Metabolically healthy obesity and risk of stroke: a meta-analysis of prospective cohort studies*, Ann. Transl. Med. **9** (3), 197 (2021).

[Maas (2013)] A. L. Maas (2013), *Rectifier nonlinearities improve neural network acoustic models.*

[Macaš *et al.* (2012)] M. Macaš, J. Kuzilek, et al. (2012), *Linear bayes classification for mortality prediction*, in *2012 Computing in Cardiology*, pp. 473–476.

[Martin *et al.* (2019)] L. Martin, B. Muller, et al., *CamemBERT: A tasty french language model*, (2019).

[Mazo *et al.* (2022)] C. Mazo, C. Aura, et al., *Application of artificial intelligence techniques to predict risk of recurrence of breast cancer: A systematic review*, J. Pers. Med. **12** (9), 1496 (2022).

[McAlearney *et al.* (2010)] A. S. McAlearney, J. Robbins, et al., *Perceived efficiency impacts following electronic health record implementation: an exploratory study of an urban community health center network*, International journal of medical informatics **79** (12), 807 (2010).

[McAlearney *et al.* (2012)] A. S. McAlearney, J. Robbins, et al., *The role of cognitive and learning theories in supporting successful ehr system implementation training: A qualitative study*, Medical Care Research and Review **69** (3), 294, pMID: 22451617, https://doi.org/10.1177/1077558711436348 (2012).

[Mikolov *et al.* (2013)] T. Mikolov, I. Sutskever, et al. (2013), *Distributed representations of words and phrases and their compositionality*, in *Advances in Neural Information Processing Systems*, Vol. 26, édité par C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Weinberger (Curran Associates, Inc.).

[Nguyen *et al.* (2014)] L. Nguyen, E. Bellucci, & L. T. Nguyen, *Electronic health records implementation: An evaluation of information system*

impact and contingency factors, Int. J. Med. Inform. **83** (11), 779 (2014).

[Nie *et al.* (2019)] D. Nie, J. Lu, et al., *Multi-channel 3D deep feature learning for survival time prediction of brain tumor patients using multimodal neuroimages*, Sci. Rep. **9** (1) (2019).

[Noblin *et al.* (2013a)] A. Noblin, K. Cortelyou-Ward, et al., *EHR implementation in a new clinic: A case study of clinician perceptions*, J. Med. Syst. **37** (4) (2013a).

[Noblin *et al.* (2013b)] A. Noblin, K. Cortelyou-Ward, et al., *EHR implementation in a new clinic: A case study of clinician perceptions*, J. Med. Syst. **37** (4) (2013b).

[Osborne *et al.* (2009)] M. Osborne, R. Garnett, & S. Roberts, *Gaussian processes for global optimization*, (2009).

[Oumer *et al.* (2021)] A. Oumer, A. Muhye, et al., *Utilization, determinants, and prospects of electronic medical records in ethiopia*, Biomed Res. Int. **2021**, 1 (2021).

[Oza *et al.* (2017)] S. Oza, D. Jazayeri, et al., *Development and deployment of the openmrs-ebola electronic health record system for an ebola treatment center in sierra leone*, J Med Internet Res **19** (8), e294 (2017).

[Ozair *et al.* (2020)] S. Ozair, G. W. Taylor, et al., *Modeling periodic behavior in time series with time2vec*, arXiv preprint arXiv:2006.03479 (2020).

[Pandey *et al.* (2022)] L. N. Pandey, R. Vashisht, & H. G. Ramaswamy, *On the interpretability of attention networks*, (2022).

[Pang *et al.* (2021a)] C. Pang, X. Jiang, et al., *CEHR-BERT: Incorporating temporal information from structured EHR data to improve prediction tasks*, (2021a).

[Pang *et al.* (2021b)] C. Pang, X. Jiang, et al. (2021b), *Cehr-bert: Incorporating temporal information from structured ehr data to improve prediction tasks*, in *Proceedings of Machine Learning for Health*, Proceedings of Machine Learning Research, Vol. 158, édité par S. Roy, S. Pfohl, E. Rocheteau, G. A. Tadesse, L. Oala, F. Falck, Y. Zhou, L. Shen, G. Zamzmi, P. Mugambi, A. Zirikly, M. B. A. McDermott, & E. Alsentzer (PMLR) pp. 239–260.

