Machine learning based novel biomarkers discovery for therapeutic use in ”pan-gyn” cancers

Elena Spirina Menand

DOCTORAT BRETAGNE LOIRE / MATHSTIC

université angers

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ D'ANGERS

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Informatique*

Par
Elena SPIRINA MENAND

« Machine learning based novel biomarkers discovery for therapeutic use in "pan-gyn" cancers »

Thèse présentée et soutenue à Angers, le 19 décembre 2022
Unité de recherche : LARIS (EA 7315)
Thèse N° : 212930

Rapporteurs avant soutenance :

Mathieu HATT, Directeur de Recherche, LaTIM, Université de Bretagne Occidentale
Marie DE TAYRAC, Professeur des Universités, CHU de Rennes, Université de Rennes 1

Composition du Jury :

Président :	Mario CAMPONE, Professeur des Universités, Directeur Général, Institut de Cancérologie de l'Ouest, Université d'Angers
Rapporteurs :	Mathieu HATT, Directeur de Recherche, LaTIM, Université de Bretagne Occidentale
	Marie DE TAYRAC, Professeur des Universités, CHU de Rennes, Université de Rennes 1
Examinateurs :	Fahed ABDALLAH, Professeur des Universités, LM2S, Université de technologie de Troyes, Université libanaise
	Jean-Marie MARION, Maître de Conférences, LARIS, Université d'Angers
Dir. de thèse :	Pierre CHAUVET, Professeur des Universités, LARIS, Université d'Angers
Co-dir. de thèse :	Alain MOREL, Professeur des Universités, Institut de Cancérologie de l'Ouest, Université d'Angers
Co-enc. de thèse :	Nisrine JRAD, Maître de Conférences, LARIS, Université d'Angers

Remerciements

"Chaos is merely order waiting to be deciphered.." ***José Saramago***

Tout d'abord je tiens à remercier tous les gens qui m'ont permis de réaliser cette thèse de doctorat et d'avancer ensemble, de me former par la recherche et pour la recherche et de tenter à mon niveau de "déchiffrer le chaos".

Je remercie mes encadrants, Pierre Chauvet, Alain Morel et Nisrine Jrad, pour leurs conseils justes et la relecture des articles. Merci particulièrement à Alain Morel pour m'avoir repérée et m'avoir tendue la main quand j'en avais besoin, merci à Pierre Chauvet pour m'avoir fait confiance et mise sur les rails à mes débuts. Merci à Nisrine Jrad pour avoir partagé sa méthodologie et sa rigueur.

Je remercie également Jean-Marie Marion pour m'avoir consulté sur les modèles de survie et m'avoir communiqué son calme et sa patience.

Merci aux membres de mon comité de suivi extérieur: Mario Campone et Fahed Abdallah pour les conseils précieux et les orientations scientifiques.

Merci aux rapporteurs de ma thèse Marie De Tayrac et Mathieu Hatt pour avoir accepté d'évaluer mon travail.

Merci à Manon De Vries et Leslie Tessier pour leur écoute et pour les cours express en oncologie et en anatomo-pathologie.

Merci à Christophe Passot et Jonathan Dauvé pour nos discussions captivantes et les explications d'expert en biologie moléculaire et bioinformatique.

Un grand merci à Lise Boussemart pour m'avoir donnée l'envie de travailler avec les données omiques et m'avoir apprise les bases.

Enfin un immense merci à ma famille et ma belle-famille pour leur soutien, mon mari pour son encouragement et son sens de l'humour, ainsi qu'à mes enfants qui me permettent de relativiser aux moments difficiles.

Table of Contents

I State of the art **9**

Introduction part I **10**

Context . . . 10

Motivation . . . 10

1 Transcriptomic data **13**

1.1 Transcriptome . . . 13

1.2 Sequencing technologies . . . 16

1.2.1 Microarrays . . . 16

1.2.2 Next generation sequencing and RNA-seq . . . 16

1.2.3 Third generation sequencing . . . 18

1.3 Bioinformatics analysis steps . . . 20

2 Survival analysis **22**

2.1 Statistical methods . . . 24

2.1.1 Non-parametric methods . . . 24

2.1.2 Parametric methods . . . 25

2.1.3 Semi-parametric method: Cox model . . . 26

2.2 Machine learning methods . . . 27

2.2.1 Machine learning paradigm . . . 27

2.2.2 Survival trees and random survival forests . . . 32

2.2.3 Artificial neural networks and survival analysis . . . 32

2.3 Evaluation criteria . . . 37

2.3.1 Concordance index . . . 37

2.3.2 Brier score . . . 40

2.3.3 Kaplan-Meier curves . . . 40

3 Survival analysis and gene expression **42**

3.1 Dimensionality reduction . . . 42

3.1.1 Feature selection . . . 42

3.1.2 Feature extraction . . . 46
3.1.3 Prior knowledge integration . . . 51
3.2 Survival analysis strategies . . . 52
3.2.1 Validation aspects . . . 59

Conclusion Part I **62**

II Contributions **65**

Introduction part II **66**

1 Comparative study **74**
1.1 Introduction . . . 74
1.2 Overview . . . 75
1.2.1 Survival analysis . . . 75
1.2.2 Gene expression and survival analysis . . . 79
1.3 Materials and Methods . . . 80
1.3.1 Gene expression and clinical data . . . 80
1.3.2 Evaluation criteria . . . 81
1.3.3 Algorithms and implementation . . . 82
1.3.4 Results . . . 83
1.4 Conclusion . . . 86

2 Transfer learning experiments **87**
2.1 Introduction . . . 87
2.2 Survival analysis and deep learning . . . 89
2.2.1 Cox proportional hazards and neural networks . . . 89
2.2.2 Regularization . . . 90
2.2.3 More data and transfer learning . . . 91
2.2.4 Automated hyperparameter optimization . . . 91
2.3 Materials and methods . . . 92
2.3.1 Gene expression and clinical data . . . 92
2.3.2 Performance metric . . . 92
2.3.3 Data pre-processing . . . 93
2.3.4 Bayesian optimization . . . 93

2.4 Results . . . 94
2.5 Conclusion . . . 95

3 Proposed method N-MTLR-Rank **96**
3.1 Introduction . . . 96
3.2 Results . . . 97
3.2.1 Training and comparing deep survival networks . . . 97
3.2.2 Validating with the external dataset . . . 98
3.2.3 Interpreting N-MTLR-Rank with PatternAttribution . . . 98
3.3 Discussion . . . 104
3.4 Methods . . . 106
3.4.1 Data . . . 106
3.4.2 Proposed deep survival model . . . 107
3.4.3 Model training and validation . . . 109
3.4.4 Model selection and interpretation . . . 110

Conclusion Part II **116**

Bibliography **119**

List of Figures **139**

List of Tables **141**

Part I

State of the art

Introduction part I

Contexte

Ovarian cancer is the most frequent pathology among gynecological cancers and the 5th cause of death in women worldwide [Cai+21]. In France, it ranks 8th in terms of frequency and 4th in terms of cancer mortality in women. For the year 2018, the estimated number of new cases of ovarian cancer in France was 5,193 and the estimated number of ovarian cancer deaths was 3,479 [Tré+20]. For comparison, its incidence was approximately 4,600 new cases in France in 2015 with 3,100 annual deaths. The prognosis for this disease remains poor, with net survival estimated at 43%, all stages combined, and the majority of deaths occur within the first two years after diagnosis.

More than 90% of adult ovarian cancers are epithelial cancers (adenocarcinomas) of which the 5 main subtypes are: high grade serous, endometrioid, clear cell, mucinous and low grade serous. Three quarters of patients are diagnosed at the advanced stage (stages IIIC and IV of International Federation of Gynecology and Obstetrics (FIGO)), i.e., the disease has spread beyond the ovaries to the entire surface of the peritoneum or at a distance.

Motivation

Despite advances in understanding the biology of ovarian cancer, patients with stage IIIC-IV disease have an overall 5-year survival of less than 20% [Pok+19]. Since the 1990s, the standard management of ovarian cancer has included cytoreduction followed by systemic treatment with paclitaxel and carboplatin. In fact, 80% of patients have a good initial response to the proposed treatments (surgery and chemotherapy), however 70% of them will present a recurrence within two years. When the recurrence occurs less than one year after the end of chemotherapy, the disease is considered "platinum-resistant", it is therefore important to understand the mechanisms of chemotherapy resistance and tumor recurrence in order to improve patient survival.

On the other hand, there is a proportion of patients who develop chronic disease with a

survival of 5 years and more. This observation allows us to imagine that there are at least 2 biological forms of the disease: a rather aggressive one and another one which would evolve in a slower way. Therefore, it is urgent to develop classifiers to separate patients into therapeutic groups and to detect new therapeutic targets. It has been shown that it is possible to identify the difference in expression from transcriptomic data and to deduce signatures in various biological processes including ovarian cancer [FSA19].

Some recent work has made it possible to stratify patients into good and poor prognosis groups according to their transcriptomic profile [Bon+08; Tot+08; Cri+09]. The availability of omics data from The Cancer Genome Atlas (TCGA) project has marked a new step in the field of cancer research. This initiative has resulted in several papers that have led to a better characterization of ovarian cancer [Bel+11; Ver+12; Way+16].

Another technological advance that has proven to be successful in different fields, deep learning, is being applied in the medical field in general, and in the use of omics data. For example, the recent work [Chi+18] has demonstrated the interest in using artificial neural networks to predict survival from the transcriptome, however not all cancer types tested among the TCGA data yield satisfactory results, in particular ovarian cancer is among the types that need further investigation [CZG18].

Chapter 1

Transcriptomic data

In this chapter we will give the few key notions important for the reading of the following chapters. Section 1.1 briefly introduces the concept of the transcriptome, section 1.2 discusses sequencing technologies that allow the quantification of transcriptome data. Finally, section 1.3 describes the bioinformatic pre-processing steps that precede gene expression analysis.

1.1 Transcriptome

The cell, the smallest independently reproducible living unit, is composed of lipids, carbohydrates, proteins and nucleic acids. Among the nucleic acids, we distinguish deoxyribonucleic acids (DNA) and ribonucleic acids (RNA). The DNA constitutes the genome of the cell in which are encoded all the information necessary for the functioning of the cell. This molecule is made up of two strands or polynucleotides, composed of 4 deoxynucleotides linked by phosphodiester bonds. These deoxynucleotides are composed of a phosphate, a deoxyribose and a heterocyclic base that differentiates them. In DNA, the bases present are Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). The two strands of DNA are paired in an inversely complementary manner by hydrogen bonds. The pairing is carried out by two hydrogen bonds between A and T and three hydrogen bonds between G and C. Thus, between the two strands the pairs A and T are complementary as the pairs G and C and the two strands encode the same information (cf Fig.1.1).

In the genome, information is structured in the form of genes that code for proteins, each gene representing several thousand deoxynucleotides. The production of a protein coded by a gene requires an intermediate molecule or RNA which will then be translated into protein. Thus, the genes are copied in a complementary way in RNA during a process called transcription. In the same way as DNA, RNA contains the genetic information, but encoded in a single strand using the alphabet of four ribonucleotides composed of a phosphate, a ribose and four bases, where the Thymine (T) is replaced by the Uracile (U)

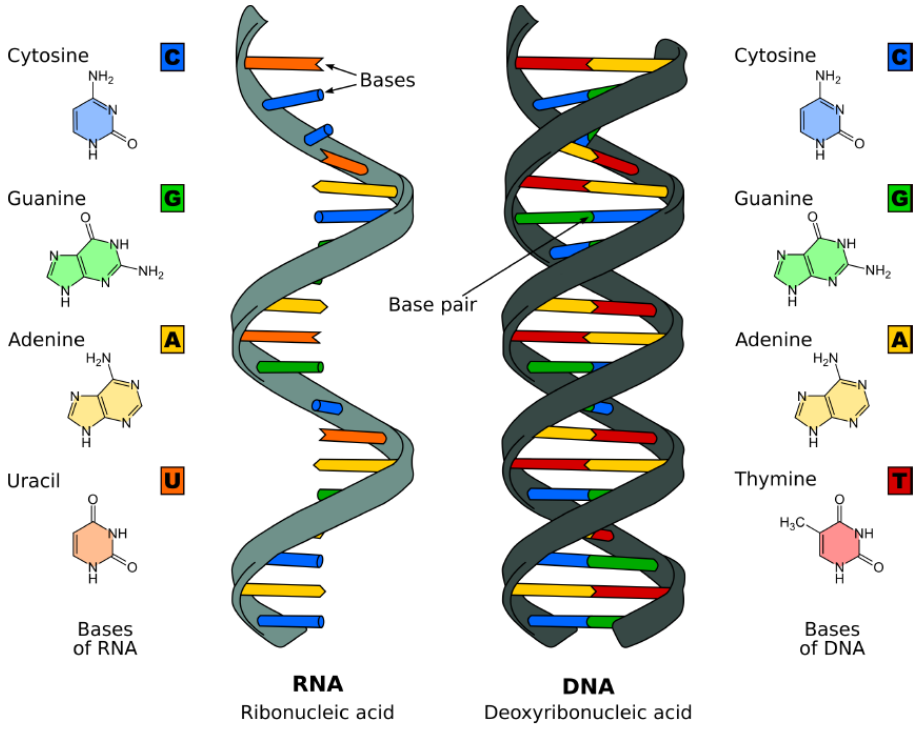


Figure 1.1: DNA and RNA alphabet bases, [Mar18].

which is paired by two hydrogen bonds with A (cf Fig.1.1). The messenger RNAs obtained in the cell by transcription are inversely complementary to the DNA strand coding for a gene. The set of expressed RNAs is called transcriptome and by extension the set of of proteins translated by the transcriptome constitutes the proteome. There are about 20,000 genes in the human genome and not all genes are transcribed at the same frequency in different cells or in a cell at a given time. Thus, each cell or homogeneous group of cells can be characterized by a transcriptome which corresponds to the identification and quantification of each expressed RNA.

The transcripts generated during transcription are called primary transcripts (pre-mRNA), they undergo a certain number of transformations to become mature transcripts. In fact, in eukaryotes, there is no genetic colinearity. In other words, the sequence information of messenger RNAs is fragmented in the genome into two sequence elements: introns and exons that are transcribed into pre-mRNAs. The most important maturation event is splicing, which keeps the exons and eliminates the introns to form the mature mRNA. From identical primary transcripts it is possible to obtain different mature transcripts

(isoforms) with different combinations of exons, this phenomenon is called alternative splicing and multiplies the number of mRNAs encoded by the genome.

The mRNAs, are recruited to be translated into proteins. The translation of mRNA or coding RNA allows the synthesis of proteins. The processes of transcription and translation are presented schematically in Fig.1.2. These RNAs carry codons, i.e. triplets of nucleotides coding for amino acids or stop codons which are signals to stop protein synthesis. There are thus 64 possible codons which code for 20 standard amino acids and 3 stop codons. There can be several codons for the same amino acid, the code is said to be degenerate. Like transcripts, proteins can undergo numerous post-translational modifications before they can perform their biological functions in the cell. This passage from DNA to protein is the basis of the fundamental theory or dogma of molecular biology. Currently it is impossible to access the exhaustive proteome with sufficient sensitivity, but the various technological breakthroughs in molecular biology in recent years have made it possible to access the transcriptome.

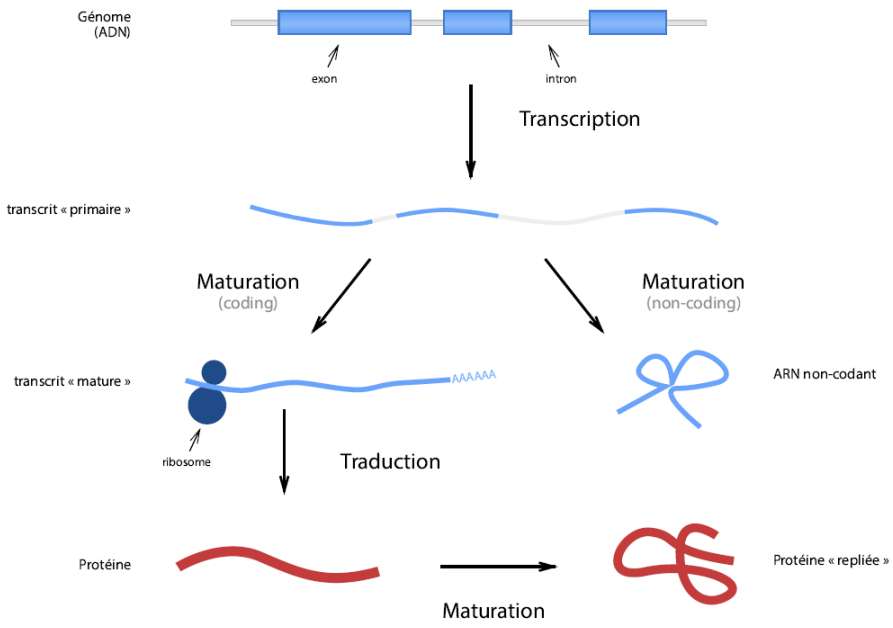


Figure 1.2: Modern molecular biology dogma, [Aud17].

In order to better understand the biology of cancers, the study of the transcriptome is essential. Indeed, investigating the transcriptome of malfunctioning cells gives the pos-

sibility to detect the mechanisms put in place by these tumor cells by which they bypass the various protections of our organism in order to proliferate in an uncontrolled way.

1.2 Sequencing technologies

1.2.1 Microarrays

The arrival of DNA chips or "microarrays" in the 90's and then the technological breakthrough access to "high throughput" sequencing methods have made it possible to establish the expression profiles of numerous tumors at the transcriptome level. The study of the transcriptome in oncology has proven to be a powerful diagnostic and prognostic tool.

DNA chips allow sequencing based on the principle of hybridization where two complementary nucleic acid fragments can associate by hydrogen bonding and dissociate in a reversible way under the action of heat and the saline concentration of the medium. They are presented in the form of a glass slide on which short DNA sequences (probes) have been deposited or synthesized. The probes have the particularity of having been chosen to be specific to a single gene.

The RNA to be analyzed, retrotranscribed into complementary DNA (cDNA) that can be visualized by incorporation of fluorochromes, are put in contact with the DNA chips where they will hybridize with the probes of which they are complementary. The reading of the chip allows to obtain the sequence spectrum of the cDNA or RNA from which they originate. We thus obtain the composition of the sample in sub-sequences of n nucleotides, where n is the size of the probes on the chip used. The computer processing of the spectrum then allows to quantitatively reconstruct the entire sequence and thus to characterize the initial transcriptome studied.

The error rate of the reading is quite low for this technology (less than 0.1%), Its main drawback is its relatively low throughput. The principle of transcriptome sequencing by DNA chips is presented in Fig.1.3.

1.2.2 Next generation sequencing and RNA-seq

Born in 2008, the technological breakthrough brought by Next Generation Sequencing (NGS) allows the digitization of the complete transcriptome by random sequencing of cDNAs and RNA copies using the RNA-seq method. The method, called sequencing

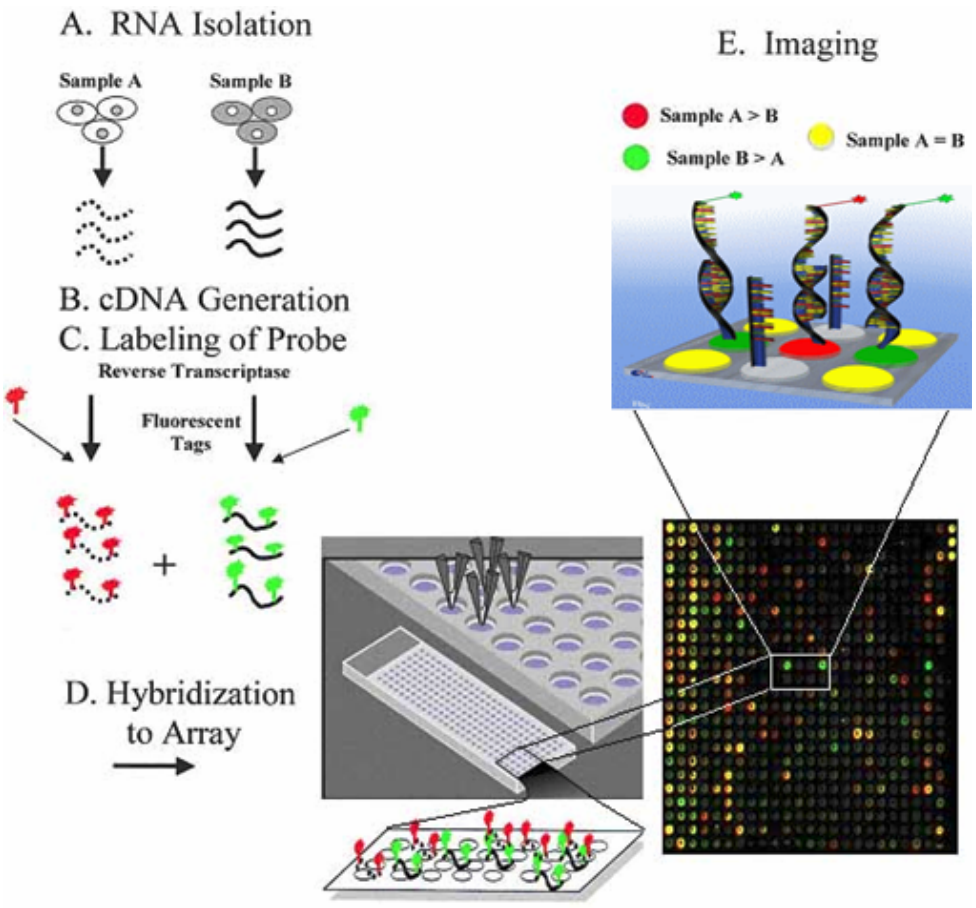


Figure 1.3: DNA chips or microarrays transcriptome sequencing principle, [Pla].

by synthesis, proposed by the company Illumina is currently the main high-throughput sequencing technology used for the transcriptome.

In RNA-seq protocols, once the RNA of interest is extracted, it is fragmented into sequences of a few hundreds of base pairs. The second step is reverse transcription (RT), which converts the RNA into complementary DNA (cDNA) and then into double-stranded cDNA that can be sequenced. The adapters represented by known short synthetic DNA sequences are linked to each end of the double-stranded cDNA fragments of unknown sequence, this preparation is called the library. The library is then deposited on a solid support, the flowcell. This flowcell is a glass surface on which complementary DNA sequences adapters are randomly arranged and allows the immobilization of the sequences present in the library. An amplification step of the hybridized sequences on the flowcell allows to increase the number of copies of each fragment, in order to form clusters of identical molecules.

A single-stranded DNA primer complementary to the adapters is added to initiate the synthesis of the DNA complementary to the single strand fixed on the flowcell. From this point on, the actual sequencing can begin. We distribute the 4 deoxyribonucleotides triphosphate each coupled to a different fluorochome on the flowcell, a complementary deoxyribonucleotide is attached to each fragment. An image capture is taken and a base is read for each cluster using specialized image processing programs. After the reading, the fluorochrome is photolyzed and a new deoxyribonucleotide is bound. Fixation, image capture and photolysis constitute a cycle of sequence by synthesis. The number of cycles is limited to 100-200 cycles depending on the speed of the sequencer and its error rate which increases with each cycle. The complete digitization of the fragments is therefore impossible and because of this limitation, we use pair-end sequencing. It allows to increase the size of the sequences read by sequencing successively the 2 ends of the fragments. The workflow of RNA-seq analysis is presented in Fig.1.4.

1.2.3 Third generation sequencing

One of the major limitations of RNA-seq technology is the fragmentation of DNA molecules, because the assembly of the read sequences is a complex problem. This limitation is even stronger for the transcript as the isoforms of a gene share large common portions. To solve this problem, another major technological revolution of sequencing is the development of third generation sequencing (TGS). It includes several technologies such as Pacific Bioscience (PacBio) single-molecule real-time (SMRT) or Oxford Nanopore Technologies

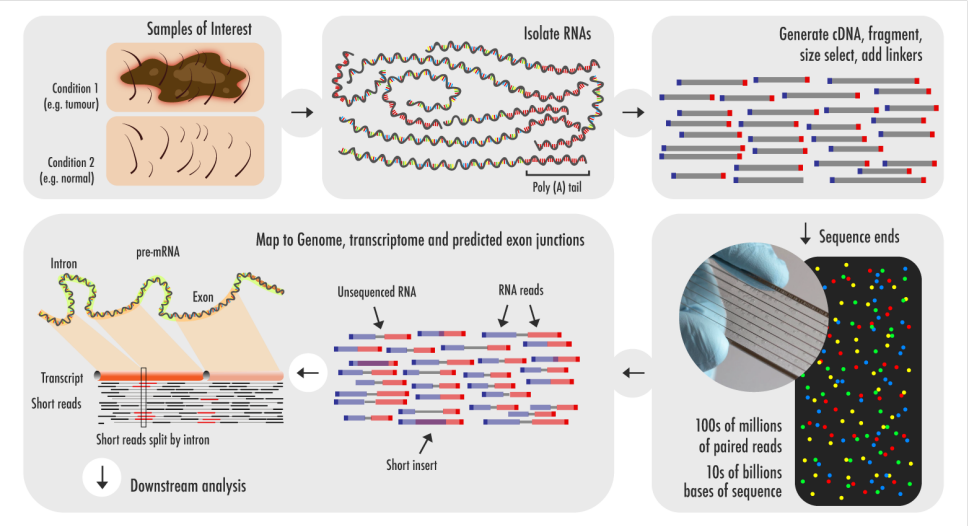


Figure 1.4: RNA-seq analysis workflow, [MN22].

(ONT) which allow direct sequencing of cDNA, without amplifications, without fragmentation and are able to digitize entire transcripts. The Oxford Nanopore technology is also the first to offer direct sequencing of RNA molecules, eliminating many of the biases associated with reverse transcription into cDNA.

The sequencing of long reads, however, poses new problems, as the error rate is higher than in the case of NGS. The combination of the two generations of sequencing should make it possible to overcome the above limitations [Aud17; Mar18; Ngu20]. The comparison of NGS and TGS technologies is presented in Fig.1.5.

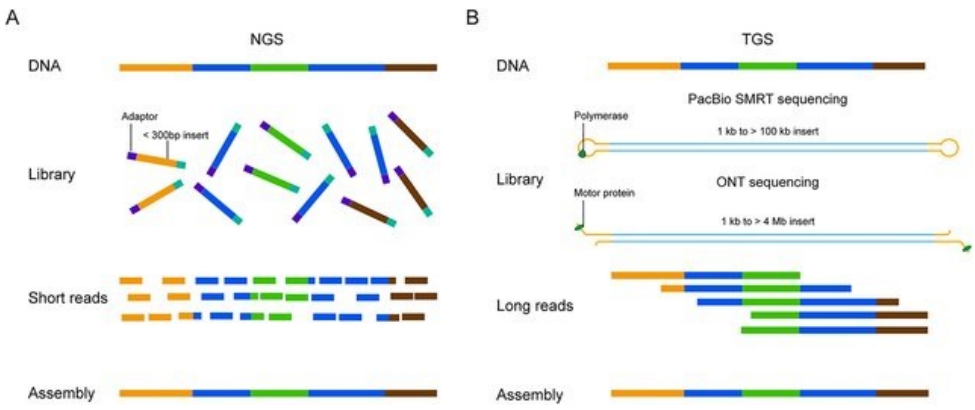


Figure 1.5: NGS (A) and TGS (B) technologies comparison, [CH21].

1.3 Bioinformatics analysis steps

The transcriptomic data used in this thesis were derived from RNA-seq technology. Therefore, we describe here the different steps of bioinformatics pre-processing and bioinformatics analysis required for this type of data (cf. Fig.1.6).

The sequences produced during RNA-seq sequencing are usually stored in the FASTQ format. This is a text file format that stores each read on four lines:

1. the unique sequence identifier;
2. the sequence read in the alphabet A, T, G, C or N (the base that could not be determined);
3. empty;
4. the quality code for each base read.

The sequencing file contains reads without any genomic or transcriptomic context. One of the steps in the analysis of RNA-seq data is therefore the alignment of the reads to the reference genome. Very greedy in computational resources, these methods are however very powerful, they allow the discovery of new genes or new transcripts, variants transcripts, splice variants, etc. Among the most used or recent genome alignment tools, we note STAR [Dob+13] ou HISAT2 [Kim+19].

An alternative to alignment to the reference genome is the strategy using the reference transcriptome. These methods have proven to be faster and more efficient, they allow the estimation of transcript and gene expression from reads in a few minutes on a desktop computer.

Finally, when a reference genome is not available or when one wishes not to introduce any bias with respect to our current knowledge of the genome or transcriptome, the so-called "de novo" approaches can be considered.

The alignment results are then used to quantify the expression of genes or transcripts. Quantification is the estimation of the abundance of transcripts using algorithms, the simplest being the counting of reads overlapping annotations like featuresCounts [LSS14] ou HTSeq-counts [APH15].

This raw read count does not accurately reflect gene expression, which makes it difficult to compare different conditions. Indeed, these values are impacted by the length of the transcripts, the total number of reads and the sequencing depth. For example, deeper

sequencing will produce more reads associated to each gene, to solve this problem several normalization strategies to correct this bias have been proposed, among the recent ones are DESeq2 [LHA14] ou TCGABiolinks [Col+16].

To study RNA-seq data, one of the most frequent analyses is the differential expression (DE) of genes. For example, it is important to find genes that are over-expressed or under-expressed in the group of patients with a treatment of interest compared to the control group. Another way to explore these large data is clustering, in this approach, we try to identify groups of patients or genes that are similar. This technique has in particular allowed the detection of several transcriptomic subtypes of ovarian cancer from the TCGA RNA-seq data [Bel+11].

Finally, this type of data also makes it possible to analyze patient survival [CZG18], in the framework of this approach we seek to discover prognostic genes and ultimately new therapeutic targets.

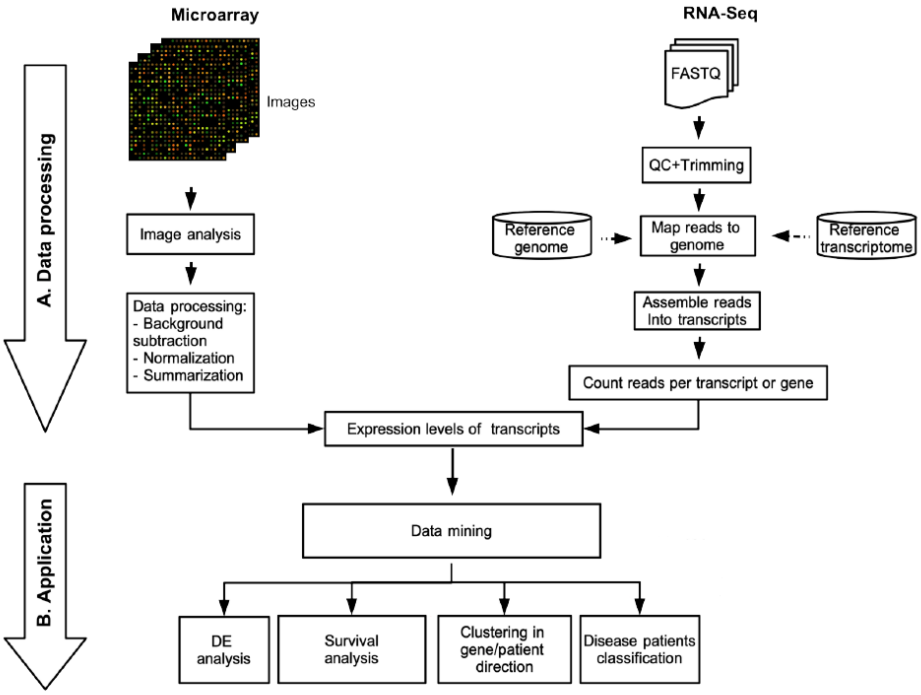


Figure 1.6: Transcriptomic data analysis pipeline, [Ngu20]. The micoriarrays and RNA-seq data are processed in 2 stages. A. Pre-processing, B. Analysis itself: differential expression, survival analysis and clusterint.

Chapter 2

Survival analysis

Survival analysis is a branch of statistics which deals with the data where the outcome is the time until the occurrence of an event of interest. One of the main challenges in this context is the presence of instances that do not experience any event during the observation period. Such a phenomenon is called censoring, the types of censoring include:

1. right-censoring, for which the observed survival time is less than or equal to the true survival time;
2. left-censoring, for which the observed survival time is greater than or equal to the true survival time;
3. interval censoring, for which we only know that the event occurs during a given time interval.

Among them, right-censoring is the most common scenario that arises in many practical problems. The applications of survival analysis in various domains are numerous, to cite a few: in healthcare to predict the disease recurrence or re-hospitalization, in reliability to prognosticate the device failure, in customer lifetime to model the purchase behavior, etc. In spite of the importance of these problems and relevance to various real-world applications, this research topic is scattered across different disciplines.

We introduce here some notations that will be necessary to describe the survival analysis methodology.

1. P - the number of input prediction features;
2. N - the number of instances or individuals;
3. X - feature matrix of size N x P
4. X_i - feature vector of individual i of size 1 x P

5. t_i - time to the event of interest for individual i

6. δ_i - event (value 1) or censoring (value 0) indicator for individual i

The survival function is the probability that the time to the event of interest is not earlier than a specified time t:

$$S(t) = Pr(T \geq t) \tag{2.1}$$

Given the input prediction features X, the goal of survival analysis is to estimate the survival time t_i, i.e. time to the event of interest for a new instance i, and to estimate the survival probability $\hat{S}(t_i)$ at the estimated survival time t_i.

On the contrary, the cumulative distribution function $F(t)$, which represents the probability that the event of interest occurs earlier than t, is defined as $F(t) = 1 - S(t)$, and probability density function can be obtained as $f(t) = \frac{d}{dt}F(t)$.

In survival analysis, another commonly used function is the hazard function $h(t)$, it is the rate of event at time t given that no event occurred before time t:

$$h(t) = \frac{f(t)}{S(t)} \tag{2.2}$$

The survival function defined in 2.1 can be rewritten as:

$$S(t) = exp(-H(t)), \tag{2.3}$$

where $H(t)$ is the cumulative hazard function (CHF).

The authors of [WLR17] reviewed the literature and created a taxonomy of the survival analysis approaches. According to this work, the survival analysis methods can be classified into two main categories: traditional statistical methods and machine learning based methods:

> ... they [statistical methods] focus more on characterizing both the distributions of the event times and the statistical properties of the parameter estimation by estimating the survival curves, while machine learning methods focus more on the prediction of event occurrence at a given time point by incorporating the traditional survival analysis methods with various machine learning techniques. Machine learning methods are usually applied to the high-dimensional problems, while statistical methods are generally developed for the low-dimensional data. In addition, machine learning methods for survival analysis offer more effective algorithms by incorporating survival

problems with both statistical methods and machine learning methods and taking advantages of the recent developments in machine learning and optimization to learn the dependencies between covariates and survival times in different ways [WLR17].

2.1 Statistical methods

The statistical methods in their turn can be divided into three groups:

1. non-parametric;
2. parametric.
3. semi-parametric;

2.1.1 Non-parametric methods

Non-parametric methods are more efficient when there is no underlying distribution for the event time or the proportional hazard assumption does not hold. In nonparametric methods, an empirical estimate of the survival function is obtained using Kaplan-Meier (KM) method or Nelson-Aalen estimator (NA).

