Heliyon 10 (2024) e29322

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

# Using machine learning methods to study the tumour microenvironment and its biomarkers in osteosarcoma metastasis

Guangyuan Liu [a,1], Shaochun Wang [b,1], Jinhui Liu [a], Jiangli Zhang [a], Xiqing Pan [a], Xiao Fan [a], Tingting Shao [c], Yi Sun [d,*]

[a] *The First Department of Orthopedic Surgery, Third Hospital of Shijiazhuang, Tiyu South Avenue No.15, Shijiazhuang, Hebei Province, China*
[b] *Department of Oncology, Shijiazhuang People's Hospital, No.365, Jian Hua Nan Da Jie, Shijiazhuang, Hebei Province, China*
[c] *Department of Pediatrics, Peking University First Hospital, 8 Xishku Street, Xicheng District, Beijing, China*
[d] *Department of Surgery, Shijiazhuang People's Hospital, No.365, Jian Hua Nan Da Jie, Shijiazhuang, Hebei Province, China*

ARTICLE INFO

*Keywords:*
Osteosarcoma
Machine learning
Feature selection
Biomarkers
Tumour microenvironment

ABSTRACT

*Background:* The long-term prognosis for patients with osteosarcoma (OS) metastasis remains unfavourable, highlighting the urgent need for research that explores potential biomarkers using innovative methodologies.

*Methods:* This study explored potential biomarkers for OS metastasis by analysing data from the Cancer Genome Atlas Program (TCGA) and Gene Expression Omnibus (GEO) databases. The synthetic minority oversampling technique (SMOTE) was employed to tackle class imbalances, while genes were selected using four feature selection algorithms (Monte Carlo feature selection [MCFS], Borota, minimum-redundancy maximum-relevance [mRMR], and light gradient-boosting machine [LightGBM]) based on the gene expression matrix. Four machine learning (ML) algorithms (support vector machine [SVM], extreme gradient boosting [XGBoost], random forest [RF], and k-nearest neighbours [kNN]) were utilized to determine the optimal number of genes for building the model. Interpretable machine learning (IML) was applied to construct prediction networks, revealing potential relationships among the selected genes. Additionally, enrichment analysis, survival analysis, and immune infiltration were performed on the featured genes.

*Results:* In DS1, DS2, and DS3, the IML algorithm identified 53, 45, and 46 features, respectively. Using the merged gene set, we obtained a total of 79 interpretable prediction rules for OS metastasis. We subsequently conducted an in-depth investigation on 39 crucial molecules associated with predicting OS metastasis, elucidating their roles within the tumour microenvironment. Importantly, we found that certain genes act as both predictors and differentially expressed genes. Finally, our study unveiled statistically significant differences in survival between the high and low expression groups of TRIP4, S100A9, SELL and SLC11A1, and there was a certain correlation between these genes and 22 various immune cells.

*Conclusions:* The biomarkers discovered in this study hold significant implications for personalized therapies, potentially enhancing the clinical prognosis of patients with OS.

[*] Corresponding authorNo.365, Jian Hua Nan Da Jie, Shijiazhuang, Hebei Province, 050011, China
*E-mail address:* 17603111132@163.com (Y. Sun).

[1] Guangyuan Liu [1#] and Shaochun Wang [2#] contributed equally.

https://doi.org/10.1016/j.heliyon.2024.e29322
Received 24 August 2023; Received in revised form 4 April 2024; Accepted 4 April 2024
Available online 5 April 2024

## 1. Introduction

**Abbreviations**

| | |
|---|---|
| OS | Osteosarcoma |
| TCGA | The Cancer Genome Atlas Program |
| GEO | Gene Expression Omnibus |
| SMOTE | Synthetic minority oversampling technique |
| MCFS | Monte Carlo feature selection |
| mRMR | minimum-redundancy maximum-relevance |
| LightGBM | Light gradient-boosting machine |
| ML, | Machine learning |
| SVM | Support vector machine |
| XGBoost | Extreme Gradient Boosting |
| RF | Random Forest |
| kNN | k-nearest neighbours |
| IML, | Interpretable machine learning |
| DS | Dataset |
| DEGs | Differentially expressed genes |
| IFS | Incremental feature selection |
| RI | Relative importance |
| CV | Cross-validation |
| MCC | Matthews correlation coefficient |
| ACC | accuracy |
| ROC, | Receiver operating characteristic |
| AUC | Area under the curve |
| RS | Rule support |
| RSLHS | Left rule support |
| RSRHS | Right rule support |
| PPI | Protein-protein interaction |
| GO | Gene ontology |
| BP | Biological processes |
| MF | Molecular functions |
| CC | Cellular components |
| KEGG | Kyoto Encyclopedia of Genes and Genomes |
| WGCNA | Weight gene co-expression network |
| SEER | Surveillance, Epidemiology, and End Results |
| SHAP | Shapley additive explanations |
| HPV | Human papillomavirus |

The most prevalent aggressive bone tumour, osteosarcoma (OS), comprises approximately 60 % of all bone cancers [1]. It typically occurs in adolescents and young adults, primarily affecting long bones like the distal femur, proximal tibia, and humerus [2]. OS is believed to originate from the malignant differentiation of mesenchymal cells [3]. Lung metastases are among the most common and perilous characteristics of OS. Approximately 15–20 % of patients have lung metastases at the time of diagnosis, and around 40 % of cases develop metastases later in the course of the disease [4]. While localized OS patients experience significant benefits from definitive surgical resection and adjuvant chemotherapy, only 20–30 % of patients with drug-resistant or metastatic disease will survive for five years [5,6]. Hence, gaining a comprehensive understanding of the tumour microenvironment and the underlying molecular mechanisms in OS, and discovering new biomarkers and therapeutic targets are crucial steps toward enhancing patient survival.

As a subset of artificial intelligence, machine learning (ML) allows computers to build algorithms and models without the need for explicit programming [7]. The ongoing advancement of multi-omics technologies has yielded more extensive biological data, rendering ML a potent tool for biomarker discovery [8]. For example, Zhang et al. used three ML models to identify eight prognostic immune genes in colorectal cancer patients and created a novel survival prediction system based on these models [9]. In another study, Zhou et al. utilized plasma lipidomic analysis and a support vector machine (SVM)-based ML algorithm to discover effective and reliable biomarkers for the metabolic detection of malignant gliomas [10]. These newly uncovered biomarkers have the potential to forecast disease onset, prognosis, and treatment outcomes, potentially enhancing our understanding of the biological processes that underlie these conditions.

Many previous studies on the OS tumour microenvironment and biomarkers have primarily relied on traditional bioinformatics methods, such as differential expression analysis or protein interactions. These methods often resulted in the screening of highly redundant genes, which hindered the construction of accurate and efficient predictive models. Furthermore, conventional ML methods often lacked interpretability, limiting their clinical applicability. Thus, in our research, we integrated statistical tests with ML techniques to enhance the reproducibility and interpretability of OS metastasis biomarker identification. Utilizing four feature ML selection algorithms and four ML classification methods, we effectively identified the optimal number of features. To address multiple dataset integration, interpretable machine learning (IML) methods were employed, enabling improved decision interpretation through visual "IF-THEN" rules and undirected networks. Through these comprehensive strategies, we aim to gain a better understanding of the

