Discover Oncology

Analysis

Advanced machine learning framework for enhancing breast cancer diagnostics through transcriptomic profiling

Mohamed J. Saadh[1] · Hanan Hassan Ahmed[2] · Radhwan Abdul Kareem[3] · Anupam Yadav[4] · Subbulakshmi Ganesan[5] · Aman Shankhyan[6] · Girish Chandra Sharma[7] · K. Satyam Naidu[8] · Akmal Rakhmatullaev[9] · Hayder Naji Sameer[10] · Ahmed Yaseen[11] · Zainab H. Athab[12] · Mohaned Adil[13] · Bagher Farhood[14]

Received: 21 November 2024 / Accepted: 10 March 2025
Published online: 17 March 2025

Abstract

Purpose This study proposes an advanced machine learning (ML) framework for breast cancer diagnostics by integrating transcriptomic profiling with optimized feature selection and classification techniques.

Materials and methods A dataset of 1759 samples (987 breast cancer patients, 772 healthy controls) was analyzed using Recursive Feature Elimination, Boruta, and ElasticNet for feature selection. Dimensionality reduction techniques, including Non-Negative Matrix Factorization (NMF), Autoencoders, and transformer-based embeddings (BioBERT, DNABERT), were applied to enhance model interpretability. Classifiers such as XGBoost, LightGBM, ensemble voting, Multi-Layer Perceptron, and Stacking were trained using grid search and cross-validation. Model evaluation was conducted using accuracy, AUC, MCC, Kappa Score, ROC, and PR curves, with external validation performed on an independent dataset of 175 samples.

Results XGBoost and LightGBM achieved the highest test accuracies (0.91 and 0.90) and AUC values (up to 0.92), particularly with NMF and BioBERT. The ensemble Voting method exhibited the best external accuracy (0.92), confirming its robustness. Transformer-based embeddings and advanced feature selection techniques significantly improved model performance compared to conventional approaches like PCA and Decision Trees.

Conclusion The proposed ML framework enhances diagnostic accuracy and interpretability, demonstrating strong generalizability on an external dataset. These findings highlight its potential for precision oncology and personalized breast cancer diagnostics.

Keywords Breast cancer · Machine learning · Transcriptomic profiling · Feature selection · Predictive modeling · Biomarkers

✉ Bagher Farhood, farhood-b@kaums.ac.ir; bffarhood@gmail.com | [1]Faculty of Pharmacy, Middle East University, Amman 11831, Jordan. [2]College of Pharmacy, Alnoor University, Mosul, Iraq. [3]Ahl Al Bayt University, Kerbala, Iraq. [4]Department of Computer Engineering and Application, GLA University, Mathura 281406, India. [5]Department of Chemistry and Biochemistry, School of Sciences, JAIN (Deemed to Be University), Bangalore, Karnataka, India. [6]Centre for Research Impact and Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, Punjab 140401, India. [7]Department of Applied Sciences-Chemistry, NIMS Institute of Engineering and Technology, NIMS University Rajasthan, Jaipur, India. [8]Department of Chemistry, Raghu Engineering College, Visakhapatnam, Andhra Pradesh 531162, India. [9]Department of Faculty Pediatric Surgery, Tashkent Pediatric Medical Institute, Bogishamol Street 223, 100140 Tashkent, Uzbekistan. [10]Collage of Pharmacy, National University of Science and Technology, Dhi Qar 64001, Iraq. [11]Gilgamesh Ahliya University, Baghdad, Iraq. [12]Department of Pharmacy, Al-Zahrawi University College, Karbala, Iraq. [13]Pharmacy College, Al-Farahidi University, Baghdad, Iraq. [14]Department of Medical Physics and Radiology, Faculty of Paramedical Sciences, Kashan University of Medical Sciences, Kashan, Iran.

1 Introduction

Breast cancer remains one of the leading causes of cancer-related deaths globally, posing significant clinical and social challenges [1–5]. While there have been notable advancements in diagnostic and treatment methods, traditional diagnostic tools like imaging and histopathological evaluations still have limitations [6]. These conventional methods often struggle to detect breast cancer at its earliest stages and fail to provide personalized insights into disease progression. Therefore, there is an increasing demand for more precise and efficient diagnostic models that can capture the complex molecular nature of breast cancer and improve patient outcomes [7–9]. In this context, machine learning (ML) has emerged as a powerful tool that can revolutionize breast cancer diagnostics by integrating high-dimensional genetic and transcriptomic data [10–12].

Transcriptomic profiling, which examines the full set of RNA transcripts in a cell, has shown promise in identifying gene expression patterns and molecular markers linked to cancer [13–16]. For breast cancer, this approach allows for the analysis of gene expression differences across various disease stages and subtypes [17]. However, the sheer volume of transcriptomic data, which often includes thousands of genes with intricate interactions, presents a challenge in identifying relevant biomarkers for building accurate diagnostic models. This is where ML, paired with advanced feature selection techniques, becomes crucial [18–20].

