BMC Bioinformatics

RESEARCH

Open Access

# Deep learning-based metabolomics data study of prostate cancer

Liqiang Sun[1], Xiaojing Fan[2], Yunwei Zhao[1], Qi Zhang[3] and Mingyang Jiang[1*]

*Correspondence:
jiang_ming_yang@163.com

[1] College of Computer Science and Technology, Inner Mongolia Minzu University, Tongliao 028000, China
[2] College of Engineering, Key Laboratory of Intelligent Manufacturing Technology, Inner Mongolia Minzu University, Tongliao 028000, China
[3] Department of Immunology and Pathogenic Biology, Yanbian University Medical College, Yanji, China

## Abstract

As a heterogeneous disease, prostate cancer (PCa) exhibits diverse clinical and biological features, which pose significant challenges for early diagnosis and treatment. Metabolomics offers promising new approaches for early diagnosis, treatment, and prognosis of PCa. However, metabolomics data are characterized by high dimensionality, noise, variability, and small sample sizes, presenting substantial challenges for classification. Despite the wide range of applications of deep learning methods, the use of deep learning in metabolomics research has not been extensively explored. In this study, we propose a hybrid model, TransConvNet, which combines transformer and convolutional neural networks for the classification of prostate cancer metabolomics data. We introduce a 1D convolution layer for the inputs to the dot-product attention mechanism, enabling the interaction of both local and global information. Additionally, a gating mechanism is incorporated to dynamically adjust the attention weights. The features extracted by multi-head attention are further refined through 1D convolution, and a residual network is introduced to alleviate the gradient vanishing problem in the convolutional layers. We conducted comparative experiments with seven other machine learning algorithms. Through five-fold cross-validation, TransConvNet achieved an accuracy of 81.03% and an AUC of 0.89, significantly outperforming the other algorithms. Additionally, we validated TransConvNet's generalization ability through experiments on the lung cancer dataset, with the results demonstrating its robustness and adaptability to different metabolomics datasets. We also proposed the MI-RF (Mutual Information-based random forest) model, which effectively identified key biomarkers associated with prostate cancer by leveraging comprehensive feature weight coefficients. In contrast, traditional methods identified only a limited number of biomarkers. In summary, these results highlight the potential of TransConvNet and MI-RF in both classification tasks and biomarker discovery, providing valuable insights for the clinical application of prostate cancer diagnosis.

**Keywords:** Prostate cancer, Metabolomics, Hybrid deep learning, Transformer, CNN, Biomarker discovery

## Introduction

Cancer remains a major health concern globally and is a leading cause of mortality. Prostate cancer is the second most prevalent cancer affecting males [1] and is recognized as a heterogeneous disease that complicates both its diagnosis and treatment. The survival

rates for individuals diagnosed with prostate cancer vary significantly, ranging from approximately 33% for cases identified at an advanced stage to nearly 100% for those diagnosed at an early stage [2]. Therefore, early detection remains a pivotal challenge in cancer care. In recent years, metabolomics has shown great promise in disease diagnosis and treatment. Huang et al. developed a pathway-based model that uses metabolomics data for disease diagnosis. This approach, when applied to blood-based metabolomics data for breast cancer, identified key metabolic pathway signatures, which could be valuable for diagnostic tests and therapeutic interventions [3]. Similarly, John et al. proposed that metabolomics is a transformative technology, enhancing our ability to predict, detect, and understand a range of cardiometabolic diseases while also monitoring therapeutic responses [4].

Prostate cancer, a condition closely linked to metabolic changes, serves as an exemplary case for the application of metabolomics. As a result, the development of methods for early detection and diagnosis of prostate cancer has become a central focus of research. Metabolomics, a key component of systems biology, allows the identification of subtle changes in molecular metabolites during tumor initiation and progression. This field seeks to profile small molecules within an organism or cell, capturing dynamic metabolic alterations to construct a comprehensive metabolic profile. This approach helps elucidate the relationship between metabolic variations and disease progression. Unlike genomics, transcriptomics, and proteomics—which focus on upstream biological processes—metabolomics offers a downstream reflection of physiological activities by detecting a wide array of small molecule end products [5]. In prostate cancer research, metabolic profiling is increasingly used to identify predictive, diagnostic, and prognostic biomarkers. However, metabolomics data are often characterized by high dimensionality, noise, and variability [6]. High dimensionality arises when the number of metabolites significantly exceeds the number of samples, which can lead to model overfitting and the curse of dimensionality. Noise is introduced through measurement errors, technical biases during sample processing, and the complexity of biological background signals, potentially obscuring critical biological information. Variability reflects substantial differences in metabolite expression levels across individuals, groups, or experimental conditions. Consequently, extracting meaningful insights from metabolomics data poses significant challenges.

Numerous machine learning algorithms have been successfully applied to tackle classification and regression tasks in metabolomics data [7]. For example, RF, a widely used machine learning algorithm based on decision tree theory, effectively addresses challenges such as data imbalance and missing values inherent in high-dimensional datasets [8]. Support vector machines (SVMs) have also emerged as a prominent machine learning algorithm for classifying metabolomics data. SVMs construct optimal linear classifiers represented as hyperplanes with maximal margins, minimizing classification errors while maximizing geometric bounds. Notably, SVMs have been used to classify healthy individuals and patients with pneumonia [9]. However, conventional machine learning methods often struggle to achieve satisfactory results, particularly when confronted with the dimensionality challenges inherent in high-dimensional and sparse datasets. Recently, advances in artificial intelligence have positioned deep learning (DL) as a transformative approach in medical research [10]. DL, a machine learning method

inspired by the architecture of artificial neural networks [11], has achieved notable success in fields such as computer vision and natural language processing [12]. Its success is largely attributed to its powerful feature learning capabilities and its ability to efficiently capture complex nonlinear relationships. Despite these achievements, the application of DL to metabolomics data analysis remains underdeveloped.

Feedforward networks have demonstrated their ability to classify breast cancer metabolomics data with high precision [13]. Yuyang Sha et al. noted that the high dimensionality and complex interrelationships of metabolomics data present significant challenges for classical machine learning algorithms. To address these challenges, they developed a deep convolutional neural network-based method, MetDIT, which effectively resolves issues such as high dimensionality, small sample size, and category imbalance in clinical metabolomics data analysis [14]. Similarly, Taeho Jo et al. introduced the circular-sliding window association test (C-SWAT), which integrates inherent biological data correlations into the learning process. C-SWAT was applied to serum metabolomics data from 997 participants of the Alzheimer's Disease Neuroimaging Initiative (ADNI) and achieved a classification accuracy of 80.8% with an AUC of 0.81 in distinguishing Alzheimer's disease (AD) cases from cognitively normal older adults [15]. Date and Kikuchi proposed an improved deep neural network (DNN)-based analytical approach, DNN-MDA, which incorporates variable importance estimation using mean decrease accuracy (MDA). This method was evaluated on a dataset of metabolic profiles from yellowfin gobies living in various rivers across Japan. The DNN-MDA approach outperformed conventional multivariate and machine learning methods, achieving the highest classification accuracy (97.8%) among the examined approaches [16]. Furthermore, Alakwaa et al. highlighted the potential of metabolomics as a novel technique for diagnosing highly heterogeneous diseases. However, despite the growing popularity of deep neural networks, their suitability for classifying metabolomics data remains uncertain [13].

The primary challenge in metabolomics data analysis lies in its unique complexities compared to other data types. First, metabolomics data are characterized by extremely high dimensionality yet are often constrained by limited sample sizes, making traditional DL models highly susceptible to overfitting. Second, the intricate relationships between metabolites, often nonlinear and multi-level, present significant challenges to a model's ability to effectively capture and represent these complexities [6]. Existing DL methods often struggle to overcome these issues, resulting in difficulties in achieving accurate and robust feature extraction. To address this, we propose a hybrid model, TransConvNet, which integrates transformer [17] and Convolutional Neural Networks (CNN) [18]. TransConvNet is designed to fully leverage the strengths of the transformer for global feature extraction while utilizing CNNs to capture local features. This hybrid approach effectively handles the complex relationships inherent in metabolomics data, thereby improving the classification of prostate cancer metabolomics data and providing robust support for biomarker identification.

In the context of prostate cancer diagnosis, biomarker identification is as crucial as the classification of metabolomics data [19]. Traditional feature selection methods often fail to account for the complex interdependencies among features, making it difficult to identify biomarkers with true diagnostic value. To address this, we propose the MI-RF

method, which combines mutual information metrics with the feature importance scores from random forest. This approach enables a more accurate identification of key features associated with prostate cancer.

