Astronomy & Astrophysics manuscript no. main
January 6, 2026

The miRNA Signature: Unveiling Breast Cancer Survival through different Cox Architectures

Matteo Bulgarelli, Giorgia Bertacchini

Università degli Studi di Modena e Reggio Emilia

January 6, 2026

ABSTRACT

Context. Accurate survival prediction is essential for personalized treatment in oncology, yet traditional clinical variables often fail to account for the complex biological heterogeneity of breast cancer. While current prognostic models primarily rely on messenger RNA (mRNA) expression profiles, the regulatory potential of microRNA (miRNA) remains an under-explored dimension that could offer more granular insights into disease progression.

Aims. This study aims to evaluate the prognostic utility of miRNA-Seq data as a robust alternative or complement to established mRNA-centric approaches. The goal is to determine if miRNA can effectively stratify patient risk and provide reliable survival probability estimations across different computational frameworks.

Methods. We compared three analytical pipelines: penalized linear regression (ElasticNet) using non-linear feature extraction (Kernel PCA), deep learning architectures (DeepSurv) paired with linear PCA, and a generative approach utilizing Cox-regularized Variational Autoencoders (CoxVAE) integrated with DeepSurv. These methods were tested across different normalization techniques, including log_2, TPM, and Quantile normalization.

Results. The analysis reveals that while mRNA remains the primary driver of absolute predictive accuracy, miRNA signatures are highly effective predictors of survivability, particularly when processed through non-linear dimensionality reduction. Furthermore, miRNA models demonstrated superior calibration in deep learning contexts compared to mRNA, suggesting a higher reliability in their survival probability estimations. We also observed that increasing model depth in DeepSurv architectures leads to diminishing returns, indicating a point of informational saturation for these datasets.

Conclusions. We conclude that mRNA continues to be the gold standard for predictive power; however, miRNA is a clinically relevant biomarker that captures unique post-transcriptional regulatory mechanisms. The study highlights that for high-dimensional genomic data with limited samples, non-linear feature extraction paired with regularized models offers a more robust path to accurate prognosis than increasing neural network complexity.

1. Introduction

Accurate survival prediction remains a cornerstone of precision oncology, providing the essential framework for personalized treatment strategies and clinical decision-making. In recent years, the integration of high-throughput sequencing with traditional clinical metadata has revolutionized our understanding of disease progression. While demographic and pathological variables—such as age, stage, and grade—offer a baseline for prognosis, they often fail to capture the underlying biological heterogeneity that dictates long-term patient outcomes.

The current State-of-the-Art (SOTA) in genomic survival analysis has largely been defined by the use of messenger RNA (mRNA) expression profiles. Though still a relatively evolving field, mRNA-based prognostic signatures have demonstrated significant success in stratifying patients into distinct risk groups, offering a more granular view of the molecular landscape than clinical data alone.

Recent benchmarks have demonstrated that the integration of mRNA-seq data with clinical metadata significantly improves survival predictions compared to models relying solely on clinical variables. For instance, R. Jardillier et al. (2022) performed an extensive study across 16 TCGA cancer types, confirming that tumor profiling consistently enhances prognostic accuracy. However, most large-scale benchmarks focus on linear prescreening and standard penalized Cox models. Our work builds upon this foundation by exploring whether more sophisticated feature selection strategies and non-linear deep learning architectures can further push the boundaries of predictive performance in Breast Cancer (BRCA).

However, the potential of other non-coding RNA species remains comparatively under-explored. MicroRNAs (miRNAs)—small, non-coding RNA molecules that post-transcriptionally regulate gene expression—present a compelling alternative. Due to their inherent stability in clinical samples and their role as master regulators of oncogenic pathways, miRNA sequencing (miRNA-Seq) data may hold untapped prognostic value that rivals or complements established mRNA-based models.

In this study, we investigate whether miRNA-Seq data, when integrated with standard clinical metadata, can achieve a level of predictive accuracy comparable to or exceeding current mRNA-centric approaches. To rigorously test this hypothesis, we implemented a graduated methodology consisting of three distinct analytical frameworks of increasing complexity.

2. Materials and Methods

2.1. Data Acquisition

The miRNA, mRNA and clinical data was all pulled from the GDC (Genomic Data Commons) Data Portal. We defined our cohort taking TCGA-BRCA for ductal and lobular neoplasms disease, with first diagnosis infiltrating duct carcinoma originating from breast. We then defined the filters:

- For miRNA we choose "miRNA-seq" as experimental strategy, "transcriptome profiling" as data category, "miRNA expression quantification" as data type, "open" access data, "tumor" as tissue type, and "primary" as tumor description
- For the clinical data we choose "clinical" as data category, "clinical supplement" for data type, and again with "open" access.
- For mRNA we defined the same filters as for the case of miRNA changing only "mRNA sequencing" for data category

The clinical files we used are XMLs, each one for a single patient, from which we extracted as features: age at initial diagnosis, vital-status (Dead/Alive), tumor stage, and follow-up data in terms of days to last follow-up in case or days to death. There were also some XML files of type OMF and TXT files that we didn't include, and so we ended up with a total of 771 useful clinical files of mixed dead and alive patients. The reason we did not include OMF (Other Malignancy Form) is that they don't keep clinical information of our interest, and do not contain laboratory or transcriptomic data, such as mRNA or RNA-Seq.

For the miRNA data we took only the files ending with "mirnas.quantification.txt", because the remaining files were mainly logs, and we used as features: folder name, file name, read count and reads per million miRNA mapped. We ended up with 767 useful files, each with 1881 genes reads.

Finally for mRNA we extracted the files ending with "rna_seq.augmented_star_gene_counts.tsv", removing, as for the miRNA, the other log files, and choose as features for only the protein coding genes: gene name, gene id, unstranded, FPKM unstranded and TPM unstranded. We ended up with 787 files, each with around ~20k genes reads.

To build the 2 raw datasets we then joined the clinical data both with the miRNA and mRNA based on the folder name, which is referred to the case ID.

2.2. Data Analysis

For Data analysis we focused particularly on clinical data, where we saw that while tissues were sampled from different sites/hospitals, they were all sent to the same institute for analysis, which is the Nationalwide Children's Hospital, and that there is quite a big gap between alive and dead patients, since we have only 75 dead patients and 696 alive[1]. We also investigated the stages of cancer of the patients we worked with to try to understand what type of survivability results we may be expecting (See appendix A.1, A.2), and since we were interested mainly in dead patients and people with high days to last follow-up, we plotted also those distributions to better visualize the data (See appendix A.3, A.4).

2.3. Data Pre-processing

In the first place we performed one-hot encoding on the categorical cancer stage column, then we removed the outlier for what concerns the age, removing those patients that fall inside a scarcely represented range, picked as patients only those that are dead or have a 'days to last follow-up', where the threshold is set to the 25th percentile of the feature value: with these first steps

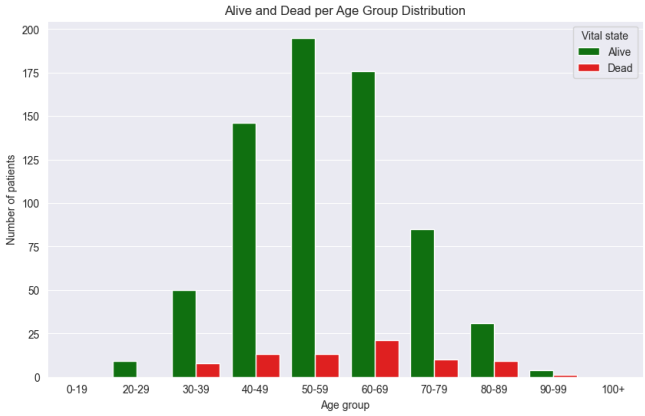

Fig. 1: Plot showing the ration between dead and alive patients grouped by age buckets of 10 years each

we drastically reduced the samples of our datasets to around 300 samples both for miRNA and mRNA.

We proceeded to apply different normalization techniques based on the type of data, miRNA or mRNA, in order to produce different datasets to try to see if different normalization techniques improved the results achieved by the models. In particular for miRNA we made 2 datasets, the first obtained by normalizing the genetic features with just a log_2, and the second by normalizing them with quantile normalization (B. Bolstad et al. (2003)). For mRNA also we made 2 distinct datasets obtained by normalizing the genetic columns, one again by simply applying a log_2 and the other one by taking the already provided TPM normalized reads. In this way we produced 4 different datasets to test and confront the models on.

The last step before effectively saving the CSV files was to remove from each a first set of columns that had variance very low, using as threshold the 50th percentile variance value of the entire dataset, to perform a first simple filtering on features and reduce them. Since the very high dimensionality of the genetic columns in each dataset, especially the mRNA ones, we took a pretty high threshold to possibly light the job of the models that would process them, in order for them to concentrate on a smaller set of features.

