Optimised Combination of Distance Metrics for Labelled Trees Using Survival Data

Laura Bispo Quintas[†][*]

Supervisors: Prof. Susana Vinga[†] and Prof. Niko Beerenwinkel[‡]

Abstract

Cancer, characterized by uncontrolled cell growth and DNA mutations, presents significant treatment challenges due to intra-tumor heterogeneity, which can be represented through mutation trees. Distance metrics, capturing diverse aspects of these trees, have been used for patient clustering and associating with clinical data to identify cancer subgroups. However, a consensus on a standard metric is yet to be reached, prompting investigations into metric combinations to leverage their combined strengths. This study critically assesses optimised combinations of distance metrics, interpreting mutation trees' characteristics across Breast Cancer and Acute Myeloid Leukaemia data sets, aiming to link patient clustering with clinical outcomes, especially survival. We employ a comprehensive approach, encompassing all available distance metrics for clustering tumor trees and integrating clinical and genotypic information using the Cox model. Our findings indicate that while these metrics offer detailed insights into tumor evolution, their predictive power for survival, individually or collectively, is not significantly superior to using clinical and genotypic covariates alone. Our results effectively challenge the notion of a single distance metric for patient clustering over a variety of cancer types by showing the inconsistent way these metrics perform across various data sets. This study underscores the complexity of cancer genomics and the challenges in deriving useful information for personalised patient prognosis. It advocates for future research with standardised data sets and diverse cancer types to further explore the potential of distance metrics and to devise reliable, tailored cancer treatment strategies.

Keywords: *Cancer, Mutation Trees, Distance Metrics, Hierarchical Clustering, Survival Analysis.*

1 Introduction

Cancer's complexity stems from its uncontrolled cell growth and genetic diversity, creating a challenge for treatment as tumors evolve and develop resistance. Traditional therapies often target the most dominant subclones, but as these are eradicated, previously suppressed or new, resistant subclones may emerge, leading to relapse. Therefore, understanding cancer requires delving into the mutation sequences that drive its progression and resistance mechanisms [1, 2].

Mutation trees are a pivotal tool for this purpose, illustrating the evolutionary relationships among subclones within a tumor. By analyzing these trees, we can gain insights into the cancer's evolutionary path and the development of drug resistance. However, the task is complicated by the lack of a standardized metric that can capture the heterogeneity of cancer tumors. Numerous metrics exist, each highlighting different features of mutation trees, yet no single metric adequately addresses all aspects of cancer's variability [3–8].

Clinical research has applied these metrics to categorize patient data into groups that correlate with clinical outcomes, such as cancer subgroups. This stratification is critical because it can reveal the effectiveness of treatments and inform personalised therapy strategies. Some researchers have proposed combining multiple metrics to better represent the complexity of tumor evolution [6]. This approach is promising as it may lead to a more nuanced understanding of how different subclones within a tumor interact and respond to therapy.

Our research is centered on optimising the use of these combined metrics to correlate more closely with survival outcomes. The goal extends beyond mere data clustering; it is to establish a robust connection between the clusters identified by our metrics and significant clinical outcomes, particularly survival. By doing so, we aim to provide a more detailed and clinically relevant map of how cancer evolves and how it can be more effectively treated, with the ultimate goal of improving patient prognosis.

2 Background

2.1 Genetics of Cancer and its Evolution

Cancer, marked by uncontrolled cell division, leads to abnormal growths that can form tumors. These can be benign (non-cancerous) or malignant (cancerous), with the latter capable of invading tissues and spreading, or metastasizing, to other parts of the body. Malignant tumors disrupt normal bodily functions and are challenging to treat [1, 2].

Cancer research has revealed that the disease involves dynamic genomic changes, contributing to its progression and the emergence of diverse cancer hallmarks. The clonal theory suggests that tumors arise through an evolutionary process, with genetic mutations providing certain cells a proliferative edge. These alterations can affect critical cellular processes, influencing cancer development [4, 9].

Intra-tumor heterogeneity, a consequence of branched evolutionary patterns, creates a complex landscape of subclones within a single tumor. This diversity complicates treatment strategies, as therapies often target the dominant subclone, potentially overlooking the multitude of other subclones that can contribute to disease progression and treatment resistance [9, 10].

At diagnosis, treatments aim to eradicate the primary subclone. However, following remission, suppressed or new, resistant subclones can proliferate, leading to relapse. Understanding the genetic diversity and evolutionary trajec-

[*] E-mail: laura.quintas@tecnico.ulisboa.pt

[†] Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal

[‡] ETH-DBSSE, Basel, Switzerland

tories of tumors is crucial for developing effective, personalized cancer therapies that can adapt to or counteract this heterogeneity, improving patient outcomes [3, 11].

2.2 Labelled Tumor Trees

Intra-tumor heterogeneity can be depicted through tumor trees, which visualize the evolutionary lineage of subclones within a tumor [4]. These trees serve as rooted, directed graphs, where each node represents a subclone characterized by specific mutations. Phylogenetic trees show actual tumor samples, with leaves corresponding to the sampled subclones and internal nodes marking mutations [5]. Mutation trees, a type of clonal tree with the finest granularity, label each vertex with distinct mutations, indicating the clone where the mutation first appeared [4].

Advancements in DNA sequencing, from bulk to single-cell technologies, have improved our ability to construct detailed evolutionary trees [12]. Bulk sequencing offers precise mutation prevalence but can miss low-frequency mutations, whereas single-cell sequencing, despite being noisier, eliminates the need for data deconvolution and captures a tumor's full mutational diversity [13, 14]. The latter's detailed mutation histories are vital for personalized cancer therapy, as they can help identify recurring mutation patterns across tumors, offering insights into cancer development and informing treatment approaches [7, 8]. Understanding the sequence of mutations, for example, the order of JAK2 and TET2 mutations in myeloproliferative neoplasms, is crucial for assessing tumor risk and guiding clinical decisions [3].

2.3 Distance Metrics for Labelled Trees

In the context of tumor evolution inference, it is imperative to have a means of quantifying the similarity between various inferred tumor histories. This is essential for tasks such as benchmarking new inference methods, deducing consensus histories, assessing the uncertainty in Bayesian methods, and identifying common patterns of tumor evolution across different patients. Understanding the structure and terminology of trees is fundamental to comprehending the concept of mutation trees distance measures, which typically involve trees with labeled mutations that indicate distinct tumor clones [7, 15, 16].

The distance metrics for comparing labeled trees take into account both the tree topology (the hierarchical structure that connects the nodes) and the labels (mutations) on the vertices. Traditional phylogenetic tree distance measures are not well-suited for mutation trees because mutation trees uniquely label both internal nodes and leaves. As a result, specialized measures have been developed. However, there is currently no single standard metric that has been universally adopted for comparing labeled trees in tumor evolution [5, 17].