To evaluate the performance of the TransConvNet model, we conducted comparative experiments against traditional methods, including transformer, Long Short-Term Memory (LSTM), CNN, Extreme Gradient Boosting (XGBoost), SVM, RF, and decision tree (DT). Through five-fold cross-validation experiments, the results demonstrated that the TransConvNet model outperformed the other algorithms in key evaluation metrics, including accuracy and AUC, for classifying prostate cancer metabolomics data. Additionally, we constructed five datasets of varying sizes and conducted classification experiments across these different sample scales. The results reveal that TransConvNet exhibits strong robustness in handling datasets of varying sizes, consistently outperforming other classification algorithms. Furthermore, we validated the generalizability of TransConvNet on a lung cancer dataset, further demonstrating its adaptability and robustness in handling diverse data sources. Additionally, we employed the MI-RF algorithm to identify four key biomarkers—serotonin, sphinganine, sarcosine, and citrate—closely associated with the development of prostate cancer from high-dimensional metabolomics data. The experimental results show that TransConvNet outperforms traditional methods in classification tasks. Moreover, the integration of mutual information with random forest-based feature selection provides valuable insights for biomarker discovery, highlighting its potential in prostate cancer diagnosis.

## Methods

### Overall architecture

We propose a novel hybrid model, TransConvNet, based on the integration of a transformer and CNN, where the transformer handles global information modeling, and the CNN is responsible for local feature extraction. The transformer leverages a self-attention mechanism to capture global relationships within the data [20], which facilitates a more comprehensive understanding of the overall data structure. The transformer model offers high flexibility, allowing for the adjustment of the network structure and parameters based on task-specific requirements. This adaptability enables the transformer to efficiently accommodate various types of metabolomics data and tasks, thereby improving classification performance. Metabolomics data are inherently complex, with intricate interrelationships among metabolites, the incorporation of CNNs within the transformer architecture enables the efficient extraction of local features through convolutional operations, allowing the model to capture correlations between metabolites more effectively, leading to a more holistic understanding of the data. TransConvNet is a hybrid architecture that combines a standard transformer encoder (spatial transformer encoder (STE)) with a 1D convolution (CONV1D). Figure 1 shows the structure of TransConvNet. TransConvNet consists of two sublayers. The first sublayer includes a local–global interaction module, a multi-head attention layer, a CNN module, and a normalization layer. The second sublayer consists of a feedforward neural network and a normalization layer. Specifically, the local–global interaction module first captures local features within the metabolomic data and then enhances the representation of global features through information fusion. The multi-head attention mechanism allows for the

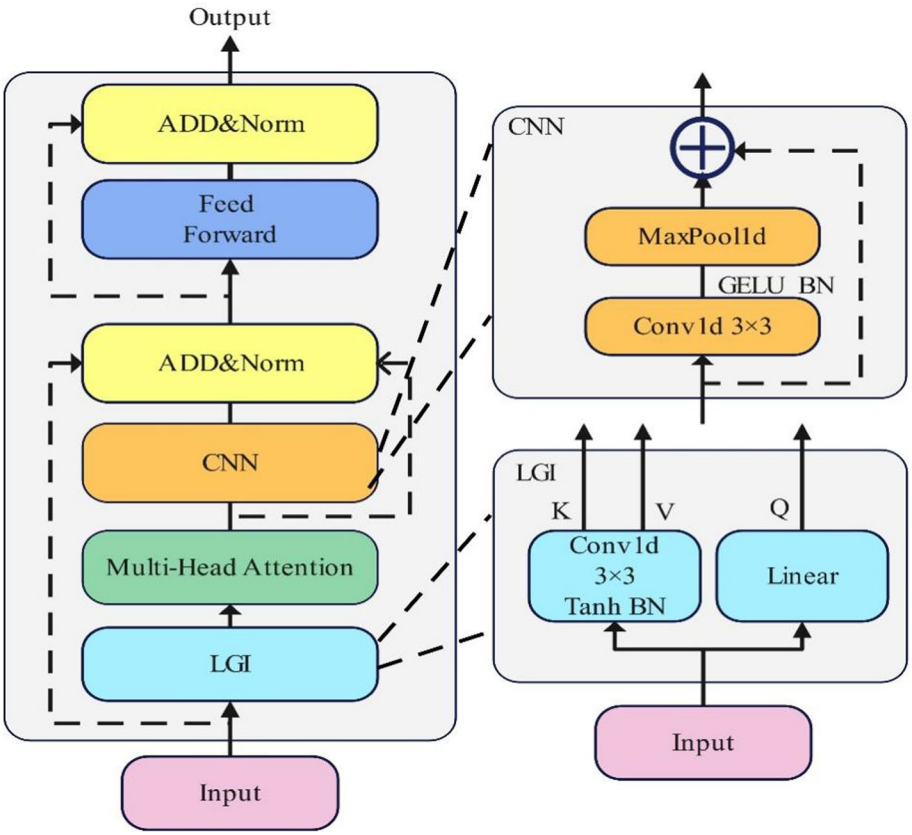


**Fig. 1** TransConvNet architecture

parallel computation of multiple attention weights, enabling the model to focus on different aspects of the input features. Each attention head learns distinct feature relationships, which allows the model to integrate and consider the multidimensional nature of metabolomic data. In the CNN module, convolutional layers process the features in the input data through local receptive fields, performing weighted summation on metabolite features to capture interactions between each feature and its neighboring features. The CNN layer's role is to extract locally discriminative information from the high-dimensional features output by the transformer model. Finally, the normalization layer ensures that the importance of each feature is balanced during training, promoting stability and convergence throughout the learning process. To address the vanishing gradient problem encountered during model training, we incorporate residual networks (ResNet) [21]. This architecture facilitates information flow across layers, enabling the network to learn effectively at greater depths and extract more complex feature representations. We implement shortcut connections for residual learning in both sublayers and the CNN module. All layers produce 400-dimensional outputs to facilitate residual connections. The details of these modules are described below.

#### Gated scaled dot product attention

The transformer is a DL model based on an attention mechanism. This mechanism allows the model to automatically focus on the most critical aspects of the input data,

thereby enhancing its ability to capture key features and complex relationships between data points. The attention weight calculation process uses a scaled dot product attention mechanism, which is known for its efficiency. Dot product attention enables the model to "automatically focus" on the most relevant parts of the data during processing. By calculating the relationships between different features, it helps the model identify which features are most important for the prediction task. The dot product attention, implemented using highly optimized matrix multiplication code, ensures faster processing and greater space efficiency [17]. In Fig. 2, the input consists of the query matrix $Q$, the key matrix $K$, and the value matrix $V$. First, the dot products of $Q$ and $K$ are computed, and then each dot product is multiplied by $\frac{1}{\sqrt{d_k}}$, where $d_k$ is the dimension of the key vector. Finally, the attention weights are obtained after the softmax operation, which is computed as follows.

$$A = softmax\left(\frac{Qk^T}{\sqrt{d_k}}\right) \cdot v$$

We introduce a gating mechanism to dynamically adjust the attention distribution, allowing finer control over the retention and output of global information. This adjustment leads to more intricate and information-rich data representations, thereby enhancing the performance of downstream tasks [22]. The gating mechanism functions as a filter, enabling the model to adjust attention weights dynamically based on the complex relationships and significance of the data. It "selects" the relevant information to process while filtering out less important data, thereby better capturing the correlations and key

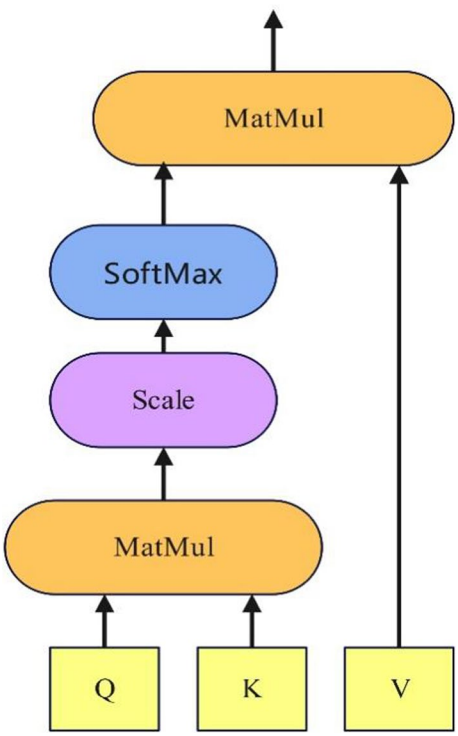


**Fig. 2** Scaled dot product attention

features within the data. This mechanism is particularly effective in identifying critical features for metabolomics data analysis and modeling. As shown in Fig. 3, the gating vector $G$ is added after computing the dot product of $Q$ and $K$, denoted as follows.

$$G = \sigma(X \cdot W + b)$$

$$A^{gate} = \left[ G \cdot softmax\left( \frac{Qk^T}{\sqrt{d_k}} \right) \right] \cdot v$$

Here, $X$ is the $d_k$-dimensional input, $W$ and $b$ are the learnable parameters, and $\sigma$ is the sigmoid activation function that compresses the linear combination of gating weights and input vectors to a value between 0 and 1.