Before passing each dataset through each model, genetic features and age at initial diagnosis of patients were also scaled with a Standard Scaler to achieve mean 0 and unitary standard deviation.

2.4. The 3 models framework

To investigate the prognostic utility of miRNA sequencing data compared to mRNA, we developed a multi-stage analytical framework consisting of three models of increasing architectural complexity. Each model was independently trained and evaluated on both data modalities to assess whether miRNA-based survival analysis achieves comparable or superior performance to established mRNA-based methods.

Penalized Cox Regression with Kernel PCA: Our baseline approach utilized an initial ElasticNet Cox model (N. Simon et al. (2011)) for feature selection, followed by Kernel Principal Component Analysis (KPCA)(B. Schölkopf et al. (1997), N. K. Speicher & N. Pfeifer (2017)) to handle non-linear structures during dimensionality reduction. The resulting components were then integrated into a final ElasticNet-penalized Cox model

to predict patient survival outcomes. We choose to use an ElasticNet instead of a standard Lasso (R. Tibshirani (1997)) or Ridge (V. H. H. Verweij PJ (1994)) penalty to overcome the 2 downsides these method have, which are: not being able to select more features than the number of samples of the dataset, and in case of Lasso, with a subset of highly correlated features, it would have randomly chosen one among the set. The ElasticNet overcomes these problems by combining the 2 penalties together by solving:

$$\arg\max_{\beta} \quad \log PL(\beta) - \alpha \left(r \sum_{j=1}^{p} |\beta_j| + \frac{1-r}{2} \sum_{j=1}^{p} \beta_j^2 \right)$$

Exploiting in this way the subset selection of Lasso, and the regularization strength of Ridge. In this first approach, after the appropriate part of data is scaled, only the genetic features are used first with an ElasticNet to perform a first step of feature selection by choosing the best penalization coefficient α: we propose 2 candidate values of alpha, one computed directly from the Randomized Cross Validation (**RCV alpha**) search we use to investigate values inside a certain range, and another one selected by a custom filter we defined (**filtered alpha**) that searches for the best alpha among the ones tested, that specifically keeps a number of active columns between 100 and 200. With the 2 alphas being selected we get the list of features with non 0 coefficient, and so the dataset is reduced to these features, which are then passed through the KPCA to further reduce them to 50 components. These components are then merged back with the clinical data, to form the actual set of data that is finally given to another ElasticNet to make the survival analysis predictions.

Linear PCA and DeepSurv: In the second configuration, we employed standard linear Principal Component Analysis (PCA) as an initial dimensionality reduction step (Y. Chen (2025), Deepali et al. (2025)). PCA performs a linear orthogonal projection that preserves maximal variance in the data, yielding a computationally efficient and interpretable low-dimensional representation. These reduced features were subsequently used as input to two DeepSurv architectures (J. Katzman et al. (2018)), featuring 3-layer and 5-layer network configurations, respectively, enabling the modeling of complex non-linear relationships between genomic features and patient mortality beyond the capacity of linear survival models.

DeepSurv is a neural extension of the classical Cox proportional hazards model in which the linear risk function is replaced by a feed-forward neural network. Specifically, the network learns a non-linear risk score $f_\theta(\mathbf{x})$ from the input covariates $\mathbf{x}$, while preserving the proportional hazards assumption. Model parameters are optimized by minimizing the negative log partial likelihood of the Cox model, defined as

$$\mathcal{L}_{\text{Cox}}(\theta) = -\sum_{i:\delta_i=1} \left(f_\theta(\mathbf{x}_i) - \log \sum_{j \in \mathcal{R}_i} \exp\left(f_\theta(\mathbf{x}_j)\right) \right),$$

where δ_i denotes the event indicator and $\mathcal{R}_i$ is the risk set at time t_i. This formulation naturally accommodates right-censored survival data and yields relative risk estimates directly comparable to those obtained from traditional Cox regression, while allowing for the learning of non-linear co-variate effects and higher-order feature interactions.

Variational Autoencoder (CoxVAE) and DeepSurv: In the final configuration, linear dimensionality reduction was replaced by a non-linear generative approach. We implemented a Cox-regularized Variational Autoencoder (CoxVAE) to learn a probabilistic latent representation by optimizing a variational lower bound while incorporating survival-related constraints. In contrast to PCA, which captures only linear correlations, the Variational Autoencoder models complex non-linear dependencies in the genomic feature space and enforces a structured latent distribution. After convergence, the encoder was frozen and used to extract low-dimensional embeddings, which were subsequently provided as input to the 3-layer and 5-layer DeepSurv networks for final survival estimation. The choice of VAEs is supported by recent developments in deep survival modeling, such as the AUTOSurv framework L. Jiang et al. (2024)), which demonstrates that VAE-based feature extraction effectively captures non-linear biological structures that linear methods fail to represent. By utilizing a VAE to compress the initial feature space of miRNA and mRNA into a lower-dimensional latent manifold, we provide a more informative and regularized input for the downstream DeepSurv architecture, thereby mitigating the risk of overfitting typically associated with deep learning in bio informatics.

2.5. Evaluation

The main metric used to evaluate the models was the Concordance index (F. E. Harrell et al. (1982)) which is a rank-based metric that measures a model's ability to correctly stratify patients by their predicted risk of an event. In survival analysis, it represents the probability that a patient with a higher predicted risk will experience the event sooner than one with a lower risk, effectively handling right-censored data. For this study, the C-index serves as the primary validation tool to determine if miRNA-based features can achieve a discriminatory power comparable to mRNA. By focusing on relative risk ranking rather than absolute time-to-event, it provides a robust benchmark for evaluating the prognostic signal across our three modeling architectures.

Another metric we used, this time only for the final validation of the models is the Brier Score. It's a metric used to evaluate the accuracy of probabilistic survival predictions at specific time points. Unlike the C-index, which focuses on ranking patients, the Brier Score measures the calibration of a model by calculating the mean squared difference between the predicted probability of survival and the actual status (1 if alive, 0 if deceased). When integrated over time, we get the Integrated Brier Score (IBS), which provides a global assessment of model performance across the entire study duration. A score of 0 represents a perfect prediction, while a score of 0.25 typically indicates a model that is no more informative than a random guess.

3. Results

3.1. ElasticNet+KPCA

Based on the results shown (see appendix table B.1), we can draw several insights regarding the impact of data type (miRNA vs. mRNA), normalization methods, and the feature selection process (RCV alpha vs. filtered alpha).

– **mRNA vs. miRNA Performance**: In general, mRNA datasets outperform miRNA datasets in terms of predictive accuracy. The mRNA datasets (both log_2 and TPM) consistently achieve higher mean C-index values, reaching up

to 0.923 and 0.987 (max) with filtered alpha. In contrast, miRNA datasets peak at a mean of 0.882. For IBS (Integrated Brier Score) lower values indicate better calibration and the mRNA models generally show lower IBS values (mean ≈ 0.074–0.086) compared to miRNA (mean ≈ 0.081–0.127), suggesting mRNA features provide a more reliable survival probability estimation. While mRNA datasets demonstrate superior overall performance, the high C-index achieved by miRNA (particularly in the log_2 filtered configuration) confirms that micro-RNA signatures remain robust and clinically relevant predictors of survivability, likely reflecting critical regulatory layers not fully captured by the transcriptome alone.

- **Impact of Feature Filtering (RCV alpha vs. filtered alpha)**: The "filtered alpha" approach leads to a significant and consistent improvement across almost all metrics and datasets. In the miRNA log_2 dataset, the mean C-index jumps from 0.686 to 0.882. A similar trend is visible in mRNA TPM, moving from 0.719 to 0.923. A further improvement can be seen in terms of stability, where the standard deviation (std) of the C-index typically decreases when moving from RCV to filtered alpha (e.g., in mRNA log_2 it drops from 0.063 to 0.040), indicating that the filtered feature set produces more robust and less volatile results.
- **Normalization Comparison (Within the same base)**: For miRNA, log_2 normalization is superior to Quantile normalization when using filtered alpha (C-index 0.882 vs 0.717). Interestingly, with RCV alpha, they perform similarly, but log_2 scales much better after filtering. Both mRNA normalizations perform exceptionally well. However, mRNA TPM achieves the highest overall mean C-index (0.923) and the lowest mean IBS (0.074) under the filtered alpha configuration, making it the top-performing setup in this study.
- **Error and Variance (std)**: The miRNA Quantile dataset appears to be the most "unstable" or difficult for the model to learn from when filtered, as it is the only case where the C-index standard deviation actually increases (from 0.087 to 0.123) and the mean performance remains relatively low. The mRNA log_2 dataset with filtered alpha shows the lowest variance in C-index (0.040), suggesting this combination is highly reproducible.

