

Translating Data Into Clinical Tools: An Integrative Strategy for Precision Biomarker Identification in Soft Tissue Sarcoma Diagnosis and Prognosis

Cancer Informatics
Volume 25: 1–17
© The Author(s) 2026
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/11769351261457913
journals.sagepub.com/home/cix


Masoume Avateffazeli¹, Rahem Rahmati^{1,2}, Abdolreza Mohammadnia³,
Maryam Hajimoradi¹, Elham Nazari⁴, and Shadi Shafaghi¹

Abstract

Objectives: Soft tissue sarcomas (STSs) are rare, heterogeneous cancers with over 70 subtypes, often diagnosed late due to diagnostic complexity, leading to poor outcomes. We aimed to identify and validate novel transcriptomic biomarkers for the diagnosis and prognosis of STS using an integrative machine learning and bioinformatics framework applied to publicly available cohorts. **Methods:** RNA-seq and clinical data from 261 STS samples were obtained from The Cancer Genome Atlas (TCGA). A multi-step analytical pipeline was implemented, including differential expression analysis, functional enrichment, protein-protein interaction network construction, and clinical correlation assessment. Machine learning algorithms were employed for feature selection and model development. Diagnostic performance was evaluated using receiver operating characteristic curve analysis, and prognostic value was assessed using Kaplan-Meier survival analysis. **Results:** We identified a 26-gene prognostic signature significantly associated with overall survival (15 upregulated and 11 downregulated genes). For diagnosis, *A1CF* alone showed an AUC of 0.70, while the combinations *A1CF-ATP6V0D2* and *A1CF-LECT2* achieved AUCs of 0.743 and 0.796, respectively. External validation confirmed dysregulation of *A1CF* in tumor tissues compared with controls. **Conclusions:** This study identifies a novel 26-gene prognostic signature and *A1CF*-based diagnostic panels for STS using computational methods. These biomarkers represent exploratory candidates for future investigation in STS diagnosis and prognosis; however, experimental and prospective clinical validation are required before their potential use in early detection, risk stratification, or personalized management.

Keywords

artificial intelligence, bioinformatics, biomarker, deep learning, diagnosis, machine learning, personalized medicine, prognosis, sarcoma

Received: 11 December 2025; Revised: 12 May 2026; Accepted: 22 May 2026

¹Lung Transplantation Research Center, National Research Institute of Tuberculosis and Lung Diseases (NRITLD), Shahid Beheshti University of Medical Sciences, Tehran, Iran

²Students Research Committee, Shahrekord University of Medical Sciences, Shahrekord, Iran

³Chronic Respiratory Diseases Research Center, National Research Institute of Tuberculosis and Lung Diseases (NRITLD), Shahid Beheshti University of Medical Sciences, Tehran, Iran

⁴Department of Health Information Technology and Management, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran

Corresponding authors:

Elham Nazari, Department of Health Information Technology and Management, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran 1956944413, Iran.

Email: Elham.Nazari@sbmu.ac.ir

Shadi Shafaghi, Lung Transplantation Research Center, National Research Institute of Tuberculosis and Lung Diseases (NRITLD), Shahid Beheshti University of Medical Sciences, Masih Daneshvari Hospital, Darabad Avenue, Shahid Bahonar Roundabout, Tehran 1956944413, Iran.

Email: shafaghishadi@yahoo.com



Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

1. Introduction

Soft tissue sarcomas (STSs) are rare and heterogeneous group of malignancies originating from mesenchymal tissue, such as muscle, fat, bone, and vascular structures, accounting for approximately 1% of adult and 15% of pediatric cancers.¹⁻³ The World Health Organization recognizes over 70 distinct histological subtypes of STS, reflecting their considerable pathological diversity.³ This pronounced heterogeneity presents significant challenges in diagnosis and clinical management, often leading to delayed detection and advanced disease at presentation.^{1,4} Therefore, the prognosis for STS patients remains suboptimal; approximately 50% of patients develop metastatic disease, and while the overall five-year survival rate is nearly 65%, it reduces to 10-16% for patients with high-grade tumors or distant metastases.^{1,5,6}

The pathogenesis of cancer, including sarcoma, involves complex molecular alterations that enable uncontrolled cellular proliferation, evasion of apoptosis, and sustained angiogenesis. Modern high-throughput technologies, such as RNA sequencing (RNA-seq), allow for the comprehensive exploration of these processes at the transcriptomic level, providing unprecedented insights into tumor biology, intratumoral heterogeneity, and dysregulated signaling pathways.⁷⁻¹⁰ Concurrently, bioinformatics has become indispensable for interpreting these vast datasets, revealing critical genetic and transcriptional differences that underlie various malignancies.

In this context, machine learning (ML), a branch of artificial intelligence (AI), offers a powerful paradigm for identifying subtle, clinically relevant patterns within complex biological data.¹¹⁻¹³ The integration of ML with bioinformatics holds exceptional promise for discovering novel molecular biomarkers, which are crucial for early diagnosis, accurate prognosis, and personalized treatment strategies.^{5,14} Such biomarkers can provide objective data for distinguishing malignancy subtypes and predicting tumor behavior.^{5,15} Pioneering efforts, such as the French Sarcoma Group's CINSARC signature a 67-gene classifier identified through differential expression and ML demonstrate the potential of this approach to predict metastatic risk in sarcomas.^{16,17}

Publicly available genomic repositories, such as The Cancer Genome Atlas (TCGA), provide a critical resource for such investigations by offering comprehensive, well-annotated molecular and clinical data across numerous cancer types, including STS.¹⁷⁻¹⁹ Leveraging this resource, the present study employs an integrative bioinformatics and ML framework to analyze a large cohort of STS samples. Our objectives are to identify robust differentially expressed genes (DEGs) and to validate novel diagnostic and prognostic biomarkers, thereby addressing a critical gap in the current management of STS.

2. Materials and Methods

2.1. Data Acquisition

We obtained RNA-seq expression data and corresponding clinical metadata for soft tissue sarcoma samples from The Cancer Genome Atlas Sarcoma cohort (TCGA-SARC) (<https://tcga-data.nci.nih.gov/tcga/>). The TCGA-SARC dataset included 261 tumor samples, comprising 51 metastatic, 121 primary, and 89 recurrent tumor samples. Because the main objective of this study was to identify prognostic biomarkers in STS, the TCGA-SARC cohort was used as the primary discovery dataset for tumor-based prognostic and survival-related analyses. Specifically, primary and metastatic tumor samples were included in the main TCGA-based analysis. For external validation, independent STS tumor and normal tissue samples were retrieved from GEO datasets GSE144190 and GSE21122. Because the objective of this study was to identify broadly relevant candidate biomarkers across the TCGA-SARC cohort, analyses were performed using a pan-STs approach. Subtype-stratified modeling was not performed because several histological subtypes had limited sample sizes, which could reduce statistical power and increase the risk of unstable estimates.

2.2. Data Preprocessing and Identification of Differentially Expressed Genes (DEGs)

Transcriptomic preprocessing and differential-expression analysis were performed independently from downstream candidate prioritization. After removal of duplicated genes and low-quality samples, remaining 20,531 genes were normalized using the Limma and edgeR packages. DEGs were identified by applying a threshold of $|\text{LogFC}| \geq 1.5$ and an adjusted $p\text{-value} < 0.05$. The top-ranked DEGs prioritized by the deep-learning model are presented in Table 1. All subsequent analyses and visualization were performed using R software (version 4.01).

2.3. Identification of Prognostic Markers

To identify diagnostic and prognostic biomarkers, we performed a bioinformatic analysis integrating differential expression screening with ML. Multiple algorithms, including Random Forest (RF), Support Vector Machine (SVM), Logistic Regression, Decision Trees (DTs), and k-Nearest Neighbors (KNN), were employed for feature selection and classification.

Table 1. The Top DEGs of TCGA Were Ranked by Deep Learning

Gene name	Coefficient	
GJB1	1	0.001557
OR2T10	0.97313	0.001515
NAT8	0.956114	0.001489
SLC26A4	0.940283	0.001464
PDZK1	0.939287	0.001463
ATP6V0D2	0.926148	0.001442
SLC17A1	0.920238	0.001433
GLYAT	0.915508	0.001426
UGT3A1	0.914528	0.001424
FGF1	0.91448	0.001424

Log fold change $|\text{LogFC}| \geq 1.5$ and P value < 0.05 .

2.4. Machine Learning Algorithms

In this study, deep learning (DL) was applied as part of the computational biomarker-discovery workflow to prioritize candidate genes from the filtered DEG set and to evaluate their predictive relevance for STS classification. After differential expression analysis and correlation-based feature filtering, selected gene-expression features were used as input variables for model development.

2.5. Computational Workflow

The DL model was configured with a learning rate of 0.01 and the Rectified Linear Unit (ReLU) activation function, and was trained for 20 epochs. The following standardized workflow was employed for model development, evaluation, and optimization.

2.5.1. Data Splitting

To enable an independent evaluation of model performance, the source dataset was partitioned into distinct training and test subsets. The training set was used for model optimization, while the held-out test set provided an unbiased assessment of its predictive capability.