ML techniques can identify complex patterns within large datasets that are difficult to detect using conventional statistical methods [21–28]. These algorithms are capable of analyzing vast numbers of genes at once, distinguishing between healthy and cancerous tissues, and identifying gene expression patterns that suggest an aggressive form of cancer [29–34]. Furthermore, advanced feature selection methods help isolate the most informative genes, reducing noise and improving the model's interpretability and accuracy. By focusing on key gene expressions and their interactions, ML-based models can facilitate early detection, enhance predictive accuracy for patient outcomes, and identify potential therapeutic targets [35–37]. Despite the vast potential of ML in breast cancer diagnostics, many current studies fail to fully leverage optimized feature selection techniques for high-dimensional genomic data. Prior research often relies on conventional methods like Principal Component Analysis (PCA) and manual selection, which, while effective for dimensionality reduction, can overlook key biomarkers due to their focus on linear transformations [38, 39]. For example, Alghanim et al. [40] applied PCA in a multi-omics context but did not assess its ability to isolate gene-level features critical for breast cancer classification. Similarly, Mirza et al. [41] utilized Recursive Feature Elimination with Cross-Validation (RFECV) but limited their scope to single-method approaches, resulting in modest predictive accuracy.

These limitations are exacerbated by the "curse of dimensionality" inherent in transcriptomic datasets, where irrelevant or redundant features can lead to overfitting and decreased model reliability [42–47]. While some studies [48–50] have introduced hybrid techniques like ElasticNet and LASSO, these methods often neglect non-linear relationships between genes, further hindering the identification of intricate patterns critical for early diagnosis. This underutilization of optimized and integrative feature selection methods highlights a significant gap in the literature, underscoring the need for comprehensive frameworks capable of isolating predictive biomarkers while preserving biological complexity.

In recent years, transformer-based models have revolutionized sequence analysis by providing context-aware embeddings that capture complex dependencies within biological data [51–55]. Unlike traditional k-mer methods or manually crafted features, these models leverage self-attention mechanisms to account for long-range relationships, enabling the extraction of nuanced patterns from genomic and RNA sequences. DNABERT, DNABERT-2, and BioBERT, as examples of such transformer architectures, have demonstrated remarkable capability in encoding sequence information into meaningful numerical representations [56–62]. These embeddings not only preserve essential sequence context but also enhance the downstream predictive performance of various machine learning classifiers. By utilizing these pretrained models, researchers can significantly improve the efficiency and accuracy of bioinformatics tasks, ranging from variant effect prediction to functional sequence annotation, without the need for extensive feature engineering. This approach exemplifies how state-of-the-art deep learning methodologies are increasingly becoming indispensable tools in genomic research [63].

Cancer diagnosis and treatment have significantly benefited from machine learning advancements, particularly in predictive modeling and biomarker discovery. Several recent studies have explored innovative computational techniques for cancer classification, antimicrobial peptide identification, and feature selection to enhance diagnostic accuracy and therapeutic decision-making. Karim et al. [64] introduced ANNprob-ACPs, a probabilistic feature fusion approach for identifying anticancer peptides (ACPs) with remarkable accuracy (93.72% in cross-validation and 90.62% on an independent test). By leveraging multiple feature encoding techniques and an ensemble of machine learning

models, the study demonstrated the effectiveness of AI in anticancer drug discovery. Additionally, SHAP analysis was employed to highlight key physicochemical properties relevant to drug interactions, reinforcing its potential for future therapeutic applications.

Shaon et al. [65] developed AMP-RNNpro, a two-stage Recurrent Neural Network (RNN) model with probabilistic feature selection, designed to enhance antimicrobial peptide (AMP) identification. With an accuracy of 97.15%, this approach proved effective in drug discovery, particularly in detecting novel AMPs. Feature importance analysis via SHAP further emphasized critical amino acid composition patterns contributing to its success.

Karim et al. [66] introduced StackAMP, a stacking-based ensemble classifier for AMP identification. This model, integrating multiple feature extraction methods, achieved near-perfect accuracy (99.97%), showcasing the power of ensemble learning in biological sequence classification and its transformative potential in biotechnology and healthcare. In cancer classification, Shaon et al. [67] conducted a comparative study of machine learning models for breast cancer prediction, employing LASSO and SHAP-based feature selection on an integrated breast cancer dataset. Their findings revealed that models incorporating SHAP-based selection achieved 99.82% accuracy, highlighting the importance of explainable AI in medical diagnostics and treatment planning. These studies collectively underscore the impact of probabilistic modeling, deep learning, ensemble learning, and explainable AI in cancer research and biomedical applications. Inspired by these advancements, our study integrates MCC, Kappa Score, ROC curve, and PR curve visualization to ensure a robust and interpretable evaluation of cancer classification models, particularly on an external test dataset, ensuring generalizability and real-world applicability.