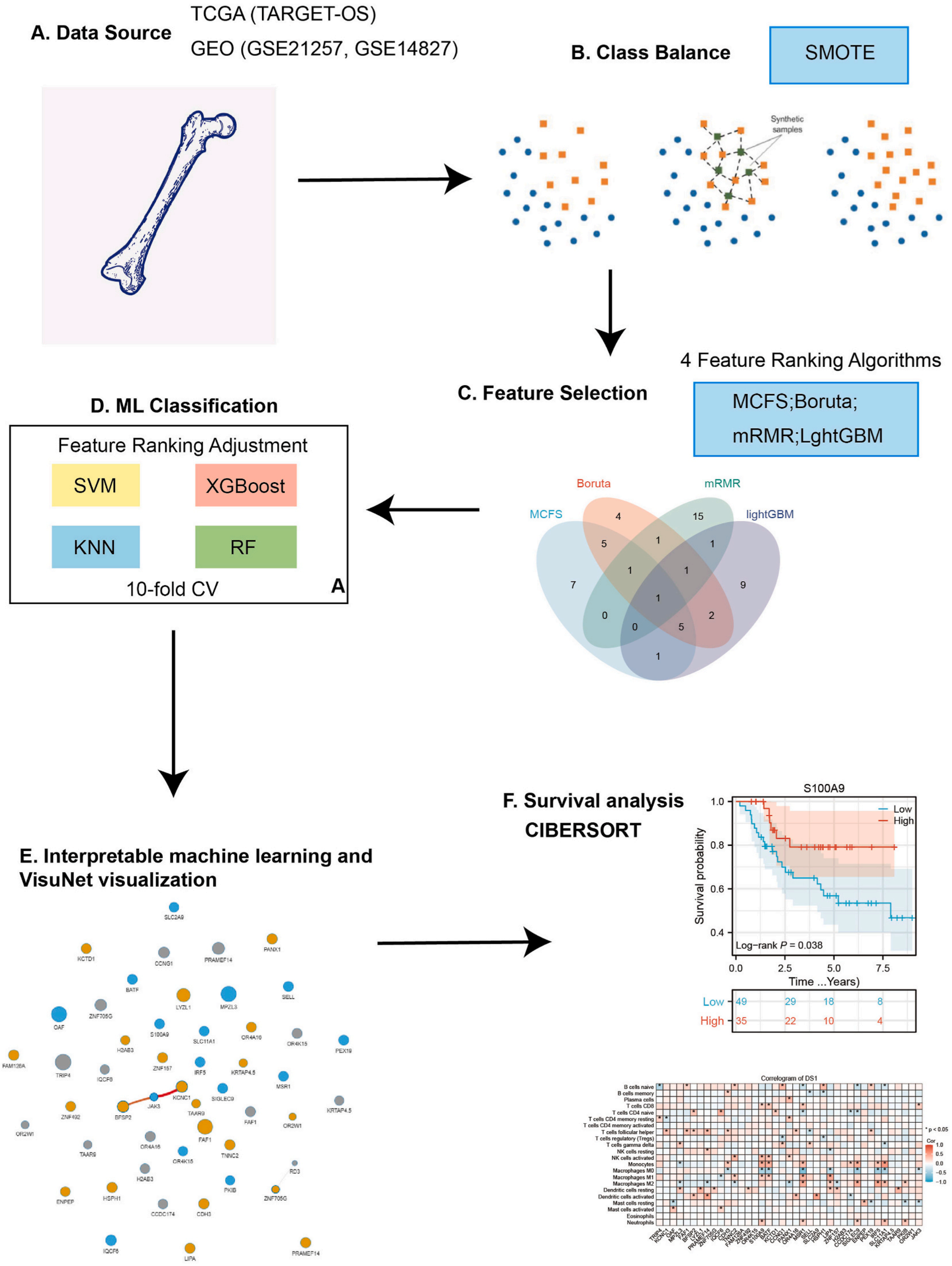


**Fig. 1.** Workflow for identifying biomarkers associated with osteosarcoma (OS) metastasis.

intricate features within the tumour microenvironment of OS and uncover relevant biomarkers. Ultimately, our goal is to reduce OS metastasis and improve treatment.

## 2. Materials and methods

Fig. 1 illustrated the procedural flow of our investigation aimed at delineating biomolecular markers linked to OS metastasis. The transcriptomes of three sets of samples were analysed by ML algorithms, involving four methods of ranking based on gene importance and four classification algorithms. In this section, the methods utilized in each step are depicted.

### 2.1. Data source

Three study datasets (DS1, DS2 and DS3) were meticulously curated from the Cancer Genome Atlas (TCGA) and the Gene Expression Omnibus (GEO). DS1, aptly referred to as the TAGET-OS dataset, was obtained by accessing the GDC website (https://portal.gdc.cancer.gov/). DS2 (GSE21257) and DS3 (GSE14827) were acquired using the GEOquery package [11].

### 2.2. Class imbalance correction

To compensate for the impact of category imbalance during model training, the synthetic minority oversampling (SMOTE) was applied. SMOTE achieved this balance by creating synthetic samples for the underrepresented categories through interpolation, using randomly selected instances and their k nearest neighbours [12]. We implemented SMOTE using the DMwR package [13].

### 2.3. Feature selection

In model training, we organized input features as lists and samples as rows. However, dimension constraints might weaken model performance. Feature selection algorithms aimed to extract a subset of features from the original set, reducing computational costs and enhancing predictive accuracy.

Four ML feature selection algorithms were employed in this study. Monte Carlo feature selection (MCFS) used supervised decision trees for iterative feature selection and assessment [14]. Boruta, inspired by random forests, identified relevant feature sets using shaded features and binomial distributions [15]. Minimum-redundancy maximum-relevance (mRMR) enhanced the feature-dependent variable relationship while minimizing overlap for the optimal subset [16]. Light gradient-boosting machine (LightGBM), a gradient-boosting decision tree algorithm, efficiently handled large samples and multi-feature data by evaluating feature importance based on their tree occurrences [17]. We implemented these algorithms with the rmcfs [18], Boruta [19], mRMRe [20], and lightgbm [21] packages, respectively.

### 2.4. Incremental feature selection

We applied the incremental feature selection (IFS) technique to determine the optimal number of features required for constructing the model [22]. Using different methods for feature selection, we created four lists of features ranked in order. From these lists, we extracted a subset of features using IFS. The IFS method added features one by one, with consideration of the relative importance (RI) diminishing at each step. It began with the initial feature ranked by four selection algorithms and continued until reaching the 100th feature. Each feature set was trained using four ML classification algorithms (support vector machine [SVM], extreme gradient boosting [XGBoost], random forest [RF], and k-nearest neighbours [kNN]) and evaluated through 10-fold cross-validation (CV). Performance evaluation utilized the Matthews correlation coefficient (MCC), accuracy (ACC), and the area under the receiver operating characteristic (ROC) curve (AUC) as key metrics.

The SVM algorithm optimized the distance between classification lines and data for accurate categorization [23]. The XGBoost algorithm efficiently trained models using decision trees, with each new tree improving on the last's predictions [24]. The RF algorithm was a composite learning algorithm that constructs a classification model using multiple tree-based classifiers [25]. The kNN algorithm assigned a category to a sample based on the majority vote of its k nearest neighbours, and calculated using Euclidean and Manhattan distances [26]. These four classification algorithms were accomplished using the e1071[27](27), XGBoost [28], randomforest [29], and class [30] packages, respectively.

### 2.5. Rule-based classification

Using rough set theory, IML handled data uncertainty and identified essential predictive features. It labelled the decision table's end columns as decision classes to find the minimum feature set needed. IML split continuous data into three parts using the equal frequency method for rule-based classification. Rule support (RS) represented the number of samples meeting the rule. The left rule support (RSLHS) was for the "IF" condition, and the right rule support (RSRHS) was for the "THEN" condition. The R. ROSETTA package built the classifier [31]. The Johnson reduction method eliminated insignificant features by iteration [32]. The function named recalculateRules updated statistical values after reducing features to discover new sets for making rules. The average rule accuracy showed the overall accuracy of the model. Also, features from three datasets were combined into a merged dataset to improve accuracy and feature variety. Rules were integrated across all models, adding up their RS and sets for frequently repeated rules.

### 2.6. Predictive network

The combined IML model was shown as a rule-based network using the VisuNet package, which allowed for visualizing shared prediction patterns among features [33]. This package had filters to highlight important parts of the network and display gene expression levels in specific node colors. This visualization helped identify frequently co-predicted genes. Prediction networks revealed central genes, indicating frequently predicted traits, and large nodes, representing features supported by many samples. Identifying these features contributed to interpret the result of IML model.

### 2.7. Protein–protein interaction analysis

The protein-protein interaction (PPI) network was created using the STRING database (http://stringdb.org) to collect data on gene interactions from the combined decision table [34].

### 2.8. Differentially expressed genes

The limma package was used to obtain differentially expressed genes (DEGs) with the p-values $<0.05$ and $|\log2 \text{ (fold-change) }| > 0$ in all three datasets.

### 2.9. Enrichment analysis

Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis were utilized to investigate OS biological functions and pathways in genes obtained from metastatic models in the merged decision table.

### 2.10. Evaluation of infiltrating immune cells

CIBERSORT (https://cibersort.stanford.edu/) was a tool used to analyse immune cell composition based on the matrix of leukocyte gene markers [35].

### 2.11. Statistical analysis

All statistical analyses were performed using R software 4.3.1. The Kaplan-Meier survival analysis was conducted utilizing the log-rank test. The correlation analysis was explored using Spearman's correlation. For all statistical analyses, p-value $<0.05$ was considered statistically significant.

## 3. Results

### 3.1. Overview of datasets

Table 1 provided a comprehensive outline of the dataset. DS1, DS2, and DS3 contained 87, 53, and 27 specimens, respectively, encompassing both metastatic and non-metastatic OS samples.

### 3.2. Class balancing

The proportions of metastatic OS samples in relation to their non-metastatic counterparts within DS1, DS2, and DS3 exhibited ratios of 22:65, 34:19, and 9:18, respectively. All three datasets were characterized by unbalanced class distributions, and DS3 had a sample size that was too small. We used SMOTE to address these issues. Table 1 presented the adjusted class distribution ratios for the three datasets: DS1, DS2, and DS3 yielded ratios of 66:65, 34:34, and 36:36, respectively.

