Current Science
CS 6 (5), 6625-6642 (2026)
Received 08 August 2026 | **Accepted 02 September** 2026 | **Online Available 03 September** 2026
https://doi.org/10.52845/CS/2026-6-5-3

Current Science

Current Science

OPEN ACCESS

ISSN(O)2795-8639

Original Article

A New Method Based on Feature Alignment for Multi-Center Machine Learning Survival Prediction in Nasopharyngeal Cancer Patients

Song Wu[1,2*], Jing Chen[3*], Wenhui Lu[3*],Yuekun Wei[3], Qiong Song[4,5#] ,Daizheng Huang[3,6#],Chao Huang[1#]

[1]School of Information and Management, Guangxi Medical University, Nanning, 530021, China

[2]Cancer Hospital & Guangxi Cancer Institute, Guangxi Medical University, Nanning, 530021, China

[3]School of Life Sciences and Medical Engineering, Guangxi Medical University, Nanning, 530021, China

[4]Key Laboratory of Longevity and Aging-related Diseases of Chinese Ministry of Education, Institute of Neuroscience and Guangxi Key Laboratory of Brain Science, School of Basic Medical Sciences, Guangxi Medical University, Nanning, 530021 China

[5]Guangxi Health Commission Key Laboratory of Basic Research on Brain Function and Disease, Nanning, 530021, China

[6]Guangxi Institute of Medical Artificial Intelligence, Nanning, 530021, China

***Corresponding Author: Qiong Song, Daizheng Huang, Chao Huang**

Abstract:

Purpose In the survival prediction of multi-center nasopharyngeal carcinoma patients, the inconsistency of data characteristics across different centers significantly affects the predictive performance of data-driven machine learning methods. This study explores how to improve the effectiveness and consistency of multi-center data through feature alignment and fusion techniques, thereby optimizing the performance of survival prediction models.

Methods We collected two sets of nasopharyngeal cancer data sets.After feature alignment, LASSO regression was used to screen the features in the aligned latent space to select the most discriminative features. Subsequently, we introduced 15 mainstream machine learning algorithms for survival prediction model training, selecting multidimensional metrics such as AUC, F1-score, Accuracy, and Recall for evaluation.Additionally, to validate the superiority of the method, we applied Pearson correlation and the Boruta algorithm to train the aforementioned 15 machine learning models separately on both datasets and compared the performance results. Combining the encoder weights of an Autoencoder, the SHAP importance is back-projected from the latent space to the original feature variables to analyze the factors influencing the survival of NPC patients.

Results All evaluation metrics (AUC, F1 score, precision) exceeding 0.97.This approach successfully mitigated multi-centre data heterogeneity while validating the robust generalisation capability and stability of these models within multi-centre data environments.SHAP interpretability analysis revealed that the feature combinations of gender + time, age + time, and primary site + age exhibited the broadest SHAP value distributions, representing the most influential feature combinations affecting survival risk prediction.

Conclusion The data heterogeneity issue associated with multi-center NPC survival prediction was effectively resolved by combining autoencoder-based feature alignment with adversarial training.

Keywords: Nasopharyngeal carcinoma; Survival prediction; Autoencoder; Adversarial training; Machine learning

Introduction

A malignant tumor that starts in the nasopharyngeal mucosa's epithelium is called nasopharyngeal carcinoma (NPC). Individual prognoses still varied greatly, with a five-year overall survival rate ranging from 60% to 80% [1], despite the fact that improvements in immunotherapy, chemotherapy, and radiation therapy have greatly increased patient survival rates. For the purpose of directing clinical decision-making and creating individualized treatment plans, accurate survival prediction is crucial in the field of NPC treatment [2].

Prior research on predicting NPC survival has relied on individual datasets, including biomarkers, targeted tumor microenvironment (TME), chemoradiotherapy, and medical imaging [3-5]. In an effort to increase patient survival rates, multiple research efforts investigate the prognostic association of nasopharyngeal carcinoma through histology, biomarkers, and chemoradiotherapy, mostly depending on the knowledge of doctors or basic statistical models. Mao et al., for instance, obtained 21 peripheral blood values from patients with NPC. They proved that peripheral blood biomarker scores can successfully differentiate between NPC patients with various prognoses using a Cox regression model [6].However, bringing together data from two sources necessitates that patient data from both hospitals be comparable, with consistent data features, since predictions based on single-center data lack population adaptability and do not reflect the model's generalizability. Direct data integration for joint predictive modeling is hampered by the substantial differences in sample feature dimensions and method of coding that result from variations in hospital protocols, data collection standards, and influencing factors.

The majority of the present study uses data on NPC patients from the SEER database, which spans several US states and provides strong representativeness and generalizability. Local hospitals, on the other hand, concentrate on patients from high-incidence areas and offer comprehensive imaging metrics like maximal short diameter and diagnostic measures like lactate dehydrogenase (LDH, ALB).Is it feasible to combine two sets of data to create a prediction model that has both local accuracy and broad adaptability? In the study by Liu et al., radiomics, deep learning, and clinical variables were combined to create a recurrence prediction model for NPC based on MRIB. An autoencoder was used to extract deep features from MRI, and the model's prediction performance set a new record [7]. An innovative method and instrument for maximizing individualized treatment plans for nasopharyngeal carcinoma is the use of autoencoders.

An unsupervised neural network model that can directly learn features from data is called an autoencoder. By learning low-dimensional representations of participants, Shen et al., for example, created an advanced deep learning autoencoder survival analysis model (AESurv) that correctly forecasts coronary heart disease by analyzing high-dimensional blood DNA methylation characteristics and conventional clinical risk factors [8].Research on NPC survival prediction is still comparatively lacking, despite the fact that autoencoders have shown notable results in a variety of data sources, including audio, text, video, and multiple medical sectors, with some success in tumor survival prediction applications. To the greatest of our knowledge, very few researchers have predicted cancer survival rates using multi-group heterogeneous data. There is disagreement in the machine learning community over the application of low-dimensional feature mapping potential space technology for NPC survival prediction. There is considerably less related research.

Data on NPC patients with different characteristics were gathered from two centers in this study. Autoencoder and adversarial training methods were utilized to align cross-center features, and 15 popular machine learning techniques were employed to train survival prediction models. In an effort to give a reference for multi-center machine learning survival prediction, we evaluated using multidimensional indicators.

2. Materials and Methods

An overview of the overall technical approach used in this study is shown in Figure 1.