2.5.2. Model Training and Validation

A 70/30 split was used to allocate data to the training and test sets, respectively. For each model, the fixed set of optimal hyperparameters was used to retrain the model on a randomly sampled training dataset, with final performance evaluated on the unused test data to estimate predictive generalization.

2.5.3. Model Evaluation

Model performance was quantitatively assessed using five key metrics: accuracy, R^2 Score, mean squared error (MSE), root mean squared error (RMSE), and the area under the receiver operating characteristic curve (AUC). These metrics comprehensively evaluate the model's predictive accuracy and discriminative power (Table 2).

2.6. Implementation and Evaluation

All machine learning analyses were implemented in Python (version 3.7) utilizing key libraries including Pandas, NumPy, Matplotlib, and Scikit-learn. Models were trained on the training data and their performance was independently evaluated on the held-out test set.

To determine the optimal data partitioning strategy, various train/test splits (ranging from 40/60 to 95/5) were evaluated and compared. A 70/30 ratio was ultimately selected for all subsequent ML models, as it provided the best balance between training data volume and reliable testing, and was consistent with the partitioning used for the PPI network analysis.

Table 2. Machine Learning Algorithm

Deep learning	
Accuracy	98.86%
MSE	9.952972E-10
R2-SCORE	0.9999999
AUC	0.99800
RMSE	3.154833E-5
Sensitivity	99.99%
Specificity	50.00%

Model performance was assessed using five distinct metrics to evaluate different aspects of predictive ability: accuracy, R^2 Score, MSE, RMSE, and the AUC. The AUC quantifies the model's overall ability to discriminate between classes across all classification thresholds. A higher AUC indicates superior discriminative performance.

Accuracy metric represents the proportion of total correct predictions (both true positives and true negatives) among the total number of cases examined, providing a general measure of classification correctness.

The R^2 score was used to evaluate the performance of regression-based analyses within the feature selection process, indicating the proportion of variance in the response variable explained by the model.²⁰⁻²²

The MSE and RMSE were used to quantify prediction inaccuracy. The RMSE, in particular, is a widely adopted standard for assessing model performance. As it preserves the units of the response variable, it offers an interpretable estimate of the average prediction error magnitude and is a robust metric when errors are normally distributed.^{23,24}

2.7. Functional and Pathway Enrichment Analysis

Functional enrichment analysis of the DEG signature was performed using the clusterProfiler package in R. Significantly enriched pathways and terms were identified with an adjusted p-value threshold of < 0.05 . The selected prognostic genes were further annotated and visualized based on the Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) databases.

2.8. Protein-Protein Interaction (PPI) Network Analysis

The PPI network for the DEGs was constructed using the STRING database (<https://string-db.org/>).²⁵ An interaction score threshold of > 0.4 was applied to define statistically significant interactions. This network provides a framework for interpreting functional genomics and elucidating relevant cellular pathways.

2.9. Correlation Between DEGs and Clinical Characteristics

After identification of the 5204 DEGs, Spearman rank correlation analysis was performed to assess the relationships among DEG expression features and relevant clinical characteristics. Correlation coefficients were calculated in R using the cor function with the Spearman method, and the resulting correlation matrix was visualized using the ggcorrplot package. In addition, correlation-based filtering was applied after DEG identification and before machine-learning model construction to reduce feature redundancy. Features were filtered using the prespecified correlation coefficient threshold of Spearman's $\rho > 0.5$.

2.10. Identification of Prognostic Markers

Prognostic potential was assessed for the top DEGs by generating Kaplan-Meier survival curves using the GDCRNATools and ggplot2 R packages. Genes were screened based on a hazard ratio (HR) > 1 and p-value < 0.05 . This analysis identified 26 candidate prognostic genes, comprising 15 up-regulated and 11 down-regulated genes.

2.11. Combined Receiver Operating Characteristic (ROC) Curve

To assess the diagnostic efficacy of the biomarkers, we employed a generalized linear model (GLM) followed by a combined ROC curve analysis. This process was executed using the *combioROC* package in R. The discrimination ability of individual and paired biomarkers was quantified by determining the AUC, sensitivity, specificity, optimal cut-off value, positive predictive value (PPV), and negative predictive value (NPV).

2.12. Validation of Biomarker Gene Expression

The expression levels of the potential biomarker genes were validated using independent datasets from the GEO and TCGA via the Broad Institute's Global Data Assembly Center (GDAC) firehose (<https://gdac.broadinstitute.org/>). Data from these repositories were obtained and pre-processed to analyze gene expression in independent STS patient cohorts.

3. Results

3.1. Patient Demographics

The study comprised 261 STS patients from TCGA. The mean age was 60.87 ± 14.652 years, with 119 males (45.6%) and 142 females (54.4%). Among the patients, 228 were White (87.4%), 18 were Black (6.9%), and 6 were Asian (2.3%). Evidence of metastasis was reported in 56 patients (21.5%), while 121 patients (46.4%) had no metastasis. The average overall survival was 863.44 days. The clinicopathological characteristics of the TCGA-SARC cohort are summarized in [Table 3](#).

3.2. Identification of DEGs

Based on the predefined differential expression criteria, a total of 5,204 DEGs were identified in the exploratory tumor-reference comparison. These DEGs were considered the initial gene-level feature set for downstream analyses, including functional annotation, survival analysis, diagnostic evaluation, and machine-learning-based prioritization. The expression patterns of the identified DEGs are shown in the heatmap presented in [Supplementary 1](#).

To reduce feature redundancy before machine-learning model development, a correlation-based filtering step was applied after DEG identification and before ML-based feature selection/model training. Specifically, Spearman rank correlation analysis was performed on the DEG expression matrix, and features were filtered using the prespecified correlation coefficient threshold (Spearman's $\rho > 0.5$). The resulting reduced feature set was then used as input for machine-learning model construction and evaluation.

The machine-learning and deep-learning models were evaluated using five performance metrics: accuracy, R^2 score, mean squared error, root mean squared error, and area under the receiver operating characteristic curve. Through this integrated DEG filtering and ML-based prioritization strategy, 26 key genes were selected for further prognostic and functional investigation. Among these, 15 genes were identified as upregulated in STS tumors, including *HIST1H1E*,

Table 3. The Clinicopathological Characteristics of STS Patients

Clinicopathological variables	No. of patients (%) / mean $\pm$ SD
Patients	261
Mean Age (years, Mean $\pm$ SD)	60.87 ± 14.652
Sex	
Male	119 (45.6)
Female	142 (54.4)
Race	
Asian	6 (2.3)
White	228 (87.4)
Black	18 (6.9)
Metastasis	
Yes	56 (21.5)
No	121 (46.4)
Dead	99 (37.9)
Alive	162 (62.1)

C20orf152, *CARTPT*, *ST7OT4*, *MAGEA8*, *HIST1H4E*, *RPA4*, *PRPS1L1*, *ZNF732*, *TAS2R10*, *ANGPTL3*, *UGT1A6*, *AQP2*, *KL*, and *CLCNKB*. Conversely, 11 genes were downregulated, including *ADAM21P1*, *C6orf163*, *GDF7*, *DSCR9*, *LECT2*, *NCRNA00169*, *SNORD1C*, *TTC9B*, *PBLD*, *APOM*, and *PIPOX*. These 26 key genes were selected for further prognostic and functional investigation.

3.3. GO, Functional Annotation, and Pathway Enrichment Analyses

To elucidate the functional implications of the key DEGs, GO functional annotation and KEGG/pathway enrichment analyses were performed. Enrichment results were considered statistically significant at an adjusted p-value threshold of $p.adjust < 0.05$. Pathway enrichment analysis revealed significant enrichment of several biologically relevant pathways, including Transmission across Chemical Synapses, GABA receptor activation, Potassium Channels, GPCR ligand binding/GPCR downstream signaling, Collagen formation, Collagen degradation, Extracellular matrix organization, and Drug ADME/Biological oxidations, all of which met the predefined adjusted significance threshold of $p.adjust < 0.05$ (Figure 1a). These enriched pathways suggest that the identified DEGs are involved in neurotransmission-related signaling, ion-channel regulation, extracellular matrix remodeling, GPCR-mediated signaling, and metabolic/xenobiotic-related processes.

GO/functional enrichment analysis further supported the involvement of the DEG set in cellular communication, matrix organization, ion transport, and signaling-related biological processes. The most relevant significantly enriched functional terms included Extracellular matrix organization, Cell junction organization, Ion channel transport, Transport of small molecules, Collagen biosynthesis and modifying enzymes, Muscle contraction/striated muscle contraction, Neuronal system/neurotransmitter receptor-related signaling, and Regulation of gene expression in beta cells. All listed GO/functional terms met the adjusted significance threshold of $p.adjust < 0.05$ (Figure 1b). Collectively, these results indicate that the DEGs are primarily associated with extracellular matrix remodeling, cell-cell communication, membrane transport, ion-channel activity, receptor-mediated signaling, and tissue-structural organization.

