RESEARCH PAPER

Integrated Transcriptomic and Machine Learning Framework Identifies a Blood-Based Biomarker Signature for Anthracycline-Induced Cardiotoxicity in Juvenile Cancer Survivors

Dr. Rajatha Maradi Hemanth Kumar[1]

[1] General Practitioner, Elova Hospitals, 27, 5th Cross, Lalbagh Main Rd, Sudhama Nagar, Bengaluru, Karnataka 560027. Email: meetrajatha@gmail.com

Received: 12th Mar, 2026 | **Revised:** 24th Mar, 2026 | **Accepted:** 14th Apr, 2026 | **Available Online:** 30th Apr, 2026

ABSTRACT

Anthracyclines are commonly used for the treatment of juvenile-onset cancers, but chronic exposure to anthracyclines is linked with cardiotoxicity that can extend to the survivor phase. The objective of this study was to discover blood transcriptomic markers of anthracycline-induced cardiotoxicity among survivors of juvenile cancers through an integrated analysis. Gene expression profiles of peripheral blood from 104 survivors, 40 with cardiomyopathy and 64 controls, were obtained from publicly available databases. Following data pre-processing, filtering, normalisation and platform correction, differential expression and prediction modelling were then carried out. Twelve genes were found to be significantly regulated, with most genes being down-regulated in cases. The best results were obtained with logistic regression of the top 25 differentially expressed genes, resulting in an area under the receiver operating characteristic curve (AUC) of 0.837, accuracy of 0.789, and F1-score of 0.726. Cross-validation indicated model stability and overlap between genes identified a 12-gene candidate biomarker panel supported by both statistical and predictive analysis. Enrichment analysis implicated signaling, developmental and metabolic processes. This suggests that blood transcriptomic profiling could help identify biomarkers for anthracycline-induced cardiotoxicity, and provides data for future validation.

Keywords: Anthracycline-induced cardiotoxicity; childhood cancer survivors; juvenile-onset cancer; blood-based biomarkers; transcriptomics; differential gene expression; logistic regression; cardiomyopathy; biomarker discovery; cardio-oncology.

How to cite this article: Hemanth Kumar RM. Integrated Transcriptomic and Machine Learning Framework Identifies a Blood-Based Biomarker Signature for Anthracycline-Induced Cardiotoxicity in Juvenile Cancer Survivors. Int J Drug Deliv Technol. 2026;16(40s): 219-230. DOI: 10.25258/ijddt.16.40s.24

Source of support: Nil.

Conflict of interest: None

1. Introduction

Anthracyclines are widely used to treat pediatric cancers, and have resulted in improved survival rates for many childhood cancers. But their use is hampered by known cardiotoxic side effects which can occur during treatment or many years later. Childhood cancer survivors who receive anthracyclines are more likely to experience heart-related issues, such as cardiomyopathy and heart failure, in adulthood [1]. Their use in curative therapy highlights the need to understand and minimise their long-term side effects. Anthracycline-induced cardiotoxicity has a significant impact, especially in long-term survivors. Cardiovascular complications are now a major source of morbidity and mortality, and may occur many years after treatment. The mechanisms underlying this include oxidative damage, mitochondrial dysfunction and direct injury to the myocardium, leading to cumulative heart damage [2]. Large-scale studies have shown a high incidence of major cardiac events in survivors treated with anthracyclines, showing the long-term effects of anthracycline-induced cardiotoxicity [3]. This is corroborated by epidemiological studies showing a significantly higher risk of developing cardiovascular disease in childhood cancer survivors as compared to the general population [4].

At a molecular level, the cardiotoxicity of anthracyclines is largely mediated through processes such as DNA damage and inhibition of topoisomerase IIβ, and disruption of cellular homeostasis, especially in cardiomyocytes. One such anthracycline, doxorubicin, has been well studied and has been shown to cause dose-dependent cardiomyopathy via these

mechanisms [5]. Molecular studies have also helped to understand the involvement of specific targets and pathways in the heart that lead to cardiotoxicity, offering new insights into the underlying disease mechanism [6]. Despite these advances, early detection remains challenging, as subclinical cardiac dysfunction often precedes overt clinical manifestations.

Despite this deeper understanding of disease onset, early detection continues to be difficult due to the presence of subclinical heart disease prior to the onset of clinical symptoms. The progressive and delayed nature of anthracycline-induced cardiotoxicity highlights the need for robust methods of early detection, and risk stratification. Recent research has highlighted the need for the integration of molecular and clinical surveillance approaches in early detection of at-risk patients before irreversible damage takes place. Recent findings have shown promise for the use of advanced imaging and molecular biomarkers in the early detection of functional changes, and improved monitoring in long-term survivors [7]. Moreover, serial monitoring with sensitive markers of cardiac function has been shown to be useful in detecting subtle changes in cardiac function over an extended period of time [8].

Objectives of the study:

- To identify blood-based transcriptomic biomarkers associated with anthracycline-induced cardiotoxicity in childhood cancer survivors
- To integrate differential gene expression and predictive modeling approaches for biomarker discovery
- To evaluate the potential of identified genes in improving early detection and risk stratification of cardiotoxicity

2. Literature Review

2.1 Anthracycline Cardiotoxicity Mechanisms

Molecular studies in recent years have broadened our knowledge of the mechanisms of anthracycline-induced cardiotoxicity in long-term survivors of childhood cancer. Blood-based transcriptomic analyses have shown that survivors with anthracycline-induced cardiomyopathy have distinct gene-expression profiles when compared to other survivors, consistent with a model of cardiotoxicity in association with systemic changes rather than isolated injury to the heart [9]. In a separate study, the haptoglobin gene was implicated in anthracycline-related cardiomyopathy, suggesting that processes related to oxidative stress, inflammation and immunology may play a role in the development of cardiac dysfunction in this group of patients [10]. These results suggest that the molecular and cellular mechanisms of anthracycline cardiotoxicity may be reflected in blood analyses.

