Cell Biol Toxicol (2025) 41:72
https://doi.org/10.1007/s10565-025-10030-w

RESEARCH

Machine learning-based identification of cuproptosis-related lncRNA biomarkers in diffuse large B-cell lymphoma

Wenhao Ouyang · Zijia Lai · Hong Huang · Li Ling

Received: 28 November 2024 / Accepted: 13 April 2025
© The Author(s) 2025

Abstract Multiple machine learning techniques were employed to identify key long non-coding RNA (lncRNA) biomarkers associated with cuproptosis in Diffuse Large B-Cell Lymphoma (DLBCL). Data from the TCGA and GEO databases facilitated the identification of 126 significant cuproptosis-related lncRNAs. Various feature selection methods, such as Univariate Filtering, Lasso, Boruta, and Random Forest, were integrated with a Transformer-based model to develop a robust prognostic tool. This model, validated through fivefold cross-validation, demonstrated high accuracy and robustness in predicting risk scores. MALAT1 was pinpointed using permutation feature importance from machine learning methods and was further validated in DLBCL cell lines, confirming its substantial role in cell proliferation. Knockdown experiments on MALAT1 led to reduced cell proliferation, underscoring its potential as a therapeutic target. This integrated approach not only enhances the precision of biomarker identification but also provides a robust prognostic model for DLBCL, demonstrating the utility of these lncRNAs in personalized treatment strategies. This study highlights the critical role of combining diverse machine learning methods to advance DLBCL research and develop targeted cancer therapies.

Wenhao Ouyang and Zijia Lai contributed equally to this work.

Highlight
1. A novel subtyping of diffuse large B-cell lymphoma (DLBCL) based on cuproptosis-related lncRNA expression was established, providing new insights into tumor heterogeneity.
2. A robust DLBCL prognostic model was constructed using multiple machine learning algorithms.
3. Experimental evidence confirmed that MALAT1 significantly promotes proliferation in DLBCL cell lines, supporting its role as a potential therapeutic target.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10565-025-10030-w.

W. Ouyang · L. Ling (✉)
Department of Neurology, Shenzhen Hospital, Southern Medical University, No.1333 Xinhu Road, Shenzhen 518000, Guangdong, China
e-mail: linglirabbit@163.com

Z. Lai
Breast Tumor Center, Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University, Guangzhou 510120, Guangdong, China

H. Huang
School of Medicine, Guilin Medical University, Guilin 541000, Guangxi, China

Published online: 21 April 2025

Keywords Cuproptosis · LncRNA · Diffuse large B-cell lymphoma · Machine learning · MALAT1

Introduction

Diffuse Large B-Cell Lymphoma (DLBCL) is the most prevalent type of non-Hodgkin lymphoma in adults, accounting for 30–40% of all case (Tong et al. 2022). This malignancy exhibits considerable heterogeneity, leading to varied pathological features and clinical outcomes, which complicates treatment and affects prognosis (Chapuy et al. 2018; Harrington et al. 2021; Schmitz et al. 2018). Despite advancements in targeted therapy and immunotherapy (Zelenetz, et al. 2023; Shadman et al. 2022), many patients still encounter resistance or relapse, highlighting the need for enhanced therapeutic strategies.

Cuproptosis, a recently defined form of regulated cell death, is mechanistically distinct from established pathways such as apoptosis and necroptosis. Although autophagy is often activated under cellular stress, it primarily serves as a survival mechanism, not a classical mode of cell death (Zhang et al. 2024). Recent studies have characterized cuproptosis as a unique cell death modality, driven by intracellular copper accumulation (Chen et al. 2022). Mechanistically, copper binds directly to lipoylated proteins in the tricarboxylic acid (TCA) cycle, inducing protein aggregation, loss of iron–sulfur cluster proteins, and consequent proteotoxic stress and cell death (Zhang et al. 2024; Li et al. 2022). This pathway underscores a direct link between copper homeostasis, mitochondrial metabolism, and cell fate, attracting increasing attention for its potential role in cancer biology.

Meanwhile, long non-coding RNAs (lncRNAs) are increasingly recognized for their regulatory functions in cancer, influencing processes such as cell proliferation, apoptosis, and therapy resistance (Lin et al. 2019). However, the specific roles of lncRNAs in DLBCL and their potential involvement in regulating cuproptosis are largely unexplored.

To date, most studies on cuproptosis-related biomarkers have focused on protein-coding genes, with the roles of non-coding RNAs—particularly lncRNAs—remaining underexplored (Yang et al. 2022). LncRNAs have emerged as crucial regulators of gene expression and cancer pathogenesis, influencing apoptosis, proliferation, and therapeutic responses. Their high tissue-specific expression and stability in body fluids also render them promising biomarker candidates (Luo et al. 2022; Zhou et al. 2017). In the context of DLBCL, less comprehensive studies have systematically identified cuproptosis-related lncRNAs with prognostic value. Thus, this study exclusively focuses on lncRNAs, aiming to discover novel, non-coding RNA-based biomarkers that could contribute to more accurate molecular classification and individualized therapeutic strategies for DLBCL.

This study employed an integrated approach using multiple machine learning algorithms to enhance the identification and prediction of cuproptosis-related lncRNAs in DLBCL. By combining various feature selection techniques—including Univariate Filtering, Lasso, Boruta, and Random Forest—with advanced deep learning models like Transformers and ensemble methods such as Bagging, we aim to develop a robust and accurate prognostic model. This multi-algorithm approach not only improves the precision of identifying key biomarkers but also provides a powerful tool for predicting patient outcomes and tailoring personalized treatment strategies. Integrating these diverse machine learning methods marks a significant advancement in the field, offering new insights and practical applications in DLBCL management.

Materials and methods

Patient cohorts in this study

RNA expression data and clinical information related to DLBCL were obtained from The Cancer Genome Atlas (TCGA) and included 29 lymphoma samples. Additional data, comprising 414 samples from GSE10846, 200 samples from GSE11318, 69 samples from GSE23501, and 119 samples from GSE53786, were downloaded from The Gene Expression Omnibus (GEO) database. Patient information is summarized in Table 1. After merging the cohorts, the ComBat algorithm was used for batch effect normalization (Leek et al. 2012; Ouyang et al. 2024). The endpoint focused on overall survival. The training sets included 443 cases from TCGA and GSE10846, while the external validation sets comprised 388 cases from GSE11318,

Table 1 Clinical characteristics of DLBCL patient cohorts

Cohort	TCGA-DLBL	GSE10846	GSE11318	GSE23501	GSE53786
Number of patients	29	420	203	69	119
Age (Mean ± SD)	56.5 ± 15.3	61.1 ± 15.4	62.1 ± 14.4	62.5 ± 15.8	61.5 ± 14.8
OS time (Mean ± SD) (years)	4.41 ± 4.55	3.17 ± 3.12	4.21 ± 4.22	2.48 ± 1.98	3.02 ± 3.49
OS status					
Alive	21	249	88	56	75
Dead	8	165	112	13	44
NA	-	6	3	-	-
Gender					
Male	13	224	112	50	66
Female	16	172	91	19	45
NA	-	24	-	-	8

GSE23501, and GSE53786. These datasets were used exclusively for external validation to assess the generalizability and robustness of our model across independent patient populations.

Identification of cuproptosis-related lncRNAs

Cuproptosis-related genes, including NFE2L2, NLRP3, ATP7B, ATP7A, SLC31A1, FDX1, LIAS, LIPT1, LIPT2, DLD, DLAT, PDHA1, PDHB, MTF1, GLS, CDKN2A, DBT, GCSH, and DLST, were identified based on the current study and databases pertinent to copper metabolism and cell death pathways. Subsequently, 1134 lncRNAs were extracted following the guidelines and protocols outlined in the GENCODE project (Frankish et al. 2023).

Using the Pearson correlation method, the interplay between cuproptosis-related genes and lncRNAs in DLBCL samples was analyzed with stringent threshold values of $|R| > 0.4$ and a p-value < 0.001 to ensure significant correlation strength and statistical relevance (Cao et al. 2020). Through this process, 126 cuproptosis-related lncRNAs that showed significant correlation with cuproptosis genes were identified. To effectively illustrate the complex relationships between these genes and their associated lncRNAs, sankey diagrams were constructed, providing a clear and comprehensive overview of potential interactions and regulatory networks. Based on the expression profiles of the cuproptosis-related lncRNAs, non-negative matrix factorization (NMF) clustering was applied to identify distinct molecular subtypes within the DLBCL cohort. The optimal number of clusters was determined using cophenetic correlation and silhouette width to ensure robust subtype classification. This unsupervised learning approach enabled the delineation of novel cuproptosis-associated molecular subtypes.

Feature selection for lncRNAs

GeneSelectR (Zhakparov, et al. 2024) was employed to identify key lncRNAs, utilizing a range of built-in feature selection methods including RandomForest, Lasso, Boruta, and Univariate Filtering (Breiman 2001; Yu et al. 2025; Kursa and Rudnicki 2010; Ouyang et al. 2022). RandomForest, an ensemble method, evaluates feature importance through impurity reduction across decision trees and captures non-linear gene interactions; Lasso regression, which applies L1 regularization, is ideal for shrinking irrelevant coefficients to zero in sparse high-dimensional genomic data; Boruta, a Random Forest-based wrapper, statistically compares features against permuted "shadow" counterparts to filter noise-resistant markers; and Univariate Filtering, which ranks genes rapidly via individual association metrics. This multi-strategy workflow combines univariate pre-screening with model-driven refinement. Permutation feature importance assesses the significance of each gene by measuring changes in model performance when gene values are randomly shuffled, providing a more accurate measure of a gene's importance by accounting for interactions between genes and the overall model complexity.