3.4. Correlation Between Clinical/Demographic Data and DEGs

Correlation analysis of clinicopathological and demographic variables showed a moderate positive correlation between race and ethnicity ($r = 0.30$, $p = 0.000056$). Gender showed a weak but statistically significant correlation with race ($r = 0.20$, $p = 0.006$) and between metastatic diagnosis and neoplasm cancer status ($r = 0.10$, $p = 0.011$). No strong correlations were identified among the evaluated variables. The correlation matrix is shown in Figure 2.

3.5. PPI Network Construction

To identify functional protein modules, a PPI network was generated by submitting the DEGs to the STRING database, applying an interaction score threshold of 0.4. Network analysis highlighted several substantial interactions, including *PDZK1-SLC17A1* and *H1-4-H4C6* (Figure 3A–C). These interactions were identified from the STRING-based computational PPI network; therefore, they should be interpreted as candidate functional associations rather than experimentally validated sarcoma-specific interactions.

3.6. Identification of Prognostic DEGs

Kaplan–Meier survival analysis was employed to evaluate the prognostic significance of key gene signatures in STS. High expression of fifteen genes (*HIST1H1E*, *C20orf152*, *CARTPT*, *ST7OT4*, *MAGEA8*, *HIST1H4E*, *RPA4*, *PRPS1L1*, *ZNF732*, *TAS2R10*, *ANGPTL3*, *UGT1A6*, *AQP2*, *KL*, and *CLCNKB*) was significantly associated with reduced overall survival. Conversely, lower expression of eleven genes (*ADAM21P1*, *C6orf163*, *GDF7*, *DSCR9*, *LECT2*, *NCRNA00169*, *SNORD1C*, *TTC9B*, *PBLD*, *APOM*, and *PIPOX*) was correlated with a poorer prognosis and decreased survival. All analyses were performed using SPSS version 20, with a statistical significance level of $p < 0.05$.

3.7. ROC Curve Analysis for Diagnostic Biomarker

ROC curve analysis was performed to evaluate the diagnostic potential of individual genes and gene combinations for distinguishing STS tumor samples from reference/non-tumor samples. The diagnostic analysis was reported using a STARD-informed framework, including the diagnostic cut-off, AUC with 95% confidence interval, sensitivity with 95% confidence interval, specificity with 95% confidence interval, PPV, NPV, reference standard, and sample composition. STARD

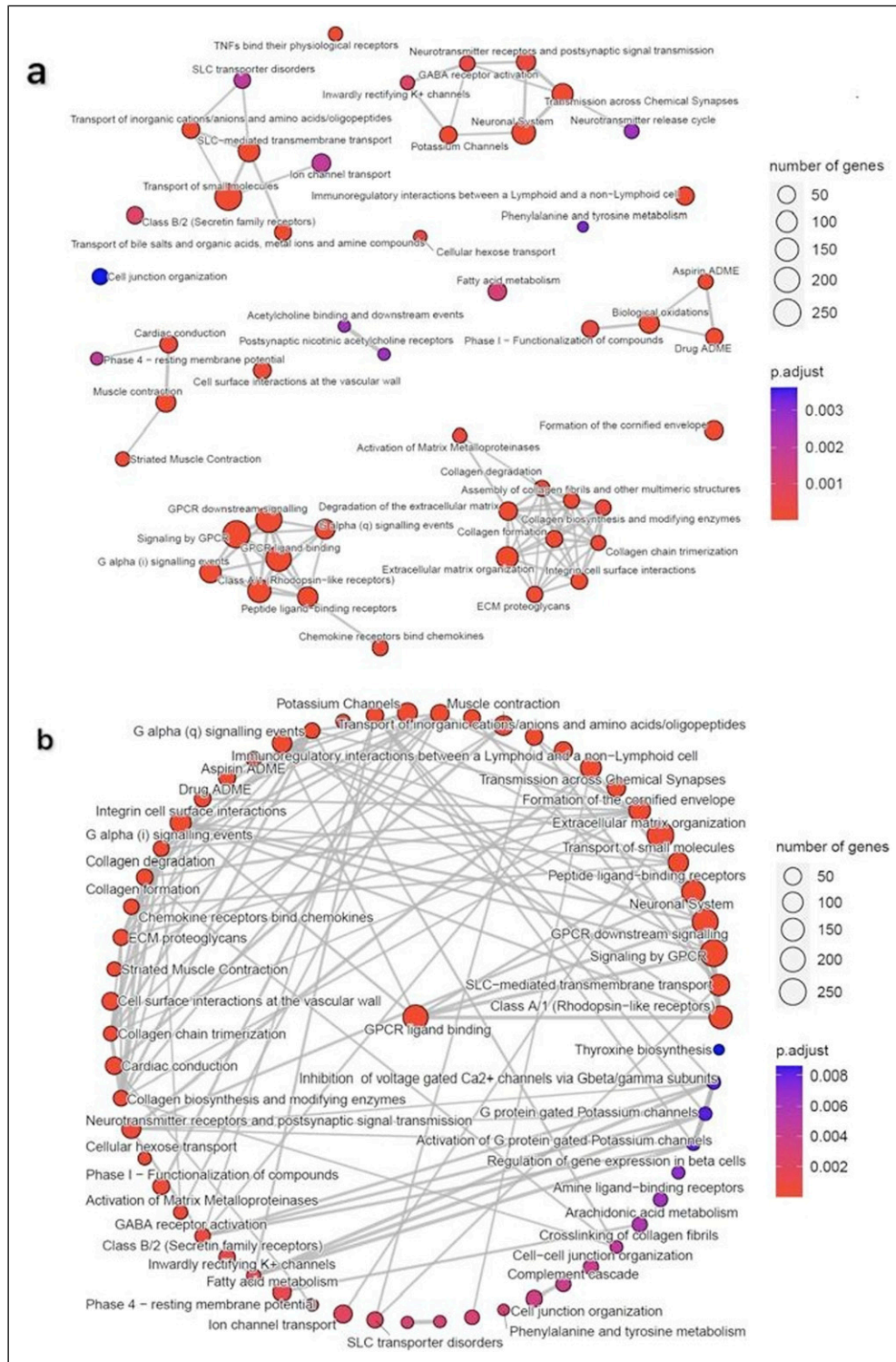


Figure 1. Functional enrichment analysis of differentially expressed genes in soft tissue sarcoma. (A) Pathway enrichment network of DEGs identified in the TCGA-SARC cohort, showing important enriched pathways including synaptic transmission, GABA receptor activation, potassium-channel activity, GPCR signaling, collagen formation/degradation, extracellular matrix organization, biological oxidation, and drug ADME-related pathways. Node size represents the number of genes, and node color indicates the adjusted p-value. (B) Functional annotation network of enriched biological processes, highlighting extracellular matrix remodeling, cell-junction organization, ion-channel transport, small-molecule transport, collagen biosynthesis, muscle contraction, and receptor-mediated signaling. DEGs, differentially expressed genes; STS, soft tissue sarcoma; GPCR, G-protein-coupled receptor; ADME, absorption, distribution, metabolism, and excretion

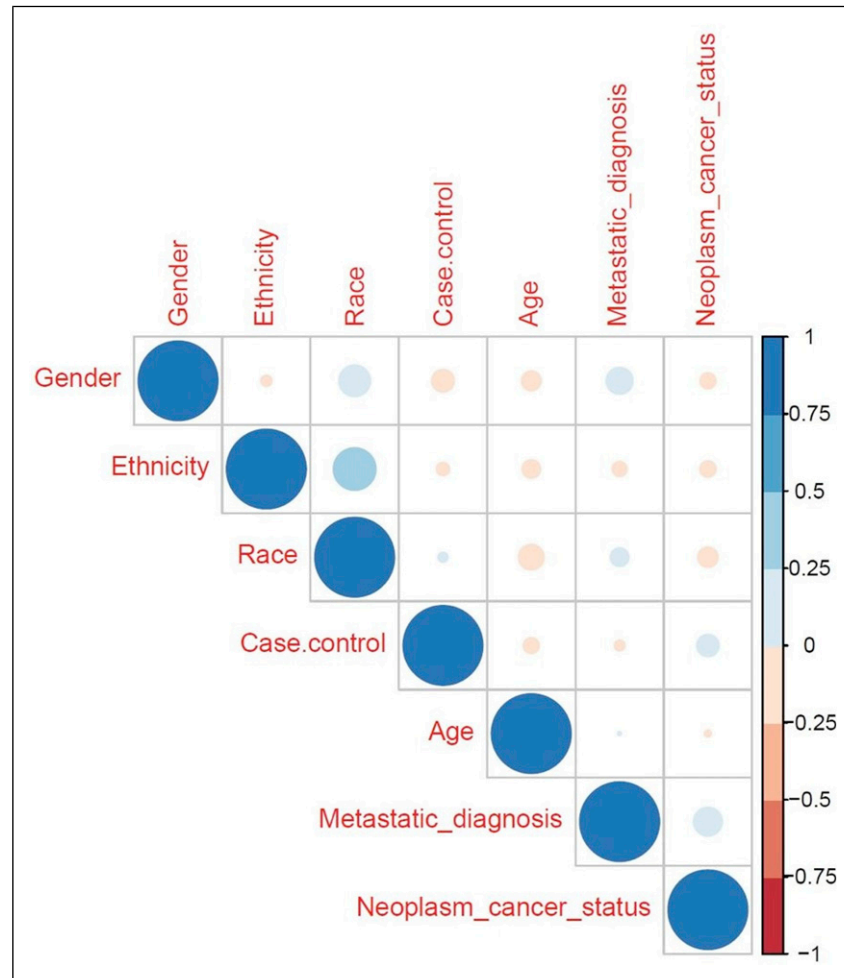


Figure 2. Correlation matrix of clinical and demographic variables in the TCGA-SARC cohort. Blue and red circles indicate positive and negative correlations, respectively. Circle size and color intensity are associated with the magnitude of the correlation coefficients. Crosses indicate non-significant correlations. The strongest observed association was between race and ethnicity, while other significant correlations were weak

2015 emphasizes transparent reporting of diagnostic accuracy studies because incomplete reporting can limit the assessment of bias, applicability, and clinical interpretability.