3.2. PCA+DeepSurv

This architecture shows different outcomes that are not always completely coherent (see appendix table C.1). As the complexity of the head increases from 3 to 5 layers, only a marginal improvement can be seen: the mRNA TPM dataset shows the most notable improvement, with the mean C-index rising from 0.706 to 0.753, but conversely for miRNA with quantile normalization there is a decrease in mean performance, suggesting maybe that a deeper model may lead only to overfitting or optimization difficulties.

Consistently with previous observations, mRNA datasets maintain a higher performance ceiling compared to miRNA, especially when the architecture is deeper, even though for IBS, mRNA actually exhibits slightly worse values, compared for example to miRNA with log_2.

Different normalization technique does not seem to particularly impact in the case of mRNA, while there's a considerable change in the case of miRNA between log_2 and quantile. A final consideration is that more depth does not always correspond to an improvement also in terms of stability.

3.3. VAE+DeepSurv

This architecture shows a more coherent and explainable behavior (see appendix table D.1). In general the increase of complexity, from 3 to 5 layers, show an improvement in performances both in terms of C-index and standard deviation, where the most significant jump is seen, for C-index, in mRNA TPM, which increases from 0.776 to 0.820 and in miRNA log_2, where the std drops from 0.072 to 0.049.

Consistent with previous analyses, mRNA datasets continue to hold a performance advantage over miRNA, although the gap is narrower in this specific test. Moreover, also the IBS is generally better in miRNA than mRNA, as in previous tests.

Also the normalization technique for each genetic source has the same impact on performances as before, where the log_2 normalization is better for miRNA, while it's TPM that for mRNA "seals the deal".

3.4. Models interpretability

To enhance model interpretability and evaluate the contribution of individual features to the survival predictions, we employed SHAP (SHapley Additive exPlanations). Specifically, for the DeepSurv architectures, SHAP values were calculated to quantify the impact of each genomic and clinical variable on the predicted risk score. This approach allows for a granular understanding of how specific miRNA or mRNA signatures drive the model's decision-making process, providing a consistent measure of feature importance that accounts for non-linear interactions within the neural network.

From our analysis the "age at initial diagnosis" is consistently at the top for the PCA+DeepSurv architecture, while in the non-linear one, it has impact only for the mRNA datasets. Other clinical data is nearly completely absent and it could be due to 3 main factors:

- Informative saturation in the genetic data: The high-dimensional genomic features (miRNA/mRNA) likely capture the intrinsic biological aggressiveness of the tumor more comprehensively than traditional clinical staging
- Age is an orthogonal variable: With respect to other clinical features that are deducible from genetic signature, age may be not, and so it's used as a complementary useful information by the model. The fact that DeepSurv quite consistently relies on age alongside genomic embeddings indicates that the model is successfully integrating two distinct dimensions of survival risk.
- By reducing the dimensionality of the vectors, the new features encapsulate more information, making less informative features being cut out.

Age at initial diagnosis emerges as a significant feature in the SHAP analysis for the mRNA dataset, whereas it is less prominent in the miRNA models. We attribute this discrepancy to the inherent predictive strength of age observed in the PCA+DeepSurv framework, where it demonstrated a markedly higher impact on survival within the mRNA cohort. Consequently, this clinical feature appears to have exhibited greater 'resilience' during the non-linear transformation performed by the Variational Autoencoder.

4. Discussion

4.1. ElasticNet vs. DeepSurv Architectures

Observing the performance across the three tables, the ElasticNet approach using Kernel PCA (KPCA) often performs equal to, or in several cases better than, the DeepSurv architectures. Specifically, the mRNA TPM dataset reaches its highest mean C-index of 0.923 under ElasticNet with filtered alpha, a value not matched by the DeepSurv models. This superiority likely stems from the scarcity of the dataset in terms of sample size; deep learning models typically require larger cohorts to generalize effectively without overfitting. Furthermore, the performance gap between Table B.1 and C.1 suggests that while DeepSurv was paired with linear PCA, the ElasticNet approach benefited from KPCA, which is better equipped to capture the complex non-linear relationships inherent in genomic features. Our results for mRNA-based survival prediction align with the findings of R. Jardillier et al. (2022) regarding the superiority of omic-integrated models over clinical-only models. However, while their benchmark reported a median C-index of approximately 0.64 for the BRCA dataset using standard Lasso-like penalization, our 'filtered alpha' ElasticNet and VAE-DeepSurv architectures achieved significantly higher performance (peaking at a mean C-index of 0.923). We attribute this out-performance to our bi-dimensional filtering strategy and the ability of non-linear latent representations (VAE and KPCA) to capture complex biological signals that traditional linear pre-screening methods may overlook. Our findings regarding the comparative performance of genomic modalities are also consistent with recent ensemble learning benchmarks. Specifically, the one of J. Yuan et al. (2025) reported that while multi-omic integration yields the highest classification accuracy for breast cancer progression, unimodal mRNA models consistently outperform miRNA-only approaches (achieving AUCs of 0.654 and 0.612, respectively). Our results mirror this hierarchy; however, our specialized architectures—such as the Filtered ElasticNet and VAE-DeepSurv—reach significantly higher performance ceilings. This suggests that while mRNA remains the more informative standalone modality, the gap between the two can be narrowed through the use of non-linear feature extraction and rigorous alpha-filtering.

4.2. Impact of Variational Autoencoders (VAE)

Comparing the two DeepSurv-based tables (C.1 and D.1), it is evident that using a VAE for dimensionality reduction instead of linear PCA consistently yields superior results. For instance, in the miRNA log_2 dataset with 5 layers, the mean C-index improves from 0.695 with linear PCA to 0.782 with VAE. This indicates that the generative and non-linear nature of VAEs provides a more informative latent representation for the downstream survival model than standard linear projections.

4.3. Architectural Complexity and Diminishing Returns

Across both DeepSurv tables (PCA and VAE-based), the transition from a 3-layer to a 5-layer head does not result in drastic performance improvements. While minor gains are observed—such as mRNA TPM moving from 0.776 to 0.820 in the VAE configuration—the overall stability (standard deviation) and calibration (IBS) remain largely comparable. This suggests that for these specific datasets, the 3-layer architecture already reaches a point of informational saturation, and additional depth may only increase the risk of overfitting given the limited sample size.

4.4. Redundancy in KPCA + ElasticNet

An important methodological observation is the synergy between KPCA and ElasticNet. Because KPCA projects data into a set of orthogonal principal components, it inherently resolves the issue of feature co-linearity. Consequently, the Ridge penalty (ℓ_2) of the ElasticNet becomes largely redundant, often tending toward zero. In this context, the model effectively collapses into a Lasso-like regressor on non-linear components, focusing primarily on feature sparsity rather than the grouping effect typically sought through the ElasticNet penalty. This means that the final ElasticNet used could be actually substituted by a standard Lasso Penalized Cox model, that would have probably returned very close or equal results in terms of performances.

4.5. Data Normalization and Type

It is worth noting that mRNA TPM consistently represents the most predictive data source across all models and pre-processing pipelines. Moreover, the "filtered alpha" strategy in Table B.1 demonstrates that aggressive feature selection is a critical driver of performance, often outweighing the choice of the survival model itself. This highlights that in the low-sample, high-feature regime of bio informatics, pre-processing and feature pruning remain as critical as model architecture.

5. Conclusions

This research evaluated the predictive performance of various genetic data treatments—specifically comparing miRNA and mRNA—using a combination of linear and non-linear feature extraction methods paired with ElasticNet and DeepSurv models.

Our findings indicate that ElasticNet, when combined with Kernel PCA (KPCA), often performs on par with or superior to complex DeepSurv architectures. This outcome suggests that for bio informatics datasets characterized by a low sample-to-feature ratio, regularized linear models remain highly effective. The performance gap between traditional and deep learning approaches in this study is likely attributed to the limited number of samples, which may prevent deep neural networks from fully converging or generalizing beyond the training data.

In the context of DeepSurv, increasing the network depth from 3 to 5 layers did not result in drastic performance improvements. Instead, we observed diminishing returns, where the 3-layer architecture was often sufficient to capture the essential patterns in the data. This suggests that for these specific genetic modalities, a moderately complex model provides the best balance between learning capacity and the risk of overfitting.

The "filtered alpha" approach consistently outperformed the RCV alpha strategy, reinforcing that rigorous feature pruning is essential for maximizing predictive accuracy in genomic survival analysis.