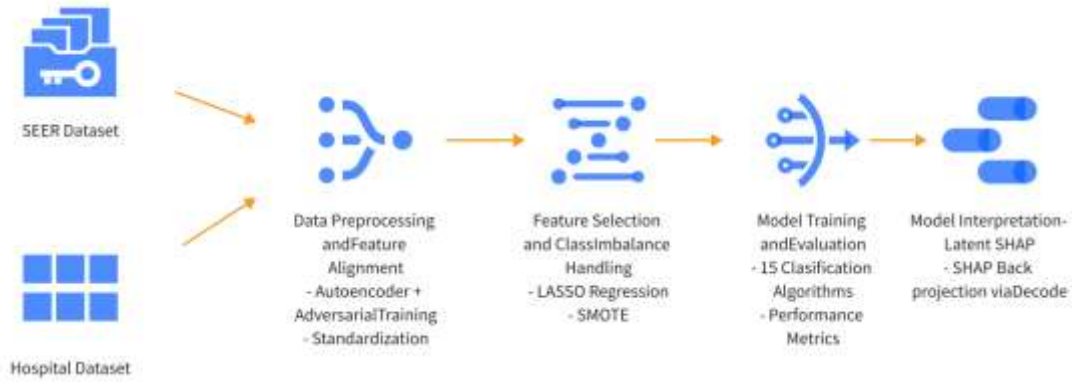

Figure 1 | Overall technical approach

2.1 Data Sources

The primary source of data includes 1,630 NPC patients who received treatment at Guangxi Medical University's Affiliated Cancer Hospital between 2015 and 2022. A rich information base for creating a precise survival prediction model is provided by the data, which includes biochemical indicators, tumor stage, treatment plan, pathological kind, basic patient information, and follow-up results. This retrospective study was approved by the Institutional Review Board of Guangxi Medical University (No.2023-KY0003). Every patient acquired an ethical evaluation from the hospital and signed informed consent documents.

The SEER database (https://seer.cancer.gov), which includes information on 1,502 eligible patients with NPC diagnoses between 2000 and 2015, served as the second data source. This database has been made accessible. The baseline characteristics of patients from the hospital and SEER databases are displayed in Tables 1 and 2, respectively.

2.2 Data Preprocessing

To determine the patients' baseline information, variables with more than 20% missing values were eliminated, and each missing data point in the dataset was imputed using the mean value. Along with demographic details like gender, age, and marital status, the information also included the tumor's primary location, histological grade, stage, and TNM staging, as well as whether chemotherapy and radiation therapy were administered during the course of treatment, the order in which they were administered, whether the tumor had spread, and the precise location of metastases. To deal with imbalance in survival status category data, apply the SMOTE approach.

2.3 Feature Alignment and Embedding

We employ encoders to transform the deep features of two datasets. As we know, the purpose of dimensionality reduction is to obtain a simpler representation of the data.Autoencoders are unsupervised neural network models—comprising an input layer, one or more hidden layers, and an output layer.This structure makes it highly flexible, allowing for the free addition or

removal of layers. Moreover, it can simultaneously capture complex relationships at both deep and surface levels within the data. Encoders operate through a clever mechanism that transforms high-dimensional input data into low-dimensional encoded representations. Typically, the dimension is significantly lower than that of the input data, thereby achieving effective dimensionality reduction.Finally, we will attempt to reconstruct the original features from this compressed encoding.The model continuously compares the discrepancy between input and output, measuring the degree of this deviation through loss functions such as mean squared error or cross-entropy. Subsequently, it adjusts its internal parameters to progressively reduce this deviation. The network weights ultimately retained essentially constitute the learned "extraction rules."

In our study, the core task of the autoencoder was to compress and map high-dimensional features from different centers—which vary like dialects—into the same low-dimensional latent space. In this space, the data no longer emphasizes its origin but instead re-expresses itself using a common "language," quietly paving the way for subsequent analysis.Commonly used loss functions include Mean Squared Error (MSE) or cross-entropy loss.

$$Loss = \sum_{n=1}^{N} \|\mathbf{bold}^n - \text{Decoder}(\text{Encoder}(\mathbf{bold}^n))\|$$

Where bold ^{n} denotes the n-th input sample (usually a bold vector), Encoder and Decoder are the encoder and decoder respectively, and $\|\cdot\|$ denotes the norm (usually the L_2 norm). Let N represent the sample size of the data. In the encoding part of the autoencoder, we designed a shared encoder—all data, regardless of which center they come from, are processed through the same encoding mechanism.However, this alone is not sufficient; underlying differences in the data may still persist. We introduced adversarial training, using a discriminator network to determine the sample source, thereby forcing the encoder to learn source-invariant common feature representations.

Ultimately, information from different data sources is mapped into a unified latent feature space and described within that space using a common "language" for model interpretation. This integrated feature space becomes the true foundation for subsequent survival prediction models to conduct analysis.

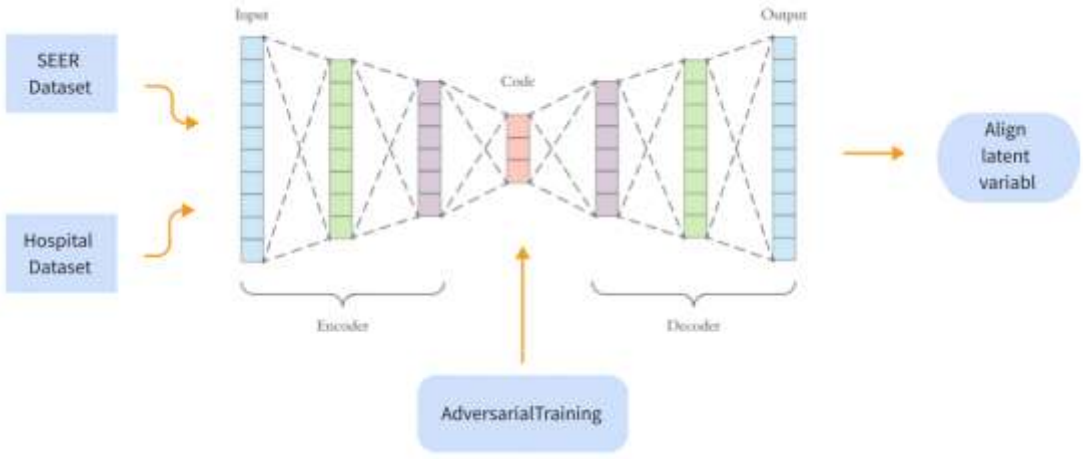

Figure 2 | The autoencoder training framework comprises two core components: an encoder and a

decoder. The encoder compresses two sets of data into a unified low-dimensional space, while the decoder reconstructs the original information from it. Simultaneously, adversarial networks are integrated to align features from different data sources.

2.4 Feature Selection

We additionally utilize LASSO regression for feature selection in the aligned latent space in order to avoid high-dimensional low-information features that could impair modeling performance. In order to reduce the cross-validation error, the regularization parameter (λ) is adjusted to select a group of latent variables with the maximum discriminative power. The final steps include restoring semantic features, visualizing selected results using feature coefficient diagrams, reflecting low-dimensional space variables using encoder weights, and using the results for further classification model building.

2.5 Machine Learning Models and Performance Evaluation

We additionally utilize LASSO regression for feature selection in the aligned latent space in order to avoid high-dimensional low-information features that could impair modeling performance. In order to reduce the cross-validation error, the regularization parameter (λ) is adjusted to select a group of latent variables with the maximum discriminative power. The final steps include restoring semantic features, visualizing selected results using feature coefficient diagrams, reflecting low-dimensional space variables using encoder weights, and using the results for further classification model building.

