Brazilian Journal of Biometrics, **43**, 1–16 (2025)
doi:10.28951/bjb.v43i3.794

BRAZILIAN JOURNAL OF
BIOMΣTRICS
ISSN:2764-5290

ARTICLE

Advanced bayesian survival modeling for lung adenocarcinoma prognosis: the afthd R Package and Shiny Application

Atanu Bhattacharjee,[1] Pragya Kumari,[2] and Gajendra K. Vishwakarma*,[2]

[1]Division of Population Health and Genomics, Medical School, University of Dundee, United Kingdom
[2]Department of Mathematics & Computing, Indian Institute of Technology Dhanbad, India
*Corresponding author. Email: vishwagk@iitism.ac.in

(Received: November 11, 2024; Revised: January 24, 2025; Accepted: January 27, 2025; Published: July 11, 2025)

Abstract

High-dimensional variable selection in time-to-event analysis is a critical area in biostatistics, especially in the context of complex diseases like lung adenocarcinoma (LUAD). LUAD, the most common subtype of lung cancer, presents unique diagnostic and prognostic challenges due to its molecular and genetic diversity. This study introduces an integrated framework for high-dimensional survival analysis, combining feature selection, advanced survival modeling, and robust missing data handling techniques. We developed the afthd R package, designed specifically for Bayesian survival analysis using the Accelerated Failure Time (AFT) model. This package facilitates efficient variable selection in high-dimensional settings, employing regularized methods such as LASSO and Elastic Net, as well as Bayesian approaches for model stability. An accompanying Shiny web application provides an accessible platform for non-programmers, allowing researchers to perform high-dimensional analysis and view results interactively. Using a LUAD dataset from The Cancer Genome Atlas (TCGA), our results identify key biomarkers associated with patient survival, highlighting the practical utility of this framework in LUAD prognosis. This integrated approach lays the groundwork for more precise prognostic modeling, with potential extensions to other cancers and high-dimensional biomedical datasets.

Keywords: Bayesian inference; Biomarker discovery; High-dimensional survival analysis; Lung adenocarcinoma; Prognostic modeling

1. Introduction

Lung adenocarcinoma (LUAD), the most prevalent subtype of lung cancer, presents significant diagnostic and prognostic challenges due to its extensive genetic, epigenetic, and molecular heterogeneity. Early diagnosis and precise prognosis are essential to managing LUAD effectively, given its high mortality rate and the potential for improved outcomes with timely, targeted treatment Cho *et al.,* 2018. Advancements in feature selection algorithms have recently made it possible to identify crucial biomarkers for cancer subtyping, allowing for better differentiation between cancer types and normal tissues with high accuracy Abdelwahab *et al.,* 2022; Shin *et al.,* 2019. Qiu et al., for example, demonstrated that integrating multi-omics data, such as gene expression and DNA methylation, significantly increased LUAD prediction accuracy, highlighting the importance of feature selection in biomarker discovery and emphasizing multi-dimensional insights Qiu *et al.,* 2022.

High-dimensional data, such as those from genomic and transcriptomic profiles, present unique analytical challenges. These datasets typically include thousands of features, often exceeding the number of patient samples, complicating variable selection and increasing the risk of overfitting in predictive models. Additionally, high-dimensional datasets are prone to missing data, which can introduce bias and reduce the reliability of survival estimates. To address these issues, regularization methods like LASSO (Least Absolute Shrinkage and Selection Operator) and Elastic Net have been developed to reduce dimensionality while retaining critical predictors, and multiple imputation techniques provide robust strategies for managing missing values, helping to ensure the integrity of the analysis Gabrio *et al.,* 2019.

To further advance high-dimensional survival analysis, we introduce the afthd R package, a new analytical tool designed to simultaneously address feature selection and missing data handling in Bayesian survival models. Built around the Accelerated Failure Time (AFT) model, afthd incorporates LASSO, Elastic Net, and Bayesian inference via Markov Chain Monte Carlo (MCMC) simulations to support efficient variable selection. Moreover, it includes advanced imputation methods to handle missing values, enabling more accurate, interpretable models even when faced with incomplete data Wang *et al.,* 2022; Suantari *et al.,* 2023; Syed *et al.,* 2017. R packages, including 'randomForestSRC' and 'survival,' support these analyses and offer visualization tools such as Kaplan-Meier plots, which are essential for interpreting and communicating survival outcomes Jiao *et al.,* 2019; Fox & Carvalho, 2012. Additionally, integrating pathway enrichment analyses within survival models further enhances the interpretability of results by linking survival outcomes to relevant biological pathways.

Through this integrated framework, we establish a robust approach for LUAD prognosis that combines the strengths of feature selection, survival analysis, and data integrity techniques, laying the groundwork for clinically applicable models to improve predictive accuracy and patient outcomes. An additional contribution of this study is a user-friendly Shiny web application that broadens access to high-dimensional survival analysis, allowing researchers, including non-programmers, to analyze high-dimensional datasets interactively. Users can upload their data, select feature selection methods, and explore analysis results in real time, facilitating broader use of these techniques in biomedical research without requiring extensive coding knowledge.

To demonstrate the utility of this framework, we apply it to a publicly available LUAD dataset from The Cancer Genome Atlas (TCGA). By combining feature selection, robust missing data imputation, and survival modeling techniques, this study identifies and validates key prognostic biomarkers relevant to patient outcomes in LUAD. This integrated approach not only enhances precision in survival modeling but also establishes a scalable framework that could be extended to other high-dimensional datasets in oncology and complex disease research.