[Pappagari *et al.* (2019)] R. Pappagari, P. Zelasko, et al. (2019), *Hierarchical transformers for long document classification*, in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 838–844.

[Park *et al.* (2022)] H. Park, Y. Vyas, & K. Shah (2022), *Efficient classification of long documents using transformers*, in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, édité par S. Muresan, P. Nakov, & A. Villavicencio (Association for Computational Linguistics) pp. 702–709.

[Pati *et al.* (2023)] S. Pati, W. Irfan, et al., *Obesity and cancer: A current overview of epidemiology, pathogenesis, outcomes, and management*, Cancers (Basel) **15** (2), 485 (2023).

[Peeken *et al.* (2019)] J. C. Peeken, T. Goldberg, et al., *Combining multimodal imaging and treatment features improves machine learning-based prognostic assessment in patients with glioblastoma multiforme*, Cancer Med. **8** (1), 128 (2019).

[Perou *et al.* (2000a)] C. M. Perou, T. Sørlie, et al., *Molecular portraits of human breast tumours*, Nature **406** (6797), 747 (2000a).

[Perou *et al.* (2000b)] C. M. Perou, T. Sørlie, et al., *Molecular portraits of human breast tumours*, Nature **406** (6797), 747 (2000b).

[Phung *et al.* (2019)] M. T. Phung, S. Tin Tin, & J. M. Elwood, *Prognostic models for breast cancer: a systematic review*, BMC cancer **19** (1), 1 (2019).

[PhysioNet (2012)] C. . PhysioNet (2012), *Physionet*, https://physionet.org/content/challenge-2012/1.0.0/, accessed: 201-04-15.

[Poissant (2005)] L. Poissant, *The impact of electronic health records on time efficiency of physicians and nurses: A systematic review*, J. Am. Med. Inform. Assoc. **12** (5), 505 (2005).

[Poletto (2023)] B. Poletto (2023), *Glossaire info cancer*, https://www.arcagy.org/infocancer/cms/glossaire, (Accessed: 2023-01-30).

[Powell-Wiley *et al.* (2021)] T. M. Powell-Wiley, P. Poirier, et al., *Obesity and cardiovascular disease: A scientific statement from the american heart association*, Circulation **143** (21) (2021).

[Priestman *et al.* (2018)] W. Priestman, S. Sridharan, et al., *What to expect from electronic patient record system implementation: lessons learned from published evidence*, BMJ Health Care Inform. **25** (2), 92 (2018).

[Qian *et al.* (2021)] X. Qian, J. Pei, et al., *Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning*, Nat. Biomed. Eng. **5** (6), 522 (2021).

[Rabinovici-Cohen *et al.* (2020)] S. Rabinovici-Cohen, A. Abutbul, et al. (2020), *Multimodal prediction of breast cancer relapse prior to neoadjuvant chemotherapy treatment*, in *Predictive Intelligence in Medicine*, édité par I. Rekik, E. Adeli, S. H. Park, & M. d. C. Valdés Hernández (Springer International Publishing, Cham) pp. 188–199.

[Rabinovici-Cohen *et al.* (2022a)] S. Rabinovici-Cohen, X. M. Fernández, et al., *Multimodal prediction of five-year breast cancer recurrence in women who receive neoadjuvant chemotherapy*, Cancers (Basel) **14** (16), 3848 (2022a).

[Rabinovici-Cohen *et al.* (2022b)] S. Rabinovici-Cohen, X. M. Fernández, et al., *Multimodal Prediction of Five-Year Breast Cancer Recurrence in Women Who Receive Neoadjuvant Chemotherapy*, Cancers **14** (16), 3848 (2022b).

[Raciti *et al.* (2020)] P. Raciti, J. Sue, et al., *Novel artificial intelligence system increases the detection of prostate cancer in whole slide images of core needle biopsies*, Mod. Pathol. **33** (10), 2058 (2020).