In this context, the proposed study introduces a multi-method feature selection approach, integrating RFE, Boruta, and ElasticNet to address these gaps and enhance predictive accuracy. This oversight can hinder the identification of important biomarkers and reduce model robustness. In high-dimensional datasets, irrelevant or redundant genes can lead to overfitting, making the model's predictions less reliable. Effective feature selection is necessary to overcome this "curse of dimensionality," allowing ML models to highlight the genes most associated with cancer and improve predictive performance[68–70].

Traditional diagnostic tools, such as imaging and histopathological evaluations, are critical in breast cancer detection but present notable limitations. For instance, imaging modalities often exhibit low sensitivity in detecting early-stage breast cancer, particularly in dense breast tissues, where malignancies can be masked by overlapping structures. Furthermore, histopathological assessments are inherently subjective, relying on the expertise of pathologists, which can lead to variability in interpretation and diagnostic accuracy. These methods also lack the capability to analyze the molecular-level alterations that underpin tumor development and progression, such as gene expression changes and transcriptomic profiles. This gap in molecular understanding limits their ability to provide personalized insights into disease mechanisms or guide tailored treatment strategies effectively.

Addressing these limitations requires advanced diagnostic approaches that combine molecular profiling with computational techniques to deliver high sensitivity, interpretability, and individualized insights. ML methods, when integrated with transcriptomic data, offer a promising avenue to overcome these challenges by enabling precise and early detection while unraveling the molecular heterogeneity of breast cancer.

This study aims to fill this gap by creating an ML framework that integrates advanced feature selection with transcriptomic profiling to enhance the accuracy and reliability of breast cancer predictions. Specifically, the focus is on refining feature selection to pinpoint the key genes whose expression patterns indicate the presence or progression of breast cancer. This approach advances computational oncology by promoting the use of ML for biomarker discovery and breast cancer diagnostics.

Major Contributions of This Study Include:

1. Development of a Robust ML-Based Feature Selection Framework: The study introduces a comprehensive feature selection process to identify critical gene expressions from high-dimensional transcriptomic data. By narrowing the dataset to only the most relevant genes, the model becomes more interpretable and accurate.
2. Enhanced Predictive Model for Early Detection and Prognosis: By integrating optimized ML algorithms, this study produces a model that surpasses traditional methods in prediction accuracy. This model supports early breast cancer detection and helps in understanding disease progression, which can aid clinical decision-making.
3. Identification of High-Potential Biomarkers: The study analyzes transcriptomic profiles in-depth to find new gene biomarkers associated with breast cancer risk and progression. These biomarkers could serve as diagnostic indicators or targets for personalized therapy, aligning with the principles of precision oncology.

4. Advancements in Personalized Medicine: By exploring the molecular basis of breast cancer, this study supports a move away from one-size-fits-all treatment plans to more tailored therapies. The biomarkers identified could help oncologists customize treatment strategies based on individual molecular profiles, potentially improving patient outcomes.
5. Incorporation of Transformer-Based Embedding Models: The study also integrates transformer-based deep learning models such as DNABERT, DNABERT-2, and BioBERT to generate high-quality embeddings from RNA sequences. These models leverage self-attention mechanisms and multi-layer architectures to capture complex local and global patterns in transcriptomic data. The resulting embeddings enhance the feature selection process, contributing to more refined and accurate predictive models.

In summary, this research marks a significant step forward in breast cancer diagnostics by leveraging ML-driven feature selection and transcriptomic profiling. By highlighting key gene expression patterns and incorporating advanced transformer embeddings, the study not only enhances predictive accuracy but also provides deeper insights into breast cancer biology. Ultimately, it underscores the transformative potential of ML approaches in identifying biomarkers, improving diagnostics, and paving the way for more targeted therapeutic interventions in precision oncology.

2 Materials and methods

2.1 Dataset and preprocessing

Figure 1 presents the multi-model machine learning framework designed for breast cancer diagnostics, integrating advanced feature selection, dimensionality reduction, and robust classifier training for enhanced predictive accuracy. This study analyzed a transcriptomic dataset containing RNA sequencing profiles from 1,759 individuals, which included 987 breast cancer patients and 772 healthy controls. The data was sourced from publicly available repositories, specifically The Cancer Genome Atlas (TCGA) [71, 72], and [73], to ensure a diverse sample and reproducibility of results. Raw RNA-Seq reads were aligned to the human reference genome, and expression levels were normalized using updated methods to reduce batch effects and variability among samples. Quality control measures were implemented to remove samples with low read counts, excessive noise, or inconsistent expression distributions, resulting in a high-quality dataset that captured the molecular heterogeneity of both breast cancer and healthy tissues.

