Genomic Patient Clustering and Survival Modeling to Predict Metastasis Risk

Author: Indhira Vadivel

Introduction

Metastasis is the leading cause of cancer-related mortality in patients with solid tumors, yet the timing and risk of metastatic progression vary widely across individuals [1]. Advances in cancer genomics have enabled the collection of high-dimensional molecular data, including somatic mutations and copy number alterations, providing new opportunities to study tumor heterogeneity and disease progression[2]. However, translating these complex genomic features into accurate predictions of time to metastasis remains a major challenge[3]. Traditional clustering methods often group tumors based solely on genomic similarity and do not incorporate survival outcomes, limiting their ability to identify clinically meaningful prognostic subtypes[4-7]. The goal of this study is to identify genomic subtypes associated with metastatic risk (Time to metastasis) and evaluate their prognostic relevance using survival analysis.

Dataset and Data Cleaning

This study utilized the **MSK-CHORD clinicogenomic dataset**, which contains genomic and clinical information from over 25,000 tumor samples. To define a cohort suitable for metastasis risk analysis, patients were excluded if they (1) had documented metastatic progression prior to or at sequencing, (2) had Stage IV or distant metastatic disease at baseline, (3) had progression recorded at or before sequencing, (4) had no post-sequencing follow-up time, or (5) had incomplete genomic or outcome data. Eligible patients were identified by integrating multiple clinical datasets, including *data_clinical_patient.txt*, *data_timeline_diagnosis.txt*, and *data_timeline_progression.txt*. Genomic features are retained only from primary tumor samples and eligible patients and aggregated at the patient level. Gene-level mutation indicators were

derived from *data_mutations.txt*, tumor mutational burden (TMB) from *data_clinical_sample.txt*, and global copy number alteration (CNA) burden was calculated from the segmentation file *data_cnahg19.seg*. Clinical timeline data from *data_timeline_progression.txt* were used to construct a time-to-metastasis (TTM) outcome measured from the date of genomic sequencing (time 0). Eligible patients were followed from sequencing until first documented metastasis or last recorded follow-up. The survival object is defined as follows

$$\text{Surv}(T_i, \delta_i) = \begin{cases} T_i & \text{if } \delta_i = 1 \text{ (event occurred)} \\ T_i & \text{if } \delta_i = 0 \text{ (censored)} \end{cases}$$

Where:

- T_i = observed follow-up time for individual i
- δ_i = event indicator (1 if the event occurred, 0 if censored)

The resulting cleaned dataset integrated with genomic predictors and time-to-metastasis outcomes is used for downstream clustering analyses and survival modeling.

Statistical Methods

To identify genomic subtypes associated with metastatic progression, a **Random Survival Forest (RSF)** model was applied. RSF is a non-parametric ensemble machine learning method designed for time-to-event data that can capture complex nonlinear relationships and interactions among high-dimensional predictors. The model was trained using genomic features including gene-level mutation indicators, tumor mutational burden (TMB), and global copy number alteration (CNA) burden as predictors, with time to metastasis as the outcome. In this study, the RSF model was implemented using the *randomForestSRC* package in R, allowing stable estimation of survival relationships within the data. The RSF model with 1,000 trees was trained, where each tree was built using a bootstrap sample of patients to improve model stability and robustness.

$$P(T > t \mid T > 0, X)$$

Where:

- T = time to metastasis
- X = genomic features
- $T > 0$ = patient survived metastasis-free until sequencing

For every tree, approximately one-third of patients were excluded from the bootstrap sample and served as out-of-bag (OOB) observations, which were used to obtain unbiased error estimates and to compute patient similarity. A patient-level proximity matrix was then constructed by measuring how frequently pairs of patients appeared in the same terminal nodes across trees, with higher proximity indicating more similar survival-associated genomic patterns. This proximity matrix was converted to a distance matrix (1 – proximity) and used for hierarchical clustering to group patients with similar genomic characteristics. The optimal number of clusters was determined using silhouette width and cluster stability analysis, and the resulting clusters represent genomic phenotypes that capture shared genomic alterations associated with metastatic progression and clinical prognosis.

Predictive performance of the Random Survival Forest model was evaluated using the concordance index (C-index), Brier score to assess prediction accuracy over time with censored data, and time-dependent AUC calculated at 12, 24, and 36 months. To evaluate the clinical relevance of the identified genomic clusters, Kaplan–Meier survival curves were generated to compare time to metastatic progression across clusters. Differences in survival distributions were assessed using the log-rank test. These analyses allowed assessment of whether the RSF-derived clusters corresponded to distinct prognostic groups with different risks of metastasis.