[Radford *et al.* (2018)] A. Radford, K. Narasimhan, et al., *Improving language understanding by generative pre-training*, OpenAI (2018).

[Radford *et al.* (2019)] A. Radford, J. Wu, et al., *Language models are unsupervised multitask learners*, OpenAI (2019).

[Ramanathan *et al.* (2022)] T. T. Ramanathan, J. Hossen, et al., *Naïve bayes based multiple parallel fuzzy reasoning method for medical diagnosis*, Journal of Engineering Science and Technology **17** (1), 0472 (2022).

[Rao *et al.* (2022)] S. Rao, M. Mamouei, et al., *Targeted-BEHRT: Deep Learning for Observational Causal Inference on Longitudinal Electronic Health Records*, IEEE Transactions on Neural Networks and Learning Systems , 1Conference Name: IEEE Transactions on Neural Networks and Learning Systems (2022).

[Rasmy *et al.* (2021)] L. Rasmy, Y. Xiang, et al., *Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction*, NPJ Digit. Med. **4** (1), 86 (2021).

[Reda *et al.* (2018)] I. Reda, A. Khalil, et al., *Deep learning role in early diagnosis of prostate cancer*, Technol. Cancer Res. Treat. **17**, 153303461877553 (2018).

[Rezaei-Dastjerdehei *et al.* (2020)] M. R. Rezaei-Dastjerdehei, A. Mijani, & E. Fatemizadeh (2020), *Addressing imbalance in multi-label classification using weighted cross entropy loss function*, in *2020 27th National*

and 5th International Iranian Conference on Biomedical Engineering (ICBME), pp. 333–338.

[Ribeiro *et al.* (2016)] M. T. Ribeiro, S. Singh, & C. Guestrin, *"why should I trust you?": Explaining the predictions of any classifier*, (2016).

[Rosenblatt (1958)] F. Rosenblatt, *The perceptron: a probabilistic model for information storage and organization in the brain*, Psychol. Rev. **65** (6), 386 (1958).

[Rupp *et al.* (2023)] M. Rupp, O. Peter, & T. Pattipaka (2023), in *Trustworthy Machine Learning for Healthcare*, Lecture notes in computer science (Springer Nature Switzerland, Cham) pp. 73–84.

[Saeed *et al.* (2011)] M. Saeed, M. Villarroel, et al., *Multiparameter intelligent monitoring in intensive care II: a public-access intensive care unit database*, Crit. Care Med. **39** (5), 952 (2011).

[Sahney & Sharma (2018)] R. Sahney & M. Sharma, *Electronic health records: A general overview*, Current Medicine Research and Practice **8** (2), 67 (2018).

[Sanh *et al.* (2019)] V. Sanh, L. Debut, et al., *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*, (2019).

[Saracci (1995)] R. Saracci, *Smoking and death*, BMJ **310** (6979), 600 (1995).

[Sedghi *et al.* (2020)] A. Sedghi, A. Mehrtash, et al., *Improving detection of prostate cancer foci via information fusion of MRI and temporal enhanced ultrasound*, Int. J. Comput. Assist. Radiol. Surg. **15** (7), 1215 (2020).

[Sejdic (2018)] T. H. Sejdic, Ervin; Falk (2018), *Signal processing and machine learning for biomedical big data.*

[Serrano & Smith (2019)] S. Serrano & N. A. Smith, *Is attention interpretable?*, (2019).

[Shrikumar *et al.* (2017)] A. Shrikumar, P. Greenside, & A. Kundaje (2017), *Learning important features through propagating activation differences*, in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17 (JMLR.org) p. 3145–3153.

[Silva *et al.* (2012)] I. Silva, G. Moody, et al., *Predicting in-hospital mortality of ICU patients: The PhysioNet/computing in cardiology challenge 2012*, Comput. Cardiol. (2010) **39**, 245 (2012).

[Skinner *et al.* (2011a)] A. Skinner, J. Windle, & L. Grabenbauer, *Electronic health record adoption – maybe it's not about the money*, Appl. Clin. Inform. **02** (04), 460 (2011a).