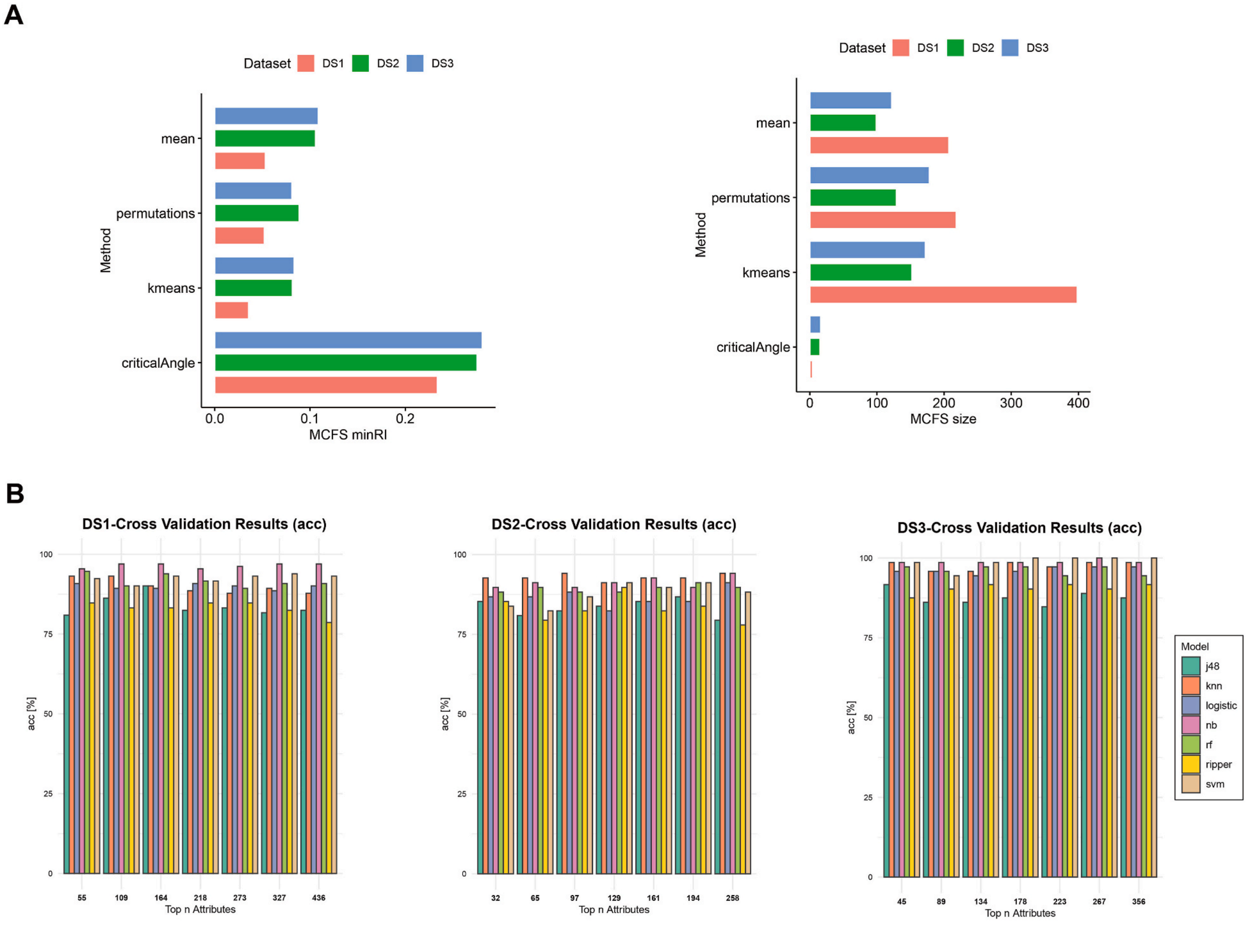
**Table 1**
Overview of the datasets.

| Overview | | | Before Class Balancing | | After Class Balancing | | No. (Genes) |
|---|---|---|---|---|---|---|---|
| Dataset | Source | Series | No. (Metastasis) | No. (Non-metastasis) | No. (Metastasis) | No. (Non-metastasis) | |
| DS1 | https://portal.gdc.cancer.gov/ | TARGET-OS | 22 | 65 | 66 | 65 | 19,545 |
| DS2 | Buddingh et al., 2011 | GSE21257 | 34 | 19 | 34 | 34 | 13,605 |
| DS3 | Kobayashi et al., 2010 | GSE14827 | 9 | 18 | 36 | 36 | 16,782 |

DS, dataset; OS, osteosarcoma.

### 3.3. Feature selection

Four different ML methods were employed for feature selection. To prevent overfitting, we conducted a 10-fold CV of the features selected by the MCFS algorithm. When compared to the other RI threshold selection methods, the critical angle method exhibited a larger minimum effective RI and a smaller number of selected features, potentially introducing noise (Fig. 2A). The outcomes of the other three methods were similar. Consequently, within the MCFS algorithm, we opted for the default permutations method. The results of the 10-fold CV illustrated the accuracy of these three datasets (Fig. 2B).



**C**

| Dataset | Confirmed | Tentative | Rejected | Recalculated |
|---|---|---|---|---|
| DS1 | 120 | 95 | 19,330 | 215 |
| DS2 | 103 | 49 | 13,468 | 137 |
| DS3 | 176 | 134 | 16,472 | 223 |

**Fig. 2.** Specific details of machine learning (ML) to select features. **(A)** Threshold selection for feature selection using four methods in Monte Carlo Feature Selection (MCFS) algorithm. The minimum relative importance (minRI) cutoff was used to select features. Four methods were evaluated: mean, k-means, critical angle, and permutations. The size of the selected features was defined as the number of features to be included. **(B)** Cross-validation (CV) results of MCFS algorithm. Histograms showing the 10-fold CV results of MCFS for the DS1, DS2, and DS3 datasets, respectively. The x-axis represents the number of top-ranked features, while the y-axis represents the accuracy. Each colored bar represents a different ML algorithm. **(C)** Feature types in the three datasets as identified by Boruta algorithm. The "Confirmed" column shows the number of features confirmed as important by Boruta, while the "Tentative column indicates the number of features that were not confirmed as important. The "Rejected" column displays the number of features that were discarded by Boruta. The "Recalculated" column is the sum of the number of features redefined as "Confirmed" by the TentativeRoughFix function and the original “Confirmed” features.

Fig. 2C presented an overview of the Boruta algorithm results. The TentativeRoughFix function within the Boruta package recalculated the features initially labelled as "Tentative" and some of them were reclassified as "Confirmed". The Boruta algorithm chose 215, 137, and 223 features for DS1, DS2, and DS3, respectively. For the mRMR and LightGBM results, we utilized the default parameters for feature selection. Tables S1–S3 displayed the top 20 feature genes for the four algorithms.

### 3.4. *Incremental feature selection*

To determine the optimal number of features required for building the model, we employed the IFS method. This method involved adding the ranked features from the Boruta algorithm based on the results of the function TentativeRoughFix. Regardless of the ML algorithm used (SVM, XGBoost, RF, or kNN), adding the first feature resulted in an average ACC ranging from 0.95 to 1.00, an average MCC ranging from 0.91 to 1.00, and an average AUC ranging from 0.94 to 1.00. Moreover, when the 20th feature was included, the average ACC ranged from 0.97 to 1.00, the average MCC ranged from 0.95 to 1.00, and the average AUC ranged from 0.95 to 1.00. These results indicated that the selected features in these datasets were of high quality. Therefore, the combination of the top 20 features from the four feature selection algorithms could be identified as the best feature set for the three datasets. All the IFS results were presented in Tables S4–S15.

### 3.5. *Classification rules*

In DS1, DS2 and DS3, 53, 45 and 46 genes were selected for the best feature set, respectively (Fig. 3A–C). Additionally, we noticed that the four algorithms shared one common feature gene (GULP1) in DS1, four common feature genes (SIGLEC9, MSR1, SYTL3, VMO1) in DS2, and one common feature gene (ZNF157) in DS3.

Combining the features from DS1, DS2, and DS3 resulted in 143, 112, and 135 genes, respectively. We observed that the VMO1 gene was common in DS2 and DS3, but there were no shared genes between DS1 and DS2 or between DS1 and DS3. We used the merged gene set to create classification rules for IML analysis. We conducted a comparison between the performance of models derived from the original gene set and the merged gene set (Table 2). Our findings showed that models based on the original gene set had an average AUC of 0.980 and an average ACC of 91.2 %. In contrast, models built on the merged gene set achieved an average AUC of 1.000 and an average ACC of 91.8 %. Furthermore, we noted that the variance in the average number of rules between the original and merged gene sets was not statistically significant.

We employed the Johnson reduction method to evaluate the rule model and the equal frequency method to categorize the data into three groups: low, medium, and high. After applying the function recalculateRules to recompute the rule set, the number of rules remained unchanged. However, both the average LHS and the average RHS increased (Table 3). Moreover, from the Johnson reduction model, we selected a total of 79 rules with Bonferroni-adjusted p-values less than 0.05 (Table 4). These selected rules can serve as interpretable predictive models for OS metastasis.

### 3.6. *Interpretable predictive network*

We analysed the predictive network using the VisuNet framework, enabling us to explore the relationships between genes in

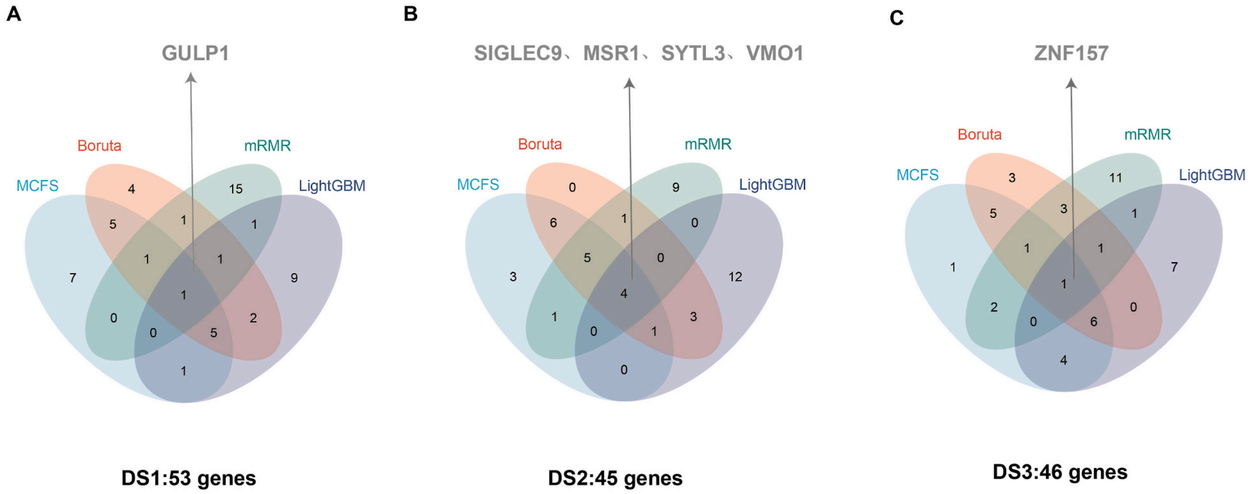


**Fig. 3.** Venn diagram showing the essential gene sets of the three datasets. The three circles represent the optimal feature sets of **(A)** DS1, **(B)** DS2, and **(C)** DS3, respectively. The number inside each oval denotes the number of genes in the corresponding feature set. The overlap between two ovals represents the common genes between the two datasets, while the intersection of all three ovals represents the shared genes in all three datasets.