AUC, F1-score, Accuracy, Recall, Specificity, Brier Score, and ROC curve are among the multi-dimensional metrics used in the test set performance evaluation to methodically assess the survival prediction skills of various machine learning algorithms. A thorough comparison of different evaluation measures is employed to choose the survival classification model with the highest overall performance score.

To further investigate the important aspects impacting patient prognosis, an interpretable analysis of the parameters affecting survival prediction in patients with nasopharyngeal cancer was conducted using the SHAP [9] model. The "semantics" of the features are lost when the autoencoder maps the original features to a low-dimensional latent space, rendering SHAP interpretability useless or challenging to define. For bettering the interpretability of the model, we backproject the SHAP significance from the latent space to the original feature variables by combining the autoencoder encoder weights.

Table 1 Baseline characteristics of nasopharyngeal carcinoma patients in a tertiary hospital in Guangxi

Characteristic	Description (n=1630)	Percentage (%)
Sex		
Male	1205	(73.9%)
Female	425	(26.1%)
Age (years)	45.80±11.07	
Age group		
<40	480	(29.4%)
40-49	531	(32.6%)
50-59	429	(26.3%)
60-69	171	(10.5%)
≥70	19	(1.2%)

T		
1	28	(1.7%)
2	385	(23.6%)
3	750	(46.0%)
4	466	(28.6%)
Missing	1	(0.1%)
N		
0	41	(2.5%)
1	545	(33.4%)
2	607	(37.2%)
3	436	(26.7%)
Missing	1	(0.1%)
stage		
3	790	48.5%
4	839	51.5%
Missing	1	0.1%
number		
<5	581	(35.6%)
≥5	1048	(64.3%)
Missing	1	(0.1%)
grouping		
0	1054	(64.7%)
1	575	(35.3%)
Missing	1	(0.1%)
Survival status		
Alive	1200	(73.6%)
Lost to follow-up	160	(9.8%)
Deceased	270	(16.6%)
Recurrence		
No	1588	(97.4%)
Yes	42	(2.6%)
Metastasis		
No	1497	(91.8%)
Yes	133	(8.2%)

Table 2 Characteristics of nasopharyngeal carcinoma patients in SEER

Category	**Group**	**n**	**%**
Age (years)			
	<40	244	14.90
	40-49	310	18.93
	50-59	480	29.30
	60-69	369	22.53
	>=70	235	14.35
Sex			
	Male	1143	69.78
	Female	495	30.22
Primary Site			
	Superior wall of nasopharynx	28	1.71
	Posterior wall of nasopharynx	210	12.82

	Lateral wall of nasopharynx	149	9.10
	Anterior wall of nasopharynx	18	1.10
	Overlapping lesion of nasopharynx	67	4.09
	Nasopharyngeal carcinoma, NOS	1166	71.18
Histological Grade			
	Well differentiated; Grade I	45	2.75
	Moderately differentiated; Grade II	204	12.45
	Poorly differentiated;Grade III	682	41.64
	Undifferentiated; Grade IV	707	43.16
Stage			
	I	155	9.46
	II	360	21.98
	III	500	30.53
	IV	623	38.03
T Stage			
	T1	593	36.20
	T2	324	19.78
	T3	338	20.63
	T4	383	23.38
N Stage			
	N0	401	24.48
	N1	538	32.84
	N2	475	29.00
	N3	224	13.68
M Stage			
	M0	1519	92.74
	M1	119	7.26
Radiotherapy			
	No	11	0.67
	Yes	1627	99.33
Chemotherapy			
	No	212	12.94
	Yes	1426	87.06
Bone Metastasis			
	No	1584	96.70
	Yes	54	3.30
Brain Metastasis			
	No	1629	99.45
	Yes	9	0.55
Liver Metastasis			
	No	1620	98.90
	Yes	18	1.10
Lung Metastasis			
	No	1597	97.50
	Yes	41	2.50
Sequence of Surgery and Radiotherapy			
	No radiotherapy or surgery	1127	68.80
	Postoperative radiotherapy	495	30.22
	Preoperative radiotherapy	13	0.79

	Both pre- and postoperative radiotherapy	3	0.18
Race			
	White	761	46.46
	Black	174	10.62
	Chinese	355	21.67
	South and Southeast Asian	174	10.62
	Pacific Islander	74	4.52
	Other Asian	100	6.11
Marital Status			
	Divorced	171	10.44
	Married	1029	62.82
	Single	352	21.49
	Widowed	86	5.25

3. Results

3.1 Feature Embedding Alignment and Feature Selection

Table 1 and Table 2 present baseline data for nasopharyngeal carcinoma patients derived from hospital records and the SEER database respectively. To integrate these heterogeneous datasets, we employ autoencoders to map the extracted features into a unified latent space.To achieve feature consistency, we introduced a conflict learning mechanism. Specifically, we trained a discriminative network to identify the origin of features in the database, thereby guiding the feature extraction network and making it harder for the discriminator to distinguish between features. This approach significantly reduced the distribution imbalance between the two datasets in the latent space.After feature alignment, we employ a random forest algorithm for modeling. This model is trained using unified latent space features obtained through feature fusion. By calculating the importance score for each feature, variables with actual influence in the model can be clearly identified.Figure 3 clearly illustrates the relative contribution ranking of each raw feature in the prediction, revealing at a glance which information maintains robustness across all database analyses.

Through nonlinear transformations, the autoencoder encodes raw features such as gender, age, and tumor size into vectors within a low-dimensional space. The original explicit clinical “semantics” are effectively lost, meaning the latent variables no longer maintain a one-to-one correspondence with the original features. They are abstracted into the form latent_x and latent_y.This may affect clinicians' judgment of key factors influencing patient prognosis.Another issue is model instability. Due to random weight initialization and variations in the learning sequence, different results are produced each time the model is run.This leads to discrepancies in the prioritization of importance during the learning iteration process. This uncertainty directly undermines the reliability of SHAP analysis and prevents its conclusions from being directly applied in clinical settings.To this end, we incorporated a random seed mechanism into the experiment to ensure consistency of results across each iteration of learning.

As shown in Figure 3a, each dimension in the latent space does not correspond to a single clinical indicator but represents a composite mapping formed by multiple raw features. For example, the dimension labeled “Gender + Survival Period + Tumor Malignancy (L34)” actually integrates the combined effects of three factors: patient gender, survival period, and tumor malignancy.The combination of these features exhibits a pronounced long-tail distribution. This implies that a very small number of dimensions

account for the overwhelming majority of importance weights. Combinations such as "gender + time + score" or "age + time + dominant hand" receive significantly higher weight scores than other variables.This may indicate that, from the model's "perspective," the combination of male gender, shortened survival time, and high tumor malignancy often correlates with poor prognosis. Conversely, advanced age, shortened survival time, and lateralization of specific tumors may constitute another set of valuable predictors. The importance scores for these underlying characteristics primarily range between 0.04 and 0.15.Next, we employed LASSO regression for feature selection, with results shown in Figure 3b. Post-screening analysis revealed that the combination of "Gender + Survival Period + Malignancy Level (L34)" demonstrated significant importance with a score exceeding 0.2. This further confirms that the feature formed by the interaction between patient gender, survival period, and tumor malignancy plays a crucial role in survival classification prediction.Additionally, several other combinations maintain relatively high importance scores. Together, these combinations form a compact and effective cluster of predictors that quietly underpin the model's classification decisions.