While mRNA consistently achieved the highest overall performance (with mRNA TPM and filtered alpha reaching a mean C-index of 0.923), miRNA demonstrated that it is a valid and robust predictive indicator. With a mean C-index peaking at 0.882 and showing superior calibration (lower IBS) in VAE-based deep learning models, miRNA represents a powerful alternative or complementary biomarker. It captures a distinct layer of post-

transcriptional regulation that provides significant insights into patient survivability.

In conclusion, the most effective pipeline for survival prediction in this study involved aggressive feature selection and the use of non-linear dimensionality reduction (KPCA or VAE). While mRNA remains the gold standard for predictive accuracy, miRNA offers highly competitive results and better model calibration in deep learning contexts, making it an invaluable asset for multi-omic survival modeling.

References

Bolstad, B., Irizarry, R., Åstrand, M., & Speed, T. 2003, A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias, Bioinformatics, 19, 185, doi: 10.1093/bioinformatics/19.2.185

Chen, Y. 2025, Applying Principal Component Analysis to Optimize Feature Selection in Gene Expression Data: A Case Study on Cancer Classification, Theoretical and Natural Science, 92, 95, doi: 10.54254/2753-8818/2025.21792

Deepali, Goel, N., & Khandnor, P. 2025, DeepOmicsSurv: a deep learning-based model for survival prediction of oral cancer, Discov. Oncol.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. 1982, Evaluating the yield of medical tests., JAMA, 247 18, 2543. https://api.semanticscholar.org/CorpusID:23344910

Jardillier, R., Koca, D., Chatelain, F., & Guyon, L. 2022, Prognosis of lasso-like penalized Cox models with tumor profiling improves prediction over clinical data alone and benefits from bi-dimensional pre-screening, BMC Cancer, 22, 1045, doi: 10.1186/s12885-022-10117-1

Jiang, L., Xu, C., Bai, Y., et al. 2024, Autosurv: interpretable deep learning framework for cancer survival analysis incorporating clinical and multi-omics data, NPJ Precis. Oncol

Katzman, J., Shaham, U., Cloninger, A., et al. 2018, DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network, BMC Medical Research Methodology, 18, doi: 10.1186/s12874-018-0482-1

Schölkopf, B., Smola, A., & Müller, K.-R. 1997, Kernel principal component analysis, in Artificial Neural Networks — ICANN'97, ed. W. Gerstner, A. Germond, M. Hasler, & J.-D. Nicoud (Berlin, Heidelberg: Springer Berlin Heidelberg), 583–588

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. 2011, Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent, Journal of Statistical Software, 39, doi: 10.18637/jss.v039.i05

Speicher, N. K., & Pfeifer, N. 2017, Towards multiple kernel principal component analysis for integrative analysis of tumor samples, J. Integr. Bioinform.

Tibshirani, R. 1997, The lasso method for variable selection in the Cox model, Stat Med

Verweij PJ, V. H. H. 1994, Penalized likelihood in Cox regression, Stat Med

Yuan, J., Xu, P., Ye, Z., & Liu, W. 2025, STmiR: A Novel XGBoost-based framework for spatially resolved miRNA activity prediction in cancer transcriptomics, PLOS ONE, 20, 1, doi: 10.1371/journal.pone.0322082

Appendix A: Data Analysis plots

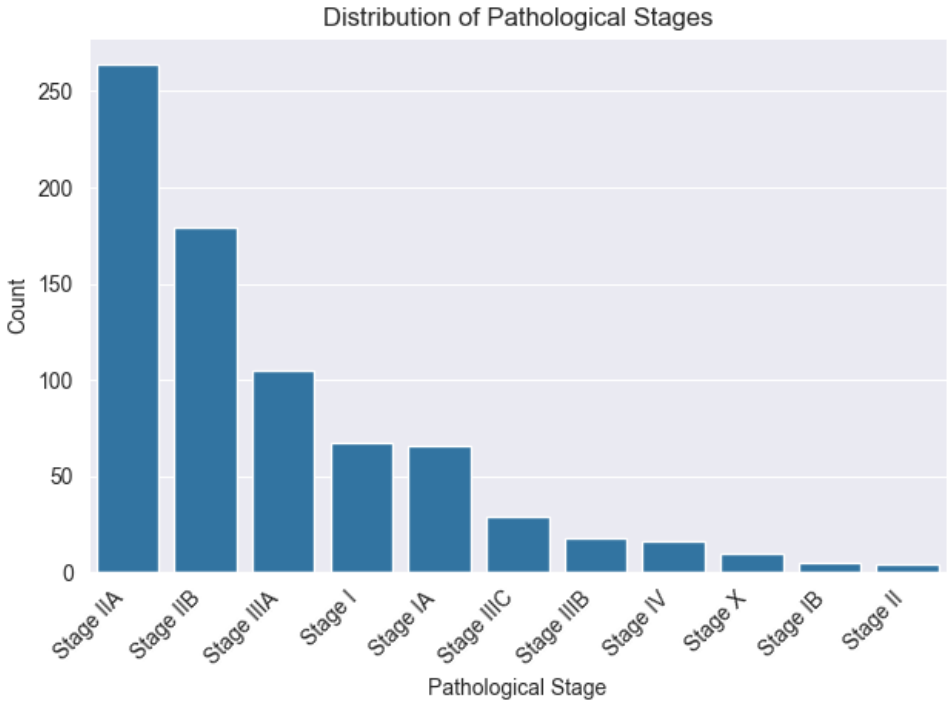

Fig. A.1: Distribution of cancer stages

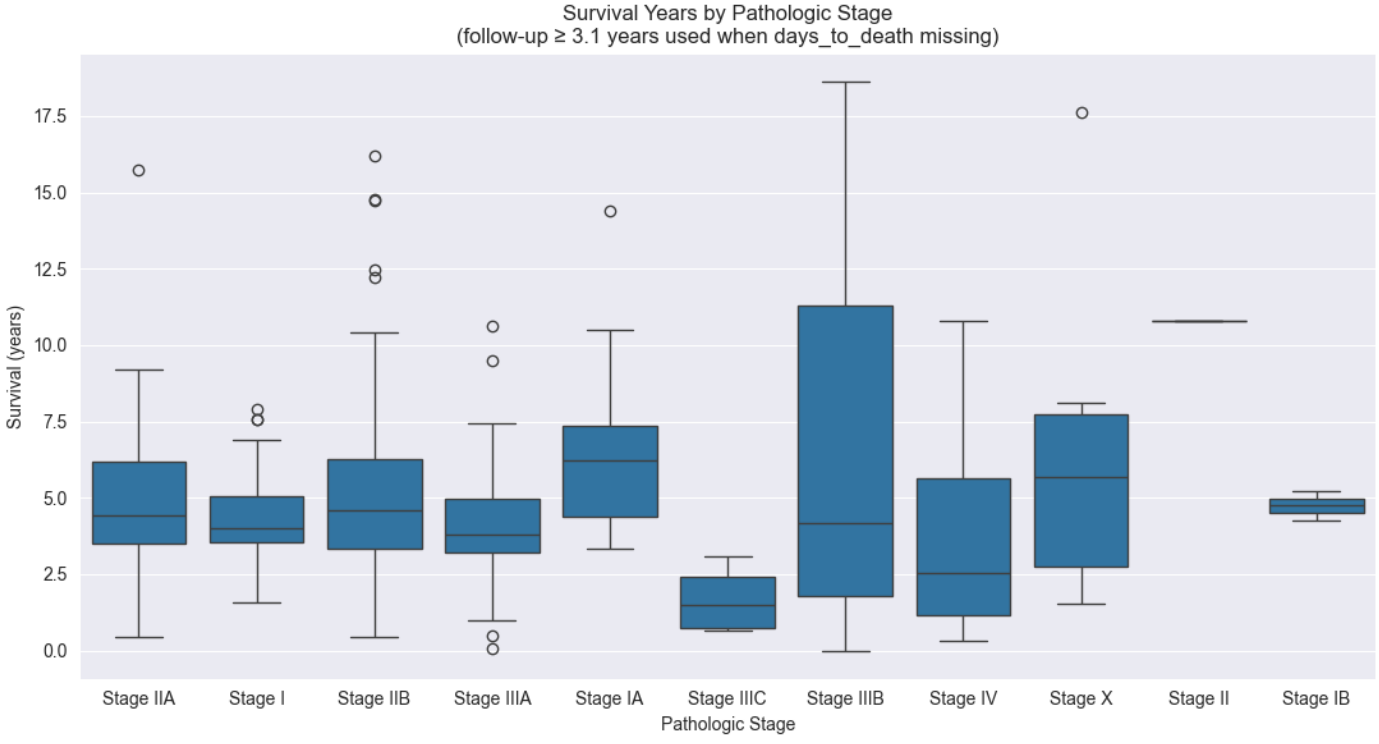

Fig. A.2: Survival time for dead patients and patients with high last days to follow-up (<3.1 years