2.2 Transcriptomic Studies in Cardiotoxicity

Despite a recent increase in transcriptomic studies of anthracycline cardiotoxicity, the body of literature is relatively small and diverse. In addition to gene expression analyses, a number of molecular studies have evaluated circulating biomarkers as surrogates of cardiac damage. Candidate biomarker studies of plasma in childhood cancer survivors have pinpointed potentially useful proteins as related to anthracycline-induced cardiomyopathy, but these results need to be confirmed in larger series [11]. Likewise, candidate proteomic studies have identified early diagnostic markers foranthracycline-induced cardiomyopathy, confirming the potential of omics screening while also re-emphasising the need for consistency and reproducibility in omics studies [12]. Circulating microRNAs have also been identified as potential markers of anthracycline cardiotoxicity, further highlighting the benefits of molecular profiling for early detection; however, variations in study design, sample type and processing/analysis methods continue to hinder study replication and comparisons across studies [13]. These study-to-study differences are also reflected in screening research in cancer survivors, where screening for cardiac dysfunction continues to be critical for clinical management but remains limited by sub-optimal sensitivity for detection of early or subclinical disease [14].

2.3 Biomarker Development in Survivors

The search for accurate biomarkers of anthracycline-induced cardiotoxicity has increasingly shifted to integrated approaches, which combine molecular markers with predictive models. In general studies of drug-induced cardiotoxicity, the combined transcriptomic/molecular machine-learning approaches have demonstrated that computer algorithms can enhance recognition of clinically meaningful cardiotoxic phenotypes when high dimensional molecular data are available [15]. Likewise, recent advances in cardiovascular precision medicine have shown the role of machine-learning methods in improving the identification of biomarkers and prediction of disease when traditional statistical analyses fail to capture complex biological relationships [16]. But such methods have yet to be fully translated to childhood cancer survivorship. Recent reviews on anthracycline-induced cardiomyopathy continue to highlight areas of need for risk prediction, prevention and treatment, and also note

that biomarker-based stratification techniques remain poorly developed and need to be better validated [17]. In particular, in the context of the survivors, secondary prevention remains based on clinical history and traditional surveillance methods rather than integrated blood transcriptomic biomarkers, suggesting a need for improvement [18].

2.4 Role of Computational Approaches

The use of computational approaches is becoming more important in cardio-oncology as they allow high-dimensional data to be explored and potential biomarkers identified that may not be obvious using traditional methods that focus on single variables. However, the existing literature indicates that these methods should be considered within a clinical context, rather than in isolation. The most recent scientific statement on cardiac toxicity in child cancer survivors advocates for the need for robust surveillance, validation over time and biologically meaningful risk stratification tools to predict the emergence of cardiac toxicity, highlighting the need for computational approaches to work in unison with well-established mechanistic and clinical knowledge [19]. Thus, there is a clear need for more research that combines transcriptomics with carefully chosen predictive approaches to discover clinically relevant biomarkers of anthracycline-induced cardiotoxicity in juvenile-onset cancer survivors.

3. Methodology

3.1 Dataset Description

We used publicly available gene expression data from children with cancer who had been treated with anthracyclines [20]. A total of 104 peripheral blood samples (40 samples from individuals with anthracycline-induced cardiomyopathy and 64 controls) were analysed for peripheral blood gene expression.

The samples were produced with two microarray platforms (GPL24676 and GPL16791), which correspond to different experimental settings and data collection methods. The cardiomyopathy status of the individuals was based on the annotations included in the original data. We used publicly available, deidentified data, and therefore no further ethical approval was necessary.

3.2 Data Acquisition and Preprocessing

We used publicly accessible microarrays data sets of anthracycline-treated childhood cancer survivors. Cases and controls were defined based on clinical labels as cardiomyopathy and non-cardiomyopathy, respectively. We combined data from two different microarray platforms to enhance the sample size and statistical significance.

Raw counts were filtered to exclude low-expressed genes which can be noisy and reduce the study's statistical power. After filtering, gene expression data were normalized based on counts per million (CPM) and then transformed to log2 scale to moderate the variance between samples. This allowed for gene expression values to be compared across samples while retaining important biological variation.

3.3 Batch Correction and Quality Control

The use of multiple platforms required correction for technical variability to adjust for batch effects. A linear modeling approach was used to correct for platform-specific effects. A principal component analysis (PCA) was used to evaluate the batch correction. PCA was performed before and after correction to assess the level of technical variability and to ensure correction led to better integration of samples across platforms without loss of biological variance between groups.

3.4 Differential Expression Analysis

A differential expression analysis was conducted between case and control groups to identify genes that might be implicated in anthracycline-induced cardiomyopathy Testing was performed with gene-wise comparisons and p-values were adjusted for multiple hypothesis testing using the Benjamini-Hochberg false discovery rate procedure. Genes with statistical and effect size criteria below user-specified thresholds were identified as differentially expressed. This produced a list of potential disease-associated genes ranked by significance.

3.5 Protein–Protein Interaction (PPI) Network Analysis

In order to investigate potential interactions between the genes, a network of protein-protein interactions (PPI) was generated using the STRING database (version at time of analysis). The gene set comprising the top 50 genes from differential expression analysis was input to the network. Interactions were sourced from known and predicted associations such as experimental evidence, curated databases, co-expression and text mining.

The network was obtained with the default parameters for Homo sapiens in STRING, and the interaction confidence thresholds were maintained to identify significant interactions. We also extracted network topology metrics, such as number of nodes, number of edges, average node degree, clustering coefficient and enrichment statistics from STRING. These were used to determine whether the interaction network was

significantly different from a random network, suggesting functional interaction between the input genes.

3.6 Feature Selection and Model Development

In order to simplify and improve model interpretability, a panel of the highest-ranking differentially expressed genes was chosen for modeling. Several gene subsets with different numbers of genes were tested, including a smaller subset of genes from the top differential gene list.