The specialized distance metrics include:

- **Parent-Child Distance (PCD)**: This metric evaluates the dissimilarity between two trees by looking at the differences in their parent-child mutation relationships [7].
- **Ancestor-Descendant Distance (AD)**: This measure quantifies the distance between trees by counting the ancestor-descendant relationships that are present in one tree but not in the other [7].
- **Path Distance (PD)**: This distance compares trees by summing the differences in the lengths of paths between pairs of mutations in the two trees [7].
- **Clonal Distance (CD)**: This metric determines the difference between two trees by considering the unique sets of clones, which represent all mutations from the root to a given node, between the two trees [7].
- **Multi-Labeled Tree Dissimilarity (MLTD)**: This is an edit distance-based metric that calculates the minimum number of modifications required to transform two trees into one common tree, reflecting differences in mutation inheritance and nodes that are multi-labeled [4].
- **Common Ancestor Set distance (CASet)** and **Distinctly Inherited Set Comparison distance (DISC)**: These metrics offer more nuanced approaches to comparing trees, with CASet focusing on the common ancestors of mutations and DISC on the mutations that are distinctly inherited [8].
- **Triplet-based Similarity Score (MP3)**: This advanced metric accounts for the recurrence and loss of mutations in tumor trees, offering a measure of the shared and distinct evolutionary paths of tumors [6].
- **Bourque Distance (BD) and k-Bourque Distance (kBD)**: Measures the topological dissimilarity between trees with different label sets, based on the idea of edge contractions [5].

These metrics cover various aspects of labeled tree comparison, from the relationships between mutations to the overall structure, mutation inheritance, and even the violation of infinite sites assumption in tumor evolution. They are critical tools for the advancement of the field of tumor evolution inference, enabling the evaluation of new algorithms and the understanding of tumor evolution patterns.

2.4 Distance Combination and Optimisation

Ciccolella et al. [6] developed the MP3 distance measure, blending two versions to enhance robustness and applicability, setting a standard for combining metrics. This synergy suggests that integrating elements from various metrics can forge a comprehensive measure, expanding possibilities in metric development and optimisation.

Optimisation algorithms, pivotal in computational fields, iteratively seek solutions to maximise or minimise a target function. They include gradient-based methods for smooth problems and metaheuristics for complex, derivative-free scenarios. Traditional methods like linear, dynamic programming, and Newton's methods face challenges such as dependency on gradients and local optima entrapment, limiting their effectiveness in intricate engineering problems [18, 19].

Metaheuristic search algorithms (MSAs) offer a robust alternative, drawing inspiration from natural processes to balance exploration and exploitation, aiming for global optima without stringent algorithmic modifications or gradient information. Their stochastic nature helps avoid local minima, making them suitable for complex optimisation tasks [19].

A notable MSA, the Differential Evolution (DE) algorithm, excels at finding the global minimum in landscapes populated with multiple local minima and is suitable for non-linear, non-differentiable, and multi-dimensional functions. The DE process unfolds through four main steps:

1. **Initialisation**: Random generation of potential solutions within predetermined boundaries.
2. **Mutation**: Creation of a mutant vector for each candidate solution, adjusting current population vectors with a scaling factor.
3. **Crossover**: Combination of mutant vector components with the original candidate to produce a trial vector, with a probabilistic approach guiding the mixing.
4. **Selection**: Adoption of the trial vector if it yields a better objective function value than the original.

The DE algorithm proceeds with these iterations, managed by minimal parameters for user-friendliness, until it fulfills the pre-established termination criteria, such as a maximum iteration count or satisfactory convergence. This ensures that only the most optimal solutions are retained [19, 20].

2.5 Clustering in Biological Data Analysis

Distance metrics for mutation trees have shown promise in cancer research by clustering trees to correlate with clinical outcomes, thus improving treatment strategies and understanding genetic mutations [6]. Clustering algorithms, which group data based on similarity, rely on measures like hierarchical clustering that uses pairwise distance metrics for tree similarity. This technique, which starts with each tree as a separate cluster and merges them, can employ different linkage methods to compare clusters [21].

Hierarchical clustering requires no assumptions about data distribution and can use any data type. The silhouette score evaluates clustering quality, with values ranging from -1 to 1, indicating the degree of cohesion within clusters. Advances in distance metric learning have led to improved methods for assessing data point similarity, crucial for clustering mutation trees and interpreting genetic mutation patterns [22].

A study applying these measures to real data clustered 36 medulloblastoma patients, using the SCITE method to infer cancer phylogenies. Hierarchical clustering based on these phylogenies, particularly with the MP3 measure, successfully grouped patients into subtypes, underscoring the potential of distance metrics in patient clustering. However, to fully grasp their value, further research with varied clinical data is essential [6].

2.6 Survival Analysis in Cancer Research

Survival analysis has gained significant attention in cancer research due to its role in assessing treatment outcomes and predicting patient survival. It involves the study of the time duration until an event of interest, such as death or relapse [23].

The Kaplan-Meier Estimation and the Cox Proportional Hazards Model are central methodologies in survival analysis. The Kaplan-Meier method provides survival function estimates and is formulated as:

$$\widehat{S}(t) = \prod_{t(i)\leq t} \frac{n_i - d_i}{n_i}, \tag{1}$$

where $\widehat{S}(t)$ represents the estimated survival probability, n_i the number at risk, and d_i the number of events at time $t(i)$ [24].

The Cox Model, which allows for covariate inclusion, is represented by the hazard function:

$$h(t \mid X_i) = h_0(t)\exp\left(X_i'\beta\right), \tag{2}$$

with $h_0(t)$ as the baseline hazard, β the covariate coefficients and where X_i, with $i = 1, \dots, n$, is the profile of the patient i over P covariates (clinical data, mutations and clusters), $X_i' = (X_{i1}, \dots, X_{iP})$ [25].

Regularisation techniques like Lasso and Ridge are used to enhance the model's performance, particularly in the context of high-dimensional data. Group Lasso is of interest for handling grouped covariate structures, leading to an optimization problem for the Cox model:

$$\min_{\beta} -l(\beta, h_0) + \lambda \sum_{l=1}^{G} \sqrt{p_l}\|\beta^{(l)}\|_2, \tag{3}$$

where $l(\beta, h_0)$ is the likelihood function the Cox Model tries to minimise, G is the total number of groups, λ is the regularisation constant and p_l is the number of covariates in group l [26].

2.7 Models Comparison and Assessment

Model evaluation is critical in survival analysis to validate model generalisability and compare different models. It involves dividing data into training and test sets, with k-fold cross-validation employed during training for a comprehensive evaluation and parameter optimisation [27].

For overall assessment, the log-rank test compares survival distributions, analysing differences in survival probabilities over time, while the Cox model's hazard ratios are examined for efficacy. This test's chi-squared statistic assesses the significance of survival curve differences between risk groups, with adjustments for multiple testing through methods like the Bonferroni Correction to avoid false positives [28, 29].