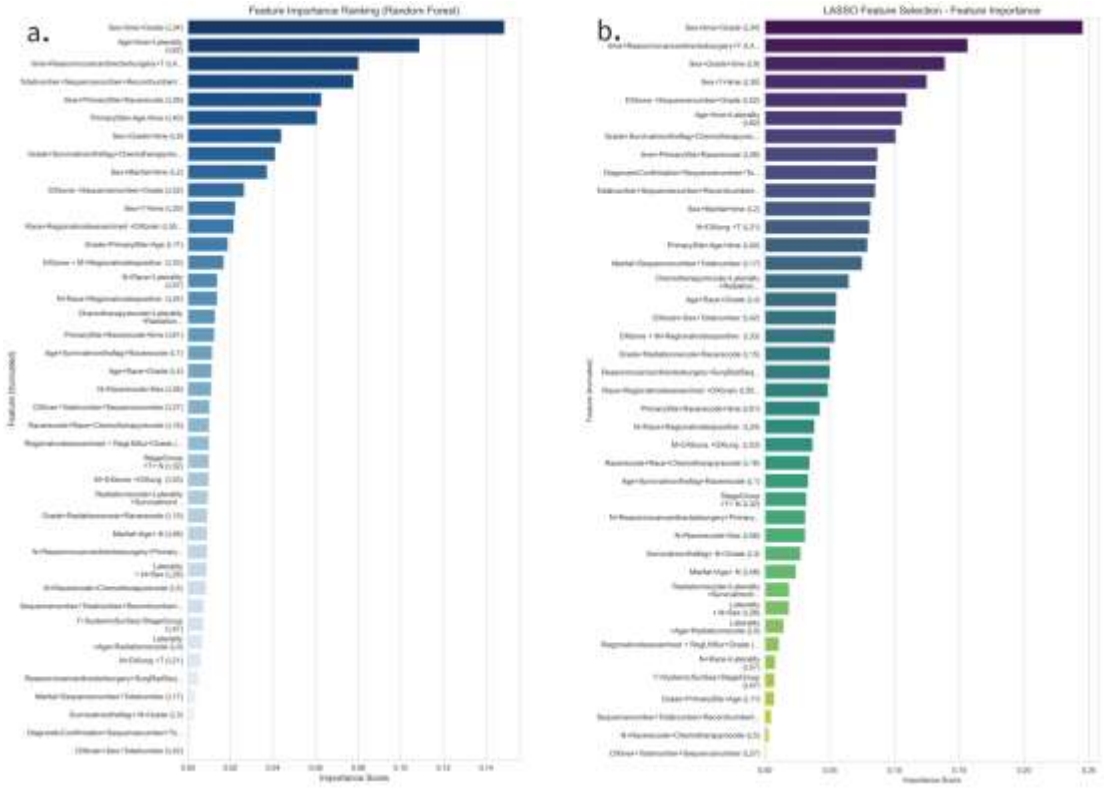

Figure 3 | a. The low-dimensional "latent vectors" in latent space reflect composite features. Feature importance bar charts generated by the random forest model indicate contribution levels through color intensity: darker colors denote higher importance of the feature in the random forest model's output. b. Feature selection via LASSO regression visualizes the importance ranking of composite features.

3.2 Classification Model Evaluation

We divided the preprocessed data features into training and test sets at a ratio of 7:3, constructing a total of 15 machine learning models. Figure 4 presents a comprehensive comparison of different

models' performance across multiple metrics. As shown in the bar chart (Figure 4a) and radar chart (Figure 4c), ensemble learning models demonstrate overall superiority over traditional models in terms of feature selection capability.Among these methods, three ensemble learning approaches (XGBoost, Boosted Trees, Random Forests) achieved particularly outstanding results: their average AUC values all exceeded 0.98, while their F1 scores and accuracy rates approached 0.97. This demonstrates that these models possess exceptional classification capabilities and stability.XGBoost demonstrated outstanding performance, achieving an astonishing AUC value of 0.995 while stabilising the F1 score and maintaining an accuracy of 0.960. Naturally, this excellent result established it as the benchmark model for subsequent analyses.By contrast, models such as Decision Tree, Naive Bayes and KNN performed poorly across all metrics.The ROC curve (Figure 4b) more clearly demonstrates the outstanding performance of the XGBoost (AUC=0.995), Extra Trees (AUC=0.998), and Voting Classifier (AUC=0.998) models. Calibration curve results (Figure 4d) indicate that the closer the curve is to the diagonal line, the higher the calibration effectiveness. The predicted probabilities obtained by Random Forest, Logistic Regression, and Support Vector Machine highly align with actual probabilities (Brier scores of 0.045, 0.024, and 0.021, respectively). This indicates that the probabilistic outputs from these models are reliable and suitable for tasks requiring probabilistic judgments, such as survival probability prediction and risk stratification.Based on a comprehensive analysis of performance metrics, XGBoost, Random Forest, and X-Tree models not only demonstrate outstanding performance but also exhibit distinct advantages in stability and probability calibration. Therefore, these models will naturally become our recommended solutions for subsequent detailed modeling and interpretability analysis.