[Skinner *et al.* (2011b)] A. Skinner, J. Windle, & L. Grabenbauer, *Electronic health record adoption – maybe it's not about the money*, Appl. Clin. Inform. **02** (04), 460 (2011b).

[Sockolow *et al.* (2012)] P. S. Sockolow, K. H. Bowles, et al., *Community-based, interdisciplinary geriatric care team satisfaction with an electronic health record*, Comput. Inform. Nurs. **30** (6), 300 (2012).

[Song *et al.* (2017)] J.-L. Song, C. Chen, et al., *The association between prognosis of breast cancer and first-degree family history of breast or ovarian cancer: a systematic review and meta-analysis*, Fam. Cancer **16** (3), 339 (2017).

[Sørlie *et al.* (2001)] T. Sørlie, C. M. Perou, et al., *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications*, Proc. Natl. Acad. Sci. U. S. A. **98** (19), 10869 (2001).

[Springenberg *et al.* (2014)] J. T. Springenberg, A. Dosovitskiy, et al., *Striving for simplicity: The all convolutional net*, (2014).

[Stuart (2015)] R. Stuart (2015), *Artificial Intelligence: A Modern Approach* (Pearson, California).

[Sun *et al.* (2019)] D. Sun, M. Wang, & A. Li, *A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data*, IEEE/ACM Transactions on Computational Biology and Bioinformatics **16** (3), 841 (2019).

[Sundararajan *et al.* (2017)] M. Sundararajan, A. Taly, & Q. Yan (2017), *Axiomatic attribution for deep networks*, in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17 (JMLR.org) p. 3319–3328.

[Sutskever *et al.* (2014)] I. Sutskever, O. Vinyals, & Q. V. Le, *Sequence to sequence learning with neural networks*, (2014).

[Taylor (1953)] W. L. Taylor, *Cloze procedure: A new tool for measuring readability*, (1953).

[Thomsen (1 30)] C. Thomsen (Accessed: 2023-01-30), *Dictionnaire Médical*, https://www.dictionnaire-medical.fr/.

[Tibshirani (1996)] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. R. Stat. Soc. Series B Stat. Methodol. **58** (1), 267 (1996).

[Tsai *et al.* (2020)] C. H. Tsai, A. Eghdam, et al., *Effects of electronic health record implementation and barriers to adoption and use: A scoping review and qualitative analysis of the content*, Life (Basel) **10** (12), 327 (2020).

[Vairavan *et al.* (2012)] S. Vairavan, L. Eshelman, et al. (2012), *Prediction of mortality in an intensive care unit using logistic regression and a hidden markov model*, in *2012 Computing in Cardiology*, pp. 393–396.

[Vapnik (1997)] V. N. Vapnik (1997), in *Lecture Notes in Computer Science*, Lecture notes in computer science (Springer Berlin Heidelberg, Berlin, Heidelberg) pp. 261–271.

[Vaswani *et al.* (2017a)] A. Vaswani, N. Shazeer, et al., *Attention is all you need*, (2017a).

[Vaswani *et al.* (2017b)] A. Vaswani, N. Shazeer, et al. (2017b), in *Advances in Neural Information Processing Systems 30*, édité par I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Curran Associates, Inc.) pp. 5998–6008.

[Verhulst (1845)] H. Verhulst, P. (1845), *Recherches mathématiques sur la loi d'accroissement de la population*.

[Vig & Belinkov (2019)] J. Vig & Y. Belinkov (2019), *Analyzing the structure of attention in a transformer language model*, in *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, édité par T. Linzen, G. Chrupała, Y. Belinkov, & D. Hupkes (Association for Computational Linguistics, Florence, Italy) pp. 63–76.

[Wang *et al.* (2020)] H. Wang, Y. Li, et al., *Prediction of breast cancer distant recurrence using natural language processing and knowledge-guided convolutional neural network*, Artificial Intelligence in Medicine **110**, 101977 (2020).

[WHO (2022a)] WHO (2022a), *Breast cancer*, `https://www.who.int/news-room/fact-sheets/detail/breast-cancer`, accessed: 2024-04-15.

[WHO (2022b)] WHO (2022b), *Cancer*, `https://www.who.int/health-topics/cancer`, accessed: 2024-04-15.