To classify samples, a logistic regression model was used because it was interpretable and had an acceptable sample size. Gene expression values were used as predictor variables, and cardiomyopathy status was the outcome variable. The coefficients of the model were used to assess the influence of each gene.

3.7 Model Evaluation

We used stratified k-fold cross-validation to evaluate model performance (ensuring equal representation of cases and controls in training and validation data). We used common metrics of classification performance, such as the area under the receiver operating characteristic curve (AUC), accuracy, precision, recall and F1-score. We also evaluated model performance using receiver operating characteristic (ROC) curves and confusion matrices. To assess stability, cross-validation was repeated, making it possible to estimate the variability in model performance.

3.8 Stability and Functional Analysis

To be considered stable predictors, model coefficients of the selected biomarkers were compared across cross-validation folds. Those genes with stable effect sizes were regarded as reliable predictors and selected for interpretation.

An enrichment analysis was conducted to contextualise the identified gene set. List of candidate genes were evaluated against curated pathway and ontology databases such as Gene Ontology, Reactome and KEGG Enrichment results were related to known mechanisms of cardiotoxicity and cardiovascular biology.

4. Results

4.1 Cohort Characteristics and Data Processing

We analysed 104 peripheral blood transcriptomic samples including 40 cases of anthracycline-induced cardiomyopathy and 64 controls. These samples were obtained from two platforms, GPL16791 and GPL24676. Initially, 58,101 genes were available following preprocessing. After low-expression filtering, 18,994 genes were used for analysis.

Table 1. Cohort characteristics and preprocessing summary

Characteristic	Value
Total samples	104
Cases	40
Controls	64
GPL16791 samples	16
GPL24676 samples	88
Initial genes	58,101
Genes retained after filtering	18,994
Genes removed after filtering	39,107
Fraction of zero entries in raw matrix	0.5511
Genes with all-zero counts	11,751

Table 1 shows the cohort characteristics and the main results of the preprocessing. This analysis confirmed that the data were ready for an integrated transcriptomic analysis after filtering and normalization.

4.2 Batch Correction Improved Cross-Platform Integration

We first used principal component analysis (PCA) to investigate the technical variability between platforms. The first two principal components (PCs) accounted for 44.9% and 8.5% of the variance, respectively, and showed some separation by platform. Following platform correction, the first two principal components explained 46.6% and 8.3% of the variance, and the extent of platform separation was diminished. PCA plots of the transcriptomic structure before correction and after platform correction, with the samples colored by platform.

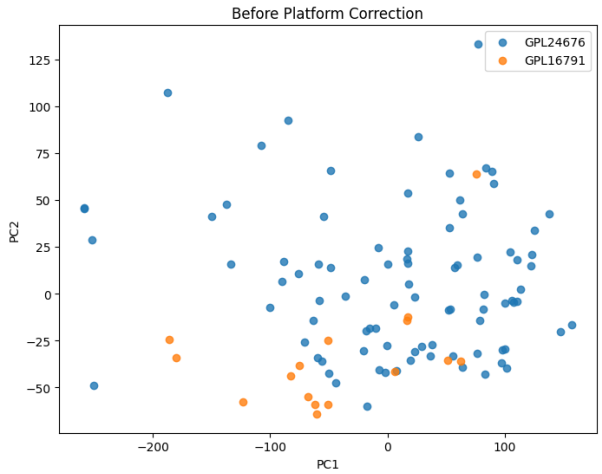

Figure 1. Principal component analysis before and after platform correction

Figure 1 shows that the integrated preprocessing approach successfully minimized the variability due to the platform used while retaining the global structure of the data. This justified the use of the normalized expression matrix for biological and predictive analyses.

4.3 Case–Control Separation Was Subtle at the Global Transcriptomic Level

To assess whether there were global transcriptomic differences associated with disease status, a principal component analysis (PCA) was produced on the corrected expression matrix with samples colored as cases and controls. The first two principal components were not completely separated by disease status, suggesting that variation associated with disease was not due to a single global expression pattern. PCA of the platform-corrected expression matrix with samples coloured as cardiomyopathy cases and controls.

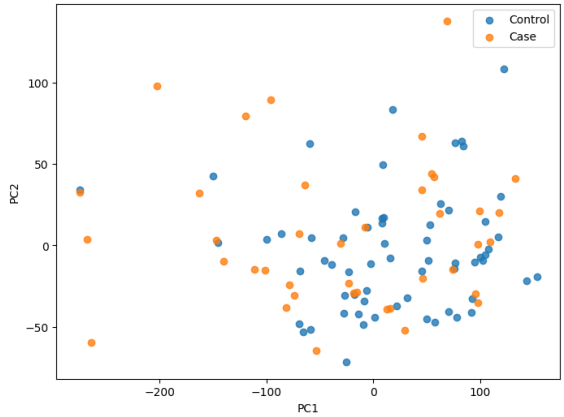

Figure 2. Principal component analysis of corrected expression profiles colored by cardiomyopathy status

Figure 2 suggests that the transcriptomic changes in peripheral blood in anthracycline-induced cardiomyopathy are subtle and therefore makes a case for an application of targeted feature selection and supervised classification, rather than unsupervised separation.

4.4 Differential Expression Analysis Identified a Compact Set of Significant Genes

Analysis of differential expression on the corrected expression matrix revealed a small but significant transcriptomic signature of cardiomyopathy. Using an adjusted p-value cut-off of <0.05 and an absolute log2 fold change cut-off of >0.2, 12 genes were found to be significantly dysregulated. 11 genes were down regulated and 1 gene up regulated in cases compared to controls. Volcano plot of the differential expression of the filtered transcriptome. Genes significantly up- and down regulated are indicated by adjusted p-value and log2 fold change cut-offs.