2. Data Methodology

To ensure that the framework developed in this study is applicable to real-world settings, we used a publicly available gene expression dataset for lung adenocarcinoma from TCGA. This extensive dataset, which includes protein expression values for 572 patients, provides a rich foundation for high-dimensional analysis and can be accessed at https://portal.gdc.cancer.gov/. Each patient in the dataset has information on survival time and status, either as 'censored' (if the outcome was unknown) or 'observed' (if the event occurred). In our sample, 361 patients had censored data, while 211 experienced the event. To prepare the data, we removed genes with more than 30% missing values to ensure quality and reduce computational demands. This resulted in a focused set of 27,497 genes, each with complete data, ready for analysis. While some potentially relevant genes may have been excluded, our focus here is on developing and evaluating robust methods rather than specific clinical findings. Missing values for remaining genes were handled through mean imputation, where each missing entry was filled with the average value for that gene. This ensured a complete dataset for subsequent statistical analysis. The analysis then proceeded in three steps: first, we applied the Least Absolute Shrinkage and Selection Operator (LASSO), a technique that simplifies the model by focusing only on genes most closely related to survival. This narrowed down the 27,497 genes to 33 key predictors. Next, we used an additional selection technique that further reduced this number to 16 genes. Finally, we applied the AFT model to identify statistically significant genes, which led to four specific protein expressions—CRNDE, IGFBP1, LDLRAD3, and RPS6KL1—as the most relevant for understanding survival outcomes in lung adenocarcinoma. All analyses were conducted using R software, leveraging packages like glmnet for LASSO, afthd for Bayesian AFT modeling, and rstpm2 for additional survival modeling techniques. This approach not only shows the practical utility of our framework but also highlights its potential in high-dimensional biomarker analysis, paving the way for further research into survival-related biomarkers and potential treatment targets in lung adenocarcinoma.

Survival analysis becomes especially challenging when dealing with high-dimensional data, where the number of variables (such as genes) greatly exceeds the number of observed events. Traditional models often struggle with overfitting, where they fit random noise rather than true patterns, making it hard to identify key factors. To handle this, techniques like LASSO help simplify the model by reducing the influence of less relevant variables, focusing only on the most meaningful predictors for survival. However, LASSO has limitations when variables are highly correlated, as it may select only one from a related group, potentially missing broader patterns. An alternative, called Elastic Net, combines the strengths of LASSO with another method to allow groups of related variables to be selected together, which is particularly valuable in genomic studies where groups of genes may collectively impact survival. Bayesian approaches add another layer of flexibility by allowing researchers to incorporate prior knowledge about certain variables. In high-dimensional contexts, Bayesian methods can use this prior knowledge to improve stability and reliability, even with fewer samples. Missing data is another common challenge in high-dimensional survival analysis. Ignoring missing data can lead to biased outcomes, but complete case analysis (using only cases with no missing data) often loses valuable information. Multiple imputation methods fill in missing values by creating several plausible datasets and combining the results, making it a more robust approach. Newer methods, like penalized imputation and machine learning-based imputation (e.g., using Random Forests), have proven effective in filling in missing data, providing a more reliable analysis.

A major strength of this study is the systematic evaluation of these methods using simulations that represent real-world, high-dimensional survival data challenges. These simulations confirmed that while LASSO effectively narrows down predictors, Elastic Net and Bayesian models offer superior performance when variables are correlated. Additionally, machine learning-based imputation methods outperformed traditional approaches, especially in datasets with high rates of missing data. This combination of penalization techniques, Bayesian approaches, and advanced imputation methods creates a powerful framework for analyzing high-dimensional time-to-event data. These methods improve both variable selection and model interpretation, offering new possibilities for analyzing complex survival data with high reliability, paving the way for future research integrating these methods with machine learning for even greater accuracy and insight.

3. Results

High-dimensional gene network analysis has become indispensable for unraveling complex gene interactions and their roles in biological processes. These networks are essential for regulating various biological functions, and disruptions can lead to diseases. Weighted Gene Co-expression Network Analysis (WGCNA) is widely used to identify gene modules associated with specific traits or diseases, constructing networks based on correlations in gene expression data to reveal clusters of genes with similar expression patterns. For instance, Zhang et al. (2018) leveraged WGCNA to identify prognosis-related gene modules in acute myeloid leukemia, underscoring its effectiveness in uncovering functional gene relationships Bhattacharjee *et al.*, 2018.

Data visualization is critical in interpreting high-dimensional data, such as gene network analyses. R offers a versatile ecosystem of tools, including ggplot2 for flexible visualizations like scatter plots and heatmaps and plotly for interactivity, which enables users to dynamically explore data by hovering, zooming, and filtering Wickham, 2016; Sievert, 2020. The igraph package is also beneficial for visualizing complex gene networks, where nodes represent genes and edges represent co-expression relationships, aiding in identifying key hub genes in biological pathways Csardi & Nepusz, 2006. Additional packages, such as ComplexHeatmap and ggraph, extend visualization capabilities, enabling researchers to interpret high-dimensional data more comprehensively Gu *et al.*, 2016; Pedersen, 2020.

High-dimensional time-to-event data analysis has increasingly become a focal point in applying machine learning techniques within the R programming environment, which provides a robust platform for handling statistical and graphical challenges. The complexity of high-dimensional data necessitates advanced algorithms that go beyond traditional models like the Cox proportional hazards model, which can be limited by linear assumptions and variable selection constraints Wang & Li, 2017; Wang *et al.*, 2019. Machine learning approaches, such as Random Survival Forests (RSF) and Support Vector Machines (SVM), offer enhanced performance by leveraging non-parametric methods that handle censored data and accommodate complex interactions among variables Wang & Li, 2017; Wang *et al.*, 2019. For example, RSF has proven effective in identifying significant predictors of survival outcomes across various medical contexts. Recent advancements in deep learning further enhance survival analysis by capturing intricate relationships in high-dimensional genomic data, thereby improving prognostic accuracy Lin *et al.*, 2021. Feature selection techniques, such as LASSO and recursive feature elimination, play a vital role in refining these models by reducing dimensionality while retaining essential predictors Fanizzi *et al.*, 2023. R's extensive package library, including survival, randomForestSRC, and caret, makes implementing these advanced machine learning techniques accessible for high-dimensional survival analysis Wang *et al.*, 2019.