Among individual genes, *AICF* showed the highest diagnostic performance, with an AUC of 0.70. In combined biomarker models, *AICF-ATP6V0D2* and *AICF-LECT2* demonstrated improved diagnostic performance, with AUC values of 0.743 and 0.796, respectively (Figure 4). These *AICF*-based diagnostic models reflect tumor-associated expression differences between STS tumor samples and normal/reference samples in the analyzed datasets. At the selected diagnostic cut-off, the models achieved a sensitivity of 1.0 and a specificity of 0.5. Therefore, although these markers showed high sensitivity, their specificity was modest, indicating a substantial false-positive rate among reference/non-tumor samples.

To improve interpretability, the diagnostic performance metrics are now reported with 95% confidence intervals and predictive values. The diagnostic summary, including log coefficient, AUC, AIC, sensitivity, specificity, is provided in Table 4. Because PPV and NPV are influenced by disease prevalence, these values were also interpreted under realistic STS prevalence scenarios rather than only within the study dataset. Accordingly, these biomarkers should be considered exploratory candidates with potential utility for screening or triage, where high sensitivity may be valuable, but they are not sufficient as standalone confirmatory diagnostic tests without further validation and improved specificity.

A generalized linear model was further fitted to assess the contribution of *AICF*-based biomarker panels to diagnostic classification. The model coefficients for *AICF*, *AICF-ATP6V0D2*, and *AICF-LECT2* are presented in Table 4, supporting their contributions to the diagnostic model. However, given the modest specificity, these findings should be interpreted as preliminary and require validation in larger, independent cohorts before clinical application.

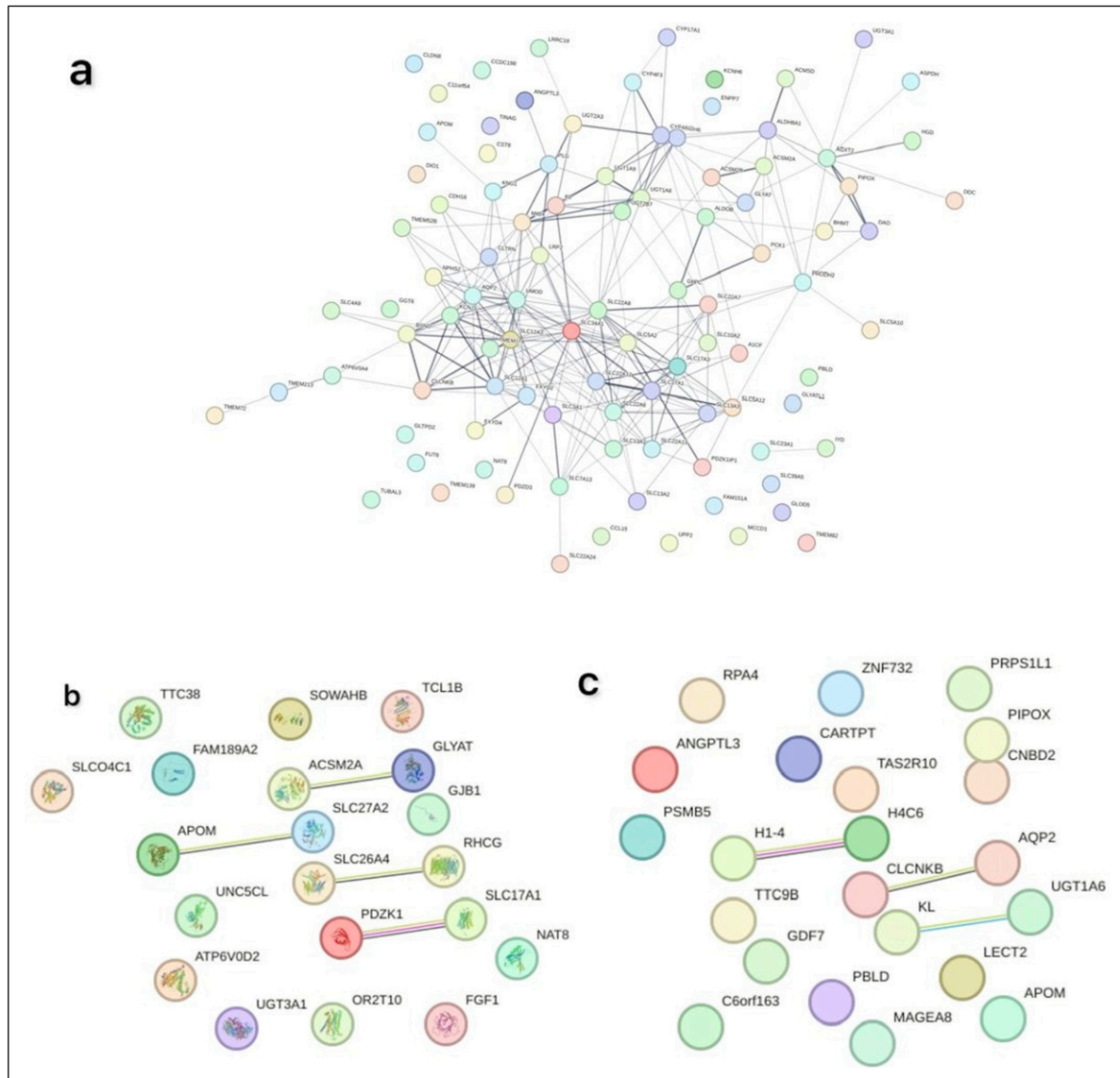


Figure 3. Protein–protein interaction network of candidate DEGs in soft tissue sarcoma. (A) STRING-based PPI network constructed from candidate DEGs using an interaction score threshold of 0.4. Nodes represent proteins, and edges represent predicted or known functional associations. (B) PPI network of top DEGs, highlighting key interactions including the PDZK1–SLC17A1 transporter/scaffold-related module. (C) PPI network of the 26 prognostic candidate genes, including the H1-4–H4C6 chromatin-associated interaction. These STRING-derived interactions should be interpreted as candidate functional associations requiring experimental validation. (PPI: protein–protein interaction; DEG: differentially expressed gene)

3.8. Validation

External validation was performed to assess *AICF* dysregulation across independent datasets. In addition to the GDAC Firehose validation analysis, two independent external datasets were included to further evaluate the expression pattern of *AICF* in STS tumor tissues compared with normal/reference controls. Detailed information regarding the validation datasets, including dataset source, sample size, sample composition, platform, and validation results, is provided in [Supplementary Table 1](#).

To further evaluate the relationship between *AICF* dysregulation and available clinicopathological variables, Spearman rank correlation analysis was performed. *AICF* expression showed no statistically significant correlation with sex, age, or tumor size. Specifically, weak and non-significant correlations were observed for sex (Spearman's $\rho = 0.0789$, $p = 0.2596$), age (Spearman's $\rho = 0.0800$, $p = 0.2810$), and tumor size (Spearman's $\rho = -0.1100$, $p = 0.1270$). These findings indicate that