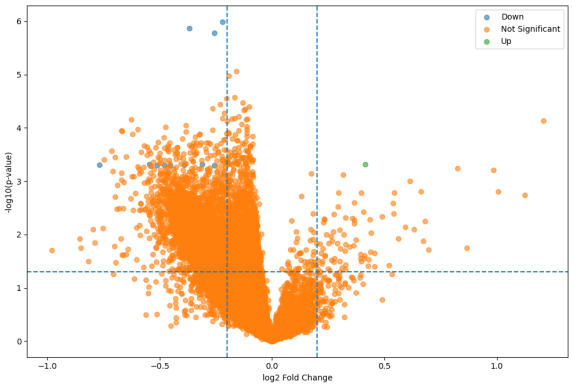

Figure 3. Volcano plot of differential expression between cardiomyopathy cases and controls

Figure 3 illustrates that only a limited number of genes passed the significance criteria, and the majority of the significant signals were down-regulated in cardiomyopathy cases. This finding supports the need for a biomarker discovery rather than a whole transcriptome classification strategy.

Table 2. Significant differentially expressed genes

Gene symbol	Gene name	log2FC	Adjusted p-value
STAG3L1	STAG3 cohesin complex component like 1	-0.255935	0.010625
EPPK1	epiplakin 1	-0.369267	0.012964
DENND2C	DENN domain containing 2C	-0.220193	0.019499

SLC1A7	solute carrier family 1 member 7	-0.390939	0.049628
SPNS1	SPNS lysolipid transporter 1	-0.310739	0.049651
TRABD2A	TraB domain containing 2A	-0.768957	0.049652
PTCH1	patched 1	-0.547502	0.049816
ITLN1	intelectin 1	0.414509	0.049845
AAK1	AP2 associated kinase 1	-0.479575	0.049947
ZNF609	zinc finger protein 609	-0.453938	0.049948
FUT2	fucosyltransferase 2	-0.257488	0.049952
LINC00402	long intergenic non-protein coding RNA 402	-0.512924	0.049986

Table 2 presents the 12 significant genes identified in the differential expression analysis. These genes formed the initial biologically informed candidate pool for downstream predictive modeling.

4.5 Protein–Protein Interaction Network Analysis

To gain insight into the interaction profile of the differentially expressed genes, we used the STRING database to build a protein-protein interaction (PPI) network. The PPI network of the top 50 differentially expressed genes is presented in Figure 4, revealing the interaction patterns among the candidate genes. The network contained 48 nodes and 150 edges, suggesting a high level of interaction among the candidate genes. The number of edges in the network was significantly greater than the expected number of edges (50), indicating non-random connectivity. The average node degree (6.25) suggested a moderate network connectivity and the average local clustering coefficient (0.414) suggested the occurrence of functional gene clusters.

Importantly, the PPI enrichment p-value was highly significant ($p < 1.0 \times 10^{-16}$), demonstrating that the observed interactions are unlikely to occur by chance. This finding supports the presence of coordinated biological processes among the identified genes. These results suggest that the differentially expressed genes are functionally interconnected and may participate in shared molecular pathways relevant to anthracycline-induced cardiotoxicity.

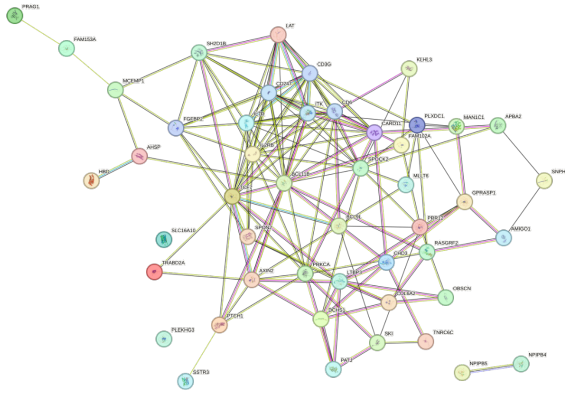

Figure 4. Protein–protein interaction network of top 50 differentially expressed genes constructed using STRING database.

The network illustrates functional associations among the selected genes, with nodes representing proteins and edges indicating known or predicted interactions. The network comprises 48 nodes and 150 edges, with an average node degree of 6.25 and a clustering coefficient of 0.414. The observed number of interactions significantly exceeds the expected number (50), with a PPI enrichment p-value $< 1.0 \times 10^{-16}$, indicating strong functional connectivity among the genes.

4.6 DE-Guided Feature Selection Improved Predictive Performance

Various approaches to classification were explored to assess the predictive value of the transcriptomic signature. Using a wide feature space (500 variable genes) resulted in moderate performance. Logistic regression with 500 variable genes performed with an AUC of 0.679, while LASSO and Random Forest classifiers performed with an AUC of 0.604 and 0.660, respectively. However, using only the top 25 variable genes resulted in the best model, with logistic regression performing with an AUC of 0.837, accuracy of 0.789 and F1-score of 0.726.

Table 3. Comparative performance of classification models

Model	Feature set	AUC	Accuracy	F1-score	Precision	Recall
Logistic	Top 500	0.679	0.635	0.508	0.528	0.500

Regression	variable genes					
LASSO	Top 500 variable genes	0.604	0.616	0.412	0.459	0.400
Random Forest	Top 500 variable genes	0.660	0.663	0.363	0.528	0.300
Logistic Regression	Top 25 DE genes	0.837	0.789	0.726	0.761	0.725

The results in Table 3 show that feature reduction by domain knowledge substantially improved classification results. Logistic regression performed best on this dataset, compared to LASSO and Random Forest, and with the top 25 differentially expressed genes. Bar plot showing the AUC of the classification models used and feature sets.

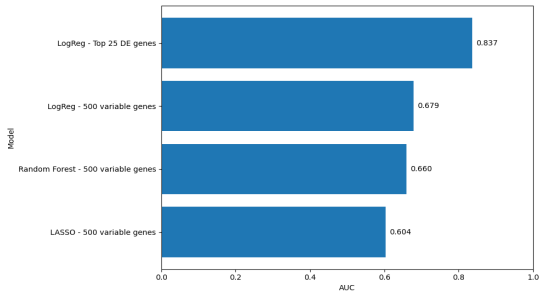

Figure 4. Comparative model performance based on area under the receiver operating characteristic curve

Figure 5 visually summarizes the superiority of the top 25 DE-gene logistic regression model relative to broader and less targeted modeling approaches.