The Likelihood Ratio Test (LLR), comparing nested models, it determines if the additional parameters in the complex model significantly improve the fit to the data, and the Akaike Information Criterion (AIC), balancing model complexity and fit, are also used [30, 31]. The Prognostic Index (PI), calculated from individual covariates and model coefficients, stratifies individuals into risk categories, assisting in the evaluation of survival differences [32].

The C-Index further appraises a model's predictive accuracy by measuring its discriminatory capability, considering censoring and assessing the concordance between predicted risks and observed survival times [33].

These methods together provide a comprehensive framework for evaluating survival models, ensuring the robustness and credibility of survival analysis in research.

3 Implementation

Breast Cancer Data

We utilised tumor trees processed and deduced by Luo et al. [34], drawn from the breast cancer data in [35], bulk sequencing data, and employing phylogenetic trees from SPRUCE [36]. The preprocessing completed by Luo et al. [34]confined the analysis to mutations observed in a minimum of 10% of patients. This resulted in 19 mutations across 1152 patients with 1232 phylogenetic trees, where trees from identical patient samples were given equal weight.

The dataset's clinical data encompassed patient details on age, tumor grade and stage, hormone receptor status, vital status, and overall survival in months.

Acute Myeloid Leukemia Data

Based on the data from Morita et al. [37], which included 154 samples from 123 AML patients with discernible somatic mutations via scDNA-seq, containing a total of 31 mutated genes, we used tumor trees processed by Luo et al. [34]. Our study's tumor trees were inferred through SCITE [15], a tool that leverages Bayesian inference to generate mutation trees aligned with observed single-cell genotypes. These trees were special due to the fact that they allowed parallel mutations, thereby violating the ISA. Nodes that were repeated in the tree became clonal nodes, which meant they included ancestral mutations. On the other hand, nodes that shared the last mutation and had a Jaccard similarity greater than 50% experienced mutation shifts to already existing mutations. The number of distinct mutations increased from the initial set of 31 to 53 as a result of this clonal node modification.

Clinical data encompassed details like age, gender, maxCCF, treatment classifications (Tx_group), specific AML treatment regimens, AML subtypes (Diagnosis), vital status, and overall survival in days calculated from the difference between 'Service Date' and 'Last Contact Date'.

Distance Metrics Implementation and Packages

Distance metrics by Govek et al. [7] and Jahn et al. [5] were implemented in Python. Packages used included:

- MLTD: `https://github.com/khaled-rahman/MLTED`
- CASet and DISC metrics: `https://bitbucket.org/oesperlab/stereodist`
- MP3 for mutation trees: `https://github.com/AlgoLab/mp3treesim`

Clustering Techniques and Packages

Hierarchical clustering was conducted using `scipy.cluster.hierarchy`. The Ward method optimised the total within-cluster variance. To identify the optimal cluster number in our analysis, we adapted the silhouette score from `sklearn.metrics` into a weighted silhouette score (WSS) and incorporated cluster size considerations, as defined by:

$$WSS = \sum_{j=1}^{K} \left(\frac{n_j}{N} \cdot \text{median}\{s(i) : i \in C_j\} \right), \tag{4}$$

where K is the cluster count, n_j the number in the j-th cluster, N the total patient number, C_j the j-th cluster sample set, and $s(i)$ the i-th sample silhouette score. We then smoothed the WSS differences across clusters using LOWESS regression via `statsmodels`, defining the optimal cluster number k^* as the lesser of k_{LOWESS}, where the smoothed WSS change fell below a threshold, and k_{WSS}, the cluster number with the highest WSS.

Cluster similarity was assessed using the Jaccard index with the aid of an online tool (`http://www.comparingpartitions.info/`).

Survival Models and Regularisation

The Cox Proportional Hazards Model was implemented using the Python `lifelines` library, with Group Lasso regularisation performed through the R package `grpreg`.

Model Comparison and Assessment

Model evaluation involved `lifelines` for the computation of pAIC and pLog-Likelihood, with Bonferroni correction for multiple comparisons. The Concordance Index (C-Index) and Prognostic Index (PI) were also calculated using `lifelines`.

Optimised Weight Combination

We employed the Differential Evolution (DE) algorithm from `scipy.optimize`. The core parameters of the DE algorithm in our study were the scaling factor F, set between 0.5 and 1, which controls the differential mutation's amplification, and the crossover probability CR, set at 0.7, dictating the likelihood of mutant vector components being chosen. We employed the 'randtobest1bin' strategy alongside the 'Sobol' initialization method, enhancing DE's efficiency by directing mutations using the best current solution and ensuring a diverse initial population distribution.

The `differential_evolution` function required the objective function and bounds as inputs, which were provided as [0,1] for all metrics, and aimed at minimizing the pAIC from the Cox model. The output included the lowest pAIC and the optimal weights for the metrics

The code with the functions used is available at: `https://github.com/laurabquintas/CombinedDistances`.

4 Results and Discussion

4.1 Breast Cancer Dataset Analysis

Mutation trees, refined to include mutations present in at least 10% of our patient cohort, resulted in an average of 2.9 nodes per tree across 1232 trees for 1152 patients. Nineteen distinct mutations were highlighted, including genes previously associated with varying risks of breast cancer. Distance metrics such as DISC, CASet, MLTD, BD, 1BD, 2BD, PD, PCD, AD, and CD were applied, generating a 1052x1052 distance matrix per metric. However, the MP3 metric produced uniform outputs of distances equal to 1 due to small tree sizes and was excluded from analysis. The distribution of distance metrics indicated a strong potential for distinct clustering (Fig. 1 (a)).

Hierarchical clustering was conducted with the Ward linkage method emerging as the most suitable due to its high silhouette scores compared to other methods. Cluster maps and dendrograms were utilised for visual assessment, revealing varying levels of cluster cohesion (Fig. 1 (b)).

The silhouette score's increasing trend with the addition of clusters posed a challenge for determining the optimal cluster count. A modified silhouette score was introduced to address the impact of smaller clusters. LOWESS smoothing was applied to identify a threshold where the score stabilized, leading to a balanced number of clusters (Fig. 1 (c)). The patient distribution across clusters was considered satisfactory, avoiding overly small clusters.