#### Local-global interaction unit

In the conventional application of a standard transformer, $Q$, $K$, and $V$ are derived through linear mappings of the input data. The attentional weights for the weighted sum of the value vectors are computed by taking the dot product of the $Q$ and $K$ vectors and then employing a softmax operation. However, linear mapping may not sufficiently capture the intricate nonlinear relationships present in metabolomics data. Simultaneously, the transformer ignores local relationships and structural information within the data. To address these limitations, we propose a local–global interaction (LGI) unit. Figure 1 illustrates a representation of $K$ and $V$ with local information and a representation of $Q$ with global information. The global and local information interact in such a way that

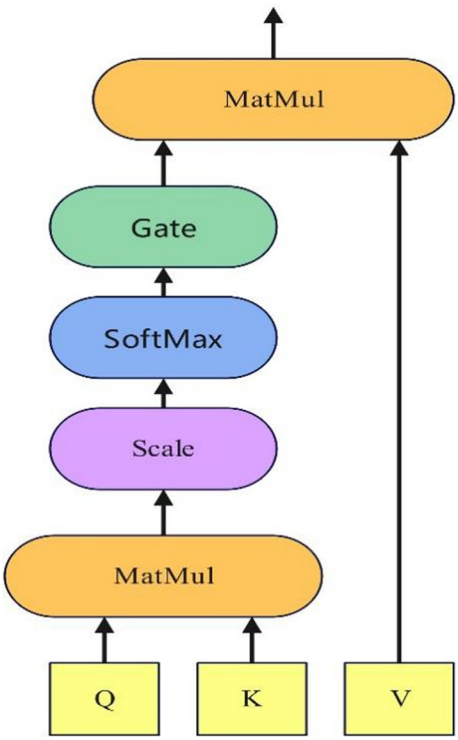


**Fig. 3** Gated scaled dot product attention

global and local correlations can be modeled, thereby improving the model's ability to capture nonlinear relationships. This is mathematically defined as follows:

$$K = BN(Tanh(Conv1d(X)))$$

$$V = BN(Tanh(Conv1d(X)))$$

$$Q = FC(X)$$

Here, $X$ represents the input data; Conv1d is a 1D convolution with a $3 \times 3$ convolution kernel; Tanh is the activation function; BN is BatchNorm1d, which is used to accelerate the neural network training and improve the stability of the model; and $FC$ represents a fully connected (linear) transformation of $X$.

#### Attention feature refinement

The attention mechanism typically improves the accuracy and performance of transformer models for prediction and classification tasks. However, these mechanisms often incur additional computational costs and may introduce redundancies in the feature representation within the encoder. We added a 1D convolution after the multi-head attention layer to extract the most relevant or representative features. By incorporating a 1D convolutional layer, the model can "refine" its outputs, focusing on specific, pertinent information. The 1D convolution operates by sliding over the outputs of the multi-head attention mechanism, extracting critical local features while compressing irrelevant information. This approach helps reduce noise, enhance feature clarity, and enables the model to focus more effectively on patterns that are most beneficial for classification or recognition tasks.

Additionally, residual learning is incorporated into the 1D convolutional network to further enhance the model's expressive power. As shown in Fig. 1, the CNN module is expressed as follows:

$$Out = Maxpool(BN(GELU(Conv1d(X)))) + X$$

Here, $X$ is the output of the multi-head attention mechanism, Conv1d() uses the GELU() activation function, the kernel width is 3, BN denotes BatchNorm1d, and max-pooling with a stride of 2 is applied to reduce the dimensionality of the data. This pooling process helps preserve important features while reducing the computational load.

## Results and discussion

#### Data set description and preprocessing

The data used in this study were sourced from the National Metabolomics Data Repository (NMDR) website under project number PR001613. The dataset comprises metabolomics data from plasma samples of adult men diagnosed with prostate cancer (n = 267) and healthy controls (n = 313). From this dataset, a total of 1169 metabolites were screened for analysis.

Raw metabolomics data are often characterized by dimensional disparities among different metabolites, which can pose challenges for subsequent analyses. To address this

issue, $z$-score normalization was applied to standardize the dataset, eliminating the effects of quantitative differences between features and rendering the data dimensionless. Specifically, the $z$-score standardization was performed for each feature by calculating the mean and standard deviation of the feature, subtracting the mean, and dividing by the standard deviation. This transformation scales the data to a standard normal distribution, with a mean of 0 and a standard deviation of 1. The formula for the transformation is as follows:

$$z = \frac{(x - \mu)}{\sigma}$$

where $x$ represents the individual value to be standardized, μ is the mean of the dataset, and σ denotes the standard deviation of the dataset.

Selecting the most relevant features from high-dimensional data is a critical step for model development. To address the challenges of high-dimensional data, effective feature selection, and dimensionality reduction techniques are essential for improving model generalization. In this study, we performed feature selection on the normalized data using the support vector machine recursive feature elimination (SVM-RFE) algorithm, as proposed by Huang et al. [23]. This method identified the top 400 most relevant features based on their contribution to the model, which were then used for further analysis.

#### Evaluation metrics

To evaluate the performance of our proposed model, we employed several widely used metrics in bioinformatics studies, including accuracy, sensitivity, specificity, and receiver operating characteristic (ROC) curves [24]. Accuracy serves as an indicator of the proportion of correct predictions made by the proposed model. Sensitivity (also referred to as the true positive rate) represents the probability of correctly identifying individuals with prostate cancer. It quantifies the likelihood of a true positive diagnosis. Specificity, also known as the true negative (TN) rate, indicates the probability of accurately diagnosing a healthy individual as negative. In this study, specificity pertains to the probability of correctly diagnosing a healthy patient. ROC curves provide a graphical representation of a classifier's performance under various classification thresholds. They are based on the trade-off between a true positive rate (sensitivity) and a false positive rate (1-specificity). The formulas calculating accuracy, sensitivity, and specificity are as follows.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

Here TP, FP, TN, and FN are the values of true positives, false positives, true negatives, and false negatives, respectively.

#### Five-fold cross-validation classification experiment

The proposed TransConvNet model for prostate cancer metabolomics classification was implemented using the PyTorch DL framework. The primary objective of this experiment was to evaluate the classification performance of the TransConvNet model, which aimed to improve the accuracy of prostate cancer diagnosis and treatment.

Given the relatively small sample size of metabolomics datasets, cross-validation is an effective technique for model validation, as it helps mitigate overfitting and ensures the model generalizes well across different subsets of data [25]. In this study, a five-fold cross-validation approach was employed, and the dataset was divided into five equal subsets. In each iteration, four subsets were used for training, while the remaining subset served as the validation set. This procedure was repeated five times, each time using a different subset as the validation set. This process was repeated until each group of data was used as a validation set. Optuna [26], an automated hyperparameter optimization framework, was employed to fine-tune the parameters of the TransConvNet architecture and maximize performance efficiency. In our experiments, we utilized the Optuna framework to optimize hyperparameters across the model's layers. Additionally, five-fold cross-validation was applied to ensure that the selected hyperparameters consistently delivered robust performance across different data partitions. During the hyperparameter tuning process, we focused on several critical parameters, including the learning rate. Through extensive experimentation, the optimal learning rate was identified as 0.0001. The Regularization parameter (Dropout rate) was adjusted through cross-validation to mitigate overfitting. Parameters related to the network architecture and depth, such as the number of convolutional layers, nodes per layer, attention heads, and the dimensions of the feedforward network, were fine-tuned based on experimental results. Furthermore, the hyperparameters of other comparative algorithms were optimized for a fair comparison.