**Table 2**
Results of interpretable machine learning (IML) models built on the original and merged feature lists.

| | Characteristic | DS1 | DS2 | DS3 |
|---|---|---|---|---|
| **Original** | No. features | 53 | 45 | 46 |
| | No. rules ($p < 0.05$) | 38 | 25 | 16 |
| | ACC | 86.3 % | 90.0 % | 97.5 % |
| | AUC | 0.939 | 1.000 | 1.000 |
| **Merged** | No. features | 143 | 112 | 135 |
| | No. rules ($p < 0.05$) | 37 | 26 | 16 |
| | ACC | 87.9 % | 90.0 % | 97.5 % |
| | AUC | 1.000 | 1.000 | 1.000 |

DS, dataset; ACC, accuracy; AUC, area under the curve.

**Table 3**
Performance evaluation of rules using the Johnson reduction method.

| | Metastasis | | Non-metastasis | |
|---|---|---|---|---|
| Rule statistics | Basic | Recalculated | Basic | Recalculated |
| Number of rules ($p < 0.05$) | 50 | 50 | 29 | 29 |
| LHS support | 40 | 45 | 37 | 44 |
| RHS support | 35 | 40 | 39 | 42 |
| Top predictors | TRIP4 | TRIP4 | LASP1 | LASP1 |

LHS, left-hand side support; RHS, right-hand side support.

metastatic OS (Fig. 4A). The co-predictive network revealed JAK3 as a co-predictor with KCNC and BSFP2, implying that these genes might be associated with similar biological processes related to OS metastasis. Moreover, numerous genes, including TRIP4, could act as independent predictors of OS metastasis.

Fig. 4B illustrated a PPI network of predictors associated with OS metastasis. Enrichment analysis conducted on these genes revealed potential biological functions linked to OS metastasis, such as the regulation of T-helper 1-type immune response and kinase regulatory activity, among others (Fig. 4C).

### 3.7. *Further exploration of the tumour microenvironment and biomarkers*

We verified the inclusion of 39 signature genes in the OS metastasis prediction model and further explored their roles in the tumour microenvironment, facilitating the search for better biomarkers. Fig. 5A demonstrates the DEGs between metastatic and non-metastatic OS in the three datasets. By combining the results of the differential expression analysis and the ML prediction model, we identified some genes, like OAF and FAF1, that may serve as both predictors and DEGs (Fig. 5B). Finally, when genes were divided into high and low-expression groups based on mean or median values, and the survival differences between these groups were compared. We observed statistically significant overall survival differences for TRIP4, S100A9, SELL, and SLC11A1 (Fig. 5C). We depicted the correlation between the expression levels of the 39 characterized genes and the abundance of the 22 immune cells in the three datasets (Fig. 6A–C).

## 4. Discussion

Previous research has utilized techniques such as bioinformatics and experiments to confirm the involvement of several crucial genes in OS metastasis. For example, Guan et al. identified essential modules using weighted gene co-expression network analysis (WGCNA) in relation to OS metastasis, leading to the discovery of four genes—ALOX5AP, HLA-DMB, HLA-DRA, and SPINT2—as potential indicators for OS metastasis [36]. Furthermore, Ma et al. conducted investigations using Transwell and Matrigel assays, which demonstrated that the inhibition of FAT10 significantly reduced the invasive and migratory capabilities of OS cells. At the same time, in vivo metastasis experiments indicated a reduced number of mice with distant metastases when FAT10 was silenced [37]. Nevertheless, most bioinformatics methodologies have primarily sought out significant molecules through singular and limited approaches. These methods involved the identification of crucial genes from PPI networks, the examination of prognosis-associated genes through COX regression, and the discovery of key modules and genes using WGCNA[38–40]. Significantly, models composed of genes selected through these methodologies demonstrated poor predictive efficacy for OS metastasis, thus constraining their practical utility in clinical practice [36,41].

The advancement in omics technology has led to an exponential increase in data. Artificial intelligence-based ML methods could utilize sophisticated statistical techniques to streamline repetitive and redundant data. This process contributed to the development of concise and finely-tuned prediction models. He et al. employed the SVM algorithm to pinpoint a classifier consisting of 64 crucial genes in OS. Their work showcased a remarkable prediction accuracy ranging from 92.6 % to 100 % for OS metastasis across multiple datasets. However, despite its impressive predictive capabilities, the ML algorithm lacked interpretability, which limited its practicality in clinical settings. Remarkably, in a study that integrated the Surveillance, Epidemiology, and End Results (SEER) database, Bai

**Table 4**
Set of rules and their statistics from the Metastasis and Non-metastasis model.

| No | Features[a] | Decision | accuracyRHS | supportRHS | coverageRHS | pValue |
|---|---|---|---|---|---|---|
| 1 | LASP1 = 3 | Non-metastasis | 0.954545 | 42 | 0.64615 | 1.24E-13 |
| 2 | RD3 = 1 | Non-metastasis | 0.931818 | 41 | 0.63077 | 4.69E-12 |
| 3 | TRIP4 = 2 | Metastasis | 0.888889 | 40 | 0.60606 | 1.87E-09 |
| 4 | KCNC1 = 1 | Non-metastasis | 0.869565 | 40 | 0.61538 | 4.38E-09 |
| 5 | TCTEX1D4 = 1 | Non-metastasis | 0.833333 | 40 | 0.61538 | 8.95E-08 |
| 6 | TRIP4 = 3 | Non-metastasis | 0.904762 | 38 | 0.58462 | 1.14E-09 |
| 7 | ASTL = 1 | Non-metastasis | 0.883721 | 38 | 0.58462 | 6.88E-09 |
| 8 | BFSP2 = 1 | Non-metastasis | 0.883721 | 38 | 0.58462 | 6.88E-09 |
| 9 | FAF1 = 1 | Non-metastasis | 0.883721 | 38 | 0.58462 | 6.88E-09 |
| 10 | MAGED2 = 3 | Non-metastasis | 0.863636 | 38 | 0.58462 | 3.54E-08 |
| 11 | KCNC1 = 3 | Metastasis | 0.826087 | 38 | 0.57576 | 1.33E-06 |
| 12 | BBS4 = 3 | Non-metastasis | 0.880952 | 37 | 0.56923 | 1.94E-08 |
| 13 | OAF = 1 | Metastasis | 0.880952 | 37 | 0.56061 | 4.10E-08 |
| 14 | MPZL3 = 1 | Metastasis | 0.860465 | 37 | 0.56061 | 2.01E-07 |
| 15 | TRAM1L1 = 3 | Non-metastasis | 0.840909 | 37 | 0.56923 | 4.20E-07 |
| 16 | SELL = 3 | Non-metastasis | 0.840909 | 37 | 0.56923 | 4.20E-07 |
| 17 | FAF1 = 3 | Metastasis | 0.840909 | 37 | 0.56061 | 8.59E-07 |
| 18 | BFSP2 = 3 | Metastasis | 0.837209 | 36 | 0.54545 | 2.13E-06 |
| 19 | LYZL1 = 3 | Metastasis | 0.941176 | 32 | 0.48485 | 1.92E-08 |
| 20 | PRAMEF14 = 2 | Metastasis | 0.964286 | 27 | 0.40909 | 2.20E-07 |
| 21 | ZNF705G = 2 | Metastasis | 0.964286 | 27 | 0.40909 | 2.20E-07 |
| 22 | CCNG1 = 1 | Non-metastasis | 1 | 26 | 0.72222 | 1.47E-10 |
| 23 | IQCF6 = 1 | Metastasis | 1 | 25 | 0.69444 | 6.30E-10 |
| 24 | ZNF157 = 1 | Non-metastasis | 1 | 24 | 0.66667 | 2.52E-09 |
| 25 | CDH3 = 3 | Metastasis | 1 | 24 | 0.66667 | 2.52E-09 |
| 26 | TNNC2 = 3 | Metastasis | 0.96 | 24 | 0.66667 | 4.79E-08 |
| 27 | FAM126A = 3 | Metastasis | 0.96 | 24 | 0.66667 | 4.79E-08 |
| 28 | ZNF492 = 3 | Metastasis | 0.96 | 24 | 0.66667 | 4.79E-08 |
| 29 | OR4K15 = 2 | Metastasis | 0.92 | 23 | 0.34848 | 6.11E-05 |
| 30 | S100A9 = 1 | Metastasis | 1 | 23 | 0.67647 | 9.28E-09 |
| 31 | BATF = 1 | Metastasis | 1 | 23 | 0.67647 | 9.28E-09 |
| 32 | PKIB = 3 | Non-metastasis | 1 | 23 | 0.67647 | 9.28E-09 |
| 33 | VMO1 = 3 | Non-metastasis | 1 | 23 | 0.67647 | 9.28E-09 |
| 34 | KCTD1 = 3 | Metastasis | 1 | 23 | 0.67647 | 9.28E-09 |
| 35 | CCNG1 = 2 | Metastasis | 1 | 23 | 0.63889 | 9.49E-09 |
| 36 | PANX1 = 3 | Metastasis | 1 | 23 | 0.63889 | 9.49E-09 |
| 37 | OR4A16 = 3 | Metastasis | 0.956522 | 22 | 0.33333 | 1.87E-05 |
| 38 | OR4A16 = 2 | Metastasis | 0.916667 | 22 | 0.33333 | 0.000137 |
| 39 | PRAMEF14 = 3 | Metastasis | 0.814815 | 22 | 0.33333 | 0.008905 |
| 40 | MSR1 = 1 | Metastasis | 1 | 22 | 0.64706 | 3.56E-08 |
| 41 | SELL = 1 | Metastasis | 0.956522 | 22 | 0.64706 | 6.14E-07 |
| 42 | LRRC25 = 3 | Non-metastasis | 0.956522 | 22 | 0.64706 | 6.14E-07 |
| 43 | SIGLEC9 = 3 | Non-metastasis | 0.956522 | 22 | 0.64706 | 6.14E-07 |
| 44 | KCNJ5 = 3 | Non-metastasis | 0.956522 | 22 | 0.64706 | 6.14E-07 |
| 45 | SLC11A1 = 3 | Non-metastasis | 0.956522 | 22 | 0.64706 | 6.14E-07 |
| 46 | OR4K15 = 1 | Metastasis | 0.916667 | 22 | 0.64706 | 5.49E-06 |
| 47 | SLC2A9 = 1 | Metastasis | 0.916667 | 22 | 0.64706 | 5.49E-06 |
| 48 | NSUN5 = 1 | Non-metastasis | 0.88 | 22 | 0.64706 | 3.39E-05 |
| 49 | CDH3 = 1 | Non-metastasis | 1 | 22 | 0.61111 | 3.39E-08 |
| 50 | HSPH1 = 3 | Metastasis | 1 | 22 | 0.61111 | 3.39E-08 |
| 51 | LIPA = 3 | Metastasis | 1 | 22 | 0.61111 | 3.39E-08 |
| 52 | ZNF157 = 3 | Metastasis | 0.956522 | 22 | 0.61111 | 5.71E-07 |
| 53 | H2AB3 = 2 | Metastasis | 0.913043 | 21 | 0.31818 | 0.000303 |
| 54 | H2AB3 = 3 | Metastasis | 0.913043 | 21 | 0.31818 | 0.000303 |
| 55 | ZNF705G = 3 | Metastasis | 0.777778 | 21 | 0.31818 | 0.045913 |
| 56 | IRF5 = 3 | Non-metastasis | 1 | 21 | 0.61765 | 1.29E-07 |
| 57 | PILRA = 3 | Non-metastasis | 1 | 21 | 0.61765 | 1.29E-07 |
| 58 | TYROBP = 3 | Non-metastasis | 0.954545 | 21 | 0.61765 | 2.08E-06 |
| 59 | CDH11 = 1 | Non-metastasis | 0.954545 | 21 | 0.61765 | 2.08E-06 |
| 60 | MIOX = 3 | Non-metastasis | 1 | 21 | 0.58333 | 1.15E-07 |
| 61 | FAF1 = 2 | Metastasis | 1 | 21 | 0.58333 | 1.15E-07 |
| 62 | CCDC174 = 2 | Metastasis | 0.954545 | 21 | 0.58333 | 1.82E-06 |
| 63 | IQCF6 = 2 | Metastasis | 0.909091 | 20 | 0.30303 | 0.000659 |
| 64 | SIGLEC9 = 1 | Metastasis | 1 | 20 | 0.58824 | 4.41E-07 |
| 65 | CD86 = 3 | Non-metastasis | 0.952381 | 20 | 0.58824 | 6.68E-06 |
| 66 | ENPEP = 3 | Metastasis | 0.952381 | 20 | 0.58824 | 6.68E-06 |
| 67 | PEX19 = 1 | Metastasis | 0.909091 | 20 | 0.58824 | 5.27E-05 |
| 68 | IRF5 = 1 | Metastasis | 0.909091 | 20 | 0.58824 | 5.27E-05 |