Jaccard index calculations showed a high correlation among clusters from different metrics (Fig. 2). No significant correlation was found between clusters and clinical or genotype data, leading to further genotype-cluster association analysis. Mutation distribution analysis within

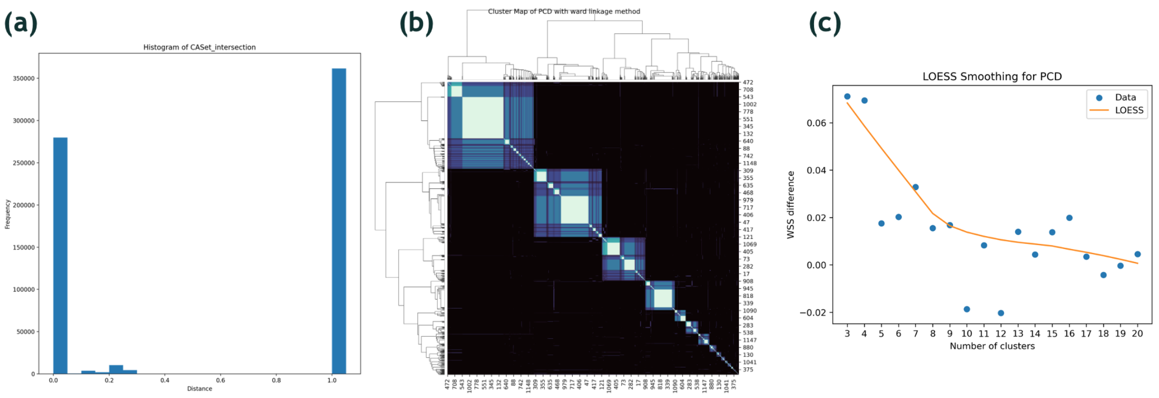

Figure 1: Analysis of Distance Metrics in Breast Cancer Data. (a) Histograms depicting the distribution of distances for the CASet ∩ metric. (b) Cluster map visualising pairwise distances for the PCD metric, ranging from dark (distance = 1) to light (distance = 0), accompanied by corresponding dendrograms. (c) Application of LOWESS regression to the WSS differences to optimise the number of clusters for PCD.

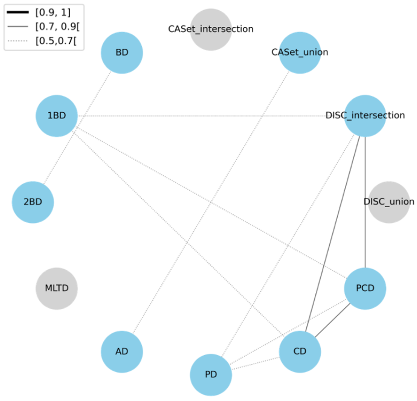

Figure 2: (Jaccard index scores assessing the similarity between clusters formed by different distance metrics.

the clusters revealed correlations between mostly one mutation and the respective cluster, some metrics, such as PCD, were able to uncover co-occurences of two mutations, *TP53* and *PIK3CA*. Additionally, clusters characterised by heterogeneous mutations were identified, explaining the lower silhouette scores previously observed.

Survival Analysis Using Cox Models

We conducted a survival analysis to evaluate if patient clusters, defined by clinical and genetic data, are significant predictors of survival. The study used 'Overall_Tumor_Grade' and 'Invasive_Carcinoma_Dx_Age' from the clinical data and included 19 mutations from the data set's mutation trees.

The data was partitioned into 70% for Cox model training and 30% for validation. The Cox model incorporated covariates from clinical data, genotype information, distance metrics clusters, and our clusters combinations strategies. Note that the distance metrics models include the clinical and genotype data.

Table 1 presents the pAIC and partial Log-Likelihood scores. The pAIC was preferred over pLog-Likelihood due to concerns about bias in smaller clusters. Furthermore Log-Likelihood ratio tests (LLR) indicated that distance metrics provided additional insights compared to 'Only Clinical', although no significance was shown when compared to the 'Genotype' model after Bonferroni correction.

Table 1: Cox regression model outcomes displaying pAIC and partial Log Likelihood scores for Breast Cancer training data. In the pAIC column, green indicates scores superior to the clinical pAIC, while grey denotes inferior results. For the Log Likelihood column, blue signifies scores better than the clinical value, with grey marking those that are worse.

	pAIC	**pLog-Likelihood**
Only Clinical	1795.1	-893.6
Genotype	1781.2	-867.6
DISC ∪	1783	-859.5
DISC ∩	1788.2	-860.1
CASet ∪	1775.5	-855.7
CASet ∩	1780.5	-864.2
BD	1789.6	-862.8
1BD	1787.2	-863.6
2BD	1787.2	-862.6
MLTD	1784.3	-867.1
AD	1774.8	-857.4
PD	1785.8	-863.9
CD	1786.3	-864.1
PCD	1784.4	-861.2
Optimal Combination	1780.6	-860.3
Group Lasso	1756	-857
Group Lasso CV	1769.1	-854.5

For optimal metric combination, the Differential Evolution (DE) algorithm was used, leveraging its time and memory efficiency. Results from 1,000 runs informed the selection of weights for the DE algorithm (Fig. 3 (a)), highlighting the importance of CASet ∪ and MLTD. Cluster analysis revealed a significant cluster involving *TP53* and *PIK3CA* mutations (Fig. 3 (b)), suggesting a better prognosis than *TP53* alone.

The study identified significant covariates using Group Lasso penalisation, with 'Invasive_Carcinoma_Dx_Age', 'Overall_Tumor_Grade', and mutations such as *PIK3CA*, *PIK3R1*, *RHOA*, *TP53*, and *TSC2* being prominent. The covariate CASet ∪ showed significant associations in multiple clusters, unlike CASet ∩, which had no impactful p-values, suggesting a confounding effect.

To prevent overfitting, a 'Group Lasso CV' with 5-fold

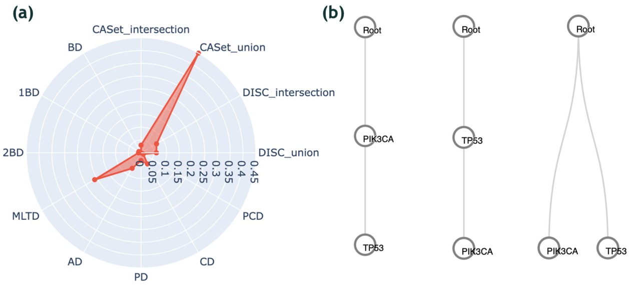

Figure 3: Optimal Combination Analysis. (a) A radar chart detailing the weight distribution of distance metrics determined by the DE algorithm for the model with the minimal pAIC value. (b) Mutation tree examples of the significant cluster in the Cox Model.

cross-validation was used, identifying PCD clusters and *MAP3K1* mutation as relevant covariates. Nonetheless, these did not reach statistical significance with p-values over 0.05. Both penalisation and cross-validation methods yielded lower pAIC values compared to previous models, as shown in Table 1.

Model Evaluation

Our evaluation of various Cox models on the test dataset, based on the Concordance Index (C-Index) and high and low risk stratification with log-rank test p-values, revealed nuanced results. The C-Index values, which measures predictive accuracy, only showed a marginal improvement in the PCD metric over the 'Only Clinical' model. This was substantiated by a slight 0.02 increase, suggesting limited enhancement from additional covariates (Table 2).