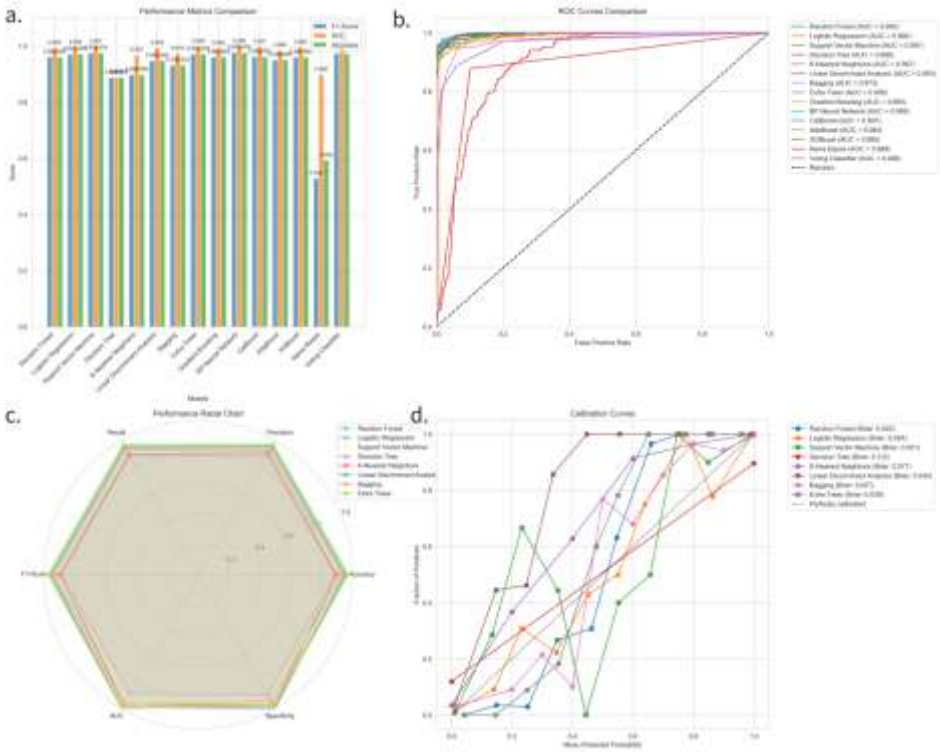

Figure 4| a. Bar chart showing F1, AUC, and accuracy metrics for 15 classification models. b. ROC curves for multiple classification models. c. Radar chart for multiple classification models; the closer to the hexagon, the more comprehensive and stable the performance. d. Calibration curve results.

We used the original features without feature transformation to do machine learning prediction of nasopharyngeal cancer in order to compare with the approach presented in this work. Initially the relationship between patient information components and survival status was examined using Pearson's correlation coefficient.The Boruta algorithm was used to choose the variables. There are now fifteen popular machine learning categorization models. The models' accuracy, F1 score, and AUC value were chosen as performance evaluation metrics. Using SEER data as an example, Figure 5 illustrates how Extra Trees, Random Forest, and CatBoost all performed well, placing in the top three.The XGBoost model achieved an AUC value of 0.767 while maintaining an accuracy of 0.721 and an F1 score of 0.688. The ROC value averaged about 0.7 in the ROC curve comparison chart. It is evident that autoencoder feature mapping in low-dimensional space has a considerably greater overall impact than machine learning techniques without feature conversion. In comparison to the SEER dataset, the hospital model performs noticeably better overall. Issues including disparities in feature representation, sample size imbalance, and data distribution could be the cause of this discrepancy.

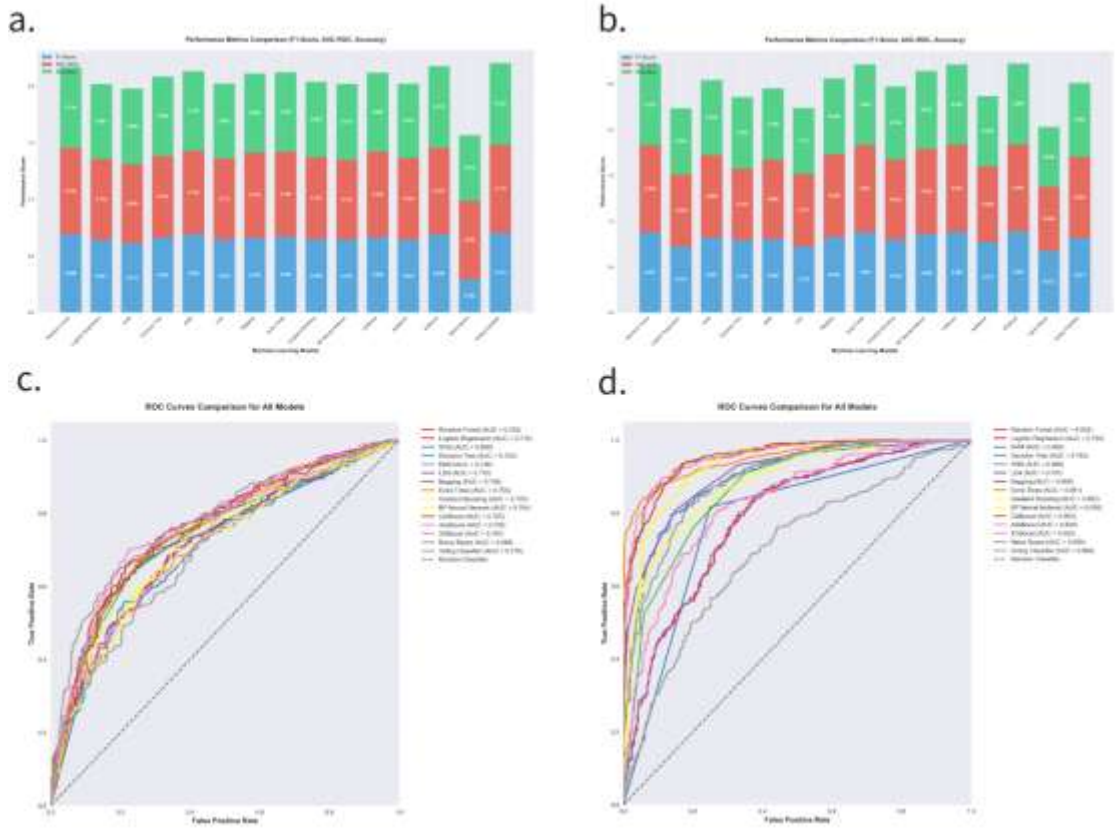

Figure 5 | Comparison of the performance of 15 models using the SEER database (a) and the corresponding AUC curves for each model (c). The same applies to (b) and (d), which show the results of data analysis related to hospital data.

The SHAP significance was back-projected from the latent space to the original feature variables by thoroughly comparing a number of evaluation indicators, which were based on the XGBoost survival prediction model and combined with the autoencoder encoder weights(Figure 6). The significance of each feature to the output of the XGBoost model was shown in an analysis chart

of key features influencing patient survival prognosis (Figure 7a). The numerical values indicate how important each feature is to the model decision, and the features are arranged by influence.The figure indicates that the model is significantly impacted by time and sex, which rank top and second, respectively. Third place goes to age, which also has a significant effect on the model because getting older is typically linked to a worse prognosis and possibly more noticeable variations in treatment response.

The high values observed in SHAP analysis for cancer staging are not surprising. This metric is a decisive factor in the model's survival risk assessment, reflecting its status as a clinically significant prognostic factor.Detailed analysis revealed that nearly all top-scoring features were associated with disease progression and sociodemographic factors. Characteristics such as primary tumor site (T), pathological malignancy (grade), clinical staging classification (stage group), marital status (marital), and ethnic background (ethnic) ranked prominently. This assessment system aligns closely with the experiential knowledge clinicians accumulate in daily practice.Figure 7b illustrates the SHAP value distributions for each 'latent variable' feature within the XGBoost survival prediction model.The model ranks the top 20 most important features by significance. Among them, the three feature combinations—gender + time, age + time, and primary site + age—exhibit the widest distribution of SHAP values, indicating greater influence. The model tends to regard these feature combinations as having a critical role in the prediction results.For example, the model indicates that the relatively short survival period observed in elderly patients may be significantly associated with poor prognosis.We have discovered that this is no longer attributable to a single indicator, but rather the result of multiple factors interacting over time. It may well be precisely the subtle risk signals that data models can capture, whose scope extends beyond what traditional experience can identify.