(*continued on next page*)

**Table 4** (*continued*)

| No | Features[a] | Decision | accuracyRHS | supportRHS | coverageRHS | pValue |
|---|---|---|---|---|---|---|
| 69 | SLC11A1 = 1 | Metastasis | 0.909091 | 20 | 0.58824 | 5.27E-05 |
| 70 | C5AR1 = 3 | Non-metastasis | 0.909091 | 20 | 0.58824 | 5.27E-05 |
| 71 | KRTAP4.5 = 2 | Metastasis | 0.95 | 19 | 0.28788 | 0.000224 |
| 72 | TAAR9 = 2 | Metastasis | 0.95 | 19 | 0.28788 | 0.000224 |
| 73 | PKIB = 1 | Metastasis | 0.904762 | 19 | 0.55882 | 0.000151 |
| 74 | TAAR9 = 3 | Metastasis | 0.947368 | 18 | 0.27273 | 0.000498 |
| 75 | KRTAP4.5 = 3 | Metastasis | 0.947368 | 18 | 0.27273 | 0.000498 |
| 76 | OR2W1 = 2 | Metastasis | 1 | 16 | 0.24242 | 0.000235 |
| 77 | JAK3 = 1, KCNC1 = 3 | Metastasis | 0.941176 | 16 | 0.24242 | 0.002359 |
| 78 | OR2W1 = 3 | Metastasis | 1 | 14 | 0.21212 | 0.001202 |
| 79 | JAK3 = 1, BFSP2 = 3 | Metastasis | 1 | 14 | 0.21212 | 0.001202 |

RHS, right-hand side support.

[a] Rules were selected based on a Bonferroni-adjusted p-value<0.05 using the recalculatedRules function. Genes were divided into three bins using the equal frequency method: 1 = low, 2 = medium, and 3 = high.

et al. achieved prediction accuracies ranging from 0.661 to 0.781 for forecasting distant metastases. The researchers obtained these results by utilizing six different machine learning algorithms, with the RF model exhibiting the most outstanding performance (71.8 % accuracy and 0.781 precision). Moreover, the integration of Shapley additive explanations (SHAP) analysis provided an interpretation that was independent of the model [42]. While this study employed IML models to predict OS metastasis, their accuracy was considerably lower than ML models constructed using genes. This difference might be attributed to the inclusion of primarily clinical variables, which were fewer in number compared to molecular variables obtained through high-throughput screening.

Taking into account the limitations of previous studies, we proposed an IML approach based on molecular data to investigate the tumour microenvironment of OS metastasis, discovered more reliable biomarkers, and developed practical predictive models. First, we addressed category imbalance by applying SMOTE to three OS datasets. We then conducted feature selection using four different ML descending dimension methods: MCFS, Boruta, mRMR, and LightGBM. Subsequently, to determine the optimal number of features for model construction, we performed IFS by combining four ML classification algorithms (SVM, XGBoost, RF, and kNN). In DS1, DS2, and DS3, we identified 53, 45, and 46 features, respectively. Next, we generated interpretable prediction networks using the merged set of genes. With the Johnson reduction method, we obtained 79 interpretable prediction rules that can be employed as a prediction model for OS metastasis. Finally, the role of 39 important molecules capable of predicting OS metastasis in the OS tumour microenvironment was further explored. Our findings revealed that certain genes serve as both predictors and DEGs. Furthermore, we observed the statistically significant survival difference between the high- and low-expression groups of TRIP4, S100A9, SELL, and SLC11A1.

Within the potential biomarkers identified in this study, several predictors were extracted, including TRIP4, KCNC1, JAK3, and BFSP2. Among them, JAK3 and KCNC1, as well as JAK and BFSP3, formed two pairs of co-predictive molecules for OS metastasis. TRIP4, a subunit of the tetrameric nuclear ASC-1 complex, plays a pro-tumorigenic role in various cancer types, such as cervical cancer and melanoma [43,44]. KCNC1 belongs to the family of membrane proteins that encode channel proteins, which regulate intracellular potassium ion permeability. Abnormalities in potassium channels within cancer cells can contribute to tumour progression, metastasis, and drug resistance [45]. The JAK family kinases are non-receptor tyrosine kinases that control the migration and invasion of osteosarcoma cells via the JAK/STAT signalling pathway [46]. While BFSP2 has not been previously investigated as a potential biomarker for OS metastasis, this study suggested its potential as such. Furthermore, survival analysis of these proteins indicated significant associations with OS prognosis for four genes: TRIP4, S100A9, SELL, and SLC11A1. S100A9, a calcium- and zinc-binding molecule of the S100 family, inhibits the growth of osteosarcoma cells by deactivating the MAPK and NF-κB signalling pathways [47]. SELL encodes Selectin-L, a molecule primarily involved in immune cell migration and inflammatory responses. It has been found to promote the progression of human papillomavirus (HPV)-positive head and neck squamous cell carcinoma [48]. SLC11A1, also known as natural resistance-associated macrophage protein-1, is a member of the lysosomal carrier family. It has been reported to be associated with a poor prognosis in various gastrointestinal tumours [49]. These potential biomarkers can enhance our understanding of molecular mechanisms and show promise as targets for future therapeutic interventions. The conducted comparative analyses validate the potential clinical significance of the genetic markers we have identified, strengthening their value as valuable indicators for both predicting OS metastasis and assessing prognosis. Nevertheless, it is crucial to further validate these biomarkers in larger cohorts and clinical settings to confirm their clinical significance and practical applicability.

