Discover Oncology

ANALYSIS Open Access

Integrating transcriptomics and hybrid machine learning enables high-accuracy diagnostic modeling for nasopharyngeal carcinoma

Hehe Wang[1], Junge Zhang[2], Peng Cheng[1], Lujie Yu[1], Chunlin Li[1*] and Yaowen Wang[1*]

*Correspondence:
Chunlin Li
lichunlin202105@163.com
Yaowen Wang
wangyaowennihao@hotmail.com
[1]Department of Otolaryngology, Head and Neck Surgery, The First Affiliated Hospital of Ningbo University, Ningbo, China
[2]Department of Anesthesiology, The First Affiliated Hospital of Ningbo University, Ningbo, China

Abstract

Background Nasopharyngeal carcinoma (NPC) lacks biomarkers demonstrating both high specificity and sensitivity for early diagnosis. This study aimed to develop robust machine learning (ML)-driven diagnostic models and identify key biomarkers through integrated analysis of multi-cohort transcriptomic data.

Methods Seven NPC transcriptomic datasets (GSE12452, GSE40290, GSE53819, and GSE64634 were merged to form the training cohort, while GSE13597, GSE34573, and GSE61218 served as independent external validation sets) were integrated and preprocessed using ComBat for batch effect correction. Differential expression analysis identified 293 differentially expressed genes (DEGs). Twelve ML algorithms (including Stepglm, glmBoost, and RF) were systematically combined into 113 distinct models to classify NPC versus normal tissues. Top-performing models underwent external validation. Immune infiltration patterns and functional enrichment were analyzed using CIBERSORT and GSEA/GSVA, respectively.

Results The Stepglm[both]-RF hybrid model demonstrated exceptional performance with AUCs of 0.999 (training set; 95% CI: 0.997–1.000), 1.000 (GSE61218/GSE34573 validation), and 0.960 (GSE13597 validation). The glmBoost-RF model showed comparable efficacy, achieving AUCs of 1.000 (training), 0.950 (GSE61218), 1.000 (GSE34573), and 0.947 (GSE13597). Single-gene analysis identified RCN1 as a promising diagnostic marker (AUC = 0.953), with elevated expression levels correlating with poor prognosis in head and neck squamous cell carcinoma (HNSCC; $p < 0.05$). Immune profiling revealed significant enrichment of M1 macrophages and concomitant reduction of memory B cells in NPC. Functional enrichment analysis associated RCN1 with cell cycle regulation and immune-related pathways.

Conclusion This study establishes two high-performance ML models (Stepglm[both]-RF and glmBoost-RF) with low variability for NPC diagnosis and identifies RCN1 as a dual-function biomarker with diagnostic and prognostic potential. The findings provide a scalable framework for early NPC detection and novel insights into immune microenvironment dysregulation.

Keywords Nasopharyngeal carcinoma, Machine learning, Diagnostic model, Biomarker discovery, RCN1

1 Background

Nasopharyngeal carcinoma (NPC) demonstrates marked ethnic and geographic disparities, exhibiting disproportionately high incidence rates in Southeast Asia, Southern China, and North Africa [1]. The disease etiology involves a well-characterized triad of Epstein-Barr virus (EBV) infection, genetic predisposition, and environmental carcinogen exposure [2]. Although advancements in radiotherapy and chemoradiotherapy have improved clinical outcomes, locally advanced NPC continues to present therapeutic challenges, with 5-year overall survival rates remaining suboptimal at approximately 60% [3, 4]. This persistent clinical challenge highlights the urgent requirement for enhanced early detection strategies to improve patient prognosis. While EBV DNA load quantification provides valuable risk stratification for NPC patients, current diagnostic approaches lack biomarkers demonstrating sufficient specificity and sensitivity for reliable early detection [5]. This diagnostic limitation has motivated extensive research efforts to identify molecular signatures capable of improving diagnostic accuracy and prognostic assessment.

Transcriptomic profiling has significantly enhanced our ability to characterize tumor-specific molecular alterations. Large-scale genomic repositories such as the Gene Expression Omnibus (GEO) provide valuable resources for investigating NPC heterogeneity through multi-cohort analyses [6]. However, technical challenges including batch effects and platform-specific variability complicate cross-dataset integration, necessitating sophisticated normalization pipelines and batch correction methodologies [7]. The inherent complexity of oncogenic processes further requires advanced computational approaches, particularly machine learning (ML) algorithms, to extract biologically relevant patterns from high-dimensional omics data. Previous investigations into NPC pathogenesis have primarily focused on discrete molecular pathways, creating a significant knowledge gap in systematically integrated frameworks that combine differential expression analysis, functional pathway enrichment, and ML-driven classification across diverse patient cohorts. Our study addresses this critical need through comprehensive analysis of seven NPC transcriptomic datasets, pursuing three principal objectives: (1) identification of robust diagnostic biomarkers, (2) development of optimized ML classification models, and (3) characterization of tumor-immune microenvironment interactions, ultimately establishing a multimodal molecular framework for NPC diagnosis and mechanistic elucidation.

To address the methodological and conceptual gaps in existing NPC research, this study integrates seven multi-cohort transcriptomic datasets (GSE12452, GSE40290, GSE53819, GSE64634 for training; GSE13597, GSE34573, GSE61218 for validation) using ComBat batch correction and hybrid machine learning (ML) frameworks. We systematically combined 12 ML algorithms into 113 combinatorial models, prioritizing parsimonious gene panels (< 10 genes) without compromising diagnostic accuracy (AUC ≥ 0.95). Differential expression analysis identified 293 high-confidence DEGs, with functional enrichment linking these genes to immune dysregulation and stromal remodeling. Through systematic evaluation, two optimized models were selected. Crucially, single-gene validation pinpointed RCN1 as a dual-function biomarker, demonstrating both diagnostic utility (AUC = 0.953) and prognostic relevance in head and neck malignancies. This integrative approach not only establishes robust ML models for NPC

detection but also uncovers novel molecular-immune interactions, providing a scalable framework for early diagnosis and mechanistic exploration.

2 Methods

2.1 Data acquisition and preprocessing

We obtained seven nasopharyngeal carcinoma (NPC) transcriptomic datasets from the Gene Expression Omnibus (GEO) database: GSE12452, GSE13597, GSE34573, GSE40290, GSE53819, GSE61218, and GSE64634 [8–16]. These datasets collectively contained gene expression profiles from both normal nasopharyngeal tissues and NPC biopsies. Raw data were converted into gene expression matrices using platform-specific preprocessing pipelines as described in the original GEO metadata. For model construction and validation, GSE12452, GSE40290, GSE53819, and GSE64634 were combined to create the training cohort, while GSE13597, GSE34573, and GSE61218 were designated as independent external validation sets (Table 1).

To address technical heterogeneity across studies, we implemented a two-step harmonization protocol. First, the integrated training cohort underwent ComBat adjustment using study identifiers as batch variables through the R sva package (version 3.48.0). Subsequently, the effectiveness of batch effect correction was systematically evaluated through distributional alignment analysis (comparative visualization of pre- and post-correction expression distributions using boxplots) and dimensionality assessment (principal component analysis [PCA] with comparative evaluation of sample clustering patterns before and after correction).

2.2 Differential expression analysis

Differentially expressed genes (DEGs) between NPC and normal tissues were identified using the limma package (v3.56.2) in R [17], implementing a linear modeling framework with empirical Bayes moderation to address cross-gene variance heterogeneity. Genes were filtered using stringent thresholds of absolute log2-fold change ($|\log2FC| > 1$ and Benjamini-Hochberg adjusted p-value (FDR) < 0.05 to ensure biological relevance while controlling false discovery rates. Significant DEGs were visualized through volcano plots (ggplot2 v3.4.2) that highlighted upregulated ($\log2FC > 1$) and downregulated ($\log2FC < -1$) genes using color-coded thresholds. To validate expression patterns, hierarchical clustering was performed on top-ranked DEGs ($|\log2FC| > 2$, FDR $< 1e-5$) using Euclidean distance and Ward's linkage, with results displayed as a z-score-normalized heatmap (pheatmap v1.0.12). All statistically significant DEGs meeting the criteria were systematically exported for subsequent analyses, ensuring reproducibility across downstream workflows.