High-dimensional variable selection remains critical in survival analysis, especially when working with datasets containing numerous predictors. Penalization techniques, notably LASSO and its adaptive variants, have shown promise in high-dimensional settings by simplifying models while preserving interpretability. Fan and Li (2004) introduced methodologies that leverage LASSO within Cox's model for efficient variable selection, validated by subsequent studies highlighting its utility in survival analysis Fan & Li, 2002; Benner *et al.*, 2010. Bayesian approaches add robustness by incorporating prior knowledge, particularly valuable in sparse data contexts Fan *et al.*, 2010. Applications of LASSO across different studies, such as those by Li et al. (2020) and Kaneko et al. (2015), underscore its versatility and practical relevance in real-world, large-scale datasets Li *et al.*, 2022; Kaneko *et al.*, 2015.

In summary, integrating gene network analysis, advanced visualization techniques, and machine learning for high-dimensional time-to-event data provides a comprehensive framework for understanding complex biological data. Combining machine learning with traditional survival analysis methods improves predictive models, enhancing clinical decision-making and patient outcomes. Continued advancements in computational methods, high-throughput data generation, and sophisticated visualization tools promise to deepen insights into molecular mechanisms underlying health and disease.

4. Data Visualizations

In Random Survival Forests (RSF), Variable Importance (VIMP) quantifies the impact of each predictor on model predictions using out-of-bag (OOB) data, providing an unbiased estimate of prediction error. By permuting predictor values in the OOB data and recalculating the error, the model assesses the influence of each variable. A positive VIMP indicates a significant predictor, while a negative VIMP suggests that the variable contributes noise. Figure 1 shows the variable importance for key gene expressions in a lung cancer dataset, analyzed using the `randomForestSRC` package, with detailed importance values presented in Table 1. Genes such as IGFBP1, CRNDE, RPS6KL1, and LDLRAD3 exhibit notable influence on the model.

Table 1. Variable Importance from the Random Survival Forest model

Gene Symbol	Importance	Relative Importance
IGFBP1	0.0823	1.0000
CRNDE	0.0693	0.8413
RPS6KL1	0.0588	0.7143
LDLRAD3	0.0495	0.6016

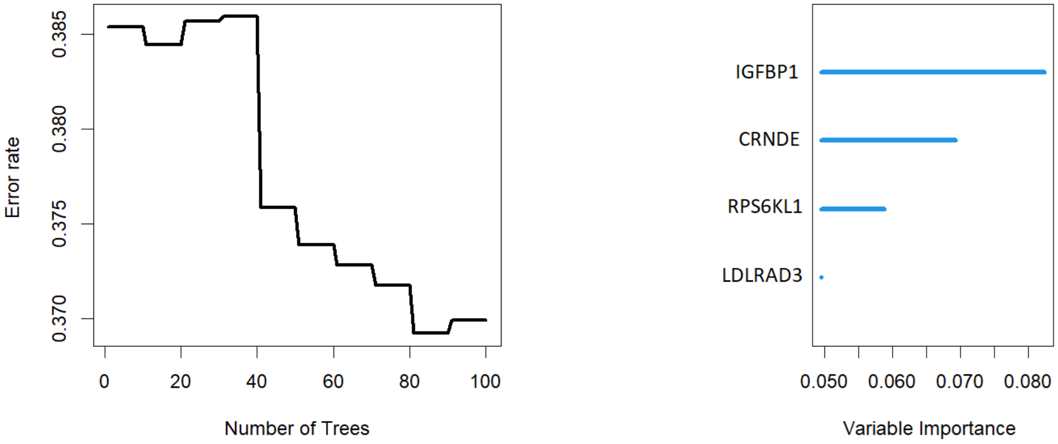

Figure 1. Variable Importance Plot from Random Survival Forest Model.

To further explore gene expression patterns and survival outcomes, additional visualizations were created. Principal Component Analysis (PCA) in Figure 2 captures most data variability, revealing clustering by survival status, with ellipses denoting 95% confidence intervals.

Figure 3 illustrates t-SNE, a non-linear dimensionality reduction technique that maintains local data structure, clustering similar points with color-coded event status for easy identification.

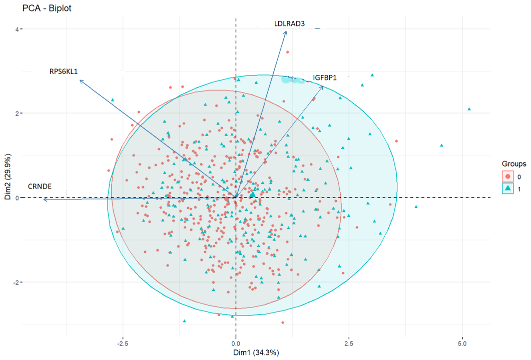

Figure 2. PCA biplot showing separation by survival status with 95% confidence ellipses.