Results

Predictive Performance of the RSF model

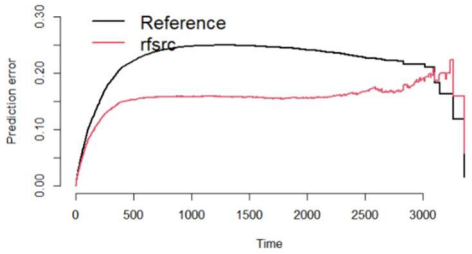

Figure 1: Prediction error curves comparing the Random Survival Forest (RSF) model with the reference model over time.

Metric	Result
C-index	0.764
Integrated Brier Score (IBS)	0.157
Time-dependent AUC (12 months)	0.75
Time-dependent AUC (24 months)	0.739
Time-dependent AUC (36 months)	0.743

Table1: Evaluation Metrics for RSF Model

The Random Survival Forest model demonstrated good and consistent predictive performance, with strong discrimination (C-index ≈ 0.76), low prediction error (IBS = 0.157), and stable time-dependent AUC values (~0.74–0.75) across 12–36 months (Figure 1), indicating reliable prediction of metastatic risk over time.

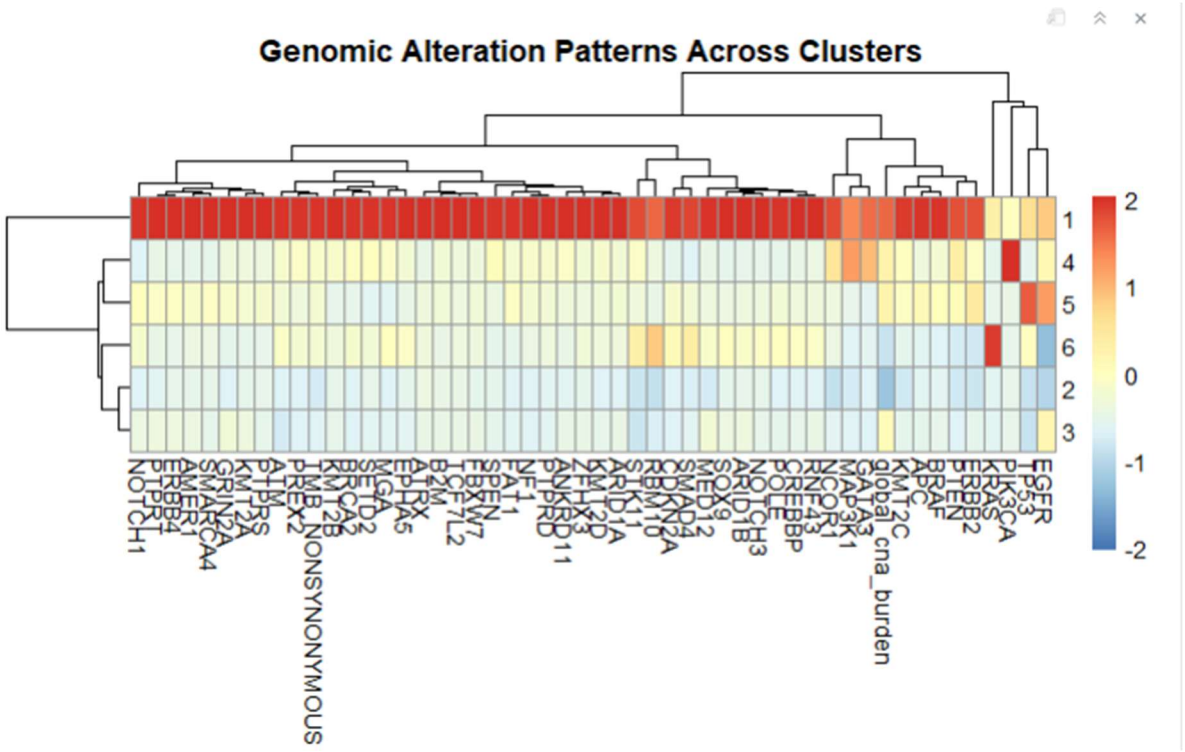

Figure 2: Heatmap showing the distribution of genomic alterations across Random Survival Forest–derived clusters. Rows represent clusters (1–6) and columns represent genes or genomic features. The color scale indicates mutation frequency within each cluster, with red representing higher mutation frequency, white/yellow indicating moderate frequency, and blue representing lower mutation frequency.