Table 1 Overview of NPC-Related gene expression datasets

GEO series	Expression type	Platform	Sample number		Type	Reference
			Normal	Tumor		
GSE12452	mRNA	GPL570	10	31	Training set	8–10
GSE13597	mRNA	GPL96	3	25	Validation cohort	11
GSE34573	mRNA	GPL570	4	16	Validation cohort	12
GSE40290	mRNA	GPL8380	8	25	Training set	/
GSE53819	mRNA	GPL6480	18	18	Training set	13
GSE61218	mRNA	GPL19061	6	10	Validation cohort	14–15
GSE64634	mRNA	GPL570	4	12	Training set	16

2.3 GO and KEGG pathway enrichment analysis

To functionally characterize the biological implications of NPC-associated DEGs, we performed Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses using the clusterProfiler package (v4.8.1) [18]. These analyses systematically interrogated three GO domains: biological processes (BP), cellular components (CC), and molecular functions (MF), while KEGG analysis delineated pathway-level dysregulation. Enrichment significance was assessed via hypergeometric testing with Benjamini-Hochberg false discovery rate (FDR) correction (adjusted $p < 0.05$), retaining terms exhibiting gene ratio > 0.1 and minimum gene count ≥ 5 to ensure biological interpretability. Results were visualized as dot plots displaying enrichment scores (-log10[FDR]), gene ratios, and associated gene counts, enabling prioritized identification of NPC-relevant mechanisms.

2.4 Machine learning model development and evaluation

We established a diagnostic framework for NPC through systematic integration of 12 machine learning algorithms into 113 combinatorial models, utilizing differentially expressed genes (DEGs) as input features. To address overfitting and improve model generalizability, we employed regularized regression methods for feature selection, including Elastic Net (Enet) [19], Ridge Regression [20], Stepwise Generalized Linear Model (StepGLM), and LASSO [21]. The classification workflow incorporated diverse methodological approaches spanning Support Vector Machines (SVM), Linear Discriminant Analysis (LDA), generalized linear model boosting (glmBoost), Random Forest, Gradient Boosting Machine (GBM), XGBoost, and Naive Bayes classifiers [22–26]. Prior to model training, all features underwent z-score normalization to ensure scale uniformity. We implemented a two-tier validation strategy comprising internal validation with 10-fold cross-validation incorporating nested feature selection, coupled with external validation across three independent transcriptomic cohorts. Model performance was comprehensively assessed using the area under the receiver operating characteristic curve (AUC), with comparative results visualized through precision-ranked heatmap analysis. Models demonstrating superior discriminative capacityn were selected for subsequent clinical translatability evaluation.

2.5 Model evaluation using ROC and confusion matrix

The optimal model demonstrating superior predictive performance (highest AUC) underwent comprehensive evaluation across both training and external validation cohorts. Systematic generation of receiver operating characteristic (ROC) curves enabled quantitative assessment of discriminative capacity between NPC and normal samples, with area under the curve (AUC) values calculated for each dataset. We employed bootstrap resampling to derive 95% confidence intervals (CIs) for AUC stability evaluation. Complementary confusion matrix analysis [27] provided detailed performance metrics through direct comparison of predicted versus actual classifications, including sensitivity (true positive rate), specificity (true negative rate), balanced accuracy, and positive/negative predictive values. This multi-dimensional evaluation confirmed robust model performance across sensitivity, specificity, and overall accuracy metrics, with systematic stratification by dataset type (training versus validation) demonstrating consistent generalizability.

2.6 Gene interaction network analysis

We constructed gene interaction networks using GeneMANIA [28] to elucidate functional relationships among prioritized genes, utilizing the human reference database. The analytical framework incorporated multimodal biological evidence encompassing experimentally validated physical interactions, co-expression patterns, and curated pathway associations. Network topology analysis identified high-confidence interactors through weighted edge scoring, followed by functional enrichment analysis focusing on kinase-related processes such as negative regulation of kinase activity and protein phosphorylation dynamics. The resulting network architecture provided mechanistic insights into selected genes' roles in NPC pathogenesis, with hub nodes identified as priority targets for experimental validation.

2.7 Immune infiltration analysis

We performed immune cell infiltration profiling using the CIBERSORT algorithm [29], which applies linear support vector regression to deconvolute bulk transcriptomic data and estimate relative proportions of 22 immune cell subtypes. Gene expression data from selected samples were analyzed against the LM22 leukocyte signature matrix with 1,000 permutations to enhance reliability. Results were filtered using permutation-derived p-values, Pearson correlation coefficients for feature stability, and root mean squared error (RMSE) metrics for deconvolution accuracy. Samples meeting stringent quality thresholds ($p < 0.05$ and RMSE < 0.15) were retained for subsequent analysis to ensure biological plausibility of immune cell fractions. This methodology enabled systematic characterization of tumor microenvironment composition across NPC and normal tissue cohorts.

2.8 Enrichment and prognostic analysis

Gene Set Enrichment Analysis (GSEA) was conducted using the Molecular Signatures Database (MSigDB) C5 (GO: gene ontology) and C2 (KEGG: Kyoto Encyclopedia of Genes and Genomes) gene sets (c5.go.v2023.1.Hs.symbols.gmt, c2.cp.kegg.v2023.1.Hs.symbols.gmt) [30] to identify pathways associated with target gene differential expression. Following z-score normalization of expression data, samples were stratified into high/low expression groups based on median target gene expression. Genes were ranked by log2-fold change (log2FC) between groups, with enrichment significance determined through 1,000 permutations (nominal $p < 0.05$, false discovery rate [FDR] < 0.25). Pathway activity quantification was performed using Gene Set Variation Analysis (GSVA) via single-sample GSEA (ssGSEA) with identical gene sets. After batch correction and control sample exclusion, normalized GSVA scores were compared between expression groups using Welch's t-tests, categorizing pathways as upregulated (FDR < 0.05, log2FC > 0), downregulated (FDR < 0.05, log2FC < 0), or nonsignificant. Results were visualized through ranked bar plots highlighting pathway dysregulation patterns. Prognostic validation was conducted through the UALCAN platform [31], integrating TCGA survival data with RNA-seq profiles to evaluate diagnostic efficiency and survival outcomes (log-rank $p < 0.05$) for prioritized genes.

2.9 Statistical analysis

All statistical analyses were performed using R software (version 4.3.1). Differential gene expression analysis employed the limma package (v3.56.2) to construct linear models with empirical Bayes moderation, applying significance thresholds of absolute log2-fold change ($|\log2FC| > 1$ and Benjamini-Hochberg adjusted p-value (FDR) < 0.05. Machine learning model evaluation incorporated the pROC (v1.18.4), caret (v6.0.94), and e1071 (v1.7.13) packages, with classifier performance quantified through receiver operating characteristic (ROC) curve analysis (AUC calculations) and confusion matrices reporting sensitivity/specificity metrics. Gene Set Enrichment Analysis (GSEA) was executed using clusterProfiler (v4.8.1) and enrichplot (v1.20.0) with predefined gene sets, retaining pathways demonstrating nominal significance ($p < 0.05$). Gene Set Variation Analysis (GSVA) implemented the GSVA (v1.48.3) and limma packages to calculate single-sample pathway activity scores, with differential pathway activities across groups assessed via Welch's t-tests (significance threshold $p < 0.05$). All data visualizations were created using ggplot2 (v3.4.2) and ggpubr (v0.6.0) packages to ensure standardized graphical outputs.