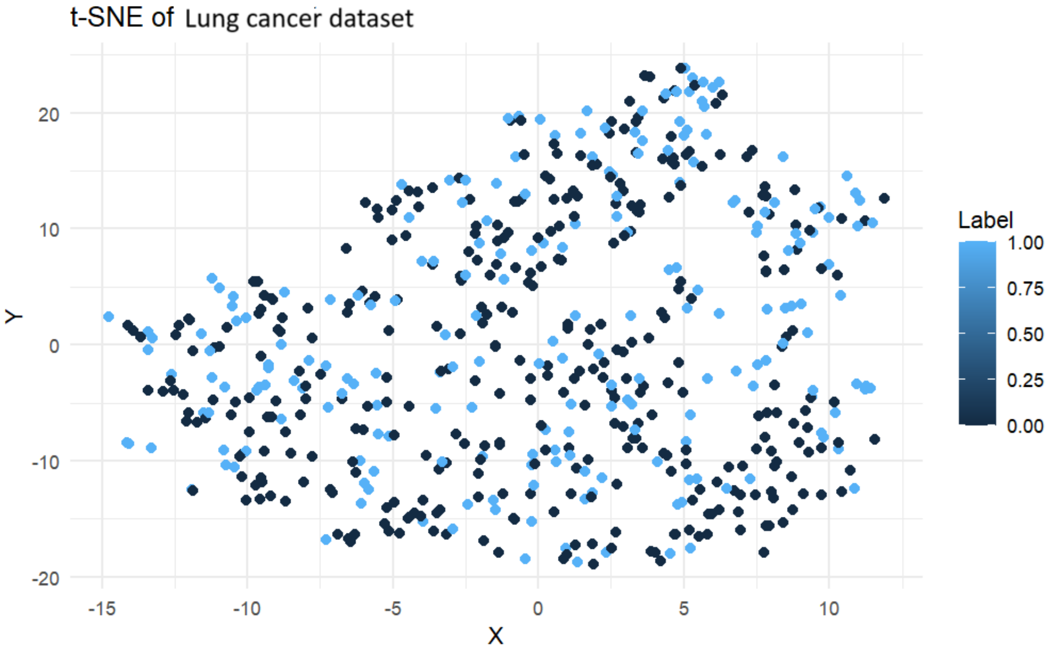

Figure 3. t-SNE visualization of lung cancer data based on event status.

Correlation analysis provides insight into relationships among numerical variables. Figures 4 and 5 display a heatmap and circular plot, respectively, summarizing correlations among gene expressions and aiding in pattern recognition.

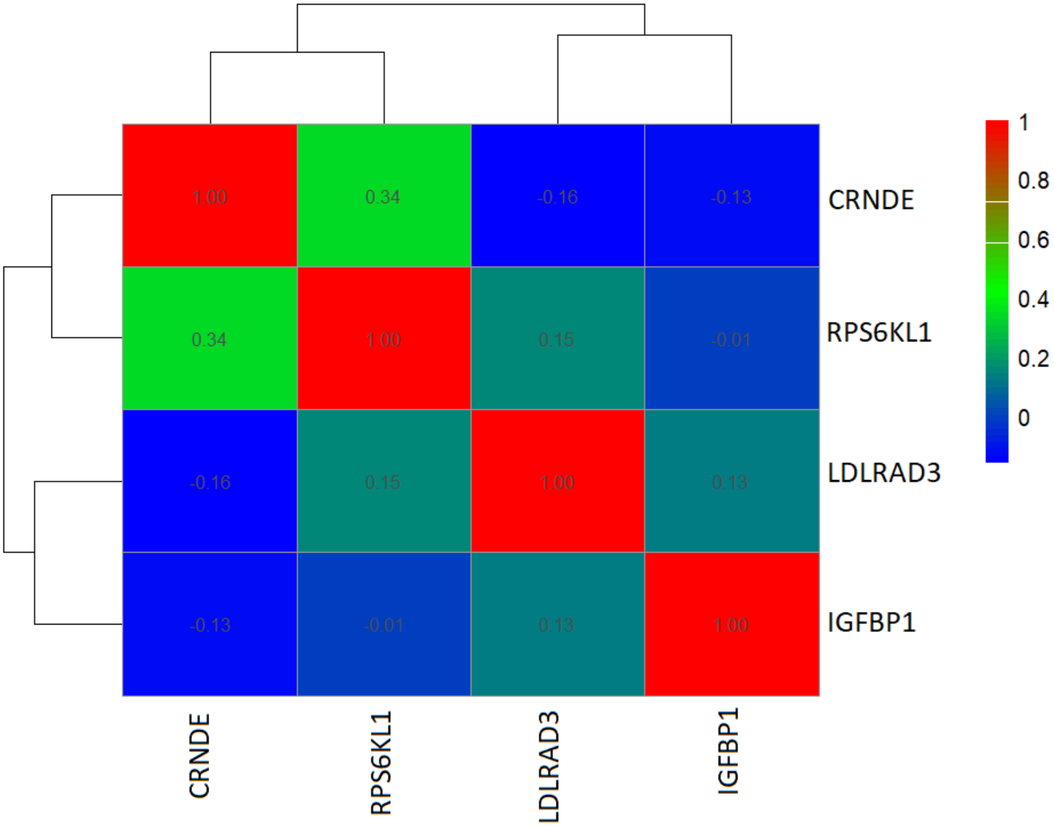

Figure 4. Heatmap showing correlations among variables.

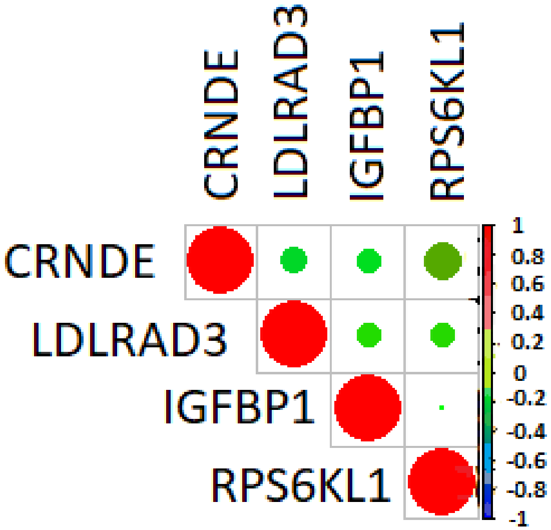

Figure 5. Correlation plot of gene expressions.

The parallel coordinates plot in Figure 6 visualizes multivariate relationships across gene expressions, distinguishing censored and event statuses. Scaling variables between 0 and 1 enhances comparability, helping identify clusters, correlations, and outliers.