We compared the TransConvNet model with seven machine learning methods: transformer, CNN, LSTM, XGBoost, SVM, RF, and DT. The average results of the evaluation metrics were computed on the test set using five-fold cross-validation. As shown in Table 1, the proposed TransConvNet model outperformed the other algorithms in

**Table 1** Five-fold cross-validation results for eight models

| Model | Sensitivity (%) | Specificity (%) | Accuracy (%) | AUC |
|---|---|---|---|---|
| Transformer | 70.26 | 76.32 | 74.47 | 0.82 |
| LSTM | 54.21 | 73.64 | 65.00 | 0.67 |
| CNN | 61.11 | 69.78 | 66.66 | 0.69 |
| XGBoost | 59.39 | 55.35 | 57.44 | 0.58 |
| SVM | 64.96 | 72.33 | 68.47 | 0.69 |
| RF | 36.43 | 70.41 | 54.47 | 0.54 |
| DT | 54.80 | 58.40 | 56.89 | 0.57 |
| TransConvNet | 81.67 | 80.72 | 81.03 | 0.89 |

terms of diagnostic sensitivity, specificity, and accuracy. The results also revealed that the TransConvNet model achieved the highest accuracy and AUC value while maintaining high sensitivity and specificity. The classification accuracy of the TransConvNet model for prostate cancer metabolomics data was 81.03%, surpassing the transformer method and demonstrating superior classification performance compared to the other seven classification methods.

These results indicate that the TransConvNet model effectively achieves interaction and fusion between local and global information, enabling the identification of critical variables that contribute to the construction of the model. In contrast, the transformer exhibited lower classification accuracy than the TransConvNet model, likely due to its neglect of local information during feature extraction. However, it is noteworthy that the transformer achieved higher classification accuracy than the other comparative algorithms, apart from TransConvNet. This underscores the transformer's effectiveness in handling high-dimensional, small sample data while highlighting the strong potential of the TransConvNet hybrid structure for further improvement.

Algorithms such as XGBoost, RF, and DT displayed the worst accuracy, likely due to their sensitivity to overfitting and limitations in addressing nonlinear problems. The experimental results demonstrate that the proposed TransConvNet model exhibits high diagnostic efficiency and strong potential for clinical applications.

The stability of a model is a critical criterion for assessing its performance and reliability. In the context of metabolomics data, which is characterized by high dimensionality, noise, and inherent variability, classification accuracy serves as a key indicator of the model's ability to handle these challenges effectively. Figure 4 presents a line graph illustrating the classification accuracy across the five-fold cross-validation for various

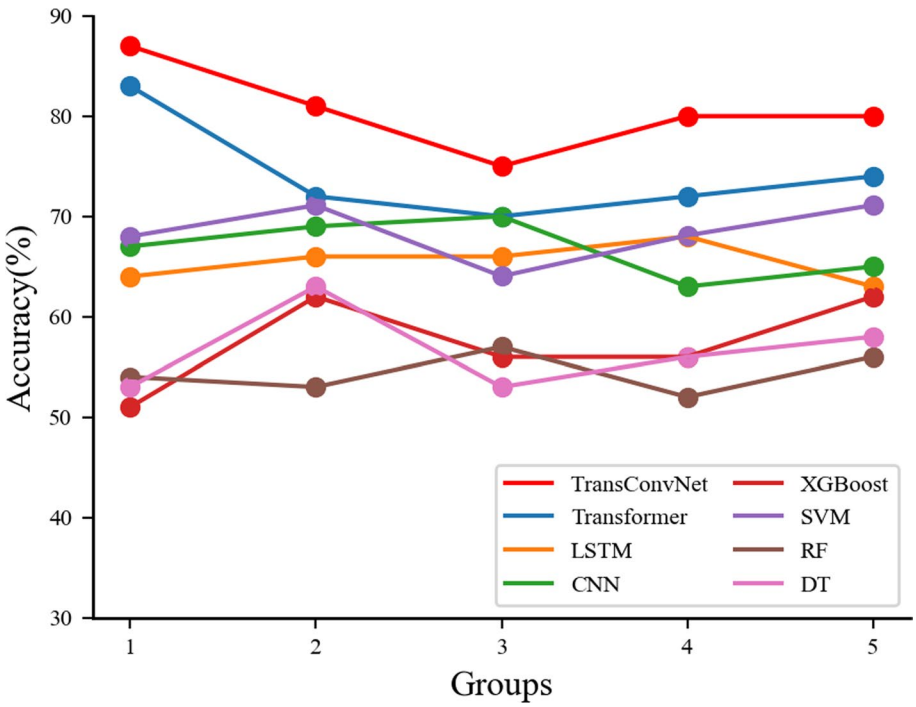


**Fig. 4** Line plot of five-fold cross-validation classification accuracy

datasets. The classification accuracies for TransConvNet ranged from 74.13% to 86.20%, with the highest accuracy observed in each fold of the validation experiments, yielding an average accuracy of 81.03%. The lowest accuracy was recorded for the third validation set. These results suggest that the model's performance is relatively stable across different training and validation sets, demonstrating good generalization ability and adaptability to new data.

In comparison, the transformer model achieved classification accuracies ranging from 70.68% to 83.62%, exhibiting a larger variation in performance. This variation may be attributed to issues such as overfitting. Although the results from the other machine learning algorithms were stable, their classification accuracies were not as satisfactory, likely due to their inability to effectively capture the complex nonlinear relationships within the data. Figure 5 illustrates the accuracy curves for each fold of the TransConvNet and transformer models during the cross-validation experiment, providing a comparison of the accuracy trends over time. The accuracy curves for TransConvNet show a quick convergence in each fold, with stable, flat curves, indicating that the model effectively learns discriminative features from the training data. In contrast, the accuracy curves for the transformer model exhibited greater fluctuations, signaling less stability and consistency in learning. Thus, from the perspective of classification accuracy, the TransConvNet model demonstrated superior stability, reliability, and suitability for classifying prostate cancer metabolomics data.

The ROC curve is a widely used tool to assess the sensitivity and specificity of a model, offering an intuitive visualization of its classification performance. In this study, the vertical axis of the ROC curve represents sensitivity, while the horizontal axis denotes specificity. A perfect classifier would reach a point where sensitivity is 1 and specificity is 0, meaning there is no misdiagnosis and no leakage in predictions [39]. As shown in Fig. 6, the ROC curve for the TransConvNet model consistently stays closer to the upper left corner across all folds of the cross-validation process compared to the transformer model. This suggests that TransConvNet demonstrates better classification performance across various thresholds. Furthermore, the TransConvNet curve remains above the transformer curve at most thresholds, signifying its superior performance and robustness in a wide range of scenarios. This enhanced performance is likely due to TransConvNet's ability to capture both global and local features, providing a more nuanced understanding of the complex, nonlinear relationships inherent in metabolomics data.

In contrast, the transformer model struggles to fully leverage these nonlinear relationships and local features, which may be attributed to its fixed embedding approach and the limitations of its model architecture. This demonstrates the advantage of the hybrid structure in TransConvNet, allowing it to outperform the transformer model by capturing intricate data features more effectively and exhibiting improved generalization ability.

During training, the TransConvNet model parameters were tuned over several iterations to minimize the loss function. The accuracy of the model increased rapidly and then stabilized. Although the loss function of the transformer gradually decreased during training, its curve showed a slower rate of decrease, requiring more training iterations to achieve comparable results. Figure 7 shows a plot of the losses recorded by the TransConvNet and transformer models for the training data for each iteration. Figure 7a