The KM [KM58] is the most widely used method for estimating survival function. If $t_1 < t_2 < ... < t_K$ is a set of distinct ordered event times observed for $N(K \leq N)$ instances, for each specific event time $t_j(j = 1; 2; ...; K)$ the number of observed events is $d_j \geq 1$. The number of instances "at risk" (their event time or censored time is greater or equal to t_j) is $r_j = r_{j-1} - d_{j-1} - c_{j-1}$, where c_{j-1} is the number of censored instances during the time period between t_{j-1} and t_j. The conditional probability of surviving beyond time t_j is:

$$P(t_j) = \frac{r_j - d_j}{r_j} \tag{2.4}$$

Based on this conditional probability, the product-limit estimate of survival function 2.1 is:

$$\hat{S}(t) = \prod_{j:t_j<t} P(t_j) = \prod_{j:t_j<t} (1 - \frac{d_j}{r_j}) \tag{2.5}$$

Among all functions, the survival function or its graphical presentation is the most widely used one (Kaplan-Meier curves, Fig.2.1).

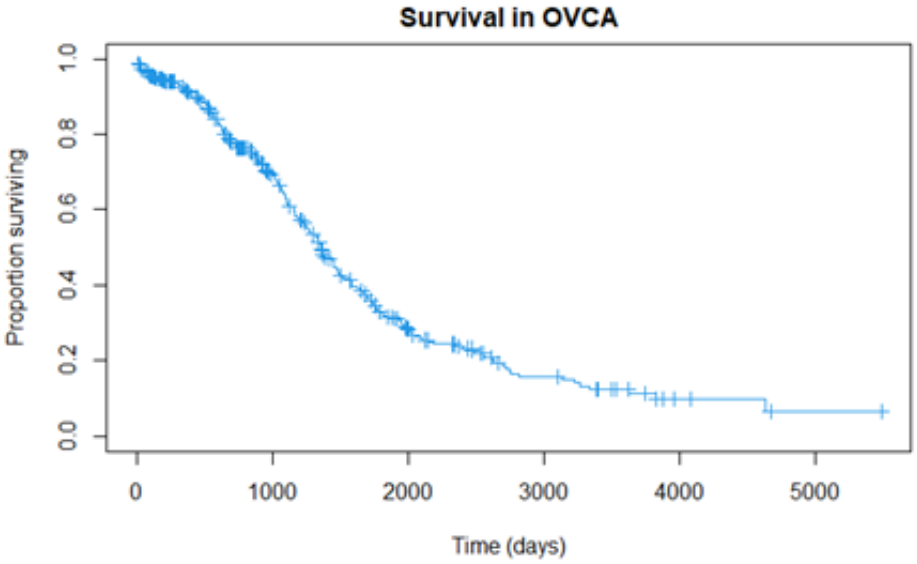


Figure 2.1: KM curve of the ovarian cancer (TCGA-OV) survival data.

The Nelson-Aalen estimator [Nel72; Aal78] is another non-parametric method. It is based on counting process approach and differs from the KM in that it estimates the CHF H for censored data and not the survival function:

$$\hat{H}(t) = \sum_{j:t_j<t} \frac{d_j}{r_j} \tag{2.6}$$

2.1.2 Parametric methods

The parametric censored regression models assume that the survival times or the logarithm of the survival times of all instances in the data follow a particular theoretical distribution. The commonly used distributions in parametric censored regression models are: normal, exponential, Weibull, logistic, log-logistic and log-normal [WLR17]. These methods were not used in this work as we preferred to focus on the survival models based on the artificial neural networks explained later in 2.2.3.

2.1.3 Semi-parametric method: Cox model

Semi-parametric models are a hybrid of the parametric and non-parametric approaches, they can obtain a more precise estimator than the non-parametric methods and more flexible estimator compared to the parametric models. Cox proportional hazards (PH) model [Cox72] is the most commonly used survival analysis method in this category. Unlike parametric methods, the knowledge of the underlying distribution of time to event of interest is not required, but the attributes are assumed to have an exponential influence on the outcome:

$$h(t, X_i) = h_0(t)exp(X_i\beta) \tag{2.7}$$

where $h_0(t)$ is the baseline hazard function (an arbitrary nonnegative function of time), and $\beta^T = (\beta_1, \beta_2, ..., \beta_P)$ is the coefficient vector. The hazard ratio between two instances:

$$\frac{h(t, X_i)}{h(t, X_j)} = exp\left[(X_i - X_j)\beta\right] \tag{2.8}$$

This hazard ratio is a constant, it is independent of the baseline hazard function $h_0(t)$ but all the subjects share the same $h_0(t)$. Because this baseline hazard function is unspecified in Cox model, it is impossible to fit the model using standard likelihood function, instead the partial likelihood is used:

$$L(\beta) = \prod_{j=1}^{N} \left[\frac{exp(X_j\beta)}{\sum_{i \in R_j} exp(X_i\beta)}\right]^{\delta_j} \tag{2.9}$$

where R_j is the set of indices, i, with $t_i \geq t_j$ (those "at risk" at time t_j). The coefficient vector is estimated by maximizing this partial likelihood, or equivalently, minimizing the negative log-partial likelihood for improving efficiency:

$$LL(\beta) = -\sum_{j=1}^{N} \delta_j \left\{X_j\beta - log\left[\sum_{i \in R_j} exp(X_i\beta)\right]\right\} \tag{2.10}$$

With the development of data collection and detection techniques, most real-world domains tend to encounter high-dimensional data. In some cases, the number of variables P in the given data is almost equal to or even exceeds the number of instances N. It is challenging to build the prediction model with all the features and the model might provide inaccurate results because of the overfitting problem. This motivates using sparsity norms to select vital features in high-dimension under the assumption that most of the features

are not significant [FHT10].

For the purpose of identifying the most relevant features to the outcome variable among tens of thousands of features, different penalty functions can be applied to a Cox model resulting in a regularized Cox models, including lasso l_1-norm, ridge l_2-norm or a combination of two (elastic net), etc. One of the examples of regularized Cox models is Coxnet proposed by [Sim+11] and integrated in a generic R package glmnet. The lasso penalty tends to choose only a few nonzero coefficients. While often desirable, this can cause problems. If two predictors are very correlated, the lasso will pick one and entirely ignore the other. On the other hand, ridge regression scales all the coefficients towards 0, but sets none to exactly zero. This helps to regularize in problems with $P > N$, but does not give a sparse solution. However, ridge regression better handles correlated predictors. If two predictors are very correlated, ridge regression will tend to give them equal weight.

The regularizer in Lasso-Cox is of the form: $\lambda \sum_{p=1}^{P} |\beta_p|$, in Ridge-Cox: $\frac{\lambda}{2} \sum_{p=1}^{P} \beta_p^2$ and elastic net Cox: $\lambda \left(\alpha \sum_{p=1}^{P} |\beta_p| + \frac{1}{2}(1-\alpha) \sum_{p=1}^{P} \beta_p^2 \right)$.

Another modification of Cox model is Cox-Boost [Bin+13]. It proposes the possibility to incorporate the mandatory features into the final model while fitting the sparse survival models on the high dimensional data. This approach estimates the coefficients of the Cox model by creating partitions, it considers one partition of candidate variables for updating in each boosting step. The partition that leads to the largest improvement in the penalized partial log likelihood is selected and in subsequent iterations, the model selects another partition and refits those variables by maximizing the penalized partial log likelihood.

2.2 Machine learning methods

This section introduces some of the machine learning concepts as well as the machine learning survival methods necessary for further explanation in Chapter 3, where the comparison of the approaches of the gene expression based survival analysis found in literature is presented. A more exhaustive list of the machine learning methods in survival analysis can be found in [WLR17].

2.2.1 Machine learning paradigm

The advantages of machine learning, such as its ability to model the non-linear relationships and the quality of their overall predictions made, have resulted in the achievement

of significant success in various practical domains in the past several years. The main challenge of the machine learning methods in survival analysis is the difficulty to appropriately deal with censoring and the time estimation of the model [WLR17].

Machine Learning (ML) is a branch of Artificial Intelligence (AI), it relates the problem of learning from data samples to the general concept of inference. There are two phases in every learning process:

1. estimation of unknown dependencies in a system from a given dataset, i.e. training phase;

2. use of estimated dependencies to predict new outputs of the system, i.e. test or validation phase.

In biomedical research, ML has turned out to be a very promising area with many applications where, using different techniques and algorithms and by searching over a P-dimensional space of biological data, an acceptable generalization is obtained. Two main common types of ML methods can be distinguished: (i) supervised learning and (ii) unsupervised learning.

In supervised learning, the training dataset has labels which are used to map the input data to the desired output. On the contrary, the unsupervised learning methods try to find the patterns or to discover the hidden groups structure in the input data without any notion of the output or labels.

The most common example of the supervised learning is a classification problem. In the classification task, a learning process categorizes the data into a set of finite classes. Two other common ML tasks are regression (supervised) and clustering (unsupervised). In the regression task, a learning function maps the data into a real-value variable. As for clustering, it tries to find the groups or clusters in order to describe the data samples. Once the training phase is finished, in the validation or test phase, the new sample can be assigned a class (classification task) or a cluster (clustering) or used to estimate the predictive variable (regression task).

A combination of supervised and unsupervised learning has been widely applied, giving birth to another type of ML methods, i.e. semi-supervised learning. It combines labeled and unlabeled data in order to construct an accurate learning model. Usually, this type of learning is used when there are more unlabeled datasets than labeled ones [Kou+15].

Once a ML model is obtained, the training and generalization errors can be estimated on the training data and test data respectively. For example, a good classification model

should fit the training set well and accurately classify the test instances. If the test error rates of a model begin to increase even though the training error rates decrease then the phenomenon of model overfitting occurs.

When a ML model is developed by means of different ML techniques, it is crucial to evaluate its performance. In general the performance analysis of each proposed model is measured in terms of sensitivity, specificity, accuracy and area under the curve (AUC). The confusion matrix presented in Tab.2.1 is a table with 4 different combinations of predicted and actual values of a classification model:

	Actual values	
Predicted values	True Positive	False Positive
	True Negative	False Negative

Table 2.1: Confusion matrix.

Sensitivity or recall or True Positive Rate (TPR) is a proportion of true positives that are correctly observed by the classifier:

$$Sensitivity = True\ Positive/(True\ Positive + False\ Negative) \tag{2.11}$$

The specificity is given by the proportion of true negatives that are correctly identified:

$$Specificity = True\ Negative/(True\ Negative + False\ Positive) \tag{2.12}$$

The quantitative metrics of accuracy and AUC are used for assessing the overall performance of a classifier. Specifically, accuracy is a measure related to the total number of correct predictions. On the contrary, AUC is a measure of the model's performance which is based on the Receiver Operating Characteristics (ROC) curve that plots the tradeoffs between sensitivity and 1-specificity (False Positive Rate or FPR) when model parameters vary. An excellent model has AUC near to 1, a poor model has an AUC near 0 and when AUC is 0.5, it means the model has no class separation capacity whatsoever.

The predictive accuracy of the model is computed from the test set which provides an estimation of the generalization errors. The training and test sets with known labels should be independent and sufficiently large in order to obtain reliable estimation of the predictive performance of a model. The initial labeled data are generally split into training and test subset using the following sampling methods: (i) Holdout Method, (ii) Random Sampling, (iii) Cross-Validation and (iv) Bootstrap [Kou+15]. The comparison of the Holdout, Cross-Validation and Bootstrap sampling is presented in Fig.2.2

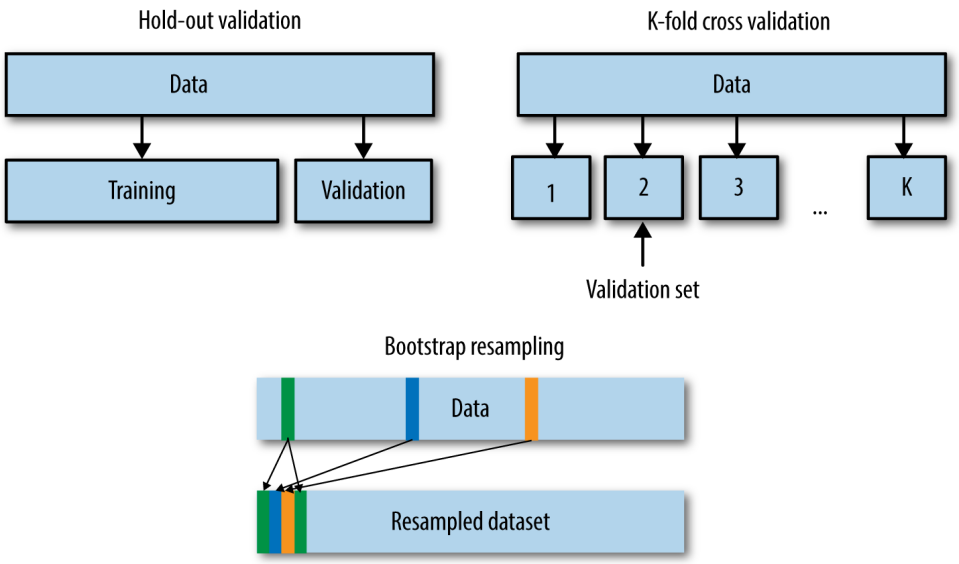


Figure 2.2: Comparison of Holdout validation, k-fold Cross-Validation, and Bootstrap sampling [Vik18].

In the Holdout method, the data samples are simply split into two separate datasets: the training and the validation or test. A classification model is then obtained from the training set while its performance is estimated on the test set. Random sampling is a similar approach to the Holdout method. In this case, for the sake of a more accurate estimation, the Holdout method is repeated several times, choosing the training and test instances randomly (cf. Fig.2.3).

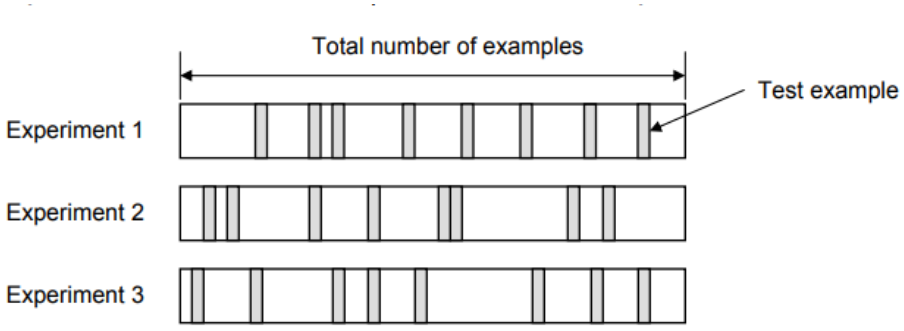


Figure 2.3: Random sampling technique [Vik18].

In the Cross-Validation approach, the original data are split into k folds, $k-1$ folds are used for training a model and 1 fold for testing (cf. Fig.2.4). As a result, the original dataset is covered successfully both in the training and in the test set. The overall accuracy

results are calculated as the average of all different validation cycles.

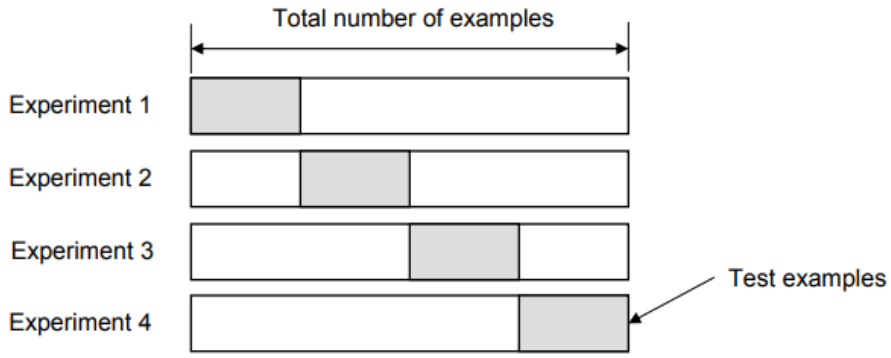


Figure 2.4: K-fold Cross-Validation sampling technique [Vik18].

In the last bootstrap approach presented in Fig.2.5, the training samples are randomly selected with replacement from the original complete dataset. The remaining examples that were not selected for training are used for testing. Unlike K-fold cross-validation, the value is likely to change from fold-to-fold and the overall error rate of the model is calculated by averaging the error rates of all the experiments.

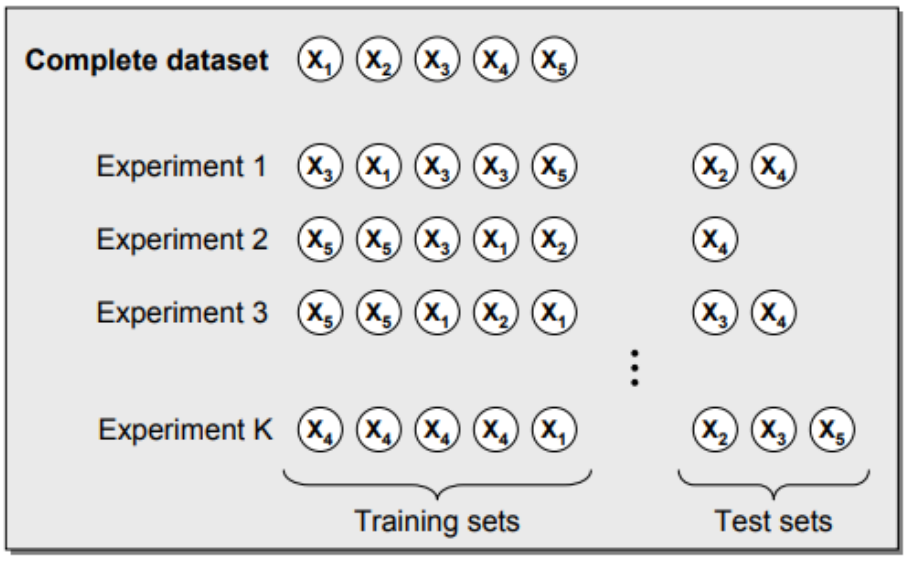


Figure 2.5: Bootstrap sampling technique [Vik18].

2.2.2 Survival trees and random survival forests

The decision tree models recursively partition the data on the basis of some splitting criterion, thus the nodes are formed by the objects similar to each other based on the event of interest. Survival trees are tailored to handle censored data and have particular splitting criteria which can be grouped into two categories: (i) maximizing between-node heterogeneity and (ii) minimizing within-node homogeneity. Another important aspect of building a survival tree is the selection of the final tree. Procedures such as backward selection or forward selection can be followed for choosing the optimal tree. However, an ensemble of trees, for example, random survival forest can avoid the problem of final tree selection with better performance compared to a single tree.

Random Survival Forests (RSF) are a tree-based, non-linear, ensemble method [Ish+08]. The graphical representation of the RSF algorithm is given in Fig.2.6. The steps in the RSF algorithm are as follows:
(i) Draw B bootstrap samples randomly from the given dataset, they will serve as training sets to grow trees. The remaining samples of the original dataset are called out-of-bag (OOB) data, they represent approximately one third of the original data and will serve as test sets to evaluate performance.
(ii) For each bootstrap sample, grow a survival tree. Select randomly a subset of features at each node, split the node by using the feature with the largest survival difference between the daughter nodes.
(iii) Grow the tree until the terminal node contains not less than a predefined positive number of uniques events.
(iv) Calculate the cumulative hazard function (CHF) H using NA estimator for each tree (cf. 2.1.1) and the ensemble CHF by averging over the trees, use the OOB data to calculate the ensemble CHF and prediction error.

2.2.3 Artificial neural networks and survival analysis

The extension of Cox regression with artificial neural networks was first proposed by Faraggi and Simon [FS95], who replaced the linear predictor of the Cox regression model, by a one hidden layer multilayer perceptron (MLP). The schema of an MLP with one hidden layer is given in Fig.2.7. It is a fully connected feedforward Artificial Neural Network (ANN), which can have multiple hidden layers. The layers are composed of neurons and the neurons implement the non-linearity, i.e. they apply the non-linear activation function

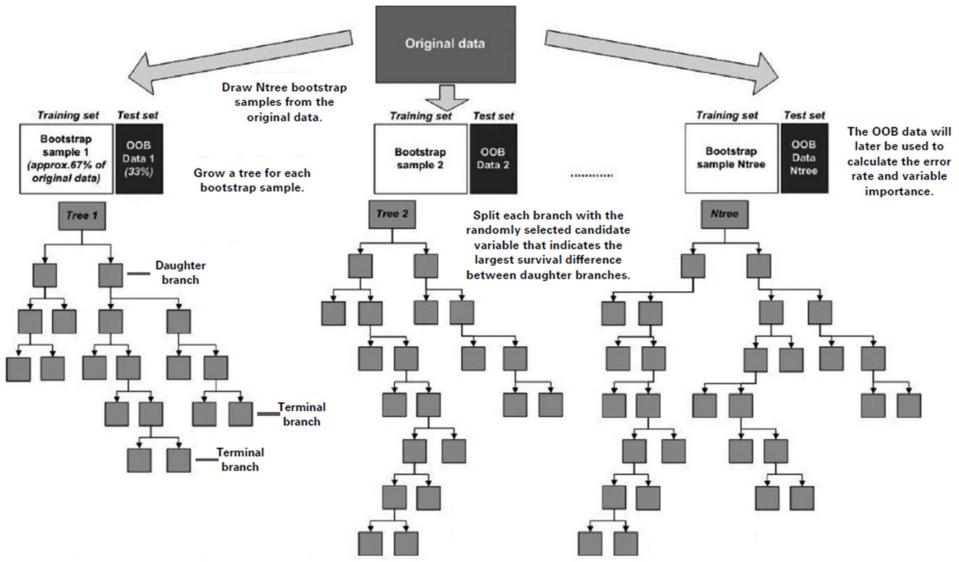


Figure 2.6: Graphical presentation of the Random Survival Forest (RSF) algorithm, [Zut+21]. OOB, out-of-bag data.

to their input. There exists various activation functions, to cite the most used ones: hyperbolic tangent (tanh), sigmoid, rectified linear unit (ReLU) [NH10], Scaled Exponential Linear Unit (SELU) [Kla+17].

The training of the MLP is done by adjusting the weights of the neurons after all input data is processed. The adjustment is based on the error between the output and the expected result and is carried out through backpropagation. This error or the loss function in the neural survival networks is generally a negative log likelihood. We give a detailed description of the survival loss functions based on the negative log likelihood and the comparison of their main characteristics in Part II, chapter 1 of this work.

Here we just introduce the reviewed survival neural network models and discuss their structure:

- Cox-nnet [CZG18];
- DeepSurv [Kat+18];
- SurvivalNet [You+17];
- Cox Case Control [KBS19];
- Cox Time [KBS19];

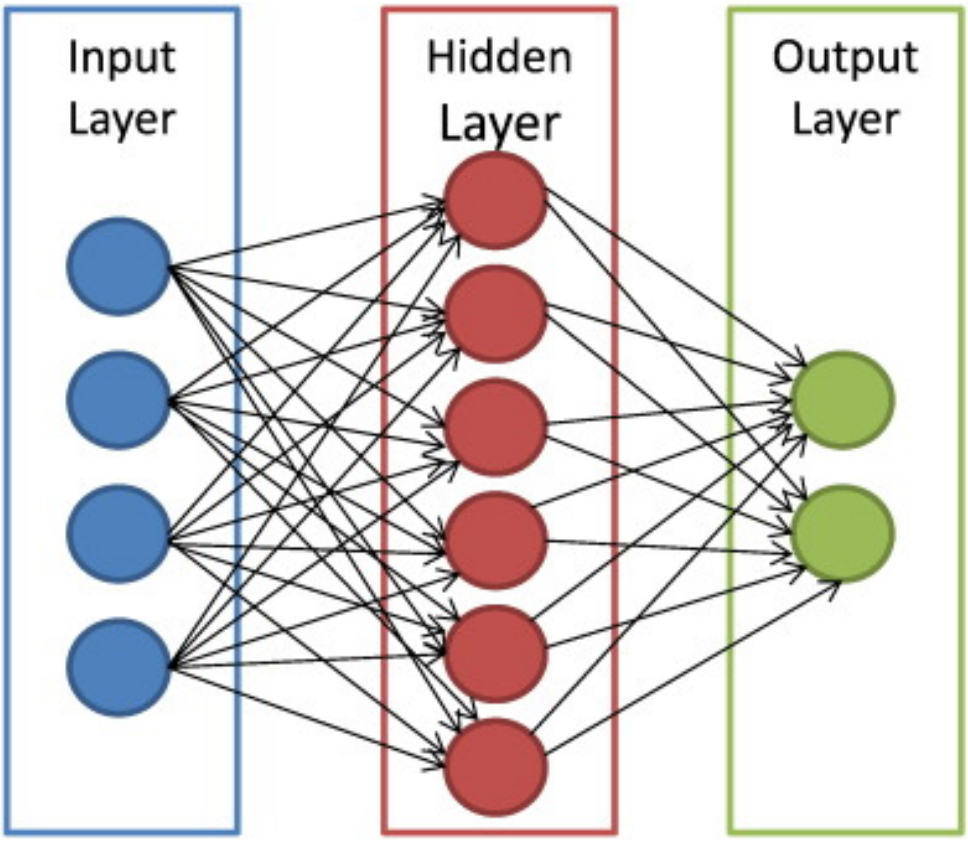


Figure 2.7: A multilayer perceptron (MLP) with 1 hidden layer, [Kou+15].

- Piecewise Constant hazard [KB19];
- Logistic Hazard [KB19];
- Nnet-survival [GN18];
- PMF [KB19];
- N-MTLR [Fot18].

The output of the reviewed networks is the survival or hazard probability, one neuron output for time-independent models and multiple neurons for time-dependent models. The schematic representation of the Cox-nnet architecture as an example of survival ANN with one neuron output and one hidden layer is given in Fig.2.8. The example of architecture with multiple neurons in output layer is N-MTLR (cf. Fig.2.9).

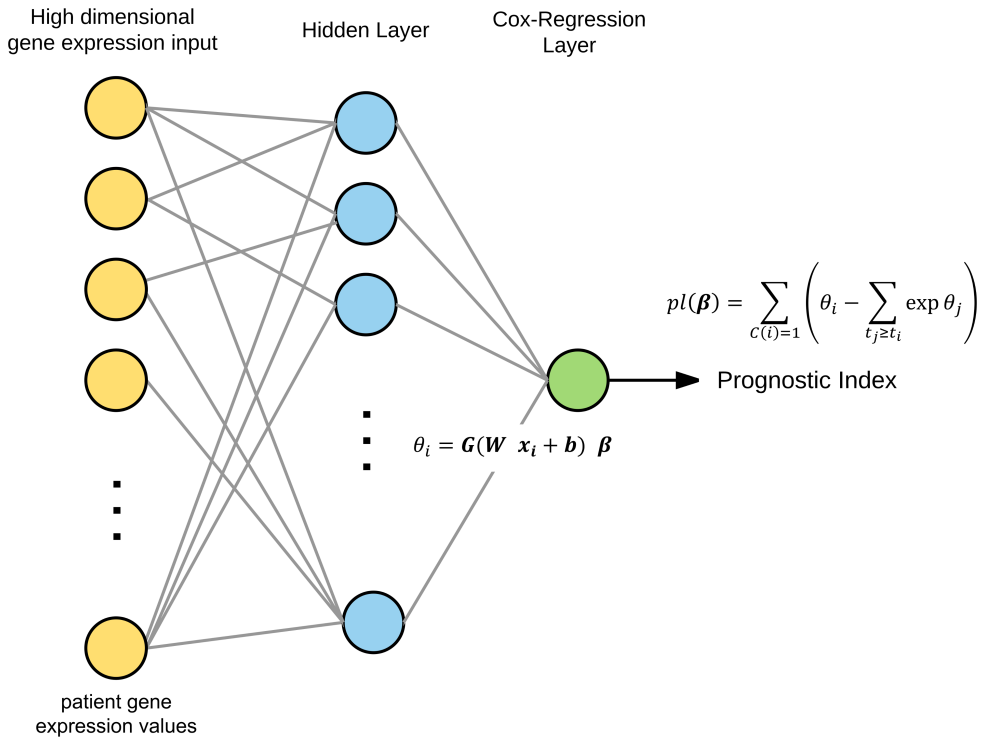


Figure 2.8: Cox-nnet architecture, representation of a 1-hidden layer transformation, [CZG18].

A more complex loss function and structure are used by following models: DeepHit [LZY18] and RNN-Surv [GNS18].

DeepHit was designed to account for competing risks (see the subnetworks in (Fig.2.10). For one competing risk, this model resembles PMF of [KBS19], but with the major difference in the loss function. Indeed, the loss function of DeepHit is composed of two parts: one based on the discrete time log likelihood (as PMF in case of one competing risk) and the second - on the ranking ability of the network which penalizes the incorrect ordering of events. We will discuss it in details in Part II, chapter 3.

RNN-Surv is based on the Long Short-Term Memory (LSTM) [HS97] cells which exploit the sequential nature of the problem (Fig.2.11). The loss function of this model is composed of two parts as well: the first one is a modified cross-entropy function able to take into account the censored data (it is in fact the negative log likelihood for Bernoulli data [Bro75; KB19]) and the second one based on the C-index (discussed in the next section).

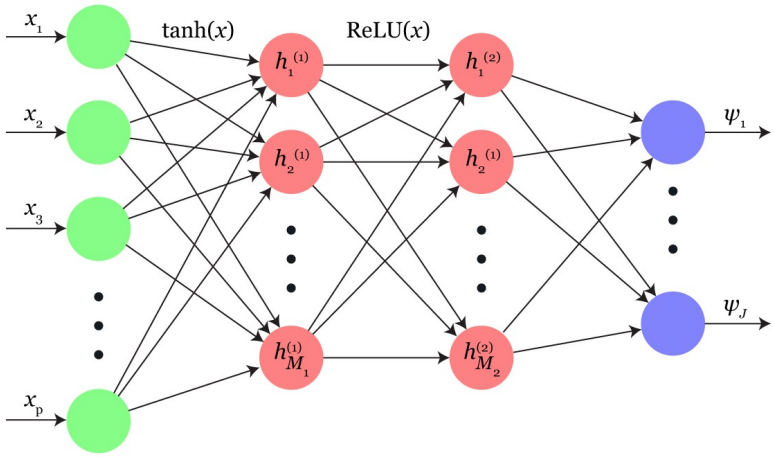


Figure 2.9: N-MTLR architecture, representation of a 2-hidden layer transformation, [Fot18].

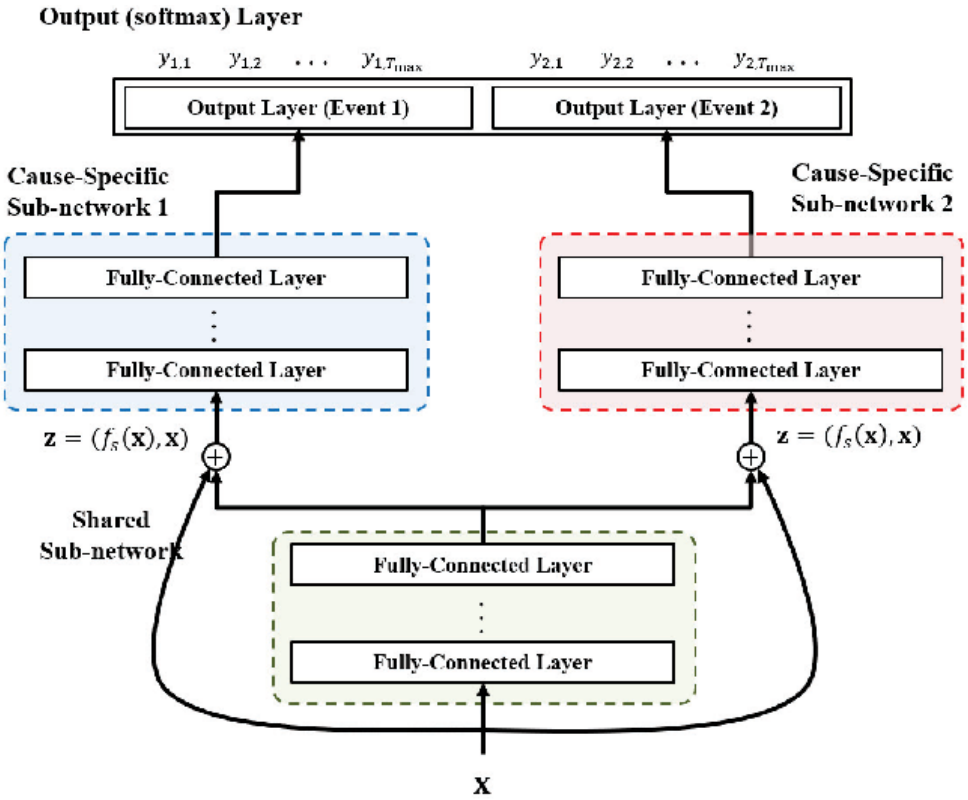


Figure 2.10: DeepHit architecture, [LZY18].

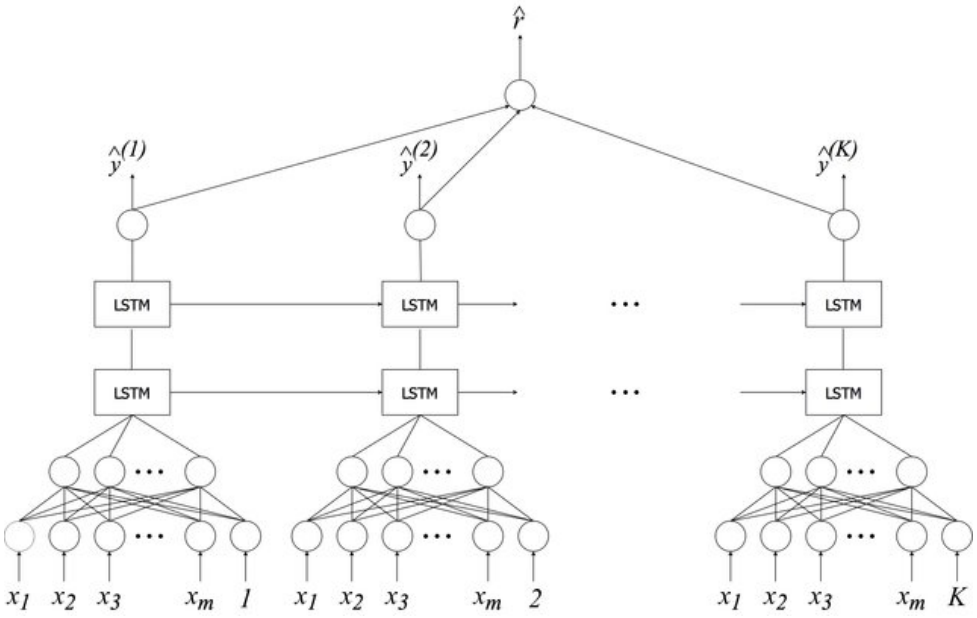


Figure 2.11: RNN-Surv architecture, [GNS18].