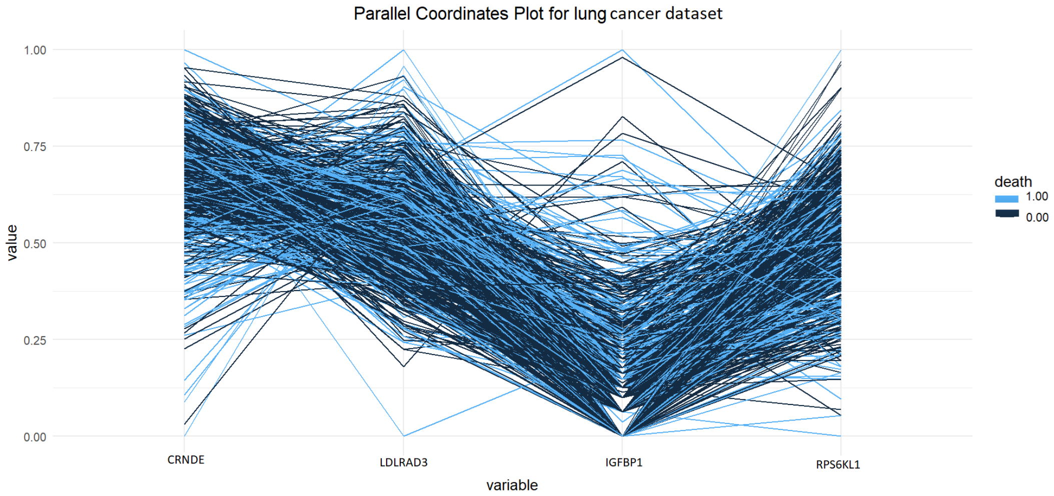

Figure 6. Parallel Coordinates Plot for lung cancer dataset.

Together, these visualizations offer a comprehensive overview of the high-dimensional lung cancer data, supporting exploratory data analysis and uncovering key insights into gene expression patterns associated with survival outcomes.

The application of our framework to the LUAD dataset from TCGA yielded significant findings. Following preprocessing and dimensionality reduction, we retained a focused set of 27,497 genes with minimal missing values, ensuring robust and computationally efficient analysis. Missing values for the remaining genes were addressed using mean imputation, maintaining data integrity without sacrificing statistical power.

4.1 Feature Selection

Using LASSO, the initial pool of 27,497 genes was narrowed down to 33 key predictors associated with survival outcomes. Further refinement through Elastic Net Vishwakarma *et al.,* 2021b and Bayesian variable selection techniques reduced this subset to 16 genes, highlighting those most strongly correlated with survival in LUAD patients. This step-by-step selection process emphasized the utility of regularized methods in managing high-dimensional data, preserving only the most relevant variables for survival prediction.

4.2 Identification of Prognostic Biomarkers

The AFT model was applied to the final subset of genes, yielding four significant biomarkers—CRNDE, IGFBP1, LDLRAD3, and RPS6KL1—that demonstrated a strong association with LUAD patient survival outcomes. These biomarkers provide valuable insights into potential therapeutic targets and markers for patient stratification, underscoring the framework's ability to identify clinically relevant variables in high-dimensional data.

4.3 Survival Analysis and Model Performance

The AFT model, supported by Bayesian inference, demonstrated robust predictive performance, effectively modeling time-to-event data. Model performance metrics indicated that this approach outperformed traditional Cox PH models Vishwakarma *et al.,* 2022, particularly in handling LUAD's high-dimensional dataset. R packages such as randomForestSRC and survival facilitated visualization of survival outcomes, with Kaplan-Meier plots illustrating the survival probabilities associated with each identified biomarker, enhancing interpretability for clinical application Jiao *et al.,* 2019; Fox & Carvalho, 2012.

4.4 Data Visualization and Interpretability

To complement the statistical analysis, data visualization techniques, including Principal Component Analysis (PCA) and t-SNE, were employed to explore clustering by survival status. These visualizations revealed clear separation between high- and low-risk groups, reinforcing the prognostic relevance of the selected biomarkers. Additional visualization tools such as heatmaps and correlation plots provided further insights into the relationships between gene expressions, aiding in the identification of potential interactions that could inform future research.

4.5 Pathway Enrichment and Biological Relevance

Pathway enrichment analysis linked the identified biomarkers to critical biological pathways involved in tumor progression and metastasis, further validating the clinical relevance of these findings. Integrating pathway enrichment analysis within the survival models demonstrated enhanced interpretability, associating survival outcomes with relevant biological mechanisms in LUAD. Overall, these results establish a robust foundation for using high-dimensional feature selection, survival analysis, and data visualization techniques to improve LUAD prognosis. The framework's successful application to TCGA data demonstrates its practical value and sets the stage for further exploration across diverse high-dimensional datasets in oncology and related fields.

5. Statistical Inference

In high-dimensional survival analysis, traditional inference methods often struggle with the complexity posed by a vast number of predictors, known as the curse of dimensionality. Our study utilized penalized regression techniques and Bayesian methods to provide reliable inference, ensuring accurate estimation, hypothesis testing, and confidence interval construction for LUAD prognostic biomarkers. Penalized Regression and Selective Inference LASSO and Elastic Net were instrumental in handling high-dimensional data, reducing the gene pool to focus on the most significant predictors for survival. Selective inference was applied to the LASSO-penalized model, producing valid p-values and confidence intervals by conditioning on selected variables. This approach provided an interpretable framework to assess the relevance of each biomarker, avoiding the bias introduced by traditional hypothesis testing in regularized models.

For instance, selective inference allowed us to identify CRNDE, IGFBP1, LDLRAD3, and RPS6KL1 as statistically significant biomarkers associated with survival outcomes. By adjusting for selection bias, these biomarkers demonstrated stable predictive power, providing clinicians with reliable markers for potential risk stratification.