Intersection of the DEGs and predictive factors revealed some DEGs that have the potential to serve as markers for distinguishing OS patients with and without metastasis according to expression patterns. Importantly, the expression of specific genes exhibited minimal variation across the three datasets, implying that the ML approach could reveal biomarkers not detected via differential expression analysis. In summary, these findings underscore the potential of ML algorithms for identifying novel biomarkers with implications for metastasis and prognosis in OS patients. Additionally, our study highlights the advantages of ML methods over conventional bioinformatics approaches for the identification of cancer-related features. Previous studies primarily conducted differential expression analysis. However, DEGs are not necessarily ideal biomarkers due to their abundance and high redundancy, which limits their suitability as biomarkers. Our ML approach, particularly the feature selection algorithm, has been demonstrated to identify the minimal number of biomarkers with the highest predictive capability, enhancing the precision and effectiveness of biomarker selection for clinical applications. The integration of ML and bioinformatics methodologies could contribute to a deeper understanding of the

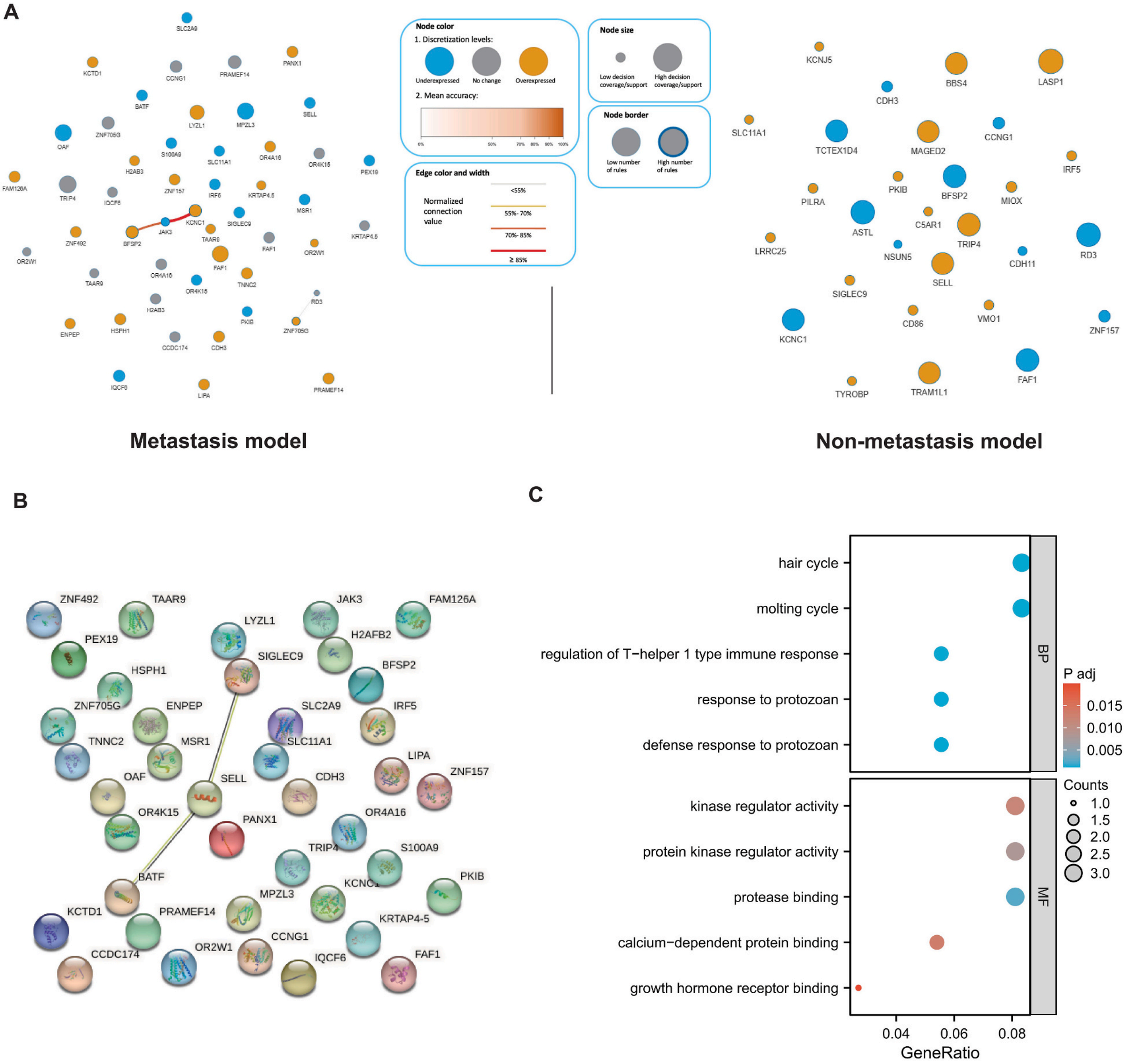
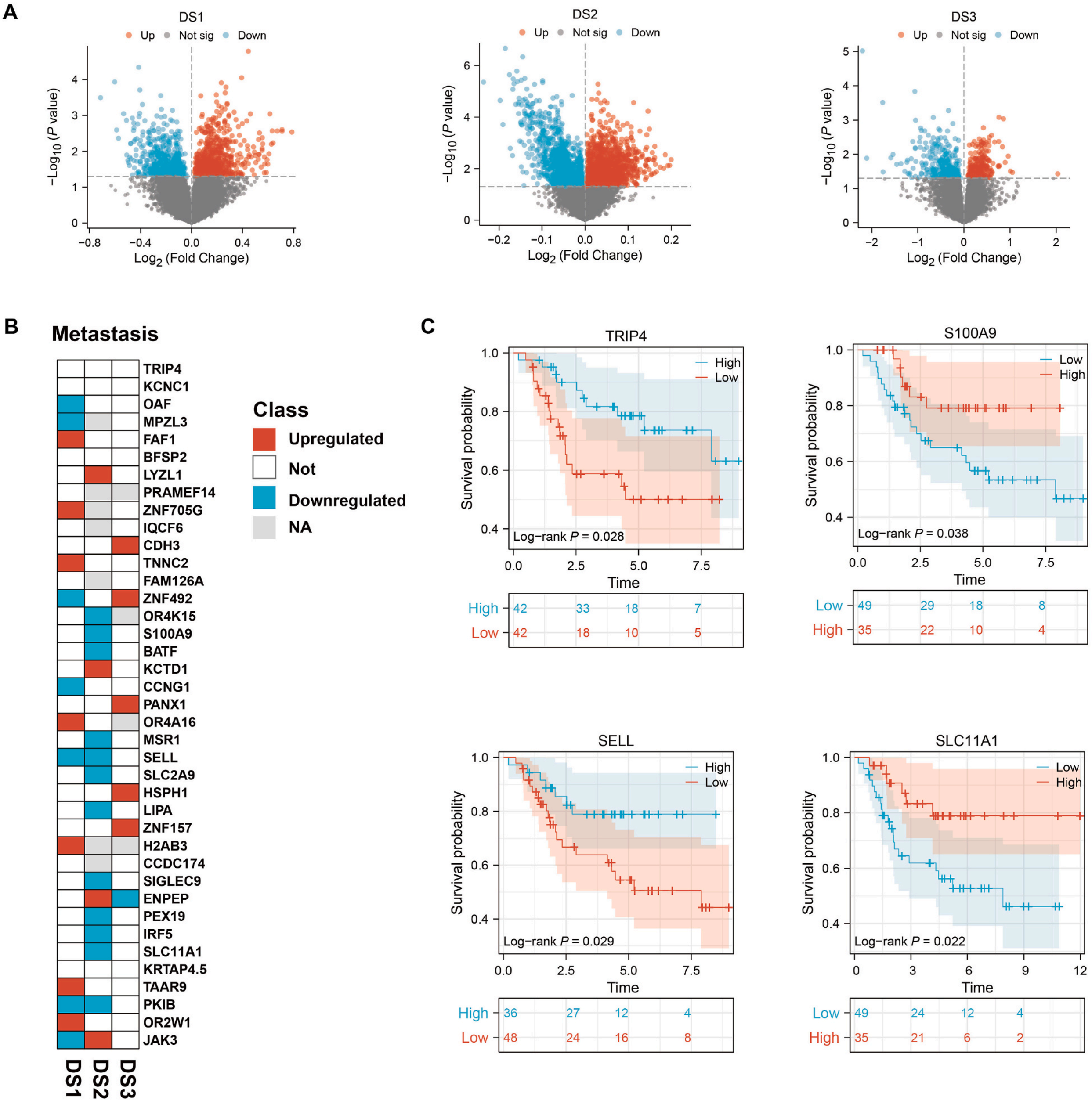


**Fig. 4.** Predictive networks constructed in accordance with osteosarcoma (OS) metastasis classification rules. **(A)** VisuNet predictive network of OS metastasis and non-metastasis samples. Each circle represents a node, where the size of the circle corresponds to the size of the support sets. The edge and node connections represent the strength of the co-prediction. **(B)** PPI network of genes based on OS metastasis model. Each node represents a protein, and the edge represents the interaction between two proteins. **(C)** Gene Ontology (GO) enrichment of genes in OS metastasis model. The top of the split side is biological processes (BP) and the bottom is molecular functions (MF).

metastatic mechanisms underlying tumours. Notably, according to our analysis, TRIP4 and BFSP2 emerged as notable predictors of OS metastasis, despite not being recognized as DEGs in the three datasets. The IML algorithm used in this study represents a potent modelling approach capable of identifying crucial predictive mechanisms and elucidating disparities between patient subgroups.

Furthermore, it is important to recognize the limitations of this study. One inherent limitation is that the continuous expression data were grouped into three intervals within the IML process, potentially resulting in the loss of certain information. Nonetheless, this step is imperative for modelling rooted in rough set theory. Another constraint lies in the limited number of genes that overlapped among the three datasets after feature selection, which could stem from the sampling of OS as well as the heterogeneity of patient lesions. Enhancing the generalizability of our findings would require the inclusion of additional datasets and larger sample sizes. Furthermore, this study exclusively concentrates on the classification of OS as positive or negative for metastasis. Subsequent studies could aim to apply ML classification for purposes such as tumour diagnosis and staging. The limitations of this study can guide future research in this domain.