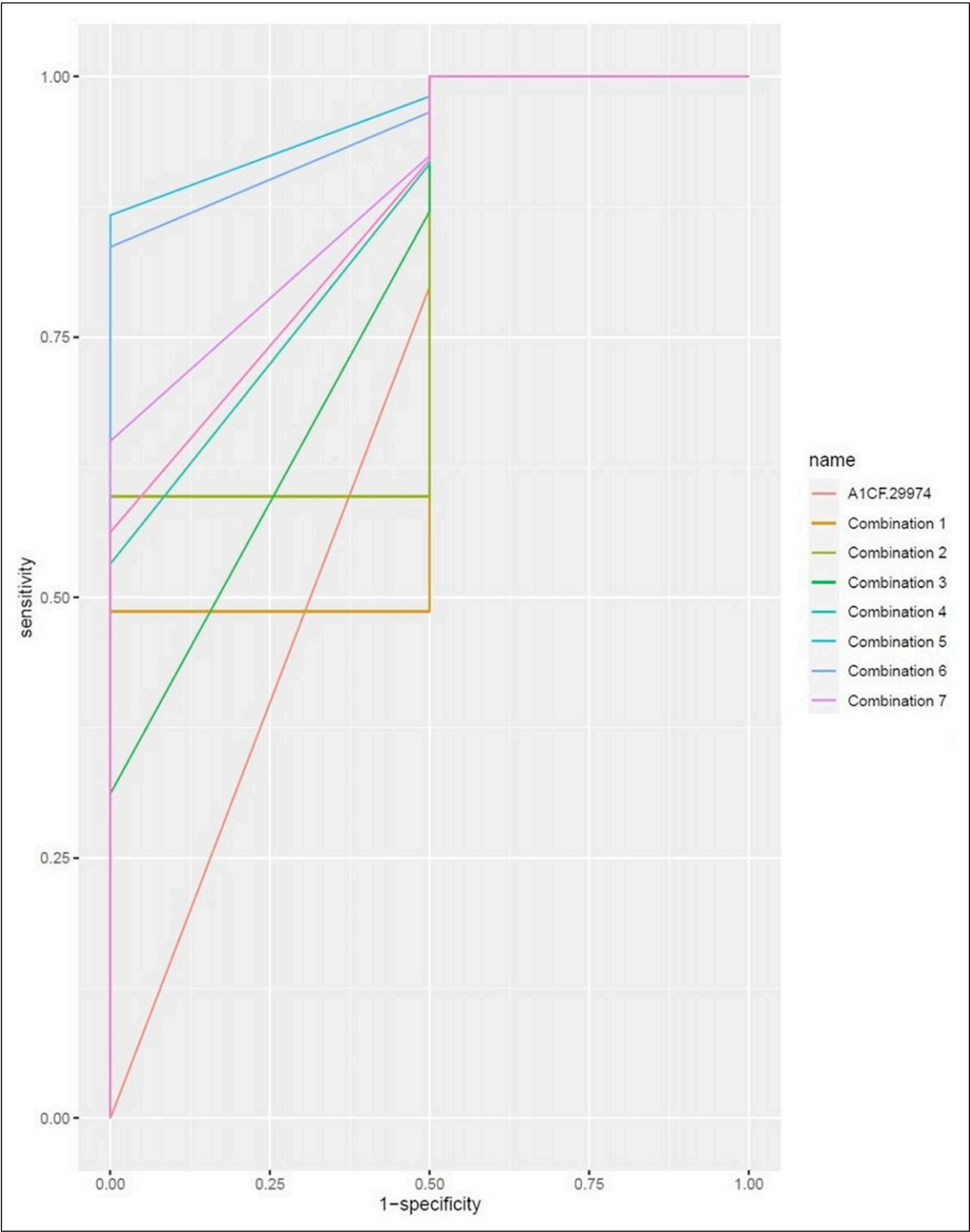


Figure 4. Receiver operating characteristic (ROC) curve analysis of A1CF and A1CF-based combinations as exploratory diagnostic biomarker candidates for STS. The curves represent A1CF alone and the following combinations: Combination 1, A1CF-ATP6V0D2; Combination 2, A1CF-C20orf152; Combination 3, A1CF-LECT2; Combination 4, A1CF-MAGEA8; Combination 5, A1CF-PRPS1L1; Combination 6, A1CF-ST7OT4; Combination 7, A1CF-TAS2R10

Table 4. GLM Analysis for Detection of More Valuable Combination

Gene	Log	AIC	Residual	AUC	95% CI for AUC	Sensitivity	95% CI	Specificity
A1CF	-1.471	18.27	14.27	0.700	0.358–1.000	1	0.832–1.000	0.5
A1CF-ATP6V0D2	-1.51390	20.27	14.27	0.743	0.431–1.000	1	0.832–1.000	0.5
A1CF-LECT2	-1.411	19.37	13.37	0.796	0.526–1.000	1	0.832–1.000	0.5

A1CF expression was not significantly associated with the evaluated clinicopathological parameters in the available cohort. These validation analyses were based on transcriptomic datasets and therefore support external expression-level validation rather than experimental biological validation.

4. Discussion

Using public transcriptomic data from the TCGA-SARC cohort, we identified a candidate gene set associated with sarcoma. Subsequent ML analysis revealed novel biomarker candidates for STS detection and prognosis. The top-performing biomarkers, *A1CF* and the combinatorial panels *A1CF-ATP6V0D2* and *A1CF-LECT2*, were subsequently confirmed in an independent validation dataset from the GDAC. These results nominate exploratory molecular signatures for STS diagnosis and prognosis and provide a basis for further clinical validation. Because STS comprises multiple histological and molecular subtypes, the findings of this study should be interpreted within the context of a pan-STS analysis. The present study was designed to identify broad candidate biomarkers across the available TCGA-SARC cohort rather than to develop subtype-specific diagnostic or prognostic classifiers. Although this approach may identify biomarkers with potential relevance across STS, it may also obscure subtype-specific molecular patterns. Therefore, the identified *A1CF*-based diagnostic panels and prognostic gene signature should be regarded as exploratory pan-STS candidates rather than subtype-specific biomarkers. Cancer develops from dysregulated cell growth at different phases of the cell cycle.²⁶ Accurate and precise diagnosis is essential for ensuring the best possible treatment strategies for improving patient outcomes. Currently, diagnosis and classification approaches mostly rely on subjective assessment by physicians, which is susceptible to human error.²⁷ Despite advancements and new treatment strategies, metastatic sarcoma survival remains poor. This might imply that administering strong chemotherapy to all patients with metastatic sarcoma without prognosis classification is not the best strategy to enhance results.²⁸ STS subtypes may appear similar but vary greatly in prognosis and often lack specific treatments. Pathologists face challenges diagnosing STS due to overlapping morphologies.¹⁷ Even with advancements in diagnostic methods, STS frequently presents with recurrences or metastasis. Its diverse histological subtypes complicate diagnosis.²⁹ Molecular diagnostic techniques can provide highly accurate and reliable approaches for tumor classification, thereby enabling personalized therapeutic strategies. However, they are not yet routinely used in clinical management.²⁷

TCGA has provided one of the most comprehensive STS sequencing studies.¹⁷ Omics data can enhance oncogenesis understanding, improving diagnosis accuracy and targeted therapy development.³⁰ Some computational techniques have been developed to extract the markers from the complex multi-omics datasets, such as ML and DL algorithms.³¹ The integration of ML and bioinformatics provides the potential for cancer diagnosis, subtyping, histology, therapeutic targeting, and prognosis.³² The French Sarcoma Group used ML on a large STS cohort to validate a 67-gene set (CINSARC) for predicting metastatic prognosis.¹⁶ In fact, ML algorithms are effective tools for understanding rare tumors. They enable deeper insight into STS biology through advanced data analysis.¹⁷ Advances in proteomics, metabolomics, transcriptomics, and genomics now enable the realization of precision medicine. Combining ML techniques with multi-omics data aims to provide a comprehensive understanding of the underlying pathophysiology. In particular, the assessment of clinical omics datasets has facilitated the development of diagnostic and prognostic models for various disorders. Bioinformatics utilizes omics data to predict disease outcomes, thereby enhancing our capacity for early detection and personalized therapy.³³ Big data and complex algorithms are used in AI-driven medical research to increase the precision of tumor prognosis. Scientists can investigate the intrinsic regulatory mechanisms underlying carcinogenesis and disease progression more effectively due to advancements in bioinformatics.³⁴ The combination of bioinformatics and statistical techniques facilitates more accurate identification of potential molecular biomarkers, while also providing a cost-effective and time-efficient alternative to traditional wet-lab experimental procedures.³⁵

One of the most popular methods for molecular analysis of diseases is transcriptomics, through applying high-throughput techniques (RNA-seq). Transcriptomics has been utilized to identify prognostic and diagnostic biomarkers, as well to comprehend the pathophysiology of illnesses and treatment targets.³³ Bioinformatics advances allow complex exploration of gene expression, cancer prognosis, and treatment responses.³²

Gene expression signatures can serve as molecular correlates of clinical characteristics of malignancies, with DEGs.²⁷ Variations in gene transcript levels can affect critical processes such as invasion, metastasis, immune evasion and tumor growth, thereby contributing to disease progression and affecting host tissue integrity through indirect pathogenic mechanisms.³⁶ Understanding these gene expression patterns facilitates the identification of biomarkers that can serve as diagnostic and prognostic biomarkers. Functional enrichment analysis suggested that the identified DEGs are involved in extracellular matrix remodeling, collagen organization, cell–cell communication, ion-channel regulation, GPCR-mediated signaling, molecular transport, and metabolic/xenobiotic-related processes. These pathways are biologically relevant to tumor invasion, stromal remodeling, and the sarcoma microenvironment; however, they do not establish causal mechanisms or subtype specificity and require experimental validation.