2.3 Evaluation criteria

Because of the presence of the censored data, the standard evaluation metrics are not applicable for measuring the performance of the survival models. The prediction performance is computed using more specialized evaluation criteria.

2.3.1 Concordance index

The Concordance index (C-index) or Harrell's index [Har+84] is widely used in survival analysis. It calculates the proportion of the concordant pairs of observations in all the comparable pairs. In other words, it estimates the probability that for a random pair of individuals, the one experiencing the event first had a worse predicted outcome.

Since the ordering of the predictions in the proportional hazards models does not change over time, the C-index is well suited for evaluation of these models. However, it is not the case of the time-dependent methods, where rather time-dependent C-index by Antolini [ABB05] is used. It still estimates the probability that observations i and j are concordant given that they are comparable but summarizes it over the observed time. The time-dependent Antolini C-index was modified by [KBS19] to account for tied event times and survival estimates, we will refer to it as adjusted Antolini C-index:

$$C_{td} = P\left\{\hat{S}(t_i|X_i) < \hat{S}(t_i|X_j)|t_i < t_j, \delta_i = 1\right\} \tag{2.13}$$

where $\hat{S}(t)$ is the estimated survival function as in 2.3 $S(t) = exp\left(-\int_0^t h(s)ds\right)$.

Hereafter we give the pseudo-code for the adjusted time-dependent Antolini C-index computation:

Algorithme for the adjusted time-dependent C-index

For each t_i do
  For each t_j do
    If $i! = j$ Then
      If ($t_i < t_j$ and $\delta_i == 1$) or ($t_i == t_j$ and ($\delta_i == 1$ or $\delta_j == 1$)) Then
        $sum\_comparable = sum\_comparable + 1$
        If ($t_i < t_j$) Then
          If ($\hat{S}_i < \hat{S}_j$) Then
            $sum\_concordant = sum\_concordant + 1$
          Else If ($\hat{S}_i == \hat{S}_j$) Then
            $sum\_concordant = sum\_concordant + 0.5$
          End If
        Else If ($t_i == t_j$) Then
          If ($\delta_i == 1$ and $\delta_j == 1$) Then
            If ($\hat{S}_i! = \hat{S}_j$) Then
              $sum\_concordant = sum\_concordant + 1$
            Else If ($\hat{S}_i == \hat{S}_j$) Then
              $sum\_concordant = sum\_concordant + 0.5$
            End If
          Else If ($\delta_i == 1$) Then
            If ($\hat{S}_i < \hat{S}_j$) Then
              $sum\_concordant = sum\_concordant + 1$
            Else If ($\hat{S}_i == \hat{S}_j$) Then
              $sum\_concordant = sum\_concordant + 0.5$
            Else If
          Else If ($\delta_j == 1$) Then
            If ($\hat{S}_i > \hat{S}_j$) Then
              $sum\_concordant = sum\_concordant + 1$
            Else If ($\hat{S}_i == \hat{S}_j$) Then
              $sum\_concordant = sum\_concordant + 0.5$
$Cindex\_td = sum\_concordant/sum\_comparable$

C-index in general is similar to the AUC and the classification accuracy [WLR17], the C-index of 1 corresponds to the best possible model. To note as well that for the proportional hazards models the time-dependent Antolini C-index is equivalent to the regular C-index.

2.3.2 Brier score

The Brier score (BS) developed by [Bri50] is designed to calculate the mean squared error of the probability estimates for binary classification problems. It was extended to the survival problems with censoring by [Gra+99] by weighting the scores by the inverse censoring distribution. The BS formula is given in [KBS19]:

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\hat{S}(t_i|X_i)^2 \mathbb{1}\left\{t_i \leq t, \delta_i = 1\right\}}{\hat{G}(t_i)} + \frac{(1-\hat{S}(t_i|X_i))^2 \mathbb{1}\left\{t_i > t\right\}}{\hat{G}(t)}\right] \tag{2.14}$$

where $\hat{G}(t) = P\left\{t_i > t, \delta_i = 0\right\}$ is the Kaplan-Meier estimate of the censoring survival function.

The BS can be extended to a time interval, giving the integrated Brier score (IBS) [KBS19]:

$$IBS = \frac{1}{Max(t_i)}\int_0^{Max(t_i)} BS(t)dt \tag{2.15}$$

In practice, this metric is approximated by the numerical integration over a predefined number of grid points, the smaller the IBS values, the better is the performance of the evaluated model.

2.3.3 Kaplan-Meier curves

Another common approach in evaluating the performance of a survival model is to define two or more groups of individuals based on the output of the survival model and to visualize the Kaplan-Meier curves of these groups (section 2.1.1). The comparison of the curves is usually done with a log-rank test, the null hypothesis being that there is no difference in survival between the groups. It is a non-parametric test, since it does not make any assumptions about the survival distributions. Essentially, the log-rank test compares the observed number of events in each group to what would be expected if the

null hypothesis were true (i.e., if the survival curves were identical) [Kas]:

$$\chi^2 = \sum_i^n \frac{(O_i - E_i)^2}{E_i}, \tag{2.16}$$

Where O_i is the number of observed events in group i, E_i is the number for expected events in group i and n is the number of groups. Usually, the threshold of 0.05 of the log-rank test p-value is considered as significant. The better visual separation of the Kaplan-Meier curves as well as the significance of the log-rank test can be interpreted as better survival model performance.

Chapter 3

Survival analysis and gene expression

An obvious trend in the field of survival analysis includes the integration of mixed data, such as clinical and genomic data. It is clear that the application of ML methods could improve the accuracy of cancer susceptibility, recurrence and survival prediction. Based on the literature reviewed by [Kou+15], the accuracy of cancer prediction outcome has significantly improved by 15%–20% the last years, with the application of ML techniques.

3.1 Dimensionality reduction

Feeding the raw high-dimensional gene expression data to the ML model, might not result in a satisfactory performance because of the so called "curse of dimensionality", where $P >> N$. Therefore, a number of different techniques and strategies are used as a preprocessing stage to extract the useful information concealed in the input data in order to eliminate irrelevant features, to reduce noise and to produce more robust learning models due to the involvement of fewer features. There are two main categories of methods for this goal, i.e., feature selection and feature extraction. We present hereafter some of the dimensionality reduction techniques which were used with gene expression data in the purpose of further survival analysis.

3.1.1 Feature selection

In feature selection, the goal is to find an inclusive subset of original features which usually either improves or maintains the accuracy or simplifies the model complexity. When there are P number of features, the total number of possible subsets are 2^P. The main idea is to come up with a reduction criterion instead of testing the exponential number of subsets. The evaluation of the subsets based on some criterion can be categorized into: (i) the

filters, (ii) the wrappers and (iii) the embedded techniques [HG15].

Filter methods

Filter methods work without taking the ML model into consideration. This makes them very computationally efficient, they can be multivariate (able to find relationships among the features) and univariate (consider each feature separately) [HG15]. In these methods various ranking mechanisms are used to grade the features (variables) and the features are then removed by setting a threshold. They are categorized as filter methods because they filter the features before feeding to a learning model. Filter methods are based on two concepts "relevance" and "redundancy", where the former is dependence (correlation) of feature with target and the latter addresses whether the features share redundant information. Correlation Criteria (CC), also known as Dependence Measure (DM), is based on the relevance (predictive power) of each feature. The predictive power is computed by finding the correlation between the independent feature X^j and the target (label). The feature with the highest correlation value will have the highest predictive power and hence will be most useful. The features are then ranked according to some correlation based heuristic evaluation function.

The majority of the reviewed papers use the univariate Cox regression analysis (section 2.1.3) to filter the survival related genes before applying further ML models [Bel+11; Kim+15; Wan+18; PCG18; Cha+18]. Each gene is used in the Cox Proportional Hazards analysis and usually a log-rank test threshold of 0.05 is applied to filter the genes with the prognostic power.

The chi-square statistic method can be used to measure the relevance of the feature to the target. The authors of [Che+15] used the chi-square value to select 10 genes related to the adjuvant chemotherapy treatment outcome in lung cancer patients.

The minimal Redundancy Maximal Revelvance (mRMR) is used in the work of [Zha+16] to diminish the number of gene expression features. This method is based on maximizing the relevance and minimizing redundancy of features. The authors obtained 130 features, consisting of 71 gene expression variables, 50 gene methylation variables, 3 miRNA expression variables and 4 gene copy numbers.

Another way to filter the genes is to use the Gene Coexpression Network (GCN) analysis. The authors of [Hua+19a] selected the genes of the calculated gene co-expression modules, they used local maximal Quasi-Clique Merger (lmQCM) algorithm [Hua+19b] which can discover densely connected gene modules across samples/patients.

Generally, a GCN is constructed by considering each gene as a node and the correlation between the expression profiles of two genes is often used to annotate the edge between them. In order to define the GCN a threshold can be applied to the correlation coefficient values to determine if there is an edge linking the nodes. If the correlation between two nodes is higher than the threshold, an edge exists between the two nodes; otherwise the two nodes are not connected. The disadvantage of this approach, is that the resulting unweighted GCN is often sensitive to the choice of the threshold. The alternative is to define a weight for each edge based on the correlation values thus constructing a weighted GCN (WGCN).

There are many ways to calculate the correlation or otherwise known Dependence Measure (DM). The commonly used Pearson correlation coefficient (PCC) is based on a linear model [Gho+19]:

$$\rho_{ij} = \frac{cov(X^i, X^j)}{\sqrt{var(X^i)var(X^j)}}, \tag{3.1}$$

where $cov(.,.)$ and $var(.,.)$ denote covariance and variance respectively, X^i and X^j are the genes expression profiles i and j.

Nonlinear metrics such as Spearman rank correlation and mutual information (MI) can also be used. Spearman correlation coefficient is defined as the PCC between the rank variables [Dow15]:

$$r_{ij} = \frac{cov(R(X^i), R(X^j))}{\sqrt{var(R(X^i))var(R(X^j))}}, \tag{3.2}$$

where $R(X^i)$ and $R(X^i)$ are the ranks of the genes X^i and X^j.

In network analysis, the dense subnetwork modules could be of different types: cliques (fully connected), quasi-cliques (densely connected) and k-core (each node has at least k edges). One of the widely used WGCN analysis tools is the WGCNA package developed by Horvath's group [LH08]. It uses the hierarchical clustering to identify the densely connected subnetworks. While it is an effective method, hierarchical clustering prevents overlaps between subnetworks even though a gene may participate in different functions and thus appear in multiple subnetworks. An alternative lmQCM algorithm was proposed by [ZH14] for mining the locally dense structures with the network weight normalization process inspired by the spectral clustering in machine learning. lmQCM is a revision of the edge-covering quasi-clique merger (eQCM) algorithm for directly mining weighted

networks based on a greedy algorithm called QCM.

Wrapper methods

Wrapper methods integrate the model within the feature subset search, while filter methods select the optimal features to be passed to the learning model, i.e., classifier, regression, etc. In this way, different subsets of features are found or generated and evaluated through the model. The fitness of a feature subset is evaluated by training and testing it on the model. Thus in this sense, the algorithm for the search of the best suboptimal subset of the feature set is essentially "wrapped" around the model. The search for the best subset of the feature set, however, is an NP-hard problem. Therefore, heuristic search methods are used to guide the search. These search methods can be divided in two categories: Sequential and Metaheurisitc algorithms [Gho+19].

Sequential feature selection algorithms access the features from the given feature space in a sequential manner. These algorithms are called sequential due to the iterative nature of the algorithms. We haven't found studies which use the sequential feature selection for the survival analysis.

The metaheuristic algorithms, also referred to as evolutionary algorithms, have low implementation complexity and can adapt to a multitude of problems. They are also less prone to get stuck in a local optima as compared to sequential methods. As examples of heuristic methods for gene expression feature selection, Genetic Algorithms (GA) are explained here. In GA, the potential solutions are represented by chromosomes, a sequence of 1 or 0. For feature selection, the genes in the GA chromosome correspond to features and can take values 1 or 0 for selection or not selection of feature, respectively. The generations of chromosomes improve by crossovers and mutations until an optimal solution is found [Whi94].

The authors of [Kim+15] used grammatical evolution (GE) for the feature selection "wrapped" into a neural network (GENN [Mot+08]). GE is a variation on Genetic Programming (GP) and is an evolutionary search algorithm, a flexible type of GP. GE uses a Backus-Naur form (BNF) grammar which is simply a set of rules for translating the array of bits (chromosome) into a NN, much like DNA is transcribed into RNA. GENN uses a GA to evolve the binary string which represents the chromosome encoding for NN structure (the input features, the weights, the activation functions, etc.). The fitness of the NN can then be evaluated, and the fittest individuals are most likely to "reproduce" in this evolutionary process. Thus, the algorithm automatically selects the appropriate network

architecture for any dataset and automatically select the appropriate input features.

Embedded methods

Embedded methods "embed" the feature selection in the learning algorithm and use its properties to guide feature evaluation. Because the embedded methods avoid the repetitive execution of the learning algorithm and examination of every feature subset, they tend to be more efficient and computationally more tractable than wrapper methods while maintaining similar performance. Like wrapper, embedded methods take into account the dependencies among features, but at the expense of making the ML model dependent selections that might not work with any other ML model. They have lower risk to overfitting compared to wrapper methods, however their performance is hindered by the computational complexity, especially in high-dimensional data [HG15].

One of the examples of the embedded methods is the lasso component in the elastic net penalty. The authors of [HB15] used the elastic net penalty with the microarray gene expression data to distinguish four lung cancer subtypes. Their classification model embedded the gene selection step into a training step.

We can cite as well the Cox proportional hazards model with an elastic net penalty used with gene expression data. Indeed, in case if the Lasso-Cox penalty (cf. section 2.1.3), this model tends to select the most prognostic features. It was used by the authors of [Wan+18] to identify 23 groups of genes after 1000 iterations. As a result, they constructed 15 genes immune related risk signature.

3.1.2 Feature extraction

In feature selection, many algorithms apply correlation metrics to find which feature correlates most to the target. These algorithms single out features and do not consider the combined effect of two or more features with the target. In other words, some features might not have individual effect but alongside other features they give high correlation to the target and increase ML algorithm performance. In the case of feature extraction, a new set of features can be created as a combination of the initial features.

The basis for this technique is the manifold hypothesis stating that the data points exist on a lower dimensional sub-manifold or subspace. This subspace is referred to as feature space (i.e., feature extraction), embedded space (i.e., embedding), encoded space (i.e., encoding), subspace (i.e., subspace learning), lower dimensional space (i.e., dimen-

sionality reduction), submanifold (i.e., manifold learning), or representation space (i.e., representation learning) in the literature [Gho+19].

The feature extraction methods can be divided into two main categories, i.e., supervised and unsupervised methods. Supervised methods take into account the labels and classes of data samples while the unsupervised methods are based on the variation and pattern of data. Another categorization of feature extraction is dividing methods into linear and non-linear. The former assumes that the data falls on a linear subspace or classes of data can be distinguished linearly, while the latter supposes that the pattern of data is more complex and exists on a non-linear sub-manifold. Hereafter, we present the unsupervised feature extraction methods used with gene expression data for the purpose of further survival analysis.

Clustering

Early methods of machine learning applied to microarray data included simple clustering methods. For example, a widely used method was hierarchical clustering, due to the flexibility of the clustering methods they became very popular among the biologists [HG15]. However, hierarchical clustering imposes a strict tree structure on the data, is highly sensitive to the metric used to assess similarity, and typically requires subjective evaluation to define clusters.

Non-negative Matrix Factorization (NMF) clustering was proposed by [Bru+04] is a natural way to cluster genes and samples, because it involves factorization into matrices with nonnegative entries. The goal of NMF is to find a small number of metagenes, each defined as a positive linear combination of the P genes. The gene expression pattern of samples is then approximated as positive linear combinations of these metagenes. Mathematically, this corresponds to factoring matrix X into two matrices with positive entries, $X \approx WH$. Matrix W represents the metagenes and matrix H is the metagenes expression patterns of the samples. Each sample is placed into a cluster corresponding to the most highly expressed metagene in the sample. NMF metagenes can overlap and thus expose the participation of a single gene in multiple pathways or processes. We found that the authors of [Bel+11] used NMF clustering of the TCGA-OV transcriptomic data obtained 4 clusters to further analyse their possible prognostic ability.

Principal Component Analysis

Principal Component Analysis (PCA) is the most well-known dimensionality reduction algorithm. It is a linear unsupervised method which tries to find the orthogonal directions which represent the variation of data the best. If $u^T X$ is a projection of data onto direction u, then the variance of this projection is $u^T cov(X, X)u$. The desired directions (columns of matrix U) are the eigenvectors of the covariance matrix of data. Using the covariance matrix and its eigenvalues and eigenvectors, PCA finds the "principal components" in the data which are uncorrelated eigenvectors each representing some proportion of variance in the data. PCA was applied as a way of reducing the dimensionality of the data in cancer gene expression data by the studies [Tan+15; Zha+18; Cha+18; PCG18].

Single Value Decomposition

Single Value Decomposition (SVD) is an alternative way to eigenvalue decomposition in PCA. The algorithm called SALMON adopts this technique to calculate the eigengene matrix derived from co-expression network analysis as input to the learning algorithm. The eigengene matrix is the expression values of each gene co-expression module summarized into the first principal component using SVD. With the first right-singular vector of each module as the summarized expression values, it projects co-expressed genes to 1-D space and thus can be treated as the "super gene." In their experiment with breast invasive carcinoma, an eigengene matrix with 57 dimensions was derived from mRNA-seq data [Hua+19a].

Among the three linear dimenionality reduction techniques presented above, PCA provides a simple way to reduce dimensionality but requires that the matrices be orthogonal, which typically requires linear combination of components with arbitrary signs. NMF is more difficult algorithmically because of the nonnegativity requirement but provides a more intuitive decomposition of the gene expression data. At the end, when using the NMF, the metagene profiles are positive, sparse, localized, and relatively independent, which makes a natural compact decomposition for interpretation.

In contrast, spectral decomposition (PCA or SVD) of expression data produces eigengene profiles that are completely independent but complex, dense, and globally supported. Despite its promising features, NMF has the limitation of somewhat greater algorithmic complexity, especially compared with the simplicity of hierarchical clustering. The challenge that remains is to provide a meaningful biological interpretation to the NMF

discovered classes when the class labels and substructure of the data set are unknown [Bru+04].

Autoencoders

Autoencoders (AEs) are a variant of ANNs, the goal of AEs is to learn compact and efficient non-linear representations from input data. Compared with commonly used feature extraction approaches such as PCA, AEs extract features in the non-linear space. It is an unsupervised model that learn the generally lower dimension representation of the original features with the minimal loss of information. It is composed of an encoder and a decoder: one to transform the input into a latent, smaller dimension representation and the other one to reconstruct the representations into output.

Denoising Autoencoders (DAE) [Vin+08] improve upon the classic autoencoder by incorporating noise during training, a procedure which generates robust features. The training objective for DAEs is to build features that reconstruct initial input data from corrupted data, i.e. input data with random noise added. Authors of [Tan+15] have evaluated the ability of DAEs applied to transcriptomic breast cancer data to extract useful latent features. A series of papers applied as well the DAEs on the survival related genes filtered based on the univariate Cox PH analysis [PCG18; Cha+18; Zha+18].

The AEs can form multiple layers resulting in the Stacked Auto-Encoder (SAE). It is trained layer by layer and the final SAE is fine-tuned afterwards. [GSL19] have integrated gene expression and transcriptome alternative splicing profiles data to identify breast cancer subtypes. They have adopted the SAE neural network to learn lower dimension features in each data type and have integrated them into another hierarchical level to learn complex representations.

Authors of [Xia+18] present a semi-supervised deep learning strategy, the Stacked Sparse Auto-Encoder (SSAE) based classification, for cancer prediction using RNA-seq data. Datasets include three types of cancers, Lung Adenocarcinoma (LUAD), Stomach Adenocarcinoma (STAD) and Breast Invasive Carcinoma (BRCA) from the TCGA project.

Variational Auto-Encoders (VAEs) first proposed by [KW14] are another deep neural network approach generating latent representations for image and text. The traditional AEs are deterministic, in contrast, VAEs are stochastic and learn the distribution of explanatory features over samples. VAEs learn two distinct latent representations, a mean and a standard deviation vector, which are reparametrized into a single vector that can be

back-propagated (cf. Fig.3.1). VAEs harness, as other types of AEs, the modeling power of deep learning without the need for accurate labels, but they are generative models, which means they learn to approximate a data generating distribution. The work in [WG17; WG18] explores the possibility to determine if VAEs can be used on gene expression data and if they can capture biologically relevant features. Authors used the TCGA pan-cancer RNA-seq data (High Grade Serous Ovarian Carcinoma, HGSOC included) to identify the patterns in the VAE learnt features and discussed the potential merits of this approach. VAE can capture signals that are able to predict gene inactivation comparably to other algorithms of dimensionality reduction (PCA and NMF).

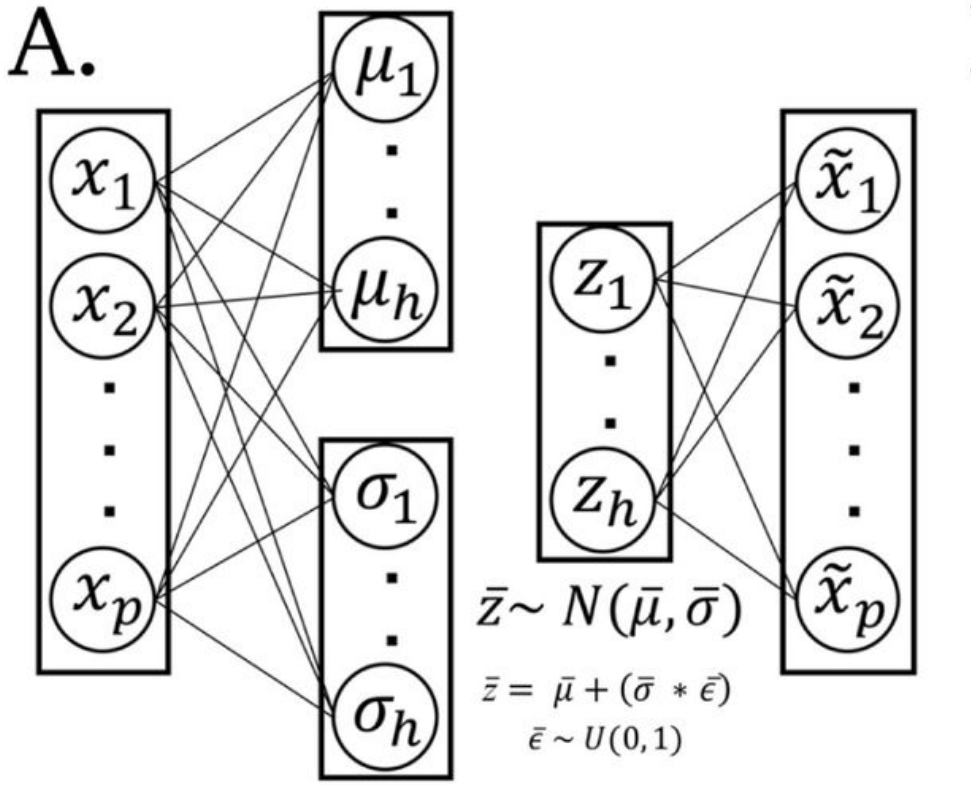


Figure 3.1: VAE schema, [WG18].

The Table 3.1 summarizes the different types of the AEs used with gene expression data for survival analysis found in the literature.

AE type	Tumour type	Input size	Hidden layers	Activation function	Loss function
DAE [Tan+15]	BRCA	2,520	100	Sigmoid	Cross-entropy
DAE [Cha+18]	HCC	15,629	500->100	Sigmoid	Logloss
DAE [PCG18]	BC	NS*	500->100	Tanh	Logloss
DAE [Zha+18]	Neuroblastoma	2,218	500->100	Tanh	Pseudo-Huber loss
VAE [WG17]	HGSOC	5,000	100	ReLU (encoder), Sigmoid (decoder)	Binary cross-entropy + KL divergence**

Table 3.1: Comparison of the AEs used with gene expression data for survival analysis. *Not specified, **Kullback–Leibler divergence

3.1.3 Prior knowledge integration

Adding prior knowledge reduces the complexity of the model and the number of parameters making analysis easier, it can be seen as a separate dimensionality reduction mechanism. Many sources of external biological information are available and can be integrated with machine learning and/or dimensionality reduction methods providing the advantage of biological connection with the output. Adding external information in gene expression data can give an insight on the functional annotation of the genes and the role they play in a disease, such as cancer.

Manual curation of the genes of interest appears as an obvious approach resembling feature filtering methods. The literature review is widely used to pre-select the genes, for example, the immune related genes as in [Kim+15]. The authors used the Immunology Database and Analysis Portal (ImmPort) [Bha+18] and pre-selected 1534 immune related genes as as candidates for signature construction.

Another example of manual genes filtering is the study of [Che+15]. It describes the usage of the Online Mendelian Inheritance in Man (OMIM) database, a comprehensive, authoritative compendium of human genes and genetic phenotypes that contain information on all known Mendelian disorders and over 12,000 genes. The authors selected the genes related to the Non-Small Cell Lung Cancer (NSCLC).

The main disadvantage of the manual gene curation is its difficulty to generalize to

other cancer types or to connect to the biological output other than pre-selected in advance.

Integrating prior knowledge can act as an embedded feature selection. Consider the usage of the biological pathways obtained from the Molecular Signatures Database (MSigDB) [Lib+15]. The authors of [Hao+19] selected the KEGG and Reactome databases pathways by excluding small pathways (i.e., less than 15 genes) and large pathways (i.e., over 300 genes), since small pathways are often redundant with other larger pathways, and large pathways are related to general biological pathways, rather than specific to a certain disease. Afterwards they investigated only the genes that were included in at least one of these pathways resulting in 5,404 genes and 659 pathways for 523 TCGA glioblastoma (TCGA-GBM) patients and 532 TCGA ovarian cancer (TCGA-OV) patients. The additional sparcity was implemented by a pathway layer in their ANN called Cox-PASNet. This pathway layer was not fully connected but only the genes in the pathway had connection to the pathway node in the pathway layer.

3.2 Survival analysis strategies

Several strategies can be observed in the recent works which deal with the gene expression data and try to detect the prognostic features within. The simpliest approach is a univariate Cox regression analysis, where each gene is tested for correlation with poor or good survival as in [Bel+11]. Having filtered the genes of interest from 489 HGSOC microarray data from the TCGA-OV project, the authors obtained a 193-gene transcriptional signature predictive of overall survival (Fig.3.2). An obvious shortcoming of this approach is the assumption that the input features are independent, which is not the case with the gene expression data.

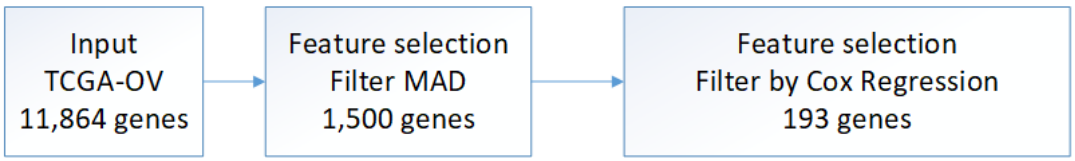


Figure 3.2: Survival analysis of TCGA-OV gene expression data by Bell et al. MAD here is Median Absolute Deviation, its purpose is the detection of intrinsically variable genes.

Another way to deal with survival analysis is to transform a survival prediction problem to a binary classification task. For example, in the study [Che+15], the authors split

the 280 lung cancer patients by the median survival of 40 months into good and poor survivors and used the ANN for classification into 2 groups based on the 10 pre-selected microarray probes (Fig.3.3). The authors of [Zha+16] followed the same strategy and divided 211 TCGA-GBM patients into short term survivors (less than 2 years survival) and long term survivors and used the algorithm called SimpleMKL (Multi Kernel Learning) for classification (Fig.3.4). The main obstacles in this kind of approaches is defining the threshold for dividing patients into groups and the censoring, indeed, the cohorts which present a large proportion of individuals lost to followup are not well suited for this type of analysis.

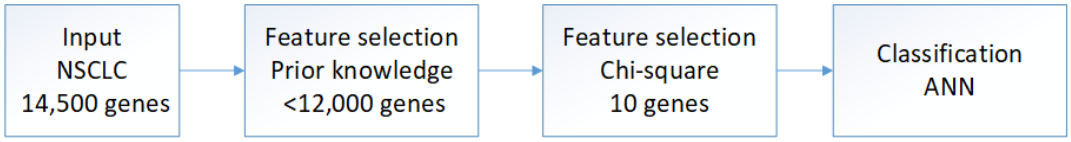


Figure 3.3: Survival analysis of NSCLC gene expression data by Chen et al. NSCLC is Non Small Cell Lung Carcinoma.

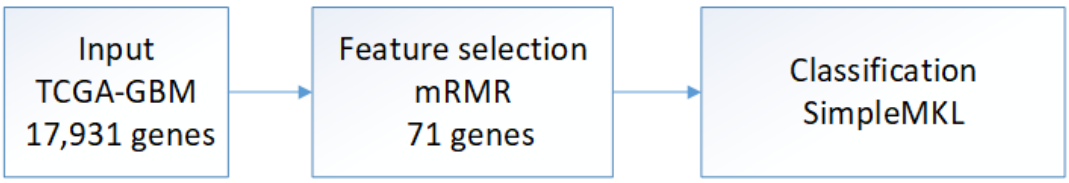


Figure 3.4: Survival analysis of TCGA-GBM gene expression data by Zhang et al.

The authors of [Kim+15] used genome, transcriptome, epigenome and proteome data to predict survival of the TCGA-BRCA patients. Their cohort was composed of 472 cases and in order to deal with the censored survival time as regression problem, the martingale residuals [TGF90] were calculated as a new continuous outcome:

$$M_i = \delta_i - H(t_i) \tag{3.3}$$

Since the martingale residuals have an exponential distribution between negative infinity and 1, the assumption of normally distributed residuals is not satisfied, thus a new fitness function based on mean absolute difference (MAD) between observed and predicted of martingale residuals was implemented and used in grammatical evolution neural network (GENN):

$$MAD = \frac{\sum_i^N |M_i - \hat{M}_i|}{\sum_i^N |M_i|} \tag{3.4}$$

Thus, the survival prediction problem is transformed into a regression task, the whole processed is shown in Fig.3.5.

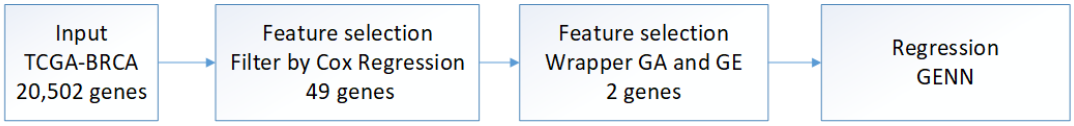


Figure 3.5: Survival analysis of TCGA-BRCA gene expression data by Kim et al.

Several studies tried to stratify the patients into the molecular subtypes without relying on survival during the process of defining subtypes. Instead, survival information was used post hoc to evaluate the clinical significance of these subtypes. The authors of [Bel+11] used the consensus NMF clustering and obtained 4 clusters, which they termed immunoreactive, differentiated, proliferative and mesenchymal on the basis of gene content in the clusters and previous observations, and they tested if the obtained clusters are related to survival afterwards (Fig.3.6). As a result, some molecular subtypes showed converging and similar survival profiles, making them redundant subtypes in terms of survival differences.

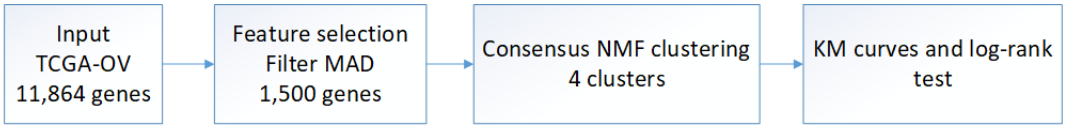


Figure 3.6: Clustering and survival analysis of TCGA-OV gene expression data by Bell et al.

Similar approach was adopted by [Tan+15] for the survival analysis of breast cancer patients from METABRIC database (2136 samples) and TCGA database (547 samples). Both datasets contained tumour and normal tissues RNA-seq data. After the DAE feature extraction step, because the distribution of activity values for each node is bimodal with one peak close to 0 and another close to 1, they separated patients into two groups based on their hidden node activity using a cutoff of 0.5. They assessed afterwards the differences of these groups for each node by Kaplan-Meier curves and the non-parametric log-rank test thus obtaining one node with the most prognostic power. To evaluate the

importance of this constructed feature, they further compared this feature with frequently used clinical markers of survival including tumor grade, molecular subtype and ER status (Fig.3.7).

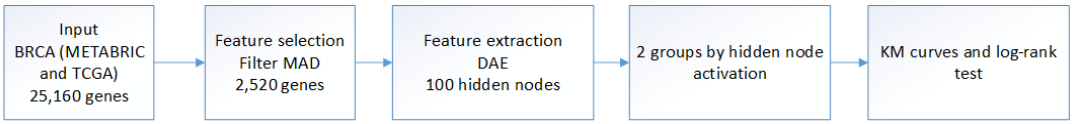


Figure 3.7: Survival analysis of METABRIC and TCGA-BRCA gene expression data by Tan et al.

Another effort in this direction is the study [WG17; WG18] in which the authors used the VAEs to extract latent features. They used the HGSOC subtypes definition of [Bel+11; Ver+12] for 490 samples to calculate the mean latent features per subtype. The prognostic difference in mesenchymal (poorer survival) and immunoreactive (better survival) histologic subtypes of HGSOC provides indirectly the survival analysis possiblity (Fig.3.8). Their results indicate as well that differential activation of glucuronidation is a strong signal distinguishing HGSOC subtypes. This observation may also help to explain increased survival in HGSOC patients with differentiated tumors.

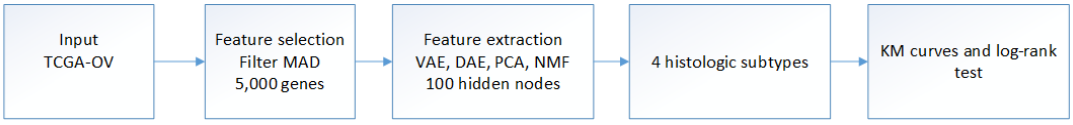


Figure 3.8: Survival analysis of TCGA-OV gene expression data by Way et Greene.

Recently, a series of 3 papers below tried to incorporate the survival information into the AEs based feature extraction. We present the description and schemas of their approaches. The authors of [Cha+18] downloaded the RNA-Seq, miRNAs and DNA methylation data of 360 TCGA hepatocellular carcinoma patients (HCC) as input features. They used the activity of the 100 nodes from the bottleneck hidden layer as new features and then conducted univariate Cox-PH regression analysis on each of the 100 features and identified 37 features significantly (log-rank $P < 0.05$) associated with survival. These 37 features were subjective to K-means clustering, with cluster number ranging from 2 to 6. Using silhouette index and the Calinski–Harabasz criterion, they found that 2 was the optimum number of clusters with the best scores for both metric. They built a supervised classification model using the SVM algorithm able to distinguish patients in 2 clusters (Fig.3.9).