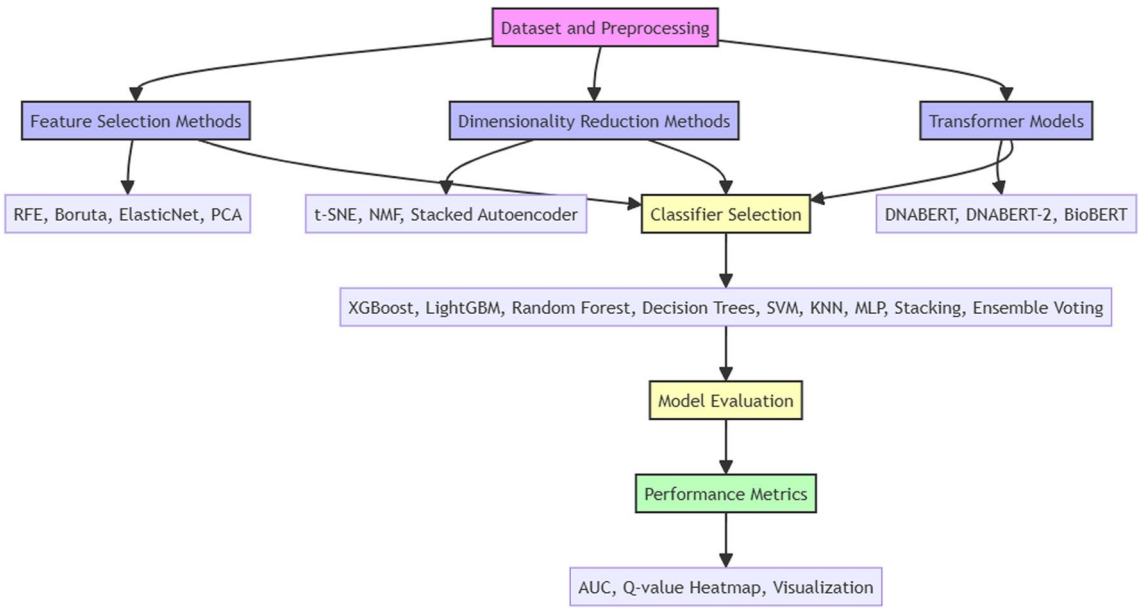

Fig. 1 Comprehensive multi-model machine learning framework for enhanced breast cancer diagnostics

2.2 Advanced feature selection and reduction

To build effective diagnostic models, it was essential to select methods and models that align with the characteristics of high-dimensional transcriptomic data and the specific challenges of breast cancer diagnostics. For feature selection, Recursive Feature Elimination (RFE) was chosen for its iterative capability to prioritize the most informative genes, improving both interpretability and classifier performance. Boruta, a robust method leveraging a random forest framework, was included to identify key gene features by comparing them with randomized "shadow" variables, thus ensuring the retention of only highly predictive features. ElasticNet regularization was employed to address multicollinearity and enhance the selection of correlated features, a critical requirement for transcriptomic datasets. Principal Component Analysis (PCA) was also used as a baseline dimensionality reduction method to evaluate its ability to reduce feature space complexity by focusing on linear transformations. These methods were specifically selected due to their proven ability to minimize redundancy and focus on biologically meaningful genes in genomic studies.

Dimensionality reduction methods were further employed to simplify the dataset while preserving critical patterns. t-Distributed Stochastic Neighbor Embedding (t-SNE) was utilized for its ability to maintain local data structures and visually distinguish cancerous from non-cancerous samples in lower dimensions, making complex interactions more interpretable. Non-negative Matrix Factorization (NMF) was incorporated to retain non-negative gene expression values, revealing biologically relevant patterns while reducing noise. PCA was compared to these advanced methods to benchmark its efficiency and interpretability. Furthermore, a stacked autoencoder was implemented for non-linear dimensionality reduction. By training the autoencoder on transcriptomic data, it learned compact, high-level representations that retained essential biological information. These embeddings served as a basis for more efficient downstream analysis, enabling more accurate predictions and deeper insights into the molecular basis of breast cancer. Together, these methods and models provided a well-rounded framework to navigate the challenges of high-dimensional data, ultimately enhancing diagnostic accuracy and interpretability.

2.3 Transformer-based embedding extraction for RNA sequences

This study leverages three cutting-edge transformer-based models—DNABERT, DNABERT-2, and BioBERT—to extract meaningful embeddings from genomic and RNA sequences. All of these models employ a transformer architecture that incorporates self-attention and multi-head attention mechanisms, enabling them to identify both local sequence motifs and broader sequence-level patterns.

2.3.1 Model architecture and embedding extraction

The underlying architecture consists of multiple transformer encoder layers. Each encoder layer employs a multi-head self-attention mechanism, a feedforward neural network, and residual connections, ensuring robust learning of contextual relationships. The attention mechanism computes alignment scores for different sequence positions, using the formula:

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

where Q, K, and V represent the query, key, and value matrices, and dk is the dimensionality of the keys. The multi-head approach allows the model to capture diverse features by attending to various aspects of the sequence simultaneously. After passing through the transformer layers, the model produces context-rich hidden states from which embeddings are extracted.