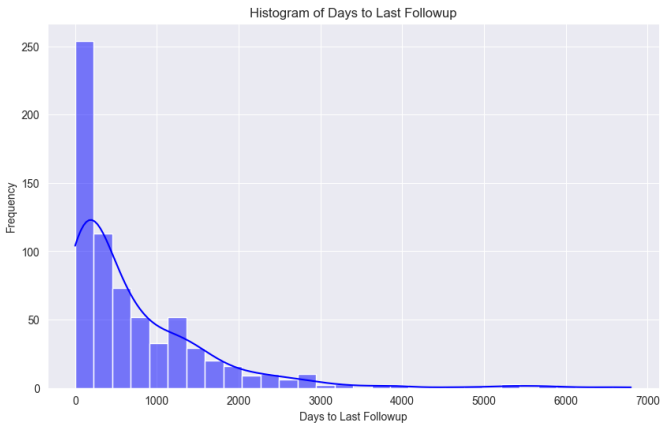

Fig. A.3: Distribution of days to last followup

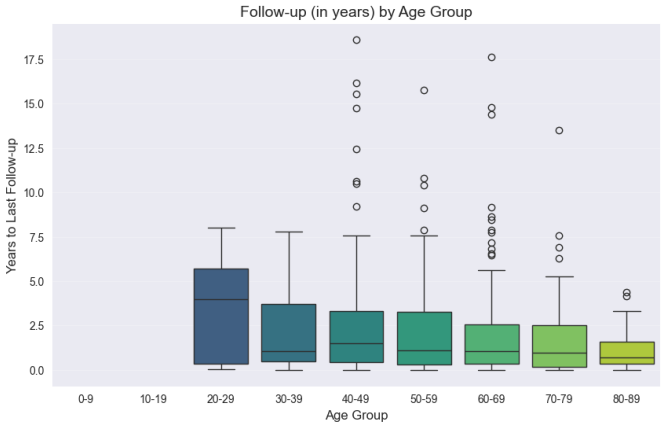

Fig. A.4: Years to last followup grouped by age in buckets of 10 years each

Appendix B: ElasticNet+KPCA tests and plots

Table B.1: Different datasets performances using ElasticNet and KernelPCA

		RCV alpha		filtered alpha	
Dataset	**Stat.**	**C-index**	**IBS**	**C-index**	**IBS**
miRNA log_2	mean	0.686	0.125	0.882	0.081
	std	0.123	0.044	0.066	0.031
	min	0.454	0.041	0.727	0.030
	max	0.904	0.183	0.958	0.139
miRNA quantile	mean	0.700	0.120	0.717	0.127
	std	0.087	0.036	0.123	0.038
	min	0.545	0.036	0.495	0.068
	max	0.841	0.179	0.865	0.186
mRNA log_2	mean	0.876	0.086	0.900	0.079
	std	0.063	0.026	0.040	0.022
	min	0.739	0.038	0.82	0.041
	max	0.961	0.120	0.961	0.111
mRNA TPM	mean	0.719	0.122	0.923	0.074
	std	0.090	0.029	0.042	0.020
	min	0.542	0.074	0.86	0.046
	max	0.843	0.178	0.987	0.111