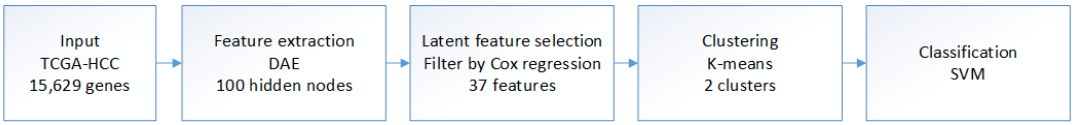


Figure 3.9: Survival analysis of TCGA-HCC gene expression data by Chaudhary et al.

Quite similar approach is adopted by [PCG18], they performed the survival analysis of the 402 bladder cancer (BC) patients based on the TCGA mRNA, miRNA and methylation data. The main difference is that they trained separate AEs for each type of omics data rather than combining all the input features together as in [Cha+18]. The outline of the processing the gene expression data is presented in Fig.3.10.

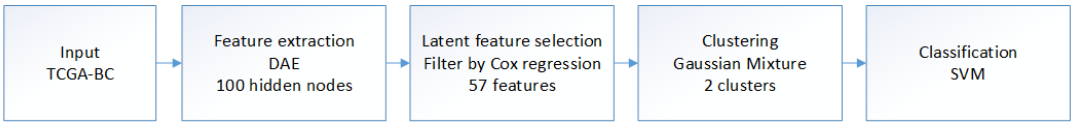


Figure 3.10: Survival analysis of TCGA-HCC gene expression data by Poirion et al.

The authors of [Zha+18] used the data of neurobalstoma patients from TARGET project (190 with gene expression and copy number alteration data) and added an extra feature selection step by filtering genes related to survival with univariate Cox regression analysis. Their results suggest that AE-based feature extraction step performs the best comparing to PCA (Fig.3.11).

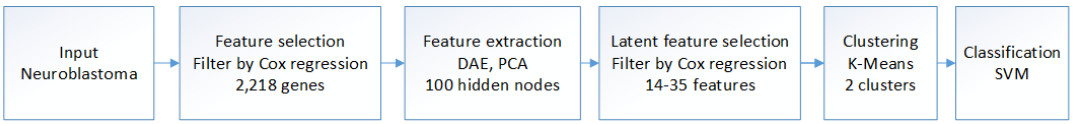


Figure 3.11: Survival analysis of neuroblastoma gene expression data by Zhang et al.

Most of the above cited approaches select genes based on a univariate summary statistics, such as p-value of log-rank test of survival groups defined by the univariate Cox PH regression analysis. As a result, these methods do not guarantee to select genes that each contribute non-redundant information, they are also difficult to generalize in order to account for additional variables, such as histological findings or patient characteristics. The elastic net was proposed by [HB15] as particularly well suited for survival analysis of the genome-scale data, which typically has many more features than observations.

The attempt to implement the multivariate Cox proportional hazards model with elastic net penalty was done by [Wan+18]. This study used the RNA-seq data of the the 285 TCGA-KIRP patients (Kidney renal papillary cell carcinoma), the authors pre-selected 1534 immune related genes from ImmPort database and, after the univariate Cox analysis, found 272 immune-related genes with predicting prognostic ability. They constructed the survival predictive model with regularized Cox model, Coxnet (the Cox proportional hazards model with an elastic net penalty [Sim+11]) and identified 15 genes stable model for construction of the immune-related risk signature (Fig.3.12).

Figure 3.12: Survival analysis of TCGA-KIRP gene expression data by Wang et al.

The main drawback of the regularized Cox model is that it is an additive model. Thus, it is difficult to capture non-linear interactions between genomic features, which might play important roles associated with survival. Deep Learning-based neural networks offer a potential solution for this problem because they are highly flexible and account for data complexity in a non-linear fashion. The advantages of learning nonlinear functions and retrieving lower dimensional representation at the same time was experimented by [You+17; CZG18]. The authors of [You+17] compared the performance of the models such as regularized Cox model (elastic net), RSF and Cox ANN (SurvivalNet) to predict the survival of the TCGA transcriptomic pan-glioma (LGG/GBM), breast (BRCA), and pan-kidney (KIPAN, consisting of chromophobe, clear cell, and papillary carcinomas) (Fig.3.13). In their experiments both Cox ANN and regularized Cox outperformed RSF models, and Cox ANN had a slight advantage over regularized Cox in LGG/GBM and KIPAN.

To note that the authors of [CZG18] used similar approach as in [You+17], a model based on the Cox ANN and compared the performance of the regularized Cox, RSF, CoxBoost and Cox ANN and revealed the advances of Deep Learning models (Cox-nnet). They analyzed 10 TCGA RNA-seq datasets with more than 300 samples: Bladder Urothelial Carcinoma (BLCA, 406 samples), Breast invasive carcinoma (BRCA, 1077 samples), Head and Neck squamous cell carcinoma (HNSC, 519 samples), Kidney renal clear cell carcinoma (KIRC, 531 samples), Brain Lower Grade Glioma (LGG, 512 samples), Liver hep-

atocellular carcinoma (LIHC, 358 samples), Lung adenocarcinoma (LUAD, 490 samples), Lung squamous cell carcinoma (LUSC, 487 samples), Ovarian serous cystadenocarcinoma (OV, 302 samples) and Stomach adenocarcinoma (STAD, 349 samples). Interestingly, the Cox ANN method worked better for some datasets (for example, KIRC) and worse for the others (OV).

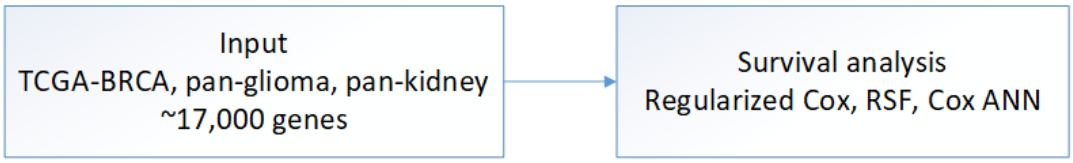


Figure 3.13: Survival analysis of TCGA pan-cancer gene expression data by Yousefi et al and Ching et al.

Two studies hereafter employed the dimensionality reduction techniques to reduce the dimension of the input transcriptomic data to improve the performance of the Cox ANNs. [Hua+19a] constructed a neural network with Cox partial log likelihood as loss function. The authors implemented Deep Learning-based networks to determine how gene expression data predicts Cox regression survival in breast cancer for 583 TCGA-BRCA patients. Rather than use raw gene expression values as model inputs, they calculated the eigengene modules from the result of gene co-expression network analysis acting as feature selection and feature extraction respectively, it greatly reduced the dimension of the original feature space. They called their algorithm SALMON for Survival Analysis Learning with Multi-Omics Neural Networks, the workflow is presented in Fig.3.14. To note that, between SALMON and the modified Cox-nnet the performance discrepancy is insignificant suggesting these two methods are comparable, the difference is that from the neural network structure perspective, SALMON enables a scalable integration of multi-omics data.

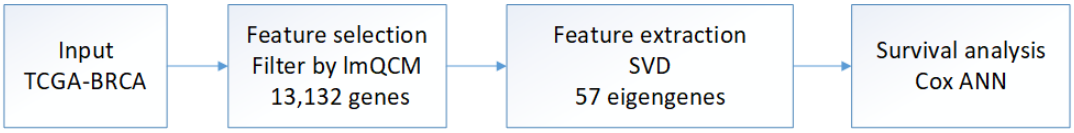


Figure 3.14: Survival analysis of TCGA-BRCA gene expression data by Huang et al.

The authors of [Hao+19] used the pathway databases (e.g., KEGG and Reactome) which contain a set of genes that are involved in a pathway, and each pathway characterizes

a biological process. The genes without pathway annotations were not considered in the analysis and the pathways were embedded into a Cox ANN based network (Cox-PASNet) by implementing the sparse connections between the gene layer and the pathway layer rather than fully-connected layers (Fig.3.15). According to the results reported for TCGA-GBM (523 patients) and TCGA-OV (532 patients) transcriptomic datasets, the authors obtained an imporved performance with their method Cox-PASNet in comparison to SurvivalNet and Cox-nnet.

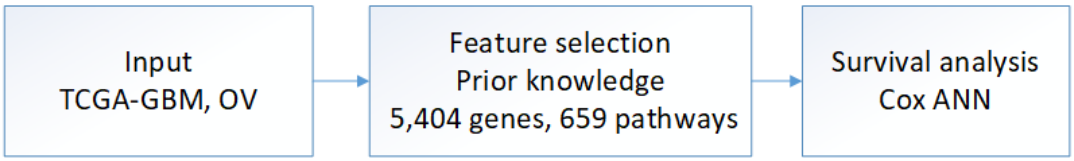


Figure 3.15: Survival analysis of TCGA-GBM and TCGA-OV gene expression data by Hao et al.

3.2.1 Validation aspects

The most present metric in evaluating the performance of the survival analysis with gene expression data was the KM curve with the log-rank test p-value (all the above cited studies except for Tan et al). C-index in its turn was reported by [You+17; CZG18; Cha+18; Zha+18; Hua+19a; Hao+19] and the Brier score was used as a metric only by Ching et al and Chaudhary et al.

The C-index and IBS comparison of the Cox-nnet, Cox PH, CoxBoost and RSF models trained with different TCGA datasets can be seen in Fig.3.16 and Fig.3.17. Cox-nnet clearly outperforms the other 3 models for the TCGA-OV dataset, but on the other hand, while having similar performance in terms of IBS, its C-index is still significantly lower that the TCGA-KIRC one, highlighting the need for more performant survival prediction models for ovarian cancer.

Another interesting aspect demonstrated by [HB15] is that training a classifier on multiple studies improved prediction when compared to training a classifier on only one study. Furthermore, data partitioning and robustness assessment was addressed by the majority of the reviewed studies: 5-fold CV in [Kim+15; Way+16; You+17; CZG18; PCG18; Cha+18; Hua+19a; Hao+19] or 10-fold in [Che+15; Tan+15; Zha+16], only [Bel+11] used the unique 50% Holdout validation. Surprisingly, none of the studies used the stratification by survival time and survival status in the training and test splitting procedure.

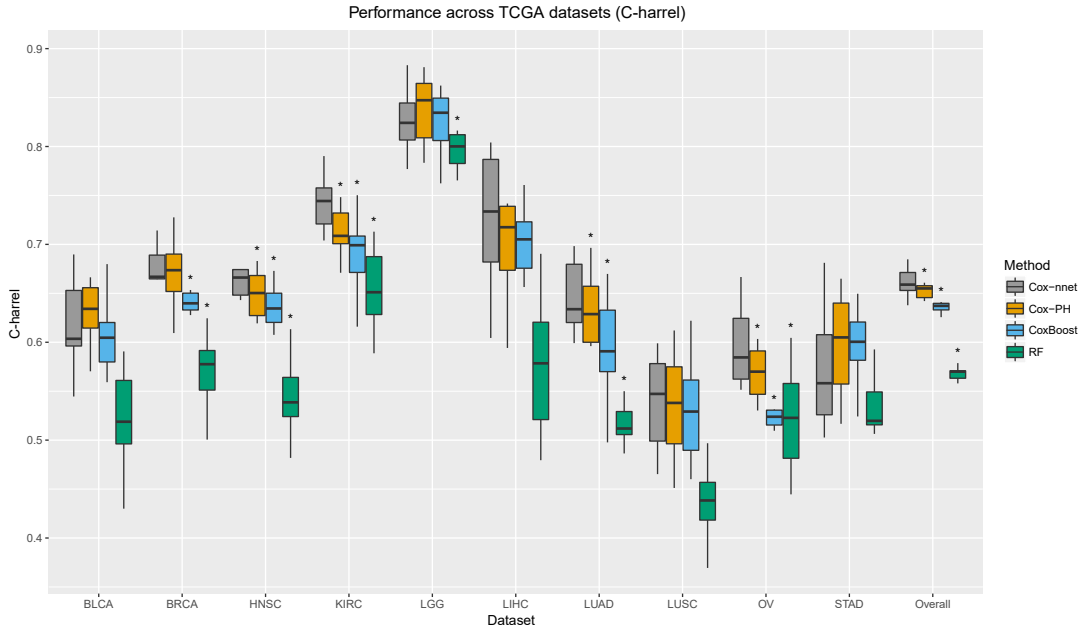


Figure 3.16: Boxplot of the C-index of the 10 TCGA datasets using the survival models Cox-nnet, Cox PH, CoxBoost and RSF, [CZG18].

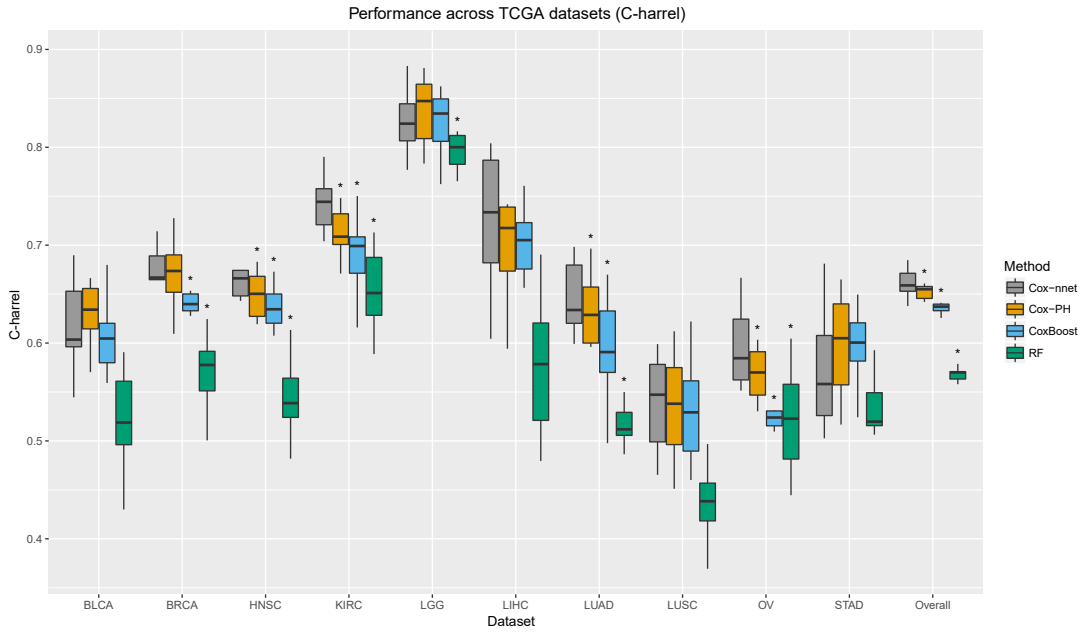


Figure 3.17: Boxplot of the IBS of the 10 TCGA datasets using the survival models Cox-nnet, Cox PH, CoxBoost and RSF, [CZG18].

However, a common problem in several works was the lack of external validation or testing regarding the predictive performance of their models. We note that the authors of [Cha+18] used 5 external independent datasets, [Bel+11] - 3 independent expression datasets, [Zha+18] - 2 independent datasets, [PCG18] - 1 dataset, [Zha+16] tested their model on the 22 new unseen TCGA patients, [Tan+15] used METABRIC for construction and TCGA for validation.

The importance of the survival models interpretation was adressed by the functional analysis provided in the reviewed papers. It included the following main categories:

- manual interpretation as in [Bel+11]
- Over-Representation Analysis (ORA) as in [Way+16]
- Gene Set Enrichment Analysis (GSEA) as in [Tan+15]

Conclusion Part I

The rapid development of high throughput sequencing technologies nowadays has produced huge amounts of cancer data that are collected and are available to the medical research community. By exploiting these data, the oncology research has made a great progress and has discovered separate subgroups within the same cancer type based on specific genetic defects that have different treatment approaches and options as well as different clinical outcomes. This is the foundation of the individualized treatment approach, in which computational techniques such as machine learning could help by identifying less costly and effectively such small groups of patients.

Ovarian cancer is a highly heterogeneous genetic disease and despite the advances already made, it still lacks successful treatment strategies. Because of the high risk of recurrence in high-grade serous ovarian carcinoma, the development of outcome predictors is important not only for patient stratification but to recognize categories of patients that are more likely to respond to particular therapies [Ver+12].

The genomic features measured at the transcriptome dimension are established to affect survival more directly than those measured at the genome or epigenome dimension [Kim+15]. Moreover, the clinical characteristics such as patient sex, age, stage and grade are now thought to be already encoded in the gene expression data [HB15; PCG18]. Thus, the survival analysis of ovarian cancer based on the gene expression data occurs to be a subject of great importance since it allows patient stratification.

In order to deal with high dimensional gene expression data and the task of survival prediction, many recent studies used the various feature selection techniques which may result in specific fluctuations concerning the creation of predictive feature lists. The filter algorithms single out features and do not consider the combined effect of two or more features with the target. The wrapper methods suffer from the size of the searched space and are relatively less used with gene expression data. The main disadvantage of the manual gene curation is its difficulty to generalize to other cancer types or to connect to the biological output other than pre-selected in advance. The database prior knowledge integration does not permit the new dependency discovery. As for the feature extraction, being a non-supervised approach, it does not integrate the survival information during

the process of defining new dimensions.

The embedded methods appear to overcome all the above cited shortcomings by discriminating the prognostic features at the same time as learning to predict the survival. The main limitation for these methods when used with traditional statistical survival models, such as penalized Cox regression, is its inability to capture non-linear interactions between genomic features, which might play important roles associated with survival.

ANN based survival networks offer a potential solution for this problem because they are highly flexible and account for data complexity in a non-linear fashion. The main drawback of this solution is its "black-box" nature, which could hamper model interpretability and further functional analysis to discover the input prognostic features and new therapy targets.

Part II

Contributions

Introduction part II

This part presents our contributions and gives the necessary explanations and supplemental material for better understanding the reasoning we followed as well as the dependencies between the different chapters. It is organized as follows:

1. Introduction part II
2. Chapter 1 "Comparative study"
3. Chapter 2 "Transfer learning experiments"
4. Chapter 3 "Proposed method N-MTLR-Rank"
5. Conclusion part II

A brief chapters description

Chapter 1 "Comparative study" is the conference paper [Men+21a], its primary objective was to present an overview of the recent neural network survival analysis techniques, apply them to the high-dimensional gene expression data and to compare their performance to predict outcome computed on high-grade serous ovarian carcinoma transcriptomic data from the TCGA project.

Indeed, the neural network based survival analysis models appeared to be an interesting research direction, and by the time of our first experiments, only one model, named Cox-nnet [CZG18], was evaluated on the TCGA ovarian cancer transcriptomic data and it outperformed other more traditional approaches. So, naturally, the idea of the comparative study of the existing neural network survival models came up. This work resulted in the best model designation: N-MTLR [Fot18]. To note that other reviewed and tested survival neural networks included DeepSurv [Kat+18], Cox CC [KBS19], Cox Time [KBS19], PC Hazard [KB19], Logistic Hazard [KB19] or Nnet-Survival [GN18], and PMF [KB19].

Chapter 2 "Transfer learning experiments" is another conference paper [Men+21b] which aimed to extend the transfer learning framework to “pan-gyn” cancers as these

gynecologic and breast cancers share a variety of characteristics being female hormone-driven cancers and could therefore share common mechanisms of progression.

We were inspired by the authors of [You+17] who used 2 additional datasets in order to augment the TCGA breast cancer dataset. Their goal was to check if this transfer learning technique could improve the performance of the proposed survival network called NetSurvival (quite similar to Cox-nnet [CZG18] and DeepSurv [Kat+18]).

Another interesting hypothesis came from the study [Ber+18] in which the authors refer to the following TCGA multi-cancer group as “pan-gyn”: high-grade serous ovarian cystadenocarcinoma (OV), uterine corpus endometrial carcinoma (UCEC), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), uterine carcinosarcoma (UCS), and invasive breast carcinoma (BRCA). They found molecular features that differed in the “pan-gyn” group and the TCGA non-gynecologic tumor types which let us hypothesize that training with this augmented "pan-gyn" group could benefit the ovarian cancer prognostication with neural networks. Interestingly, the original paper [You+17] did not report any significant improvement for the breast cancer outcome prediction, on the contrary, our study let to conclude that the whole "pan-gyn" group could be profitable for ovarian cancer survival prediction.

Lastly, Chapter 3 "Proposed method N-MTLR-Rank" is the paper which is currently in the process of submission. It exploits the results of Chapters 1 and 2. It proposes a new deep learning survival model which we called N-MTLR-Rank and trained using ovarian cancer clinical and molecular data from TCGA project. We used transfer learning to overcome over-fitting and generalization issues and we sought to validate our deep learning survival model on an independent clinical and molecular datatset. We illustrated as well the way our model can be interpreted, by calculating the contributions of the input features to the network outputs. We demonstrated how these contributions can be related to the molecular pathways to uncover biological processes associated with ovarian cancer patients survival.

Supplemental material

Folds generation

As mentioned in Part I, none of the reviewed gene expression based survival analysis studies used the stratification by survival time and survival status in the training and test

splitting procedure. For our tests, we split our TCGA-OV dataset into 5 folds using R package MTLR [Yu+11] thus constructing 5 different splits into training and test sets with respectively 80% and 20% of samples. The split was done using the stratification in order to have similar distributions of survival times and censoring in training and test sets. To compare the survival of training and test set splits, we plotted Kaplan-Meier curves and calculated the log-rank test p-value and concluded that the difference of survival between our generated training and test sets was not significant. In Fig.1, Fig.2, Fig.3, Fig.4 and Fig.5 we present the generated graphics which were not included in the chapters below.

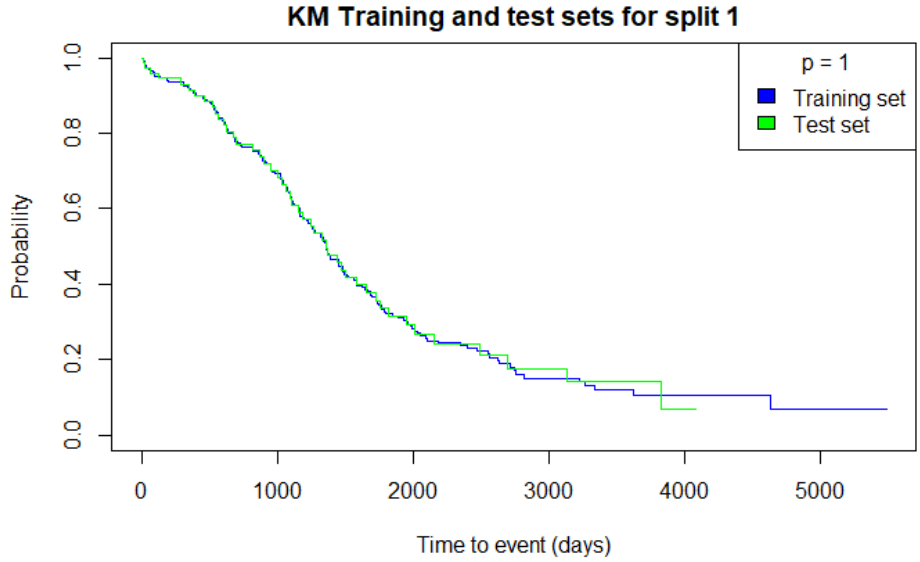


Figure 1: KM curves for training/test sets, fold 1.

Dimensionality reduction results

Another part of our results which was not included in the chapters below is the dimensionality reduction experimets.

In order to evaluate the impact of reducing data dimension, we employed a popular feature extraction method, the Principal Component Analysis (PCA). We kept 256 principal components for training the neural networks as they explain more than 97% of variability of the training sets.

Other dimensionality reduction techniques tested were the DAE as described in [Tan+15] and VAE used in [WG18] which aim to learn high-level latent features from the input data.

We used the best hyperparameters found in the original studies and tested 100, 256, 512 and 1024 hidden units. Extracted features were then used to train the survival neural network models.

The obtained C-index and IBS metrics for different dimensionality reduction techniques with survival neural networks are presented in Fig.6 and Fig.7. We did not observe any substantial gain of performance when using the feature extraction. Plus, for the sake of interpretability of the model, we made a choice of keeping the trained survival networks simple and left dimensionality reduction option aside.

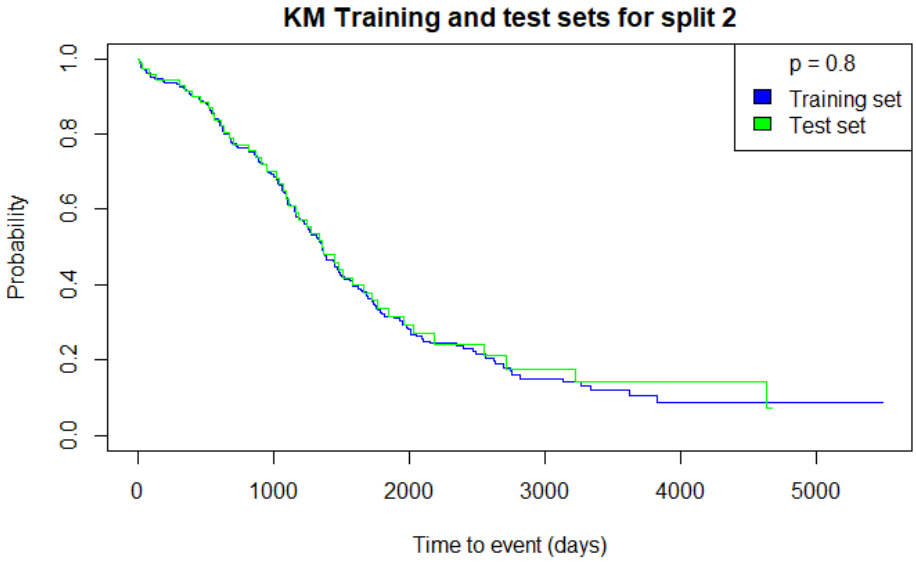


Figure 2: KM curves for training/test sets, fold 2.

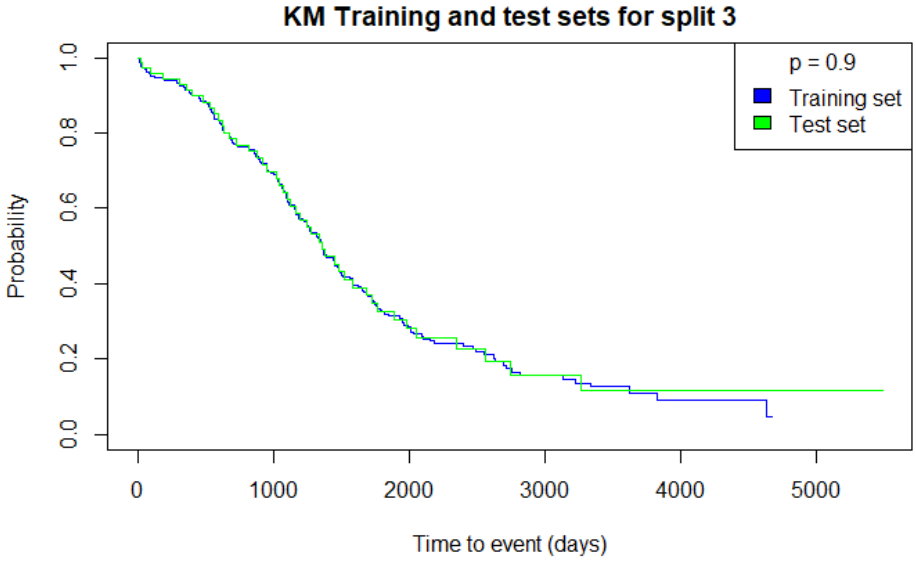


Figure 3: KM curves for training/test sets, fold 3.

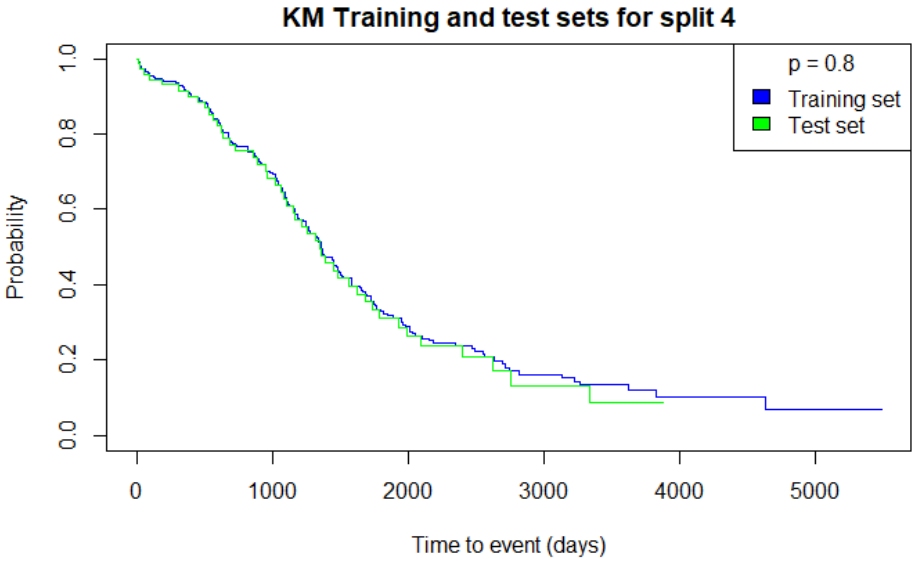


Figure 4: KM curves for training/test sets, fold 4.

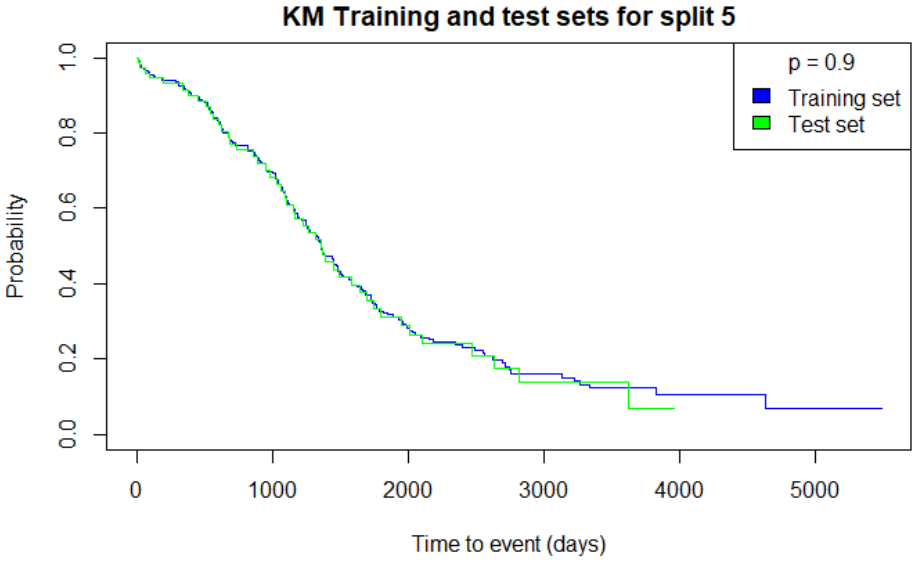


Figure 5: KM curves for training/test sets, fold 5.

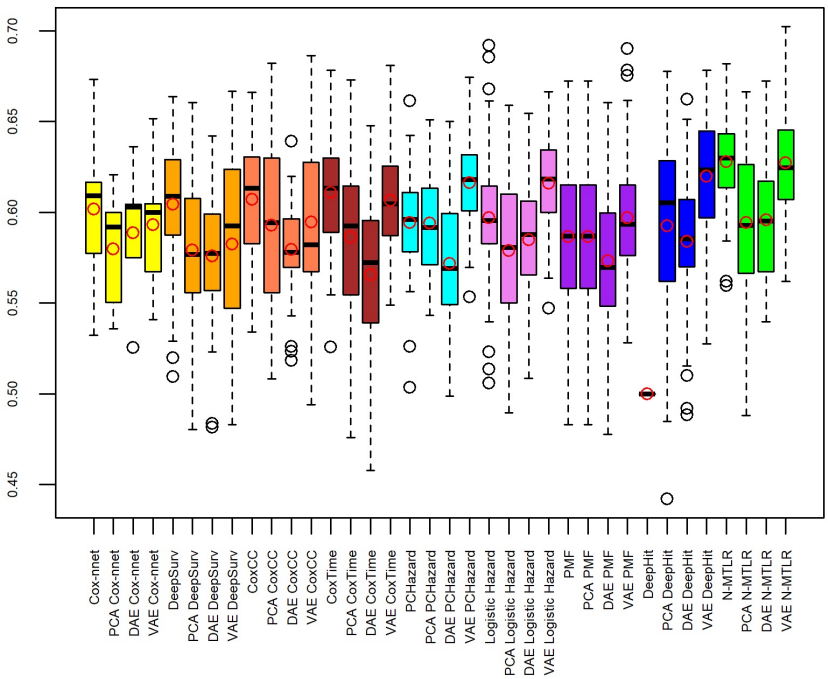


Figure 6: From left to right are the boxplots of the obtained 5-fold cross validation C-index for the dimensionality reduction experiments. The horizontal bars in the boxes represent the median values, the boundaries of the boxes delimite lower and upper quartiles, the values outside the boxes are the lowest and the highest observations and the red circles represent the mean values.

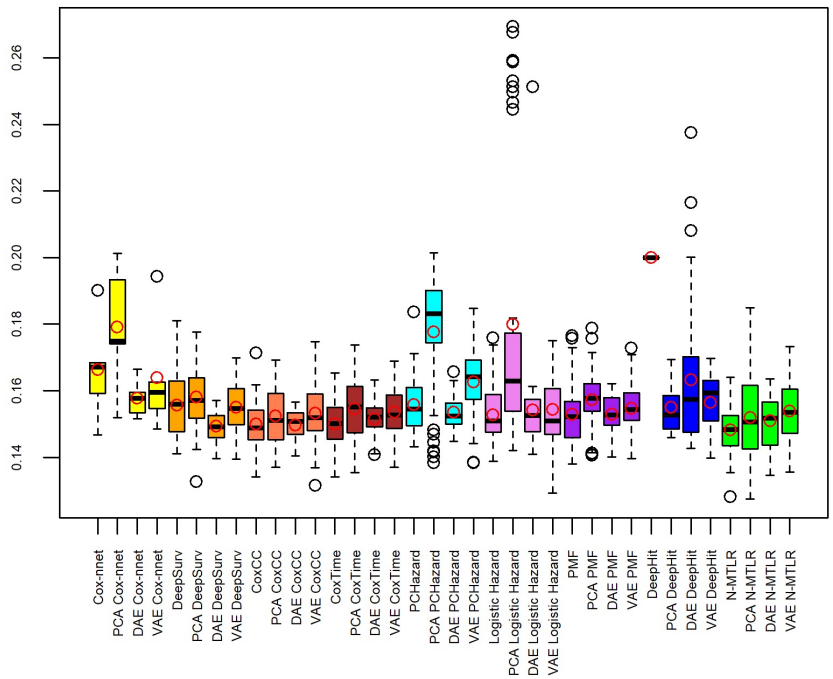


Figure 7: From left to right are the boxplots of the obtained 5-fold cross validation IBS for the dimensionality reduction experiments. The horizontal bars in the boxes represent the median values, the boundaries of the boxes delimite lower and upper quartiles, the values outside the boxes are the lowest and the highest observations and the red circles represent the mean values.