4.7 The Final Top 25-Gene Logistic Regression Model Showed Good Discrimination

The best-performing classifier was the logistic regression model trained on the top 25 differentially expressed genes. Cross-validated ROC analysis demonstrated a mean AUC of 0.835 with a standard deviation of 0.041 across folds, indicating good and relatively consistent discrimination. Cross-validated ROC curves for the final logistic regression model based on the top 25 DE genes. The mean ROC curve and fold-specific AUC values are shown.

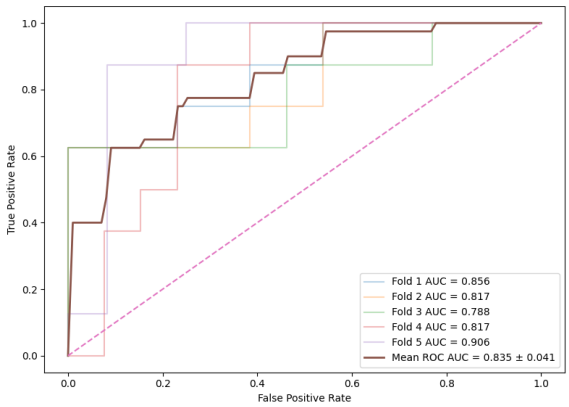

Figure 6. ROC curve for the logistic regression model using the top 25 differentially expressed genes

Figure 6 confirms that the final model provided a strong discriminative signal, substantially outperforming the initial models built on the larger 500-gene feature set.

To further characterize classification performance, a confusion matrix was generated using cross-validated predictions from the final model. The model correctly classified 53 of 64 controls and 29 of 40 cases, corresponding to balanced class-wise performance. Confusion matrix based on cross-validated predictions from the final logistic regression model.

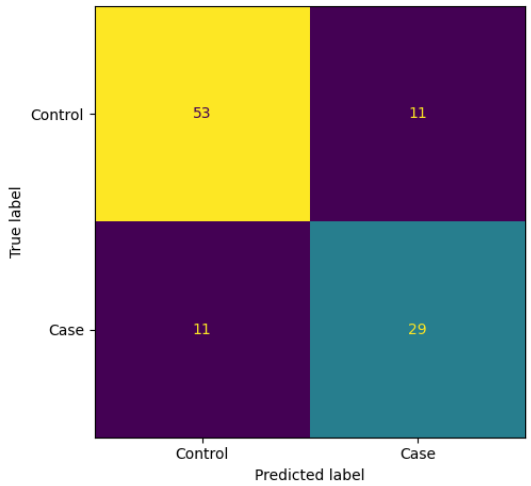

Figure 7. Confusion matrix for the logistic regression model using the top 25 differentially expressed genes

The corresponding classification metrics showed a precision of 0.72 and recall of 0.72 for cardiomyopathy cases, indicating that the classifier retained balanced sensitivity and specificity rather than favoring the majority class.

4.8 Stable Predictive Features Were Identified Across Cross-Validation Folds

To assess feature robustness, model coefficients were examined across cross-validation folds. Several genes demonstrated consistently high mean absolute coefficients, indicating stable contribution to model discrimination. The most stable genes included AAK1, DENND2C, EPPK1, GJA3, STAG3L1, and SLC1A7. Bar plot showing the top genes ranked by mean absolute logistic regression coefficient across cross-validation folds.

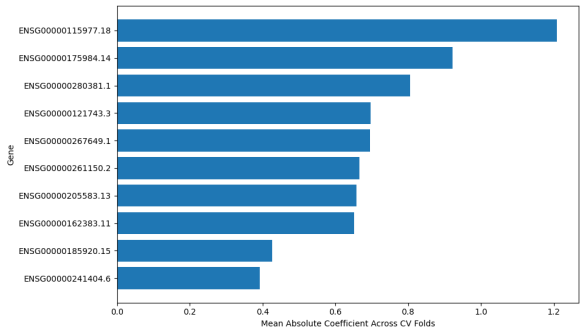

Figure 8. Stable gene contributors across cross-validation folds

Figure 8 shows the stability analysis supports the reproducibility of the final gene panel and suggests that the predictive performance was not driven by unstable or fold-specific features.

4.10 Integration of Differential Expression and Predictive Modeling Identified a Core Candidate Biomarker Panel

To prioritize the most biologically and analytically relevant genes, overlap analysis was performed between significant differentially expressed genes and the genes contributing to the final predictive model. This yielded 12 overlapping genes, representing the core candidate biomarker set.

Table 4. Overlapping genes identified by both differential expression and predictive modeling

Gene symbol	Gene name	Logistic regression coefficient	log2FC	Adjusted p-value
STAG3L1	STAG3 cohesin complex component like 1	-0.691363	-0.255935	0.010625
EPPK1	epiplakin 1	-0.731764	-0.369267	0.012964
DENND2C	DENN domain containing 2C	-0.956778	-0.220193	0.019499
SLC1A7	solute carrier family 1 member 7	-0.661481	-0.390939	0.049628
SPNS1	SPNS lysolipid transporter 1	0.005559	-0.310739	0.049651
TRABD2A	TraB domain containing 2A	0.166857	-0.768957	0.049652
PTCH1	patched 1	0.444462	-0.547502	0.049816
ITLN1	intelectin 1	0.330339	0.414509	0.049845
AAK1	AP2 associated kinase 1	1.378090	-0.479575	0.049947
ZNF609	zinc finger protein 609	0.172808	-0.453938	0.049948
FUT2	fucosyltransferase 2	0.237113	-0.257488	0.049952
LINC00402	long intergenic non-protein coding RNA 402	-0.164955	-0.512924	0.049986

Table 4 represents the most important transcriptomic findings of the study, as these genes were supported by both statistical differential expression and predictive modeling. This integrated panel forms the main candidate blood-based biomarker signature identified in the present analysis.