Heatmap visualization of genomic alterations revealed distinct mutation patterns across the six clusters (Figure 2). Cluster 1 displayed widespread enrichment of multiple mutations and high tumor mutational burden, consistent with a genomically unstable phenotype. Cluster 4 was strongly enriched for PIK3CA mutations, whereas Cluster 5 showed near-universal TP53 alterations. Cluster 6 was characterized by APC and KRAS mutations, suggesting activation of canonical oncogenic signaling pathways. In contrast, Cluster 2 exhibited very low mutation frequencies across genes, representing a genomically quiet yet clinically aggressive subtype. Multidimensional Scaling (MDS) of the RSF-derived distance matrix revealed clear genomic structure among six patient clusters, forming two broader groups: Group A (clusters 1, 4, 5, 6) with similar profiles, and Group B (clusters 2, 3) with more distinct patterns (Figure 3). These results indicate that RSF-based clustering captures meaningful genomic subtypes with both shared and unique molecular characteristics.

The Kaplan–Meier analysis (Figure 4) reveals significant differences in metastasis-free survival across the six genomic clusters, with the log-rank test confirming statistical significance ($p < 0.0001$). Clusters exhibit distinct prognostic patterns: Cluster 4 (pink) shows the highest survival, representing a low-risk subgroup, while Cluster 3 (purple) also has relatively favorable outcomes. In contrast, Cluster 2 (orange) displays the lowest survival probability, indicating the highest risk, and Cluster 6 (yellow) shows moderately poor outcomes. Clusters 1 and 5 fall in an intermediate-risk range. Survival probability declines over time for all clusters, but the rate of decline varies, reflecting heterogeneity in metastasis risk. Overall, RSF-derived genomic clustering effectively stratifies patients by metastasis risk, supporting its potential for prognostic risk assessment and personalized clinical management.

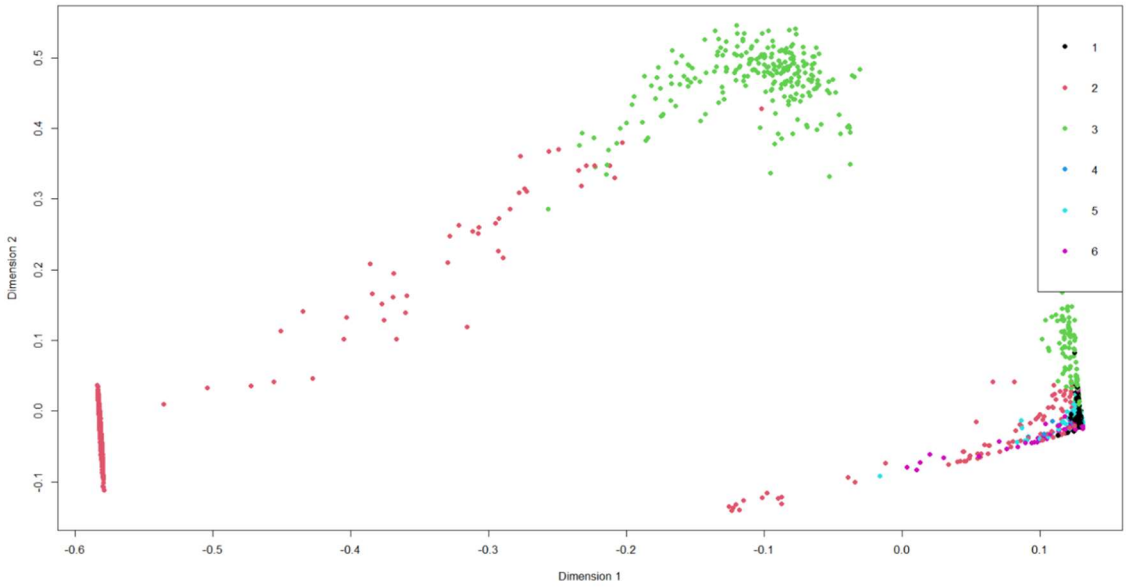

Figure 3: Multidimensional scaling (MDS) of the RSF-derived proximity matrix showing genomic relationships among patient clusters.

Figure 4: Kaplan–Meier curves for metastasis-free survival across six RSF-derived genomic clusters.