Multi-algorithm estimation of TME composition and immune scoring

TME composition was inferred using various algorithms integrated within IOBR (Zeng et al. 2021), including CIBERSORT (Newman et al. 2015), TIMER (Li et al. 2016), xCell (Aran et al. 2017), MCPcounter (Becht et al. 2016), EPIC (Racle, et al. 2017), and quanTIseq (Finotello et al. 2019), to estimate the proportions of different cell types within the TME. Immune scoring was performed to evaluate the immune status of the samples using ESTIMATE (Yoshihara et al. 2013) and IPS (Charoentong et al. 2017).

Transformer-based model construction

Initially, in the data processing stage, features were standardized using StandardScaler to eliminate issues caused by differences in feature magnitudes and to ensure uniform scaling across all features. The standardized data were subsequently converted into PyTorch tensors for model training and inference.

A Transformer-based model was developed to predict sample risk scores. The model comprises embedding layers, multiple Transformer encoder layers, fully connected layers, and regularization layers. Specifically, the embedding layer maps the input features $\boldsymbol{X}$ to a 128-dimensional space, denoted by $\boldsymbol{E} = \boldsymbol{Embedding}(\boldsymbol{X})$. Each Transformer encoder layer consists of multi-head self-attention mechanisms and feed-forward neural networks. In our model, eight attention heads are utilized, each attention mechanism being computed as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where Q, K, *and* V represent the query, key, and value matrices, respectively, and d_k is the dimensionality of the keys. The multi-head attention mechanism extends this computation:

$$\text{MultiHead}(\text{Q}, \text{K}, \text{V}) = \text{Concat}\left(\text{head}_1, \ldots, \text{head}_\text{h}\right)\text{W}^\text{O} \tag{2}$$

where each head is defined as:

$$\text{head}_\text{i} = \text{Attention}\left(\text{QW}_\text{i}^\text{Q}, \text{KW}_\text{i}^\text{K}, \text{VW}_\text{i}^\text{V}\right) \tag{3}$$

The feed-forward network in each encoder layer consists of two linear transformations with a ReLU activation between them, represented as:

$$\text{FFN}(\text{x}) = \text{ReLU}\left(\text{xW}_1 + \text{b}_1\right)\text{W}_2 + \text{b}_2 \tag{4}$$

Six such encoder layers were stacked to form a deep feature extraction structure.

Following the encoder layers, the outputs were further processed by the fully connected layers. Initially, the data passed through a 64-dimensional linear layer for preliminary feature extraction, followed by a dropout layer to prevent overfitting, and finally, through an output layer to generate the risk score. To improve the model's stability and generalization ability, Layer Norm and a 30% dropout were included, and the Bagging ensemble method was employed. Bagging enhances performance by training multiple base models and averaging their outputs, thus reducing dependency on a single training set and mitigating the risk of overfitting. In this study, ten base models were trained and their results averaged.

Model training and validation were conducted using fivefold cross-validation. The dataset was randomly split into five subsets, using each subset once as the validation set while the remaining subsets served as the training set. This process was repeated five times to ensure that all the data points were used for both training and validation. Cross-validation helps reduce variability in model performance due to data partitioning and maximizes data utilization, aiding in identifying whether the model is overfitting or underfitting.

During training, the AdamW (Loshchilov and Hutter 2017) optimizer and OneCycleLR (Smith and Topin 2019) learning rate scheduler were utilized. The AdamW optimizer combines adaptive learning rates with weight decay, promoting faster convergence and stability, while the OneCycleLR scheduler dynamically adjusts the learning rate to facilitate rapid convergence in the early stages and fine-tuning later. Additionally, early stopping was implemented to halt training if the validation loss did not decrease within a specified number of epochs, thereby preventing overfitting.

After completing the training for each fold, the model parameters were saved, and various evaluation metrics for the validation set were recorded. By comparing the performance of different folds, the model with the highest concordance index (c-index)

was selected as the final risk-scoring model. To ensure optimal model performance, hyperparameters were optimized, including the embedding dimension (128), number of heads in multi-head self-attention (8), number of encoder layers (6), dropout rate (0.3), number of base models (10), learning rate (0.001), weight decay (1e- 5), number of training epochs (100), and early stopping patience (10).

The performance of the model was comprehensively evaluated using metrics such as accuracy, precision, recall, F1 score, area under the receiver operating characteristic curve (AUC), and concordance index (C-index), ensuring its robustness and effectiveness.

After completing the cross-validation, the model with the highest c-index was selected to calculate the risk scores for all samples. The risk score for each sample was calculated through forward propagation by inputting the features of all samples into the selected best model. Forward propagation efficiently computes the model predictions for each sample, ensuring the accuracy and speed of the calculation process.

DLBCL cell lines culture

This investigation utilized various cell lines, including DLBCL models and a benchmark lymphoblastoid cell line sourced from ATCC. The lymphoblastoid cell line (B-LCL) and DLBCL cell lines (SU-DHL- 4, OCI-LY10, and DB) were cultured in RPMI- 1640 medium (Gibco, USA), supplemented with 10% fetal bovine serum (FBS) and 1% penicillin–streptomycin (Gibco, USA). These cells were maintained under a controlled environment at 37 °C with 5% CO2 humidity, ensuring their exponential phase growth through frequent media renewals and cell passaging.

RNA isolation, cDNA synthesis, and qPCR analysis

RNA was isolated from the cell lines B-LCL, SU-DHL- 4, OCI-LY10, and DB utilizing TRIzol reagent from Thermo Fisher Scientific, USA. A NanoDrop spectrophotometer was employed to ascertain RNA quality and quantity. cDNA synthesis was executed using the HiScript III First Strand cDNA Synthesis Kit (Vazyme, Nanjing, China), adhering strictly to the kit's guidelines. qPCR analysis was conducted on a QuantStudio TMDx system (Applied Biosystems, USA) using ChamQ Universal SYBR qPCR Master Mix (Vazyme, Nanjing, China). The cycling conditions were an initial activation at 95 °C for 30 s, followed by 40 cycles of 95 °C for 5 s and 60 °C for 30 s. Relative lncRNA expression was calculated using the $2^{-\Delta\Delta CT}$ method, normalized to a β-actin gene. Primers for each gene were summarized in Table S1.

siRNA transfection targeting MALAT1

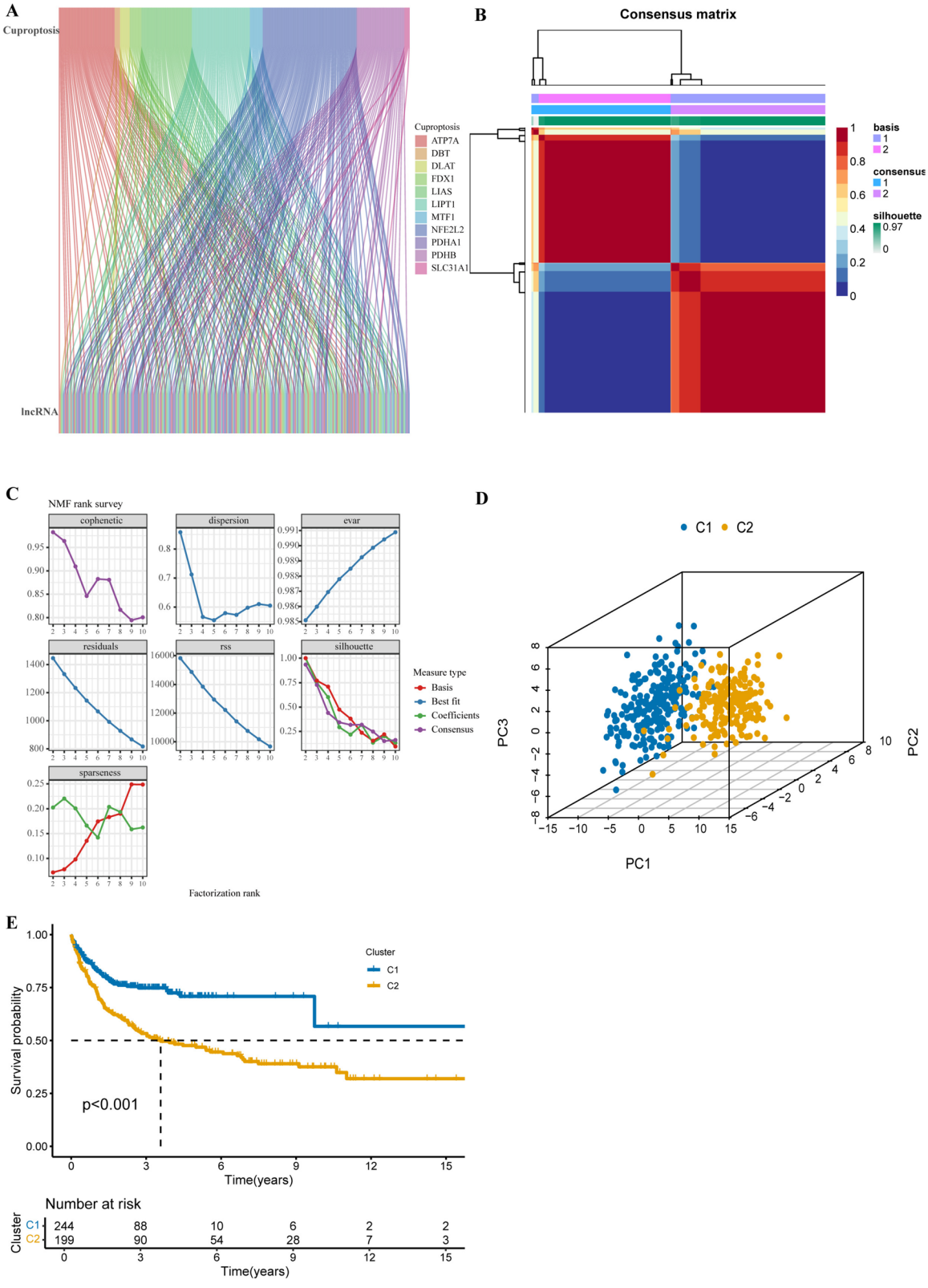
The OCI-LY10 and DB cell lines underwent transfection with siRNA specifically targeting MALAT1, employing Lipofectamine 3000 from Thermo Fisher Scientific, USA, following the provided protocol. Cells were plated in 6-well plates at 2×10^5 cells per well and cultured overnight. Both siRNA and Lipofectamine 3000 were separately mixed with Opti-MEM Reduced Serum Medium and incubated at room temperature for 5 min, followed by a 20-min incubation to form transfection complexes. These were then introduced to the cells and maintained at 37 °C in a 5% CO2 atmosphere. Post 48 h, cells were collected to assess MALAT1 knockdown efficacy, with specific siRNA sequences detailed in Table S2. To ensure specificity, two distinct siRNA sequences for MALAT1 (si-MALAT1#1 and si-MALAT1#2) were employed alongside a scrambled siRNA serving as a negative control. All procedures were conducted in biological triplicates.