5.1 Bayesian Credible Intervals and Inference

The Bayesian framework in afthd allowed for the use of prior information, improving the stability and reliability of estimates in this high-dimensional context. Posterior distributions enabled the construction of credible intervals, offering a probabilistic interpretation of biomarker significance. For example, a 95% credible interval for each selected biomarker indicated the range within which the true effect size lies with 95% probability, given the observed data and prior knowledge Kelter, 2020. This Bayesian approach, particularly useful when dealing with smaller sample sizes in high-dimensional settings, enhances the interpretability of findings by allowing more nuanced probability-based conclusions.

5.2 Handling Missing Data and Model Robustness

Missing data in high-dimensional survival analysis can lead to biased estimates and unreliable inference, compromising the accuracy of survival predictions and model performance. By incorporating mean imputation within the afthd package's functions, we minimized potential biases, ensuring robust statistical inference for survival outcomes. The model demonstrated resilience against the incomplete data typically encountered in genomic studies, enhancing the reliability of our prognostic estimates for LUAD Gabrio *et al.,* 2019; Vishwakarma *et al.,* 2021a Vishwakarma *et al.,* 2023.

5.3 Interpretive Insights and Clinical Implications

The interpretive power of our framework is further augmented through visualizations, such as Kaplan-Meier plots and pathway enrichment analyses, which link survival outcomes to underlying biological mechanisms. These tools aid clinicians in visualizing patient stratification and understanding how each biomarker impacts survival, enhancing decision-making in LUAD management. Additionally, Bayesian credible intervals and selective inference provide a framework to confidently interpret significant biomarkers, ensuring that the findings are clinically meaningful and potentially applicable to precision oncology.

In summary, our framework successfully integrates selective inference, Bayesian credible intervals, and missing data imputation, enabling accurate, interpretable outcomes in high-dimensional survival analysis. By addressing the complexities of high-dimensional inference, this study lays a foundation for further investigations that incorporate advanced modeling and visualization techniques, ultimately supporting the development of reliable, clinically applicable models for LUAD and other cancers.

5.4 Interpretation of Outcomes

Interpreting results in high-dimensional models requires careful consideration due to the complex relationships between predictors and outcomes. Penalization methods often shrink smaller coefficients toward zero, introducing potential bias that must be accounted for when interpreting these estimates. Techniques like debiased LASSO and selective inference help address this issue, offering more accurate estimates and ensuring valid inferential statements.

Bayesian methods add the advantage of a probabilistic interpretation, enabling statements such as, "there is a 95% probability that the true value of the parameter lies within the credible interval." This differs from traditional confidence intervals in frequentist inference, which describe the behavior of intervals over repeated sampling rather than providing a direct probability statement about the parameter.

In summary, recent advances in high-dimensional inference enable reliable hypothesis testing, valid confidence interval construction, and meaningful interpretation of model parameters, even with a large number of predictors relative to observed events.

6. Shiny Application and R Package Overview

This study presents the `afthd` R package, a comprehensive tool for high-dimensional survival analysis using Bayesian AFT models. Designed to address the challenges of high-dimensional data, `afthd` integrates advanced feature selection methods, robust missing data handling, and survival modeling capabilities. By incorporating regularization techniques such as LASSO and Elastic Net alongside Bayesian inference through Markov Chain Monte Carlo (MCMC) simulations, `afthd` provides a reliable, flexible framework tailored to high-dimensional survival data, especially relevant in genomic and clinical research contexts.

6.1 Key Capabilities of the afthd R Package

The `afthd` package includes various tools for high-dimensional survival analysis:

- **Variable Selection**: Utilizing LASSO, Elastic Net, and Bayesian variable selection methods, `afthd` effectively reduces the feature space, identifying the most relevant biomarkers while controlling for overfitting. This approach is particularly advantageous in genomic datasets, where the number of predictors often exceeds the sample size.
- **Missing Data Handling**: Recognizing the prevalence of missing data in clinical datasets, `afthd` incorporates multiple imputation methods to preserve data integrity and minimize potential biases. This functionality ensures robust estimates in survival analysis, even when confronted with incomplete data.
- **Survival Modeling**: Based on the AFT model, `afthd` provides a more flexible alternative to the Cox Proportional Hazards (Cox PH) model, accommodating complex data structures and allowing for more precise survival time predictions in high-dimensional settings. Bayesian inference further enhances model stability by incorporating prior knowledge, yielding credible intervals that offer interpretable probabilistic estimates.

In addition to these core capabilities, the `afthd` package supports a range of parametric distributions, including log-normal, Weibull, and log-logistic models, catering to diverse survival analysis needs. With built-in diagnostic plots for MCMC convergence and posterior distributions, `afthd` facilitates a thorough examination of model performance and stability.

6.2 Shiny Web Application: Enhancing Accessibility for Non-Programmers

To make high-dimensional survival analysis more accessible, we developed a Shiny web application that interfaces with `afthd`. This app is designed to accommodate researchers, clinicians, and analysts who may not have extensive programming knowledge, allowing them to perform complex survival analyses through a user-friendly interface. The Shiny application can be accessed at https://atanu.shinyapps.io/app2/.

Key features of the Shiny application include:

- **Data Upload and Preprocessing**: Users can upload high-dimensional datasets in CSV format directly into the app. Once uploaded, data preprocessing options are provided to ensure data quality before analysis.
- **Interactive Feature Selection**: The Shiny app offers interactive options for selecting variable selection methods, such as LASSO or Elastic Net, enabling users to fine-tune the model for their specific datasets.
- **Survival Analysis and Visualization**: With the click of a button, users can run AFT models and visualize survival outcomes through Kaplan-Meier plots and other interactive graphics. Pathway enrichment analysis is also integrated to help link survival outcomes to biological pathways, further enhancing result interpretability.
- **Real-Time Results and Interpretation**: The Shiny app generates results in real-time, including model coefficients, p-values, and credible intervals, allowing users to easily interpret and export their findings. Interactive visualizations support a more in-depth exploration of the results, aiding in understanding the relationships between predictors and survival outcomes.