3 Results

3.1 Differential expression and batch effect correction

Integrated analysis of the training datasets (GSE12452, GSE40290, GSE53819, GSE64634) revealed substantial technical variability necessitating batch correction. Initial preprocessing assessment identified dataset-specific expression distribution biases through boxplot visualization (Fig. 1A), supported by principal component analysis (PCA) demonstrating platform-driven sample clustering patterns (Fig. 1C). Application of the ComBat algorithm (sva package v3.48.0) effectively mitigated cross-platform variation, producing normalized expression distributions (Fig. 1B) and biologically meaningful PCA clustering (Fig. 1D). Subsequent differential expression analysis identified 293 high-confidence DEGs ($|\log2FC| > 1$; Benjamini-Hochberg FDR-adjusted $p < 0.05$), with hierarchical clustering analysis showing clear segregation between NPC and normal tissues (Fig. 1E). Expression patterns were further characterized using a color-graded volcano plot (Fig. 1F), where gradient transitions from blue (downregulated genes) to red (upregulated genes) encoded both log2-fold change magnitudes and statistical significance.

3.2 Functional enrichment analysis of the DEGs

Gene Ontology (GO) analysis revealed significant enrichment of these differentially expressed genes in biological processes including humoral immune response and leukocyte migration, cellular components such as collagen-containing extracellular matrix, and molecular functions related to receptor-ligand activity (Fig. 2A). KEGG pathway analysis further demonstrated their involvement in cytokine-cytokine receptor interaction, amoebiasis pathogenesis, chemokine signaling, and muscle cell cytoskeletal regulation (Fig. 2B), collectively highlighting the multi-faceted biological relevance of these transcriptional alterations in NPC progression.

3.3 Evaluation of top-performing machine learning models

A cohort of 113 machine learning models was systematically developed and validated using integrated training datasets (GSE12452, GSE40290, GSE53819, GSE64634) and

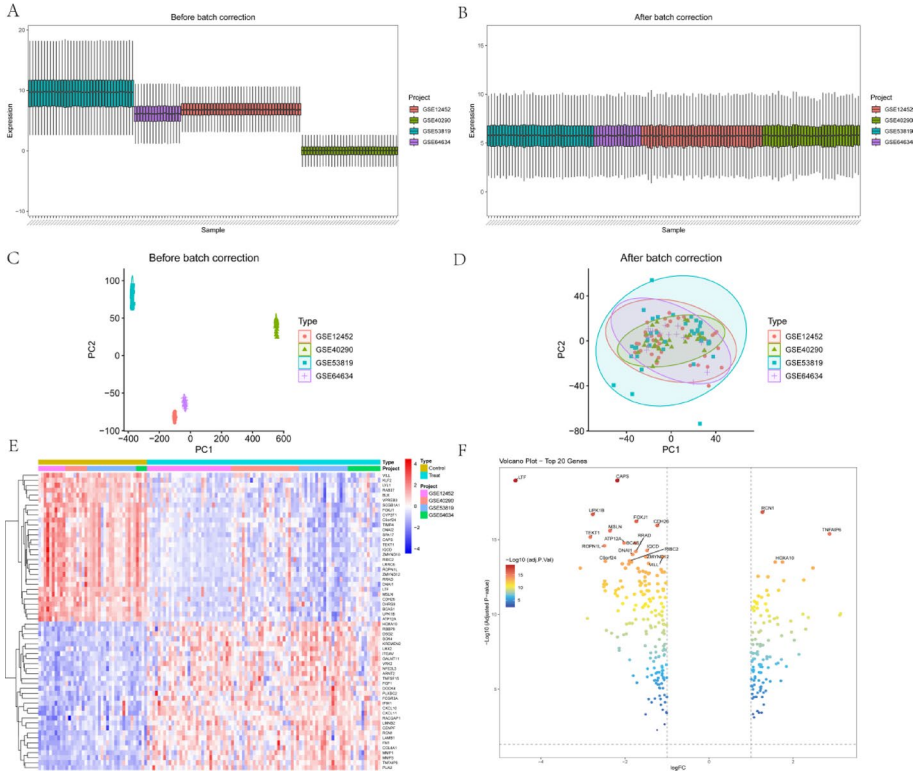

Fig. 1 Batch Effect Correction and Differential Gene Expression Analysis. **A** Pre-correction boxplots showing inter-dataset variability across GSE12452, GSE40290, GSE53819, and GSE64634. **B** Post-ComBat harmonization demonstrating aligned expression distributions. **C** PCA plot revealing batch-driven clustering prior to correction. **D** PCA plot after batch correction, illustrating improved alignment and reduced batch effects. **E** Heatmap of top DEGs (rows) distinguishing NPC (columns) from normal tissues. **F** Gradient volcano plot depicting expression dynamics: leftward blue gradient (downregulated genes, log2FC < -1) to rightward red gradient (upregulated genes, log2FC > 1), with significance threshold (FDR < 0.05)

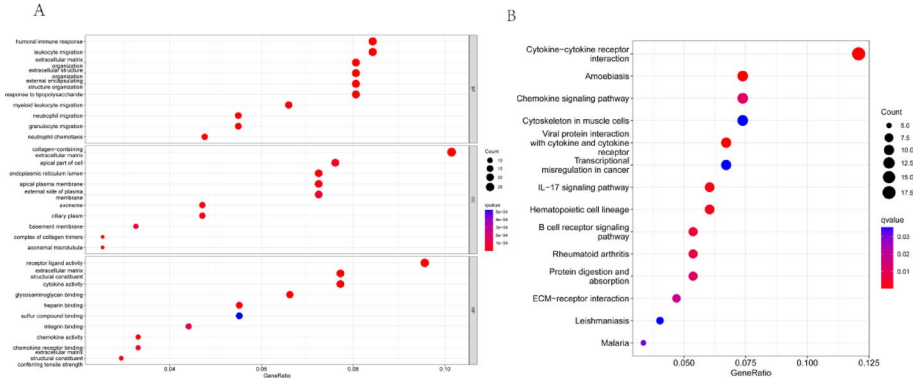

Fig. 2 Functional Enrichment Analysis. **A** GO enrichment of consensus DEGs, highlighting humoral immune response, leukocyte migration, collagen-containing extracellular matrix, and receptor-ligand activity. **B** KEGG pathway analysis revealing cytokine-cytokine receptor interaction, amoebiasis pathogenesis, chemokine signaling, and muscle cell cytoskeleton regulation

external validation cohorts (GSE13597, GSE34573, GSE61218) (Fig. 3A). All models exhibited strong discriminatory capacity in distinguishing NPC from normal tissues. To enhance clinical translation potential, model selection prioritized parsimonious feature sets without compromising diagnostic precision. For example, while the RF-plsRglm