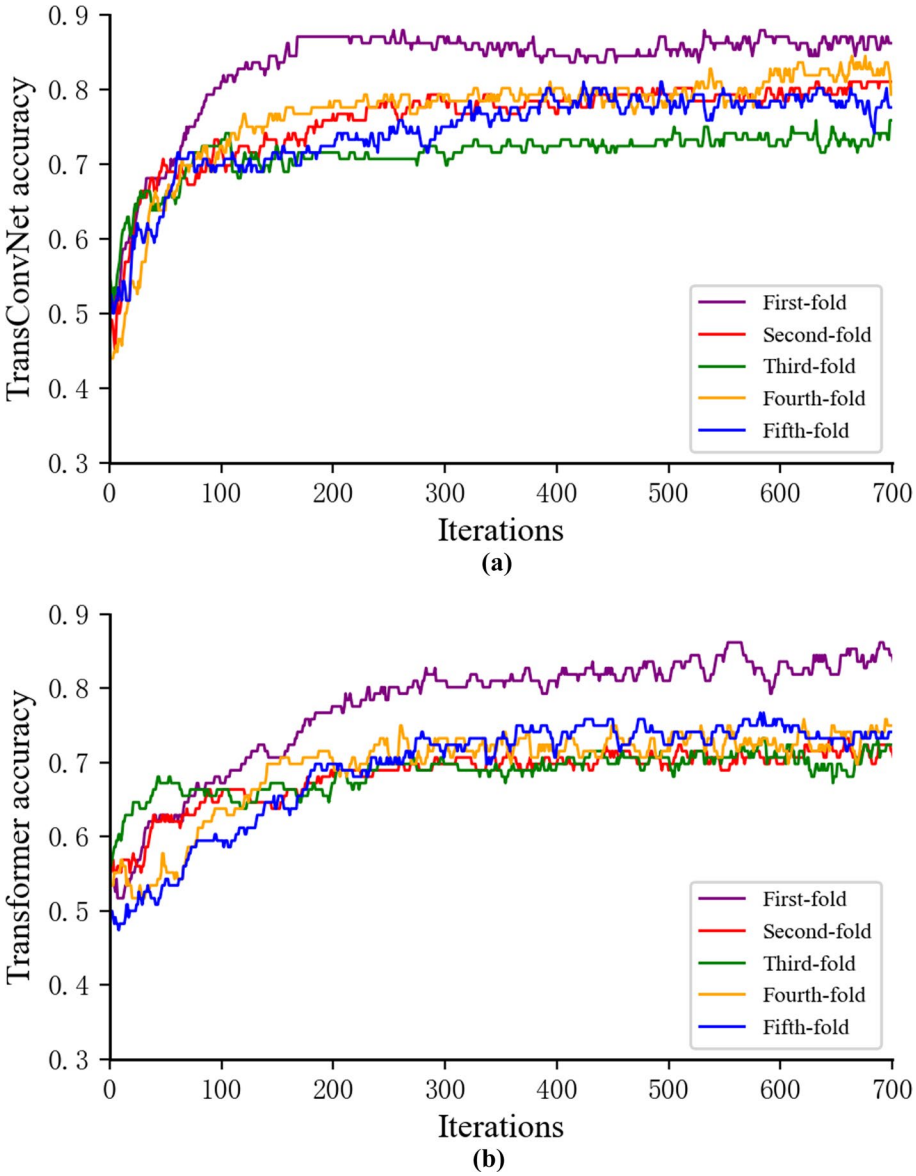


**Fig. 5** **a** Accuracy curves of the TransConvNet model. **b** Accuracy curves of the transformer model

demonstrates that the TransConvNet model's loss value gradually decreased during each fold of training, eventually reaching a steady state. This indicates that the model converged quickly to the optimal solution, fitting the data effectively and enabling it to accurately classify metabolomics data. In contrast, the transformer model struggled to fit the data effectively, likely due to its inability to capture local features, which prevented the loss value from converging to 0.

#### Classification experiments with samples of different sizes

The complexity of classifying high-dimensional, noisy, and small sample metabolomics data is significantly increased by the inherent challenges these data present. To further evaluate the classification performance of different models for metabolomics data and

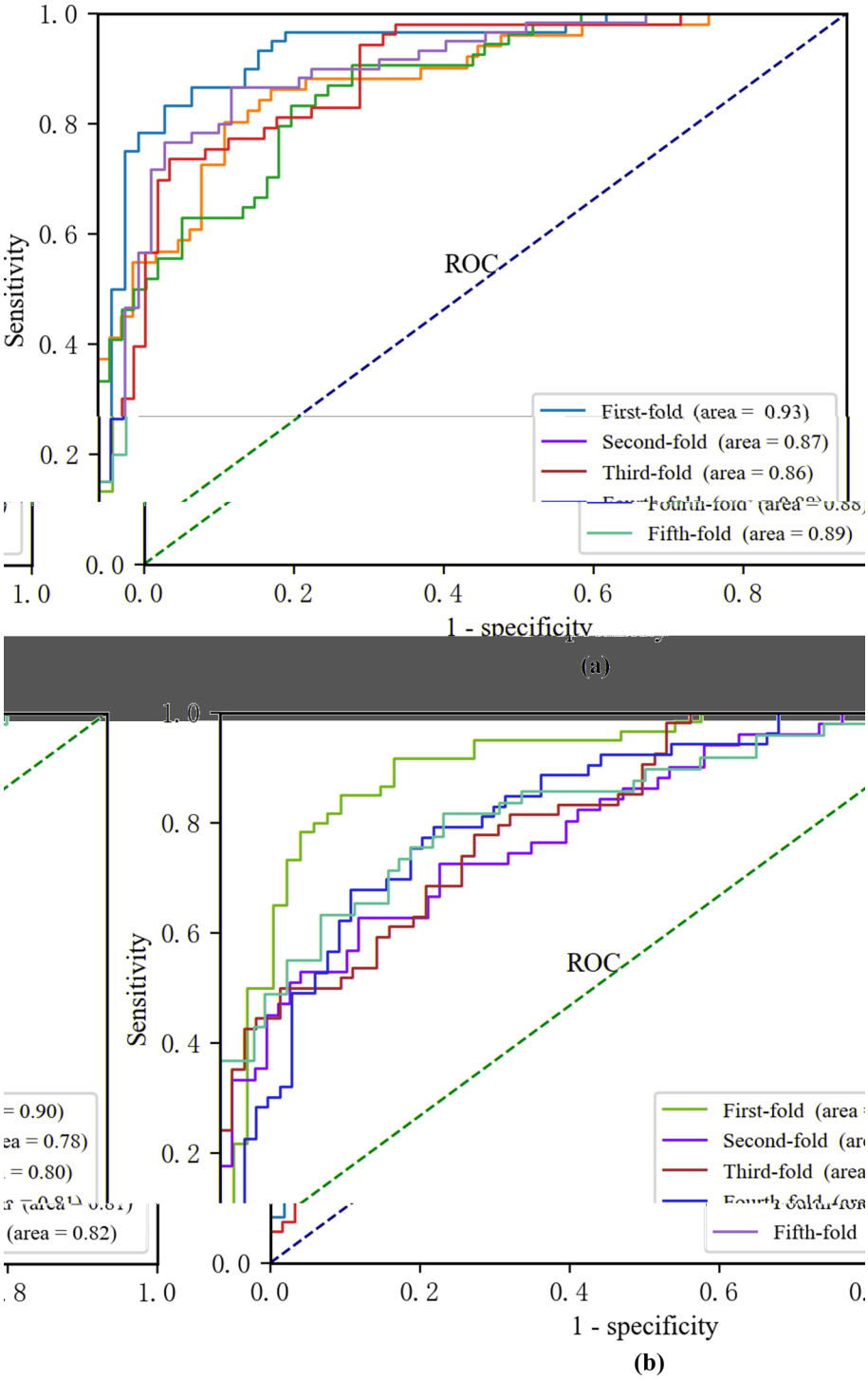


**Fig. 6** **a** ROC curves of five-fold cross-validation for the TransConvNet. **b** ROC curves of five-fold cross-validation for the transformer

the robustness of the proposed TransConvNet model, we divided the training and test sets of the metabolomics data into five groups for classification experiments with different sample sizes.