2.3.2 Preprocessing and tokenization

Raw RNA sequences were tokenized into overlapping k-mers using DNABERT's specialized tokenizer. This method ensures that each sequence fragment aligns with the pre-trained vocabulary. These k-mer tokens are then mapped into dense vectors through a learned embedding matrix and input into the transformer's encoder stack. Positional encodings are added to maintain sequence order, a critical factor for understanding sequential biological data.

2.3.3 Implementation and training

The transformer models were initialized from pre-trained checkpoints available through the Hugging Face Transformers library and fine-tuned on specific genomic and RNA sequence datasets. Training utilized NVIDIA GPUs to handle the computational demands of deep transformer layers. Batch normalization, dropout regularization, and layer normalization ensured stable and effective model training. The final embeddings were extracted from the top encoder layer, providing highly informative, sequence-contextualized features for downstream tasks.

By applying advanced transformer-based architectures, this study achieved high-quality embeddings that reflect complex sequence relationships. These embeddings were subsequently employed as input features for a variety of machine learning models, yielding improved performance metrics across multiple biological sequence analysis tasks.

2.4 Classifier selection and model training

To handle the complexity of the dataset, classifiers were selected based on their compatibility with the problem's requirements. XGBoost and LightGBM were chosen for their superior handling of structured, high-dimensional data and their ability to model complex feature interactions effectively. Random Forest was employed for its ensemble approach, which reduces overfitting while maintaining robust performance. Simpler models such as Decision Trees were included for their interpretability, despite their tendency to overfit, and Support Vector Machines (SVM) were evaluated for their effectiveness with high-dimensional data through kernel transformations. K-Nearest Neighbors (KNN) was also tested to explore proximity-based classification. Multi-Layer Perceptron (MLP), a deep learning-based model, was incorporated to capture non-linear relationships and complex patterns in the dataset, leveraging its multi-layer architecture for enhanced learning. Stacking, another ensemble technique, was added to integrate predictions from multiple classifiers through a meta-model, maximizing performance by combining the strengths of diverse models. Finally, an ensemble voting model was used to aggregate predictions across classifiers, leveraging their collective strengths for improved accuracy and robustness.

2.5 Model evaluation metrics, statistical analysis and visualization

Model performance was evaluated using balanced accuracy, which accounts for class imbalances to ensure fair assessment across both cancerous and non-cancerous samples. The area under the curve (AUC) was also measured to evaluate the models' discriminatory power in distinguishing between the two sample types. A Q-value heatmap was created to visualize statistical significance across feature selection and classifier combinations. The heatmap, using the Benjamini–Hochberg procedure for multiple comparison adjustments, allowed identification of significant differences, with lower Q-values indicating greater significance. This helped highlight the most effective feature-classifier pairs. Python libraries played a central role in this analysis. Scikit-Learn was used for model training, feature selection, and tuning, while XGBoost and LightGBM provided efficient handling of high-dimensional data. NumPy and Pandas were essential for data manipulation, and SciPy supported statistical testing and Q-value adjustments. Visualizations, including heatmaps, were generated using Matplotlib and Seaborn to display model accuracy and AUC scores across training and testing sets. This comprehensive approach allowed for identifying top-performing model and feature selection combinations and understanding trends in predictive performance. In addition to standard evaluation metrics, MCC, Kappa score, ROC curve, and PR curve (visualization) were employed to comprehensively assess model performance. These metrics ensure a balanced evaluation, particularly in the presence of class imbalance, and validate the robustness of the models on the external test dataset.

The hyperparameter settings and computational requirements for each model were carefully optimized to ensure robust performance and reproducibility. Hyperparameter tuning was performed using grid search with cross-validation, covering a range of values for key parameters. For instance, XGBoost and LightGBM models were tuned for learning rate (0.01–0.3), maximum depth (3–10), and number of estimators (100–500). ElasticNet regularization involved balancing L1 and L2 penalties with alpha values ranging from 0.1 to 1.0. Voting and Stacking ensemble models incorporated base learners optimized individually before integration. Dimensionality reduction techniques like t-SNE were parameterized with perplexity values (5–50) and learning rates (200–1000). Computationally, the experiments were conducted on a system with an Intel Core i9 processor, 64 GB RAM, and an NVIDIA RTX 3090 GPU, ensuring efficient handling of high-dimensional data. These details provide transparency and facilitate reproducibility of the proposed framework.

3 Results

Figures 2 through 5 provide an overview of how different machine learning models performed when combined with feature selection and dimensionality reduction methods. These heatmaps illustrate both accuracy and AUC scores on training and testing datasets, showcasing how each model and feature selection combination performed in breast cancer prediction using transcriptomic data.