CCK-8 assay

Cell proliferation post-siRNA transfection was quantified using the Cell Counting Kit- 8 (CCK- 8) from Dojindo Laboratories, Japan. Cells were re-plated in 96-well plates at 5×10^3 cells per well after 48 h of transfection. The assay was performed at intervals of 24, 48, and 72 h by adding 10 μL of CCK- 8 solution to each well and incubating at 37 °C for 2 h before measuring the absorbance at 450 nm. Viability percentages were compared against controls, with all measurements repeated in triplicate to ensure experimental validity.

Analytical tools and package versions

All statistical computations and the construction of models were executed using R (v4.2.2) and Python (v3.9).


A
Cuproptosis
lncRNA
Cuproptosis
ATP7A
DBT
DLAT
FDX1
LIAS
LIPT1
MTF1
NFE2L2
PDHA1
PDHB
SLC31A1
B
Consensus matrix
basis
consensus
silhouette
C
NMF rank survey
cophenetic
dispersion
evar
residuals
rss
silhouette
sparseness
Measure type
Basis
Best fit
Coefficients
Consensus
Factorization rank
D
C1
C2
PC1
PC2
PC3
E
Cluster
C1
C2
Survival probability
p<0.001
Time(years)
Number at risk
C1 244 88 10 6 2 2
C2 199 90 54 28 7 3
Time(years)

◂**Fig. 1** NMF clustering for constructing DLBCL cuproptosis-related lncRNA subtypes. **A** The Sankey diagram illustrates the lncRNAs related to cuproptosis. **B** Consensus matrix displaying the clustering of samples into two distinct groups based on the NMF results. **C** NMF rank survey evaluating various factorization ranks using cophenetic, dispersion, evar, residuals, rss, silhouette, and sparseness measures to determine the optimal rank. **D** 3D Principal Component Analysis (PCA) plot showing the separation of patients into two clusters (C1 in blue and C2 in yellow). (E) Kaplan–Meier survival curves comparing the overall survival between the two clusters (C1 and C2), with a significant difference ($p < 0.001$)

The development of machine learning algorithms was facilitated by employing the scikit-learn (v1.3.0) and PyTorch (v2.0.1). Analyses of immune deconvolution and the tumor microenvironment were performed utilizing the IOBR package (v0.99.9) in R (Table S3).

Computational resources

The computational model was trained utilizing a robust high-performance computing setup, which included an NVIDIA RTX 2080 Ti GPU with 11 GB of VRAM and a system memory of 128 GB. The entire training process, spanning 100 epochs, was completed in roughly 30 minutes.

Results

Patient cohorts

All cohorts of 831 DLBCL patients was analyzed, consisting of 29 from the TCGA dataset and 414 from the GSE10846 dataset as training groups; and 200 from GSE11318, 119 from GSE53786, and 69 from GSE23501 as validation groups (Supplementary Fig. 1A). The 'ComBat' function from the 'sva' package was employed to address batch effects among these groups, with pre- and post-correction states indicating successful normalization (Supplementary Fig. 1B & C).

Identification of novel molecular subtypes via non-negative matrix factorization clustering

We analyzed 1,134 lncRNAs along with 17 genes linked to cuproptosis. Co-expression analysis highlighted 126 lncRNAs significantly correlated with genes involved in cuproptosis (Fig. 1A, Table S4).

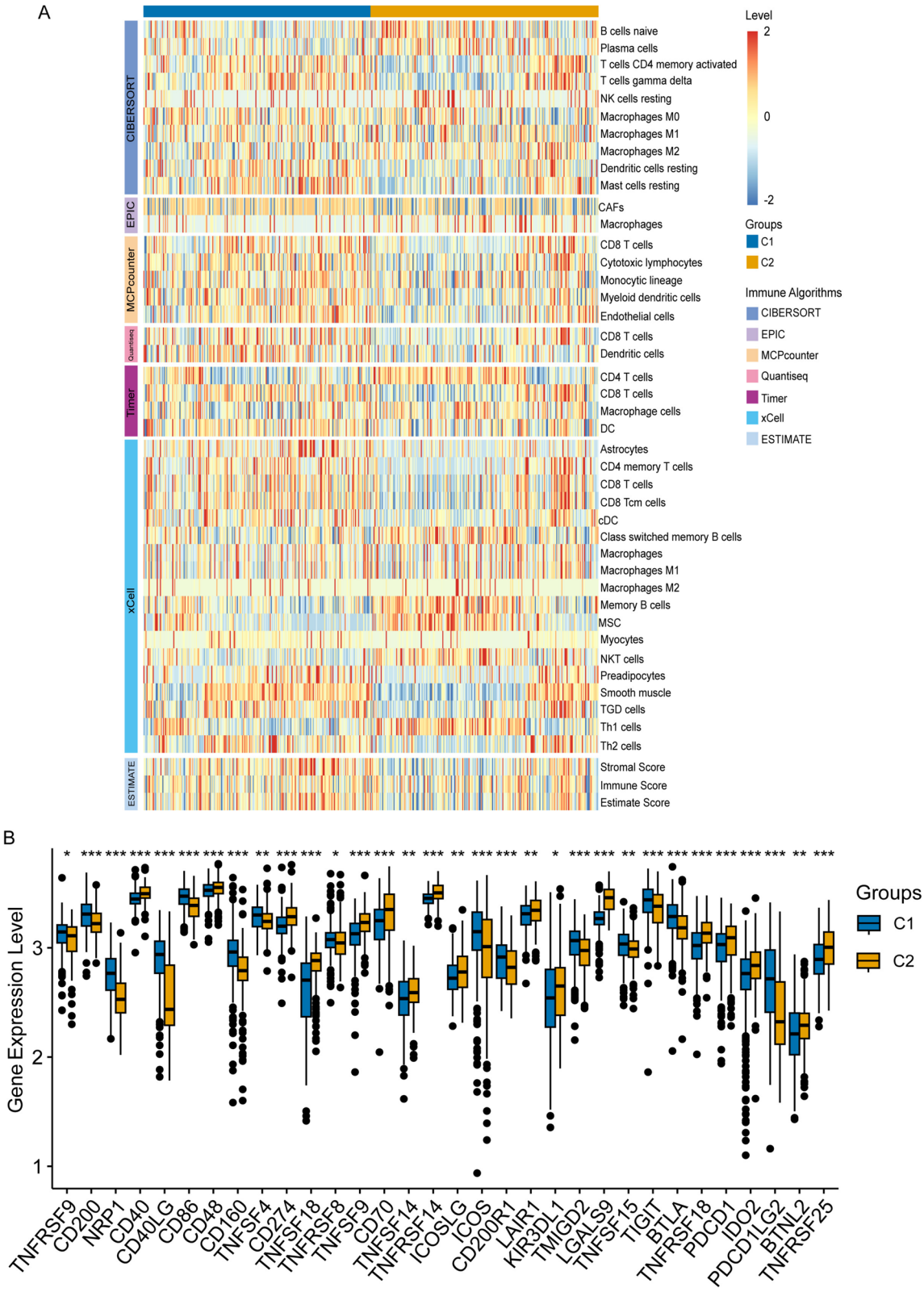
To elucidate the roles and interrelations of these lncRNAs in DLBCL, Non-negative Matrix Factorization (Wang and Zhang 2012) (NMF) was utilized for cluster analysis. After thorough validation, a model with two clusters was found to be most effective. The clustering consensus matrix revealed distinct sample separation into two primary groups, as evidenced by deeper color intensities indicating stronger agreement. The cophenetic correlation coefficient peaked at approximately 0.989 with a factorization rank of 2, signifying a highly stable and consistent clustering framework (Fig. 1B and C). The supporting low residual sum of squares and the near-perfect silhouette score further validated the efficacy of the two-cluster model.

To corroborate the clinical relevance of the NMF clustering outcomes, Principal Component Analysis (PCA) and survival studies were conducted. The PCA exhibited distinct gene expression separations between the clusters (Fig. 1D), and the survival analysis indicated significant prognostic differences between them (Fig. 1E). These combined results support the rationality and clinical significance of classifying cuproptosis-related lncRNAs in DLBCL into two distinct clusters, laying a robust foundation for future biomarker identification and tailored therapeutic approaches.