Chapter 1

Comparative study

Survival analysis of ovarian cancer is a subject of great importance since it allows patient stratification. The objective of this paper is to present an overview of the recent neural network survival analysis techniques, apply them to high-dimensional gene expression data and to compare their performance to predict outcome computed on high-grade serous ovarian carcinoma transcriptomic data. The Cancer Genome Atlas (TCGA) data were used to evaluate different methods. The obtained results were promising.

1.1 Introduction

Over the past few decades, with the high throughput sequencing technology development and the different machine learning techniques application, the oncology research has made a great progress based on genomic profiles. At the same time, while the high-dimensional data generated, such as RNA-seq, keep growing, a real need for appropriate machine learning techniques has appeared. These techniques should be able to effectively deal with mass data in order to make accurate medical decisions.

Ovarian cancer is a complex, heterogeneous genetic disease. Because of the high risk of recurrence in high-grade serous ovarian carcinoma (HGS-OvCa), the development of outcome predictors is important for patient stratification. In addition to predicting survival, the potential of prognostic classifiers lies in the ability to recognize categories of patients that are more likely to respond to particular therapies [Ver+12].

The lack of successful treatment strategies for ovarian cancer led The Cancer Genome Atlas (TCGA) researchers to analyze 489 cases of HGS-OvCa using copy number, expression and methylation arrays, and exonic sequencing data. Their work aimed to identify molecular abnormalities that influence pathophysiology, affect outcome and constitute therapeutic targets [Bel+11].

Gene expression profiles are considered to reflect the cancer progression driven by mutations and epigenetic modifications. The comprehension of these gene expression pat-

terns can serve to distinguish between normal and cancer tissue, classify cancer subtypes and stages. These profiles have been established to be associated with overall survival and the study [Bel+11] developed the prognostic signatures for ovarian cancer based on the TCGA microarray gene expression profiles using a univariate Cox regression analysis, and validated them on external datasets.

Recently, artificial neural networks (ANN) caught attention to solve problems with genomic profiles. Singh, Bapi, et Vinod [SBV18] used ANN to classify early and late stage of Papillary Renal Cell Carcinoma (PRCC). Chen et al [Che+15] described the use of ANNs to classify the patients with non-small cell lung cancer (NSCLC) for Adjuvant Chemotherapy (ACT) benefit strategy.

The objective of this work is first to make a survey of the existing ANN based techniques for survival analysis. Second, we seek to compare these techniques and evaluate them on the up-to-date harmonized (aligned to hg38) RNA-sequencing (RNA-seq) data from the TCGA-OV project in order to detect prognostic features. The outline of the paper will be as follows: section 1.2 presents the recent survival analysis techniques based on neural networks and the possibility to use them with gene expression data, section 1.3 describes materials, methods and results and section 1.4 discusses the future work.

1.2 Overview

1.2.1 Survival analysis

Survival analysis is a subfield of statistics modeling the data where the outcome is the time-to-event. One of the major difficulties in this context is censoring, i.e. the outcome is unobservable after a certain time period. Wang et al [WLR17] created a taxonomy of different approaches for survival analysis. They distinguish the traditional statistical and machine learning methods. One of the commonly used statistical method is a semi-parametric Cox regression or Cox proportional hazards. Each data instance is described by a triplet (X_i, t_i, δ_i), where $X_i = (x_{i1}, x_{i2}, ..., x_{iP})$ is the feature vector for instance i, t_i is the observed time, time of failure if δ_i is 1 or right-censoring if δ_i is 0. We note here the number of observations N and the number of features P. In this framework the rate of event at time t given that no event occurred before time t, i.e. the hazard function is:

$$h(t, X_i) = h_0(t)exp(X_i\beta) \tag{2.7}$$

where $h_0(t)$ is the baseline hazard function (an arbitrary nonnegative function of time), and $\beta^T = (\beta_1, \beta_2, ..., \beta_P)$ is the coefficient vector. The Cox model is a semi-parametric algorithm since the features are assumed to have an exponential influence on the outcome but the baseline hazard function, $h_0(t)$, is unspecified which makes it impossible to fit the model using standard likelihood function, instead the partial likelihood is used:

$$L(\beta) = \prod_{j=1}^{N} \left[\frac{exp(X_j\beta)}{\sum_{i \in R_j} exp(X_i\beta)} \right]^{\delta_j} \tag{2.9}$$

where R_j is the set of indices, i, with $y_i \geq t_j$ (those at risk at time t_j). The coefficient vector is estimated by maximizing this partial likelihood, or equivalently, minimizing the negative log-partial likelihood for improving efficiency:

$$LL(\beta) = -\sum_{j=1}^{N} \delta_j \left\{ X_j\beta - log \left[\sum_{i \in R_j} exp(X_i\beta) \right] \right\} \tag{2.10}$$

The extension of Cox regression with artificial neural networks was first proposed by Faraggi and Simon [FS95], who replaced the linear predictor of the Cox regression model, by a one hidden layer multilayer perceptron (MLP). This work was further explored by Ching et al [CZG18] (Cox-nnet) and Katzman et al [Kat+18] (DeepSurv) who proposed to incorporate the advances of deep learning framework and demonstrated that their methods outperform the classical Cox method. The linear predictors in their models become:

$$\theta_i = G(WX_i + b)^T\beta \tag{1.1}$$

where W is the coefficient weight matrix between the input and hidden layer of size H x P, H is the number of neurons in the hidden layer, b is the bias vector of size H and G is the activation function. In the paper [CZG18] the partial log likelihood is written as:

$$PL(\beta, W) = \sum_{\delta_j=1} \left\{ \theta_j - log \left[\sum_{i \in R_j} exp(\theta_i) \right] \right\} \tag{1.2}$$

And the ridge regularization term with the partial log likelihood gives the following cost function:

$$cost(\beta, W) = PL(\beta, W) + \lambda(||W||_2 + ||\beta||_2) \tag{1.3}$$

where $||.||_2$ designates L_2-norm penalty function and λ is a regularization coefficient.

Ching et al [CZG18] experimented only hyperbolic tangent (tanh) activation function, whereas Katzman et al [Kat+18] proposed rectified linear unit (ReLU) [NH10] and Scaled Exponential Linear Unit (SELU) [Kla+17] activation functions.

More recently, [KBS19] have proposed 2 new extensions to Cox model with neural nets, namely Cox Case Control (Cox CC) and Cox Time. Cox CC is based on the nested case control methodology and uses the simplified partial log-likelihood as the loss function. For neural net implementation the loss is:

$$loss = \frac{1}{N} \sum_{i:\delta_i=1} log\left(1 + exp\left[\theta(X_j) - \theta(X_i)\right]\right), j \in R_i \setminus \{i\} \tag{1.4}$$

where N is the number of events in the dataset, j is one sampled individual from the risk set R_i.

The authors of [KBS19] demonstrated that the loss 1.4 is a good approximation of the Cox partial log likelihood.

Cox Time is a non-proportional hazards extension, it integrates time as a parameter into a loss function in order to overcome the proportionality assumption constraint of the Cox model. The loss function for this model with neural net implementation is:

$$loss = \frac{1}{N} \sum_{i:\delta_i=1} log\left(\sum_{j \in \tilde{R}_i} \left[\theta(t_i, X_j) - \theta(t_i, X_i)\right]\right) \tag{1.5}$$

where $\tilde{R}_i$ is a subset of the patients at risk R_i.

The piecewise constant (PC-Hazard) method proposed by [KB19] recently is a continuous-time method as well, it parametrizes the hazard in the loss function, but requires defined intervals in which the hazard is constant:

$$loss = -\frac{1}{N} \sum_{i=1}^{N} \left(\delta_i log\tilde{\eta}_{k(t_i)}(X_i) - \tilde{\eta}_{k(t_i)}(X_i)p(t_i) - \sum_{j=1}^{k(t_i)-1} \tilde{\eta}_j(X_i)\right) \tag{1.6}$$

where time intervals are modeled as $0 = \tau_0 < \tau_1 < \tau_2 < ...\tau_m$, m is the number of intervals, $k(t_i)$ denotes the duration index of individual i's event time, i.e. $t_i = \tau_k(t_i)$, $\tilde{\eta}_j = \eta_j \Delta\tau_k$, $\Delta\tau_k = \tau_k - \tau_{k-1}$, $p(t) = \frac{t - \tau_{k(t)-1}}{\Delta\tau_{k(t)}}$

$$h(t) = \eta_{k(t)} \tag{1.7}$$

for set of non-negative constants $\eta_1, ..., \eta_m$.

The above cited methods work with continuous time. There are other models con-

structed for the discrete time. The authors of [KB19] explore the possibility to create time intervals for the continuous time observations and compare hazard rate or probability mass function (PMF) parametrization of the loss function.

According to [KB19], Lee et al. [LZY18] were the 1st to apply neural networks to the discrete-time likelihood for right-censored time-to-event data. The proposed method, denoted DeepHit, estimates the probability mass function with a neural net and combines the log-likelihood with a ranking loss. The authors of [KB19] used essentially the same negative log-likelihood but for one type of event to implement the method they called PMF:

$$loss = -\frac{1}{N}\sum_{i=1}^{N}\left(\delta_i log\left[\sigma_{k(t_i)}(\phi(X_i))\right] - (1-\delta_i) log\left[\sum_{k=k(t_i)+1}^{m+1}\sigma_k(\phi(X))\right]\right) \tag{1.8}$$

where $\sigma_j\left(\phi(X)\right) = \frac{exp[\phi_j(X)]}{1+\sum_{k=1}^{m} exp[\phi_k(X)]}$ is the softmax function and $\phi(X)$ is the neural network.

Yu et al. [Yu+11] proposed the multi-task logistic regression (MTLR), which is a generalization of the binomial log-likelihood, to jointly model a sequence of binary labels representing event indicators. Fotso et al. [Fot18] later applied this framework to neural networks and called their method neural multi-task logistic regression (N-MTLR). As shown by [KB19] N-MTLR is equivalent to PMF method in 1.8 but where $\phi_j(X)$ is the (reverse) cumulative sum of the output of the network $\psi(X)$.

In statistical survival analysis, it is, however, more common to express the likelihood by the discrete-time hazard rate. Gensheimer and Narasimhan [GN18] used this form of the likelihood and parameterized the hazard rates with a neural network:

$$loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{k(t_i)}\left(y_{ij} log\left[h(\tau_j|X_i)\right] + (1-y_{ij}) log\left[1-h(\tau_j|X_i)\right]\right) \tag{1.9}$$

where $y_{ij} = \mathbb{1}\left\{\tau_j = t_i, \delta_i = 1\right\}$ and the parametrized discrete hazard rate:

$$h(\tau_j|X_i) = \frac{1}{1+exp(-\phi_j(X))} \tag{1.10}$$

For a better readability the comparison of the main characteristics of the above cited loss functions is given in the Table 1.1.

Characteristics				
Technique	*Time*	*Parametrization*	*PH**	*Particularity*
Cox-nnet, DeepSurv	Continuous	Hazard function	Yes	Cox partial log-likelihood used
Cox CC	Continuous	Hazard function	Yes	Simplified loss function based on the nested case control
Cox Time	Continuous	Hazard function	No	Time dependent Cox CC loss function
PC Hazard	Continuous	Hazard function	No	Hazard function is constant in predefined intervals
Logistic Hazard	Discrete	Hazard function	No	Discrete partial log-likelihood used
PMF	Discrete	PMF	No	Equivalent to DeepHit with one type of event
N-MTLR	Discrete	PMF	No	Reverse cumulative sum of the output of the network is used

Table 1.1: Comparison of ANN based survival analysis loss functions. *Proportional hazards assumption.

1.2.2 Gene expression and survival analysis

In the field of medical research there are different definitions of survival [Liu+18]. Liu et al underlined the necessity to have sufficient observation period in order to catch enough events and thus to provide enough power for statistical tests. They analyzed all TCGA clinical data and stated that Overall Survival (OS) and Progression Free Interval (PFI) could be relatively accurately calculated from available data. They have also derived Disease Free Interval (DFI) and have judged it reasonably accurate. As for Disease Specific Survival (DSS), they have concluded that it could only be estimated for most cases. Even if the primary goal of the TCGA program was not the survival analysis, the study [Liu+18] demonstrated that the survival plots are similar to previous independent studies for most cancer types. One of the best examples is TCGA outcomes for Ovarian Cancer (OV).

The authors of [HB15] proposed to use the elastic net for a meta-analysis of lung cancer gene expression. Their primary purpose was to distinguish between lung cancer subtypes but the approach can also be applied to survival analysis by using the regularized Cox model.

The authors of [CZG18] used the high-throughput transcriptomic data of different cancer types from TCGA and compared survival methods such as regularized Cox model,

Random Survival Forests, CoxBoost and the proposed Cox-nnet method. They discussed as well that it is possible to use the weights of the hidden units of the trained neural network in order to interpret biological meaning of the received results. Their suggested approach Cox-nnet gave satisfactory results for some cancer types, especially for TCGA Kidney Renal Cell Carcinoma (KIRC), and insufficient results for other types, for example, in the OV dataset. These two datasets are comparable in terms of data available, regularized Cox model and Cox-nnet results are much worse in the TCGA-OV dataset than in TCGA-KIRC. This confirms the need of new machine learning approaches for survival analysis in ovarian cancer. To our knowledge, other survival analysis methods based on neural networks presented in this paper were not tested on the high-dimensional transcriptomic data.

1.3 Materials and Methods

1.3.1 Gene expression and clinical data

TCGA RNA-sequencing data and clinical data were downloaded from Genomics Data Commons (GDC) portal (https://portal.gdc.cancer.gov/) using the pipeline of the R/ Bioconductor package TCGAbiolinks [Col+16]. The harmonized RNA-seq data (HTSeq-counts) were normalized using the existing TCGAbiolinks normalization function which is recommended for differential expression analysis.

Supplemental survival data were downloaded from the standardized dataset named the TCGA Pan-Cancer Clinical Data Resource (TCGA-CDR) [Liu+18]. We merged the OV survival data from TCGA-CDR with the GDC clinical data. We made a choice to perform our tests on OS endpoint. The corresponding TCGA-CDR columns included OS for status and OS.time for time-to-event data. OS column contained the value 0 encoding for alive (censored) status and 1 for deceased (failure) and OS.time contained numbers of days from the date of diagnosis to either the date of last follow up if OS was 0 or time to death if OS was 1.

A total number of 379 RNA-seq samples were obtained for OV (TCGA-OV project), 5 of which were recurrent tumors and 374 primary tumor samples. After merging RNA-seq and clinical data, we obtained 374 cases among which we discarded 2 cases without survival data. As a result, our complete dataset included normalized expression with $P = 17401$ genes and $N = 372$ samples.

1.3.2 Evaluation criteria

The widely used in survival analysis Concordance-index (C-index) measures the concordance between predicted risk score and observed survival time. This measure is computed for all comparable pairs in the test set and the number of times the predictions are concordant is summarized.

As the C-index only depends on the ordering of the predictions, it is very useful for evaluating proportional hazards models. Another metric for the time-dependent methods is C-index by Antolini et al. [ABB05], which estimates the probability that observations i and j are concordant given that they are comparable. It was modified by [KBS19] to account for tied event times and survival estimates, we will refer to it as adjusted Antolini C-index:

$$C_{td} = P\left\{\hat{S}(t_i|X_i) < \hat{S}(t_i|X_j)|t_i < t_j, \delta_i = 1\right\} \tag{2.13}$$

where $\hat{S}(t)$ is the estimated survival function $\hat{S}(t) = exp\left(-\int_0^t h(s)ds\right)$.

The C-index value of 0.5 is equivalent to random guess and 1 is the perfect concordance, so higher C-index means better model performance. To note that this metric has a close relationship to classification accuracy and AUC [Ish+08] and, for the proportional hazards models, it is equivalent to the regular C-index.

Kvamme et al. [KB19] showed that, by only considering concordance, DeepHit, for example, has excellent discriminating performance at the cost of poorly calibrated survival estimates. In this perspective, it is important to calculate another metric, the Brier score (BS), the mean squared error of the probability estimates. In order to calculate it, we get the binary outcomes from time-to-event data, choose a fixed time t and label data according to whether or not an individual's event time is shorter or longer than t. This score can be generalized to account for censoring by weighting the scores by the inverse censoring distribution:

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\hat{S}(t|X_i)^2 \mathbb{1}\left\{t_i \leq t, \delta_i = 1\right\}}{\hat{G}(t_i)} + \frac{(1-\hat{S}(t|X_i))^2 \mathbb{1}\left\{t_i > t\right\}}{\hat{G}(t)}\right] \tag{2.14}$$

where $\hat{G}(t) = P\left\{t_i > t, \delta_i = 0\right\}$ is the Kaplan-Meier estimate of the censoring survival function.

The BS can be extended from a single duration t to an interval by computing the

integrated Brier score (IBS) [KBS19], for this metric, smaller values signify better performance:

$$IBS = \frac{1}{Max(t_i)} \int_0^{Max(t_i)} BS(t)dt \tag{2.15}$$

1.3.3 Algorithms and implementation

For our tests, we (log2+1) transformed the normalized values and split our dataset into 5 folds using R package MTLR [Yu+11] thus constructing 5 different splits into training and test sets with respectively 80% and 20% of samples. The split was done using the stratification by the OS.time and OS features in order to have similar distributions of survival times and censoring in training and test sets. To compare the survival of training and test set splits, we plotted Kaplan-Meier curves and calculated the log-rank test p-value and concluded that the difference of survival between our generated training and test sets was not significant.

In order to facilitate the training procedure, the training data were standardized to zero-mean and unit-variance to comply with best practices for training deep learning algorithms.

We used the Cox-nnet implementation by [CZG18] but corrected the C-index computation to account for tied events. The methods DeepSurv, Cox CC, Cox Time, PC Hazard, Logistic Hazard or Nnet-Survival, PMF and N-MTLR were tested with help of pycox Python package [KB19].

For discrete time methods, the quantile discretization scheme and constant density interpolation were used as recommended in [KB19] for smaller datasets.

Hyperparameter search was performed with C-index evaluation criterion for Cox-nnet and adjusted Antolini C-index for all the other methods. The hyperparameter search space is given in the Table 1.2. We did not use dropout [Sri+14] for Cox-nnet as the authors did not report any benefit from this technique for Cox-nnet. Only 1 hidden layer was used in the tested neural net architectures for all the methods.

Each training set was further split into 5 different combinations of optimization and validation datasets following the same stratification strategy as for training/test split. The hyperparameter grid search was done by training the optimization datasets separately; for each training set the best hyperparameter combination was selected on the highest mean validation test C-index, i.e. nested 5-fold cross-validation for hyperparameter optimization.

Hyperparameter	Values
Layers	1
Nodes per layer	32, 64, 132, 256, 512, 1024
Dropout (except Cox-nnet)	0., 0.1
Cox-nnet L2	-3., -2.67, -2.34, -2.01, -1.68, -1.35, -1.02, -0.69, -0.36, -0.03, 0.3, 0.63, 0.96, 1.29, 1.62
Weight decay (except Cox-nnet)	0.1, 0.01, 0.
Batch size (except Cox-nnet)	128, 256
L1 (CoxCC and CoxTime)	0.1, 0.01, 0.001, 0.
Num. durations (discrete time methods)	2, 5, 10, 20

Table 1.2: Hyperparameter search space.

Afterwards, each best model was trained on the training dataset and test datasets evaluation metrics are reported resulting in a 5-fold cross validation on test datasets. The IBS was computed over 100 equidistant points between the minimum and maximum observed times in the test sets.

For the DeepSurv, Cox CC, Cox Time, PC Hazard, Logistic Hazard or Nnet-Survival, PMF, and N-MTLR methods the best hyperparameter configuration on each fold was fitted 10 times, and we report all the C-index and IBS values of the 10 repetitions.

1.3.4 Results

Results of the algorithms, each applied to TCGA-OV transcriptomic data are reported in figure 1.1 for C-index and in figure 1.2 for IBS.

Cox-nnet for ovarian cancer was reported by [CZG18] and it serves us the baseline for comparison, we evaluate this technique on the up-to-date TCGA gene expression data for ovarian cancer (302 samples versus 372 in our study). Our corrected version of C-index computation was used, we demonstrated that it affects the training procedure and the overall accuracy evaluation and thus gives better results on our dataset. Even if the authors of [CZG18] did not report any improvements for dropout with tanh activation function, we notice that DeepSurv method with ReLU activation function [NH10] and dropout [Sri+14] works slightly better than Cox-nnet; it is more evident on IBS metric. The performance of DeepSurv, in its turn, is comparable to Cox CC method which proves that the proposed simplified loss function 1.4 is a good approximation of the classical Cox partial log-likelihood and Cox CC even outperforms DeepSurv in terms of IBS. We

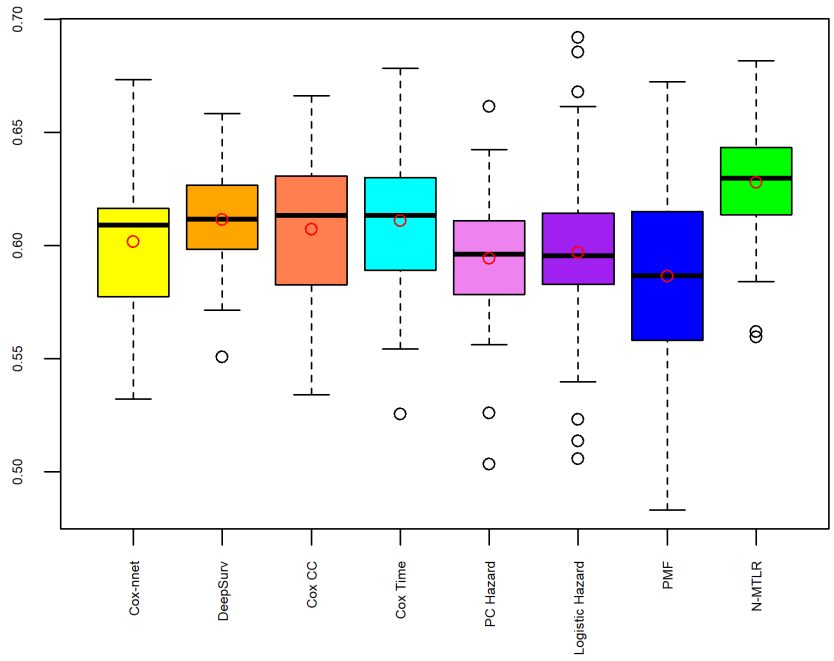


Figure 1.1: C-index comparison of different deep survival models. From left to right are the boxplots of the obtained 5-fold cross validation C-index on the test datasets. Higher C-index means better model performance. The horizontal bars in the boxes represent the median values, the boundaries of the boxes delimit lower and upper quartiles, the values outside the boxes are the lowest and the highest observations and the red circles represent the mean values.

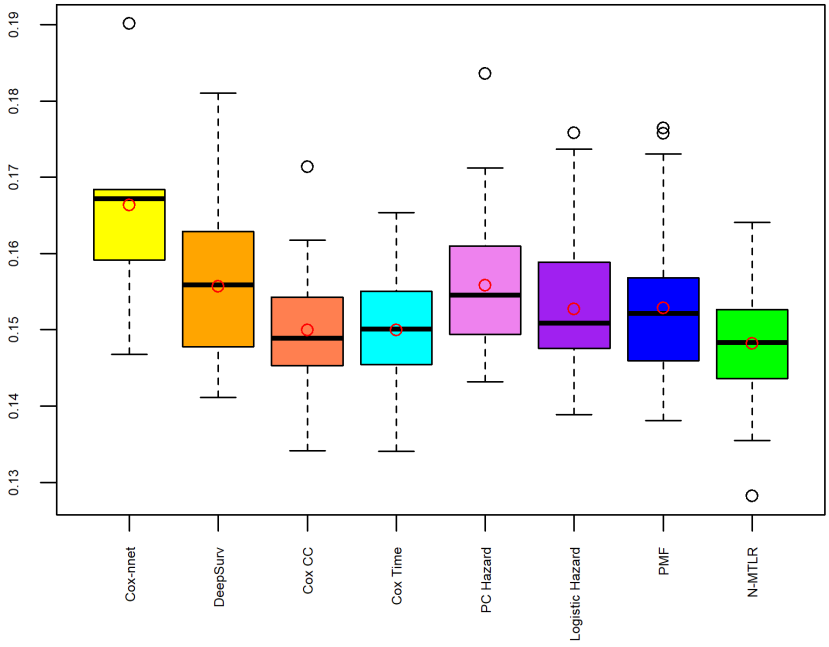


Figure 1.2: IBS comparison of different survival models. From left to right are the boxplots of the obtained 5-fold cross validation IBS on the test datasets. For this metric, smaller values signify better performance. The horizontal bars in the boxes represent the median values, the boundaries of the boxes delimit lower and upper quartiles, the values outside the boxes are the lowest and the highest observations and the red circles represent the mean values.

note as well that Cox Time, which overcomes proportional hazards assumption, gives the best results for continuous time methods, PC Hazard not resulting in any substantial improvements over Cox Time. As for the discrete time methods, the authors of [KB19] underlined that hazard parametrization works better than PMF parametrization. We do not observe any important difference for PMF and Logistic Hazard methods, the only noticeable advantage of the hazard parametrization being robustness on the C-index criterion. The overall best performance was obtained with N-MTLR for C-index and IBS as well as for 5-fold cross-validation robustness, even if this method is in fact a PMF parametrization as shown by [KB19].

1.4 Conclusion

In this paper, we have discussed ANN based survival analysis techniques adaptable to deal with the high-dimensional gene expression data. We have explored the survival methodology built on neural networks for continuous and discrete time data, in particular Cox-nnet, DeepSurv, Cox CC, Cox Time, PC-Hazard, Logistic Hazard or Nnet-Survival, PMF, N-MTLR. Since neither traditional regularized Cox model nor Cox-nnet produced satisfactory results in transcriptomic TCGA-OV dataset, it was important to search for other survival analysis techniques capable to deal with ovarian RNA-seq data. According to our experiments, the N-MTLR model appears as the most effective and promising one outperforming all the other ANN based techniques found in literature.

As a future work, we plan to integrate multiple data types (other omics data, whole-slide images, etc.) to construct performant models for survival prediction based on the N-MTLR model. There is as well a strong need to interpret the obtained results and link them to the information with the biological meaning to be applicable in clinical decision-making.

Chapter 2

Transfer learning experiments

With the advent of high-throughput sequencing technologies, the genomic platforms generate a vast amount of high dimensional genomic profiles. One of the fundamental challenges of genomic medicine is the accurate prediction of clinical outcomes from these data. Gene expression profiles are established to be associated with overall survival in cancer patients, and this perspective the univariate Cox regression analysis was widely used as primary approach to develop the outcome predictors from high dimensional transcriptomic data for ovarian cancer patient stratification.

Recently, the classical Cox proportional hazards model was adapted to the artificial neural network implementation and was tested with The Cancer Genome Atlas (TCGA) ovarian cancer transcriptomic data but did not result in satisfactory improvement, possibly due to the lack of datasets of sufficient size. Nevertheless, this methodology still outperforms more traditional approaches, like regularized Cox model, moreover, deep survival models could successfully transfer information across diseases to improve prognostic accuracy. We aim to extend the transfer learning framework to "pan-gyn" cancers as these gynecologic and breast cancers share a variety of characteristics being female hormone-driven cancers and could therefore share common mechanisms of progression.

Our first results using transfer learning show that deep survival models could benefit from training with multi-cancer datasets in the high-dimensional transcriptomic profiles.

2.1 Introduction

The recent development of high-throughput sequencing technology and machine learning methodology resulted in a great progress in the field of oncology research based on genomic profiles. However, while the high-dimensional data generated, such as RNA-seq, keep growing, a real need for appropriate machine learning techniques, capable of dealing with mass data, has appeared.

Ovarian cancer is a complex, heterogeneous genetic disease. Because of the high risk

of recurrence in high-grade serous ovarian carcinoma (HGS-OvCa), the development of outcome predictors is important not only for patient stratification but also to recognize categories of patients that are more likely to respond to particular therapies [Ver+12]. The lack of successful treatment strategies for ovarian cancer led The Cancer Genome Atlas (TCGA) researchers to gather the HGS-OvCa genomic profiles in order to identify molecular abnormalities that influence pathophysiology, affect outcome and constitute therapeutic targets [Bel+11].

Gene expression profiles are considered to reflect the cancer progression driven by mutations and epigenetic modifications. These profiles were established to be associated with overall survival and the study [Bel+11] developed the prognostic signatures for ovarian cancer based on the TCGA microarray gene expression profiles using a univariate Cox regression analysis, and validated them on external datasets.

Recently, artificial neural networks (ANN) caught attention to solve problems with genomic profiles. [You+17; CZG18] used ANN to construct survival models using the TCGA gene expression data. The authors of [CZG18] used the high-throughput transcriptomic data of the different TCGA cancer types and compared survival methods such as regularized Cox model, Random Survival Forests, CoxBoost and the proposed Cox-nnet method. Their approach Cox-nnet gave satisfactory results for some cancer types, especially for TCGA Kidney Renal Cell Carcinoma (KIRC), and insufficient results for other types, for example, in the OV dataset. The study of [You+17] applied the ANN to the survival analysis of the TCGA-BRCA transcriptional and integrated features datasets, exploring at the same time the benefits of transfer learning with multi-cancer datasets.

The objective of this work is to experiment the transfer learning strategy in the task of ovarian cancer prognostication with the up-to-date harmonized (aligned to hg38) RNA-sequencing (RNA-seq) data from the TCGA-OV project in order to detect significant prognostic features. The outline of the paper is as follows: section 2.2 presents the Cox survival analysis technique based on neural networks and the different aspects of deep learning, section 2.3 describes materials and methods and section 2.4 present the results and discusses the future work.

2.2 Survival analysis and deep learning

2.2.1 Cox proportional hazards and neural networks

Survival analysis, one of the statistics subfields, deals with the time-to-event as outcome. When the outcome is unknown during the observation period, it is called censoring and it is one of the major difficulties in survival analysis. Recently Wang et al [WLR17] have created a taxonomy of different approaches in this branch of statistics, distinguishing the traditional statistical and machine learning methods. One of the commonly used statistical method is a semi-parametric Cox regression or Cox proportional hazards. In this model each data instance is described by a triplet (X_i, t_i, δ_i), where $X_i = (x_{i1}, x_{i2}, ..., x_{iP})$ is the feature vector for instance i, t_i is the observed time, time of failure if δ_i is 1 or right-censoring if δ_i is 0. We note here the number of observations N and the number of features P. In this framework the rate of event at time t given that no event occurred before time t, i.e. the hazard function (recall equation 2.7).

$$h(t, X_i) = h_0(t)exp(X_i\beta) \tag{2.7}$$

where $h_0(t)$ is the baseline hazard function (an arbitrary nonnegative function of time), and $\beta^T = (\beta_1, \beta_2, ..., \beta_P)$ is the coefficient vector. To note that the features are assumed to have an exponential influence on the outcome but the baseline hazard function, $h_0(t)$, is unspecified, thus resulting in a semi-parametric model. This makes it impossible to fit the model using standard likelihood function, instead the partial likelihood is used (recall equation 2.9):

$$L(\beta) = \prod_{j=1}^{N} \left[\frac{exp(X_j\beta)}{\sum_{i \in R_j} exp(X_i\beta)} \right]^{\delta_i} \tag{2.9}$$

where R_j is the set of indices, i, with $y_i \geq t_j$ (those at risk at time t_j). The coefficient vector is estimated by maximizing this partial likelihood, or equivalently, minimizing the negative log-partial likelihood for improving efficiency (recall equation 2.10):

$$LL(\beta) = -\sum_{j=1}^{N} \delta_j \left\{ X_j\beta - log \left[\sum_{i \in R_j} exp(X_i\beta) \right] \right\} \tag{2.10}$$

The extension of Cox regression with artificial neural networks was first proposed by [FS95], who replaced the linear predictor of the Cox regression model, by a one hidden

layer multilayer perceptron (MLP). This work was further explored by [You+17] (SurvivalNet), [CZG18] (Cox-nnet), [Kat+18] (DeepSurv) and who proposed to incorporate the advances of deep learning framework and demonstrated that their methods outperform the classical Cox method. The linear predictors in their models become (recall equation 1.1):

$$\theta_i = G(WX_i + b)^T\beta \tag{1.1}$$

where W is the coefficient weight matrix between the input and hidden layer of size H x P, H is the number of neurons in the hidden layer, b is the bias vector of size H and G is the nonlinear activation function. The partial log likelihood 2.10 can be written as (recall equation 1.2):

$$PL(\beta, W) = \sum_{\delta_j=1} \left\{ \theta_j - log \left[\sum_{i \in R_j} exp(\theta_i) \right] \right\} \tag{1.2}$$

2.2.2 Regularization

When applied to high-dimensional transcriptomic data, the major issue of this model is overfitting which can be overcome with the help of different regularization techniques such as ridge regularization, dropout, early stopping and to a lesser extent batch normalization. Adding the ridge regularization term to the partial log likelihood 1.2 gives the following cost function (recall equation 1.3):

$$cost(\beta, W) = PL(\beta, W) + \lambda(||W||_2 + ||\beta||_2) \tag{1.3}$$

where $||.||_2$ designates L_2 norm penalty function and λ is a regularization coefficient leading to a weight decay.

In addition to ridge regularization, when using ANN it is common to employ dropout regularization [Sri+14]. During training, this approach randomly zeroes some of the elements of the input with probability p (dropout rate or fraction). This has proven to be an effective technique for regularization and preventing the co-adaptation of neurons as described in the paper [Hin+12].

Early stopping means stopping the training as soon as performance on a validation set starts to get worse. If regularization methods like weight decay that update the loss function to encourage less complex models are considered "explicit" regularization, then

early stopping may be thought of as a type of "implicit" regularization, much like using a smaller network that has less capacity [Zha+17].

Batch normalization (also known as batch norm) is a method used to accelerate the training of artificial neural networks. It draws its power from normalizing activations, and from incorporating this normalization in the network architecture itself. It was proposed by [IS15] and offers small regularization effect as well.

2.2.3 More data and transfer learning

Another possibility to deal with a substantial generalization error is to get more data and apply transfer learning strategy as in the study of [You+17]. Indeed, gynecologic cancers share a variety of characteristics, their development is influenced by female hormones, and they are managed by a particular medical specialty, gynecologic oncology as underlined by [Ber+18]. In this study, the authors refer to the following multi-cancer group as "pan-gyn" and focus on five TCGA tumor types: high-grade serous ovarian cystadenocarcinoma (OV), uterine corpus endometrial carcinoma (UCEC), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), uterine carcinosarcoma (UCS), and invasive breast carcinoma (BRCA). They found molecular features that differed in the "pan-gyn" group and the TCGA non-gynecologic tumor types.

This lets us hypothesize that augmenting OV training data with other datasets from the "pan-gyn" group could improve OV prognostication. The transfer learning rule of thumb being that while adding more training data, the validation and test sets should still come from the same target distribution, OV cancer in our case.

2.2.4 Automated hyperparameter optimization

Deep neural networks' prediction accuracy is highly dependable on many hyperparameters (number of layers, number and type of activation functions in each layer, and choices for optimization/regularization techniques). These details of algorithm tuning are crucial to judging whether a given technique is genuinely better, or simply better tuned.