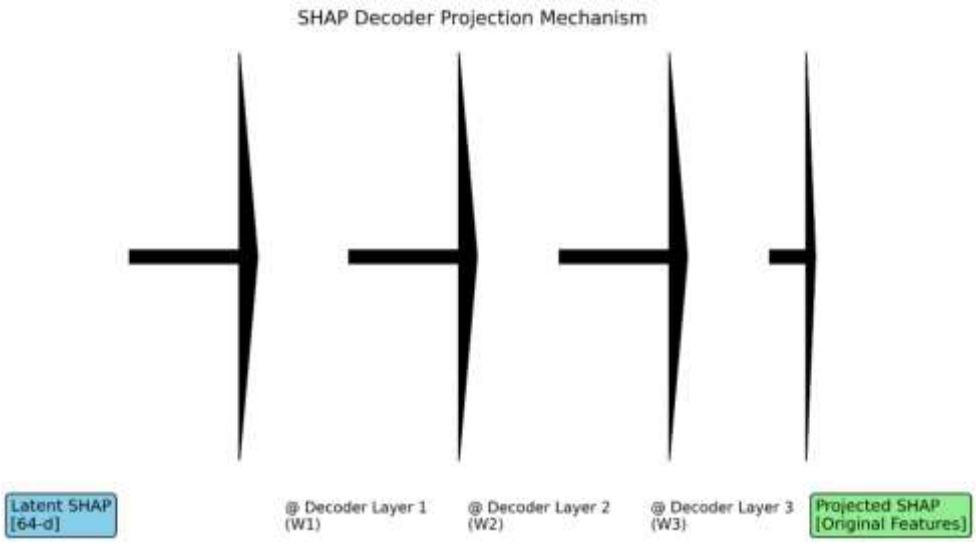

Figure 6 | Decoder-based SHAP value backprojection mechanism. SHAP values in the 64-dimensional latent space are projected back to the original input feature space through the decoder's weight matrix W1–W3 to restore the interpretability of clinical features.

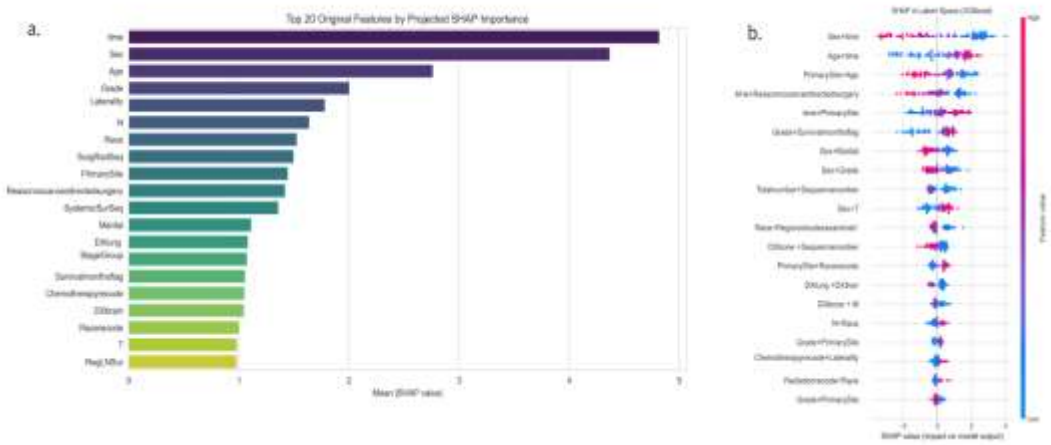

Figure 7 | a. Average SHAP value analysis chart based on XGBoost (average impact on model output amplitude). b. SHAP value analysis chart based on XGBoost (impact on model output). The distribution of positive and negative SHAP values reflects the directional impact of different combination values on risk prediction.

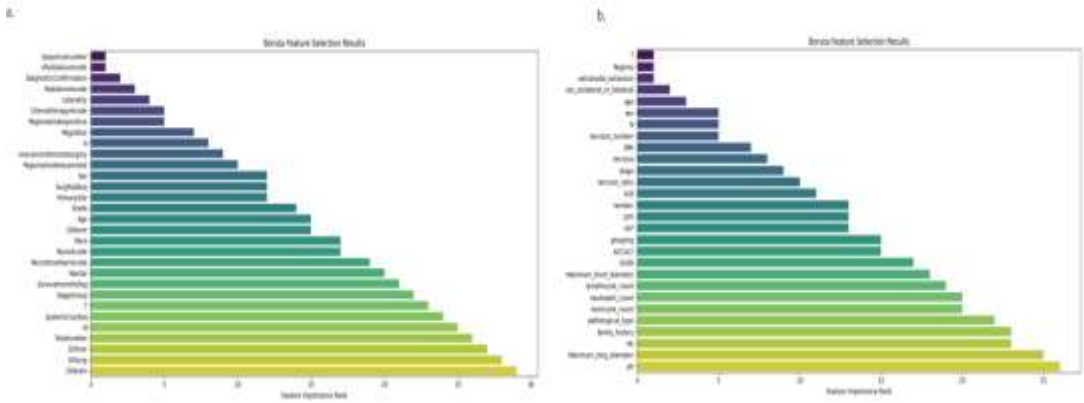

Figure 8 | Results of feature variable screening using the Boruta algorithm, left (a) for the SEER public dataset and right (b) for the hospital's own data.

4. Discussion

The fields of disease detection and prediction have found machine learning technology to be an essential tool in the current wave of big data-driven medical informatization. Despite their distinct benefits, ensemble learning techniques have proven to be very applicable in the prediction of complicated illnesses, such as chronic conditions.The SEER dataset has been used to assess the prognosis of young NPC patients both domestically and abroad [11], analyze survival rates in NPC based on histology [12], and detect early beginnings of nasopharyngeal carcinoma (NPC) in NPC [10]. Our research also demonstrates that clinical text aspects can be used to differentiate between disease survival forecasts for NPC patients.We discovered that the autoencoder can resolve the issue of heterogeneity in multi-center data by mapping the characteristics of the data to a low-dimensional unified latent space and then reconstructing the features. This approach works better than straight feature categorization, according to experiments. With no apparent connection to disease prognosis, the XGBoost model outperformed conventional techniques in the data preprocessing stage of mapping to a low-dimensional space, maintaining

a high F1 score (0.963) and accuracy (0.953) while attaining an AUC value of 0.996.