Tumor Microenvironment (TME) and immune checkpoints analysis

In recent years, research has demonstrated that DLBCL is influenced not only by the autonomous proliferation of tumor cells but also by signals from the TME. To explore this further, TME-related analyses were conducted within our two-cluster classification using the IOBR package, which employs eight deconvolution methods to decode the TME and conduct a comprehensive analysis simultaneously, including CIBERSORT, TIMER, xCell, MCPcounter, ESITMATE, EPIC, IPS, and quanTIseq (Supplementary Fig. 2).

Despite variations in TME composition inference across different algorithms, significant differences in certain immune cell populations were identified between clusters (Fig. 2A). Notably, differences were observed in several cell types, including naive B cells, plasma cells, activated CD4 memory T cells, gamma delta T cells, and resting NK cells. For example,


A
CIBERSORT
EPIC
MCPcounter
Quantiseq
Timer
xCell
ESTIMATE
B cells naive
Plasma cells
T cells CD4 memory activated
T cells gamma delta
NK cells resting
Macrophages M0
Macrophages M1
Macrophages M2
Dendritic cells resting
Mast cells resting
CAFs
Macrophages
CD8 T cells
Cytotoxic lymphocytes
Monocytic lineage
Myeloid dendritic cells
Endothelial cells
CD8 T cells
Dendritic cells
CD4 T cells
CD8 T cells
Macrophage cells
DC
Astrocytes
CD4 memory T cells
CD8 T cells
CD8 Tcm cells
cDC
Class switched memory B cells
Macrophages
Macrophages M1
Macrophages M2
Memory B cells
MSC
Myocytes
NKT cells
Preadipocytes
Smooth muscle
TGD cells
Th1 cells
Th2 cells
Stromal Score
Immune Score
Estimate Score
Level
2
0
-2
Groups
C1
C2
Immune Algorithms
CIBERSORT
EPIC
MCPcounter
Quantiseq
Timer
xCell
ESTIMATE
B
Gene Expression Level
1
2
3
Groups
C1
C2
TNFRSF9
CD200
NRP1
CD40
CD40LG
CD86
CD48
CD160
TNFSF4
CD274
TNFSF18
TNFRSF8
TNFSF9
CD70
TNFSF14
TNFRSF14
ICOSLG
ICOS
CD200R1
LAIR1
KIR3DL1
TMIGD2
LGALS9
TNFSF15
TIGIT
BTLA
TNFRSF18
PDCD1
IDO2
PDCD1LG2
BTNL2
TNFRSF25

◂**Fig. 2** Differential immune cell and immune checkpoint expression between two subtypes. **A** Heatmap showing the differential expression of various immune cell types between the two subtypes (C1 in blue and C2 in yellow) using multiple immune algorithms (CIBERSORT, EPIC, MCPcounter, Quantiseq, Timer, xCell, ESTIMATE). **B** Box plots comparing the expression levels of immune checkpoint genes between the two subtypes (C1 and C2), indicating significant differences across several genes

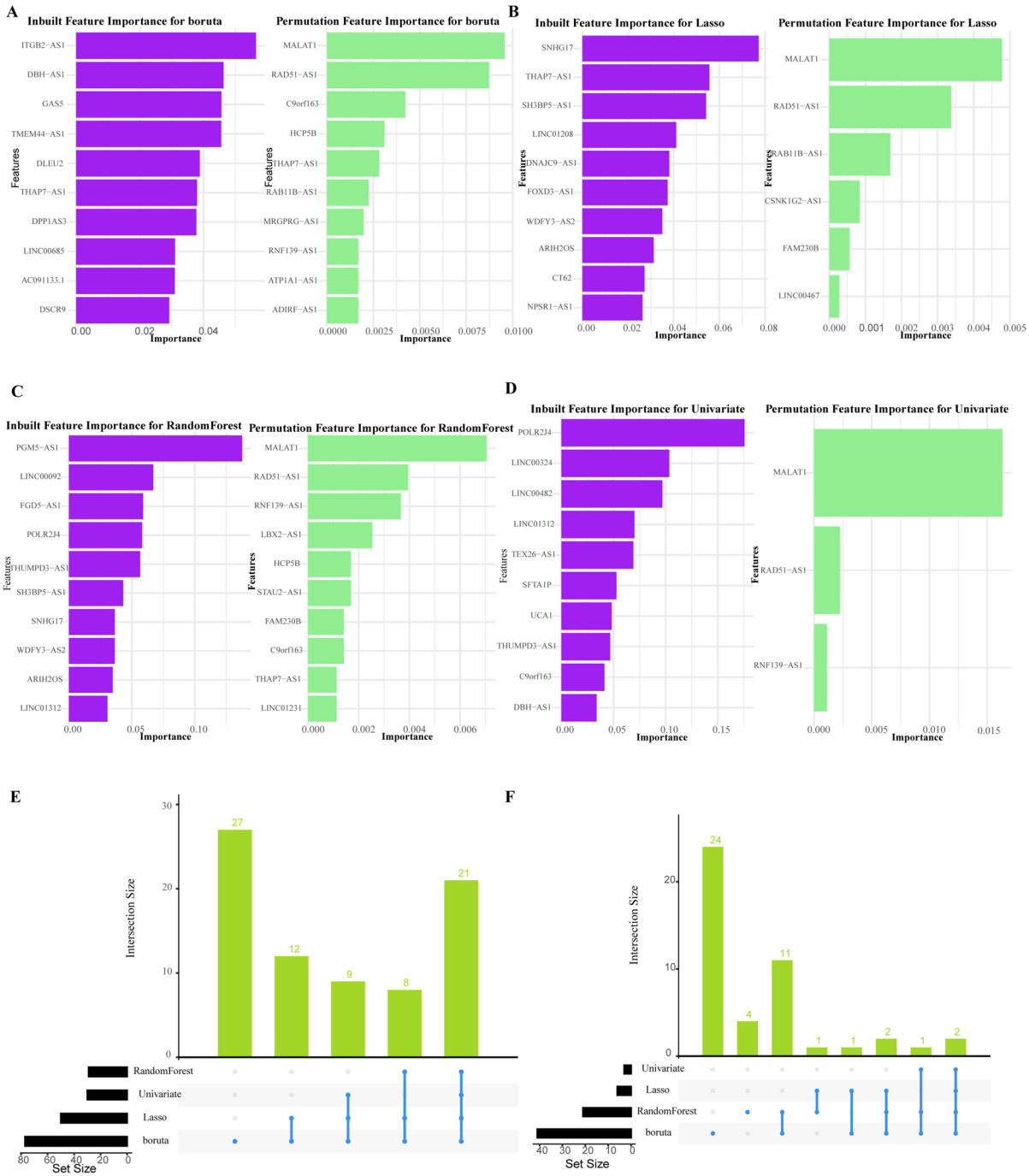
cluster 1 exhibited a higher presence of activated CD4 memory T cells and gamma delta T cells, indicative of a more active immune response. In contrast, cluster 2 displayed higher levels of naive B cells and resting NK cells, suggesting a less active or more suppressive immune environment. These findings remained consistent across multiple deconvolution methods such as CIBERSORT, EPIC, and xCell. For patients in cluster 1, therapies that enhance the immune response may be advantageous. Conversely, for patients in cluster 2, strategies that mobilize specific immune cells may aid in overcoming the suppressive TME.

Immunotherapy has become a significant treatment modality over the past decade, especially for relapsed and refractory DLBCL. Building on our TME analysis, the expression levels of common immune checkpoints were examined in these two clusters. Given the critical role of immune checkpoints in regulating immune responses, the expression of key checkpoints such as PD- 1, PD-L1, CTLA- 4, and others was analyzed within the two clusters (Fig. 2B). The varied expression of these checkpoints provides insights into the potential responsiveness of each cluster to immunotherapy. For instance, Cluster 1, with its more active immune profile, might exhibit higher expression levels of certain checkpoints, suggesting a potential for a better response to checkpoint inhibitors. Conversely, Cluster 2, with its more suppressive immune environment, might show different checkpoint profiles, indicating alternative immunotherapeutic strategies. By correlating immune checkpoint expression with clinical outcomes, we aimed to identify biomarkers that can predict the response to immunotherapy and guide treatment decisions. This approach could improve the precision of immunotherapy, ensuring that patients receive the most effective treatment based on their specific TME and immune checkpoint characteristics.

Multi-algorithm feature selection strategy

To further identify significant gene features among the 126 lncRNAs, four complementary feature selection techniques were employed: Univariate feature selection, Lasso, Boruta, and Random Forest. Each method identified a different number of key lncRNAs: Univariate feature selection identified 30 lncRNAs, Lasso identified 50 lncRNAs, Boruta identified 77 lncRNAs, and Random Forest identified 29 lncRNAs (Table S5). For example, Univariate feature selection highlighted POLR2 J4, LINC00324, and LINC00482 (Fig. 3A); Lasso highlighted SNHG17, THAP7-AS1, and SH3BP5-AS1 (Fig. 3B); Boruta identified ITGB2-AS1, DBH-AS1, GAS5, and TMEM44-AS1 (Fig. 3C); and Random Forest identified PGM5-AS1, LINC00092, and FGD5-AS1 (Fig. 3D).