Table 2: Breast Cancer test data results for pAIC and C-Index using pre-fitted Cox models. In the C-Index column, green signifies values superior to the clinical model, while grey indicates those that are inferior.

	pAIC	C-Index
Only Clinical	668.589	0.669
Genotype	738.678	0.648
DISC ∪	766.333	0.631
DISC ∩	760.369	0.655
CASet ∪	762.64	0.641
CASet ∩	748.108	0.643
BD	756.544	0.657
1BD	749.031	0.657
2BD	751.346	0.662
MLTD	741.86	0.644
AD	755.022	0.63
PD	749.363	0.662
CD	749.864	0.654
PCD	752.37	0.671
Optimal Combination	755.605	0.647
Group Lasso	714.777	0.655
Group Lasso CV	736.242	0.653

The pAIC values across all models, unexpectedly, did not surpass the 'Only Clinical' model on the validation set, contrasting with the training data results. This indicates that the inclusion of distance metrics did not significantly improve model performance in predicting patient survival.

To address the limitations of the C-Index, we examined survival curves for high-risk and low-risk groups and conducted log-rank tests for a more in-depth assessment. The survival curves, particularly for the 'Optimal Combination' model, showed clearer distinctions than what the C-Index suggested, although the additional distance metrics did not markedly impact the survival outcomes.

Figure 4 illustrates the survival curves for selected models, Optimal Combination shows a better stratification than Only Clinical' whereas compared to 'Genotype' the difference is not so noticeable.

In summary, while the survival analysis incorporating distance metrics and genotype information provided nuanced insights, it did not significantly outperform the clinical and genotype model alone in the test dataset.

4.2 Acute Myeloid Leukaemia Dataset Analysis

Our study extended to Acute Myeloid Leukaemia (AML), analyzing 154 mutation trees from 123 patients with an average of 4.8 nodes, which included parallel mutations. The unique mutations in our dataset increased from 31 to 53, encompassing significant AML the clonal nodes label. We applied a comprehensive range of distance metrics, including all MP3 variants, to generate 123x123 distance matrices. Our histograms and hierarchical clustering with Ward linkage revealed distributions skewed to the right, translating into higher dissimilarities among the patient mutation trees (Fig. 5 (a)).

Hierarchical clustering using distance metrics revealed distinct patterns of cluster cohesion and separation. The most pronounced differentiation between clusters was observed with the DISC ∩ metric (Fig. 5 (b)). However other distances only showed similarity near the diagonal, revealing small cohesive clusters.

LOWESS regression was employed as was in the Breast Cancer data set to determine the optimal number of clusters, with DISC ∩ supporting a greater number of clusters, 14, aligning with its cluster map clarity, while most metrics selected 3 clusters only.

Through Jaccard index evaluation, we found strong correlations between simpler metrics (CD and PCD) and between union and sigmoid variants of MP3 (Fig. 5 (c)). No correlation was seen with the clinical a mutation analysis when employing this index.

The indepth mutation distribution analysis yielded mixed results with regard to clustering AML mutation trees. Metrics like PD showed the separation of clusters involving topology differences, while most showed separate mutation majority with no indication of topology.

Survival Analysis Using Cox Models

In our survival analysis using Cox models, clinical variables 'Age', 'Gender', and 'maxCCF', were evaluated along with 15 gene mutations identified in at least 10% of patients. Among the metrics assessed, 'Group Lasso' and 1BD showed improvement over the 'Only Clinical' model according to the pAIC, even though all metrics indicated better partial Log-Likelihood values. The metrics' performance is concisely presented in Table 3.

Despite the 'Group Lasso' and '1BD' metrics performing better in some respects, the LLR did not reveal a substan-

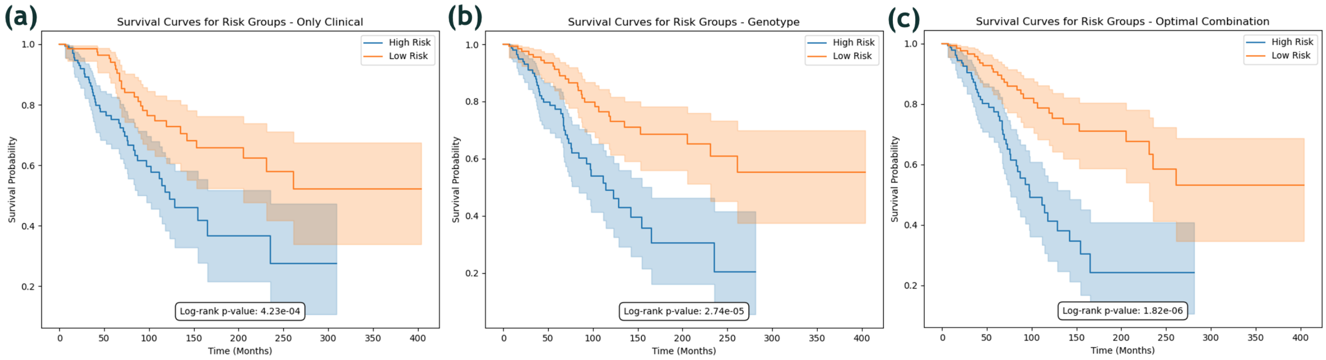

Figure 4: Survival curves for the high and low risk for (a) 'Only Clinical', (b) 'Genotype', and (c) 'Optimal Combination'

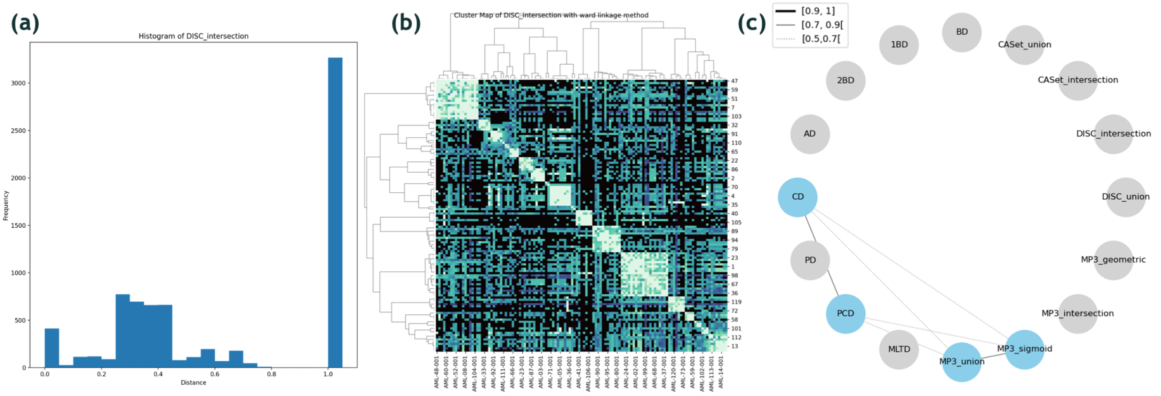

Figure 5: Analysis of Distance Metrics in AML Data. (a) Histograms depicting the distribution of distances for the DISC ∩ metric. (b) Cluster map visualising pairwise distances for the DISC ∩ metric, ranging from dark (distance = 1) to light (distance = 0), accompanied by corresponding dendrograms. (d) Jaccard index scores assessing the similarity between clusters formed by different distance metrics.