The naïve approach of the exhaustive grid search of the hyperparameters space is time consuming, so other, more intelligent strategies have appeared recently for automated hyperparameter optimization using Bayesian optimization supported by Sequential Model-Based Global Optimization (SMBO) methodology [BYC13; Mar14].

SMBO algorithms have been used in many applications where evaluation of the fitness

function is expensive. In an application where the true fitness function, as PL in our case, is costly to evaluate, model-based algorithms approximate it with a surrogate that is cheaper to evaluate. A point that maximizes the surrogate becomes the proposal for where the true function PL should be evaluated, thus resulting in a fewer fitness function evaluations and a faster hyperparameter optimization [BYC13].

2.3 Materials and methods

2.3.1 Gene expression and clinical data

TCGA RNA-sequencing data and clinical data were downloaded from Genomics Data Commons (GDC) portal (https://portal.gdc.cancer.gov/) using the pipeline of the R/ Bioconductor package TCGAbiolinks [Col+16]. The harmonized RNA-seq data (HTSeq-counts) were normalized using the existing TCGAbiolinks normalization function which is recommended for differential expression analysis.

Supplemental survival data were downloaded from the standardized dataset named the TCGA Pan-Cancer Clinical Data Resource (TCGA-CDR) [Liu+18]. We merged the OV survival data from TCGA-CDR with the GDC clinical data. We made a choice to perform our tests on OS endpoint. The corresponding TCGA-CDR columns included OS for status and OS.time for time-to-event data. OS column contained the value 0 encoding for alive (censored) status and 1 for deceased (failure) and OS.time contained numbers of days from the date of diagnosis to either the date of last follow up if OS was 0 or time to death if OS was 1.

We downloaded the RNA-seq data for the following TCGA projects: TCGA-OV, TCGA-BRCA, TCGA-UCEC, TCGA-CESC, TCGA-UCS. After merging RNA-seq and clinical data and discarding cases without survival information, we obtained 372 samples for OV, 1076 for BRCA, 541 for UCEC, 291 for CESC, 55 for UCS. All the datasets contained 17,000 + gene expression features in common.

2.3.2 Performance metric

The widely used in survival analysis Concordance-index (C-index) measures the concordance between predicted risk score and observed survival time. This measure is computed for all comparable pairs in the test set and the number of times the predictions are concordant is summarized. The C-index value of 0.5 is equivalent to random guess and 1 is

the perfect concordance, so higher C-index means better model performance.

2.3.3 Data pre-processing

For our tests, we (log2+1) transformed the normalized values and split our dataset into 5 folds using R package MTLR [Yu+11] thus constructing 5 different splits into training and test sets with respectively 80% and 20% of samples for a further 5-fold cross-validation. As the accuracy obtained on one test set could be very different from the accuracy obtained for a different test set, the widely used K-fold cross-validation technique ensures that each fold is used as a test set at some point and provides the solution to the reliability problem. The split was done using the stratification by the OS.time and OS features in order to have similar distributions of survival times and censoring in training and test sets. To compare the survival of training and test set splits, we plotted Kaplan-Meier curves and calculated the log-rank test p-value and concluded that the difference of survival between our generated training and test sets was not significant. In order to facilitate the training procedure, the training data were standardized to zero-mean and unit-variance to comply with best practices for training deep learning algorithms. The training data included the samples from OV-only and different combinations of OV and the datasets among BRCA, UCEC, CESC and UCS. For our tests we used the DeepSurv implementation of the Python package pycox [KB19].

2.3.4 Bayesian optimization

For each of the 5 training sets, we performed 4-fold cross-validation for hyperparameter automated search, only the OV dataset samples were used in the validation sets and 16 different combinations of cancer types as optimization sets. We used python library hyperopt [BYC13] for Bayesian optimization with adaptive Tree of Parzen Estimators algorithm and the following search space: number of layers (1–8), layer width (8–2048), dropout rate (0–0.6), weight decay (0-0.9), learning rate for Adam optimizer [KB17] (0.00001-0.1) and activation function among ReLU [NH10], SELU [Kla+17], hyperbolic tangent (tanh), sigmoid function and a maximum of 200 trials. The best network design was then used to re-train a deep survival model using the optimization and validation samples, and the C-index of this best model is reported using the held-out OV testing samples. We repeated this procedure 10 times for each test dataset. To compare the C-index values in different experiments, we performed Wilcoxon rank-sum tests and report the significant (<0.05)

p-values.

2.4 Results

Transfer learning experiments showed that ANN survival models could benefit from training with multi-cancer datasets in the high-dimensional transcriptional data. The results of our tests are presented in the figure 2.1. Training with four combined datasets $OV + BRCA + UCS$, $OV + CESC + UCS$, $OV + BRCA + UCEC + UCS$, and $OV + BRCA + UCEC + CESC + UCS$ resulted in the small but significant improvements to the ANN survival model (Wilcoxon rank-sum p-values respectively of 0.018, 0.02, 0.0033 and 0.0045). Among these results, the best C-index gain of 2.1% was with OV+BRCA+UCEC+UCS combined dataset.

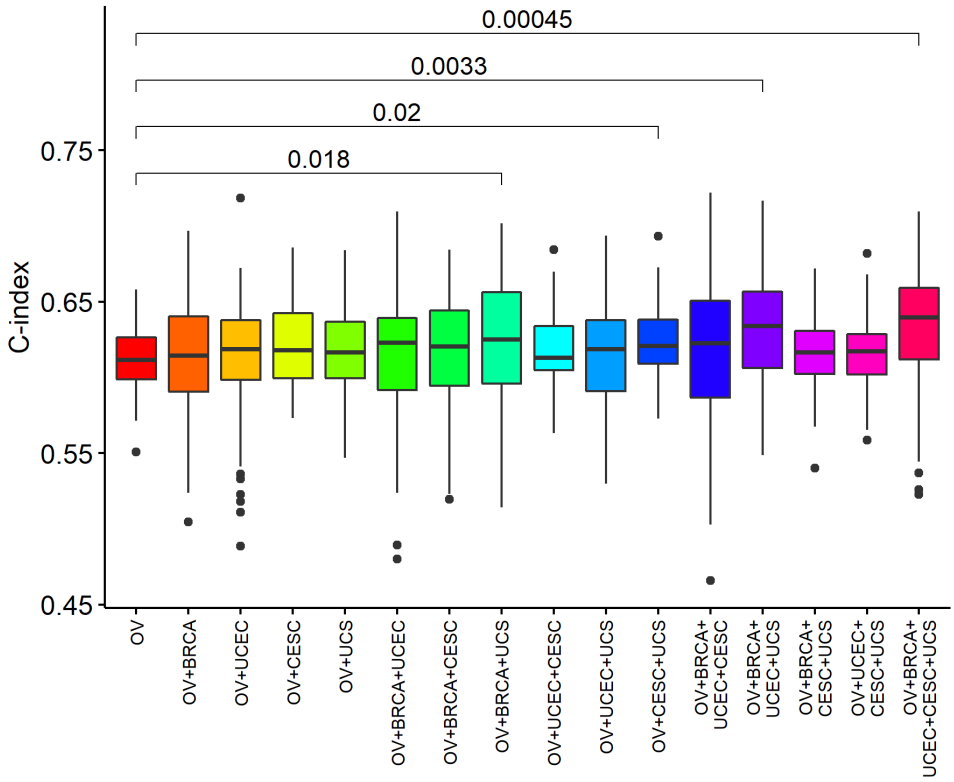


Figure 2.1: C-index comparison of transfer learning experiments. From left to right are the boxplots of the obtained 5-fold cross validation C-index on the OV test datasets. Higher C-index means better model performance. The horizontal bars in the boxes represent the median values, the boundaries of the boxes delimit lower and upper quartiles, the values outside the boxes are the lowest and the highest observations. The brackets show the significant Wilcoxon rank-sum test p-values.

The authors [You+17] noticed that prediction accuracy generally decreases as the proportion of right-censored samples in a dataset increases. We measured the censoring proportion in our datasets: OV (38.44%), BRCA (85.97%), UCEC (83.18%), CESC (75.26%) and UCS (38.18%). Interestingly, the UCS dataset with the smallest right-censoring proportion being present in all the four combined datasets with improved C-index, the best or the most significant gains are still obtained with the datasets with bigger censoring proportions than the target OV dataset itself. We hypothesize that although genetic alterations and expression patterns are often strongly associated with primary disease site, the "pan-gyn" group is likely to share common mechanisms of progression and the improved performance of the deep survival models with augmented datasets could provide additional evidence of these mechanisms.

As a future work, there is as a strong need to interpret the biological meaning of the transcriptional features contributing to the survival patient stratification. However, it is important to understand, as underlined by [Ber+18], that the "pan-gyn" project possibilities should be considered as hypothesis-generators and are to be tested and validated in the follow-up studies.

2.5 Conclusion

In this paper, we have presented the Cox proportional hazards methodology and its implementation with the artificial neural networks. We have discussed the different deep learning techniques such as regularization, automated optimization, meant to overcome the obstacles when dealing with the high-dimensional gene expression data and survival analysis. Since more data is another option to prevent the neural networks from overfitting, we have explored the transfer learning framework applied to the deep survival analysis with the TCGA ovarian RNA-seq data. According to our experiments, the deep survival models could benefit from training with the augmented multi-cancer datatsets, and more data could further improve the survival network performance.

Chapter 3

Proposed method N-MTLR-Rank

3.1 Introduction

Ovarian cancer is one of the most common female malignant tumours and the fifth leading cause of cancer-related mortality in women worldwide [SMJ20]. Because it is usually diagnosed at a late stage and currently lacks effective treatment options, the five-year survival rate for advanced stage is as low as 30% [SMJ20]. The first-line therapy of ovarian cancer patients consists of cytoreductive surgery and platinum-based chemotherapy, although 80% of newly diagnosed patients respond to the first-line therapy, approximately 75% with advanced stages experience disease relapse [Pok+19].

Currently, the additional major therapeutic regimen is a targeted Poly(ADP-ribose) polymerase (PARP) inhibitor. It is a maintenance therapy in first line for BRCA mutated high grade serous ovarian cancer (HGSOC) stages III and IV after partial or complete response to platinum salts. Additionally, in second line and onward for platinum sensitive relapsed high grade serous or endometrioid ovarian cancer [Tur+21]. Unfortunately, it is either restricted to 10% of patients who have BRCA mutations and/or hampered by the resistance phenomenon [Lu+22]. Hence, there is an urgent need to identify and validate novel, highly sensitive, and specific molecular biomarkers for prognosis, monitoring, and therapy improvement.

The development of outcome predictors is important not only for patient stratification but also to recognize categories of patients that are more likely to respond to particular therapies [Ver+12]. Multiple studies have attempted to establish molecular signatures based on gene expression to predict survival of ovarian cancer patients [Bel+11; CZG18]. However, only a small number of prognostic signatures have been developed, and none have been directly applied in clinical practice [ZH20].

The recent advances in neural networks have led to the development of the deep learning survival models [You+17; CZG18; KB19; KBS19]. These models are feed-forward artificial neural networks of the densely connected layers which transform the inputs into

more predictive lower dimension disease or biological features. A fundamental challenge in deep learning is to find the best network hyperparameters, i.e. the network design that provides the best prediction accuracy. Other common issues of the deep learning techniques are the over-fitting and generalization failure. When applied to high-dimensional transcriptomic data, the deep learning survival models show good performance on the traininig datasets and fail to generalize well on the test datasets or transfer the learned features to the independent dataset. Finally, the difficulty of deconstructing these black-box models into explainable biological processes is a big obstacle on the way of their adoption.

This paper extends our preliminary studies exploring deep learning for solving prognostic problems with high-dimensional ovarian cancer genomic profiles [Men+21a; Men+21b]. We propose a new deep learning survival model which we called N-MTLR-Rank and trained using ovarian cancer clinical and molecular data from The Cancer Genome Atlas (TCGA). We use the Bayesian optimization techniques [Ber+15] to automatically search the hyperparameter space, the different regularization techniques and transfer learning to overcome over-fitting and generalization issues. We seek to validate our deep learning survival model on an independent clinical and molecular datatset. We illustrate as well the way our model can be interpreted, by calculating the contributions of the input features to the network outputs. We demonstrate how these contributions can be related to the molecular pathways to uncover biological processes associated with ovarian cancer patients survival.

3.2 Results

3.2.1 Training and comparing deep survival networks

Our previous experiments [Men+21b] let us demonstrate the ability of deep survival models to benefit from training with data from multiple cancer types. In this work the survival models have been trained using all five TCGA datasets of the "pan-gyn" group [Ber+18]: high-grade serous ovarian cystadenocarcinoma (OV), uterine corpus endometrial carcinoma (UCEC), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), uterine carcinosarcoma (UCS), and invasive breast carcinoma (BRCA). The survival networks were evaluated for their accuracy in predicting OV outcomes and the transfer learning strategy used is shown in Fig.3.1.

The deep survival models use the negative log-likelihood to adapt the weights of the neural network to transform molecular features into lower dimension latent variables to explain survival. We compared the performance of the survival network called DeepHit [LZY18] which combines the negative discrete log-likelihood with a ranking loss and our proposed method. Our recent experiments [Men+21a] showed the promising results of the survival model N-MTLR (Neural Multi-Task Logistic Regression) [Fot18] when predicting the survival with ovarian high dimensional gene expression profiles. As in DeepHit, we add the ranking loss to the N-MTLR model and name our proposed model N-MTLR-Rank. The schema of the N-MTLR-Rank model is presented in Fig.3.2.

The obtained results (see in Fig.3.3 and 3.4) confirm that DeepHit provides good performance in terms of Concordance index (C-index) but at the cost of poorly calibrated survival estimates [KBS19]. Indeed, ranking ability of the DeepHit given by the C-index is slightly better than N-MTLR-Rank one (Wilcoxon rank-sum p=0.036), nevertheless N-MTLR-Rank provides much greater improvements of the Integrated Brier Score (IBS) (Wilcoxon rank-sum p=1.4e-07) thus overcoming the survival estimates calibration problem of DeepHit.

3.2.2 Validating with the external dataset

We explored further the prognostic accuracy of our proposed model N-MTLR-Rank on the external independent dataset. For this purpose, we selected twelve High Grade Serous Ovarian Carcinoma (HGSOC) patients treated within the Institut de Cancérologie de l'Ouest (ICO), we will refer to this dataset as ICO-OV. The results of this comparison is presented in Fig.3.5 and Fig.3.6. We observed that our model N-MTLR-Rank can generalize rather well on the new unseen data as the resulting performance on the ICO-OV dataset stays acceptable, especially for C-index metric. We hypothesize as well that poorer IBS results could be due to the small dataset size of ICO-OV. In order to compare the clinical characteristics of TCGA-OV and ICO-OV datasets we generated the descriptive statistics (see Table 3.1 and Table 3.2).

3.2.3 Interpreting N-MTLR-Rank with PatternAttribution

The machine-learning methods apply complex transformations to input features thus making the interpretation of these models difficult. Among the machine learning methods, the deep neural networks are especially seen as "black-box" since the input features in them

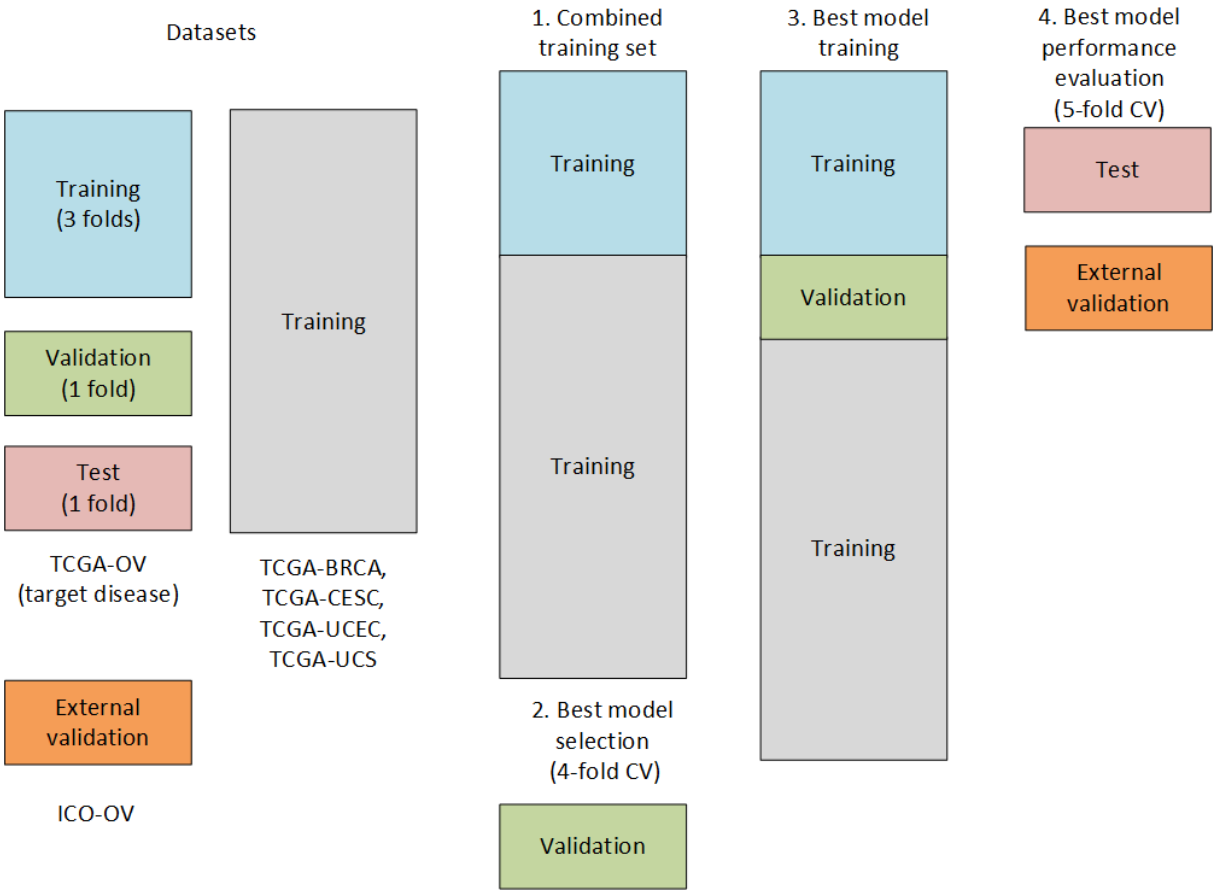


Figure 3.1: Transfer learning strategy. The target disease being ovarian cancer, the validation and test datasets from the TCGA-OV split serve to select the best model hyperparameters and evaluate the performance respectively. The combined training dataset is composed of the training TCGA-OV and the four other "pan-gyn" datasets: TCGA-BRCA, TCGA-CESC, TCGA-UCEC and TCGA-UCS. The external validation is performed on the ICO-OV dataset.

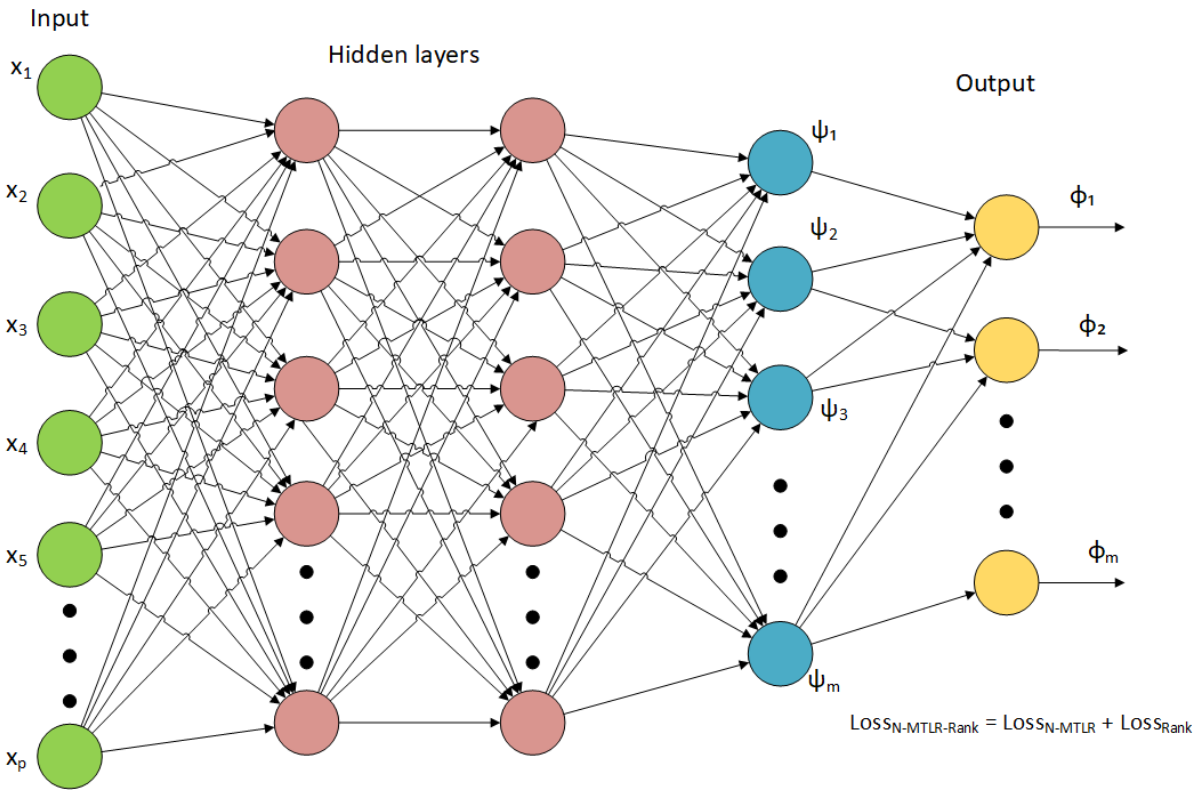


Figure 3.2: Proposed N-MTLR-Rank model architecture. The particularity of the network is that the last layer is not fully connected, instead, the reverse cumulative sum is implemented as in N-MTLR model, plus, the loss function is composed of two parts: N-MTLR loss and ranking loss, similar to PMF model.

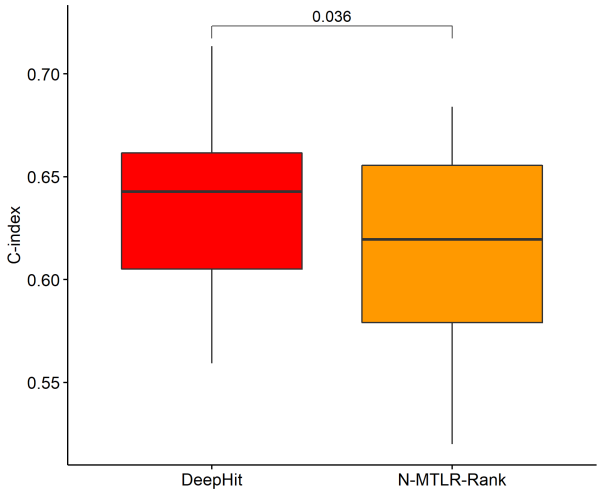


Figure 3.3: C-index comparison of DeepHit and our N-MTLR-Rank models. The boxplots of the obtained 5-fold cross validation C-index on the test TCGA-OV dataset. Higher C-index means better model performance. The horizontal bars in the boxes represent the median values, the boundaries of the boxes delimit lower and upper quartiles, the values outside the boxes are the lowest and the highest observations.

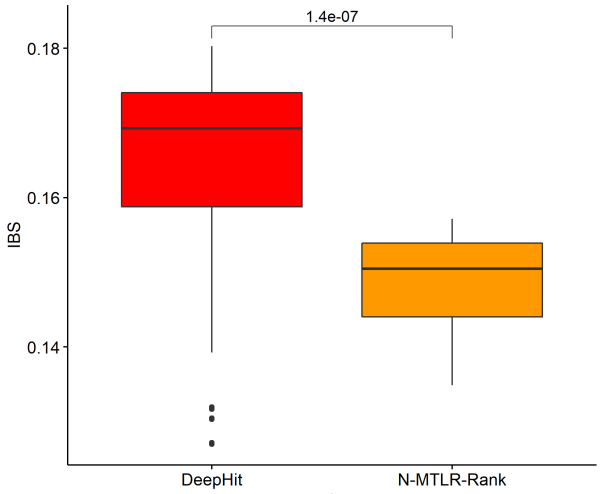


Figure 3.4: IBS comparison of DeepHit and our N-MTLR-Rank models. The boxplots of the obtained 5-fold cross validation IBS on the test TCGA-OV dataset. For this metric, smaller values signify better performance. The horizontal bars in the boxes represent the median values, the boundaries of the boxes delimit lower and upper quartiles, the values outside the boxes are the lowest and the highest observations.

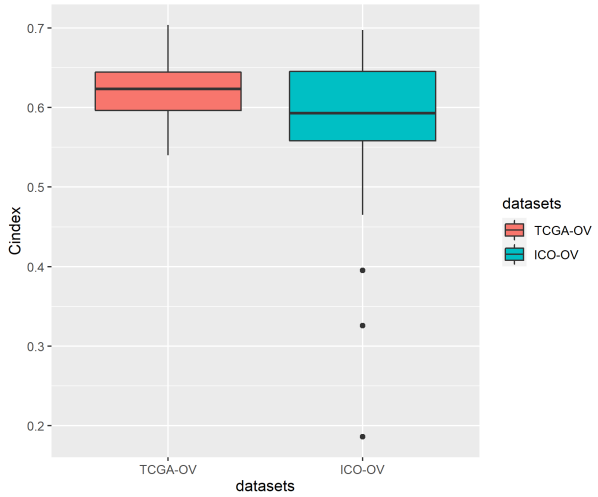


Figure 3.5: C-index comparison of our N-MTLR-Rank model generalization behavior. The boxplots of the obtained 5-fold cross validation C-index on the test TCGA-OV and ICO-OV datasets. Higher C-index means better model performance. The horizontal bars in the boxes represent the median values, the boundaries of the boxes delimit lower and upper quartiles, the values outside the boxes are the lowest and the highest observations.

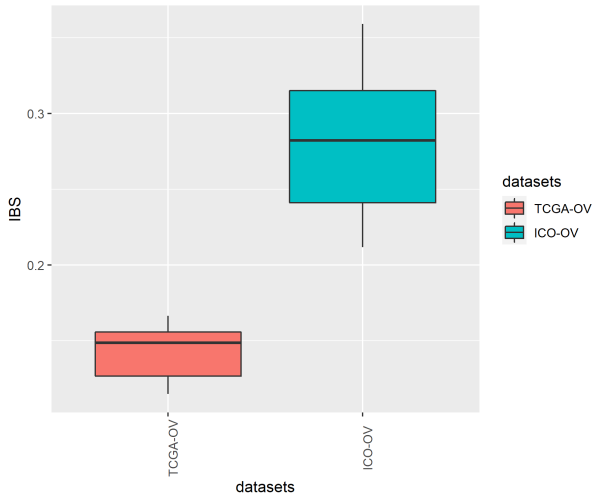


Figure 3.6: IBS comparison of our N-MTLR-Rank model generalization behavior. The boxplots of the obtained 5-fold cross validation IBS on the test TCGA-OV and ICO-OV datasets. For this metric, smaller values signify better performance. The horizontal bars in the boxes represent the median values, the boundaries of the boxes delimit lower and upper quartiles, the values outside the boxes are the lowest and the highest observations.

undergo numerous nonlinear transformations. In order to provide a clear interpretation of the prognostic significance of individual features, we used the method called PatternAttribution [Kin+17]. The authors of this method demonstrated that the direction of the model gradients does not necessarily provide an estimate for the signal in the data. Instead it reflects the relation between the signal direction and the distracting noise contributions and proposed a new decomposition method PatternAttribution by taking the data distribution into account. The measure of how much the input dimensions contribute to the output through the layers in this method is referred to as the attribution. We generated the attributions for all the input gene expressions for each patient in the TCGA-OV and ICO-OV datasets in order to measure how the input features contribute to different outputs of our networks, i.e. survival probability mass function predictions.

To investigate the molecular pathways related to ovarian prognosis, we performed the attributions Gene-Set Enrichment Analysis (GSEA) [Sub+05] of our N-MTLR-Rank model using the Molecular Signatures Database (MSigDB) [Lib+15]. The GSEA focuses on coordinated differential expression of annotated groups of genes, or gene-sets, and produces results that can more easily be interpreted in terms of the relevant biological processes. We analyzed the attributions obtained for the endpoints close to 1 year (372 days) and close to 5 years (1919 days) and run the GSEA for all the patients in TCGA-OV and ICO-OV datasets. The hallmark collections (H) used for GSEA contains 50 gene-sets and the oncogenic signature collection (C6) - 189 gene-sets, the Table 3.3 presents the overall number of significantly enriched pathways (p-value < 0.05) found for each collection in different datasets. We report as well the most frequent pathways found for patients of the TCGA-OV ($> 5\%$ at 5 years endpoint) and the presence of the corresponding pathways in ICO-OV dataset (see Table 3.4 and Table 3.5). We noted that 4 pathways out of 6 most frequent pathways in TCGA-OV are also detected in ICO-OV dataset.

Finally, these molecular pathways can be used to identify the high-risk or low-risk individuals based on their molecular pathway attribution enrichment scores. We constructed the Kaplain-Meier curves for the significantly enriched pathways of TCGA-OV cases (see Fig.3.7) separating patients into 2 groups based on the enriched or not enriched criterium. Among the 6 tested pathways IL2 STAT5 SIGNALING, ALLOGRAFT REJECTION, EG2F TARGETS, G2M CHECKPOINT, MTORC1 SIGNALING AND STK33 DN result in significant survival stratification of TCGA-OV patients.

3.3 Discussion

We proposed a new deep survival model and evaluated its ability to learn from the high dimensional transcriptomic profiles to predict the clinical outcomes. Our model N-MTLR-Rank overcomes the time-invariant covariates effect requirement of the Cox Proportional Hazards models [KBS19] providing the survival estimates for multiple time endpoints. It uses as well the censored and uncensored data during the training to adapt the neural network weights which helps to defeat the uncensored proportion drawback of the simpler models [You+17; CZG18]. We argue that its predicting accuracy performance comes from the fact that its probability estimate at time t is a function of the probability estimates at times $t'>t$. Interestingly, it is the opposite of the RNN-Surv method proposed by [GNS18] and based on the Long Short-Term Memory (LSTM) [HS97] cells which exploit the sequential nature of the problem but nevertheless N-MTLR-Rank still performs well.

N-MTLR-Rank model gave slightly worse results in terms of C-index in comparison to the similar DeepHit model, but provides significant improvements in terms of IBS criterium. We demonstrated that it is capable as well to generalize on the new unseen data, coming from RNA-sequencing of the archival Formalin Fixed Paraffin Embedded (FFPE) samples opening the possibility to exploit other retrospective cohorts.

We also experimented the new method for model interpretation, e.g. PatternAttribution [Kin+17]. This methodology let us calculate the gene expression attributions that the best trained neural network model uses to predict the outcomes. We performed the GSEA analysis using these input feature attributions and reported the most frequent enriched pathways of the MSigDB hallmark (H) and oncogenic signatures collections (C6). Among these pathways IL2 STAT5 SIGNALING, ESTROGEN RESPONSE EARLY, G2M CHECKPOINT, MTORC1 SIGNALING were already reported as prognostically enriched by the authors [You+17].

The generated Kaplan-Meier curves showed that the immune activation pathways, such as ALLOGRAFT REJECTION (p=1e-22) and IL2 STAT5 SIGNALING (p=6e-16) are strongly associated with better survival. This observation agrees with the results of another study where the authors clustered the TCGA-OV patients into immune subtypes [Lu+22] based on the manually curated immune-related genes and reported that these pathways had higher activation in immune subtype 1 associated with better survival.

The ESTROGEN RESPONSE EARLY pathway was not reported significantly prognostic (p=0.23), it is a set of genes defining an early response to estrogen. Given that

the estrogen receptor alpha (ER-alpha, ESR1) is known to be the major mediator of the estrogen response [Lan+20], we hypothesize that the TCGA-OV subpopulation detected as enriched with this pathway by our survival model could serve a plausible foundation for a future anti-estrogen therapies biomarker study.

Our findings that the proliferation pathways G2M CHECKPOINT (p=1e-04) and E2F TARGETS (p=8e-07) were associated with poorer survival, are in line with the study [ZH20] where the authors found these pathways significantly enriched in ovarian cancer suggesting that they might play a critical role in the development of ovarian cancer. The prior findings in the literature reviewed by [ZH20] suggest that the E2F family is crucial for cancer initiation, progression, and resistance to therapy. The signalling pathway MTORC1 SIGNALING (p=0.11), while not significantly prognostic, is often activated in ovarian tumors and plays an important role in tumor metabolism [Plo+21] and in the differentiation and function of immune cells [Zou+20]. Therefore, the mTOR signaling pathway is a hot target in anti-tumor therapy research.

Among the detected significantly enriched C6 gene-sets, only STK33 DN was significantly associated with the overall survival (p=4e-53). This gene- set consists of the genes downregulated in KRAS mutant cells after knockdown of STK33. Indeed, the study [Sch+09] demonstrated that STK33 is preferentially required by cells that rely on mutant KRAS for their survival and proliferation, this suggests that upregulated genes of this pathway might lead to a poorer survival which agrees with our results: the TCGA-OV patients with negatively enriched STK33 DN pathway were found to be in a high-risk group.

By contrast, the TCGA-OV patients whose attributions were enriched in KRAS.50 UP.V1 DN pathway were not associated with significantly different prognosis (p=0.29), nevertheless, this pathway includes the downregulated genes in epithelial cells over-expressing an oncogenic form of KRAS while inhibiting TBK1, KRAS synthetic lethal partner [Bar+09]. Thus, this pathway could still be an alternative method of targeting oncogenic KRAS-driven cancers.

Finally, all of the ICO-OV individuals were reported as ALLOGRAFT REJECTION and IL2 STAT5 SIGNALING enriched at 5 years endpoint, our deep survival network "see" them as low-risk group, which is in line with the overall better survival in ICO-OV dataset. The absence of the ICO-OV individuals attributions enriched in MTORC1 SIGNALING, G2M CHECKPOINT and E2F TARGETS is rather consistent as well with overall survival difference in TCGA-OV and ICO-OV datasets.

Understanding and advancing the treatment of the ovarian cancer is conditioned by the identifying the underlying biology and molecular pathogenesis of this disease. Although our study extends the insights into the use of deep learning for survival modeling, the found prognostic associated molecular pathways of ovarian cancer patients represent an interesting point for future research, it is worth underlying that these pathways should be considered hypothesis generators and more detailed in vitro experiments and follow-up clinical studies are required.

3.4 Methods

3.4.1 Data

TCGA gene expression and clinical data

TCGA RNA-sequencing and clinical data were downloaded from Genomics Data Commons (GDC) portal using the pipeline of the R/Bioconductor package TCGAbiolinks [Col+16]. Supplemental survival data were downloaded from the standardized dataset named the TCGA Pan-Cancer Clinical Data Resource (TCGA-CDR) [Liu+18]. We merged the survival data from TCGA-CDR with the GDC clinical data. We performed our tests on overall (OS) endpoint. The corresponding TCGA-CDR columns included OS for status and OS.time for time-to-event data. OS column contained the value 0 encoding for alive (censored) status and 1 for deceased (failure) and OS.time contained numbers of days from the date of diagnosis to either the date of last follow up if OS was 0 or time to death if OS was 1.