The UpSet plots demonstrated both intersections and unique attributes pinpointed by each method, with one diagram revealing the overlap and distinct traits based on inbuilt importance scores (Fig. 3E), and another illustrating the intersections derived from permutation importance scores (Fig. 3F). Considering the advantages of each technique and aiming for a comprehensive feature set, the intersecting genes recognized by both inbuilt and permutation assessments across all four machine learning methods were selected. By integrating these intersecting attributes, the intention was to harness the consensus of diverse methods, guaranteeing the relevance and significance of the chosen genes for our analytical endeavors. Performance evaluation of these feature selection methods was conducted using several metrics: CV Score, F1 Score, Recall, Precision, and Accuracy (Fig. 4, Table S6). Boruta exhibited strong performance in CV Score and Recall, indicating its robustness and high sensitivity. Lasso demonstrated the highest Precision, reflecting its accuracy in positive predictions. Random Forest excelled in Accuracy and F1 Score, suggesting its overall reliability and balance in precision and recall. Heatmaps (Supplementary Fig. 3) further illustrate the similarity and robustness of these methods. The built-in overlap coefficient heatmap showed high consistency between the methods, particularly between Lasso and Boruta. However, the permutation overlap coefficient heatmap exhibited lower consistency, highlighting the sensitivity of these methods to data perturbations.



Transformer-based model for score prediction

A Bagging Ensemble model utilizing Transformer architecture was deployed to enhance predictive analysis accuracy (Fig. 5A). Initially, the dataset was prepared by loading, preprocessing, and standardizing features and labels using StandardScaler, which ensured uniformity and enhanced model performance. The Transformer architecture included an embedding layer, a multi-layer Transformer encoder, fully connected layers, and a Sigmoid activation function (Fig. 5B). This configuration enabled the model to capture complex dependencies and

◂**Fig. 3** Feature importance analysis using various machine learning methods. **A** Inbuilt and permutation feature importance for the Boruta method. The left panel shows inbuilt feature importance, while the right panel shows permutation feature importance. **B** Inbuilt and permutation feature importance for the Lasso method. The left panel shows inbuilt feature importance, while the right panel shows permutation feature importance. **C** Inbuilt and permutation feature importance for the RandomForest method. The left panel shows inbuilt feature importance, while the right panel shows permutation feature importance. **D** Inbuilt and permutation feature importance for the Univariate method. The left panel shows inbuilt feature importance, while the right panel shows permutation feature importance. **E** Upset plot displaying the intersections of significant features among different methods, with the total number of significant features identified by each method shown on the bars. **F** Upset plot displaying the intersections of top features among different methods, with the total number of top features identified by each method shown on the bars

generate binary classification results. During the training phase, a bagging strategy was utilized to capitalize on the strengths of ensemble learning. The ensemble comprised ten Transformer model instances, each trained on distinct bootstrap samples of the training dataset. Predictions from these base models were aggregated to produce the final output, thus minimizing variance and enhancing both generalization and hyperparameter optimization as outlined in Table S7.

To ensure robustness, fivefold cross-validation was employed and early stopping was implemented based on validation loss to avert overfitting. The top-performing model, chosen based on the highest c-index, attained an accuracy of 0.93, precision of 0.76, recall of 0.76, F1 score of 0.76, AUC of 0.93, and c-index of 0.93. The predictive efficacy of this model proved substantially superior to that of the conventional Cox regression risk model (Table S8, Supplementary Fig. 4A-C).

Model performance evaluation

Employing the optimal model, risk scores were generated for both the training and external validation sets, offering a quantitative assessment of risk for each sample. In the training sets, ROC analysis revealed AUC values of 0.790, 0.829, and 0.822 for 1-year, 3-year, and 5-year survival predictions, respectively (Fig. 6A), displaying superior predictive performance compared to the traditional Cox regression risk model (Supplementary Fig. 4D). The substantial AUC values indicate the model's effectiveness in discriminating between high-risk and low-risk individuals. This assertion is corroborated by the Kaplan–Meier survival curves, which exhibited a significant survival disparity between the high- and low-risk groups ($p < 0.0001$, Fig. 6B).

In the external validation sets, ROC analysis yielded AUC values of 0.662, 0.678, and 0.688 for 1-year, 3-year, and 5-year survival predictions, respectively (Fig. 6C), indicating moderate predictive performance. The decline in AUC values within the validation cohort, as opposed to the training cohort, is anticipated given that the external validation data are novel to the model and pose a stringent test of its generalizability. Despite these moderate AUC values, Kaplan–Meier analysis continued to show a significant survival difference between the high- and low-risk groups ($p < 0.0001$, Fig. 6D), verifying that the model maintains effective patient stratification capability outside the training dataset. This validation is essential, as it affirms the model's robustness and its prospective applicability across diverse patient populations.

Independent risk analysis using forest plots demonstrated that the risk score was a highly significant survival predictor in both cohorts. In the training cohort, the risk score exhibited a hazard ratio of 9.423 (95% CI: 5.836–15.213, $p < 0.001$), underscoring a potent correlation with survival outcomes (Fig. 6E). In the validation cohort, the risk score continued to be a significant predictor, with a hazard ratio of 8.750 (95% CI: 5.413–14.143, $p < 0.001$), reaffirming its predictive power (Fig. 6F). Age proved to be another significant predictor, whereas gender did not, suggesting that the risk score effectively encapsulates the key factors impacting survival.

The Transformer-based Bagging Ensemble model demonstrated strong predictive capabilities, with consistent performance across different evaluations. The prominent role of the risk score as a significant predictor in both training and validation cohorts emphasizes its essential function in forecasting survival. These findings highlight the model's generalizability and potential clinical utility in precisely stratifying patients by risk, thereby facilitating personalized treatment planning.

Expression analysis of key prognostic lncRNAs in DLBCL cell lines

To elucidate the molecular mechanisms underlying DLBCL, quantitative PCR (qPCR) was employed to

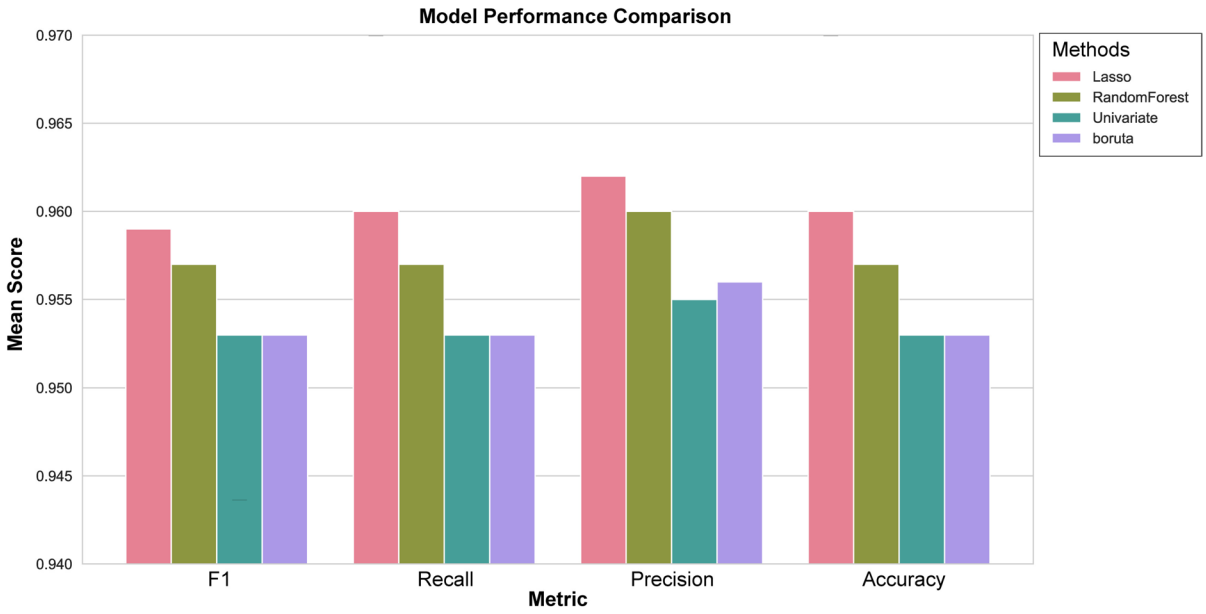


Fig. 4 Performance metrics of different machine learning methods. Bar plot showing the different cross-validation (CV) mean scores among F1 score, recall values, precision values, accuracy values for Boruta, Lasso, RandomForest, and Univariate methods

assess the expression patterns of pivotal lncRNAs across various DLBCL cell lines including B-LCL, SU-DHL- 4, OCI-LY10, and DB. These lncRNAs were identified using their importance scores derived from machine learning algorithms, with a particular focus on MALAT1, which was highly ranked based on permutation importance. Our analysis indicated that MALAT1 expression was significantly elevated in the DLBCL cell lines compared to the control lymphoblastoid cell line (Fig. 7A).