The PPI network analysis highlighted PDZK1–SLC17A1 and H1-4–H4C6 as prominent computationally prioritized interactions. Because these interactions were derived from a STRING-based PPI network, they should be considered as candidate functional associations rather than validated sarcoma-specific protein–protein interactions.²⁵ The PDZK1–SLC17A1 interaction may reflect transporter related and scaffold related regulation. PDZK1 is a multi-PDZ-domain scaffold protein involved in organizing membrane-associated transporter complexes, while SLC17A1/NPT1 is a solute carrier transporter involved in renal phosphate and urate transport.^{37,38} PDZK1 has been reported to influence the processes of cancer, including proliferation, migration, invasion, apoptosis, and cell-cycle regulation, and clinical outcome in renal cell carcinoma and hepatocellular carcinoma.^{39,40} However, direct evidence supporting a sarcoma-specific role for the PDZK1–SLC17A1 interaction is currently limited. Therefore, this interaction should be considered a candidate transporter/scaffold-related module requiring further experimental validation.

The H1-4–H4C6 interaction likely reflects chromatin and nucleosome associated regulation. H1-4 is a linker histone, whereas H4C6 encodes a core histone H4 protein. Linker and core histones regulate chromatin organization, transcriptional control, DNA replication, DNA repair, and genome stability, all of which are frequently dysregulated in cancer.^{41,42} Thus, the H1-4–H4C6 interaction may suggest chromatin-level dysregulation within the STS transcriptomic profile. Nevertheless, because direct evidence of these in sarcoma is lacking, this finding should be interpreted as a chromatin-associated signal.

As diagnosis of STS is challenging, which is critical for timely initiation of appropriate treatment and can significantly influence patient outcomes and survival rates, we obtained the data of STS patients available in the TCGA database,^{43,44} and it enabled molecular analyses through RNA-seq and ML algorithms to study sarcoma patients.⁴⁵

Biomarkers have revolutionized cancer care, improving outcomes. They may be predictive, prognostic, or diagnostic.²⁹ They can reduce the impact of histological variability.⁴⁶ Determining prognosis remains a major challenge in cancer management, and accurate diagnosis is essential for selecting proper treatment strategies.²⁷ Prognostic biomarkers can help with risk stratification and treatment planning, indicating favorable or poor outcomes.⁴⁶ The best approach involves a combination of molecular biomarkers with conventional clinical and pathological prognostic variables. Achievement of favorable outcomes in cancer research requires the application of several methodologies and comprehensive data analysis approaches.²⁷ Identification of new biomarkers for early diagnosis and appropriate treatment is critically important, as early detection leads to lower mortality rates and improved prognosis.⁴⁷ Early detection biomarkers are key factors to prevent poor prognosis.⁴⁸

Several studies have examined biomarkers expressed in sarcomas that are relevant to treatment, including CD3, CD4, CD8, FOXP3, and CD20. However, most involved small samples due to the rarity of sarcomas. More research is needed, especially on specific subtypes and biomarkers.⁴⁶ Reliable molecular markers are needed to differentiate subtypes and identify new targets.⁴⁹

AICF (Apolipoprotein-B 100 Complementing Factor) is an RNA-binding protein and a component of the APOBEC complex, where the C-to-U conversion happens during apolipoprotein-B (ApoB) mRNA editing. ApoB100 has multiple functions in the regulation of metabolism. *AICF* increases the RNA-modifying capacity of *APOBEC1*. The domains of *AICF* determine the protein substrate specificity by identifying an 11 nt docking region on modified transcripts.⁵⁰ In this study, we investigated *AICF* as an independent diagnostic factor for sarcoma patients.

The function of *AICF* in sarcoma remains unclear, and no previous study has directly defined its mechanistic role in STS. However, evidence from other solid tumors suggests that *AICF* may influence tumor-associated phenotypes through post-transcriptional RNA regulation. In glioma, *AICF* was reported to participate in the *AICF*–FAM224A–miR-590-3p–ZNF143 regulatory loop, which promoted malignant biological behaviors of glioma cells.⁵¹ In renal cell carcinoma, *AICF* facilitated cell migration by promoting nuclear translocation of SMAD3.⁵² In breast cancer cells, *AICF* regulated migration and apoptosis through effects on the 3' untranslated region of DKK1,⁵³ while in Wilms tumor-derived cells, the *AICF*–Axin2 axis influenced apoptosis and migration through the Wnt/ β -catenin pathway.⁵⁴ A recent review also summarized emerging evidence linking *AICF* to several malignancies, including hepatocellular carcinoma, renal cell carcinoma, breast cancer, endometrial cancer, lung cancer, and glioma.⁵⁵ Based on these findings, we hypothesize that downregulation of *AICF* in STS may contribute to sarcoma pathogenesis by altering RNA-binding-dependent regulation of transcripts involved in

proliferation, survival, migration, inflammatory signaling, and extracellular-matrix remodeling. Nevertheless, this interpretation remains exploratory, and functional studies in sarcoma-specific models are required to determine whether A1CF has a causal role in STS initiation, progression, or metastatic behavior.

Dysregulation of most of these prognostic biomarkers, such as *GPR160*, *ST7*, *MAGEA8*, *HIST1H4E*, *TAS2R10*, *ANGPTL*, *UGT1A*, and *AQP2*, is associated with different malignancies.⁵⁶⁻⁶³ *RPA4*, *PRPS*, and *ZNF* are involved in DNA synthesis or regulation.⁶⁴⁻⁶⁶

The implementation of medical data in ML systems has made significant advances in resolving clinical challenges such as accurate diagnosis and management of debilitating diseases, especially cancers.^{33,45} RNA sequence analysis is accepted as a standard method for transcriptomic studies.³³ Our data suggest the exploratory diagnostic relevance of the *A1CF* and *A1CF-ATP6V0D2* and *A1CF-LECT2* combinations in STS. In addition, the results of survival analysis showed that overexpression of 15 genes and lower expression of 11 genes were associated with poorer survival outcomes in STS. These genes may have potential relevance for future biomarker development in STS diagnosis and prognosis, but their clinical utility requires validation in larger independent and prospective cohorts. Our findings are consistent with prior research in the field of malignancies. However, validating the clinical usage of these genes in diagnosis, prognosis, and personalized treatment necessitates further clinical investigations involving larger patient cohorts. Personalized management and patient education are interdependent components of modern, high-quality cancer care.⁶⁷ These candidate biomarkers may provide crucial insights into STS diagnosis and prognosis, particularly when evaluated in combination; however, their clinical and therapeutic relevance remains to be established.

5. Limitations and Future Directions

While big data analytics present considerable benefits for enhancing healthcare quality and operational efficiency, their implementation is accompanied by notable limitations. A primary impediment lies in the decentralized and often heterogeneous architecture of clinical data repositories, which complicates comprehensive integration. Furthermore, the absence of prospective clinical validation underscores a broader translational gap inherent in purely computational investigations. Methodological inconsistencies, particularly in extraction protocols, sample acquisition, and the use of reference controls further undermine cross-study comparability. This underscores the necessity for standardized workflows to ensure reproducible quantification of these biomarkers. A further limitation is the modest specificity of the A1CF-based diagnostic panels. Although high sensitivity may be useful in an exploratory screening or triage setting, the specificity of 0.5 limits their clinical utility as standalone diagnostic tests. Larger independent studies with appropriate reference controls and subtype-level annotation are required to determine whether specificity can be improved and whether these markers have reproducible clinical value.

As a follow-up translational step, we plan to validate A1CF and selected genes from the prognostic signature in an independent cohort of patients with soft tissue sarcoma from our referral center. Transcript-level validation may be performed using quantitative real-time PCR, and protein-level validation may be assessed by immunohistochemistry when suitable tissue specimens and validated antibodies are available. These analyses will also aim to evaluate associations between biomarker expression and clinicopathological features, including histological subtype, tumor grade, metastatic status, recurrence, treatment response, and survival outcomes. Furthermore, given the marked histological and molecular heterogeneity of STS, this pan-STs approach may obscure subtype-specific transcriptomic patterns. Because several TCGA-SARC subtypes have limited sample sizes, subtype-stratified machine-learning, diagnostic, and survival analyses were not performed. Therefore, the identified biomarkers should be interpreted as exploratory pan-STs candidates, and future studies should validate them in larger subtype-annotated cohorts.

6. Conclusion

In conclusion, our analysis identified that dysregulation of 26 genes, as well as *A1CF* and its combinations (*A1CF-ATP6V0D2* and *A1CF-LECT2*) might be novel prognostic and diagnostic biomarkers for STS, through bioinformatics and ML frameworks applied to TCGA-SARC data. This research supports the application of bioinformatics and ML as discovery tools for generating testable biomarker hypotheses in rare cancers using public datasets. The discovery of these biomarkers provides exploratory evidence for potential diagnostic and prognostic relevance in STS, but their clinical utility remains to be established through experimental validation and prospective evaluation in independent patient cohorts. Subsequent clinical validation will be crucial to translating these findings into improved patient care.

Acknowledgements

In the composition of this work, artificial intelligence was utilized as an editorial tool to critique prose, suggest structural improvements, and ensure formal academic style, while all original research, analysis, and intellectual contributions remain our own.