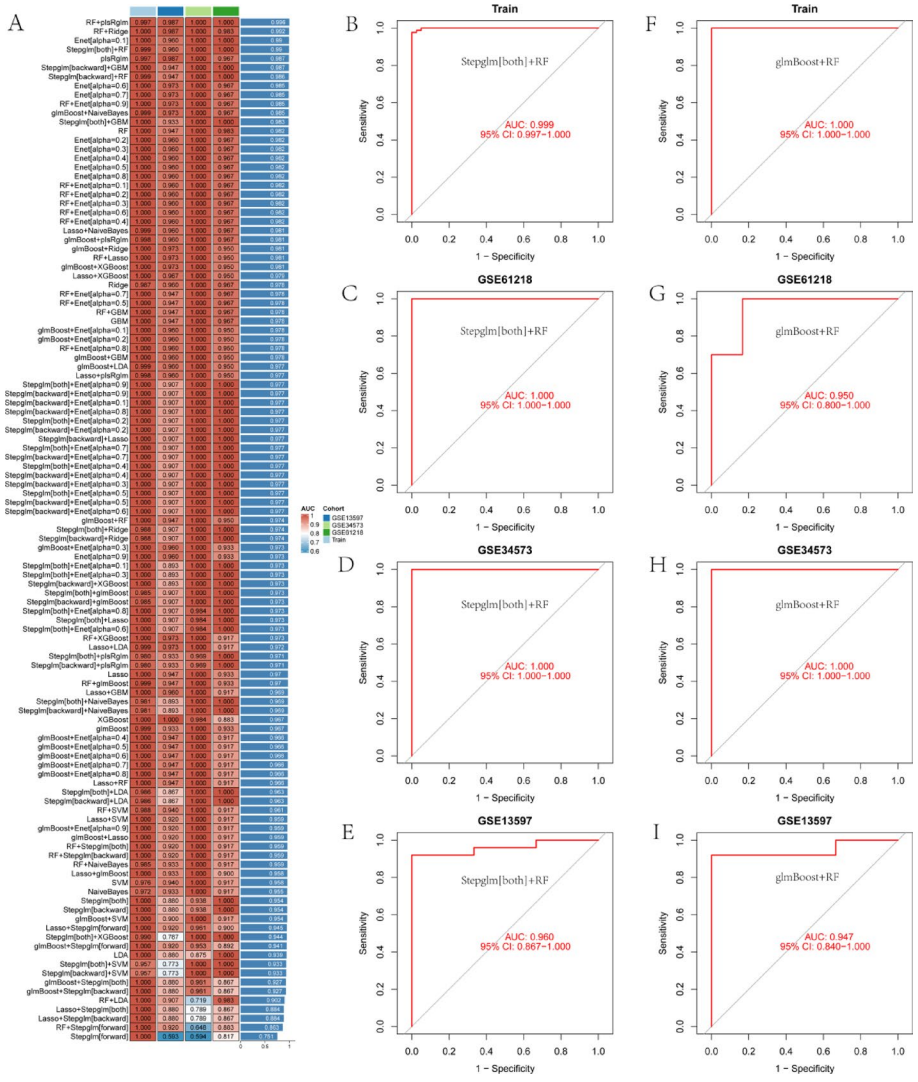

Fig. 3 Machine Learning Model Evaluation for NPC Classification. (B-E) Stepglm[both]-RF Model Performance. (F-I) glmBoost-RF Model Performance. **A** Heatmap of AUC performance for 113 models across training (GSE12452/GSE40290/GSE53819/GSE64634) and validation cohorts (GSE13597/GSE34573/GSE61218). **B** Training set: AUC = 0.999 (95%CI: 0.997-1.000). **C** GSE61218 validation: AUC = 1.000 (1.000–1.000). **D** GSE34573 validation: AUC = 1.000 (1.000–1.000). **E** GSE13597 validation: AUC = 0.960 (0.887-1.000). **F** Training set: AUC = 1.000 (1.000–1.000). **G** GSE61218 validation: AUC = 0.950 (0.800-1.000). **H** GSE34573 validation: AUC = 1.000 (1.000–1.000). **I** GSE13597 validation: AUC = 0.947 (0.840-1.000)

model achieved high accuracy, its requirement of 94 features limited practicality (Supplementary Document 1). Through systematic evaluation, two optimized models were selected, with comprehensive performance metrics detailed in Table 2 and visualized in Fig. 3B-I.

3.4 Model validation through confusion matrix analysis

Confusion matrices were constructed to quantify classification accuracy across training and validation cohorts (Fig. 4A-H). For the Stepglm[both]-RF model, the training set (Fig. 4A) correctly classified 40 controls and 86 NPC samples, with two controls misclassified as NPC. External validation demonstrated consistent performance: GSE61218 (Fig. 4B) and GSE34573 (Fig. 4C) achieved perfect classification, while GSE13597

Table 2 Diagnostic performance of Stepglm[both]-RF and glmBoost-RF models

Model	Stepglm[both]-RF	glmBoost-RF
Feature Genes	CXCL10, ARNT2, DHRS9, DLEC1, KCNE1	RCN1, LTF, UPK1B, FOXJ1, LYL1, BLK, TNFSF15
Training Set	AUC = 0.999 (0.997–1.000) [Fig. 3B]	AUC = 1.000 (1.000–1.000) [Fig. 3F]
External validation		
GSE61218	AUC = 1.000 (1.000–1.000) [Fig. 3C]	AUC = 0.950 (0.800–1.000) [Fig. 3G]
GSE34573	AUC = 1.000 (1.000–1.000) [Fig. 3D]	AUC = 1.000 (1.000–1.000) [Fig. 3H]
GSE13597	AUC = 0.960 (0.887–1.000) [Fig. 3E]	AUC = 0.947 (0.840–1.000) [Fig. 3I]

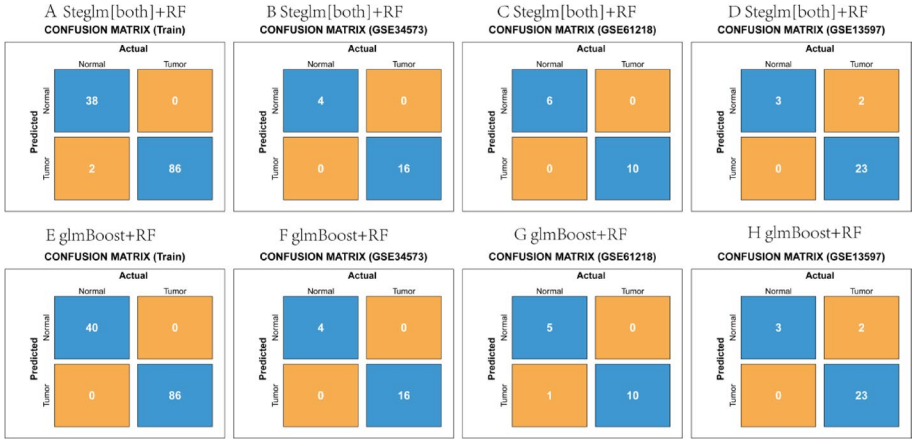

Fig. 4 Confusion Matrix Analysis of Diagnostic Models (**A**) Training Set (Stepglm[both]-RF): 40 controls and 86 NPCs correctly classified; 2 controls misclassified as NPC. **B** GSE61218 (Stepglm[both]-RF): Flawless NPC/control classification. **C** GSE34573 (Stepglm[both]-RF): Perfect accuracy (0 misclassifications). **D** GSE13597 (Stepglm[both]-RF): All controls correct; 2 NPCs misclassified. **E** Training Set (glmBoost-RF): 40 controls and 86 NPCs correctly classified (0 NPC errors). **F** GSE61218 (glmBoost-RF): 1 control misclassified as NPC; all NPCs correct. **G** GSE34573 (glmBoost-RF): Perfect classification. **H** GSE13597 (glmBoost-RF): All controls correct; 2 NPCs misclassified

(Fig. 4D) misclassified two NPC samples as controls. The glmBoost-RF model showed equivalent training set accuracy (Fig. 4E: 40 controls, 86 NPCs), with validation results indicating: one control misclassified in GSE61218 (Fig. 4F), perfect classification in GSE34573 (Fig. 4G), and two NPC misclassifications in GSE13597 (Fig. 4H). Both models maintained robust diagnostic performance across all datasets, as evidenced by ROC curve analysis and confusion matrices, confirming high accuracy and generalizability for NPC detection.