Table 3: Cox regression model outcomes displaying pAIC and Log Likelihood scores for AML training data. In the pAIC column, green indicates scores superior to the clinical pAIC, while grey denotes inferior results. For the pLog-Likelihood column, blue signifies scores better than the clinical value, with grey marking those that are worse.

	pAIC	pLog-Likelihood
Only Clinical	388	-191.0
Genotype	394.2	-179.1
DISC ∪	398.1	-179.0
DISC ∩	399.7	-169.8
CASet ∩	399.7	-178.8
CASet ∪	396.7	-178.3
BD	395.2	-177.6
1BD	388.5	-174.2
2BD	397.6	-178.8
AD	395.2	-177.6
CD	397.4	-178.7
PD	400.1	-175.1
PCD	397.1	-178.5
MLTD	394.9	-177.4
MP3 ∪	397.5	-178.8
MP3 σ	396.9	-178.5
MP3 ∩	398.1	-179.1
MP3 *geo*	397.5	-178.8
Optimal Combination	395.8	-178.9
Group Lasso	371.5	-178.7

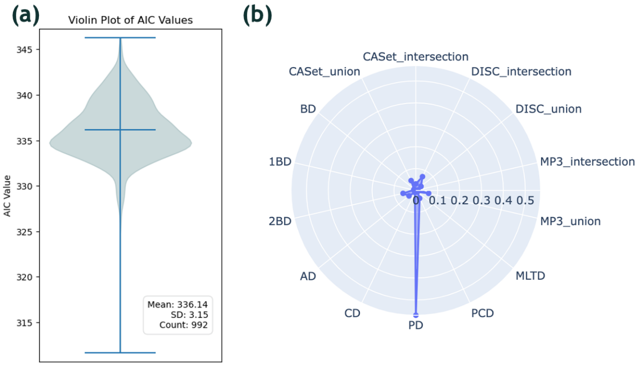

Figure 6: Outcomes of the Differential Evolution Algorithm Analysis. (a) A violin plot showing the range and distribution of penalized Akaike Information Criterion (pAIC) values over 1000 iterations. (b) A radar chart detailing the weight distribution of distance metrics determined by the DE algorithm for the model with the minimal pAIC value.

tial advantage for any metric compared to the clinical or genotype models when adjusted for multiple comparisons with Bonferroni.

Through the optimisation with the Differential Evolution algorithm, we achieved 992 runs with the lowest pAIC being an outlier, indicating a potential for further optimisation. The distributions of pAIC values for all runs are shown in Fig. 6 (a). The optimal pAIC weight distribution is revealed in Fig. 6 (b). However, no significant clusters emerged from the Cox Regression Model analysis with this combination, pointing to a unsuccessful combination of our approach for this dataset.

The Group Lasso approach highlighted the importance of 'maxCCF', several mutations, and the 1BD distance metric. A detailed examination of the data subsets associated with significant clusters, particularly cluster 2, indicated a favorable combination of mutations, *IDH2*, *SRSF2* and *RUNX1*, compared to mutation *NPM1* existence, which is the most common mutation, that could be relevant to patient outcomes. This interaction is explored through the analysis of the mutation trees depicted in Fig. 7.

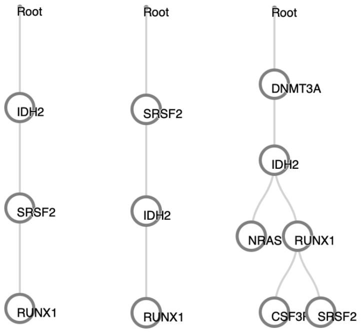

Figure 7: Analysis of the clusters of the 1BD: illustrative trees from the significant cluster of the applied Cox model.

Model Evaluation

We assessed the performance of Cox proportional hazards models on a test dataset using pAIC and C-Index as metrics. The 'Only Clinical' model was not outperformed by any other models in terms of pAIC. However, the 1BD and DISC ∩ models showed improvements in the C-Index, Table 4. The 'Group Lasso' model's underperformance was not as anticipated since the 1BD had a higher value in the C-Index and was selected by this method, this was possibly due to not including significant mutations such as *NRAS* important in the test data according to the Cox model for 1BD.

Table 4: AML test data results for pAIC and C-Index using prefitted Cox models. In the C-Index column, green signifies values superior to the clinical model, while grey indicates those that are inferior.

	pAIC	C-Index
Only Clinical	143.956	0.561
Genotype	185.607	0.498
DISC ∪	189.548	0.5
DISC ∩	323.555	0.586
CASet ∩	193.234	0.509
CASet ∪	184.331	0.541
BD	197.63	0.496
1BD	189.725	0.619
2BD	190.351	0.502
AD	192.94	0.522
CD	192.527	0.504
PD	218.256	0.474
PCD	191.387	0.53
MLTD	192.263	0.489
MP3 ∪	191.016	0.517
MP3 σ	191.155	0.506
MP3 ∩	189.719	0.502
MP3 *geo*	190.878	0.522
Optimal Combination	189.874	0.498
Group Lasso	158.933	0.541

Kaplan-Meier survival curves did not reveal significant differences in patient stratification into high and low-risk categories for most models (Fig. 8). An exception was the DISC ∩ model, which provided a better separation although not significant still, shown in 8 (c), suggesting the potential importance of mutations *IDH2* and *WT1* as driver mutations compared to *FLT3-ID* displayed in Fig. 9, tree examples portraying the significant clusters in the Cox Model. These results suggest a nuanced role of driver mutations in the prognosis of AML, although the limited data and cluster size warrants cautious interpretation.

4.3 Discussion

Our analysis aimed to enhance prognostic predictions in cancer by investigating whether clustering based on distance metrics could reveal unique tumor characteristics. We explored this in Breast Cancer and AML, focusing on the potential for these metrics to improve patient stratification and reveal cancer-specific risk factors.

In Breast Cancer, robust cluster selection was possible due to a large patient cohort, while in AML, the smaller dataset size limited the effectiveness of the clustering approach, often reducing the number of clusters to a suboptimal level. Despite this, some metrics like DISC ∩ still identified a significant number of clusters in AML. This suggests that cluster-based methodologies may not capture all the intricate details of the data, pointing to the need for alternative approaches that can utilise the full complexity of tumor genomics.

When combining metrics, the DE algorithm provided contrasting results between the two cancer types. For Breast Cancer, there was consistency in the metric weights identified for the optimal pAIC solution. However, in AML, the optimal solution deviated from this pattern, indicating that the DE algorithm might miss the true optimal solution, suggesting a need for more advanced or exhaustive search strategies.

Interestingly, the Group Lasso method identified different metrics as important for AML compared to the DE approach, with some metrics like PD being selected despite its theoretical limitations in capturing the full biological significance of mutation patterns.