This study utilized data from 1,630 nasopharyngeal carcinoma patients at a tertiary hospital in Guangxi Zhuang Autonomous Region and 1,638 nasopharyngeal carcinoma patients from the SEER database. Both datasets exhibited similarities and differences in demographic characteristics, clinical staging, and treatment modalities.The SEER nasopharyngeal carcinoma dataset offers the advantage of encompassing multiple racial groups. Specifically, the data includes baseline information for 761 Caucasians, 174 African Americans, 355 Asians, as well as South Asians, Pacific Islanders, and other ethnic groups.This diversity is undoubtedly valuable for studying disease characteristics across different populations.However, the SEER database also has limitations, such as frequently lacking complete records detailing specific treatment regimens like radiation therapy or chemotherapy, which may significantly impact the analysis of patient survival outcomes.On the other hand, the hospital dataset (n = 1,630) exhibits high comprehensiveness, primarily focusing on regions with high incidence rates in southern China, where the ethnic composition is relatively homogeneous. However, its core value lies in its depth and dynamic nature. It comprehensively documents the entire patient journey from diagnosis and treatment to follow-up, accurately reflecting disease progression and subtle changes in patient physiological status through continuous dynamic monitoring and biochemical indicators. This dataset effectively compensates for the lack of clinical detail in SEER, enabling the development of more precise survival prediction models.However, this diversity itself presents practical challenges in application. When training a model simultaneously on two entirely distinct datasets, how can we ensure it captures common patterns while adapting to the unique characteristics of each data source? This may well be the fundamental contradiction we must confront and resolve in multicenter collaborative research.

To clarify differences between multi-institutional datasets, this study employed the Boruta algorithm to screen features across both datasets.As shown in Figure 8, the feature importance rankings in the XGBoost model constructed using SEER data and hospital data exhibit significant differences.In the SEER dataset, survival status recoding, age, histological grade of malignancy, and lymph node metastasis (N stage) are the primary predictors.This aligns with the predictive logic of the traditional TNM classification system, reflecting the database's greater emphasis on foundational demographic and pathological information. However, hospital data encompass richer biomarker characteristics, including laboratory indicators (body mass index (BMI), albumin (ALB), lactate dehydrogenase (LDH), etc.) and imaging findings (extracapsular lesions, etc.). These results reflect the heterogeneity of multicenter data and the complementary potential of multi-source information in predicting prognosis. In other words, the SEER system provides a reference model for large-scale populations, while regional data enriches it with detailed clinical information.Through efficient adjustment and integration of features, we aim to develop more versatile predictive models. These models maintain stability across diverse populations while better aligning with specific requirements in regional clinical settings, achieving a balance between universality and specificity.

In our analytical methodology, we first employ mean imputation to address missing values. Subsequently, we utilise autoencoders and adversarial training techniques to align the embedding vectors from both data sources within the latent space. This approach not only achieves dimensionality reduction but also mitigates the risk of overfitting to a certain extent.We

employed LASSO regression to screen key features and utilised the SMOTE method to address data imbalance, thereby enhancing model training stability. Through comprehensive comparison across multiple metrics using 15 classification models, the optimal predictive tool was ultimately selected. However, in actual clinical practice, clinicians focus more on factors influencing disease progression rather than survival outcomes. We applied latent space SHAP combined with a reverse projection mechanism to map the importance of latent features back to the original variables for analysis.Analysis of SHAP results indicates that the impact of TNM staging on prognosis aligns with clinical experience, confirming the biological plausibility of this model. This analysis reveals the contribution levels of primary risk factors such as TNM staging, age, and treatment sequence, with their ranked importance closely matching clinical observations. This provides an interpretable basis for developing personalised treatment strategies.Concurrently, this enables public health policymakers to better comprehend and apply model outcomes. Such analyses not only assist clinicians in understanding the inherent logic of the models but also empower public health decision-makers to utilise model results more effectively. Consequently, an interpretable link is established between data and decision-making, laying the groundwork for policy formulation, resource allocation, and patient care.

To improve feasibility, we employed the Boruta algorithm for variable screening, created a classification model with survival status as the output, established a correlation between patient information factors and survival status for the two sets of data using Pearson's method, and compared the models' performance. According to the trial, the autoencoder low-dimensional technology's survival prediction effect was more noticeable.

To assess external validity, the methodological framework was applied to diabetes patient cohorts derived from the UK Biobank (UKB) and the National Health and Nutrition Examination Survey (NHANES) databases for survival outcome prediction. The model achieved robust performance, with AUC exceeding 0.98 and F1-score above 0.97, demonstrating its ability to generalize across distinct disease entities and heterogeneous data sources.

In order to enable the features to use the "same language" for training 15 classification models, we employed multi-data autoencoders to reconstitute the features following low-dimensional mapping. During data preprocessing, autoencoders maintained important survival-related properties, as seen by the best model AUC values, which were approximately 0.97. Prior research has demonstrated that autoencoder technology has resolved the problems of multi-data heterogeneity and high-dimensional data.In an effort to lower the dimensionality of high-dimensional neural characteristics in the field of BCI, Ran et al. suggested a hybrid autoencoder model based on CNN and LSTM. This model achieves more efficient decoding by integrating various brain signal frequency bands while retaining task-related information [13]. In order to fully characterize miRNAs and diseases, Dai et al. conducted model correlation prediction for miRNAs and diseases, used autoencoders for dimension reduction to obtain the optimal feature space, and eventually obtained good model prediction results [14]. By embedding mapping consensus on multimodal data and retaining information, Sun et al. were able to get a semantic encoding vector. In order to lower reconstruction error and ultimately provide for the mutual recovery of image-text pairs, feature projections were restored to the source data [15].The above findings further support our belief that data preparation via dimension reduction mapping can improve accuracy in prediction by addressing clinical concerns with heterogeneous patient data

sources.

However,Despite the employment of adversarial networks and autoencoders for feature mapping and alignment, shifts in the latent space may also result from information disparities between the two sets of data. For instance, the ability to generalize may be strong on SEER data but weak on hospital data. To make the model more robust, we will also add more hospital data. Capturing data with linear relationships is the primary application of LASSO. Enhancing the attention mechanism's screening capabilities or employing a variety of feature selection techniques is worthwhile. Although SHAP value analysis makes the model easier to understand, its complicated nature still conflicts with the need for clinical decision-making to be straightforward, and additional validation is necessary for clinical translation. Converting prediction data into useful treatment routes requires interdisciplinary cooperation.

Future investigations may integrate multimodal data streams – encompassing genomics, radiomics, and clinical variables – to develop comprehensive survival prediction frameworks. For instance, radiomic features derived from CT or MRI scans, including texture descriptors and heterogeneity metrics, capture intratumoral spatial variation and thus complement anatomy-based TNM staging. Concurrently, longitudinal profiling of circulating tumor DNA (ctDNA) serves as a liquid biopsy-based sentinel, enabling early detection of minimal residual disease or incipient recurrence prior to radiological manifestation. Employing multi-task learning architectures, such models could jointly estimate overall survival (OS), progression-free survival (PFS), and quality-of-life (QoL) outcomes, thereby providing a multidimensional evaluative platform for precision oncology.