MALAT1 promotes DLBCL cell proliferation

In cell lines exhibiting high levels of MALAT1, specifically OCI-LY10 and DB, MALAT1 knockdown was executed. The knockdown of siMALAT1 achieved

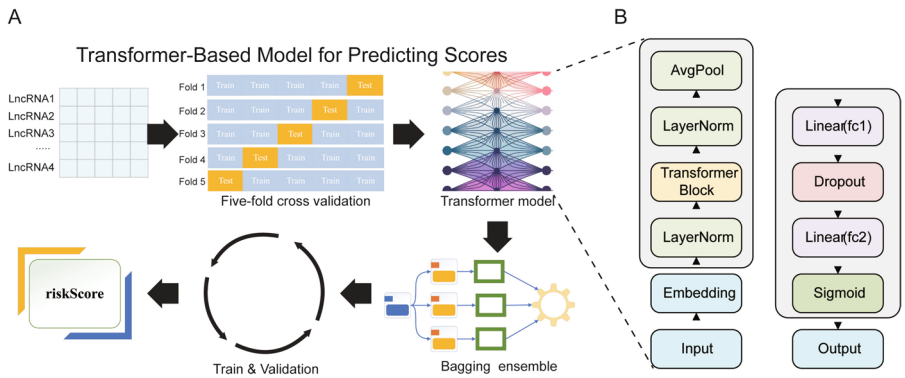


Fig. 5 Transformer-based model for predicting scores. **A** Overview of transformer-based model workflow for predicting scores using lncRNA data. The lncRNA matrix was subjected to five-fold cross-validation, with each fold alternately used for training and testing. The trained transformer model was then applied, followed by bagging ensemble techniques to generate risk scores through an iterative training and validation process. **B** Detailed architecture of transformer model used in this study. The model includes an input embedding layer, multiple transformer blocks with LayerNorm, AvgPool, and dropout layers, followed by linear layers (fc1 and fc2) and a sigmoid output layer

substantial efficiency in both OCI-LY10 and DB cell lines (Fig. 7B and C). Subsequent assessments of MALAT1's role in DLBCL were conducted through CCK- 8 assays. These assays demonstrated that cell proliferation in the MALAT1 knockdown groups was diminished compared to the control groups in both OCI-LY10 and DB cell lines (Fig. 7D and E). This evidence underscores MALAT1's significant function in enhancing DLBCL cell proliferation and highlights its potential as a therapeutic target in DLBCL treatment strategies.

Discussion

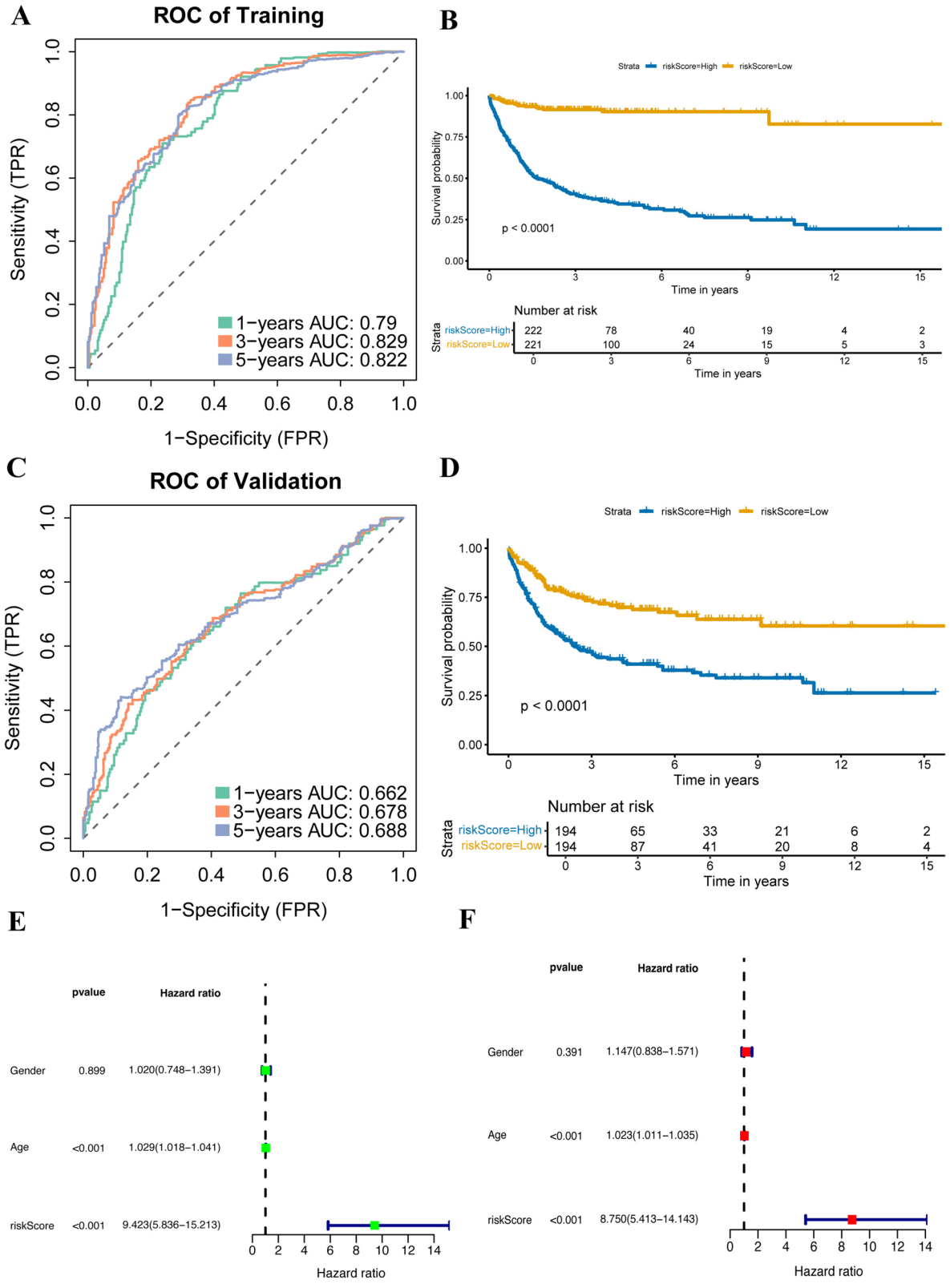
This research explored the prognostic value of cuproptosis-related lncRNAs in DLBCL, aiming to construct a dependable prognostic framework. By harnessing these lncRNAs through sophisticated feature selection techniques, we established a Transformer-based prognostic model, augmented by the Bagging ensemble approach, which proficiently forecasts patient outcomes. The efficacy of this innovative risk model in segregating patients into distinct prognostic categories highlights the promise of these lncRNAs as biomarkers for DLBCL.

Cuproptosis, or copper-induced cell death, emphasizes the pivotal role of copper metabolism, mitochondrial dynamics, and protein alterations in cancer development. Initially proposed by Tsvetkov et al. (Tsvetkov et al. 2022), this concept has inspired bioinformatic investigations to identify markers of cuproptosis by analyzing genes associated with this process. The link between copper levels and cancer is well-established, as tumors generally require more copper than normal tissues. Specifically, cancers such as melanoma, breast cancer, and leukemia, especially those exhibiting stem cell-like traits or resistance to treatment (Liu et al. 2024; Zheng et al. 2022; Lu et al. 2024; Ning et al. 2023), are characterized by heightened mitochondrial metabolism and increased aerobic respiration.

Advancements in molecular and genetic sequencing technologies have markedly refined the classification of lymphomas, enhancing personalized treatment plans (Jaffe 2019). The identification of DLBCL driver genes and pathways has deepened our understanding of its biological behavior, facilitating better risk stratification and prognostic accuracy, and paving the way for tailored therapeutic approaches. Schmitz et al. (Schmitz et al. 2018) identified four genetic subtypes of DLBCL, each associated with distinct prognostic implications. Wright et al. (Wright, et al. 2020) further divided DLBCL into seven genetic subtypes, introducing the LymphGen molecular classification algorithm, which elucidates the heterogeneity of DLBCL and its variable responses to immune therapy. Lacy et al. (Lacy et al. 2020) investigated the genetic mutation profiles in DLBCL patients, uncovering distinct prognostic impacts for each subtype and underscoring the influence of genetic mutations on prognosis. These contributions have refined the molecular categorization of DLBCL, steering new directions for precision therapy.

Despite the advantages of NGS and related technologies, several limitations hinder their broader clinical implementation. Firstly, the high cost of NGS technology may restrict its extensive clinical use (Yin et al. 2021). Secondly, the complexity of data analysis necessitates specialized bioinformatics expertise and technical support. Thirdly, variations in standardization across different research institutions can affect the reproducibility and consistency of results. Additionally, molecular classification methods that rely predominantly on specific gene mutations or expressions might not adequately represent the full heterogeneity and dynamic nature of tumors. Finally, although molecular classification provides crucial insights for therapy, further empirical research and clinical trials are essential to validate these methods for routine clinical use and to ensure they accurately reflect patient prognoses.