Comparing with clinical and genotype data, each cancer type presented a different scenario. For Breast Cancer, all methods outperformed the clinical model based on pAIC, suggesting a better fit with the inclusion of additional covariates. In AML, only the Group Lasso method improved upon the clinical model's pAIC, and the optimal combination of metrics was not as effective as individual metrics.

Evaluation of the C-Index showed that different metrics might be more representative of different cancer types, with some improvements in patient stratification observed in Breast Cancer. However, for AML, only two metrics showed a clear improvement over the clinical model. Notably, the survival curves for the highest performing metric, DISC ∩, indicated better patient stratification and highlighted the potential prognostic importance of driver mutations.

This study underscores the complex nature of cancer genomics and the challenges in developing universally ef-

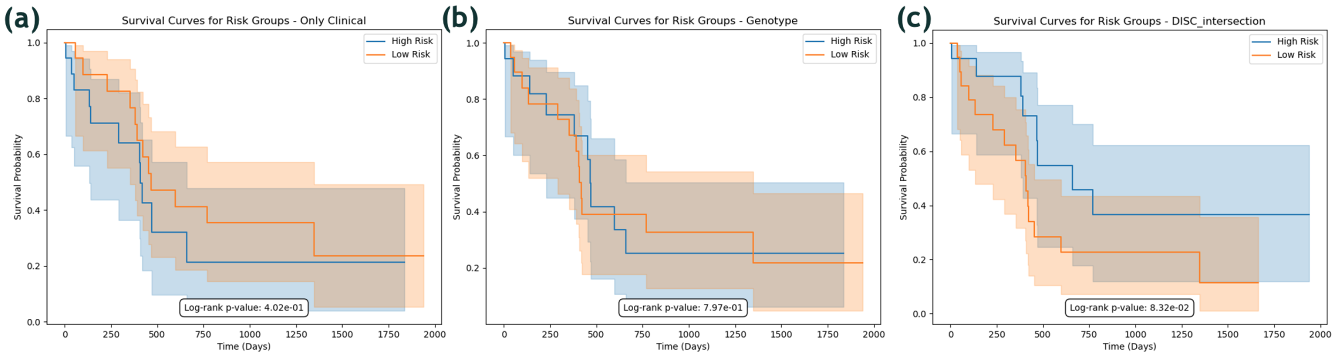

Figure 8: Survival curves for high and low risk patients for (a) 'Only Clinical', (b) 'Genotype', and (c) DISC ∩.

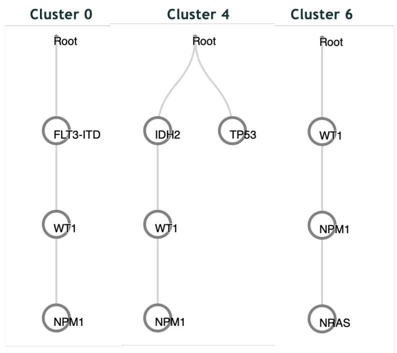

Figure 9: Tree examples for clusters 0, 4, and 6 of DISC ∩.

fective prognostic tools. It emphasizes the need for tailored approaches that consider the unique characteristics of each cancer type and patient cohort.

5 Conclusion

This study has expanded the understanding of distance metrics in cancer analysis by offering a comprehensive framework for processing tree-structured data and calculating distance matrices. Our work is particularly effective in large patient datasets, where the established methodologies successfully navigated complexities such as optimal cluster determination, avoiding the pitfalls of including overly small clusters.

Despite leveraging distance metrics clustering to improve survival predictions, the metrics did not consistently outperformed clinical or genotypical data. However, they provided valuable insights into tumor evolution. The findings were dataset-specific, emphasising the need for careful data set preparation and consideration in future studies.

The research challenges the concept of a single universal distance metric suitable for all cancer types. In Breast Cancer, metrics like the Optimal Combination and PCD were highlighted, whereas in AML, 1BD and DISC ∩ stood out, reinforcing the idea of developing cancer-specific combined metrics. This study suggests that no single metric consistently excels in prediction across different cancers.

We introduced and assessed two methods for optimising metric combinations: the Group Lasso and a novel application of the Differential Evolution (DE) algorithm. While both methods were effective, the DE algorithm, in particular, showcased strategic advantages in metric weighting, although it faces limitations in its exhaustive search capability.

This work serves as a foundation for future research, integrating distance metrics with survival data to discern prognostic tumor patterns. While the outcomes from these specific datasets were inconclusive, the approach of patient stratification using Cox Regression model coefficients derived from clustering shows potential. This study underscores the intricate challenge of improving cancer survival predictions and contributes to the advancement of personalised medicine by informing treatment decisions based on individual patient tumor characteristics.

Future Work

Our research on mutation trees in cancer prognosis delineates several directions for future research. Upcoming studies should contemplate alternatives to hierarchical clustering to more fully exploit the data within distance matrices, which may reveal insights that our current methods do not capture. Further work should also test these metrics against cancers with known mutation sequences that affect prognosis, to validate the metrics' ability to detect significant genomic distinctions. Lastly, there is a critical need for standardisation in sequencing techniques to minimise biases and enhance the reliability of comparative studies across various cancer types. Such uniformity is crucial to truly assess the effectiveness of these metrics for personalised cancer treatment strategies.

Acknowledgements

This document was written as an institutional academic requirement and as a part of the evaluation of the MSc thesis in Biomedical Engineering of the author at Instituto Superior Técnico. The work presented in this thesis was performed at ETH - Department of Biosystems Science and Engineering in Basel for 5 months during the period March-October 2023, under the supervision of Prof. Niko Beerenwinkel with the support of Xiang Ge Luo and Monica Dragan. The thesis was supervised at INESC-ID and Instituto Superior Técnico by Prof. Susana Vinga. This project was within the frame of and has received funding from the *Erasmus*+ programme and from the European Union's

Horizon 2020 research and innovation programme under grant agreement No 951970 (project OLISSIPO).

References

[1] *What is Cancer?* https://www.cancer.gov.

[2] *Cancer Research UK*. https://www.cancerresearchuk.org/about-cancerl.

[3] Ortmann, Christina A., David G. Kent, and Jyoti Nangalia: *Effect of mutation order on myeloproliferative neoplasms*. NE J. of Med., 372(7):601–612, feb 2015. https://doi.org/10.1056%2Fnejmoa1412098.

[4] Karpov, Nikolai, Salem Malikic, Md. Khaledur Rahman, and S. Cenk Sahinalp: *A multi-labeled tree dissimilarity measure for comparing "clonal trees" of tumor progression*. Alg. for Mol. Bio., 14:17, December 2019, ISSN 1748-7188.

[5] Jahn, Katharina, Niko Beerenwinkel, and Louxin Zhang: *The bourque distances for mutation trees of cancers*. Alg. for Mol. Bio., 16:9, December 2021, ISSN 1748-7188.