By facilitating the functionalities of `afthd`, the Shiny application allows non-programmers to perform high-dimensional survival analyses, visualize results, and explore prognostic markers without needing to write code. This accessibility broadens the use of advanced survival analysis techniques in clinical research, making it a practical tool for various users, from academic researchers to healthcare practitioners.

In summary, the `afthd` R package and Shiny app together offer a versatile, accessible solution for high-dimensional survival analysis. By combining robust variable selection, missing data handling, and survival modeling capabilities with an intuitive, user-friendly interface, this framework supports meaningful, interpretable analyses applicable to LUAD and other complex diseases.

7. Discussion

This study presents a comprehensive analytical framework tailored for the prognostic analysis of LUAD, the most prevalent subtype of lung cancer. Given lung cancer's position as one of the leading causes of cancer-related mortality worldwide, accurate prognosis and early diagnosis are essential to improve patient outcomes. However, LUAD's molecular and genetic heterogeneity complicates the identification of reliable biomarkers, which are crucial for distinguishing it from other lung cancer subtypes and for predicting survival outcomes. Our framework tackles this issue by combining high-dimensional feature selection, survival modeling, and advanced imputation techniques, aiming to make prognostic modeling more robust and clinically meaningful.

The contribution of this work extends to the development of the afthd R package, specifically designed for high-dimensional survival data typical of LUAD and other cancers. This package provides a range of penalized and Bayesian feature selection methods that enable precise identification of relevant biomarkers, despite the high-dimensional nature of genomic and transcriptomic data associated with cancer research. By focusing on techniques like LASSO, Elastic Net, and Bayesian methods, afthd reduces model complexity while retaining critical prognostic factors, thereby enhancing interpretability and clinical relevance.

Moreover, the user-friendly Shiny application developed in this study allows easy access to complex survival analysis, making high-dimensional prognostic modeling accessible to a wider audience, including clinicians and researchers without programming expertise. This application enables users to conduct high-dimensional data analyses, apply multiple feature selection methods, and visualize results interactively, thus bridging the gap between complex statistical modeling and practical, user-centered analysis. We acknowledge that the afthd package and its shiny application are designed exclusively for time-to-event data and are not applicable to other types of data, such as longitudinal data. Additionally, we note that the posterior estimates in the multivariable case are currently limited to scenarios involving up to five covariates due to computational constraints.

In the context of lung cancer, where early detection and precise prognosis remain challenges, our study's contribution lies in its adaptable framework that leverages robust variable selection, missing data imputation, and advanced survival modeling. The integration of these methodologies helps address the specific demands of LUAD research, enabling improved identification of prognostic biomarkers and enhancing the potential for personalized treatment strategies. By developing tools that facilitate the application of these methods in real-world clinical datasets, we hope to contribute to more targeted, effective interventions for LUAD patients and potentially extend this framework to other cancer types with similar analytical challenges.

8. Conclusion

This study introduces an integrated framework that leverages high-dimensional feature selection, robust survival modeling, and advanced missing data handling to improve the prognostic accuracy of LUAD. By developing the afthd R package and a Shiny web application, we provide researchers with powerful tools for high-dimensional survival analysis, enabling precise identification of biomarkers essential for LUAD prognosis. The inclusion of LASSO, Elastic Net, and Bayesian inference within afthd, along with robust imputation methods, addresses common challenges in analyzing complex datasets, thereby enhancing both model interpretability and reliability.

The application of this framework to the LUAD dataset from TCGA demonstrates its utility in identifying key biomarkers linked to survival outcomes. These findings underscore the framework's potential for clinical applications, paving the way for personalized treatment strategies in LUAD and contributing to the broader field of precision oncology. The Shiny app further democratizes access to high-dimensional survival analysis, allowing a broader range of users, including non-programmers, to interactively explore and analyze survival data.

Future research can build upon this work by extending the afthd framework to incorporate additional machine learning techniques, such as deep learning, to capture non-linear relationships and further enhance predictive accuracy. Additionally, expanding the framework's applicability to other cancers and complex diseases would validate its generalizability and adaptability. Integrating pathway enrichment analysis within survival models and exploring multi-omics data integration offer promising directions for deepening insights into disease mechanisms and improving clinical decision-making. Through these ongoing advancements, the framework presented in this study contributes to more accurate and interpretable prognostic modeling in LUAD and beyond.

Acknowledgments

The authors are deeply indebted to the Co-Editor-in-Chief Prof. João Domingos Scalon and learned reviewers for their valuable comments improving the original manuscript's contents and presentation.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization: BATTACHARJEE, A.; KUMARI, P.; VISHWAKARMA, G. K. **Data curation**: BATTACHARJEE, A.; KUMARI, P.; VISHWAKARMA, G. K. **Formal analysis**: BATTACHARJEE, A.; KUMARI, P.; VISHWAKARMA, G. K. **Funding acquisition**: BATTACHARJEE, A.; KUMARI, P.; VISHWAKARMA, G. K. **Investigation**: BATTACHARJEE, A.; KUMARI, P.; VISHWAKARMA, G. K. **Methodology**: BATTACHARJEE, A.; KUMARI, P.; VISHWAKARMA, G. K. **Project administration**: BATTACHARJEE, A.; KUMARI, P.; VISHWAKARMA, G. K. **Software**: BATTACHARJEE, A.; KUMARI, P.; VISHWAKARMA, G. K. **Resources**: BATTACHARJEE, A.; KUMARI, P.; VISHWAKARMA, G. K. **Supervision**: BATTACHARJEE, A.; KUMARI, P.; VISHWAKARMA, G. K. **Validation**: BATTACHARJEE, A.; KUMARI, P.; VISHWAKARMA, G. K. **Visualization**: **Writing – original draft**: BATTACHARJEE, A.; KUMARI, P.; VISHWAKARMA, G. K. **Writing – review and editing:** BATTACHARJEE, A.; KUMARI, P.; VISHWAKARMA, G. K.