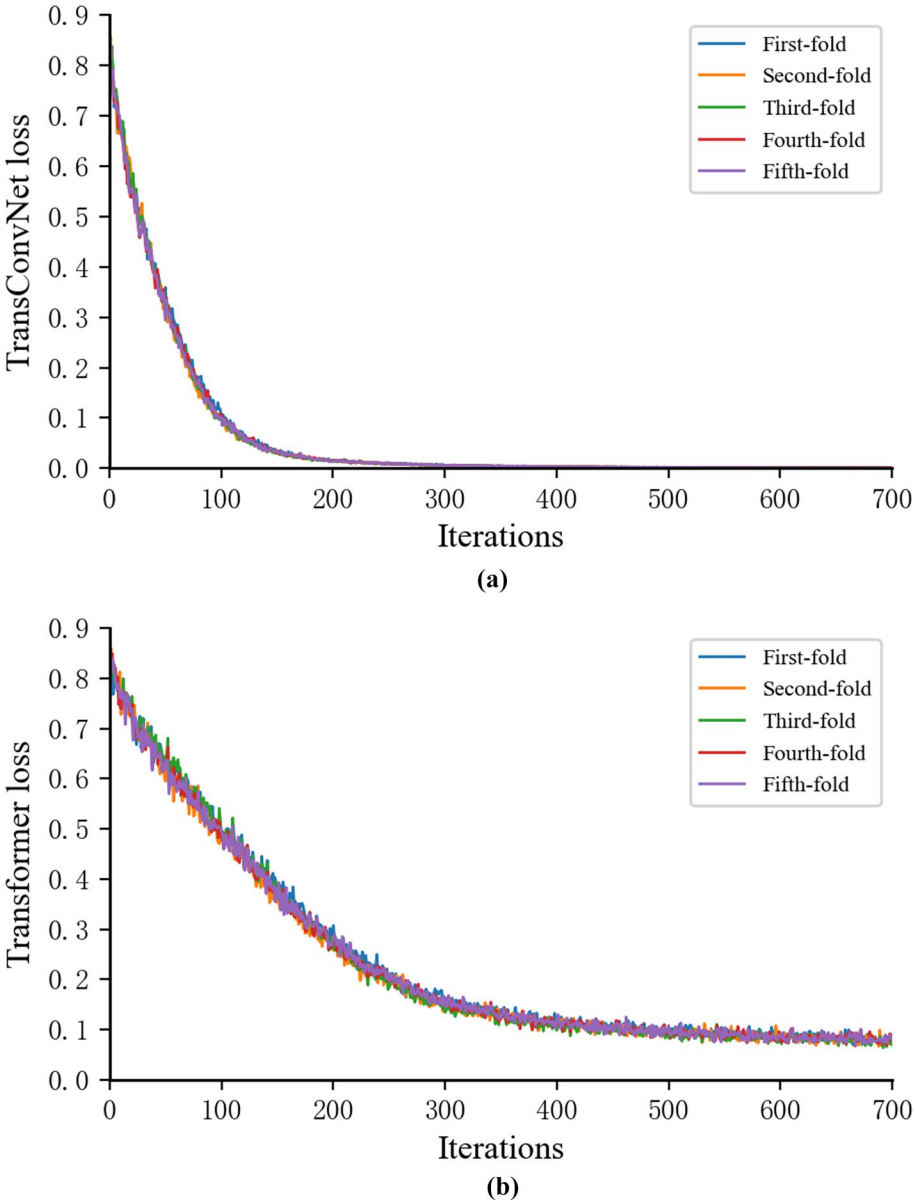


**Fig. 7** **a** Loss curves of the TransConvNet. **b** Loss curves of the transformer

**Table 2** Classification accuracies of the different training and test sets (%)

| Train | Test | Trans former | LSTM | CNN | XGBoost | SVM | RF | DT | Trans ConvNet |
|---|---|---|---|---|---|---|---|---|---|
| 174 | 406 | 60.83 | 57.60 | 59.11 | 53.68 | 59.95 | 53.25 | 53.30 | 61.87 |
| 232 | 348 | 62.22 | 58.85 | 59.02 | 52.01 | 61.26 | 54.31 | 52.18 | 64.82 |
| 290 | 290 | 68.34 | 60.06 | 63.70 | 52.14 | 65.25 | 52.06 | 52.34 | 70.82 |
| 348 | 232 | 69.36 | 61.98 | 63.88 | 53.70 | 66.78 | 54.39 | 50.95 | 72.70 |
| 406 | 174 | 74.24 | 62.87 | 65.06 | 52.53 | 69.65 | 55.40 | 51.38 | 80.36 |

Table 2 demonstrates that when the number of samples in the training set is limited, all models face difficulties in sufficiently learning the intricate relationships within the data, particularly in the high-dimensional feature space. In such cases, models may fail to capture potentially useful features, leading to lower classification accuracy across all the models. As the training set size increases, the transformer, LSTM, CNN, SVM, and TransConvNet models are able to learn more features from the data, thereby improving their classification accuracy. However, the classification outcomes of XGBoost, RF, and DT remain unsatisfactory and exhibit minimal improvement, suggesting that these three models may struggle to effectively capture the nonlinear relationships in metabolomics data and may lack robustness when confronted with problems characterized by small sample sizes. Thus, DL models like the transformer, LSTM, CNN, SVM, and TransConvNet are more effective at extracting useful feature representations from limited samples.

The experimental results show that TransConvNet, which combines CNN and transformer-based components, outperforms other classification algorithms when classifying metabolomics data of varying sizes. These results suggest that TransConvNet has strong robustness and generalization performance, making it well-suited to handle datasets of different sizes.

We are committed to continuously improving and updating our models as we iterate on our DL algorithms. Subsequent studies will witness the integration of additional pretrained models into our DL framework aimed at facilitating the screening of metabolic markers. Our goal is to provide clinicians with more precise guidance that will ultimately benefit patients.

#### Five-fold cross-validation classification experiment for lung cancer

To validate the generalization capability of the TransConvNet model, we conducted experiments using metabolomics data from lung cancer patients. The dataset consists of serum and plasma data from 181 named metabolites obtained from the NMDR website under project number ST000369. We compared the plasma dataset from the first independent case–control study (ADC1), which includes 51 adenocarcinoma lung cancer samples and 32 healthy controls. In this study, we applied the same recursive feature elimination (SVM-RFE) algorithm described in Sect. "Data set description and preprocessing" for feature selection to optimize the input feature set for the model. Through recursive feature elimination, we removed features that contributed minimally to the model's predictive ability and whose exclusion did not significantly affect the model's performance. The top 180 features were retained for subsequent analysis. Additionally, due to the small sample size, we employed the SMOTE data augmentation algorithm to enhance the dataset. This method doubled the number of both positive and negative samples, resulting in a final sample size of 166. Similarly, we conducted the same comparative experiments as described in Sect. "Five-fold cross-validation classification experiment", recording the average results of the evaluation metrics computed on the test set during five-fold cross-validation.

As shown in Table 3, validation of the lung cancer dataset demonstrated that the proposed TransConvNet model exhibited stable performance. Specifically, the model achieved favorable results in key performance metrics, such as accuracy and AUC,

**Table 3** Five-fold cross-validation results for eight models on lung cancer data

| Model | Sensitivity (%) | Specificity (%) | Accuracy (%) | AUC |
|---|---|---|---|---|
| Transformer | 78.31 | 72.95 | 76.21 | 0.78 |
| LSTM | 79.37 | 58.84 | 71.36 | 0.73 |
| CNN | 76.36 | 66.54 | 72.57 | 0.77 |
| XGBoost | 59.59 | 55.77 | 58.01 | 0.58 |
| SVM | 79.37 | 55.64 | 70.11 | 0.68 |
| RF | 72.35 | 55.26 | 65.95 | 0.64 |
| DT | 69.27 | 47.82 | 61.03 | 0.59 |
| TransConvNet | 82.13 | 82.69 | 82.31 | 0.84 |

while also showing strong discriminative ability in terms of sensitivity and specificity. The improvement in the AUC value, in particular, highlights the model's superior ability to distinguish between different classes. These results suggest that the proposed method performs well in classifying lung cancer metabolomics data, effectively differentiating between distinct sample categories.

#### Screening of key biomarkers for prostate cancer

The screening of biomarkers is also particularly important compared to the classification of metabolomics data. In this paper, our proposed TransConvNet model has good classification performance for high-dimensional, high-noise, and high-variability metabolomics data, and through the self-attention mechanism, the model is capable of generating the attention weights for each feature, which reflects the importance of different features in prostate carcinogenesis. We take the feature weights obtained by TransConvNet as a research object and introduce a MI-RF algorithm. This approach utilizes MI-RF to effectively screen for potential biomarkers in prostate cancer metabolomics data.

The MI-RF algorithm for biomarker selection consists of three steps: Firstly, the mutual information between the feature weights X and the target variable Y is computed to evaluate the relationship between the features and the target, with a mutual information score assigned to each feature. Subsequently, an RF model is trained, and the relative importance score for each feature is computed based on the model's training results. Finally, by performing a weighted fusion of the mutual information scores and the feature importance scores obtained from the RF model, a combined weight coefficient for each feature is derived. These features are then ranked, and the top 10 features with the highest diagnostic potential are selected. The formula is as follows:

$$coefficients(X_i) = MI(X_i; Y) \cdot RF_{importance}(X_i)$$

Here $MI(X_i; Y)$ represents the mutual information score between the feature $X_i$ and the target variable $Y$, quantifying their dependency. $RF_{importance}(X_i)$ denotes the feature importance score assigned to $X_i$ by the RF model, reflecting its contribution to the predictive task. This formula multiplies the two scores to obtain the comprehensive weight coefficient $coefficients(X_i)$ for each feature.