3.5 Identification and validation of key prognostic markers using ridge regression model

The top-performing models (Stepglm[both]-RF and glmBoost-RF) were further applied to identify key prognostic markers in NPC. Differential expression analysis identified 286 DEGs, from which 12 genes were selected for diagnostic model construction, as visualized in volcano plots (Fig. 5A and B). Comparative expression analysis of these 12 genes between NPC and control tissues revealed distinct differential patterns (Fig. 5C), with all genes showing statistically significant differences ($p < 0.001$). Correlation analysis (Fig. 5D) identified notable associations, including a strong positive correlation between LYL1 and BLK. ROC curve evaluation (Fig. 5E) demonstrated diagnostic potential for all markers, with RCN1 achieving the highest AUC (0.953). GeneMANIA network

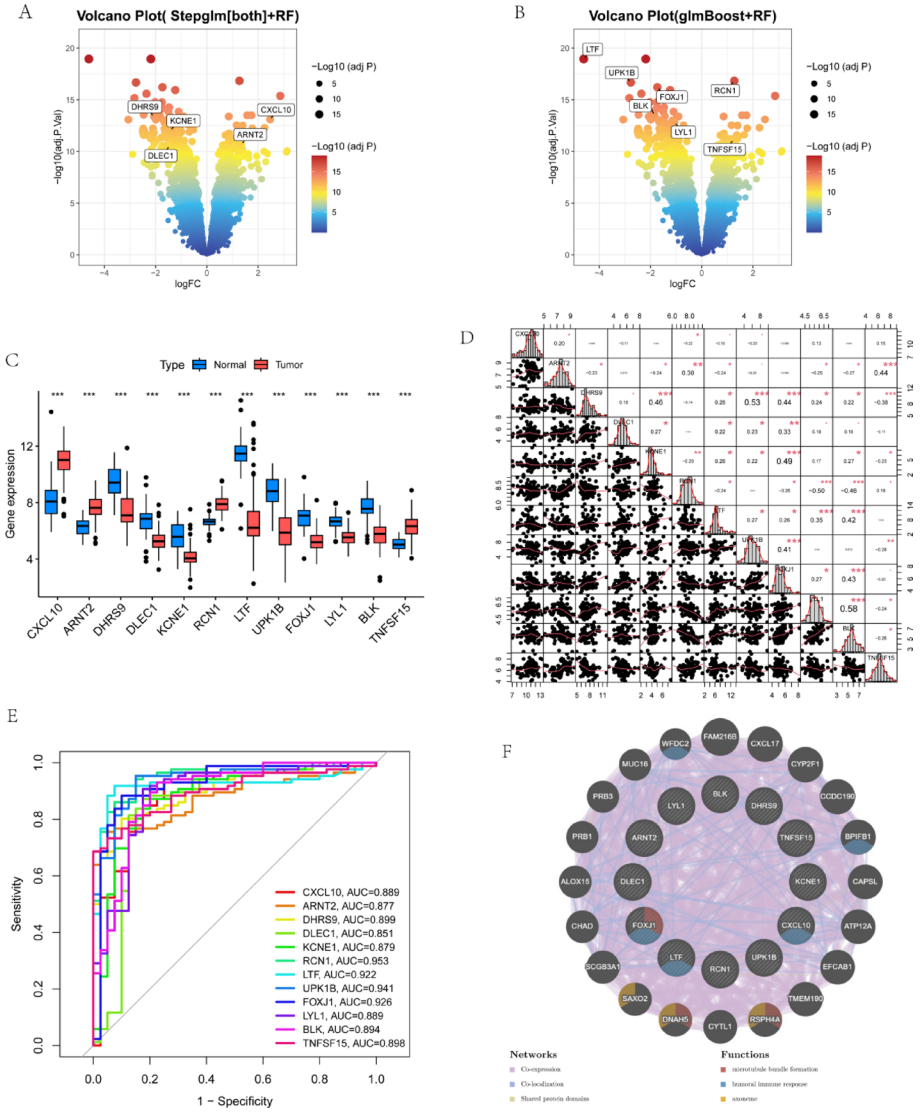

Fig. 5 Identification and Validation of Key NPC-Associated Genes. **A** Gradient volcano plot of genes used in the Stepglm[both]-RF model: DHRS9, DLEC1, KCNE1 (downregulated); CXCL10, ARNT2 (upregulated). **B** Gradient volcano plot for glmBoost-RF model genes: LTF, UPK1B, FOXJ1, LYL1, BLK (downregulated); RCN1, TNFSF15 (upregulated). **C** Expression profiles of selected genes in NPC vs. controls (***$p < 0.001$, **$p < 0.01$). **D** Correlation matrix depicting pairwise correlations (coefficients and significance) among genes. **E** ROC curves showing diagnostic potential (AUC values), with RCN1 achieving peak performance (AUC = 0.953). **F** GeneMANIA interaction network implicating genes in microtubule bundle formation and humoral immunity via physical interactions, co-expression, and shared pathways

analysis (Fig. 5F) revealed functional interactions involving microtubule bundle formation and humoral immune response, supported by physical interactions, co-expression patterns, and shared pathways among the genes, confirming their collective role in NPC pathogenesis.

3.6 Immune infiltration analysis in NPC

The relative proportions of 22 immune cell types in NPC and control tissues were quantified using the CIBERSORT algorithm. Figure 6A displays the differential distribution of immune cell subtypes between groups, revealing distinct compositional shifts in NPC samples characterized by elevated M0/M1 macrophage infiltration and reduced memory

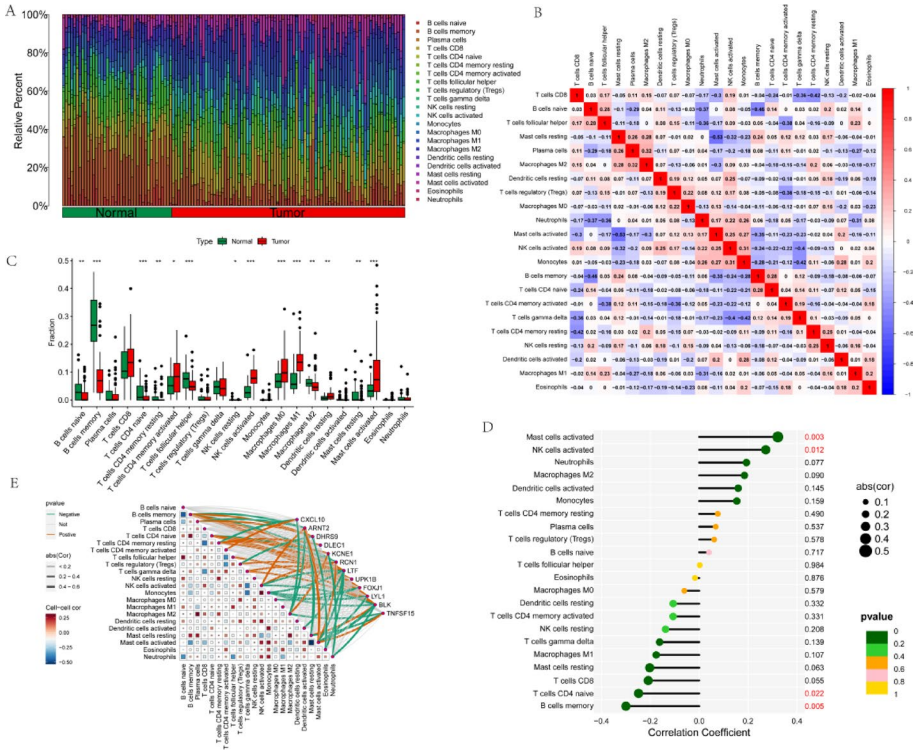

Fig. 6 Immune Infiltration Analysis in NPC. **A** Bar plot comparing relative proportions of immune cell types between NPC (treatment) and control groups using CIBERSORT analysis, revealing distinct immune landscape alterations in NPC. **B** T Heatmap displaying correlation coefficients among immune cell populations, demonstrating both positive associations and mutual exclusivity patterns. **C** Box plots showing significant differences in immune cell infiltration levels between groups, particularly in M1 macrophages and memory B cells. **D** Lollipop plot ranking immune cell types by their correlation with key NPC gene expression, identifying mast cells (activated), NK cells (activated), memory B cells, and naive CD4 + T cells as having strongest associations. **E** Cell-cell interaction network mapping significant positive/negative associations between key genes and immune cell populations, revealing complex immune microenvironment dynamics in NPC

B cell populations. Intercellular correlation analysis (Fig. 6B) identified significant negative associations between activated and resting mast cells, as well as between memory B cells and naïve B cells, suggesting dynamic regulatory interplay within the tumor microenvironment.