[Wiegreffe & Pinter (2019)] S. Wiegreffe & Y. Pinter (2019), *Attention is not not explanation*, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*,

édité par K. Inui, J. Jiang, V. Ng, & X. Wan (Association for Computational Linguistics, Hong Kong, China) pp. 11–20.

[Wishart *et al.* (2010)] G. C. Wishart, E. M. Azzato, et al., *PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer*, Breast Cancer Res. **12** (1) (2010).

[Witteveen *et al.* (2015)] A. Witteveen, I. M. H. Vliegen, et al., *Personalisation of breast cancer follow-up: a time-dependent prognostic nomogram for the estimation of annual risk of locoregional recurrence in early breast cancer patients*, Breast Cancer Research and Treatment **152** (3), 627 (2015).

[Woldemariam & Jimma (2023)] M. T. Woldemariam & W. Jimma, *Adoption of electronic health record systems to enhance the quality of healthcare in low-income countries: a systematic review*, BMJ Health Care Inform. **30** (1), e100704 (2023).

[Wolf *et al.* (2019)] T. Wolf, L. Debut, et al. (2019), *Hugging face's transformers: State-of-the-art natural language processing*, `https://github.com/huggingface/transformers`, accessed: 2024-06-19.

[Wu *et al.* (2017)] X. Wu, Y. Ye, et al., *Personalized prognostic prediction models for breast cancer recurrence and survival incorporating multidimensional data*, J. Natl. Cancer Inst. **109** (7) (2017).

[Xia *et al.* (2012)] H. Xia, B. J. Daley, et al. (2012), *A neural network model for mortality prediction in icu*, in *2012 Computing in Cardiology*, pp. 261–264.

[Xiong *et al.* (2021)] Y. Xiong, Z. Zeng, et al. (2021), *Nyströmformer: A nyström-based algorithm for approximating self-attention*, in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, pp. 14138–14148.

[Yang *et al.* (2021)] P.-T. Yang, W.-S. Wu, et al., *Breast cancer recurrence prediction with ensemble methods and cost-sensitive learning*, Open Medicine **16** (1), 754, publisher: De Gruyter Open Access (2021).

[Yao *et al.* (2022a)] Y. Yao, Y. Lv, et al., *ICSDA: a multi-modal deep learning model to predict breast cancer recurrence and metastasis risk by integrating pathological, clinical and gene expression data*, Brief. Bioinform. **23** (6) (2022a).

[Yao *et al.* (2022b)] Y. Yao, Y. Lv, et al., *Icsda: a multi-modal deep learning model to predict breast cancer recurrence and metastasis risk by integrating pathological, clinical and gene expression data*, Briefings in Bioinformatics **23** (6), 10.1093/bib/bbac448 (2022b).

[Yau *et al.* (2011)] G. L. Yau, A. S. Williams, & J. B. Brown, *Family physicians' perspectives on personal health records: qualitative study*, Can. Fam. Physician **57** (5), e178 (2011).

[You *et al.* (2018)] S. H. You, B. J. Chae, et al., *Clinical differences in triple-positive operable breast cancer subtypes in korean patients: An analysis of korean breast cancer registry data*, J. Breast Cancer **21** (4), 415 (2018).

[Youlden *et al.* (2014)] D. R. Youlden, S. M. Cramb, et al., *Incidence and mortality of female breast cancer in the Asia-Pacific region*, Cancer Biol. Med. **11** (2), 101 (2014).

[Zaheer *et al.* (2020a)] M. Zaheer, G. Guruganesh, et al., *Big bird: Transformers for longer sequences*, (2020a).

[Zaheer *et al.* (2020b)] M. Zaheer, G. Guruganesh, et al. (2020b), *Big bird: Transformers for longer sequences*, in *Advances in neural information processing systems*, Vol. 33, pp. 17283–17297.

[Zeiler *et al.* (2011)] M. D. Zeiler, G. W. Taylor, & R. Fergus (2011), *Adaptive deconvolutional networks for mid and high level feature learning*, in *2011 International Conference on Computer Vision*, pp. 2018–2025.