4.11 Functional Enrichment Analysis Provided Supportive Biological Context

We used the overlapping gene list for functional enrichment analysis to gain biological insight. Gene Ontology analysis identified enrichment terms associated with epithelial cell proliferation,

development and fucose metabolism. Reactome analysis showed enrichment in Hedgehog signaling, ligand-receptor interactions, immune system signaling, and KEGG analysis revealed enrichment in glycosphingolipid biosynthesis, Hedgehog signaling, synaptic vesicle cycle, and glutamatergic synapses. Figure 9 Bar plot of the top enriched KEGG pathways from the overlapping genes of differentially expressed genes and model-selected genes.

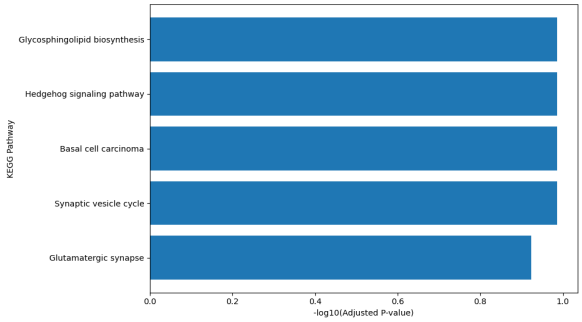

Figure 9. Top enriched KEGG pathways for the overlapping candidate gene set

Although enrichment significance was modest and several pathways were driven by single genes, the results provide biologically relevant context for the identified transcriptomic signature and support its potential relevance to anthracycline-related cardiotoxicity.

5. Discussion

A small transcriptomic signature in the bloodstream has been found to be related to anthracycline cardiotoxicity in juvenile cancer survivors in the present study. This study applied an integrated approach to show that although global transcriptomic differences between the cases and controls were relatively small, differential expression and biologically relevant feature selection significantly enhanced classification. In particular, the final logistic regression model with the top 25 differentially expressed genes performed better than models with larger feature sets, highlighting the importance of feature selection in this context. These results are in line with the general observation that anthracycline exposure is a risk factor for long-term cardiovascular disease in childhood cancer survivors and that molecular approaches may help to further refine clinical scrutiny beyond traditional clinical evaluation [1], [3].

The current findings are generally consistent with previous studies that have demonstrated that anthracycline-induced cardiotoxicity is related to a complex of molecular events, rather than any single pathway. Laboratory studies have demonstrated a role for anthracycline-induced cardiac injury via reactive oxygen species, DNA damage, mitochondrial dysfunction, and topoisomerase IIβ-mediated injury, which may lead to subsequent cardiac damage and dysfunction [6]. More recent transcriptomic studies of childhood cancer survivors have also shown peripheral blood gene-expression changes in childhood cancer survivors with anthracycline-induced cardiomyopathy, suggesting the importance of systemic molecular changes in this disease. These studies provide a rationale for the current gene panel and add to the notion that peripheral blood may reflect key molecular changes associated with a disease process, such as cardiac injury. The biomarker set identified in this study is also consistent with the growing body of literature on biomarker discovery in survivors, although it is difficult to make direct comparisons due to the different platforms and biospecimens used in previous studies. Plasma studies have helped to identify candidate circulating proteins that may be associated with anthracycline-related cardiomyopathy, while microRNA studies have also indicated that minimally invasive molecular markers may be of translational value in cardio-oncology [11], [13]. In that respect, the current analysis contributes to this area of research by specifically investigating blood-based transcriptomic biomarkers, and by further suggesting that a set of genes that show both statistical differential expression and predictive power may be a better candidate set than either type of gene alone. Such an approach may help to decrease false-positive biomarker discovery and enhance interpretability in small sample size transcriptomic studies. Biologically, a number of genes selected in the overlap of differential expression and predictive modeling are interesting. The enrichment and candidate gene selection of PTCH1 suggest that Hedgehog (and other) developmental regulatory pathways may be involved in cardiac repair or remodeling responses. FUT2 and enrichment for glycosylation processes highlight that metabolic and cell surface modification processes may also play a role in the cardiotoxic response. Similarly, genes such as AAK1, DENND2C, EPPK1 and SLC1A7 may represent more general changes to cellular trafficking and structure regulation and signal transduction pathways. While these results are preliminary, they support the concept of a multiple-biological basis of anthracycline-induced cardiotoxicity and widespread, coordinated dysfunction of multiple pathways. This view is in line with recent reviews that highlight the complexity of the anthracycline-induced

cardiomyopathy and the importance of holistic approaches to risk prediction and understanding of mechanisms [17].

The results of the modeling have implications for methods. The poorer performance seen with the larger feature set (500 genes) versus the much better performance seen with the top 25 ranked genes (which were differentially expressed) suggests that biologically driven feature selection is important in small-to-moderate sized transcriptomic data sets. The results are in line with the computational literature that indicates the predictive power of omics data is not only dependent on the model chosen, but also feature prioritisation and interpretability. In the current analysis, logistic regression was more appropriate than other more sophisticated models for the final list of biomarkers, most likely due to its ability to interpretability and stability. The additional stability analysis of coefficients also provided additional confidence that the final model was not fit solely to unstable fold-specific "artifacts". From a clinical perspective, the current results indicate that transcriptomic profiling of blood may be potentially useful, in combination with other strategies, to assess childhood cancer survivors at risk of developing anthracycline-induced cardiotoxicity. This may be important as conventional screening techniques, while crucial, may not detect early molecular events that undergo a transition to dysfunction. An interpretable and short transcriptomic profile may eventually be added to echocardiographic surveillance or other biomarker techniques to aid in early risk prediction. However, the results need to be taken with a grain of salt. Recent scientific recommendations to improve childhood cancer survivorship stress that models for the assessment of cardiovascular risk need to be easily interpretable, validated over time and integrated with current surveillance practices before they can be applied in clinical practice [19]. As such, the current model should be considered as a potential biomarker model rather than a diagnostic instrument. There are a number of caveats. First, a single public data set was used, and we lacked an independent data set to assess the performance of the transcriptomic signature. Second, while repeated cross-validation suggested reproducibility, variability in the model performance from repeated splits suggests that the small sample size is still a limiting factor. Third, the pathway analyses were biologically plausibly but not strong, with multiple terms having few genes, and thus should be viewed as complementary. Fourth, the study used whole blood transcriptomics and it is possible that these may not entirely reflect the molecular events in the heart. Despite these caveats, the study has several strengths including the combination of differential expression analysis, interpretable prediction modeling, prioritization of biomarkers based on overlap, and stability of the model. In conclusion, the findings support the presence of a consistent transcriptomic signature in blood associated with anthracycline-induced cardiotoxicity, and set the stage for future studies in larger cancer survivor cohorts.