3.1 Test accuracy and test AUC analysis

The heatmap for test accuracy highlights significant differences in model performance across various feature selection and dimensionality reduction methods. XGBoost achieved the highest test accuracy of 0.90, particularly when paired with NMF and BioBERT. LightGBM closely followed, reaching a peak accuracy of 0.89 using ElasticNet and BioBERT features. These results demonstrate the strong compatibility of gradient-boosting algorithms with high-dimensional transcriptomic data. Voting and MLP also showed strong accuracy, with Voting achieving a maximum accuracy of 0.91 using BioBERT, while MLP reached 0.90 under the same feature selection method.

In contrast, simpler models like Decision Trees and KNN exhibited consistently lower accuracy. Decision Trees peaked at 0.82 using BioBERT, while KNN achieved a maximum accuracy of 0.86, also with BioBERT. These results suggest that simpler models struggle to capture the complexity of the dataset compared to ensemble and gradient-boosting methods. Feature-wise, BioBERT, NMF, and RFE emerged as the most effective feature extraction techniques, contributing to consistently high accuracy across a range of models. In comparison, methods like t-SNE and PCA often resulted in lower accuracy, underscoring their limitations in handling high-dimensional genomic data (Fig. 2).

The AUC results provide additional insights into the discriminatory power of the models. LightGBM and XGBoost achieved the highest AUC scores of 0.92, both with PCA and NMF embeddings. This demonstrates the ability of these advanced classifiers to effectively model complex feature interactions and distinguish between cancerous and non-cancerous samples. Random Forest also performed well, achieving an AUC of 0.92 when paired with DNABERT embeddings, showcasing its ensemble strength.

Simpler models like Decision Trees and KNN exhibited weaker AUC performance, with peak values of 0.85 and 0.87, respectively, when paired with BioBERT. Among advanced feature extraction methods, BioBERT and NMF consistently led to higher AUC values across most models, further validating their capability to retain biologically relevant patterns. Ensemble methods like Voting achieved a strong AUC of 0.91 with NMF, while Stacking reached 0.90 under the same feature extraction approach. In summary, the results underscore the superior performance of advanced models like LightGBM XGBoost, and Voting, especially when combined with cutting-edge feature extraction techniques such as BioBERT, NMF, and DNABER (Fig. 3).

3.2 Train accuracy and train AUC analysis

The train accuracy heatmap demonstrates the performance of classifiers when trained on the dataset using various feature selection and dimensionality reduction techniques. XGBoost and LightGBM emerged as the best-performing classifiers, consistently achieving high train accuracy across features. XGBoost achieved its peak train accuracy of 0.91 when paired with NMF and BioBERT, showcasing its ability to model complex interactions in transcriptomic data. LightGBM also excelled, reaching its highest train accuracy of 0.90 with BioBERT and t-SNE, further confirming its suitability for high-dimensional genomic datasets.

Ensemble methods such as Voting and Stacking also performed well in training, with Voting achieving the highest train accuracy of 0.92 with BioBERT, and MLP following closely with a maximum accuracy of 0.91. In contrast, simpler models such as Decision Trees and KNN demonstrated lower train accuracy, with Decision Trees peaking at 0.83 using BioBERT, and KNN achieving a maximum train accuracy of 0.87 with the same feature set. These findings indicate the limited capacity of simpler models to effectively fit the training data. Among feature selection methods, BioBERT contributed significantly to high training accuracy across most models, underscoring their ability to extract meaningful biological patterns. Conversely, methods like t-SNE and PCA resulted in comparatively lower training accuracy, indicating their limitations in retaining sufficient feature information for model training.

The train AUC heatmap highlights the ability of classifiers to distinguish between classes effectively during training. LightGBM and XGBoost consistently demonstrated the highest train AUC values, both achieving a peak AUC of 0.93 with