Finally, the future implications of this study extend to significantly improving the diagnosis and treatment of OS through ML



| | | | |
|---|---|---|---|
| □ | □ | □ | TRIP4 |
| □ | □ | □ | KCNC1 |
| 🟦 | □ | □ | OAF |
| 🟦 | ▨ | □ | MPZL3 |
| 🟥 | □ | □ | FAF1 |
| □ | □ | □ | BFSP2 |
| □ | 🟥 | □ | LYZL1 |
| □ | ▨ | ▨ | PRAMEF14 |
| 🟥 | ▨ | □ | ZNF705G |
| □ | ▨ | □ | IQCF6 |
| □ | □ | 🟥 | CDH3 |
| 🟥 | □ | □ | TNNC2 |
| □ | ▨ | □ | FAM126A |
| 🟦 | □ | 🟥 | ZNF492 |
| □ | 🟦 | ▨ | OR4K15 |
| □ | 🟦 | □ | S100A9 |
| □ | 🟦 | □ | BATF |
| □ | 🟥 | □ | KCTD1 |
| 🟦 | □ | □ | CCNG1 |
| □ | □ | 🟥 | PANX1 |
| 🟥 | □ | ▨ | OR4A16 |
| □ | 🟦 | □ | MSR1 |
| 🟦 | 🟦 | □ | SELL |
| □ | 🟦 | □ | SLC2A9 |
| □ | □ | 🟥 | HSPH1 |
| □ | 🟦 | □ | LIPA |
| □ | □ | 🟥 | ZNF157 |
| 🟥 | ▨ | ▨ | H2AB3 |
| □ | ▨ | □ | CCDC174 |
| □ | 🟦 | □ | SIGLEC9 |
| □ | 🟥 | 🟦 | ENPEP |
| □ | 🟦 | □ | PEX19 |
| □ | 🟦 | □ | IRF5 |
| □ | 🟦 | □ | SLC11A1 |
| □ | □ | □ | KRTAP4.5 |
| 🟥 | □ | □ | TAAR9 |
| 🟦 | 🟦 | □ | PKIB |
| 🟥 | □ | □ | OR2W1 |
| 🟦 | 🟥 | □ | JAK3 |
| DS1 | DS2 | DS3 | |

**Fig. 5.** Differential expression analysis and survival analysis of biomarkers based on OS metastasis model. **(A)** Volcano plots of differential expression genes (DEGs) in three datasets. Red dots represent upregulated DEGs in metastatic OS samples; green dots represent downregulated DEGs; grey dots represent non-differentially expressed genes. **(B)** Summary of predictors associated with OS metastasis. The heatmap displays the predictors identified by the ML algorithm, along with their corresponding expression changes in the three datasets. Red squares represent upregulated DEGs in metastatic OS samples; blue squares represent downregulated DEGs; white squares represent non-differentially expressed genes; grey squares indicate missing data. **(C)** Kaplan-Meier survival analysis of genes associated with OS metastasis patient. The x-axis represents time in years, and the y-axis represents the percentage of surviving patients. The blue line represents patients with high gene expression, while the red line represents patients with low gene expression. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

methodologies. By identifying precise biomarkers, ML can offer a roadmap for personalized medicine, enabling treatments tailored to individual patient profiles. To maximize these advancements, it is crucial to educate physicians on ML techniques. This involves integrating ML training into medical education, focusing on its application in oncology, to ensure physicians are adept at interpreting ML-derived insights for clinical decision-making, ultimately enhancing patient outcomes in OS care.

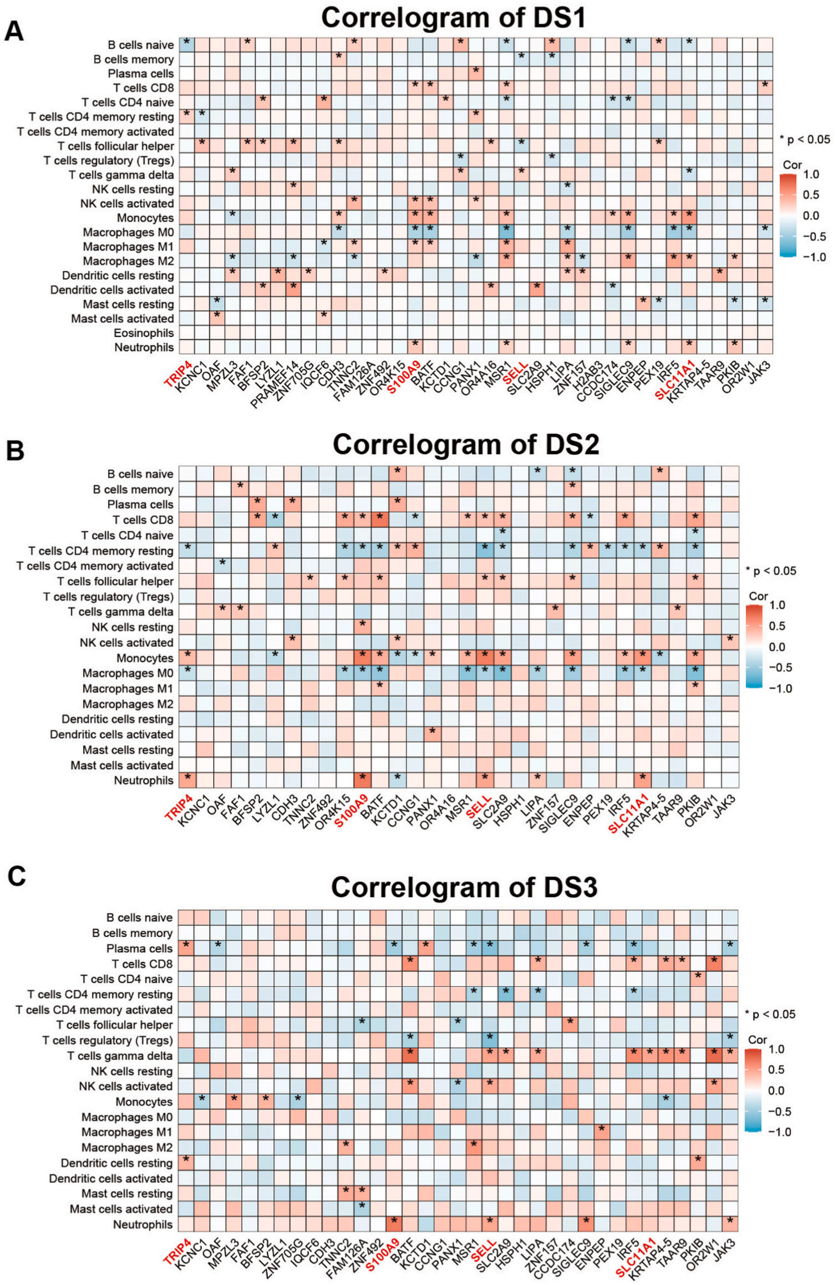


**Fig. 6.** Correlation of gene expression levels of OS metastasis model with 22 immune-infiltrating cell abundances in **(A)** DS1, **(B)** DS2 and **(C)** DS3. The correlation coefficients between the two are indicated using gradient colors, where darker red indicates a positive correlation and darker blue indicates a negative correlation. p-values indicate that the correlation is statistically significant, and asterisks are used to denote $p < 0.05$. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

## 5. Conclusions

In this study, the IML algorithm was used to explore the characteristics of the tumour microenvironment in OS metastasis, to identify potential biomarkers, and to construct a practical model with strong predictive efficacy. This study demonstrated the power of ML algorithms in cancer research and provided a completely new direction for medical and biological research. We recommend that omics data other than transcriptomics data be analysed in the future, even in combination with clinical data, to provide a strong theoretical basis for the diagnosis and treatment of OS.

## Funding statement

This study was funded by the Hebei Provincial Medical Science Research Project Program (grant no. 20231645).

## Data availability statement

The data associated with this study have been deposited into a publicly available repository. All the data are available from the TCGA database (https://portal.gdc.cancer.gov/) and GEO database (https://www.ncbi.nlm.nih.gov/geo/).

## Ethics approval and consent to participate

As this work benefited from the public database, informed consent was not applicable.

## CRediT authorship contribution statement

**Guangyuan Liu:** Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Shaochun Wang:** Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation. **Jinhui Liu:** Writing – review & editing, Visualization, Supervision, Formal analysis, Data curation. **Jiangli Zhang:** Writing – original draft, Software, Resources, Investigation, Data curation. **Xiqing Pan:** Formal analysis, Data curation. **Xiao Fan:** Resources, Formal analysis, Data curation. **Tingting Shao:** Writing – review & editing, Data curation. **Yi Sun:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Funding acquisition, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e29322.

## References

[1] H.C. Beird, S.S. Bielack, A.M. Flanagan, J. Gill, D. Heymann, K.A. Janeway, et al., Osteosarcoma, Nat. Rev. Dis. Prim. 8 (1) (2022) 77.

[2] M.S. Isakoff, S.S. Bielack, P. Meltzer, R. Gorlick, Osteosarcoma: current treatment and a collaborative pathway to success, J. Clin. Oncol. : official journal of the American Society of Clinical Oncology 33 (27) (2015) 3029–3035.