In this study, we proposed a novel molecular classification for DLBCL based on cuproptosis-related lncRNAs. From the dataset, we extracted 1,134 lncRNAs and 17 genes associated with cuproptosis, identifying 126 significantly co-expressed lncRNAs through co-expression analysis. Employing the NMF algorithm for clustering analysis, we established two optimal models that demonstrated strong consistency and distinct classification structures among the samples. Subsequent PCA and survival analysis indicated marked differences in gene expression and prognosis between the two identified molecular subtypes. This study not only introduces a new molecular classification for DLBCL but also underscores the profound clinical implications of these classifications


A
ROC of Training
Sensitivity (TPR)
1−Specificity (FPR)
1−years AUC: 0.79
3−years AUC: 0.829
5−years AUC: 0.822
B
Strata riskScore=High riskScore=Low
Survival probability
p < 0.0001
Time in years
Number at risk
riskScore=High 222 78 40 19 4 2
riskScore=Low 221 100 24 15 5 3
C
ROC of Validation
Sensitivity (TPR)
1−Specificity (FPR)
1−years AUC: 0.662
3−years AUC: 0.678
5−years AUC: 0.688
D
Strata riskScore=High riskScore=Low
Survival probability
p < 0.0001
Time in years
Number at risk
riskScore=High 194 65 33 21 6 2
riskScore=Low 194 87 41 20 8 4
E
pvalue
Hazard ratio
Gender 0.899 1.020(0.748−1.391)
Age <0.001 1.029(1.018−1.041)
riskScore <0.001 9.423(5.836−15.213)
Hazard ratio
F
pvalue
Hazard ratio
Gender 0.391 1.147(0.838−1.571)
Age <0.001 1.023(1.011−1.035)
riskScore <0.001 8.750(5.413−14.143)
Hazard ratio

◂**Fig. 6** Evaluation of predictive performance of the riskscore model. **A** ROC curve for the risk score model predicting 1-year, 3-year, and 5-year survival with AUC values of 0.79, 0.829, and 0.822, respectively. **B** Kaplan–Meier survival curves comparing high and low risk score groups, with a significant difference in survival ($p < 0.0001$). The numbers at risk at different time points are shown in the plot below. **C** ROC curve for the risk score model in a different dataset, predicting 1-year, 3-year, and 5-year survival, with AUC values of 0.662, 0.678, and 0.688, respectively. **D** Kaplan–Meier survival curves comparing high and low risk score groups in the different datasets, with a significant difference in survival ($p < 0.0001$). The numbers at risk at different time points are shown in the plot below. **E** Forest plot showing the hazard ratios and p-values for gender, age, and risk score in the primary dataset. The risk score is a significant predictor of survival. **F** Forest plot showing the hazard ratios and *p*-values for gender, age, and risk score in the different datasets. The risk score remains a significant predictor of survival

in prognostication, laying a solid groundwork for future biomarker research and personalized treatment approaches.

Our multi-machine learning strategy effectively pinpointed critical lncRNAs that predict DLBCL molecular subtypes and prognoses. By integrating Univariate feature selection, Lasso, Boruta, and Random Forest, we achieved a thorough selection of significant features. This robust approach revealed both overlapping and unique lncRNAs, enhancing the precision of our predictive models. The transformer-based bagging ensemble model, constructed from these selected features, exhibited exceptional efficacy in predicting DLBCL outcomes, highlighting the utility of our multi-algorithm strategy in refining molecular classification and prognosis prediction. This ensemble model demonstrated high performance across both internal and external validation cohorts. During fivefold cross-validation, the model maintained moderate performance on independent external datasets, suggesting its generalizability across different patient groups. These findings affirm the model's potential clinical value for DLBCL risk stratification.

Then, we focused on MALAT1 (Skeparnias et al. 2024), an lncRNA with significant prognostic implications in DLBCL. MALAT1 has been implicated in various cellular processes including cell proliferation, migration, and drug resistance (Wang et al. 2019). Recent research has highlighted its role in regulating cell death pathways, particularly copper-induced cell death (Tan et al. 2019). MALAT1 has been shown to modulate cell survival and proliferation in several cancers. For example, in lung cancer (Bhat et al. 2024), MALAT1 was found to promote tumor growth and resistance to chemotherapy by affecting cellular stress responses and apoptosis pathways. In breast cancer (Kim et al. 2018), it has been linked to metastasis and poor prognosis through its influence on cell cycle regulation and epithelial-mesenchymal transition.

Despite the lack of research into the role of MALAT1 in copper-induced apoptosis in cells, its potential influence cannot be underestimated. Copper is a known promoter of oxidative stress and apoptosis in oncogenic cells, possibly impacting these mechanisms through the modulation of copper metabolism or the oxidative stress response. This underscores the importance of dissecting the function of MALAT1 in this framework to elucidate its therapeutic relevance and its role in the pathophysiology of DLBCL.

Although traditionally classified as a housekeeping lncRNA due to its pervasive expression across various tissues, recent evidence suggests MALAT1's active participation in the regulation of several cancers, including lymphomas. In our study, MALAT1 consistently emerged as a prominent feature in various machine learning models, indicating a deliberate and significant selection rather than random chance. Furthermore, its elevation in DLBCL cell lines, coupled with its demonstrated influence on cell proliferation via siRNA-mediated silencing, underscores its integral role in cell proliferation.Our findings contest the notion that the prominence of MALAT1 in our models is simply an artifact of its inherent expression stability. Rather, the applied machine learning techniques appear to have effectively distinguished its specific pathogenic relevance through consistent expression disparities and its verified functional impact in DLBCL. This accentuates the utility of computational methods in not only pinpointing biomarkers of statistical relevance but also in confirming their pathobiological importance.

Within the spectrum of lncRNAs evaluated, MALAT1 was identified as a key element consistently across diverse machine learning techniques, with validations at both the expression and functional strata. Recognized as an oncogenic lncRNA, MALAT1 is involved in several types of cancers, including those of the lung, breast, and blood (Amodio et al. 2018). Specifically, in DLBCL, our findings demonstrate MALAT1's facilitation of cell proliferation,

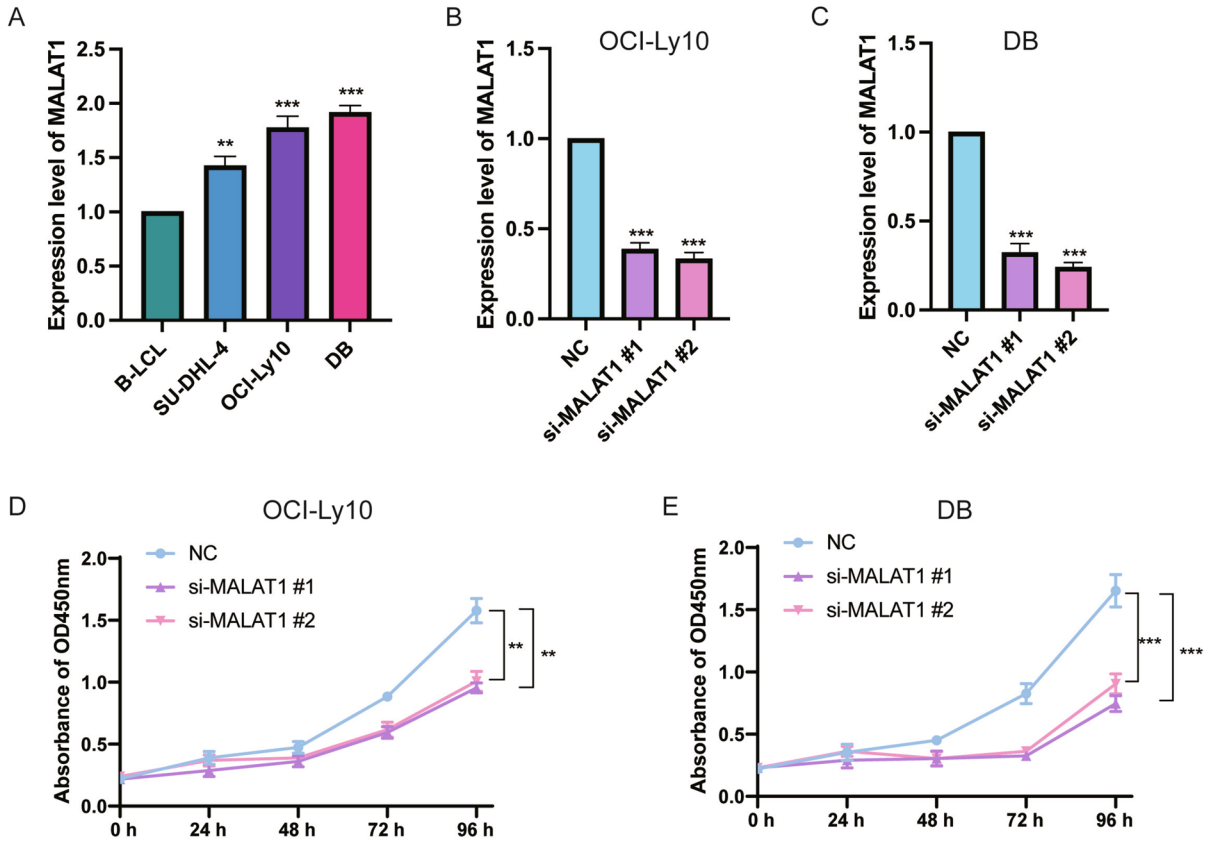


Fig. 7 Expression and proliferation analysis of MALAT1 in DLBCL cell lines. **A** MALAT1 expression levels in various cell lines, including B-LCL, SU-DHL- 4, OCI-LY10, and DB, measured by qPCR. **B** Knockdown efficiency of siMALAT1 in OCI-LY10 cells, showing a significant reduction in MALAT1 expression with two siMALAT1 constructs (si-MALAT1 #1 and si-MALAT1 #2) compared with the negative control (NC). **C** Knockdown efficiency of siMALAT1 in DB cells, demonstrating a marked decrease in MALAT1 expression with si-MALAT1 #1 and si-MALAT1 #2 compared with NC. **D** Cell proliferation assessed by CCK- 8 assay in OCI-LY10 cells following MALAT1 knockdown with si-MALAT1 #1 and si-MALAT1 #2. The absorbance at 450 nm was measured at 0, 24, 48, 72, and 96 h post-transfection. **E** Cell proliferation assessed by CCK- 8 assay in DB cells following MALAT1 knockdown with si-MALAT1 #1 and si-MALAT1 #2. The absorbance at 450 nm was measured at 0, 24, 48, 72, and 96 h post-transfection. $^{**}p < 0.01$, $^{***}p < 0.001$

indicating its probable influence on tumoral advancement. With emerging interest in cuproptosis, it's conceivable that MALAT1 might interact with pathways involved in cuproptosis, potentially via lncRNA-miRNA-mRNA interactions or by impacting copper homeostasis and oxidative stress. Future investigations are essential to substantiate these propositions and to further elucidate the roles of MALAT1 and other lncRNAs related to cuproptosis in the pathology and therapeutic resistance of DLBCL.