Mutual information (MI) [27] quantifies the dependency between a feature $X_i$ and the target variable $Y$. Larger values of mutual information indicate stronger dependence between $X_i$ and Y. The formula is as follows:

$$MI(X_i; Y) = \sum_{x_i \in X_i} \sum_{y \in Y} P(x_i, y) log \frac{P(x_i, y)}{P(x_i)P(y)}$$

Here $MI(X_i; Y)$ is the mutual information between the feature $X_i$ and the target variable $Y$, $P(x_i, y)$ represents the joint probability distribution of the feature $X_i$ and the target variable $Y$, quantifying the likelihood of their co-occurrence within the given dataset. $P(x_i)$ and $P(y)$ represent the marginal probability distributions of the feature $X_i$ and the target variable $Y$, respectively, reflecting the individual likelihood of each variable occurring independently.

To evaluate the effectiveness of the MI-RF algorithm in selecting key biomarkers for prostate cancer, we compared its performance with traditional feature selection methods, including RF, SVM, and the Least Absolute Shrinkage and Selection Operator (Lasso). The focus was on examining the number and consistency of significant prostate cancer biomarkers selected by each method. The experimental results are shown in Table 4.

The experimental results reveal that among the top 10 identified biomarkers, the MI-RF method successfully pinpointed four known significant biomarkers associated with prostate cancer: serotonin, sphinganine, sarcosine, and citrate. In contrast, the Lasso method identified a single biomarker, citrate; RF selected two biomarkers, sphinganine and serotonin, while the SVM recognized one biomarker, sphinganine. Furthermore, the four prostate cancer biomarkers identified by MI-RF have been extensively validated in existing literature and are closely associated with the onset and progression of prostate cancer [28]. To provide a clearer representation of the MI-RF method's selection results, Fig. 8 presents the top 10 key biomarkers identified by the MI-RF method.

**Table 4** Top 10 biomarkers selected by each algorithm

| Rank | MI-RF | Lasso | RF | SVM |
|---|---|---|---|---|
| 1 | 2,3-dihydroxy-2-methylbutyrate | 2,3-diphosphoglycerate | 1-(1-enyl-palmit-oyl)-2-linoleoyl-GPE | Oleoyl-linoleoyl-glycerol |
| 2 | Pantothenate | Hexadecenedioate | 2,3-dihydroxy-2-methylbutyrate | Dihomo-linoleoylcarnitine |
| 3 | Serotonin | 2-aminoheptanoate | Pantothenate | 1-methyl-4-imidazoleacetate |
| 4 | 1-(1-enylpalmitoyl)-2-linoleoyl-GPE | Hexadecasphingosine | Sphinganine | Phenyllactate |
| 5 | Sphinganine | Maleate | N-acetylphenyla-lanine | 4-acetylcatechol sulfate |
| 6 | Sarcosine | Octadecadienedioate | Serotonin | Sphingomyelin |
| 7 | 1-methyl-4-Imida-zoleacetate | Caproate | N-acetyl-isoputre-anine | Sphinganine |
| 8 | N-acetylphenylal-anine | N-palmitoylserine | 1-methyl-4-imidazo-leacetate | 1-oleoyl-GPG |
| 9 | Citrate | Citrate | X–26,109 | 3-methoxytyramine sulfate |
| 10 | X–26,109 | Heptenedioate | 3-hydroxydodec-anedioate | Stearidonate |

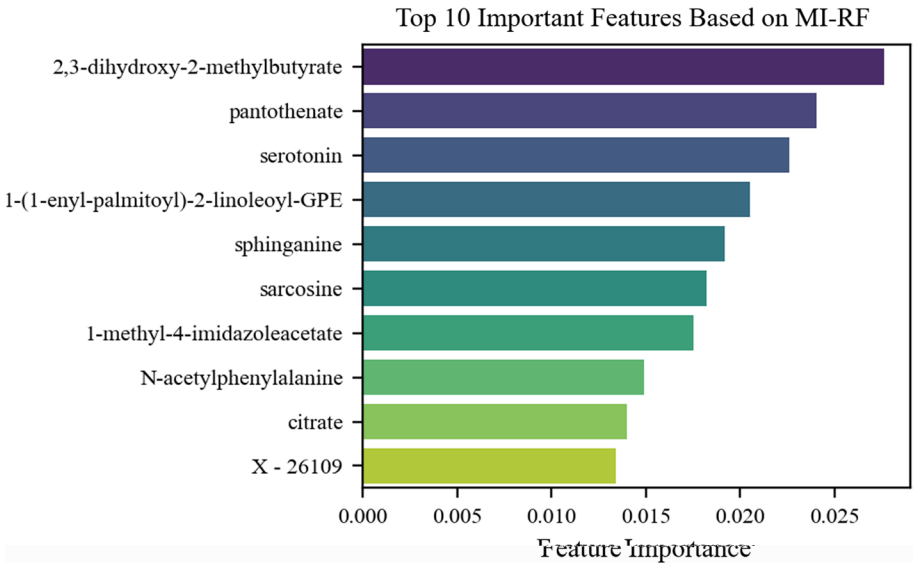


**Fig. 8** The top 10 biomarkers identified by the MI-RF method

These results demonstrate that the MI-RF method exhibits a clear advantage over traditional approaches such as Lasso, RF, and SVM in identifying known biomarkers associated with prostate cancer. In particular, by combining mutual information metrics with the feature importance scores from random forests, the MI-RF method can more comprehensively capture the complex relationships between features and the target variable. The ensemble learning nature of random forests further enhances the model's robustness to noise and its ability to discriminate between features, thereby identifying additional key metabolites closely associated with the occurrence of prostate cancer. Lasso, being a feature selection method based on linear regression models, may be limited in handling high-dimensional, nonlinear relationships, resulting in the selection of only one prostate cancer biomarker. Although RF are capable of capturing nonlinear relationships between features, their feature importance ranking can be influenced by model instability, leading to a relatively smaller set of selected biomarkers. SVM are powerful classification models, but they typically require extensive feature engineering and data preprocessing and are sensitive to hyperparameters, which contributed to their relatively weaker performance in this study compared to MI-RF.

In summary, MI-RF, by integrating mutual information metrics with the random forest algorithm, demonstrates its superiority in identifying key biomarkers for prostate cancer. Compared to other methods, MI-RF can select a greater number of clinically significant biomarkers, thereby offering valuable insights for early disease diagnosis and personalized treatment. Future research could further explore the application of MI-RF in other cancer types or diseases, expanding its potential to contribute to precision medicine.

## Conclusions

Prostate cancer is one of the most prevalent malignancies in men, highlighting the critical importance of early prediction and diagnosis. In this paper, we propose a hybrid model, TransConvNet, to classify prostate cancer metabolomics data. We employed

five-fold cross-validation to assess the model's accuracy, AUC, and other key metrics. When compared with seven other common machine learning methods—transformer, CNN, LSTM, XGBoost, SVM, RF, and DT, TransConvNet demonstrated a significant advantage across all evaluation metrics. Specifically, TransConvNet achieved an accuracy of 81.03%, an AUC of 0.89, and a sensitivity of 81.67%. In contrast, the traditional support vector machine model exhibited an accuracy of 68.47%, the DT model had a sensitivity of 54.80%, and the XGBoost model recorded an AUC of 0.58. These results highlight that TransConvNet, when applied to high-dimensional, noisy metabolomics data, significantly enhances classification performance and model stability, making it more suitable for metabolomics data classification. To further validate the classification performance of the TransConvNet algorithm for metabolomics data and the robustness of the model, we conducted classification experiments using different training and test datasets. The experimental results indicate that TransConvNet can discern superior features from metabolomics data, even with a reduced training dataset. In addition, we validated the generalization ability of the TransConvNet model using a lung cancer metabolomics dataset. The results demonstrate that the TransConvNet model maintains stable performance on the lung cancer metabolomics dataset, with its classification accuracy and AUC values significantly outperforming traditional comparison methods. This result indicates that TransConvNet can effectively handle metabolomics data with different sources and feature distributions. Despite the challenges of high dimensionality, noise, and complexity in metabolomics data, TransConvNet is able to maintain strong classification performance.