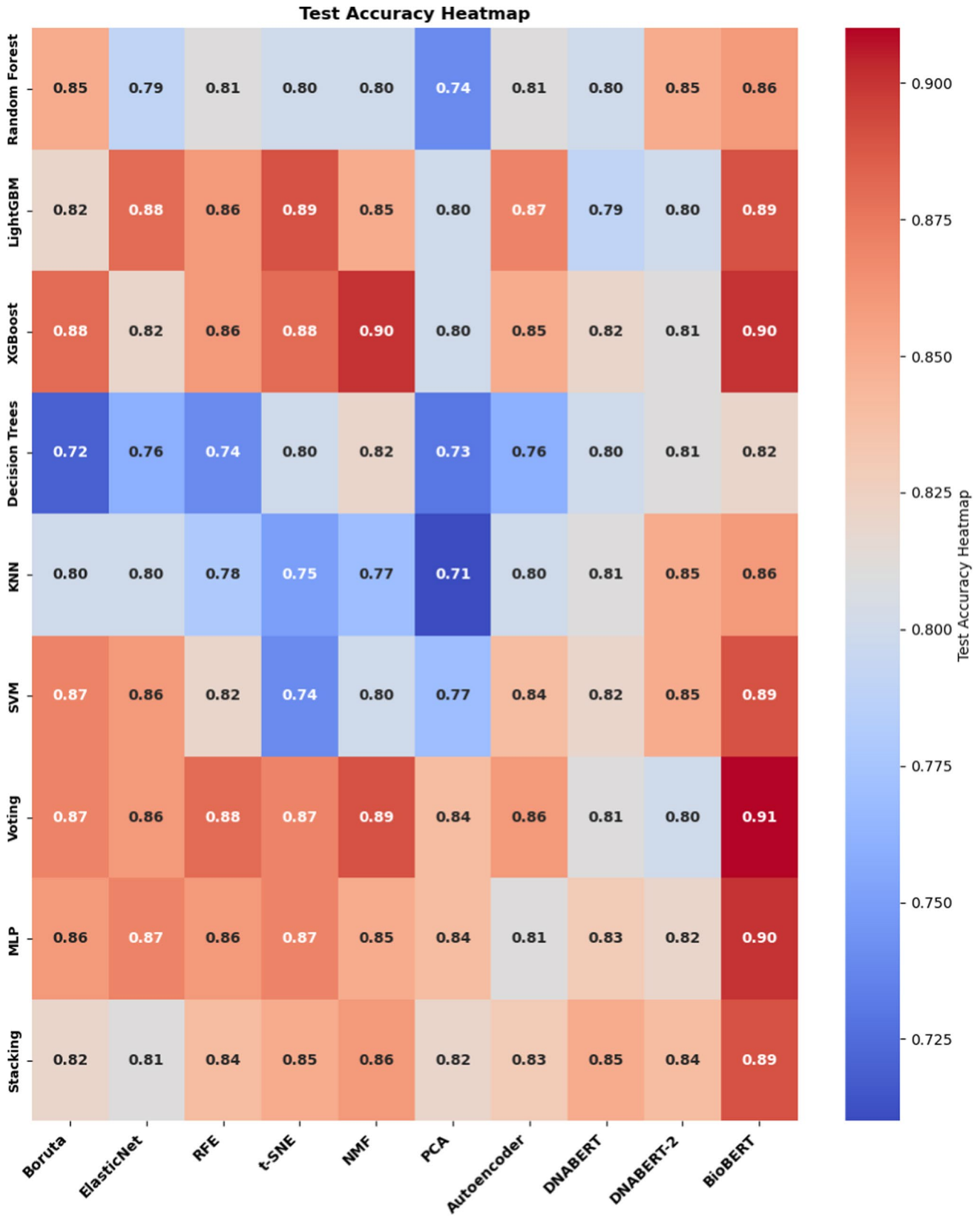

Fig. 2 Test accuracy heatmap

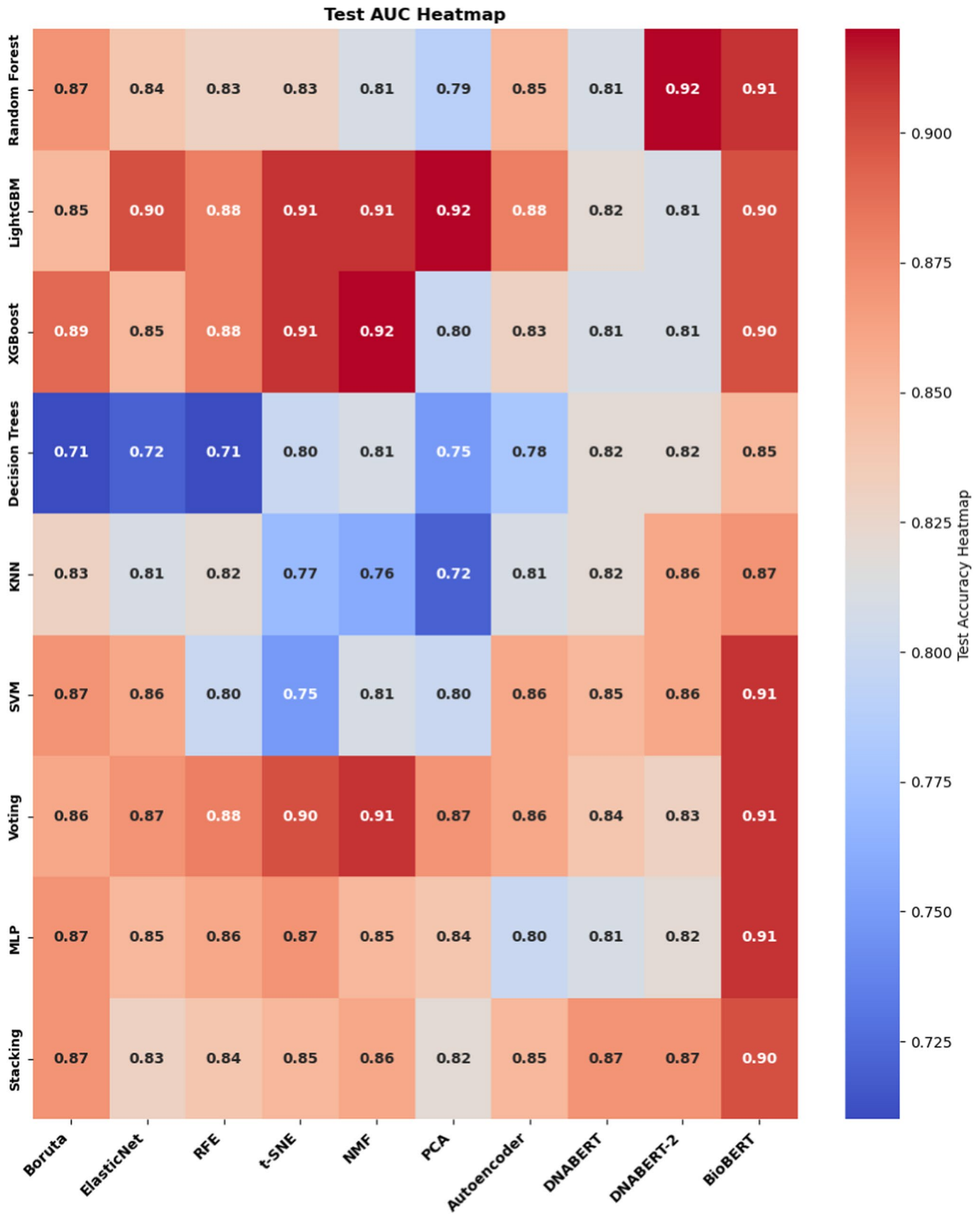

Fig. 3 Test AUC-ROC heatmap

PCA and NMF. These results underline the strength of gradient-boosting classifiers in capturing complex relationships and maximizing discriminatory power. XGBoost also achieved strong AUC values of 0.92 with BioBERT, while LightGBM reached a similar score with BioBERT, confirming the complementary strengths of these advanced classifiers. Ensemble models, particularly Voting, performed exceptionally well in training, achieving the highest AUC of 0.93 with BioBERT, followed closely by MLP, which reached a peak AUC of 0.92 with BioBERT. Simpler models like Decision Trees and KNN showed limited AUC performance, with Decision Trees peaking at 0.86 using BioBERT, and KNN achieving a maximum AUC of 0.89 with BioBERT. These results highlight the inability of simpler models to effectively leverage the complex feature sets provided by advanced feature extraction techniques.

In terms of feature selection, NMF, BioBERT, and RFE proved to be the most effective, consistently contributing to higher AUC scores across models. Transformer-based embeddings such as BioBERT and DNABERT were particularly effective in combination with gradient-boosting and ensemble classifiers, providing a robust feature space for class distinction. In contrast, methods like PCA and t-SNE generally resulted in lower AUC scores, reflecting their limited capacity to capture the full complexity of the transcriptomic dataset (Figs. 4 and 5).

3.3 Q-value heatmap analysis: key model-feature selection combinations

The Q-value heatmap highlights the statistical significance of model-feature selection combinations, emphasizing the reliability of certain pairings in breast cancer diagnostics. Among the classifiers, Stacking and Voting demonstrated the lowest Q-values across various feature selection methods, particularly when combined with advanced techniques like