[3] V.K. Sarhadi, R. Daddali, R. Seppänen-Kaijansinkko, Mesenchymal stem cells and extracellular vesicles in osteosarcoma pathogenesis and therapy, Int. J. Mol. Sci. 22 (20) (2021).

[4] C. Meazza, P. Scanagatta, Metastatic osteosarcoma: a challenging multidisciplinary treatment, Expet Rev. Anticancer Ther. 16 (5) (2016) 543–556.

[5] A. Biazzo, M. De Paolis, Multidisciplinary approach to osteosarcoma, Acta Orthop. Belg. 82 (4) (2016) 690–698.

[6] C. Yang, Y. Tian, F. Zhao, Z. Chen, P. Su, Y. Li, et al., Bone microenvironment and osteosarcoma metastasis, Int. J. Mol. Sci. 21 (19) (2020).

[7] K. Swanson, E. Wu, A. Zhang, A.A. Alizadeh, J. Zou, From patterns to patients: advances in clinical machine learning for cancer diagnosis, prognosis, and treatment, Cell 186 (8) (2023) 1772–1791.

[8] M. Picard, M.P. Scott-Boyer, A. Bodein, O. Périn, A. Droit, Integration strategies of multi-omics data for machine learning analysis, Comput. Struct. Biotechnol. J. 19 (2021) 3735–3746.

[9] Z. Zhang, L. Huang, J. Li, P. Wang, Bioinformatics analysis reveals immune prognostic markers for overall survival of colorectal cancer patients: a novel machine learning survival predictive system, BMC Bioinf. 23 (1) (2022) 124.

[10] J. Zhou, N. Ji, G. Wang, Y. Zhang, H. Song, Y. Yuan, et al., Metabolic detection of malignant brain gliomas through plasma lipidomic analysis and support vector machine-based machine learning, EBioMedicine 81 (2022) 104097.

[11] S. Davis, P.S. Meltzer, GEOquery: a bridge between the gene expression Omnibus (GEO) and BioConductor, Bioinformatics 23 (14) (2007) 1846–1847.

[12] R. Blagus, L. Lusa, SMOTE for high-dimensional class-imbalanced data, BMC Bioinf. 14 (2013) 106.

[13] M. Hao, Y. Wang, S.H. Bryant, An efficient algorithm coupled with synthetic minority over-sampling technique to classify imbalanced PubChem BioAssay data, Anal. Chim. Acta 806 (2014) 117–127.

[14] M. Draminski, A. Rada-Iglesias, S. Enroth, C. Wadelius, J. Koronacki, J. Komorowski, Monte Carlo feature selection for supervised classification, Bioinformatics 24 (1) (2008) 110–117.

[15] M.B. Kursa, A. Jankowski, W.R. Rudnicki, Boruta–a system for feature selection, Fundam. Inf. 101 (4) (2010) 271–285.

[16] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.

[17] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, et al., Lightgbm: a highly efficient gradient boosting decision tree, Adv. Neural Inf. Process. Syst. 30 (2017).

[18] M. Dramiński, J. Koronacki, rmcfs: an R package for Monte Carlo feature selection and interdependency discovery, J. Stat. Software 85 (2018) 1–28.

[19] M.B. Kursa, W.R. Rudnicki, Feature selection with the Boruta package, J. Stat. Software 36 (2010) 1–13.

[20] N. De Jay, S. Papillon-Cavanagh, C. Olsen, N. El-Hachem, G. Bontempi, B. Haibe-Kains, mRMRe: an R package for parallelized mRMR ensemble feature selection, Bioinformatics 29 (18) (2013) 2365–2368.

[21] D. Wang, Y. Zhang, Y. Zhao (Eds.), LightGBM: an Effective miRNA Classification Method in Breast Cancer Patients. Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics, 2017.

[22] H. Liu, R. Setiono, Incremental feature selection, Appl. Intell. 9 (1998) 217–230.

[23] M. Bhasin, G. Raghava, ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST, Nucleic Acids Res. 32 (suppl_2) (2004) W414–W419.

[24] W. Li, Y. Yin, X. Quan, H. Zhang, Gene expression value prediction based on XGBoost algorithm, Front. Genet. 10 (2019) 1077.

[25] Subcellular localisation of proteins in fluorescent microscope images using a random forest, in: A.Z. Kouzani (Ed.), 2008 IEEE International Joint Conference on Neural Networks, IEEE World Congress on Computational Intelligence), IEEE, 2008.

[26] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theor. 13 (1) (1967) 21–27.

[27] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingesse, F. Leisch, C. Chang, et al., The e1071 package: Misc. functions of the department of statistics, 2014.

[28] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, et al., Xgboost: extreme gradient boosting, R package version 04-2 1 (4) (2015) 1–4.

[29] S. RcolorBrewer, M.A. Liaw, Package 'randomforest', University of California, Berkeley: Berkeley, CA, USA, 2018.

[30] B. Ripley, W. Venables, M.B. Ripley, Package 'class', The Comprehensive R Archive Network 11 (2015).

[31] M. Garbulowski, K. Diamanti, K. Smolińska, N. Baltzer, P. Stoll, S. Bornelöv, et al., ROSETTA: an interpretable machine learning framework, BMC Bioinf. 22 (2021) 1–18.

[32] D.S. Johnson (Ed.), Approximation Algorithms for Combinatorial Problems. Proceedings of the Fifth Annual ACM Symposium on Theory of Computing, 1973.

[33] K. Smolinska, M. Garbulowski, K. Diamanti, X. Davoy, S.O.O. Anyango, F. Barrenäs, et al., VisuNet: an Interactive Tool for Rule Network Visualization of Rule-Based Learning Models, 2021.

[34] D. Szklarczyk, R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, et al., The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest, Nucleic Acids Res. 51 (D1) (2023) D638–D646.

[35] A.M. Newman, C.L. Liu, M.R. Green, A.J. Gentles, W. Feng, Y. Xu, et al., Robust enumeration of cell subsets from tissue expression profiles, Nat. Methods 12 (5) (2015) 453–457.

[36] X. Guan, Z. Guan, C. Song, Expression profile analysis identifies key genes as prognostic markers for metastasis of osteosarcoma, Cancer Cell Int. 20 (2020) 104.

[37] C. Ma, Z. Zhang, Y. Cui, H. Yuan, F. Wang, Silencing FAT10 inhibits metastasis of osteosarcoma, Int. J. Oncol. 49 (2) (2016) 666–674.

[38] J. Liang, J. Chen, S. Hua, Z. Qin, J. Lu, C. Lan, Bioinformatics analysis of the key genes in osteosarcoma metastasis and immune invasion, Transl. Pediatr. 11 (10) (2022) 1656–1670.

[39] M.D. Cao, Y.C. Song, Z.M. Yang, D.W. Wang, Y.M. Lin, H.D. Lu, Identification of osteosarcoma metastasis-associated gene biomarkers and potentially targeted drugs based on bioinformatic and experimental analysis, OncoTargets Ther. 13 (2020) 8095–8107.

[40] J.S. Wang, Y.G. Wang, Y.S. Zhong, X.D. Li, S.X. Du, P. Xie, et al., Identification of co-expression modules and pathways correlated with osteosarcoma and its metastasis, World J. Surg. Oncol. 17 (1) (2019) 46.

[41] H. Zhang, L. Guo, Z. Zhang, Y. Sun, H. Kang, C. Song, et al., Co-expression network analysis identified gene signatures in osteosarcoma as a predictive tool for lung metastasis and survival, J. Cancer 10 (16) (2019) 3706–3716.

[42] B.L. Bai, Z.Y. Wu, S.J. Weng, Q. Yang, Application of interpretable machine learning algorithms to predict distant metastasis in osteosarcoma, Cancer Med. 12 (4) (2023) 5025–5034.

[43] Y. Che, Y. Li, F. Zheng, K. Zou, Z. Li, M. Chen, et al., TRIP4 promotes tumor growth and metastasis and regulates radiosensitivity of cervical cancer by activating MAPK, PI3K/AKT, and hTERT signaling, Cancer Lett. 452 (2019) 1–13.

[44] J. Hao, H. Xu, M. Luo, W. Yu, M. Chen, Y. Liao, et al., The tumor-promoting role of TRIP4 in melanoma progression and its involvement in response to BRAF-targeted therapy, J. Invest. Dermatol. 138 (1) (2018) 159–170.

[45] S. Chen, L. Xiao, H. Peng, Z. Wang, J. Xie, Methylation gene KCNC1 is associated with overall survival in patients with seminoma, Oncol. Rep. 45 (5) (2021).

[46] K.H. Lu, H.H. Wu, R.C. Lin, Y.C. Lin, P.W. Lu, S.F. Yang, et al., Curcumin analogue L48H37 suppresses human osteosarcoma U2OS and MG-63 cells' migration and invasion in culture by inhibition of uPA via the JAK/STAT signaling pathway, Molecules 26 (1) (2020).

[47] S. Cheng, X. Zhang, N. Huang, Q. Qiu, Y. Jin, D. Jiang, Down-regulation of S100A9 inhibits osteosarcoma cell growth through inactivating MAPK and NF-κB signaling pathways, BMC Cancer 16 (2016) 253.

[48] S. Wang, X. Pang, L. Tong, H. Fan, J. Jiang, M. Zhao, et al., LncRNA SELL/L-selectin promotes HPV-positive HNSCC progression and drives fucoidan-mediated therapeutic strategies, Acta Biomater. 167 (2023) 436–448.

[49] Y. Ma, L. Zhan, J. Yang, J. Zhang, SLC11A1 associated with tumor microenvironment is a potential biomarker of prognosis and immunotherapy efficacy for colorectal cancer, Front. Pharmacol. 13 (2022) 984555.