In biomarker selection, the MI-RF algorithm demonstrated excellent performance. By combining mutual information scores with feature importance scores from random forests, we successfully identified several key biomarkers closely related to prostate cancer. Compared to traditional selection methods, MI-RF is more effective at capturing nonlinear relationships between features, allowing it to identify more potential biomarkers. This finding not only confirms the advantages of TransConvNet in metabolomics classification tasks but also further supports the potential of MI-RF in metabolomics data applications.

In summary, the TransConvNet model proposed in this study outperforms traditional methods in prostate cancer metabolomics data classification. Comparisons with conventional approaches demonstrate that the TransConvNet model effectively handles high-dimensional, high-noise, and high-variability metabolomics data. it exhibits strong classification performance and robust generalization capabilities across different datasets. Furthermore, the integration of the MI-RF model for biomarker selection successfully identified key features associated with prostate cancer, providing valuable insights for early clinical diagnosis. This method not only performed excellently in experimental data but also holds significant potential for application in clinical practice. TransConvNet has the potential to assist healthcare professionals in extracting meaningful biomarkers from complex metabolomics data, supporting early detection and personalized treatment of prostate cancer. In the future, the TransConvNet model is expected to expand its application scope by integrating more clinical data and multi-omics information, providing a more intelligent and efficient tool for early warning, precise diagnosis, and optimization of treatment strategies for prostate cancer.

**Acknowledgements**
The authors would like to thank the study participants.

**Author contributions**
Liqiang Sun conceived the study, trained models, and drafted the manuscript. Yunwei Zhao and Qi Zhang collected the datasets. Xiaojing Fan reviewed relevant literature and helped refine the research methodology. Mingyang Jiang guided the model design, analyzed the results, and edited the manuscript. All authors read and approved the final draft.

**Funding**
This work was supported by the National Natural Science Foundation of China (62162049), the Innovative Research Team in Universities of Inner Mongolia Autonomous Region (NMGIRT2417), the Science and Technology Projects of Inner Mongolia Autonomous Region (2020GG0190), the Natural Science Foundation of Inner Mongolia Autonomous Region of China (2021LHMS06007), Inner Mongolia Autonomous Region Universities Basic Research (GXKY22127), and the Inner Mongolia University for Nationalities doctoral research start fund project (BS543).

**Availability of data and material**
The utilized data is publicly available; Respective links can be found in the paper.

## Declarations

**Ethics approval and consent to participate**
Not applicable. The data used in this study were obtained from publicly available datasets, and no human participants were directly involved.

**Consent for publication**
Not applicable.

**Competing interests**
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.

Received: 29 May 2024 Accepted: 16 December 2024
Published online: 26 December 2024

**References**

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics. CA: Cancer J Clin. 2022;72:733.
2. Jemal A, Tiwari RC, Murray T, Ghafoor A, Samuels A, Ward E, Feuer EJ, Thun MJ. Cancer statistics. CA: Cancer J Clin. 2004;54:8–29.
3. Huang S, Chong N, Lewis NE, Jia W, Xie G, Garmire LX. Novel personalized pathway-based metabolomics models reveal key metabolic pathways for breast cancer diagnosis. Genom Med. 2016;8:34.
4. Ussher JR, Elmariah S, Gerszten RE, Dyck JRB. The emerging role of metabolomics in the diagnosis and prognosis of cardiovascular disease. J Am Coll Cardiol. 2016;68:2850–70.
5. Vandergrift LA, Decelle EA, Kurth J, Wu S, Fuss TL, DeFeo EM, Halpern EF, Taupitz M, McDougal WS, Olumi AF, Wu C-L, Cheng LL. Metabolomic prediction of human prostate cancer aggressiveness: magnetic resonance spectroscopy of histologically benign tissue. Sci Rep. 2018;8:4997.
6. Su B, Luo P, Yang Z, Yu P, Li Z, Yin P, Zhou L, Fan J, Huang X, Lin X, Qiao Y, Xu G. A novel analysis method for biomarker identification based on horizontal relationship: identifying potential biomarkers from large-scale hepatocellular carcinoma metabolomics data. Anal Bioanal Chem. 2019;411:6377–86.
7. Truong Y, Lin X, Beecher C, Learning a complex metabolomic dataset using random forests and support vector machines. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, (2004), pp. 835–840.
8. Huang J-H, Yan J, Wu Q-H, Duarte Ferro M, Yi L-Z, Lu H-M, Xu Q-S, Liang Y-Z. Selective of informative metabolites using random forests based on model population analysis. Talanta. 2013;117:549–55.
9. Mahadevan S, Shah SL, Marrie TJ, Slupsky CM. Analysis of metabolomic data using support vector machines. Anal Chem. 2008;80(19):7562–70.
10. Aisha A-S, Mahmoud A-A, Yaser J, Fumie C. Visual question answering in the medical domain based on deep learning approaches: a comprehensive study. Pattern Recogn Lett. 2021;150:57–75.
11. Bengio Y. Learning deep architectures for AI. Found Trends® Mach Learn. 2009;2(1):1–127.
12. Chen P, Li Y, Zhou H, Liu B, Liu P. Detection of small ship objects using anchor boxes cluster and feature pyramid network model for SAR imagery. J Mar Sci Eng. 2020;8:112.
13. Alakwaa FM, Chaudhary K, Garmire LX. Deep learning accurately predicts estrogen receptor status in breast cancer metabolomics data. J Proteome Res. 2017;17:337–47.
14. Yuyang S, Weiyu M, Gang L, Xiaobing Z, Henry HYT, Yuefei W, Kefeng L. MetDIT: transforming and analyzing clinical metabolomics data with convolutional neural networks. Anal Chem. 2024;96:2949–57.
15. Jo T, Kim J, Bice P, Huynh K, Wang T, Meikle PJ, Kaddurah-Daouk R, Nho K, Saykin AJ. Circular-SWAT for deep learning based diagnostic classification of Alzheimer's disease: application to metabolome data. EBioMedicine. 2023;97:104820.
16. Date Y, Kikuchi J. Application of a deep neural network to metabolomics studies and its performance in determining important variables. Anal Chem. 2018;90:1805–10.

17. Ashish V, Noam S, Niki P, Jakob U, Llion J, Aidan NG, Lukasz K, Illia P Attention is all you need, arXiv - CS - machine learning, pp. 6000–6010 (2017)
18. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017;60:84–90.
19. He Z, Liu Z, Gong L. Biomarker identification and pathway analysis of rheumatoid arthritis based on metabolomics in combination with ingenuity pathway analysis. Proteomics. 2021;21:2100037.
20. Choi SR, Lee M. Transformer architecture and attention mechanisms in genome data analysis: a comprehensive review. Biology. 2023;12:1033.
21. He K, Zhang X, Ren S, Sun J Deep residual learning for image recognition, In: 2016 IEEE Conference on computer vision and pattern recognition (CVPR), (2016), pp. 770–778.
22. Zhou H, Tan W, Shi S. DeepGpgs: a novel deep learning framework for predicting arginine methylation sites combined with Gaussian prior and gated self-attention mechanism. Brief Bioinf. 2023;24:bbad018.
23. Huang M-L, Hung Y-H, Lee WM, Li RK, Jiang B-R. SVM-RFE based feature selection and Taguchi parameters optimization for multiclass SVM classifier. Sci World J. 2014;2014: 795624.
24. Kha Q-H, Tran T-O, Nguyen T-T-D, Nguyen V-N, Than K, Le NQK. An interpretable deep learning model for classifying adaptor protein complexes from sequence information. Methods. 2022;207:90–6.
25. Soper DS. Greed is good: rapid hyperparameter optimization and model selection using greedy k-fold cross validation. Electronics. 2021;10:1973.
26. Takuya A, Shotaro S, Toshihiko Y, Takeru O, Masanori K, Optuna: anext-generation hyperparameter optimization framework, arXiv - CS - Machine learning, (2019).
27. Zhou H, Wang X, Zhu R. Feature selection based on mutual information with correlation coefficient. Appl Intell. 2021;52:5457–74.
28. Lin X, Lécuyer L, Liu X, Triba MN, Deschasaux-Tanguy M, Demidem A, Liu Z, Palama T, Rossary A, Vasson M-P, et al. Plasma metabolomics for discovery of early metabolic markers of prostate cancer based on ultra-high-performance liquid chromatography-high resolution mass spectrometry. Cancers. 2021;13:3140.

## Publisher's Note