We downloaded the RNA-seq data for the following TCGA projects: TCGA-OV, TCGA-BRCA, TCGA-UCEC, TCGA-CESC, TCGA-UCS. The harmonized GRCh38 aligned RNA-seq data (HTSeq-counts[APH15]) were normalized per TCGA project using the TCGAbiolinks normalization function which is recommended for differential expression analysis. After merging RNA-seq and clinical data and discarding cases without survival information, we obtained 372 samples for OV, 1087 for BRCA, 549 for UCEC, 291 for CESC, 55 for UCS. All the datasets contained 16673 gene expression features in common.

External validation dataset ICO-OV

We curated 12 ICO patients diagnosed with HGSOC between 2007 and 2016, their retrospective electronic records data were collected: date of birth, date of pathologic diagnosis,

clinical stage, histological grade, date of death or date of last follow-up, the age at pathological diagnosis, OS and OS.time were derived.

We extracted RNA from the corresponding archival FFPE slides with COVARIS ME220 Focused-ultrasonicator, to allow a high quantity and high quality of RNA extracted. RNA libraries were prepared with the SureSelect XT HS2 RNA Reagent kit and the SureSelectXT Human All Exon V6 +UTR probes from Agilent. All libraries were sequenced on an Illumina NextSeq550 in paired-end mode (2 x 75bp) with a target depth of 20 million fragments per sample. Sequenced reads were trimmed with fastp v0.20.1 and mapped to GRCh38 using HISAT2 v2.1.0 both with default parameters. Reads overlapping genomic features were counted with featureCounts v2.0.0 [LSS14] from the Subread package and Ensembl v99. Only uniquely mapped and not duplicated reads were counted. Multiple overlaps of unique genomic feature were not counted.

The obtained raw featureCounts [LSS14] of the ICO-OV dataset were further normalized using the TCGAbiolinks normalization function which resulted in 15521 genes in common between ICO-OV and TCGA "pan-gyn" group. In order to account for “batch effect”, we used the method ComBat-seq [ZPJ20] particularly suited for RNA-seq data.

Our study was approved by the ethics committee of the university hospital center of Angers (2021-102) and done in accordance with ethical standards of the 1964 Helsinki Declaration and its later amendments. Patients provided signed informed consent in accordance with their respective trial protocols.

3.4.2 Proposed deep survival model

Deep survival models are multi-layer artificial neural networks with different output layers that use various negative log-likelihood based loss functions. The papers [KBS19; KB19] give the overview of the different feed forward survival models and their corresponding loss functions. The authors note that the discrete-time models may be used as approximations of models in continuous time subdividing time into m intervals. Among the discrete-time methods they used the following negative log-likelihood and called the method PMF:

$$loss_{PMF} = -\frac{1}{N}\sum_{i=1}^{N}\left(\delta_i log[\sigma_{k(t_i)}(\phi(X_i))] + (1-\delta_i) log[\hat{S}(k(t_i)|X_i)]\right), \tag{3.1}$$

where $\hat{S}(k(t_i)|X) = 1 - \sum_{k=1}^{i}\sigma_k(\phi(X))$ is the estimated survival function,

$\sigma_i(\phi(X)) = -\frac{exp[\phi_i(X)]}{1+\sum_{k=1}^{m} exp[(\phi_k(X)]}$ is the softmax function,

$\phi(X)$ is the neural network,

$k(t_i)$ is the duration index of individual time t_i among m intervals,

δ_i is the censoring indicator,

X_i are the features of individual i,

X is a feature matrix of N individuals and P features.

Another discrete-time method reviewed in [KBS19] is the Neural Multi-Task Logistic Regression (N-MTLR) proposed by Fotso et al [Fot18]. This work is the neural network adaptation of Multi-Task Logistic Regression (MTLR) by Yu et al [Yu+11] and, as shown by [KBS19], N-MTLR is equivalent to PMF method in (3.1) but where:

$\phi_j(X) = \sum_{k=j}^{m} \psi(X_k)$ is the (reverse) cumulative sum of the output of the network $\psi(X_k)$.

According to this work, Lee et al. [LZY18] were the 1st to apply neural networks to the discrete-time likelihood for right-censored time-to-event data. The proposed method, denoted DeepHit, estimates the probability mass function (PMF) with a neural net and combines the log-likelihood with a ranking loss, for one type of event:

$$loss_{DeepHit} = \alpha loss_{PMF} + (1-\alpha) loss_{rank} \tag{3.2}$$

$$loss_{rank} = \sum_{i,j} \delta_i \mathbb{1}(t_i < t_j) exp\left(\frac{\hat{S}(k(t_i)|X_i) - \hat{S}(k(t_i)|X_j)}{\beta}\right), \tag{3.3}$$

where α and β are the hyperparameters of the network.

We have recently reported in [Men+21a] that N-MTLR performance is better than PMF for TCGA-OV transciptome based prognostication. So we propose a new method which we called N-MTLR-Rank as it combines the discrete-time negative log-likelihood of N-MTLR and the ranking loss of DeepHit:

$$loss_{N-MTLR-Rank} = \alpha loss_{N-MTLR} + (1-\alpha) loss_{rank} \tag{3.4}$$

3.4.3 Model training and validation

For our tests, we (log2+1) transformed all the normalized values and split only TCGA-OV dataset into 5 folds using R package MTLR [Yu+11] thus constructing 5 different splits into training and test sets with respectively 80% and 20% of samples for a further 5-fold cross-validation. The split was done using the stratification by the OS.time and OS features in order to have similar distributions of survival times and censoring in training and test sets. For the sake of facilitating the training procedure, the training data were standardized to zero-mean and unit-variance to comply with best practices for training deep learning algorithms. To note that in order to benefit from the multi-cancer transfer learning strategy, all the training hereafter included BRCA+CESC+UCEC+UCS along with TCGA-OV samples.

For each of the five 80% TCGA-OV training sets, we performed the best hyperparameters combination search based on the Bayesian optimization technique. We further split the 80% training set into 60% optimization and 20% validation with the aim to train the networks with the optimization set and evaluate on validation set. We used python library hyperopt [Ber+15] for Bayesian optimization with adaptive Tree of Parzen Estimators algorithm, the maximum number of trials of 400 and the following search space:

- number of layers: 1–8
- layer width: 8–2048
- learning rate for Adam optimizer [KB17]: 0.00001-0.1
- weight decay [LH19]: 0-0.9
- dropout rate [Sri+14]: 0–0.6
- activation function: ReLU [NH10], SELU [Kla+17], hyperbolic tangent (tanh), sigmoid
- discretization scheme: equidistant or Kaplan-Meier quantiles [KB19]
- interpolation scheme: constant density interpolation (CDI) or constant hazard interpolation (CHI) [KB19]
- α, ranking loss parameter in (3.3): 0-1

- β, ranking loss parameter in (3.3): 0.1-100

The best network design was then used to re-train a deep survival model using the 80% training data and the 20% test set to evaluate C-index and IBS. We repeated this procedure 10 times for each test set. We reported the time-dependent C-index, which estimates the probability that observations i and j are concordant given that they are comparable. The C-index value of 0.5 is equivalent to random guess and 1 is the perfect concordance. As for IBS, it is an extension of Brier Score (BS) over an interval of time, where BS is the mean squared error of the probability estimates. For this metric, smaller values signify better performance, for more details see [KB19]. For our experiments we used DeepHit implementation in the Python package pycox [KB19]. For N-MTLR-Rank we used N-MTLR implementation along with the ranking loss implementation of DeepHit of the same python package.

3.4.4 Model selection and interpretation

The model used for interpretation was created by identifying the best performing model configuration for TCGA-OV and ICO-OV experiments. This configuration was then used to re-train a model using all available TCGA-OV samples. Feature attributions were calculated using the PatternAttribution method [Kin+17] implemented in PyTorch and available at https://github.com/KnurpsBram/PyTorch-PatternNet.

Feature attributions were analyzed using the R/Bioconductor package clusterProfiler [Wu+21] for the GSEA using (MSigDB) [Lib+15] hallmark collections (H) and oncogenic signature collection (C6) gene-sets.

Variable		Overall
Age at pathologic diagnosis		
Count		372
Mean (SD)		59.60 (11.38)
Median (IQR)		59.00 (17.00)
Q1, Q3		51.00, 68
Min, Max		30.00, 87
Missing		0
Clinical stage		
Count (%)		372
Stage IC		1 (0.27%)
Stage IIA		3 (0.81%)
Stage IIB		3 (0.81%)
Stage IIC		15 (4.03%)
Stage IIIA		7 (1.88%)
Stage IIIB		13 (3.49%)
Stage IIIC		270 (72.58%)
Stage IV		57 (15.32%)
Missing		3 (0.81%)
Histological grade		
Count (%)		372
G1		1 (0.27%)
G2		42 (11.29%)
G3		319 (85.75%)
G4		1 (0.27%)
GB		2 (0.54%)
GX		5 (1.34%)
Missing		2 (0.54%)
OS		
Count (%)		372
0		143 (38.44%)
1		229 (61.56%)
Missing		0
OS.time		
Count		372
Mean (SD)		1187.17 (943.74)
Median (IQR)		1024.00 (1141.75)
Q1, Q3		517.25, 1659
Min, Max	111	8.00, 5481
Missing		0

Table 3.1: TCGA-OV clinical descriptive statistics.

Variable	Overall
Age at pathologic diagnosis	
Count	12
Mean (SD)	63.17 (12.64)
Median (IQR)	67.50 (20.50)
Q1, Q3	50.50, 71
Min, Max	46.00, 86
Missing	0
Clinical stage	
Count (%)	12
IB	1 (8.33%)
II	1 (8.33%)
IIIA	2 (16.67%)
IIIC	7 (58.33%)
IV	1 (8.33%)
Missing	0
Histological grade	
Count (%)	12
G3	319 (100%)
Missing	0
OS	
Count (%)	12
0	7 (58.33%)
1	5 (41.67%)
Missing	0
OS.time	
Count	12
Mean (SD)	2132.75 (1609.92)
Median (IQR)	1677.50 (2786.50)
Q1, Q3	954.00, 3740.5
Min, Max	8.00, 4646
Missing	0

Table 3.2: ICO-OV clinical descriptive statistics.

Dataset, collection	1 year	5 years
TCGA-OV, hallmark	6	12
ICO-OV, hallmark	4	6
TCGA-OV, C6	10	10
ICO-OV, C6	8	2

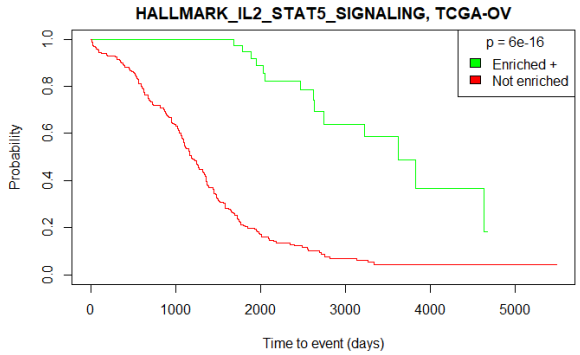
Table 3.3: The overall number of significantly enriched pathways found in hallmark and C6 MSigDB collections (p-value < 0.05).

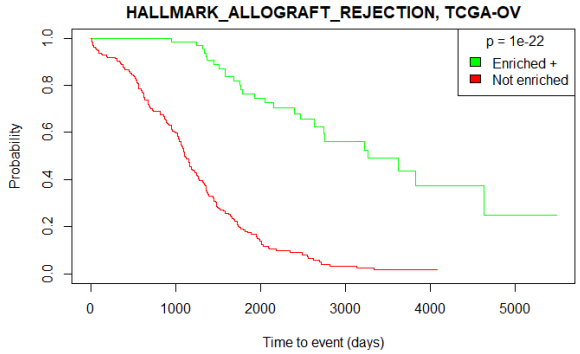
Collection	TCGA-OV, 1 y.	ICO-OV, 1 y.	TCGA-OV, 5 y.	ICO-OV, 5 y.
HALLMARK_IL2_STAT5_SIGNALING	2.96%	91.67%	20.97%	100%
HALLMARK_ALLOGRAFT_REJECTION	0.54%	91.67%	25.81%	100%
HALLMARK_ESTROGEN_RESPONSE_EARLY	0.81%	50%	6.72%	25%
HALLMARK_E2F_TARGETS	-	-	13.98%	-
HALLMARK_G2M_CHECKPOINT	-	-	8.33%	-
HALLMARK_MTORC1_SIGNALING	-	-	5.91%	-

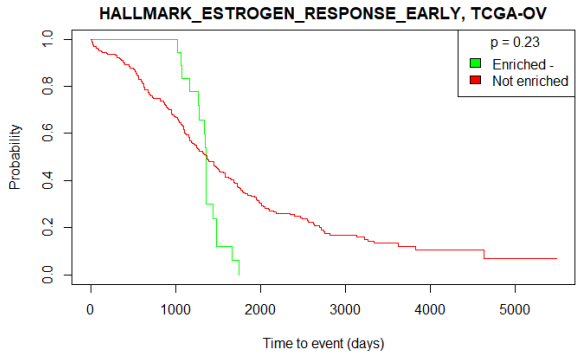
Table 3.4: The most frequent hallmark pathways found for patients of the TCGA-OV and the presence of the corresponding pathways in ICO-OV dataset ($> 5\%$ at 5 years endpoint).

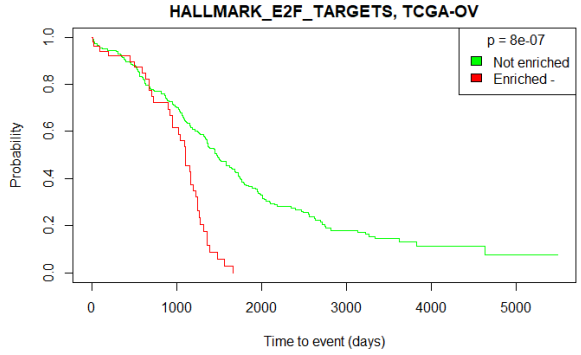
Collection	TCGA-OV, 1 y.	ICO-OV, 1 y.	TCGA-OV, 5 y.	ICO-OV, 5 y.
KRAS.50_UP.V1_DN	22.04%	25%	1.08%	-
STK33_DN	1.34%	-	10.22%	-

Table 3.5: The most frequent C6 collection pathways found for patients of the TCGA-OV and the presence of the corresponding pathways in ICO-OV dataset ($> 5\%$ at 5 years endpoint).


HALLMARK_IL2_STAT5_SIGNALING, TCGA-OV
p = 6e-16
Enriched +
Not enriched
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
1000
2000
3000
4000
5000
Time to event (days)


HALLMARK_ALLOGRAFT_REJECTION, TCGA-OV
p = 1e-22
Enriched +
Not enriched
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
1000
2000
3000
4000
5000
Time to event (days)


HALLMARK_ESTROGEN_RESPONSE_EARLY, TCGA-OV
p = 0.23
Enriched -
Not enriched
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
1000
2000
3000
4000
5000
Time to event (days)


HALLMARK_E2F_TARGETS, TCGA-OV
p = 8e-07
Not enriched
Enriched -
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
1000
2000
3000
4000
5000
Time to event (days)

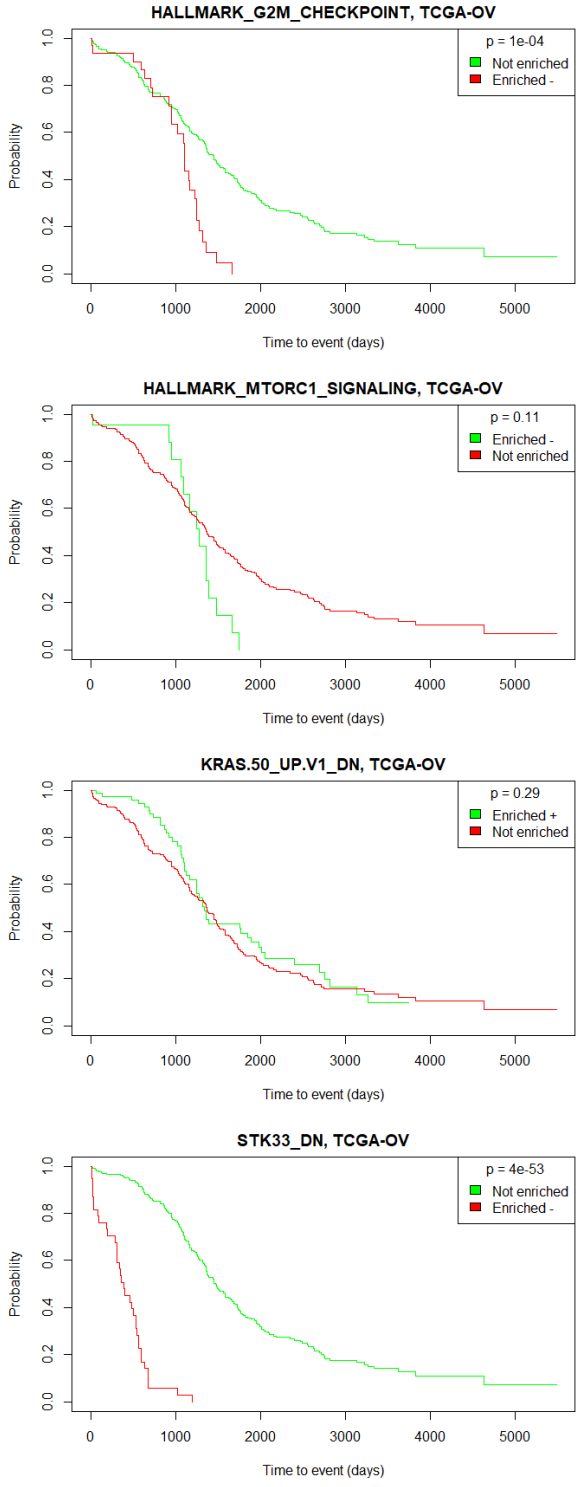


Figure 3.7: KM curves of the high-risk and low-risk TCGA-OV groups. The two groups were defined on the basis of the individual molecular pathway attribution enrichment scores.

Conclusion Part II

Since neither traditional regularized Cox model nor Cox-nnet produced satisfactory results in transcriptomic TCGA-OV dataset, it was important to search for other survival analysis techniques capable to deal with ovarian RNA-seq data. The results obtained in the Chapter 1 "Comparative study" show that the N-MTLR model appears as the most effective and promising one outperforming all the other ANN based techniques found in literature. We have reviewed the neural networks based survival analysis techniques adaptable to deal with the high-dimensional gene expression data and have benchmarked the following methods built on the neural networks for continuous and discrete time data: Cox-nnet, DeepSurv, Cox CC, Cox Time, PC-Hazard, Logistic Hazard or Nnet-Survival, PMF, N-MTLR.

According to our transfer learning experiments presented in Chapter 2, the deep survival models could benefit from training with the augmented multi-cancer datatsets, and more data could further improve the survival network performance. We have discussed the different deep learning techniques such as regularization, automated optimization, meant to overcome the obstacles when dealing with the high-dimensional gene expression data and survival analysis. In order to prevent the neural networks from overfitting, we have explored the transfer learning framework applied to the deep survival analysis with the TCGA ovarian RNA-seq data and have showed that the whole "pan-gyn" group is profitable in the ovarian cancer prognostication task.

Based on the previously obtained results, as explained in Chapter 3, we have proposed a new deep survival model and evaluated its ability to learn from the high dimensional transcriptomic profiles to predict the clinical outcomes. Our model N-MTLR-Rank overcomes the time-invariant covariates effect requirement of the Cox Proportional Hazards models providing the survival estimates for multiple time endpoints. It is capable as well to generalize on the new unseen data, coming from RNA-sequencing of the archival FFPE samples opening the possibility to exploit other retrospective cohorts.

We have also experimented the new method for model interpretation and reported the found enriched MSigDB pathways: IL2 STAT5 SIGNALING, ALLOGRAFT REJECTION, ESTROGEN RESPONSE EARLY, G2M CHECKPOINT, E2F TARGETS,

MTORC1 SIGNALING, STK33 DN and KRAS.50 UP.V1 DN. We advocate that these molecular pathways represent an interesting point for future research with more detailed in vitro experiments and follow-up clinical studies.

As a future work, the integration of multiple data types (other omics data, whole-slide images, etc.) to construct performant models for survival prediction based on the N-MTLR model appear to be an intresting direction. There is as well a strong need to interpret the obtained results and link them to the information with the biological meaning to be applicable in clinical decision-making.

Primary sources

[Aal78] Odd Aalen, "Nonparametric Inference for a Family of Counting Processes", *in*: *The Annals of Statistics* 6.*4* (July 1978), ISSN: 0090-5364, DOI: 10.1214/aos/1176344247, URL: https://projecteuclid.org/journals/annals-of-statistics/volume-6/issue-4/Nonparametric-Inference-for-a-Family-of-Counting-Processes/10.1214/aos/1176344247.full (visited on 09/23/2022).

[ABB05] Laura Antolini, Patrizia Boracchi, and Elia Biganzoli, "A time-dependent discrimination index for survival data", en, *in*: *Statistics in Medicine* 24.*24* (Dec. 2005), pp. 3927–3944, ISSN: 0277-6715, 1097-0258, DOI: 10.1002/sim.2427, URL: http://doi.wiley.com/10.1002/sim.2427 (visited on 05/22/2020).

[APH15] S. Anders, P. T. Pyl, and W. Huber, "HTSeq–a Python framework to work with high-throughput sequencing data", en, *in*: *Bioinformatics* 31.*2* (Jan. 2015), pp. 166–169, ISSN: 1367-4803, 1460-2059, DOI: 10.1093/bioinformatics/btu638, URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu638 (visited on 04/01/2022).

[Aud17] Jérôme Audoux, "A l'assaut du puzzle transcriptomique: optimisations, applications et nouvelles méthodes d'analyse pour le RNA-Seq", Français, Médecine humaine et pathologie, Université Montpellier, 2017.

[Bar+09] David A. Barbie et al., "Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1", en, *in*: *Nature* 462.*7269* (Nov. 2009), pp. 108–112, ISSN: 0028-0836, 1476-4687, DOI: 10.1038/nature08460, URL: http://www.nature.com/articles/nature08460 (visited on 06/02/2022).

[Bel+11] D. Bell et al., "Integrated genomic analyses of ovarian carcinoma", *in*: *Nature* 474.*7353* (June 2011), pp. 609–615, ISSN: 0028-0836, 1476-4687, DOI: 10.1038/nature10166, URL: http://www.nature.com/doifinder/10.1038/nature10166 (visited on 12/07/2018).

[Ber+15] James Bergstra et al., "Hyperopt: a Python library for model selection and hyperparameter optimization", en, *in*: *Computational Science & Discovery* 8.*1* (July 2015), p. 014008, ISSN: 1749-4699, DOI: 10.1088/1749-4699/8/

1/014008, URL: https://iopscience.iop.org/article/10.1088/1749-4699/8/1/014008 (visited on 03/19/2021).

[Ber+18] Ashton C. Berger et al., "A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers", en, *in*: *Cancer Cell* 33.*4* (Apr. 2018), 690–705.e9, ISSN: 15356108, DOI: 10.1016/j.ccell.2018.03.014, URL: https://linkinghub.elsevier.com/retrieve/pii/S1535610818301193 (visited on 11/23/2018).

[Bha+18] Sanchita Bhattacharya et al., "ImmPort, toward repurposing of open access immunological assay data for translational and clinical research", en, *in*: *Scientific Data* 5.*1* (Dec. 2018), p. 180015, ISSN: 2052-4463, DOI: 10.1038/sdata.2018.15, URL: http://www.nature.com/articles/sdata201815 (visited on 07/29/2022).

[Bin+13] H. Binder et al., "Tailoring sparse multivariable regression techniques for prognostic single-nucleotide polymorphism signatures", en, *in*: *Statistics in Medicine* 32.*10* (May 2013), pp. 1778–1791, ISSN: 02776715, DOI: 10.1002/sim.5490, URL: https://onlinelibrary.wiley.com/doi/10.1002/sim.5490 (visited on 07/29/2022).

[Bon+08] T. Bonome et al., "A Gene Signature Predicting for Survival in Suboptimally Debulked Patients with Ovarian Cancer", en, *in*: *Cancer Research* 68.*13* (July 2008), pp. 5478–5486, ISSN: 0008-5472, 1538-7445, DOI: 10.1158/0008-5472.CAN-07-6595, URL: http://cancerres.aacrjournals.org/cgi/doi/10.1158/0008-5472.CAN-07-6595 (visited on 02/21/2020).

[Bri50] Glenn W. Brier, "VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY", en, *in*: *Monthly Weather Review* 78.*1* (Jan. 1950), pp. 1–3, ISSN: 0027-0644, 1520-0493, DOI: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2, URL: http://journals.ametsoc.org/doi/10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2 (visited on 09/23/2022).

[Bro75] Charles C. Brown, "On the Use of Indicator Variables for Studying the Time-Dependence of Parameters in a Response-Time Model", *in*: *Biometrics* 31.*4* (Dec. 1975), p. 863, ISSN: 0006341X, DOI: 10.2307/2529811, URL: https://www.jstor.org/stable/2529811?origin=crossref (visited on 09/05/2022).

[Bru+04] Jean-Philippe Brunet et al., "Metagenes and molecular pattern discovery using matrix factorization", en, *in*: *Proceedings of the National Academy of Sciences* 101.*12* (Mar. 2004), pp. 4164–4169, ISSN: 0027-8424, 1091-6490, DOI: 10.1073/pnas.0308531101, URL: https://pnas.org/doi/full/10.1073/pnas.0308531101 (visited on 07/29/2022).

[BYC13] J Bergstra, D Yamins, and D D Cox, "Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures", en, *in*: (2013), p. 9.

[Cai+21] Shurui Cai et al., "Bioinformatics analysis of miRNAs identifies enrichment of axon guidance pathway genes in ovarian cancer stem cells", en, *in*: *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Houston, TX, USA: IEEE, Dec. 2021, pp. 2415–2422, ISBN: 978-1-66540-126-5, DOI: 10.1109/BIBM52615.2021.9669299, URL: https://ieeexplore.ieee.org/document/9669299/ (visited on 01/28/2022).

[CH21] Zhiao Chen and Xianghuo He, "Application of third-generation sequencing in cancer research", en, *in*: *Medical Review* 1.*2* (Dec. 2021), pp. 150–171, ISSN: 2749-9642, DOI: 10.1515/mr-2021-0013, URL: https://www.degruyter.com/document/doi/10.1515/mr-2021-0013/html (visited on 09/05/2022).

[Cha+18] Kumardeep Chaudhary et al., "Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer", en, *in*: *Clinical Cancer Research* 24.*6* (Mar. 2018), pp. 1248–1259, ISSN: 1078-0432, 1557-3265, DOI: 10.1158/1078-0432.CCR-17-0853, URL: http://clincancerres.aacrjournals.org/lookup/doi/10.1158/1078-0432.CCR-17-0853 (visited on 06/20/2019).

[Che+15] Yen-Chen Chen et al., "Cancer adjuvant chemotherapy strategic classification by artificial neural network with gene expression data: An example for non-small cell lung cancer", en, *in*: *Journal of Biomedical Informatics* 56 (Aug. 2015), pp. 1–7, ISSN: 15320464, DOI: 10.1016/j.jbi.2015.05.006, URL: https://linkinghub.elsevier.com/retrieve/pii/S1532046415000854 (visited on 02/07/2019).

[Chi+18] Travers Ching et al., "Opportunities and obstacles for deep learning in biology and medicine", en, *in*: *Journal of The Royal Society Interface* 15.*141* (Apr. 2018), p. 20170387, ISSN: 1742-5689, 1742-5662, DOI: 10.1098/rsif.

2017.0387, URL: https://royalsocietypublishing.org/doi/10.1098/rsif.2017.0387 (visited on 01/17/2020).

[Col+16] Antonio Colaprico et al., "*TCGAbiolinks* : an R/Bioconductor package for integrative analysis of TCGA data", en, *in*: *Nucleic Acids Research* 44.*8* (May 2016), e71–e71, ISSN: 0305-1048, 1362-4962, DOI: 10.1093/nar/gkv1507, URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1507 (visited on 03/08/2019).

[Cox72] D. R. Cox, "Regression Models and Life-Tables", en, *in*: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.*2* (Jan. 1972), pp. 187–202, ISSN: 00359246, DOI: 10.1111/j.2517-6161.1972.tb00899.x, URL: https://onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1972.tb00899.x (visited on 06/24/2022).

[Cri+09] Anne P. G Crijns et al., "Survival-Related Profile, Pathways, and Transcription Factors in Ovarian Cancer", en, *in*: *PLoS Medicine* 6.*2* (Feb. 2009), ed. by Steven Narod, e1000024, ISSN: 1549-1676, DOI: 10.1371/journal.pmed.1000024, URL: https://dx.plos.org/10.1371/journal.pmed.1000024 (visited on 02/10/2022).

[CZG18] Travers Ching, Xun Zhu, and Lana X. Garmire, "Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data", en, *in*: *PLOS Computational Biology* 14.*4* (Apr. 2018), ed. by Florian Markowetz, e1006076, ISSN: 1553-7358, DOI: 10.1371/journal.pcbi.1006076, URL: https://dx.plos.org/10.1371/journal.pcbi.1006076 (visited on 06/20/2019).

[Dob+13] Alexander Dobin et al., "STAR: ultrafast universal RNA-seq aligner", en, *in*: *Bioinformatics* 29.*1* (Jan. 2013), pp. 15–21, ISSN: 1460-2059, 1367-4803, DOI: 10.1093/bioinformatics/bts635, URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts635 (visited on 04/01/2022).

[Dow15] Allen B. Downey, *Think Stats: exploratory data analysis*, eng, 2. ed, Beijing Köln: O'Reilly, 2015, ISBN: 978-1-4919-0733-7.

[FHT10] Jerome Friedman, Trevor Hastie, and Robert Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent", en, *in*: *Journal of Statistical Software* 33.*1* (2010), ISSN: 1548-7660, DOI: 10.18637/jss.v033.i01, URL: http://www.jstatsoft.org/v33/i01/ (visited on 01/25/2019).

[Fot18] Stephane Fotso, "Deep Neural Networks for Survival Analysis Based on a Multi-Task Framework", *in*: *arXiv:1801.05512 [cs, stat]* (Jan. 2018), arXiv: 1801.05512, URL: http://arxiv.org/abs/1801.05512 (visited on 10/10/2019).

[FS95] David Faraggi and Richard Simon, "A neural network model for survival data", en, *in*: *Statistics in Medicine* 14.*1* (Jan. 1995), pp. 73–82, ISSN: 02776715, 10970258, DOI: 10.1002/sim.4780140108, URL: http://doi.wiley.com/10.1002/sim.4780140108 (visited on 02/18/2021).

[FSA19] FRANCOGYN, SFOG, and ARCAGY-GINECO, *Conduites à tenir initiales devant des patientes atteintes d'un cancer épithélial de l'ovaire / Synthèse*, Nov. 2019.

[Gho+19] Benyamin Ghojogh et al., "Feature Selection and Feature Extraction in Pattern Analysis: A Literature Review", *in*: (2019), DOI: 10.48550/ARXIV.1905.02845, URL: https://arxiv.org/abs/1905.02845 (visited on 06/30/2022).

[GN18] Michael F. Gensheimer and Balasubramanian Narasimhan, "A Scalable Discrete-Time Survival Model for Neural Networks", *in*: *arXiv:1805.00917 [cs, stat]* (May 2018), arXiv: 1805.00917, URL: http://arxiv.org/abs/1805.00917 (visited on 10/10/2019).

[GNS18] Eleonora Giunchiglia, Anton Nemchenko, and Mihaela van der Schaar, "RNN-SURV: A Deep Recurrent Model for Survival Analysis", en, *in*: *Artificial Neural Networks and Machine Learning – ICANN 2018*, ed. by Věra Kůrková et al., vol. 11141, Series Title: Lecture Notes in Computer Science, Cham: Springer International Publishing, 2018, pp. 23–32, DOI: 10.1007/978-3-030-01424-7_3, URL: http://link.springer.com/10.1007/978-3-030-01424-7_3 (visited on 03/19/2020).

[Gra+99] Erika Graf et al., "Assessment and comparison of prognostic classification schemes for survival data", en, *in*: *Statistics in Medicine* 18.*17-18* (Sept. 1999), pp. 2529–2545, ISSN: 0277-6715, 1097-0258, DOI: 10.1002/(SICI)1097-0258(19990915/30)18:17/18<2529::AID-SIM274>3.0.CO;2-5, URL: https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-0258(19990915/30)18:17/18%3C2529::AID-SIM274%3E3.0.CO;2-5 (visited on 09/23/2022).

[GSL19] Yang Guo, Xuequn Shang, and Zhanhuai Li, "Identification of cancer subtypes by integrating multiple types of transcriptomics data with deep learning in breast cancer", en, *in*: *Neurocomputing* 324 (Jan. 2019), pp. 20–30, ISSN: 09252312, DOI: 10.1016/j.neucom.2018.03.072, URL: https://linkinghub.elsevier.com/retrieve/pii/S0925231218306222 (visited on 12/07/2018).

[Hao+19] Jie Hao et al., "Interpretable deep neural network for cancer survival analysis by integrating genomic and clinical data", en, *in*: *BMC Medical Genomics* 12.*S10* (Dec. 2019), ISSN: 1755-8794, DOI: 10.1186/s12920-019-0624-2, URL: https://bmcmedgenomics.biomedcentral.com/articles/10.1186/s12920-019-0624-2 (visited on 03/19/2020).

[Har+84] Frank E. Harrell et al., "Regression modelling strategies for improved prognostic prediction", en, *in*: *Statistics in Medicine* 3.*2* (Apr. 1984), pp. 143–152, ISSN: 02776715, 10970258, DOI: 10.1002/sim.4780030207, URL: https://onlinelibrary.wiley.com/doi/10.1002/sim.4780030207 (visited on 09/06/2022).

[HB15] Jacob J. Hughey and Atul J. Butte, "Robust meta-analysis of gene expression using the elastic net", en, *in*: *Nucleic Acids Research* 43.*12* (July 2015), e79–e79, ISSN: 0305-1048, 1362-4962, DOI: 10.1093/nar/gkv229, URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv229 (visited on 01/04/2019).

[HG15] Zena M. Hira and Duncan F. Gillies, "A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data", en, *in*: *Advances in Bioinformatics* 2015 (June 2015), pp. 1–13, ISSN: 1687-8027, 1687-8035, DOI: 10.1155/2015/198363, URL: https://www.hindawi.com/journals/abi/2015/198363/ (visited on 06/30/2022).

[Hin+12] Geoffrey E. Hinton et al., "Improving neural networks by preventing co-adaptation of feature detectors", *in*: *arXiv:1207.0580 [cs]* (July 2012), arXiv: 1207.0580, URL: http://arxiv.org/abs/1207.0580 (visited on 03/25/2021).