Comparative quantification of immune cell fractions (Fig. 6C) confirmed pronounced increases in M1 macrophages and decreases in memory B cells in NPC specimens. Gene-immune correlation analysis (Fig. 6D) revealed strong associations between activated mast cells, NK cells, memory B cells, and naïve CD4 + T cells with specific transcriptional profiles. Notably, integrative network analysis (Fig. 6E) demonstrated significant co-variation between key genes (RCN1, UPK1B, FOXJ1, LTF) and immune cell infiltration, with RCN1/ARNT2 expression showing positive correlations with activated NK cells. These findings delineate an immunosuppressive microenvironment in NPC, characterized by coordinated interactions between dysregulated immune cell populations and tumor-associated gene expression patterns.

3.7 Functional enrichment analysis of RCN1 and prognosis

To investigate the functional implications of RCN1 expression in NPC, we performed Gene Set Enrichment Analysis (GSEA) and Gene Set Variation Analysis (GSVA) on

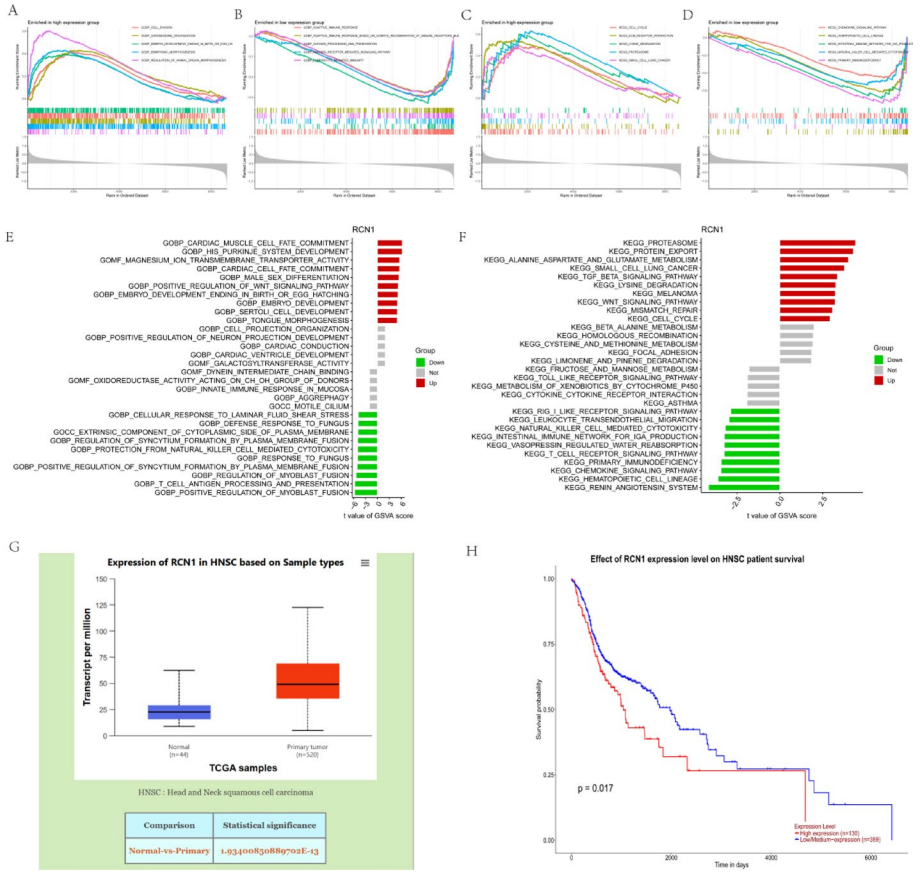

Fig. 7 Gene Set Enrichment Analysis (GSEA) and Gene Set Variation Analysis (GSVA) for RCN1 in NPC. **A** GSEA of high-RCN1-expressing samples showing significant enrichment in mucin transport, cell division, chromosome organization, and embryonic morphogenesis. Enrichment plots demonstrate strong pathway associations. **B** GSEA of low-RCN1-expressing samples revealing enrichment in adaptive immune response and antigen processing/presentation, suggesting tumor microenvironment modulation. **C** KEGG analysis of high-RCN1 group highlighting cell cycle, ECM-receptor interaction, and lysine degradation pathways. **D** KEGG analysis of low-RCN1 group identifying chemokine signaling, hematopoietic cell lineage, and intestinal immune network for IgA production. **E** GSVA heatmap displaying RCN1-associated pathway activity differences across samples. **F** GSVA t-value plot categorizing upregulated/downregulated pathways, emphasizing RCN1's regulatory role in NPC progression. **G** Significant RCN1 upregulation in HNSCC versus normal tissues ($p < 0.01$). **H** Kaplan-Meier curve showing reduced overall survival in high-RCN1 patients ($p < 0.05$)

pathways associated with differential RCN1 expression. Significant enrichment was observed in both high- and low-RCN1 expression groups. GSEA results for GO terms (Fig. 7A, B) revealed that the high-expression group showed enrichment in cell division, chromosome organization, and embryonic morphogenesis (Fig. 7A), while the low-expression group exhibited enrichment in adaptive immune response and antigen processing and presentation (Fig. 7B). KEGG pathway analysis further confirmed these associations: high RCN1 expression correlated with cell cycle, ECM-receptor interaction, and lysine degradation (Fig. 7C), whereas low expression was linked to chemokine signaling pathway, hematopoietic cell lineage, and intestinal immune network for IgA production (Fig. 7D). GSVA results (Fig. 7E, F) revealed distinct pathway activity differences between the groups. The t-value bar plots demonstrated RCN1's significant influence on biological processes, particularly cardiac muscle cell fate commitment, proteasome function, and protein export. These findings collectively indicate that RCN1

modulates key pathways in NPC, with differential expression driving distinct tumor biological outcomes.

To assess RCN1's clinical relevance in head and neck squamous cell carcinoma (HNSCC), we analyzed TCGA data via UALCAN. RCN1 expression was significantly elevated in HNSCC tissues compared to normal controls ($p < 0.01$; Fig. 7G). Kaplan-Meier analysis showed significantly shorter overall survival (OS) in patients with high RCN1 expression ($p < 0.05$; Fig. 7H), suggesting poor prognosis associated with elevated RCN1 levels.

4 Discussion

This investigation systematically integrates multi-cohort transcriptomic data to identify robust diagnostic biomarkers for nasopharyngeal carcinoma (NPC) while elucidating molecular-immune microenvironment interactions. Through advanced machine learning (ML) analysis of seven independent GEO datasets, we identified two gene panels demonstrating high diagnostic accuracy in distinguishing NPC from normal tissues. The Stepglm[both]-RF and glmBoost-RF models achieved demonstrated high accuracy (AUC: 0.999-1.000) across training and validation cohorts, surpassing conventional EBV DNA biomarkers in specificity. Characterized by minimal feature requirements and strong generalizability, these models propose a paradigm shift in NPC diagnostics, enabling cost-effective early detection in high-risk populations.

CIBERSORT-based immune profiling revealed an NPC microenvironment dominated by M1 macrophages with concomitant depletion of memory B cell populations. This immunosuppressive signature aligns with established mechanisms of EBV-associated immune evasion, including impaired antigen presentation and T cell exhaustion [32–34]. The observed M1 macrophage predominance correlates with chronic inflammatory states in viral oncogenesis [35]– [36], while memory B cell depletion mirrors patterns in head and neck squamous cell carcinoma, suggesting conserved immune escape strategies. Notably, the inverse correlation between mast cell activation states (Fig. 6D) indicates dynamic stromal remodeling requiring mechanistic exploration. The identified gene-immune correlations—particularly RCN1/UPK1B associations with activated NK cells—suggest transcriptional regulation of immune cell trafficking. RCN1's dual association with extracellular matrix remodeling and lysine degradation pathways implicates its role in stromal barrier formation, potentially mediating immune exclusion. Its prognostic significance in HNSCC further positions RCN1 as a therapeutic target with dual diagnostic utility.