ORCID iDs

Masoume Avateffazeli  <https://orcid.org/0000-0002-9723-4676>

Rahem Rahmati  <https://orcid.org/0000-0001-7822-8759>

Shadi Shafaghi  <https://orcid.org/0000-0002-9960-8854>

Ethical Considerations

The data analyzed in this study were obtained from publicly available, de-identified transcriptomic and clinical datasets, including The Cancer Genome Atlas Sarcoma cohort (TCGA-SARC) and Gene Expression Omnibus (GEO) datasets. All patient-related ethical approvals and consent procedures for data collection and public data sharing were managed by the original data-generating institutions and repositories in accordance with their respective ethical guidelines.

This study was approved by the Ethics Committee of Masih Daneshvari Hospital, Shahid Beheshti University of Medical Sciences, Tehran, Iran. The ethical approval number is IR.SBMU.NRITLD.REC.1403.050.

Consent to Participate

No new human participants were recruited, and no identifiable patient, guardian, or participant information was accessed in this study. Therefore, additional patient/guardian/participant informed consent was not required for the present secondary analysis of publicly available de-identified data.

Author Contributions

MA: Investigation; Visualization; Writing – original draft; Writing – review and editing. **RR:** Writing – original draft; Writing – review and editing. **AM:** Investigation; Validation; Visualization. **MH:** Validation; Visualization; Writing – review and editing. **EH:** Conceptualization; Methodology; Software; Formal analysis; Investigation; Validation; Visualization; Supervision. **ShSh:** Methodology; Investigation; Validation; Visualization; Writing – review and editing; Supervision. All authors read and approved the final manuscript.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Availability Statement

The dataset analysed during the current study is available in the TCGA dataset, openly available in the following URL: <https://tcga-data.nci.nih.gov/tcga/>.

Supplemental Material

Supplemental material for this article is available online.

References

1. Qi L, Jiang K, Zhao FF, Ren P, Wang L. Identification of therapeutic targets and prognostic biomarkers in the Siglec family of genes in tumor immune microenvironment of sarcoma. *Sci Rep.* 2024;14(1):577.
2. Lu J, Li T, Liao Z, et al. The efficacies and biomarker investigations of anti-programmed death-1 (anti-PD-1)-based therapies for metastatic bone and soft tissue sarcoma. *Cancer Biol Med.* 2022;19(6):910-930.
3. Grünewald TG, Alonso M, Avnet S, et al. Sarcoma treatment in the era of molecular medicine. *EMBO Mol Med.* 2020;12(11):e11131.
4. Brennan MF, Antonescu CR, Moraco N, Singer S. Lessons learned from the study of 10,000 patients with soft tissue sarcoma. *Ann Surg.* 2014;260(3):416-421.
5. Wei D, Lan X, Huang Z, et al. Pyroptosis-Related Gene Signature Is a Novel Prognostic Biomarker for Sarcoma Patients. *Dis Markers.* 2021;2021(1):9919842.
6. Bae JY, Choi KU, Kim A, et al. Evaluation of immune-biomarker expression in high-grade soft-tissue sarcoma: HLA-DQA1 expression as a prognostic marker. *Exp Ther Med.* 2020;20(5):107.

7. Aremu AO, Ekundina VO, Enye LA, Kehinde AO, Ogunlayi AC. Assessing the Diagnostic Impact of p53, p16, Retinoblastoma and bcl-2 Proteins in Human Papillomavirus-associated Squamous Cell Carcinoma of the Cervix. *J Prev Diagn Treat Strateg Med.* 2024; 3(2):115-121.
8. Eclarin PR, Yan PA, Paliza CL, et al. Benchmarking the distribution coefficient of anticancer lead compounds using the predicted log D values of clinically approved chemotherapeutic drugs. *J Prev Diagn Treat Strateg Med.* 2022;1(2):143-152.
9. Hong M, Tao S, Zhang L, et al. RNA sequencing: new technologies and applications in cancer research. *J Hematol Oncol.* 2020; 13:1-16.
10. Zhang L, Guan M, Zhang X, Yu F, Lai F. Machine-learning and combined analysis of single-cell and bulk-RNA sequencing identified a DC gene signature to predict prognosis and immunotherapy response for patients with lung adenocarcinoma. *J Cancer Res Clin Oncol.* 2023;149(15):13553-13574.
11. Nazari E, Aghemiri M, Avan A, Mehrabian A, Tabesh H. Machine learning approaches for classification of colorectal cancer with and without feature selection method on microarray data. *Gene Rep.* 2021;25:101419.
12. Nazari E, Biviji R, Roshandel D, et al. Decision fusion in healthcare and medicine: a narrative review. *mHealth.* 2022;8:8.
13. Nazari E, Chang H-CH, Deldar K, et al. A comprehensive overview of decision fusion technique in healthcare: A systematic scoping review. *Iran Red Crescent Med J.* 2020;22(10).
14. Ma X, Huang S, Shi H, et al. Identification of ACBD3 as a new molecular biomarker in pan-cancers through bioinformatic analysis: a preclinical study. *Eur J Med Res.* 2023;28(1):590.
15. Doddawad VG, Bannimath G, Shivakumar S, Bannimath N. Biomarkers of oral cancer: A current views and directions. *Biomed Biotechnol Res J.* 2022;6(1):33-39.
16. Chibon F, Lagarde P, Salas S, et al. Validated prediction of clinical outcome in sarcomas and multiple types of cancer on the basis of a gene expression signature related to genome complexity. *Nat Med.* 2010;16(7):781-787.
17. van IJzendoorn DG, Szuhai K, Briaire-de Bruijn IH, Kostine M, Kuijjer ML, Bovée JV. Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas. *PLoS Comput Biol.* 2019;15(2):e1006826.
18. Khalili-Tanha G, Mohit R, Asadnia A, et al. Identification of ZMYND19 as a novel biomarker of colorectal cancer: RNA-sequencing and machine learning analysis. *J Cell Commun Signal.* 2023;17(4):1469-1485.
19. Chen Y, Zhu R, Chen M, et al. Prognostic value of a three-DNA methylation biomarker in patients with soft tissue sarcoma. *J Oncol.* 2020;2020:1-11.
20. Fergus P, Chalmers C. *Performance evaluation metrics. Applied Deep Learning: Tools, Techniques, and Implementation.* Springer; 2022:115-138.
21. Varoquaux G, Colliot O. Evaluating machine learning models and their diagnostic value. *Machine Learning for Brain Disorders.* 2023:601-630.
22. Dinga R, Penninx BW, Veltman DJ, Schmaal L, Marquand AF. *Beyond accuracy: Measures for assessing machine learning models, pitfalls and guidelines.* bioRxiv; 2019.743138.
23. Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE). *Geosci Model Dev Discuss.* 2014;7(1): 1525-1534.
24. Köksoy O. Multiresponse robust design: Mean square error (MSE) criterion. *Appl Math Comput.* 2006;175(2):1716-1729.
25. Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 2019;47(D1):D607-D613.
26. Aran V, Devalle S, Meohas W, et al. Osteosarcoma, chondrosarcoma and Ewing sarcoma: Clinical aspects, biomarker discovery and liquid biopsy. *Crit Rev Oncol Hematol.* 2021;162:103340.
27. Mirza Z, Ansari MS, Iqbal MS, et al. Identification of novel diagnostic and prognostic gene signature biomarkers for breast cancer using artificial intelligence and machine learning assisted transcriptomics analysis. *Cancers.* 2023;15(12):3237.
28. Aggerholm-Pedersen N, Maretty-Kongstad K, Keller J, Safwat A. Serum biomarkers as prognostic factors for metastatic sarcoma. *Clin Oncol.* 2019;31(4):242-249.
29. Pillozzi S, Bernini A, Palchetti I, et al. Soft tissue sarcoma: an insight on biomarkers at molecular, metabolic and cellular level. *Cancers.* 2021;13(12):3044.
30. Min L, Choy E, Tu C, Hornicek F, Duan Z. Application of metabolomics in sarcoma: From biomarkers to therapeutic targets. *Crit Rev Oncol Hematol.* 2017;116:1-10.
31. Dhillon A, Singh A, Bhalla VK. A systematic review on biomarker identification for cancer diagnosis and prognosis in multi-omics: from computational needs to machine learning and deep learning. *Arch Comput Methods Eng.* 2023;30(2):917-949.
32. Rahmati R, Zarimeidani F, Ahmadi F, et al. Identification of novel diagnostic and prognostic microRNAs in sarcoma on TCGA dataset: bioinformatics and machine learning approach. *Sci Rep.* 2025;15(1):7521.
33. Bostanci E, Kocak E, Unal M, Guzel MS, Acici K, Asuroglu T. Machine learning analysis of RNA-seq data for diagnostic and prognostic prediction of colon cancer. *Sensors.* 2023;23(6):3080.