[HS97] Sepp Hochreiter and Jürgen Schmidhuber, "Long Short-Term Memory", en, *in*: *Neural Computation* 9.*8* (Nov. 1997), pp. 1735–1780, ISSN: 0899-7667, 1530-888X, DOI: 10.1162/neco.1997.9.8.1735, URL: https://direct.mit.edu/neco/article/9/8/1735-1780/6109 (visited on 06/10/2022).

[Hua+19a] Zhi Huang et al., "SALMON: Survival Analysis Learning With Multi-Omics Neural Networks on Breast Cancer", *in*: *Frontiers in Genetics* 10 (Mar. 2019), ISSN: 1664-8021, DOI: 10.3389/fgene.2019.00166, URL: https://www.frontiersin.org/article/10.3389/fgene.2019.00166/full (visited on 06/14/2019).

[Hua+19b] Zhi Huang et al., *TSUNAMI: Translational Bioinformatics Tool Suite For Network Analysis And Mining*, en, preprint, Bioinformatics, Sept. 2019, DOI: 10.1101/787507, URL: http://biorxiv.org/lookup/doi/10.1101/787507 (visited on 07/07/2022).

[Ins] National Cancer Insitute, *mRNA Analysis Pipeline*, URL: https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/.

[IS15] Sergey Ioffe and Christian Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift", *in*: *arXiv:1502.03167 [cs]* (Mar. 2015), arXiv: 1502.03167, URL: http://arxiv.org/abs/1502.03167 (visited on 03/25/2021).

[Ish+08] Hemant Ishwaran et al., "Random survival forests", en, *in*: *The Annals of Applied Statistics* 2.*3* (Sept. 2008), pp. 841–860, ISSN: 1932-6157, DOI: 10.1214/08-AOAS169, URL: http://projecteuclid.org/euclid.aoas/1223908043 (visited on 06/20/2019).

[Kas] Alboukadel Kassambara, *Statistical tools for high-throughput data analysis*, URL: http://www.sthda.com/english/wiki/survival-analysis-basics#log-rank-test-comparing-survival-curves-survdiff.

[Kat+18] Jared L. Katzman et al., "DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network", en, *in*: *BMC Medical Research Methodology* 18.*1* (Dec. 2018), ISSN: 1471-2288, DOI: 10.1186/s12874-018-0482-1, URL: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-018-0482-1 (visited on 06/20/2019).

[KB17] Diederik P. Kingma and Jimmy Ba, "Adam: A Method for Stochastic Optimization", *in*: *arXiv:1412.6980 [cs]* (Jan. 2017), arXiv: 1412.6980, URL: http://arxiv.org/abs/1412.6980 (visited on 03/26/2021).

[KB19] Håvard Kvamme and Ørnulf Borgan, "Continuous and Discrete-Time Survival Prediction with Neural Networks", en, *in*: *Lifetime Data Analysis 27* (Oct. 2019), arXiv: 1910.06724, URL: http://arxiv.org/abs/1910.06724 (visited on 05/22/2020).

[KBS19] Håvard Kvamme, Ørnulf Borgan, and Ida Scheel, "Time-to-Event Prediction with Neural Networks and Cox Regression", en, *in*: *Journal of Machine Learning Research* (Sept. 2019), arXiv: 1907.00825, URL: http://arxiv.org/abs/1907.00825 (visited on 10/26/2020).

[Kim+15] Dokyoon Kim et al., "Predicting censored survival data based on the interactions between meta-dimensional omics data in breast cancer", en, *in*: *Journal of Biomedical Informatics* 56 (Aug. 2015), pp. 220–228, ISSN: 15320464, DOI: 10.1016/j.jbi.2015.05.019, URL: https://linkinghub.elsevier.com/retrieve/pii/S1532046415001070 (visited on 12/07/2018).

[Kim+19] Daehwan Kim et al., "Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype", en, *in*: *Nature Biotechnology* 37.*8* (Aug. 2019), pp. 907–915, ISSN: 1087-0156, 1546-1696, DOI: 10.1038/s41587-019-0201-4, URL: http://www.nature.com/articles/s41587-019-0201-4 (visited on 04/01/2022).

[Kin+17] Pieter-Jan Kindermans et al., "Learning how to explain neural networks: PatternNet and PatternAttribution", *in*: *arXiv:1705.05598 [cs, stat]* (Oct. 2017), arXiv: 1705.05598, URL: http://arxiv.org/abs/1705.05598 (visited on 10/09/2021).

[Kla+17] Günter Klambauer et al., "Self-Normalizing Neural Networks", *in*: *arXiv:1706.02515 [cs, stat]* (Sept. 2017), arXiv: 1706.02515, URL: http://arxiv.org/abs/1706.02515 (visited on 02/06/2021).

[KM58] E. L. Kaplan and Paul Meier, "Nonparametric Estimation from Incomplete Observations", en, *in*: *Journal of the American Statistical Association* 53.*282* (June 1958), pp. 457–481, ISSN: 0162-1459, 1537-274X, DOI: 10.1080/01621459.1958.10501452, URL: http://www.tandfonline.com/doi/abs/10.1080/01621459.1958.10501452 (visited on 06/16/2022).

[Kou+15] Konstantina Kourou et al., "Machine learning applications in cancer prognosis and prediction", en, *in*: *Computational and Structural Biotechnology Journal* 13 (2015), pp. 8–17, ISSN: 20010370, DOI: 10.1016/j.csbj.2014.11.005, URL: https://linkinghub.elsevier.com/retrieve/pii/S2001037014000464 (visited on 12/20/2018).

[KW14] Diederik P. Kingma and Max Welling, "Auto-Encoding Variational Bayes", *in*: *arXiv:1312.6114 [cs, stat]* (May 2014), arXiv: 1312.6114, URL: http://arxiv.org/abs/1312.6114 (visited on 01/09/2020).

[Lan+20] Simon P. Langdon et al., "Estrogen Signaling and Its Potential as a Target for Therapy in Ovarian Cancer", en, *in*: *Cancers* 12.*6* (June 2020), p. 1647, ISSN: 2072-6694, DOI: 10.3390/cancers12061647, URL: https://www.mdpi.com/2072-6694/12/6/1647 (visited on 06/10/2022).

[LH08] Peter Langfelder and Steve Horvath, "WGCNA: an R package for weighted correlation network analysis", en, *in*: *BMC Bioinformatics* 9.*1* (Dec. 2008), p. 559, ISSN: 1471-2105, DOI: 10.1186/1471-2105-9-559, URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-559 (visited on 07/08/2022).

[LH19] Ilya Loshchilov and Frank Hutter, "Decoupled Weight Decay Regularization", *in*: *arXiv:1711.05101 [cs, math]* (Jan. 2019), arXiv: 1711.05101, URL: http://arxiv.org/abs/1711.05101 (visited on 03/25/2021).

[LHA14] Michael I Love, Wolfgang Huber, and Simon Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2", en, *in*: *Genome Biology* 15.*12* (Dec. 2014), ISSN: 1474-760X, DOI: 10.1186/s13059-014-0550-8, URL: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8 (visited on 05/14/2020).

[Lib+15] Arthur Liberzon et al., "The Molecular Signatures Database Hallmark Gene Set Collection", en, *in*: *Cell Systems* 1.*6* (Dec. 2015), pp. 417–425, ISSN: 24054712, DOI: 10.1016/j.cels.2015.12.004, URL: https://linkinghub.elsevier.com/retrieve/pii/S2405471215002185 (visited on 02/04/2022).

[Liu+18] Jianfang Liu et al., "An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics", en, *in*: *Cell* 173.*2* (Apr. 2018), 400–416.e11, ISSN: 00928674, DOI: 10.1016/j.cell.2018.02.052, URL: https://linkinghub.elsevier.com/retrieve/pii/S0092867418302290 (visited on 06/07/2019).

[LSS14] Y. Liao, G. K. Smyth, and W. Shi, "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features", en, *in*: *Bioinformatics* 30.*7* (Apr. 2014), pp. 923–930, ISSN: 1367-4803, 1460-2059, DOI: 10.1093/bioinformatics/btt656, URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt656 (visited on 04/01/2022).

[Lu+22] Weihong Lu et al., "Immune Subtypes Characterization Identifies Clinical Prognosis, Tumor Microenvironment Infiltration, and Immune Response in Ovarian Cancer", *in*: *Frontiers in Molecular Biosciences* 9 (Mar. 2022), p. 801156, ISSN: 2296-889X, DOI: 10.3389/fmolb.2022.801156, URL: https://www.frontiersin.org/articles/10.3389/fmolb.2022.801156/full (visited on 05/13/2022).

[LZY18] Changhee Lee, William Zame, and Jinsung Yoon, "DeepHit: A Deep Learning Approach to Survival Analysis with Competing Risks", en, *in*: *Thirty-Second AAAI Conference on Artificial Intelligence* (2018), p. 8.

[Mar14] Ruben Martinez-Cantin, "BayesOpt: A Bayesian Optimization Library for Nonlinear Optimization, Experimental Design and Bandits", en, *in*: (Nov. 2014), p. 5.

[Mar18] Camille Marchet, "From reads to transcripts: de novo methods for the analysis of transcriptome second and third generation sequencing.", Anglais, Bioinformatics [q-bio.QM], Université Rennes 1, 2018.

[Men+21a] Elena Spirina Menand et al., “Gene expression RNA-sequencing survival analysis of high-grade serous ovarian carcinoma: a comparative study”, *in*: *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Houston, TX, USA: IEEE, Dec. 2021, pp. 3414–3419, ISBN: 978-1-66540-126-5, DOI: 10.1109/BIBM52615.2021.9669562, URL: https://ieeexplore.ieee.org/document/9669562/ (visited on 05/13/2022).

[Men+21b] Elena Spirina Menand et al., “Predicting Clinical Outcomes of Ovarian Cancer Patients: Deep Survival Models and Transfer Learning”, en, *in*: *Proceedings of the 31st European Safety and Reliability Conference (ESREL 2021)*, Research Publishing Services, 2021, pp. 371–375, ISBN: 978-981-18201-6-8, DOI: 10.3850/978-981-18-2016-8_505-cd, URL: https://rpsonline.com.sg/proceedings/9789811820168/html/505.xml (visited on 10/01/2021).

[MN22] Ruairi J Mackenzie and Technology Networks, “RNA-Seq: Basics, Applications and Protocol”, en, *in*: (Apr. 2022), p. 10.

[Mot+08] Alison A. Motsinger-Reif et al., “Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology”, en, *in*: *Genetic Epidemiology* 32.*4* (May 2008), pp. 325–340, ISSN: 07410395, 10982272, DOI: 10.1002/gepi.20307, URL: https://onlinelibrary.wiley.com/doi/10.1002/gepi.20307 (visited on 07/08/2022).

[Nel72] Wayne Nelson, “Theory and Applications of Hazard Plotting for Censored Failure Data”, en, *in*: *Technometrics* 14.*4* (Nov. 1972), pp. 945–966, ISSN: 0040-1706, 1537-2723, DOI: 10.1080/00401706.1972.10488991, URL: http://www.tandfonline.com/doi/abs/10.1080/00401706.1972.10488991 (visited on 09/17/2022).

[Ngu20] Thi Ngoc Ha Nguyen, “Combining machine learning and reference-free transcriptome analysis for the identification of prostate cancer signatures”, Anglais, Bioinformatics [q-bio.QM], Université Paris-Saclay, 2020.

[NH10] Vinod Nair and Geoffrey E Hinton, “Rectified Linear Units Improve Restricted Boltzmann Machines”, en, *in*: *Proceedings of the 27th International Conference on Machine Learning (ICML-10),* 2010, p. 8.

[PCG18] Olivier B. Poirion, Kumardeep Chaudhary, and Lana X. Garmire, "Deep Learning data integration for better risk stratification models of bladder cancer", eng, *in*: *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science* 2017 (2018), pp. 197–206, ISSN: 2153-4063.

[Pla] Plateforme transcriptome IFR 88, *Qu'est ce que les puces à ADN ?*, fr, URL: https://www.imm.cnrs.fr/transcriptome/spip.php?rubrique11.

[Plo+21] Phyllis van der Ploeg et al., "The effectiveness of monotherapy with PI3K/AKT/mTOR pathway inhibitors in ovarian cancer: A meta-analysis", en, *in*: *Gynecologic Oncology* 163.*2* (Nov. 2021), pp. 433–444, ISSN: 00908258, DOI: 10.1016/j.ygyno.2021.07.008, URL: https://linkinghub.elsevier.com/retrieve/pii/S0090825821005382 (visited on 09/07/2022).

[Pok+19] Ruchika Pokhriyal et al., "Chemotherapy Resistance in Advanced Ovarian Cancer Patients", en, *in*: *Biomarkers in Cancer* 11 (Jan. 2019), p. 1179299X1986081, ISSN: 1179-299X, 1179-299X, DOI: 10.1177/1179299X19860815, URL: http://journals.sagepub.com/doi/10.1177/1179299X19860815 (visited on 04/21/2022).

[SBV18] Noor Pratap Singh, Raju S. Bapi, and P.K. Vinod, "Machine learning models to predict the progression from early to late stages of papillary renal cell carcinoma", en, *in*: *Computers in Biology and Medicine* 100 (Sept. 2018), pp. 92–99, ISSN: 00104825, DOI: 10.1016/j.compbiomed.2018.06.030, URL: https://linkinghub.elsevier.com/retrieve/pii/S0010482518301781 (visited on 12/20/2018).

[Sch+09] Claudia Scholl et al., "Synthetic Lethal Interaction between Oncogenic KRAS Dependency and STK33 Suppression in Human Cancer Cells", en, *in*: *Cell* 137.*5* (May 2009), pp. 821–834, ISSN: 00928674, DOI: 10.1016/j.cell.2009.03.017, URL: https://linkinghub.elsevier.com/retrieve/pii/S009286740900316X (visited on 02/04/2022).

[Sim+11] Noah Simon et al., "Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent", en, *in*: *Journal of Statistical Software* 39.*5* (2011), ISSN: 1548-7660, DOI: 10.18637/jss.v039.i05, URL: http://www.jstatsoft.org/v39/i05/ (visited on 01/25/2019).

[SMJ20] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal, "Cancer statistics, 2020", en, *in*: *CA: A Cancer Journal for Clinicians* 70.*1* (Jan. 2020), pp. 7–30, ISSN: 0007-9235, 1542-4863, DOI: 10.3322/caac.21590, URL: https://onlinelibrary.wiley.com/doi/abs/10.3322/caac.21590 (visited on 06/02/2022).

[Sri+14] Nitish Srivastava et al., "Dropout: A Simple Way to Prevent Neural Networks from Overfitting", en, *in*: *Journal of Machine Learning Research* (June 2014), p. 30.

[Sub+05] Aravind Subramanian et al., "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles", en, *in*: *Proceedings of the National Academy of Sciences* 102.*43* (Oct. 2005), pp. 15545–15550, ISSN: 0027-8424, 1091-6490, DOI: 10.1073/pnas.0506580102, URL: https://pnas.org/doi/full/10.1073/pnas.0506580102 (visited on 03/11/2022).

[Tan+15] Jie Tan et al., "Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders", eng, *in*: *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (2015), pp. 132–143, ISSN: 2335-6936.

[TGF90] T. M. Therneau, P. M. Grambsch, and T. R. Fleming, "Martingale-based residuals for survival models", en, *in*: *Biometrika* 77.*1* (Mar. 1990), pp. 147–160, ISSN: 0006-3444, 1464-3510, DOI: 10.1093/biomet/77.1.147, URL: https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/77.1.147 (visited on 07/08/2022).

[Tot+08] R. W. Tothill et al., "Novel Molecular Subtypes of Serous and Endometrioid Ovarian Cancer Linked to Clinical Outcome", en, *in*: *Clinical Cancer Research* 14.*16* (Aug. 2008), pp. 5198–5208, ISSN: 1078-0432, 1557-3265, DOI: 10.1158/1078-0432.CCR-08-0196, URL: http://clincancerres.aacrjournals.org/cgi/doi/10.1158/1078-0432.CCR-08-0196 (visited on 02/21/2020).

[Tré+20] Brigitte Trétarre et al., *Survie des personnes atteintes de cancer en France métropolitaine 1989-2018 – Ovaire*, fr, tech. rep., Boulogne-Billancourt, Sept. 2020, p. 10, URL: https://www.e-cancer.fr.

[Tur+21] Margherita Turinetto et al., "The Role of PARP Inhibitors in the Ovarian Cancer Microenvironment: Moving Forward From Synthetic Lethality", *in*: *Frontiers in Oncology* 11 (June 2021), p. 689829, ISSN: 2234-943X, DOI: 10.3389/fonc.2021.689829, URL: https://www.frontiersin.org/articles/10.3389/fonc.2021.689829/full (visited on 09/02/2022).

[Ver+12] Roel G.W. Verhaak et al., "Prognostically relevant gene signatures of high-grade serous ovarian carcinoma", en, *in*: *Journal of Clinical Investigation* (Dec. 2012), ISSN: 0021-9738, DOI: 10.1172/JCI65833, URL: http://www.jci.org/articles/view/65833 (visited on 02/21/2020).

[Vik18] Mored Vikas, *Machine learning Model Validation techniques*, https://moredvikas.wordpress.com/2018/10/10/machine-learning-model-validation-techniques/, Oct. 2018.

[Vin+08] Pascal Vincent et al., "Extracting and composing robust features with denoising autoencoders", en, *in*: *Proceedings of the 25th international conference on Machine learning - ICML '08*, Helsinki, Finland: ACM Press, 2008, pp. 1096–1103, ISBN: 978-1-60558-205-4, DOI: 10.1145/1390156.1390294, URL: http://portal.acm.org/citation.cfm?doid=1390156.1390294 (visited on 07/15/2022).

[Wan+18] Zhongyu Wang et al., "Construction of immune-related risk signature for renal papillary cell carcinoma", en, *in*: *Cancer Medicine* (Dec. 2018), ISSN: 20457634, DOI: 10.1002/cam4.1905, URL: http://doi.wiley.com/10.1002/cam4.1905 (visited on 12/14/2018).

[Way+16] Gregory P. Way et al., "Comprehensive Cross-Population Analysis of High-Grade Serous Ovarian Cancer Supports No More Than Three Subtypes", en, *in*: *G3: Genes|Genomes|Genetics* 6.*12* (Dec. 2016), pp. 4097–4103, ISSN: 2160-1836, DOI: 10.1534/g3.116.033514, URL: http://g3journal.org/lookup/doi/10.1534/g3.116.033514 (visited on 04/14/2020).

[WG17] Gregory P. Way and Casey S. Greene, "Evaluating deep variational autoencoders trained on pan-cancer gene expression", *in*: *arXiv:1711.04828 [q-bio]* (Nov. 2017), arXiv: 1711.04828, URL: http://arxiv.org/abs/1711.04828 (visited on 02/23/2020).

[WG18] Gregory P Way and Casey S Greene, "Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders", en, *in*: (2018), p. 15.

[Whi94] Darrell Whitley, "A genetic algorithm tutorial", en, *in*: *Statistics and Computing* 4.*2* (June 1994), ISSN: 0960-3174, 1573-1375, DOI: 10.1007/BF00175354, URL: http://link.springer.com/10.1007/BF00175354 (visited on 07/21/2022).

[WLR17] Ping Wang, Yan Li, and Chandan K. Reddy, "Machine Learning for Survival Analysis: A Survey", *in*: *arXiv:1708.04649 [cs, stat]* (Aug. 2017), arXiv: 1708.04649, URL: http://arxiv.org/abs/1708.04649 (visited on 12/18/2019).

[Wu+21] Tianzhi Wu et al., "clusterProfiler 4.0: A universal enrichment tool for interpreting omics data", en, *in*: *The Innovation* 2.*3* (Aug. 2021), p. 100141, ISSN: 26666758, DOI: 10.1016/j.xinn.2021.100141, URL: https://linkinghub.elsevier.com/retrieve/pii/S2666675821000667 (visited on 12/09/2021).

[Xia+18] Yawen Xiao et al., "A semi-supervised deep learning method based on stacked sparse auto-encoder for cancer prediction using RNA-seq data", en, *in*: *Computer Methods and Programs in Biomedicine* 166 (Nov. 2018), pp. 99–105, ISSN: 01692607, DOI: 10.1016/j.cmpb.2018.10.004, URL: https://linkinghub.elsevier.com/retrieve/pii/S0169260718304553 (visited on 07/18/2019).

[You+17] Safoora Yousefi et al., "Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models", en, *in*: *Scientific Reports* 7.*1* (Dec. 2017), ISSN: 2045-2322, DOI: 10.1038/s41598-017-11817-6, URL: http://www.nature.com/articles/s41598-017-11817-6 (visited on 10/26/2020).

[Yu+11] Chun-Nam Yu et al., "Learning Patient-Specific Cancer Survival Distributions as a Sequence of Dependent Regressors", en, *in*: *Advances in Neural Information Processing Systems 24* (2011), p. 10.

[ZH14] Jie Zhang and Kun Huang, "Normalized ImQCM: An Algorithm for Detecting Weak Quasi-Cliques in Weighted Graph with Applications in Gene Co-Expression Module Discovery in Cancers", en, *in*: *Cancer Informatics* 13s3

(Jan. 2014), CIN.S14021, ISSN: 1176-9351, 1176-9351, DOI: 10.4137/CIN.S14021, URL: http://journals.sagepub.com/doi/10.4137/CIN.S14021 (visited on 07/07/2022).

[ZH20] Xinnan Zhao and Miao He, "Comprehensive pathway-related genes signature for prognosis and recurrence of ovarian cancer", en, *in*: *PeerJ* 8 (Dec. 2020), e10437, ISSN: 2167-8359, DOI: 10.7717/peerj.10437, URL: https://peerj.com/articles/10437 (visited on 06/02/2022).

[Zha+16] Ya Zhang et al., "Improve Glioblastoma Multiforme Prognosis Prediction by Using Feature Selection and Multiple Kernel Learning", *in*: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 13.*5* (Sept. 2016), pp. 825–835, ISSN: 1545-5963, DOI: 10.1109/TCBB.2016.2551745, URL: http://ieeexplore.ieee.org/document/7448848/ (visited on 11/23/2018).

[Zha+17] Chiyuan Zhang et al., "Understanding deep learning requires rethinking generalization", *in*: *arXiv:1611.03530 [cs]* (Feb. 2017), arXiv: 1611.03530, URL: http://arxiv.org/abs/1611.03530 (visited on 03/26/2021).

[Zha+18] Li Zhang et al., "Deep Learning-Based Multi-Omics Data Integration Reveals Two Prognostic Subtypes in High-Risk Neuroblastoma", *in*: *Frontiers in Genetics* 9 (Oct. 2018), ISSN: 1664-8021, DOI: 10.3389/fgene.2018.00477, URL: https://www.frontiersin.org/article/10.3389/fgene.2018.00477/full (visited on 06/20/2019).

[Zou+20] Zhilin Zou et al., "mTOR signaling pathway and mTOR inhibitors in cancer: progress and challenges", en, *in*: *Cell & Bioscience* 10.*1* (Dec. 2020), p. 31, ISSN: 2045-3701, DOI: 10.1186/s13578-020-00396-1, URL: https://cellandbioscience.biomedcentral.com/articles/10.1186/s13578-020-00396-1 (visited on 09/07/2022).

[ZPJ20] Yuqing Zhang, Giovanni Parmigiani, and W Evan Johnson, "ComBat-seq: batch effect adjustment for RNA-seq count data", en, *in*: *NAR Genomics and Bioinformatics* 2.*3* (Sept. 2020), lqaa078, ISSN: 2631-9268, DOI: 10.1093/nargab/lqaa078, URL: https://academic.oup.com/nargab/article/doi/10.1093/nargab/lqaa078/5909519 (visited on 10/14/2021).

[Zut+21] Moniek van Zutphen et al., “Identification of Lifestyle Behaviors Associated with Recurrence and Survival in Colorectal Cancer Patients Using Random Survival Forests”, en, *in*: *Cancers* 13.*10* (May 2021), p. 2442, ISSN: 2072-6694, DOI: 10.3390/cancers13102442, URL: https://www.mdpi.com/2072-6694/13/10/2442 (visited on 09/16/2022).

Secondary sources

[Ale+13] Ludmil B. Alexandrov et al., "Deciphering Signatures of Mutational Processes Operative in Human Cancer", en, *in*: *Cell Reports* 3.*1* (Jan. 2013), pp. 246–259, ISSN: 22111247, DOI: 10.1016/j.celrep.2012.12.008, URL: https://linkinghub.elsevier.com/retrieve/pii/S2211124712004330 (visited on 12/07/2018).

[ASB15] Dvir Aran, Marina Sirota, and Atul J. Butte, "Systematic pan-cancer analysis of tumour purity", en, *in*: *Nature Communications* 6.*1* (Dec. 2015), ISSN: 2041-1723, DOI: 10.1038/ncomms9971, URL: http://www.nature.com/articles/ncomms9971 (visited on 11/15/2019).

[Che+17] Ke Chen et al., "Towards In Silico Prediction of the Immune-Checkpoint Blockade Response", en, *in*: *Trends in Pharmacological Sciences* 38.*12* (Dec. 2017), pp. 1041–1051, ISSN: 01656147, DOI: 10.1016/j.tips.2017.10.002, URL: https://linkinghub.elsevier.com/retrieve/pii/S0165614717301967 (visited on 12/28/2018).

[Jim+17] Alejandro Jiménez-Sánchez et al., "Heterogeneous Tumor-Immune Microenvironments among Differentially Growing Metastases in an Ovarian Cancer Patient", en, *in*: *Cell* 170.*5* (Aug. 2017), 927–938.e20, ISSN: 00928674, DOI: 10.1016/j.cell.2017.07.025, URL: https://linkinghub.elsevier.com/retrieve/pii/S0092867417308322 (visited on 12/07/2018).

[KY17] Zura Kakushadze and Willie Yu, "*K-means and cluster models for cancer signatures", en, *in*: *Biomolecular Detection and Quantification* 13 (Sept. 2017), pp. 7–31, ISSN: 22147535, DOI: 10.1016/j.bdq.2017.07.001, URL: https://linkinghub.elsevier.com/retrieve/pii/S2214753517302061 (visited on 12/07/2018).

[LWN16] Yifeng Li, Fang-Xiang Wu, and Alioune Ngom, "A review on machine learning principles for multi-view biological data integration", en, *in*: *Briefings in Bioinformatics* (Dec. 2016), bbw113, ISSN: 1467-5463, 1477-4054, DOI: 10.1093/bib/bbw113, URL: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbw113 (visited on 11/23/2018).

[Mar+14] Filipe C Martins et al., "Combined image and genomic analysis of high-grade serous ovarian cancer reveals PTEN loss as a common driver event and prognostic classifier", en, *in*: *Genome Biology* 15.*12* (Dec. 2014), ISSN: 1474-760X, DOI: 10.1186/s13059-014-0526-8, URL: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0526-8 (visited on 12/07/2018).

[Mou+19] Mohamed Mounir et al., "New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx", en, *in*: *PLOS Computational Biology* 15.*3* (Mar. 2019), ed. by Edwin Wang, e1006701, ISSN: 1553-7358, DOI: 10.1371/journal.pcbi.1006701, URL: http://dx.plos.org/10.1371/journal.pcbi.1006701 (visited on 05/03/2020).

[Sal+18] Joel Saltz et al., "Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images", en, *in*: *Cell Reports* 23.*1* (Apr. 2018), 181–193.e7, ISSN: 22111247, DOI: 10.1016/j.celrep.2018.03.086, URL: https://linkinghub.elsevier.com/retrieve/pii/S2211124718304479 (visited on 12/07/2018).

[Tho+18] Vésteinn Thorsson et al., "The Immune Landscape of Cancer", en, *in*: *Immunity* 48.*4* (Apr. 2018), 812–830.e14, ISSN: 10747613, DOI: 10.1016/j.immuni.2018.03.023, URL: https://linkinghub.elsevier.com/retrieve/pii/S1074761318301213 (visited on 12/07/2018).

[TTG16] Jeffrey A. Thompson, Jie Tan, and Casey S. Greene, "Cross-platform normalization of microarray and RNA-seq data for machine learning applications", en, *in*: *PeerJ* 4 (Jan. 2016), e1621, ISSN: 2167-8359, DOI: 10.7717/peerj.1621, URL: https://peerj.com/articles/1621 (visited on 03/05/2020).

[Var+17] Hebert Alberto Vargas et al., "Radiogenomics of High-Grade Serous Ovarian Cancer: Multireader Multi-Institutional Study from the Cancer Genome Atlas Ovarian Cancer Imaging Research Group", en, *in*: *Radiology* 285.*2* (Nov. 2017), pp. 482–492, ISSN: 0033-8419, 1527-1315, DOI: 10.1148/radiol.2017161870, URL: http://pubs.rsna.org/doi/10.1148/radiol.2017161870 (visited on 12/07/2018).

List of Figures

1.1 DNA and RNA alphabet bases. . . . 14
1.2 Modern molecular biology dogma. . . . 15
1.3 DNA chips or microarrays transcriptome sequencing principle. . . . 17
1.4 RNA-seq analysis workflow. . . . 19
1.5 NGS and TGS technologies comparison. . . . 19
1.6 Transcriptomic data analysis pipeline. . . . 21

2.1 KM curve of the ovarian cancer (TCGA-OV) survival data. . . . 25
2.2 Comparison of sampling techniques . . . 30
2.3 Random sampling schema . . . 30
2.4 K-fold Cross-Validation sampling schema . . . 31
2.5 Bootstrap sampling schema . . . 31
2.6 Random Survival Forest. . . . 33
2.7 MLP with 1 hidden layer. . . . 34
2.8 Cox-nnet architecture, representation of a 1-hidden layer transformation. . . 35
2.9 N-MTLR architecture, representation of a 2-hidden layer transformation. . 36
2.10 DeepHit architecture. . . . 36
2.11 RNN-Surv architecture. . . . 37

3.1 VAE schema. . . . 50
3.2 Survival analysis strategy by Bell et al. . . . 52
3.3 Survival analysis strategy by Chen et al. . . . 53
3.4 Survival analysis strategy by Zhang et al. . . . 53
3.5 Survival analysis strategy by Kim et al. . . . 54
3.6 Survival analysis strategy by Bell et al. . . . 54
3.7 Survival analysis strategy by Tan et al. . . . 55
3.8 Survival analysis strategy by Way et Greene. . . . 55
3.9 Survival analysis strategy by Chaudhary et al. . . . 56
3.10 Survival analysis strategy by Poirion et al. . . . 56

3.11 Survival analysis strategy by Zhang et al. 56
3.12 Survival analysis strategy by Wang et al. 57
3.13 Survival analysis strategy by Yousefi et al and Ching et al. 58
3.14 Survival analysis strategy by Huang et al. 58
3.15 Survival analysis strategy by Hao et al. 59
3.16 Boxplot of the C-index of the 10 TCGA datasets using the survival models by Ching et al. 60
3.17 Boxplot of the IBS of the 10 TCGA datasets using the survival models by Ching et al. 60
1 KM curves for training/test sets, fold 1 68
2 KM curves for training/test sets, fold 2 70
3 KM curves for training/test sets, fold 3 70
4 KM curves for training/test sets, fold 4 71
5 KM curves for training/test sets, fold 5 71
6 C-index for the dimensionality reduction experiments. 72
7 IBS for the dimensionality reduction experiments. 73

1.1 C-index comparison of different deep survival models 84
1.2 IBS comparison of different survival models 85

2.1 C-index comparison of transfer learning experiments 94

3.1 Transfer learning strategy for ovarian pronostication. 99
3.2 Proposed N-MTLR-Rank model architecture. 100
3.3 C-index comparison of DeepHit and our N-MTLR-Rank models. 101
3.4 IBS comparison of DeepHit and our N-MTLR-Rank models. 101
3.5 C-index comparison of our N-MTLR-Rank model generalization behavior. . 102
3.6 IBS comparison of our N-MTLR-Rank model generalization behavior. . . . 102
3.7 Kaplan-Meier curves of the high-risk and low-risk TCGA-OV groups. . . . 115

List of Tables

2.1 Confusion matrix. 29

3.1 Comparison of the AEs used with gene expression data for survival analysis. *Not specified, **Kullback–Leibler divergence 51

1.1 Comparison of ANN based survival analysis loss functions. *Proportional hazards assumption. 79
1.2 Hyperparameter search space. 83

3.1 TCGA-OV clinical descriptive statistics. 111
3.2 ICO-OV clinical descriptive statistics. 112
3.3 The overall number of significantly enriched pathways found in hallmark and C6 MSigDB collections (p-value < 0.05). 113
3.4 The most frequent hallmark pathways found for patients of the TCGA-OV and the presence of the corresponding pathways in ICO-OV dataset ($> 5\%$ at 5 years endpoint). 113
3.5 The most frequent C6 collection pathways found for patients of the TCGA-OV and the presence of the corresponding pathways in ICO-OV dataset ($> 5\%$ at 5 years endpoint). 113

Titre : La recherche de nouveaux bio-marqueurs pour les séquences thérapeutiques des cancers "pan-gyn" grâce à l'apprentissage automatique

Mot clés : Cancer de l'ovaire, TCGA, RNA-seq, analyse de survie, réseaux de neurones artificiels (ANN)

Résumé : L'expression des gènes est connue pour être associée à la survie globale chez les patients avec un cancer. L'analyse de survie pour le cancer de l'ovaire permet potentiellement non seulement la stratification des patientes mais également la recherche des nouvelles cibles thérapeutiques. Ce travail présente l'étude des techniques d'analyse de survie récentes basées sur des réseaux de neurones artificiels (ANN) et compare la performance des ces modèles sur la base des données RNA-seq des tumeurs ovariennes. Il souligne également le fait que ces modèles d'apprentissage profond sont capables de transférer la connaissance à travers le groupe "pan-gyn" dans le but d'améliorer la précision des prédictions au niveau du cancer de l'ovaire car ce groupe des cancers gynécologiques et du sein partage de caractéristiques communes. Cette thèse propose un nouveau modèle de survie basé sur l'apprentissage profond, appelé N-MTLR-Rank, ce dernier a été entrainé avec les données du TCGA (The Cancer Genome project) et validé avec un dataset indépendant. De plus, elle démontre comment ce modèle peut être apparenté à des voies biologiques en lien avec la survie des patientes avec le cancer de l'ovaire.

Title: Machine learning based novel biomarkers discovery for therapeutic use in "pan-gyn" cancers

Keywords: Ovarian cancer, TCGA, RNA-seq, survival analysis, artificial neural networks (ANN)

Abstract: Gene expression is established to be associated with overall survival in cancer patients. Survival analysis of ovarian cancer could allow not only patient stratification but possible discovery of new therapeutic targets. This work presents an overview of the recent artificial neural network (ANN) survival analysis techniques and benchmarks these models on the basis of ovarian cancer RNA-seq data. It also highlights that deep survival models could successfully transfer information across "pan-gyn" group to improve the ovarian cancer prognostic accuracy as these gynecologic and breast cancers share a variety of characteristics. This thesis proposes a new deep learning survival model called N-MTLR-Rank, trained using The Cancer Genome project (TCGA) data and validated on an independent datatset. Additionnaly, it demonstrates how this model can be related to the molecular pathways to uncover biological processes associated with ovarian cancer patients survival.