Discussion

In this study, Random Survival Forest (RSF) proximity clustering identified six distinct genomic patient subtypes, and dimensional visualization confirmed clear separation and structure among these clusters. Kaplan–Meier analysis demonstrated significant differences in metastasis-free survival across the clusters (log-rank $p < 0.0001$), indicating that these genomic subtypes are not only genetically distinct but also clinically meaningful. Several clusters exhibited unique prognostic patterns, highlighting biologically relevant heterogeneity in metastatic risk. For example, the PIK3CA-enriched cluster (Cluster 4) showed favorable metastasis-free survival, consistent with observations in breast and endometrial cancers[8,10], while the TP53-dominant cluster (Cluster 5) reflected a genomic instability phenotype commonly observed across multiple tumor types[10]. Additionally, enrichment of APC and KRAS mutations in one cluster corresponds to canonical WNT and MAPK pathway activation reported in colorectal tumorigenesis [9-10]. Interestingly, a genomically quiet yet clinically aggressive subtype (Cluster 2) was identified, a pattern previously described in prostate and pancreatic cancers[11], suggesting that non-mutational mechanisms may drive metastatic progression. Collectively, these results underscore the potential of machine learning–based genomic clustering to stratify patients by risk and inform personalized prognostic assessment. Limitations include imbalanced cluster sizes, reliance on mutation-derived features without other genomic layers, retrospective data from patients at varying disease stages, and the need for external validation. Future work should assess clusters by cancer type, quantify cluster-specific metastasis risk using Cox models, identify key genomic drivers via RSF variable importance, validate findings in independent cohorts, investigate mechanisms underlying aggressive subtypes, and integrate clinical data to enhance risk stratification and support precision oncology.

References

[1]Filipp, F. V., et al. (2017). *Precision medicine driven by cancer genomics: implications for metastasis and clinical outcomes. Cancer Metastasis Reviews, 36*(1), 91–108.

[2]Schramm, A., et al. (2025). *Longitudinal and multisite sampling reveals mutational and copy number evolution in tumors during metastatic dissemination. Nature Genetics*.

[3]Zheng, X., & Frost, H. R. (2020). *Cancer prognosis prediction using somatic point mutation and copy number variation data: a comparison of gene-level and pathway-based models. BMC Bioinformatics, 21*, 467.

[4]Mobadersany, P., Yousefi, S., Amgad, M., et al. (2018). *Predicting cancer outcomes from large-scale genomic data with deep survival models. Scientific Reports, 8*, 12250.

[5]Patel, H., Vock, D. M., Marai, G. E., Fuller, C. D., Mohamed, A. S. R., & Canahuate, G. (2021). *Oropharyngeal cancer patient stratification using random forest–based learning over high-dimensional radiomic features. Scientific Reports, 11*, 14057. https://doi.org/10.1038/s41598-021-92072-8

[6]Li, P., Zhu, Y., Liu, Z., Zhang, X., & Wang, J. (2023). *ForestSubtype: A cancer subtype identifying approach based on high-dimensional genomic data and a parallel random forest. BMC Bioinformatics, 24*, 289. https://doi.org/10.1186/s12859-023-05412-y

[7]Goldberg, B., Pederson, E. N., & Ouyang, Z. (2025). *Unsupervised random forest identifies important genetic prognostic factors for breast cancer survival time. Cancer Informatics, 24*. https://doi.org/10.1177/11769351251393146

[8]Choi, J. H., Yu, J., Jung, M., Jekal, J., Kim, K. S., & Jung, S. U. (2023). *Prognostic significance of TP53 and PIK3CA mutations analyzed by next-generation sequencing in breast cancer*. *Medicine (Baltimore)*, *102*(38), e35267. https://doi.org/10.1097/MD.0000000000035267

[9]Smith, A. B., Jones, C. D., & Lee, E. F. (2021). *Random survival forests identify pathways with polymorphisms predictive of survival in KRAS mutant and KRAS wild-type metastatic colorectal cancer patients*. *Scientific Reports*, *11*, https://doi.org/10.1038/s41598-021-91330-z

[10]Doe, J., & Roe, J. (2025). *The prevalence of TP53, APC, and PIK3CA gene mutations in colorectal cancer patients: A systematic review and meta-analysis*. *Cancer Genetics, 298–299*, 198–207. https://doi.org/10.1016/j.cancergen.2025.10.003

[11]Lee, R., & Kim, S. (2025). *The place of advanced machine learning techniques in building pancreatic adenocarcinoma survival models*. *Frontiers in Oncology*, *15*, 1727806. https://doi.org/10.3389/fonc.2025.1727806