[Zeng *et al.* (2019a)] Z. Zeng, L. Yao, et al., *Identifying breast cancer distant recurrences from electronic health records using machine learning*, J. Healthc. Inform. Res. **3** (3), 283 (2019a).

[Zeng *et al.* (2019b)] Z. Zeng, L. Yao, et al., *Identifying breast cancer distant recurrences from electronic health records using machine learning*, Journal of Healthcare Informatics Research **3** (3), 283–299 (2019b).

[Zhang *et al.* (2020)] D. Zhang, C. Yin, et al., *Combining structured and unstructured data for predictive models: a deep learning approach*, BMC Med. Inform. Decis. Mak. **20** (1), 280 (2020).

[Zhang *et al.* (2018)] H. Zhang, I. Goodfellow, et al., *Self-attention generative adversarial networks*, (2018).

[Zhang *et al.* (2016)] N. Zhang, W. Lou, et al., *Low molecular weight heparin and cancer survival: clinical trials and experimental mechanisms*, J. Cancer Res. Clin. Oncol. **142** (8), 1807 (2016).

[Zhang *et al.* (2012)] Y. Zhang, P. Yu, & J. Shen, *The benefits of introducing electronic health records in residential aged care facilities: A multiple case study*, Int. J. Med. Inform. **81** (10), 690 (2012).

[Zhou *et al.* (2019)] L. Zhou, J. Zhang, & C. Zong (2019), *Synchronous bidirectional neural machine translation*.

[Zhu *et al.* (2015)] Y. Zhu, R. Kiros, et al. (2015), *Aligning books and movies: Towards story-like visual explanations by watching movies and reading books*, in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 19–27.

[Zitnik *et al.* (2019a)] M. Zitnik, F. Nguyen, et al., *Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities*, Inf. Fusion **50**, 71 (2019a).

[Zitnik *et al.* (2019b)] M. Zitnik, F. Nguyen, et al., *Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities*, Inf. Fusion **50**, 71 (2019b).

RÉSUMÉ

Le cancer du sein est l'un des cancers les plus fréquents dans le monde, représentant 12,5 % des nouveaux cas annuels. En 2022, environ 2,3 millions de femmes ont été diagnostiquées, avec plus de 666 000 décès. Bien que les dossiers médicaux électroniques (DME) aient révolutionné la recherche clinique en fournissant des données précieuses, les études sur le cancer du sein exploitent rarement les rapports médicaux en texte libre, qui contiennent pourtant des informations cruciales. Cette thèse propose de développer des modèles d'apprentissage automatique et profond pour prédire les statuts de survie du cancer du sein en utilisant des données multimodales (rapports textuels en français, résultats de laboratoire et descripteurs cliniques) d'une vaste cohorte de l'Institut Curie. Des modèles ont été construits pour analyser séparément puis conjointement ces modalités. Les résultats montrent que l'intégration des données textuelles et structurées améliore la prédiction des statuts de survie des patientes. De plus, l'analyse des facteurs predictifs des statuts de survie des patients ouvre de nouvelles perspectives pour une meilleure compréhension des méchanismes du cancer du sein et par conséquent, l'amélioration des soins.

MOTS CLÉS

Cancer du sein, Apprentissage automatique, Données multimodales, Dossiers médicaux électroniques

ABSTRACT

Breast cancer is one of the most common cancers worldwide, accounting for 12.5% of new cases each year. In 2022, around 2.3 million women were diagnosed, with over 666,000 deaths. Although electronic health records (EHRs) have revolutionized clinical research by providing valuable data, breast cancer studies rarely exploit free-text medical reports, which nonetheless contain crucial information. This thesis proposes to develop machine and deep learning models to predict breast cancer outcomes using multimodal data (French text reports, laboratory results, clinical descriptors) from a large Institut Curie cohort. Models were built to analyze these modalities separately and then jointly. Results show that the integration of textual and structured data improves the prediction of patients' survival status. Moreover, the analysis of predictive factors for patients' survival status opens up new perspectives for a better understanding of underlying mechanisms in breast cancer, and thus, for improving care.

KEYWORDS

Breast Cancer, Machine learning, Multimodal data, Electronic Health Records