34. He T, Huang L, Li J, Wang P, Zhang Z. Potential prognostic immune biomarkers of overall survival in ovarian cancer through comprehensive bioinformatics analysis: a novel artificial intelligence survival prediction system. *Front Med.* 2021;8:587496.
35. Alam MS, Sultana A, Reza MS, Amanullah M, Kabir SR, Mollah MNH. Integrated bioinformatics and statistical approaches to explore molecular biomarkers for breast cancer diagnosis, prognosis and therapies. *PloS One.* 2022;17(5):e0268967.
36. Hossain MA, Islam SMS, Quinn JM, Huq F, Moni MA. Machine learning and bioinformatics models to identify gene expression patterns of ovarian cancer associated with disease progression and mortality. *J Biomed Inform.* 2019;100:103313.
37. Higashino T, Matsuo H, Sakiyama M, et al. Common variant of PDZ domain containing 1 (PDZK1) gene is associated with gout susceptibility: a replication study and meta-analysis in Japanese population. *Drug Metab pharmacokinet.* 2016;31(6):464-466.
38. Sun H-l, Wu Y-w, Bian H-g, et al. Function of uric acid transporters and their inhibitors in hyperuricaemia. *Front Pharmacol.* 2021;12:667753.
39. Tao T, Yang X, Zheng J, et al. PDZK1 inhibits the development and progression of renal cell carcinoma by suppression of SHP-1 phosphorylation. *Oncogene.* 2017;36(44):6119-6131.
40. Gao H, Gao Z, Liu X, et al. miR-101-3p-mediated role of PDZK1 in hepatocellular carcinoma progression and the underlying PI3K/Akt signaling mechanism. *Cell Division.* 2024;19(1):9.
41. Amatori S, Tavolaro S, Gambardella S, Fanelli M. The dark side of histones: genomic organization and role of oncohistones in cancer. *Clin Epigenetics.* 2021;13(1):71.
42. Zhang Y, Zhang Q, Zhang Y, Han J. The role of histone modification in DNA replication-coupled nucleosome assembly and cancer. *Int J Mol Sci.* 2023;24(5):4939.
43. Wang X-W, Sun Q, Xu S-B, et al. A 3-DNA methylation signature as a novel prognostic biomarker in patients with sarcoma by bioinformatics analysis. *Medicine.* 2021;100(20):e26040.
44. Yang K, Mooney C, Spahlinger G, et al. DR6 as a diagnostic and predictive biomarker in adult sarcoma. *PloS One.* 2012;7(5):e36525.
45. Patton A, Dermawan JK. Current updates in sarcoma biomarker discovery: emphasis on next-generation sequencing-based methods. *Pathology.* 2024;56:274.
46. Zhu MM, Shenasa E, Nielsen TO. Sarcomas: Immune biomarker expression and checkpoint inhibitor trials. *Cancer Treat Rev.* 2020;91:102115.
47. Fattahi S, Kosari-Monfared M, Golpour M, et al. LncRNAs as potential diagnostic and prognostic biomarkers in gastric cancer: a novel approach to personalized medicine. *J Cell Physiol.* 2020;235(4):3189-3206.
48. Daher M, Zalaquett Z, Chalhoub R, et al. Molecular and biologic biomarkers of Ewing sarcoma: A systematic review. *J Bone Oncol.* 2023;40:100482.
49. Lou S, Balluff B, de Graaff MA, et al. High-grade sarcoma diagnosis and prognosis: Biomarker discovery by mass spectrometry imaging. *Proteomics.* 2016;16(11-12):1802-1813.
50. Liu Y, Yang J, Weng D, Xie Y. A1CF Binding to the p65 Interaction Site on NKRF Decreased IFN- β Expression and p65 Phosphorylation (Ser536) in Renal Carcinoma Cells. *Int J Mol Sci.* 2024;25(7):3576.
51. Song Y, Shao L, Xue Y, et al. Inhibition of the aberrant A1CF-FAM224A-miR-590-3p-ZNF143 positive feedback loop attenuated malignant biological behaviors of glioma cells. *J Exp Clin Cancer Res.* 2019;38(1):248. doi:[10.1186/s13046-019-1200-5](https://doi.org/10.1186/s13046-019-1200-5).
52. Xia H, Liu Y, Xu L, Mao H, Zhou Q, Xie Y. APOBEC1 complementation factor facilitates cell migration by promoting nucleus translocation of SMAD3 in renal cell carcinoma cells. *In Vitro Cell Dev Biol Anim.* 2021;57(5):501-509. doi:[10.1007/s11626-021-00589-z](https://doi.org/10.1007/s11626-021-00589-z).
53. Yan X, Li Q, Ni D, et al. Apobec-1 complementation factor regulates cell migration and apoptosis through Dickkopf1 by acting on its 3' untranslated region in MCF7 cells. *Tumour Biol.* 2017;39(6):1010428317706218. doi:[10.1177/1010428317706218](https://doi.org/10.1177/1010428317706218).
54. Ni D, Liu J, Hu Y, et al. A1CF-Axin2 signal axis regulates apoptosis and migration in Wilms tumor-derived cells through Wnt/ β -catenin pathway. *In Vitro Cell Dev Biol Anim.* 2019;55(4):252-259.
55. Wang L, Cheng Q. APOBEC-1 Complementation Factor: From RNA Binding to Cancer. *Cancer Control.* 2024;31:10732748241284952. doi:[10.1177/10732748241284952](https://doi.org/10.1177/10732748241284952).
56. Owe-Larsson M, Pawłasek J, Piecha T, Sztokfisz-Ignasiak A, Pater M, Janiuk IR. The role of cocaine-and amphetamine-regulated transcript (CART) in cancer: a systematic review. *Int J Mol Sci.* 2023;24(12):9986.
57. Vincent JB, Petek E, Thevarkunnel S, et al. The RAY1/ST7 tumor-suppressor locus on chromosome 7q31 represents a complex multi-transcript system. *Genomics.* 2002;80(3):283-294.
58. Wan Q, Tang J, Lu J, et al. Six-gene-based prognostic model predicts overall survival in patients with uveal melanoma. *Cancer Biomark.* 2020;27(3):343-356.
59. Jia B, Zhao X, Wang Y, Wang J, Wang Y, Yang Y. Prognostic roles of MAGE family members in breast cancer based on KM-Plotter Data. *Oncol Lett.* 2019;18(4):3501-3516.
60. Kang W, Wang Y, Li J, et al. TAS2R supports odontoblastic differentiation of human dental pulp stem cells in the inflammatory microenvironment. *Stem Cell Res Ther.* 2022;13(1):374.

61. Koyama T, Ogawara K, Kasamatsu A, et al. ANGPTL3 is a novel biomarker as it activates ERK/MAPK pathway in oral cancer. *Cancer Med.* 2015;4(5):759-769.
62. Li M, Liu G, Jin X, et al. Micropeptide MIAC inhibits the tumor progression by interacting with AQP2 and inhibiting EREG/EGFR signaling in renal cell carcinoma. *Mol Cancer.* 2022;21(1):181.
63. Feng L, Wang Y, Qin J, et al. UGT1A Gene Family Members Serve as Potential Targets and Prognostic Biomarkers for Pancreatic Cancer. *Biomed Res Int.* 2021;2021(1):6673125.
64. Hong K, Yang Q, Yin H, Wei N, Wang W, Yu B. Comprehensive analysis of ZNF family genes in prognosis, immunity, and treatment of esophageal cancer. *BMC cancer.* 2023;23(1):301.
65. Xie H, Wang W, Sun F, et al. Proteomics analysis to reveal biological pathways and predictive proteins in the survival of high-grade serous ovarian cancer. *Sci Rep.* 2017;7(1):9896.
66. Haring SJ, Humphreys TD, Wold MS. A naturally occurring human RPA subunit homolog does not support DNA replication or cell-cycle progression. *Nucleic Acids Res.* 2010;38(3):846-858.
67. Ramezannezhad P, Rahmati R, Zarimeidani F, et al. Personalized Patient Education: Patient's Perspective. *Iran Biomed J.* 2024;28(0):66. doi:[10.61186/ibj.25th-11th-IACRTIMSS](https://doi.org/10.61186/ibj.25th-11th-IACRTIMSS).

Appendix

List of Abbreviations

AI	Artificial Intelligence
AIC	Akaike Information Criterion
AUC	Area Under the Curve
DEGs	Differentially Expressed Genes
DL	Deep Learning
DT	Decision Tree
GDAC	Global Data Assembly Center
GLM	Generalized Linear Model
GO	Gene Ontology
HR	Hazard Ratio
KEGG	Kyoto Encyclopedia of Genes and Genomes
KNN	k-Nearest Neighbors
ML	Machine Learning
MSE	Mean Squared Error
NPV	Negative Predictive Value
PPI	Protein-Protein Interaction
PPV	Positive Predictive Value
ReLU	Rectified Linear Unit
RF	Random Forest
RNA-seq	Ribonucleic Acid Sequencing
ROC	Receiver Operating Characteristic
RMSE	Root Mean Squared Error
STS	Soft Tissue Sarcoma
SVM	Support Vector Machine
TCGA	The Cancer Genome Atlas.