Fig. B.1: miRNA log_2 plots

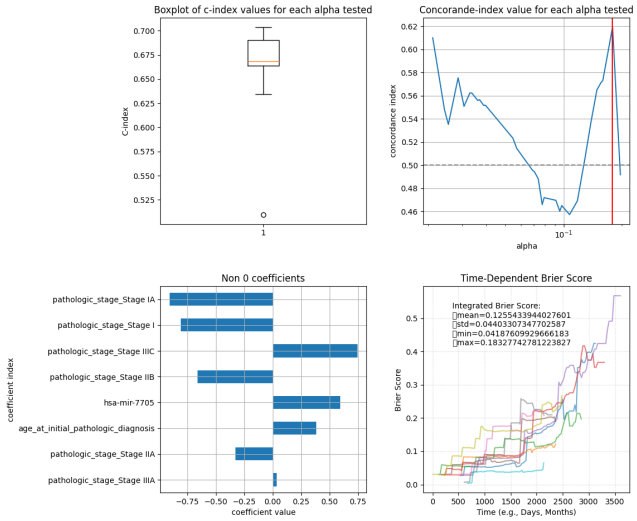

Fig. B.2: with RCV α

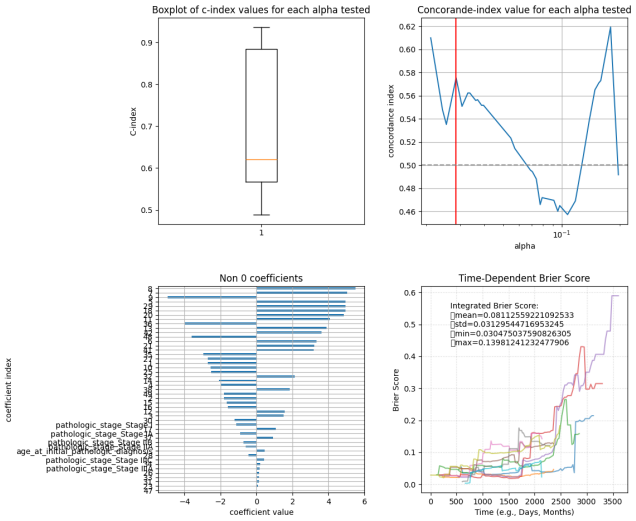

Fig. B.3: with filtered α

Fig. B.4: miRNA quantile plots

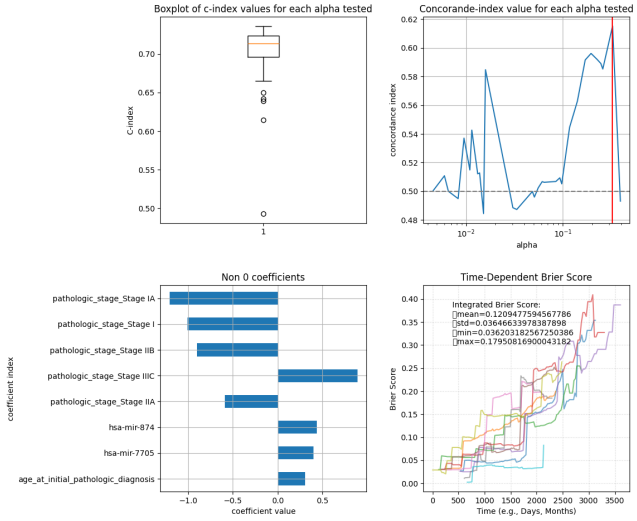

Fig. B.5: with RCV α

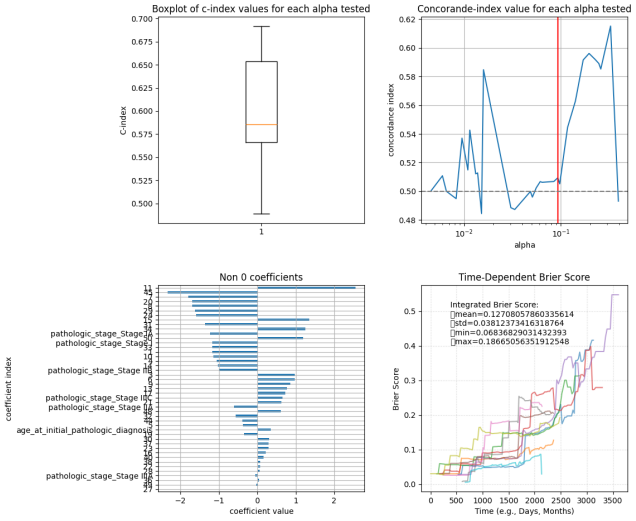

Fig. B.6: with filtered α

Fig. B.7: mRNA log_2 plots

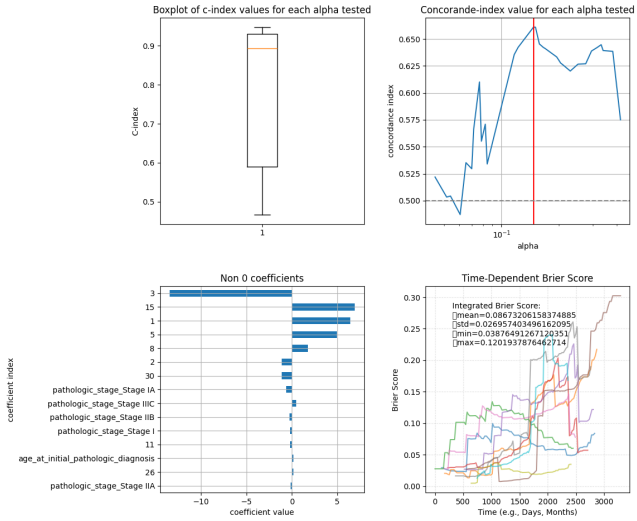

Fig. B.8: with RCV α

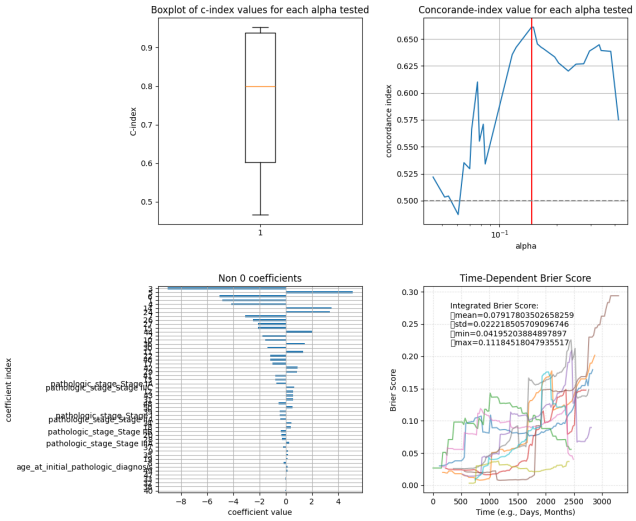

Fig. B.9: with filtered α

Fig. B.10: mRNA TPM plots

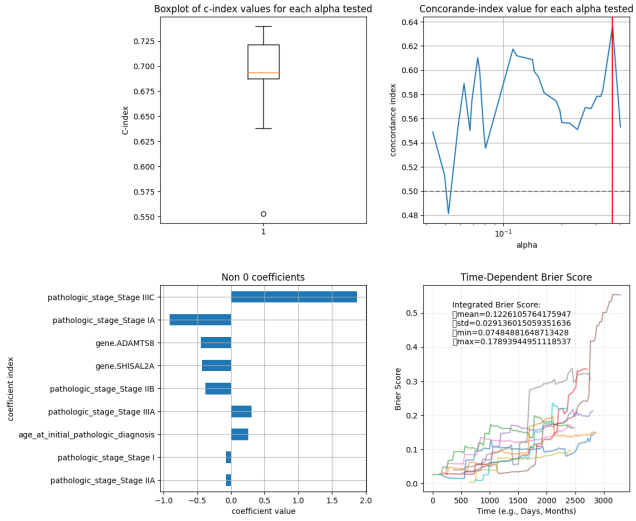

Fig. B.11: with RCV α

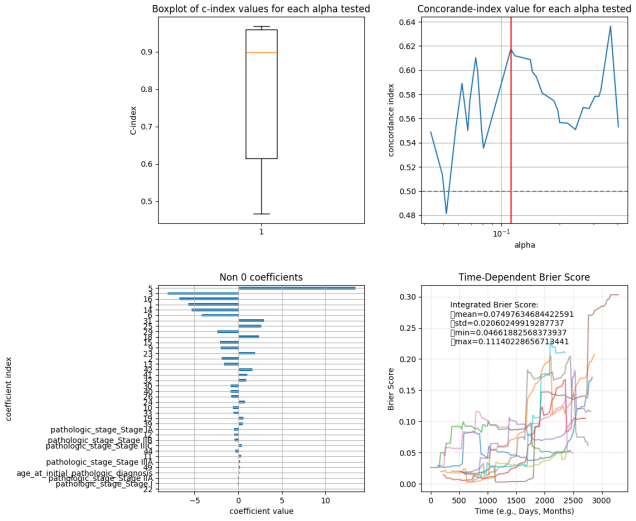

Fig. B.12: with filtered α

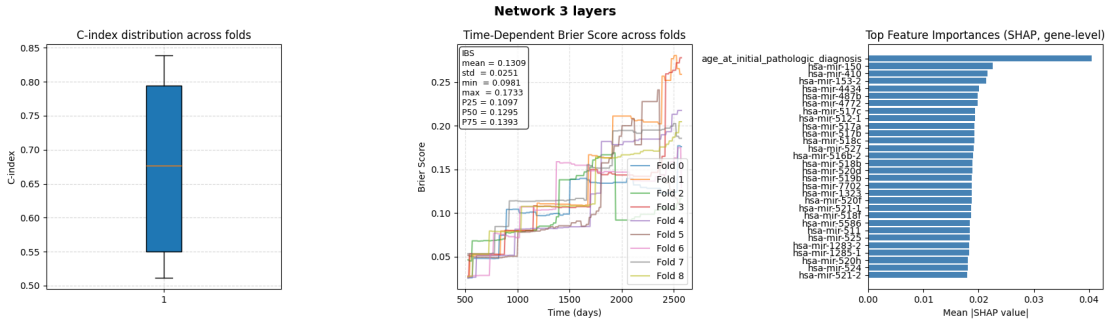

Fig. C.1: miRNA log_2 3-layers

Appendix C: PCA+DeepSurv results and plots

Table C.1: Different datasets performances using linear PCA and DeepSurv networks

		3-layers		5-layers	
Dataset	**Stat.**	**C-index**	**IBS**	**C-index**	**IBS**
miRNA log_2	mean	0.679	0.130	0.695	0.130
	std	0.126	0.025	0.122	0.025
	min	0.511	0.098	0.500	0.094
	max	0.838	0.173	0.896	0.184
miRNA quantile	mean	0.656	0.142	0.612	0.135
	std	0.176	0.045	0.145	0.037
	min	0.375	0.074	0.429	0.059
	max	0.864	0.220	0.907	0.186
mRNA log_2	mean	0.684	0.155	0.701	0.147
	std	0.073	0.032	0.068	0.032
	min	0.585	0.104	0.602	0.125
	max	0.786	0.227	0.818	0.234
mRNA TPM	mean	0.706	0.158	0.753	0.151
	std	0.120	0.052	0.125	0.042
	min	0.468	0.117	0.550	0.102
	max	0.888	0.293	0.886	0.250