[6] Ciccolella, Simone, Giulia Bernardini, and Luca Denti: *Triplet-based similarity score for fully multilabeled trees with poly-occurring labels*. Bioinf., 37:178–184, April 2021, ISSN 1367-4803.

[7] Govek, Kiya, Camden Sikes, and Layla Oesper: *A consensus approach to infer tumor evolutionary histories*. Proc. 2018 ACM ICB, Comp. Bio. and Heal. Inf., pages 63–72, August 2018.

[8] DiNardo, Zach and Kiran Tomlinson: *Distance measures for tumor evolutionary trees*. Bioinf., 36:2090–2097, April 2020, ISSN 1367-4803.

[9] Jamieson, Nigel B. and David K. Chang: *Cancer genetics and implications for clinical management*. Surg. Clin. of N. Am., 95:919–934, October 2015, ISSN 00396109.

[10] Pallikonda, Husayn Ahmed and Samra Turajlic: *Predicting cancer evolution for patient benefit: Renal cell carcinoma paradigm*. BBA, 1877(5):188759, sep 2022. https://doi.org/10.1016%2Fj.bbcan.2022.188759.

[11] Cajal, Santiago Ramón y and Marta Sesé: *Clinical implications of intratumor heterogeneity: challenges and opportunities*. J. of Mol. Med., 98(2):161–177, jan 2020. https://doi.org/10.1007%2Fs00109-020-01874-2.

[12] Ciccolella, Simone, Camir Ricketts, and Mauricio Soto Gomez: *Inferring cancer progression from Single-Cell Sequencing while allowing mutation losses*. Bioinf., 37(3):326–333, August 2020, ISSN 1367-4803. https://doi.org/10.1093/bioinformatics/btaa722.

[13] Griffith, Malachi and Christopher A. Miller: *Optimizing cancer genome sequencing and analysis*. Cell Systems, 1(3):210–223, sep 2015. https://doi.org/10.1016%2Fj.cels.2015.08.015.

[14] Kuipers, Jack, Katharina Jahn, and Niko Beerenwinkel: *Advances in understanding tumour evolution through single-cell sequencing*. BBA, 1867(2):127–138, apr 2017. https://doi.org/10.1016%2Fj.bbcan.2017.02.001.

[15] Jahn, Katharina, Jack Kuipers, and Niko Beerenwinkel: *Tree inference for single-cell data*. Gen. Bio., 17:86, December 2016, ISSN 1474-760X.

[16] El-Kebir, Mohammed, Layla Oesper, and Hannah Acheson-Field: *Reconstruction of clonal trees and tumor composition from multi-sample sequencing data*. Bioinf., 31(12):i62–i70, jun 2015. https://doi.org/10.1093%2Fbioinformatics%2Fbtv261.

[17] Robinson, D F and L R Foulds: *Comparison of phylogenetic trees*.

[18] Kochenderfer, Mykel J. and Tim A. Wheeler: *Algorithms for Optimization*. The MIT Press, 2019, ISBN 0262039427.

[19] Ahmad, Mohamad Faiz, Nor Ashidi Mat Isa, and Wei Hong Lim: *Differential evolution: A recent review based on state-of-the-art works*. Alex. Eng. J., 61(5):3831–3872, may 2022. https://doi.org/10.1016%2Fj.aej.2021.09.013.

[20] Storn, Rainer and Kenneth Price: *Minimizing the real functions of the icec'96 contest by differential evolution*. In *Proc. of IEEE ICEC*, pages 842–844. IEEE, 1996.

[21] Frades, Itziar and Rune Matthiesen: *Overview on techniques in cluster analysis*. Bioinf. met. in clin. res., pages 81–107, 2010.

[22] Rousseeuw, Peter J: *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*. J. of computational and applied mathematics, 20:53–65, 1987.

[23] Clark, T G, M J Bradburn, S B Love, and D G Altman: *Survival analysis part i: Basic concepts and first analyses*. Brit. J. of Can., 89(2):232–238, jul 2003. https://doi.org/10.1038%2Fsj.bjc.6601118.

[24] Goel, Manish Kumar, Pardeep Khanna, and Jugal Kishore: *Understanding survival analysis: Kaplan-meier estimate*. Int. J. of Ayurveda res., 1(4):274, 2010.

[25] Johnson, Laura Lee and Joanna H Shih: *An introduction to survival analysis*. In *Principles and practice of clinical research*, pages 273–282. Elsevier, 2007.

[26] Malenová, Gabriela, Daniel Rowson, and Valentina Boeva: *Exploring pathway-based group lasso for cancer survival analysis: A special case of multi-task learning*. Frontiers in Genetics, 12, nov 2021. https://doi.org/10.3389%2Ffgene.2021.771301.

[27] Hastie, Trevor, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman: *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[28] Kamarul Imran, Wan Nor Arifin, Tengku Muhammad Hanis Tengku Mokhtar: *Chapter 11 Survival Analysis: Kaplan-Meier and Cox Proportional Hazard (PH) Regression | Data Analysis in Medicine and Health using R — bookdown.org*, 2022. [Accessed 10-10-2023].

[29] Jafari, Mohieddin and Naser Ansari-Pour: *Why, when and how to adjust your p values?* Cell J. (Yakhteh), 20(4):604, 2019.

[30] Neyman, Jerzy and Egon S Pearson: *On the use and interpretation of certain test criteria for purposes of statistical inference: Part i*. Biometrika, pages 175–240, 1928.

[31] Akaike, Hirotugu: *Factor analysis and aic*. Psychometrika, 52:317–332, 1987.

[32] Longato, Enrico, Martina Vettoretti, and Barbara Di Camillo: *A practical perspective on the concordance index for the evaluation and selection of prognostic time-to-event models*. J. of Biomedical Informatics, 108:103496, aug 2020. https://doi.org/10.1016%2Fj.jbi.2020.103496.

[33] Harrell, Frank E, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati: *Evaluating the yield of medical tests*. Jama, 247(18):2543–2546, 1982.

[34] Luo, Xiang Ge, Jack Kuipers, and Niko Beerenwinkel: *Joint inference of exclusivity patterns and recurrent trajectories from tumor mutation trees*. Nature Communications, 14(1):3676, 2023.

[35] Razavi, Pedram, Matthew T Chang, Guotai Xu, Chaitanya Bandlamudi, Dara S Ross, Neil Vasan, Yanyan Cai, Craig M Bielski, Mark TA Donoghue, Philip Jonsson, *et al.*: *The genomic landscape of endocrine-resistant advanced breast cancers*. Cancer cell, 34(3):427–438, 2018.

[36] El-Kebir, Mohammed, Gryte Satas, and Layla Oesper: *Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures*. Cell Systems, 3(1):43–53, jul 2016. https://doi.org/10.1016%2Fj.cels.2016.07.004.

[37] Morita, Kiyomi and Feng Wang: *Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics*. Nature, 11(1):5327, 2020.