Advancing further, the integration of reinforcement learning could support adaptive treatment decision-making. Such a system would enable dynamic optimization of therapeutic strategies—including chemotherapy and radiation dosing—in response to real-time patient-specific physiological and treatment response data, thereby establishing a self-refining, closed-loop framework for personalized care. Ultimately, the clinical utility of any model must be empirically verified; therefore, prospective cohort studies are warranted to rigorously assess the real-world effectiveness of these predictive tools in routine healthcare environments.A pragmatic implementation strategy would integrate the optimized XGBoost model within the hospital information infrastructure to facilitate real-time risk stratification and automated alerts for intensified monitoring or intervention in high-risk cohorts. Concurrently, decision-curve analysis should be employed to quantify the incremental clinical net benefit of the model relative to conventional prognostic schemas such as AJCC staging—specifically, whether it meaningfully enhances the timeliness and precision of clinical decision-making. This comparative utility constitutes a critical metric for evaluating its translational impact.

5. Conclusion

The present study investigated a survival prediction paradigm centered on enhanced data preprocessing. Through the projection of high-dimensional clinical variables into a condensed latent representation via autoencoders, we achieved both improved prognostic accuracy in nasopharyngeal carcinoma and reduced sensitivity to inter-center heterogeneity. The proposed framework demonstrates inherent extensibility and is conceptually transferable to analogous predictive modeling challenges across diverse oncological contexts.A distinctive aspect of this work lies in extending the deep learning-based feature abstraction process beyond opaque prediction. By integrating latent-space SHAP interpretation with back-projection to original clinical variables, we bridge the abstract

representation with medically meaningful predictors. This design not only preserves model performance but also introduces an interpretable interface—effectively enabling the model to "communicate" its reasoning in clinically intelligible terms.Clinically speaking, significant work remains ahead. Yet this approach hints at a promising direction: bridging high-dimensional data and clinical decisions need not force a trade-off between predictive accuracy and interpretability. Persistent exploration along this line may ultimately yield methodologies that harmonize model performance with clinical trust—a dual achievement essential for real-world integration.

Acknowledgments

This research was supported by Guangxi Key Research and Development Program No. FN2600640239, the National Natural Science Foundation of China (Grant No. 62341601 and Grant No. 82160482).

The authors declare no conflict of interest.

Author Contributions

Song Wu: Conceptualization, Data curation, Formal analysis, Investigation, Methodology. Jing Chen: Conceptualization, Data curation, Formal analysis. Wenhui Lu: Project administration, Resources, Supervision, Validation. Yuekun Wei: Visualization, Writing – review & editing. Daizheng Huang: Investigation, Writing – review & editing. Chao Huang: Investigation, Writing – review & editing.Qiong Song:Investigation, Writing – review

Data Availability Statement

The data that support the findings of this study are openly available in the Surveillance, Epidemiology, and End Results (SEER) database at https://seer.cancer.gov. The clinical data from Guangxi Medical University's Affiliated Cancer Hospital are not publicly available due to patient privacy restrictions but may be requested from the corresponding author upon reasonable request and with permission from the institution's ethics committee.

Ethics Statement

This retrospective study was approved by the Institutional Review Board of Guangxi Medical University (No.2023-KY0003).

References

1. Lee AW, et al. Management of nasopharyngeal carcinoma: current practice and future perspective. J Clin Oncol. 2015; 33 (29):3356-3364.
2. Chen Zihong, Zhong Qiang, Gao Jianquan, et al. Survival prognosis curve for elderly nasopharyngeal carcinoma patients based on the SEER database [J]. Chinese General Practice, 2022, 20(03): 487-492.
3. Luo HD, Xia FJ, Wu JH, Yi B. Efficacy of chemoradiotherapy in survival of stage IV nasopharyngeal carcinoma and establishment of a prognostic model. Oral Oncol. 2022; 13 1:105927. doi:10.1016/j.oraloncology.2022.105927
4. Chen YP, Yin JH, Li WF, et al. Single-cell transcriptomics reveals regulators underlying immune cell diversity and immune subtypes associated with prognosis in nasopharyngeal carcinoma. Cell Res. 2020;30(11):1024-1042. doi:10.1038/s41422-020-0374-x
5. Wang K, Zhu L, Gong H, et al. ANXA6 expression as a potential indicator of tumor diagnosis, metastasis and immunity in nasopharyngeal carcinoma. Int J Biol Macromol. 2024;283(Pt 4):137809. doi:10.1016/j.ijbiomac.2024.137809
6. Mao JR, Lan KQ, Liu SL, et al. Can the prognosis of individual patients with nasopharyngeal carcinoma be predicted using a routine blood test at admission?. Radiother Oncol. 2023;179:109445. doi:10.1016/j.radonc.2022.109445
7. Liu Y, Wang X, Li J, et al. Prediction of Recurrence using a Stacked Denoising

Autoencoder and Multifaceted Feature Analysis of Pretreatment MRI in Patients with Nasopharyngeal Carcinoma. Curr Radiopharm. 2025;18(3):224-243. doi:10.2174/0118744710384129250327060846

8. Shen Y, Domingo-Relloso A, Kupsco A, et al. AESurv: autoencoder survival analysis for accurate early prediction of coronary heart disease. Brief Bioinform. 2024;25(6):bbae479. doi:10.1093/bib/bbae479
9. Luo Yan, Wang Cong, and Ye Wenling. An interpretable predictive model for acute kidney injury based on XGBoost and SHAP [J]. Journal of Electronics and Information Technology, 2022, 44(01): 27-38.
10. Lee AW, Sou A, Patel M, Guzman S, Liu L. Early onset of nasopharyngeal cancer in Asian/Pacific Islander Americans revealed by age-specificanalysis.AnnEpidemiol.2023;80:25-29.doi:10.1016/j.annepidem.2023. 02.006
11. Zhu Y, Song X, Li R, Quan H, Yan L. Assessment of Nasopharyngeal Cancer in Young Patients Aged ≤ 30 Years. Front Oncol. 2019;9:1179. Published 2019 Nov 6. doi:10.3389/fonc.2019.01179
12. Guo LF, Hong JG, Wang RJ, Chen GP, Wu SG. Nasopharyngeal carcinoma survival by histology in endemic and non-endemic areas. Ann Med. 2024;56(1):2425066. doi:10.1080/07853890.2024.2425066
13. Ran X, Chen W, Yvert B, Zhang S. A hybrid autoencoder framework of dimensionality reduction for brain-computer interface decoding. Comput Biol Med. 2022;148:10 5871. doi:10.1016/j.compbiomed.2022.105871.
14. Dai Q, Chu Y, Li Z, et al. MDA-CF: Predicting MiRNA-Disease associations based on a cascade forest model by fusing multi-source information. Comput BiolMed. 2021;136:104706.doi: 10.1016/j.compbiomed.2021.104706
15. Sun S, Guo B, Mi Z, Zheng Z. Cross-modal semantic autoencoder with embedding consensus. Sci Rep. 2021;11(1):20319. Published 2021 Oct 13. doi:10.1038/s41598-021-92750-7