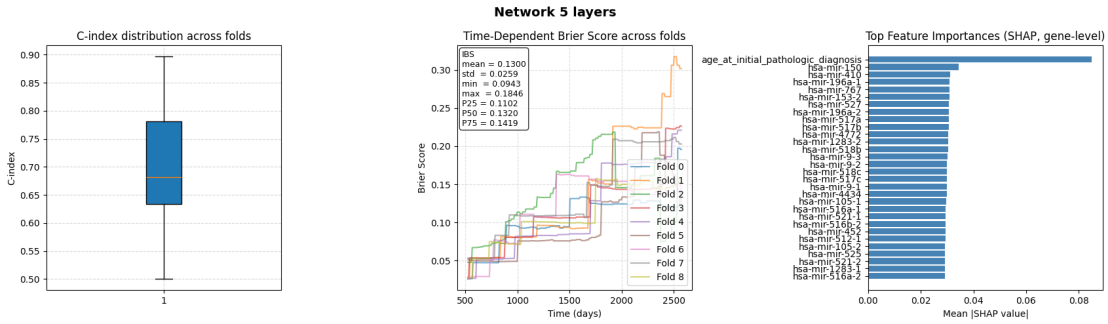

Fig. C.2: miRNA log_2 5-layers

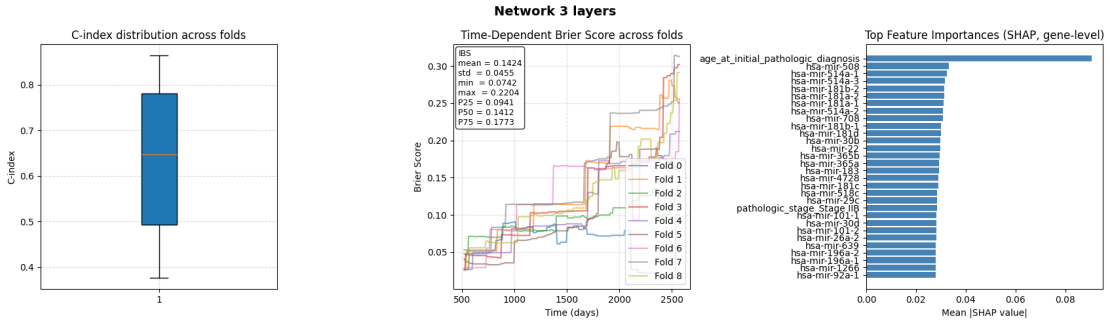

Fig. C.3: miRNA quantile 3-layers

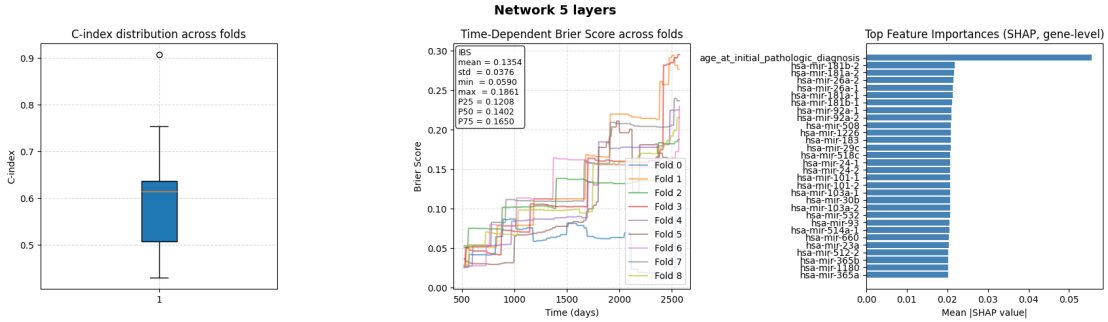

Fig. C.4: miRNA quantile 5-layers

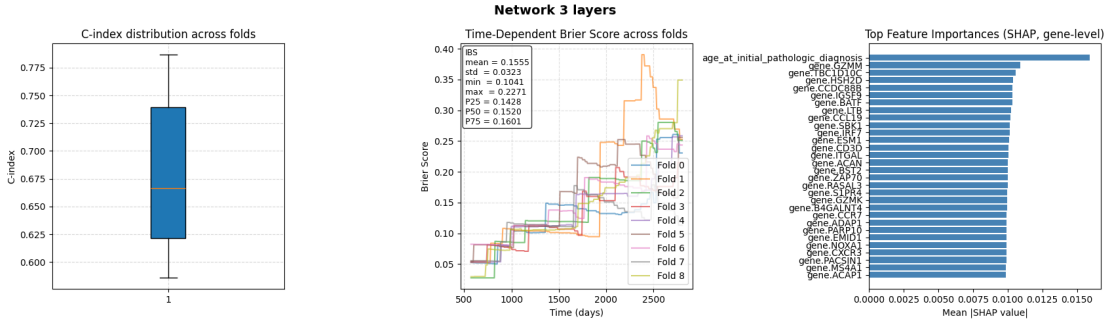

Fig. C.5: mRNA log_2 3-layers

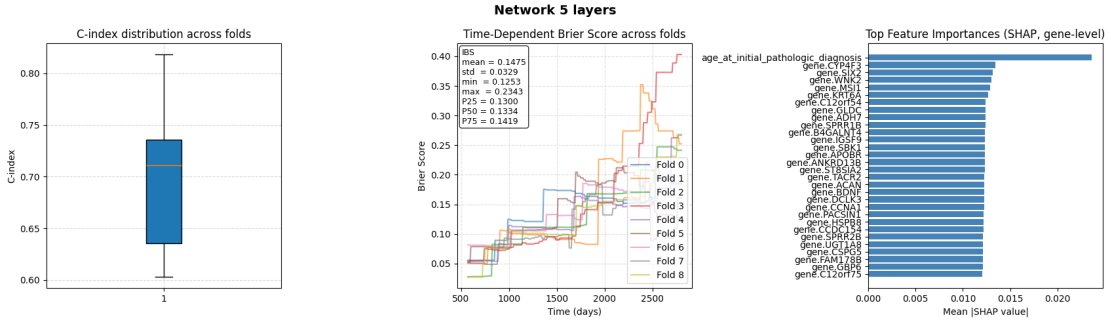

Fig. C.6: mRNA log_2 5-layers

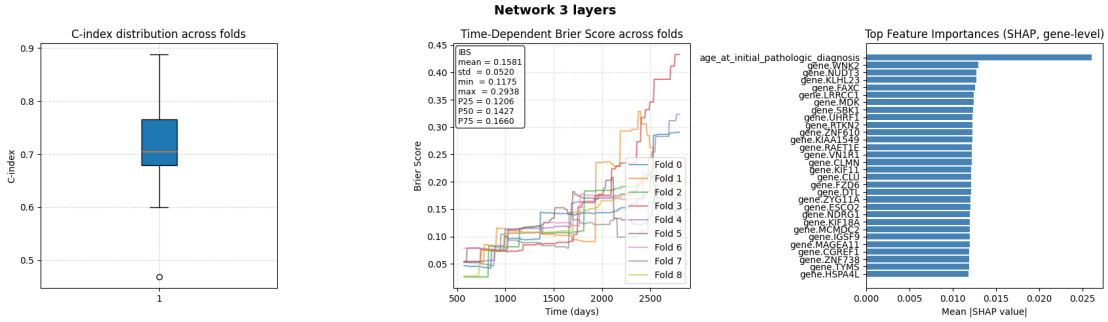

Fig. C.7: mRNA TPM 3-layers

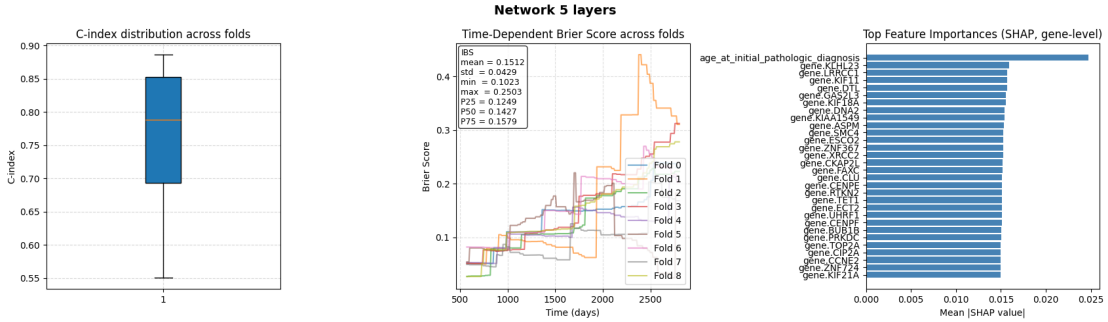

Fig. C.8: mRNA TPM 5-layers

Appendix D: VAE+DeepSurv results and plots

Table D.1: Different datasets performances using linear PCA and DeepSurv networks

		3-layers		5-layers	
Dataset	**Stat.**	**C-index**	**IBS**	**C-index**	**IBS**
miRNA log_2	mean	0.780	0.120	0.782	0.122
	std	0.072	0.018	0.049	0.027
	min	0.654	0.095	0.711	0.077
	max	0.903	0.153	0.838	0.172
miRNA quantile	mean	0.705	0.146	0.725	0.144
	std	0.102	0.031	0.071	0.030
	min	0.483	0.113	0.610	0.107
	max	0.838	0.214	0.812	0.195
mRNA log_2	mean	0.754	0.145	0.782	0.142
	std	0.115	0.050	0.105	0.048
	min	0.550	0.071	0.595	0.081
	max	0.878	0.252	0.878	0.259
mRNA TPM	mean	0.776	0.142	0.820	0.131
	std	0.101	0.044	0.093	0.037
	min	0.603	0.080	0.675	0.070
	max	0.872	0.227	0.931	0.190