Our investigation establishes a comprehensive framework for the identification of lncRNA biomarkers linked to cuproptosis in DLBCL, though several limitations persist. The cohorts used for validation were retrospective and publicly available, highlighting the imperative for prospective clinical trials. While MALAT1's functional role was corroborated, the involvement of other lncRNAs awaits further experimental validation. Our model relied exclusively on transcriptomic data, potentially restricting the biological depth of the findings. Future studies integrating multi-omic data, including genomic, epigenomic, and proteomic layers, could enhance the granularity of DLBCL's heterogeneity and augment the predictive accuracy of these models. Additionally, the observed decrease in AUC values in external cohorts suggests potential batch effects, differences in platforms, and variability among patients.

Although ComBat normalization was employed to mitigate these issues, its influence on feature selection must be acknowledged. Nonetheless, batch correction remains crucial for preserving consistency across studies and facilitating meaningful external validations.

In conclusion, our study capitalizes on multiple machine learning methodologies to refine the molecular profiling and prognostic accuracy of DLBCL. Through the integration of various algorithmic approaches and the deployment of a sophisticated Transformer-based model, we significantly enhanced the stratification of risk. This multifaceted machine learning strategy is pivotal in advancing our comprehension of DLBCL and in guiding the development of targeted treatment modalities.

Author contributions W.O: Conceptualization, Validation, Investigation, Resources, Writing—Original Draft, Writing—Review & Editing, Visualization, Supervision. Z.L: Conceptualization, Methodology, Software, Data Curation, Formal analysis, Writing—Original Draft, Writing—Review & Editing, Visualization. H.H: Data Curation, Formal analysis, Writing—Review & Editing, Visualization. L.L: Writing—Review & Editing, Supervision, Funding acquisition. All authors reviewed the manuscript.

Funding This study was supported by the Science and Technology Project of Shenzhen (JCYJ20210324131614038).

Data availability We have uploaded the model code to Github: https://github.com/ZijiaLai/Tmodel.git.

Declarations

Ethical approval This study was permitted by the Ethics Committee of Shenzhen Hospital, Southern Medical University.

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

References

Amodio N, et al. MALAT1: a druggable long non-coding RNA for targeted anti-cancer approaches. J Hematol Oncol. 2018;11(1):63.

Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. 2017;18:1–14.

Becht E, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol. 2016;17:1–20.

Bhat AA, et al. MALAT1: a key regulator in lung cancer pathogenesis and therapeutic targeting. Pathol Res Pract. 2024;253:154991.

Breiman L. Random forests. Mach Learn. 2001;45:5–32.

Cao R, et al. Immune-related long non-coding RNA signature identified prognosis and immunotherapeutic efficiency in bladder cancer (BLCA). Cancer Cell Int. 2020;20:276.

Chapuy B, et al. Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. Nat Med. 2018;24(5):679–90.

Charoentong P, et al. Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. Cell Rep. 2017;18(1):248–62.

Chen L, Min J, Wang F. Copper homeostasis and cuproptosis in health and disease. Signal Transduct Target Ther. 2022;7(1):378.

Finotello F, et al. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. Genome Med. 2019;11:1–20.

Frankish A, et al. GENCODE: reference annotation for the human and mouse genomes in 2023. Nucleic Acids Res. 2023;51(D1):D942-d949.

Harrington F, et al. Genomic characterisation of diffuse large B-cell lymphoma. Pathology. 2021;53(3):367–76.

Jaffe ES. Diagnosis and classification of lymphoma: impact of technical advances. Semin Hematol. Elsevier; 2019. https://doi.org/10.1053/j.seminhematol.2018.05.007.

Kim J, et al. Long noncoding RNA MALAT1 suppresses breast cancer metastasis. Nat Genet. 2018;50(12):1705–15.

Kursa MB, Rudnicki WR. Feature selection with the Boruta package. J Stat Softw. 2010;36:1–13.

Lacy SE, et al. Targeted sequencing in DLBCL, molecular subtypes, and outcomes: a Haematological Malignancy Research Network report. Blood. 2020;135(20):1759–71.

Leek JT, et al. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28(6):882–3.
Li B, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biol. 2016;17:1–16.
Li SR, Bu LL, Cai L. Cuproptosis: lipoylated TCA cycle proteins-mediated novel cell death pathway. Signal Transduct Target Ther. 2022;7(1):158.
Lin T, et al. Emerging roles of p53 related lncRNAs in cancer progression: a systematic review. Int J Biol Sci. 2019;15(6):1287–98.
Liu Y-T, et al. Dysregulated Wnt/β-catenin signaling confers resistance to cuproptosis in cancer cells. Cell Death Differ. 2024;31(11):1452–66.
Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101. 2017. https://doi.org/10.48550/arXiv.1711.05101.
Lu X, et al. Elesclomol loaded copper oxide nanoplatform triggers cuproptosis to enhance antitumor immunotherapy. Advanced Science. 2024;11(18):2309984.
Luo Y, et al. Non-coding RNAs in breast cancer: implications for programmed cell death. Cancer Lett. 2022;550:215929.
Newman AM, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12(5):453–7.
Ning S, et al. Type-I AIE photosensitizer loaded biomimetic system boosting cuproptosis to inhibit breast cancer metastasis and rechallenge. ACS Nano. 2023;17(11):10206–17.
Ouyang W, et al. A prognostic risk score based on hypoxia-, immunity-, and epithelialto-mesenchymal transition-related genes for the prognosis and immunotherapy response of lung adenocarcinoma. Front Cell Dev Biol. 2022;9:758777.
Ouyang W, et al. Unraveling the unfolded protein response signature: implications for tumor immune microenvironment heterogeneity and clinical prognosis in stomach cancer. Aging (Albany NY). 2024;16(9):7818–44.
Racle J, et al. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. eLife. 2017;6:e26476.
Schmitz R, et al. Genetics and pathogenesis of diffuse large B-cell lymphoma. N Engl J Med. 2018;378(15):1396–407.
Shadman M, et al. Autologous transplant vs chimeric antigen receptor T-cell therapy for relapsed DLBCL in partial remission. Blood. 2022;139(9):1330–9.
Skeparnias I, et al. Structural basis of MALAT1 RNA maturation and mascRNA biogenesis. Nat Struct Mol Biol. 2024;31(11):1655–68.
Smith LN, Topin N. Super-convergence: very fast training of neural networks using large learning rates. In: Artificial intelligence and machine learning for multi-domain operations applications. SPIE. 2019. https://doi.org/10.48550/arXiv.1708.07120.
Tan M, et al. Dihydromyricetin induced lncRNA MALAT1-TFEB-dependent autophagic cell death in cutaneous squamous cell carcinoma. J Cancer. 2019;10(18):4245.
Tong X, et al. Targeting cell death pathways for cancer therapy: recent developments in necroptosis, pyroptosis, ferroptosis, and cuproptosis research. J Hematol Oncol. 2022;15(1):174.
Tsvetkov P, et al. Copper induces cell death by targeting lipoylated TCA cycle proteins. Science. 2022;375(6586):1254–61.
Wang Y-X, Zhang Y-J. Nonnegative matrix factorization: a comprehensive review. IEEE Trans Knowl Data Eng. 2012;25(6):1336–53.
Wang Q-M, et al. LncRNA MALAT1 promotes tumorigenesis and immune escape of diffuse large B cell lymphoma by sponging miR-195. Life Sci. 2019;231:116335.
Wright GW, et al. A probabilistic classification tool for genetic subtypes of diffuse large B cell lymphoma with therapeutic implications. Cancer Cell. 2020;37(4):551-568. e14.
Yang M, et al. A novel signature to guide osteosarcoma prognosis and immune microenvironment: cuproptosis-related lncRNA. Front Immunol. 2022;13:919231.
Yin Y, Butler C, Zhang Q. Challenges in the application of NGS in the clinical laboratory. Hum Immunol. 2021;82(11):812–9.
Yoshihara K, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013;4(1):2612.
Yu Y, et al. Artificial intelligence-based multi-modal multi-tasks analysis reveals tumor molecular heterogeneity, predicts preoperative lymph node metastasis and prognosis in papillary thyroid carcinoma: a retrospective study. Int J Surg. 2025;111(1):839–56.
Zelenetz AD, et al. NCCN guidelines(R) insights: B-cell lymphomas, version 6.2023. J Natl Compr Canc Netw. 2023;21(11):1118–31.
Zeng D, et al. IOBR: multi-omics immuno-oncology biological research to decode tumor microenvironment and signatures. Front Immunol. 2021;12:687975.
Zhakparov D, et al. GeneSelectR: an R package workflow for enhanced feature selection from RNA sequencing data. bioRxiv. 2024. https://doi.org/10.1101/2024.01.22.576646.
Zhang C, Huang T, Li L. Targeting cuproptosis for cancer therapy: mechanistic insights and clinical perspectives. J Hematol Oncol. 2024;17(1):68.
Zheng P, et al. Elesclomol: a copper ionophore targeting mitochondrial metabolism for cancer therapy. J Exp Clin Cancer Res. 2022;41(1):271.
Zhou M, et al. Discovery and validation of immune-associated long non-coding RNA biomarkers associated with clinically molecular subtype and prognosis in diffuse large B cell lymphoma. Mol Cancer. 2017;16(1):16.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.