References

1. Abdelwahab, O., Awad, N., Elserafy, M. & Badr, E. A feature selection-based framework to identify biomarkers for cancer diagnosis: A focus on lung adenocarcinoma. *Plos One* **17,** e0269126 (2022).

2. Benner, A., Zucknick, M., Hielscher, T., Ittrich, C. & Mansmann, U. High-dimensional Cox models: the choice of penalty as part of the model building process. *Biometrical Journal* **52,** 50–69 (2010).

3. Bhattacharjee, A., Vishwakarma, G. K. & Thomas, A. Bayesian state-space modeling in gene expression data analysis: An application with biomarker prediction. *Mathematical biosciences* **305,** 96–101 (2018).

4. Cho, H.-J., Lee, S., Ji, Y. G. & Lee, D. H. Association of specific gene mutations derived from machine learning with survival in lung adenocarcinoma. *PLoS One* **13,** e0207204 (2018).
5. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *Inter-Journal, Complex Systems* **1695.** https://igraph.org (2006).
6. Fan, J., Feng, Y. & Wu, Y. A Bayesian approach to variable selection in high-dimensional survival data. *Biometrika* **97,** 691–703 (2010).
7. Fan, J. & Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96,** 1348–1360 (2002).
8. Fanizzi, C., De Marco, M. & De Santis, A. Machine learning survival models trained on clinical data to identify high-risk patients with hormone responsive HER2 negative breast cancer. *Scientific Reports* **13,** e8575 (2023).
9. Fox, J. & Carvalho, M. S. The RcmdrPlugin. survival package: Extending the R Commander interface to survival analysis. *Journal of Statistical Software* **49,** 1–32 (2012).
10. Gabrio, A., Mason, A. J. & Baio, G. A full Bayesian model to handle structural ones and missingness in economic evaluations from individual-level data. *Statistics in medicine* **38,** 1399–1420 (2019).
11. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32,** 2847–2849 (2016).
12. Jiao, Y., Li, Y., Jiang, P., Han, W. & Liu, Y. PGM5: a novel diagnostic and prognostic biomarker for liver cancer. *PeerJ* **7,** e7070 (2019).
13. Kaneko, S., Hirakawa, A. & Hamada, C. Enhancing the lasso approach for developing a survival prediction model based on gene expression data. *Computational and Mathematical Methods in Medicine* **2015,** 259474 (2015).
14. Kelter, R. *Statistical Rethinking: A Bayesian course with examples in R and STAN. Taylor & Francis Group.* 2020.
15. Li, R., Chang, C., Justesen, J. M., Tanigawa, Y., Qian, J., Hastie, T., Rivas, M. A. & Tibshirani, R. Fast Lasso method for large-scale and ultrahigh-dimensional Cox model with applications to UK Biobank. *Biostatistics* **23,** 522–540 (2022).
16. Lin, Y., Chen, Y. & Zhang, H. Deep learning-based survival analysis for high-dimensional survival data. *Mathematics* **9,** 1244 (2021).
17. Pedersen, T. L. *ggraph: An implementation of grammar of graphics for graphs and networks* R package version 2.0.5. 2020. https://CRAN.R-project.org/package=ggraph.
18. Qiu, W.-R., Qi, B.-B., Lin, W.-Z., Zhang, S.-H., Yu, W.-K. & Huang, S.-F. Predicting the lung adenocarcinoma and its biomarkers by integrating gene expression and DNA methylation data. *Frontiers in Genetics* **13,** 926927 (2022).
19. Shin, B., Park, S., Hong, J. H., An, H. J., Chun, S. H., Kang, K., Ahn, Y.-H., Ko, Y. H. & Kang, K. Cascaded Wx: A novel prognosis-related feature selection framework in human lung adenocarcinoma transcriptomes. *Frontiers in genetics* **10,** 662 (2019).
20. Sievert, C. *Interactive data visualization for the web* (O'Reilly Media, 2020).
21. Suantari, N. G. A. P. P., Fitrianto, A. & Sartono, B. Comparative study of survival support vector machine and random survival forest in survival data. *BAREKENG: Jurnal Ilmu Matematika dan Terapan* **17,** 1495–1502 (2023).
22. Syed, H., Jorgensen, A. L. & Morris, A. P. SurvivalGWAS_SV: software for the analysis of genome-wide association studies of imputed genotypes with "time-to-event" outcomes. *BMC Bioinformatics* **18,** 1–6 (2017).

23. Vishwakarma, G. K., Bhattacharjee, A. & Banerjee, S. Handling missingness value on jointly measured time-course and time-to-event data. *Communications in Statistics-Simulation and Computation* **52,** 126–141 (2023).

24. Vishwakarma, G. K., Bhattacharjee, A. & Kumar, N. Missing data handling techniques in joint modeling context. *Biomedical Engineering Advances* **2,** 100012 (2021).

25. Vishwakarma, G. K., Kumari, P. & Bhattacharjee, A. Thresholding of prominent biomarkers of breast cancer on overall survival using classification and regression tree. *Cancer Biomarkers* **34,** 319–328 (2022).

26. Vishwakarma, G. K., Thomas, A. & Bhattacharjee, A. A weight function method for selection of proteins to predict an outcome using protein expression data. *Journal of Computational and Applied Mathematics* **391,** 113465 (2021).

27. Wang, H. & Li, R. A selective review on random survival forests for high dimensional data. *Quantitative Bio-Science* **36,** 85–95 (2017).

28. Wang, Y., Li, J. & Zhang, Y. Machine learning for survival analysis. *ACM Computing Surveys* **52** (2019).

29. Wang, Y., Gao, X., Ru, X., Sun, P. & Wang, J. Identification of gene signatures for COAD using feature selection and Bayesian network approaches. *Scientific Reports* **12,** 8761 (2022).

30. Wickham, H. *ggplot2: Elegant graphics for data analysis* (Springer-Verlag, 2016).