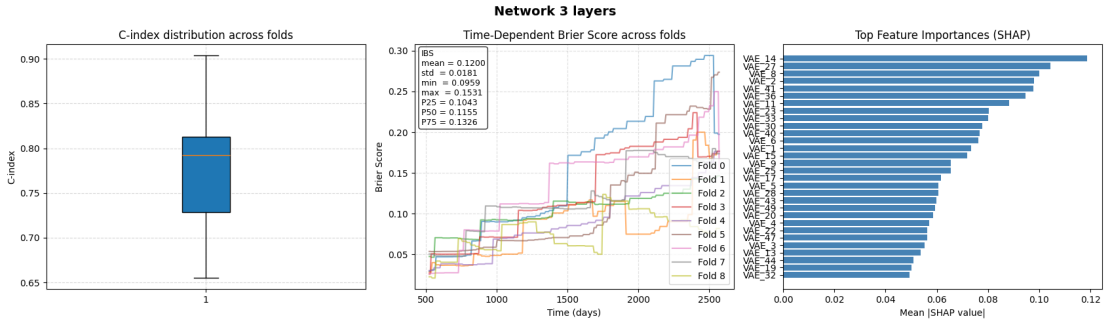

Fig. D.1: miRNA log_2 3-layers

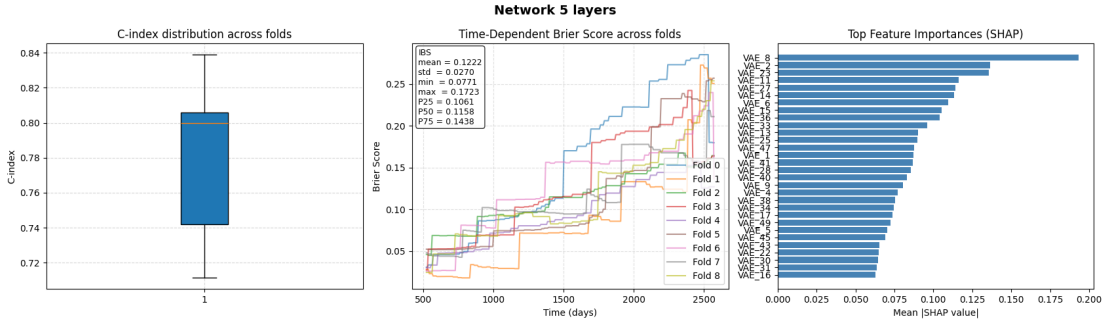

Fig. D.2: miRNA log_2 5-layers

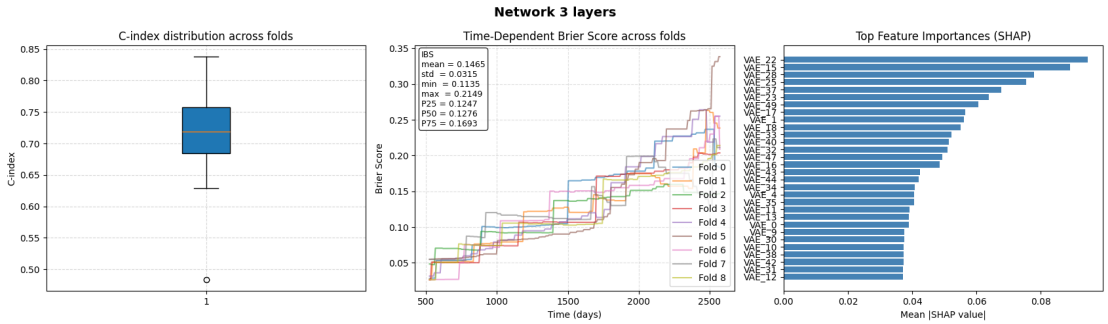

Fig. D.3: miRNA quantile 3-layers

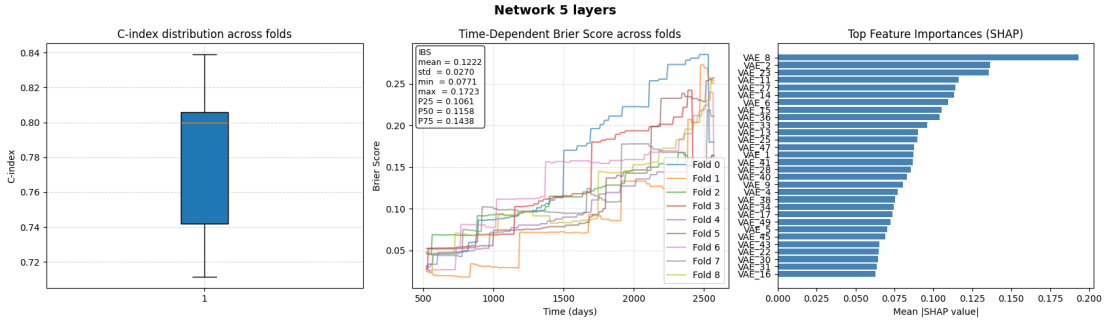

Fig. D.4: miRNA quantile 5-layers

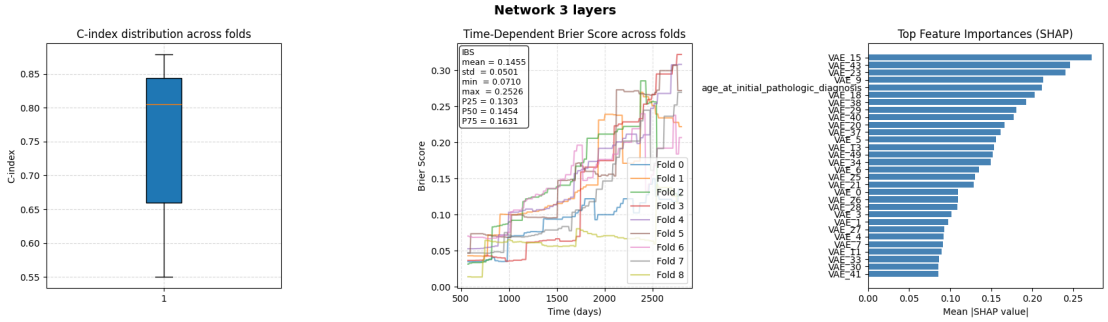

Fig. D.5: mRNA log_2 3-layers

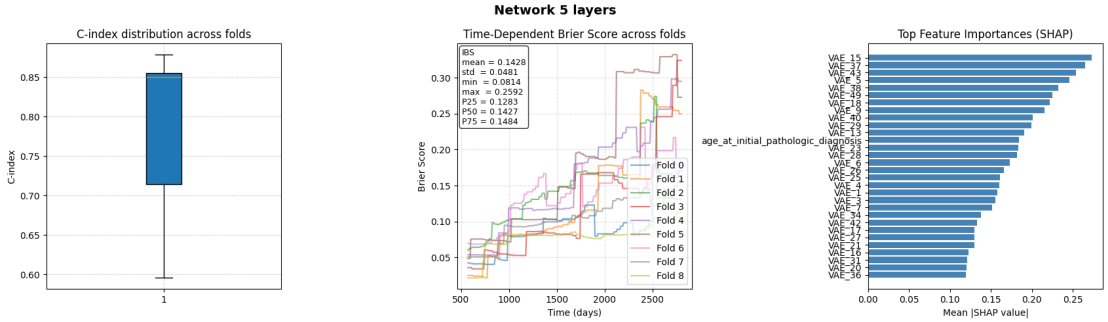

Fig. D.6: mRNA log_2 5-layers

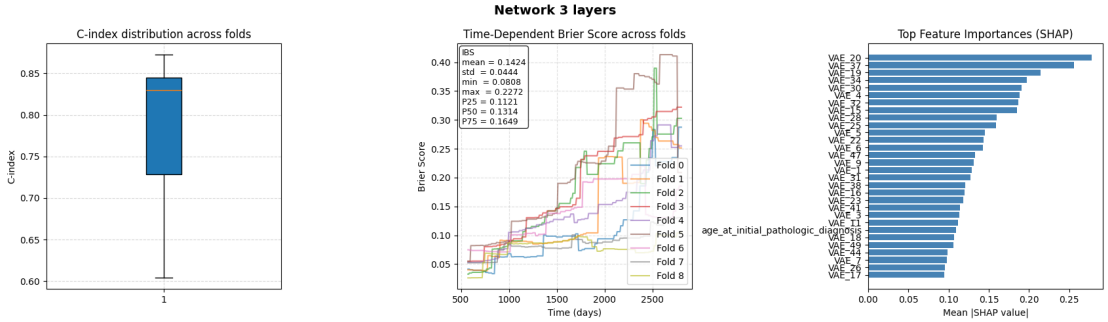

Fig. D.7: mRNA TPM 3-layers

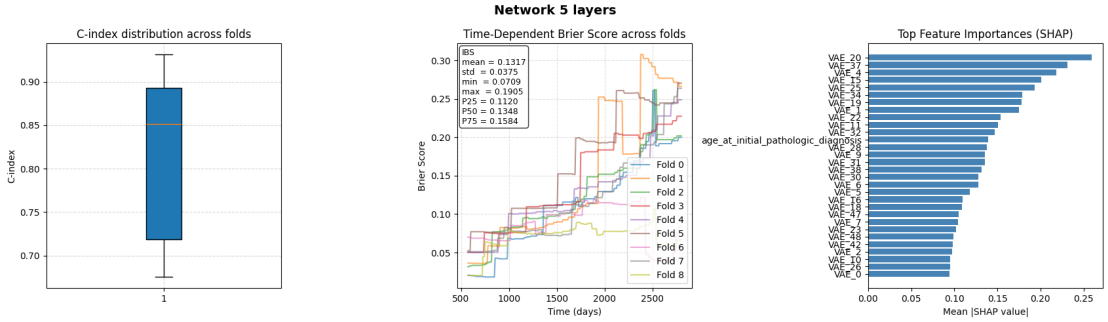

Fig. D.8: mRNA TPM 5-layers