RCN1 is a calcium-binding protein that resides within the lumen of the endoplasmic reticulum (ER) and plays a significant role in maintaining intracellular calcium homeostasis. Its involvement in various cellular processes, including cell proliferation and apoptosis, makes it a critical factor in the pathogenesis of several cancers, including NPC. In NPC, RCN1 is significantly upregulated, which correlates with tumor progression and poor prognosis. This upregulation is associated with the modulation of calcium signaling pathways that are crucial for the survival and proliferation of cancer cells [37]–[38]. RCN1's role in NPC is further highlighted by its ability to interact with other proteins involved in calcium homeostasis and ER stress response. For instance, RCN1 has been shown to interact with inositol 1,4,5-trisphosphate receptor type 1 (IP3R1), thereby influencing calcium release from the ER. This interaction is crucial for the regulation of

intracellular calcium levels, which, when dysregulated, can lead to enhanced cell survival and resistance to apoptosis, contributing to the malignancy of NPC [39]– [40]. Additionally, RCN1's role extends to influencing the tumor microenvironment, particularly through its impact on tumor-associated macrophages (TAMs). RCN1 has been shown to promote the polarization of macrophages towards the M2 phenotype, which is associated with tumor progression and immune evasion [37]– [38]. This effect on TAMs further contributes to the aggressive nature of NPC and highlights RCN1 as a potential target for therapeutic intervention.

Functional enrichment analysis revealed NPC-associated DEG enrichment in tumorigenic pathways including cytokine signaling and chemokine interactions, consistent with EBV-driven oncogenic processes [41]– [42]. Our work extends current understanding by identifying novel pathway associations—microtubule bundle formation and humoral immune regulation—that may drive NPC-specific pathobiology. Integrated GSEA-GSVA analysis uncovered RCN1's context-dependent roles: high expression correlates with cell cycle progression, while low expression enhances adaptive immunity, revealing therapeutic vulnerabilities [43]– [44].

While plasma EBV DNA remains the gold-standard biomarker for NPC diagnostics, it exhibits critical limitations in clinical practice: (1) up to 29% of NPC patients in endemic regions show undetectable pretreatment EBV DNA levels, particularly in early-stage disease; (2) transient EBV reactivation in healthy individuals leads to false positives (5–7% in screening cohorts); and (3) EBV DNA alone cannot resolve molecular heterogeneity for personalized therapeutic strategies. Our multi-omics gene panel directly addresses these gaps by integrating complementary biomarkers beyond viral load. Future studies will explore the potential of this panel to reduce NPC diagnostic false-positive rates, either independently or in combination with EBV DNA [45]– [46].

Study limitations include inherent biases from retrospective GEO data analysis, partially addressed through ComBat correction but necessitating validation in prospective cohorts. The inflated performance metrics in small validation datasets underscore the need for multi-ethnic validation. Future research should prioritize mechanistic investigations of identified biomarkers, particularly their interactions with EBV latency programs and immune checkpoint regulation.

5 Limitations

While this study establishes a robust computational framework for NPC biomarker discovery and diagnostic modeling, several limitations should be acknowledged. First, the retrospective nature of publicly sourced transcriptomic datasets may inherently introduce selection biases and technical heterogeneity, despite our implementation of ComBat-based batch correction. While bioinformatics validation demonstrated strong model performance, the diagnostic utility of prioritized biomarkers and the mechanistic basis of model-selected gene signatures require experimental confirmation through in vitro functional assays and in vivo preclinical models. Second, the near-perfect AUC values observed in limited validation cohorts may inflate real-world applicability estimates, necessitating validation through larger multi-center prospective cohorts encompassing diverse ethnic populations. Furthermore, immune microenvironment characterization via CIBERSORT deconvolution, while computationally robust, could be refined through spatially resolved transcriptomics or single-cell RNA sequencing to resolve

cell-type-specific interactions. Future investigations integrating mechanistic wet-lab experiments with multi-omics validation will be essential to bridge computational predictions and clinical translation.

6 Conclusions

In summary, we integrated multi-cohort transcriptomic data with multiple machine learning algorithms to develop two minimal-feature, high-accuracy diagnostic models for NPC (Stepglm[both]-RF and glmBoost-RF models). Our findings position RCN1 as a promising dual-purpose biomarker for NPC diagnosis and prognosis, demonstrating robust diagnostic performance. Future studies will focus on clinical cohort validation and model optimization to establish a scalable, cost-effective strategy for early NPC detection.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s12672-025-02932-2.

Supplementary Material 1

Acknowledgements
Not applicable.

Author contributions
Hehe Wang and Junge Zhang contributed equally to this work. Hehe Wang: Conceptualization, Methodology, Formal analysis, Data curation, Writing – original draft, Visualization. Junge Zhang: Software, Validation, Investigation, Formal analysis, Writing – original draft. Peng Cheng: Resources, Data curation, Validation, Writing – review & editing. Lujie Yu: Methodology, Visualization, Writing – review & editing. Chunlin Li (Corresponding Author): Supervision, Funding acquisition, Project administration, Writing – review & editing. Yaowen Wang (Corresponding Author): Supervision, Resources, Validation, Writing – review & editing. All authors critically reviewed and approved the final manuscript.

Funding
This research was founded by "The effect and molecular mechanism of IL-41 regulating Th17/Treg cell imbalance on allergic rhinitis." Ningbo Natural Science Foundation. Project number: 2024J040.

Data availability
The data used in this study are all from the public database GEO. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE12452;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi? acc=GSE13597;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi? acc=GSE34573;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi? acc=GSE40290;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi? acc=GSE53819;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi? acc=GSE61218;https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi? acc=GSE64634.

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare no competing interests.

Supplementary Information
See the supplementary material for the images and data that support the findings of this study.

Received: 20 March 2025 / Accepted: 5 June 2025
Published online: 12 June 2025

References

1. Jin-Zhang C, Jun-Jun C, Kai X, et al. Clinicopathologic and prognostic significance of VEGF, JAK2 and STAT3 in patients with nasopharyngeal carcinoma. Cancer Cell Int. 2018;18(110):1–9.
2. Min T, Xin D, Lanbo X, et al. CPT1A-mediated fatty acid oxidation promotes cell proliferation via nucleoside metabolism in nasopharyngeal carcinoma. Cell Death Dis. 2022;13(4):331.
3. Prawira A, Oosting S, Chen T, et al. Systemic therapies for recurrent or metastatic nasopharyngeal carcinoma: a systematic review. Br J Cancer. 2017;117(12):1743–52.