6. Conclusion

Overall, this study discovered a small transcriptomic signature in peripheral blood linked to anthracycline cardiotoxicity in childhood cancer survivors. Using a combination of differential expression and interpretable predictive modeling, the study showed that biologically relevant gene selection resulted in superior case-control discrimination than larger sets of genes. The final panel of candidates were genes significant both statistically and as model features, potentially useful as predictors of cardiotoxicity. Pathway analyses also suggested that these genes may be associated with biologically significant pathways such as signaling, development and metabolism. Despite the predictive model's strong performance, the results should be considered exploratory as they were based on a single public dataset without validation. However, the findings suggest that blood gene expression data include information relevant to anthracycline-induced heart damage. The study adds to the growing body of work aimed at better predicting, diagnosing and stratifying risk in childhood cancer survivors and the need for validation studies in independent samples and clinical trials.

References

[1] V. I. Franco, J. M. Henkel, T. L. Miller, and S. E. Lipshultz, “Cardiovascular Effects in Childhood Cancer Survivors Treated with Anthracyclines,” *Cardiology Research and Practice*, vol. 2011, pp. 1–13, 2011, doi: 10.4061/2011/134679.

[2] S. E. Lipshultz *et al.*, “Long-term Cardiovascular Toxicity in Children, Adolescents, and Young Adults Who Receive Cancer Therapy: Pathophysiology, Course, Monitoring, Management, Prevention, and Research Directions: A Scientific Statement From the American Heart Association,” *Circulation*, vol. 128, no. 17, pp. 1927–1995, Oct. 2013, doi: 10.1161/CIR.0b013e3182a88099.

[3] D. A. Mulrooney *et al.*, “Major cardiac events for adult survivors of childhood cancer diagnosed between

1970 and 1999: report from the Childhood Cancer Survivor Study cohort," *BMJ*, vol. 368, p. l6794, Jan. 2020, doi: 10.1136/bmj.l6794.

[4] S. H. Armenian *et al.*, "Cardiovascular Disease in Survivors of Childhood Cancer: Insights Into Epidemiology, Pathophysiology, and Prevention," *J Clin Oncol*, vol. 36, no. 21, pp. 2135–2144, Jul. 2018, doi: 10.1200/JCO.2017.76.3920.

[5] K. Chatterjee, J. Zhang, N. Honbo, and J. S. Karliner, "Doxorubicin Cardiomyopathy," *Cardiology*, vol. 115, no. 2, pp. 155–162, Dec. 2009, doi: 10.1159/000265166.

[6] Dr. Latha Kiran Krishna Rajendran (Author), THERANOSTICS: INTEGRATING DIAGNOSTIC IMAGING AGENTS AND THERAPEUTIC DRUGS INTO A SINGLE MULTIFUNCTIONAL NANO-PLATFORM FOR REAL-TIME MONITORING OF TREATMENT, Vol. 53 No. 2 (2025): April-June 2025, Power System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/article/view/305 , DOI: https://doi.org/10.46121/pspc.53.2.31

[7] Dr. Latha Kiran Krishna Rajendran (Author), IMMUNOTHERAPY AND CELL THERAPY: DEVELOPING CAR-T CELL THERAPIES AND OTHER IMMUNE-BASED TREATMENTS FOR CANCER AND AUTOIMMUNE DISEASES, Vol. 51 No. 2 (2023): April-June 2023, Power System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/article/view/304 , DOI: https://doi.org/10.46121/pspc.51.2.7

[8] Dr. Latha Kiran Krishna Rajendran (Author), STRICT LIABILITY OR FAULT-BASED REGIMES FOR AI-CAUSED HARM? A DOCTRINAL ANALYSIS ACROSS COMMON LAW AND CIVIL LAW SYSTEMS, Vol. 52 No. 4 (2024): October-December 2024, Power System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/article/view/312, DOI: https://doi.org/10.46121/pspc.52.4.13

[9] Dr. Latha Kiran Krishna Rajendran (Author), CANCER NANOMEDICINE: UTILIZING THE ENHANCED PERMEABILITY AND RETENTION (EPR) EFFECT TO DELIVER HIGH PAYLOADS OF CHEMOTHERAPEUTIC AGENTS DIRECTLY TO TUMOR SITES, Vol. 52 No. 2 (2024): April-June 2024, Power System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/article/view/311, DOI: https://doi.org/10.46121/pspc.52.2.12

[10] Dr. Latha Kiran Krishna Rajendran (Author), MECHANISMS DRIVING IMMUNOTHERAPY RESISTANCE IN COLORECTAL CANCER LIVER METASTASES, Vol. 52 No. 1 (2024): January-March 2024 , Power System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/article/view/303, DOI: https://doi.org/10.46121/pspc.52.1.5

[11] S. Zhang *et al.*, "Identification of the molecular basis of doxorubicin-induced cardiotoxicity," *Nat Med*, vol. 18, no. 11, pp. 1639–1642, Nov. 2012, doi: 10.1038/nm.2919.