4. Xue-Song SS-L, Mei-Juan L. The association between the development of radiation therapy, image technology, and chemotherapy, and the survival of patients with nasopharyngeal carcinoma: a cohort study from 1990 to 2012. Int J Radiat Oncol Biol Phys. 2019;105(3):581–90.
5. Wang H, Wang W, Fan S. Emerging roles of LncRNA in nasopharyngeal carcinoma and therapeutic opportunities. Int J Biol Sci. 2022;18(7):2714–28.
6. Marino GB, Ngai M, Clarke DJB, Fleishman RH, Deng EZ, Xie Z, Ahmed N. Ma'ayan A. GeneRanger and targetranger: processed gene and protein expression levels across cells and tissues for target discovery. Nucleic Acids Res. 2023;51(0):213–24.
7. Agnieszka G, Shadi DS, Kacper Domżał, et al. Celloscope: a probabilistic model for marker-gene-driven cell type deconvolution in spatial transcriptomics data. Genome Biol. 2023;24(1):120.
8. Dodd LE, Sengupta S, Chen IH, den Boon JA, et al. Genes involved in DNA repair and nitrosamine metabolism and those located on chromosome 14q32 are dysregulated in nasopharyngeal carcinoma. Cancer Epidemiol Biomark Prev. 2006;15(11):2216–25.
9. Sengupta S, den Boon JA, Chen IH, Newton MA, et al. Genome-wide expression profiling reveals EBV-associated Inhibition of MHC class I expression in nasopharyngeal carcinoma. Cancer Res. 2006;66(16):7999–8006.
10. Hsu WL, Tse KP, Liang S, Chien YC, et al. Evaluation of human leukocyte antigen-A (HLA-A), other non-HLA markers on chromosome 6p21 and risk of nasopharyngeal carcinoma. PLoS ONE. 2012;7(8):e42767.
11. Bose S, Yap LF, Fung M, Starzcynski J, et al. The ATM tumour suppressor gene is down-regulated in EBV-associated nasopharyngeal carcinoma. J Pathol. 2009;217(3):345–52.
12. Hu C, Wei W, Chen X, Woodman CB, et al. A global view of the oncogenic landscape in nasopharyngeal carcinoma: an integrated analysis at the genetic and expression levels. PLoS ONE. 2012;7(7):e41055.
13. Bao YN, Cao X, Luo DH, Sun R, et al. Urokinase-type plasminogen activator receptor signaling is critical in nasopharyngeal carcinoma cell growth and metastasis. Cell Cycle. 2014;13(12):1958–69.
14. Fan C, Wang J, Tang Y, Zhang S, et al. Upregulation of long non-coding RNA LOC284454 May serve as a new serum diagnostic biomarker for head and neck cancers. BMC Cancer. 2020;20(1):917.
15. Fan C, Xiong F, Tang Y, Li P, et al. Construction of a lncRNA-mRNA co-expression network for nasopharyngeal carcinoma. Front Oncol. 2022;12:809760.
16. Bo H, Gong Z, Zhang W, Li X, et al. Upregulated long non-coding RNA AFAP1-AS1 expression is associated with progression and poor prognosis of nasopharyngeal carcinoma. Oncotarget. 2015;6(24):20404–18.
17. Sanne N, Martin Z, Dirk R et al. Prognostic impact of t(16;21)(p11;q22) and t(16;21)(q24;q22) in pediatric AML: a retrospective study by the I-BFM study group. Blood, 132(15):1584–92.
18. JinHui L, ShuLin Z, SiYue L, et al. Eleven genes associated with progression and prognosis of endometrial cancer (EC) identified by comprehensive bioinformatics analysis. Cancer Cell Int. 2019;19(0):136.
19. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B. 2005;67:301–20.
20. Wan S, Mak M-W, Kung S-Y. R3P-Loc: a compact multi-label predictor using ridge regression and random projection for protein subcellular localization. J Theor Biol. 2014;360(0):34–45.
21. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B. 1996;58:267–88.
22. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016).
23. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonor thogonal problems. Technometrics. 1970;12(1):55–67.
24. Friedman JH. Greedy function approximation: a gradient boosting Ma Chine. Ann Statist. 2001;29(5):1189–232.
25. Ho TK. Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition.1. Manhattan, NY:IEEE;1995: 278–82.
26. John GH, Langley P. Estimating continuous distributions in bayesian classifiers. In: Proceedings of the Eleventh conference on Uncertainty in artificial intelligence (UAI'95). San Francisco: Morgan Kaufmann Publishers; 1995: 338–45.
27. Wiltschko AB, Tsukahara T, Zeine A, Anyoha R, Gillis WF, Markowitz JE, Peterson RE, Katon J, Johnson MJ, Datta SR. Revealing the structure of pharmacobehavioral space through motion sequencing. Nat Neurosci. 2020;23(11):1433–43.
28. Sánchez BJ, Mubaid S, Busque S, de Los Santos YL, Ashour K, Sadek J, Lian XJ, Khattak S, Di Marco S, Gallouzi IE. The formation of HuR/YB1 complex is required for the stabilization of target mRNA to promote myogenesis. Nucleic Acids Res. 2023;51(3):1375–92.
29. Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013;4(0):2612.
30. Qiang Q, Linyu Y, Yunyu F, et al. HDAC i/iib selective inhibitor purinostat mesylate combined with GLS1 Inhibition effectively eliminates CML. Stem Cells Bioact Mater. 2022;21(0):483–98.
31. Chandrashekar DS, et al. UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. Neoplasia. 2017;19:649–58.
32. Thol K, Pawlik P, McGranahan N. Therapy sculpts the complex interplay between cancer and the immune system during tumour evolution. Genome Med. 2022;14(1):137.
33. Garrido F, Schirrmacher V, Festenstein H. H–2-like specificities of foreign haplotypes appearing on a mouse sarcoma after vaccinia virus infection. Nature. 1976;259(5540):228–30.
34. Kang SH, Keam B, Ahn YO, Park HR, Kim M, Kim TM, Kim DW, Heo DS. Inhibition of MEK with Trametinib enhances the efficacy of anti-PD-L1 inhibitor by regulating anti-tumor immunity in head and neck squamous cell carcinoma. Oncoimmunology. 2018;8(1):e1515057.
35. Thompson LD. Update on nasopharyngeal carcinoma. Head Neck Pathol. 2007;1:81–6.
36. Peterson BR, Nelson BL. Nonkeratinizing undifferentiated nasopharyngeal carcinoma. Head Neck Pathol. 2013;7:73–5.
37. Yao H, Zhang S, Xie H, et al. RCN2 promotes nasopharyngeal carcinoma progression by curbing calcium flow and mitochondrial apoptosis. Cell Oncol (Dordr). 2023;46(4):1031–48.
38. Guo H, Shu J, Hu G, et al. Downregulation of RCN1 inhibits esophageal squamous cell carcinoma progression and M2 macrophage polarization. PLoS ONE. 2024;19(5):e0302780.
39. Xu S, Xu Y, Chen L, Fang Q, Song S, Chen J, Teng J. RCN1 suppresses ER stress-induced apoptosis via calcium homeostasis and PERK-CHOP signaling. Oncogenesis. 2017;6(3):e304.

40. Sasaki-Osugi K, Imoto C, Takahara T, Shibata H, Maki M. Nuclear ALG-2 protein interacts with Ca2 + homeostasis Endoplasmic reticulum protein (CHERP) Ca2+-dependently and participates in regulation of alternative splicing of inositol trisphosphate receptor type 1 (IP3R1) pre-mRNA. J Biol Chem. 2013;288(46):33361–75.
41. Matt L, Volker H, Schartinger,Christopher D, Steele, et al. Somatostatin receptor 2 expression in nasopharyngeal cancer is induced by epstein barr virus infection: impact on prognosis, imaging and therapy. Nat Commun. 2021;12(1):117.
42. Lanqi G, Jie L, Yu Z et al. Nasopharyngeal carcinoma cells promote regulatory T cell development and suppressive activity via CD70-CD27 interaction.Nat Commun, 2023, 14(10):1912.
43. Chen W, Lin Y, Jiang M, Wang Q, Shu Q. Identification of LARS as an essential gene for osteosarcoma proliferation through large-Scale CRISPR-Cas9 screening database and experimental verification. J Transl Med. 2022;20(1):355.
44. Xiaofei L, Nianzhao Z, Dawei W, et al. Downregulation of reticulocalbin-1 differentially facilitates apoptosis and necroptosis in human prostate cancer cells. Cancer Sci. 2018;109(4):1147–57.
45. Kanakry J, Ambinder R. The biology and clinical utility of EBV monitoring in blood. Curr Top Microbiol Immunol. 2015;391:475–99.
46. Lee AWM, Lee VHF, Ng WT, et al. A systematic review and recommendations on the use of plasma EBV DNA for nasopharyngeal carcinoma. Eur J Cancer. 2021;153:109–22.

Publisher's note