[12] Y. Qiu, P. Jiang, and Y. Huang, "Anthracycline-induced cardiotoxicity: mechanisms, monitoring, and prevention," *Front. Cardiovasc. Med.*, vol. 10, Dec. 2023, doi: 10.3389/fcvm.2023.1242596.

[13] Y. Hosono *et al.*, "Assessment of anthracycline-induced cardiotoxicity in childhood cancer survivors during long-term follow-up using strain analysis and intraventricular pressure gradient measurements," *Heart Vessels*, vol. 39, no. 2, pp. 105–116, Feb. 2024, doi: 10.1007/s00380-023-02312-2.

[14] P. Singh *et al.*, "Altered Peripheral Blood Gene Expression in Childhood Cancer Survivors With Anthracycline-Induced Cardiomyopathy – A COG-ALTE03N1 Report," *Journal of the American Heart Association*, vol. 12, no. 19, p. e029954, Oct. 2023, doi: 10.1161/JAHA.123.029954.

[15] P. Singh *et al.*, "Haptoglobin Gene Expression and Anthracycline-Related Cardiomyopathy in Childhood Cancer Survivors," *JACC: CardioOncology*, vol. 5, no. 3, pp. 392–401, Jun. 2023, doi: 10.1016/j.jaccao.2022.09.009.

[16] Rajendran, L. K. K. (2026). Integrative pharmacogenomic analysis of drug response heterogeneity across cancer cell lines: Insights from large-scale GDSC data. Scientific Culture, 12(4), 7537–7546. https://doi.org/10.5281/zenodo.12426762

[17] J. M. Leerink *et al.*, "Candidate Plasma Biomarkers to Detect Anthracycline-Related Cardiomyopathy in Childhood Cancer Survivors: A Case Control Study in the Dutch Childhood Cancer Survivor Study," *Journal of the American Heart Association*, vol. 11, no. 14, p. e025935, Jul. 2022, doi: 10.1161/JAHA.121.025935.

[18] J. M. Leerink *et al.*, "Targeted Proteomics to Identify Early Diagnostic Biomarkers for Anthracycline-Induced Cardiomyopathy in Long-Term Survivors of Childhood Cancer," *Journal of Cardiac Failure*, vol. 25, no. 8, p. S21, Aug. 2019, doi: 10.1016/j.cardfail.2019.07.061.

[19] Rajendran, L. K. K. (2026). Evaluating the association of cancer-related risk factors with multisystem health: Insights into fertility, cardiovascular, and renal indicators. Scientific Culture, 12(4), 7520–7527. https://doi.org/10.5281/zenodo.12426760

[20] H. M. Boen *et al.*, "Circulating MicroRNA as Biomarkers of Anthracycline-Induced Cardiotoxicity," *JACC: CardioOncology*, vol. 6, no. 2, pp. 183–199, Apr. 2024, doi: 10.1016/j.jaccao.2023.12.009.

[21] Rajendran, L. K. K. (2026). From prediction to precision: An externally validated deep learning–based survival and adjuvant therapy recommendation system for resected stage III non-small cell lung cancer. International Journal of Drug Delivery Technology, 16(30), 430–438. https://doi.org/10.25258/ijddt.16.30.41

[22] S. H. Armenian *et al.*, "Screening for Cardiac Dysfunction in Anthracycline-Exposed Childhood Cancer Survivors," *Clin Cancer Res*, vol. 20, no. 24, pp. 6314–6323, Dec. 2014, doi: 10.1158/1078-0432.CCR-13-3490.

[23] P. Mamoshina, A. Bueno-Orovio, and B. Rodriguez, "Dual Transcriptomic and Molecular Machine Learning Predicts all Major Clinical Forms of Drug Cardiotoxicity," *Front. Pharmacol.*, vol. 11, May 2020, doi: 10.3389/fphar.2020.00639.

[24] Rajendran, L. K. K. (2026). From prediction to practice: A machine learning–based clinical decision support tool for bevacizumab risk stratification in oncology. International Journal of Drug Delivery Technology, 16(30s), 414–429. https://doi.org/10.25258/ijddt.16.30s.40

[25] W. DeGroat, H. Abdelhalim, K. Patel, D. Mendhe, S. Zeeshan, and Z. Ahmed, "Discovering biomarkers associated and predicting cardiovascular disease with high accuracy using a novel nexus of machine learning techniques for precision medicine," *Sci Rep*, vol. 14, no. 1, p. 1, Jan. 2024, doi: 10.1038/s41598-023-50600-8.

[26] I. Fabiani, M. Chianca, C. M. Cipolla, and D. M. Cardinale, "Anthracycline-induced cardiomyopathy: risk prediction, prevention and treatment," *Nat Rev Cardiol*, vol. 22, no. 8, pp. 551–563, Aug. 2025, doi: 10.1038/s41569-025-01126-1.

[27] J. M. Leerink and E. A. M. Feijen, "Secondary prevention of anthracycline cardiotoxicity in childhood cancer survivors," *The Lancet Oncology*, vol. 25, no. 2, pp. 154–156, Feb. 2024, doi: 10.1016/S1470-2045(24)00001-9.

[28] Rajendran, L. K. K. (2026). Impact of treatment modalities on fertility, sexual function, and psychological outcomes in testicular cancer survivors: A comprehensive review. International Journal of Drug Delivery Technology, 16(30s), 447–453. https://doi.org/10.25258/ijddt.16.30s.43

[29] T. D. Ryan *et al.*, "Cardiovascular Toxicity in Patients Treated for Childhood Cancer: A Scientific Statement From the American Heart Association," *Circulation*, vol. 151, no. 15, pp. e926–e943, Apr. 2025, doi: 10.1161/CIR.0000000000001308.

[30] S. Bhatia, "Haptoglobin Gene Expression and Anthracycline-related Cardiomyopathy in Childhood Cancer Survivors," *NCBI Gene Expression Omnibus*, GEO Series accession GSE218276, Jan. 31, 2023. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE218276