BioBERT and DNABERT embeddings. For instance, Stacking paired with BioBERT achieved a Q-value of 0.010, marking it as a highly significant and robust combination. Similarly, LightGBM with NMF yielded a Q-value of 0.007, underscoring its consistency and effectiveness in high-dimensional transcriptomic analysis. Transformer-based embeddings (BioBERT, DNABERT) and feature reduction techniques like NMF and RFE emerged as the most statistically reliable features, consistently achieving low Q-values across classifiers. In contrast, simpler methods such as PCA and Boruta produced higher Q-values, reflecting lower statistical significance and reproducibility. These findings reinforce the importance of leveraging advanced classifiers and feature selection techniques to achieve statistically robust and reliable diagnostic models. (Fig. 6).

3.4 External dataset accuracy and AUC analysis

Voting also excelled, reaching the highest external accuracy of 0.92 and AUC of 0.91 when paired with BioBERT, demonstrating the robustness of ensemble methods. In contrast, simpler models like Decision Trees and KNN showed lower performance, with maximum accuracy and AUC values of 0.81 and 0.85, respectively, indicating their limitations in handling complex transcriptomic data. Among features, BioBERT were the most effective, consistently contributing to high scores across top classifiers, while PCA and t-SNE lagged in performance. These results confirm the importance of combining advanced feature extraction methods with robust classifiers for reliable performance on external datasets. (Figs. 7 and 8).

The ROC curves for the training dataset illustrate the classification performance of various models. High AUC values across models such as LightGBM, XGBoost, and Stacking indicate their strong ability to distinguish between positive and negative classes. The steep curves approaching the top-left corner suggest that these models have learned well from the training data. However, slight overfitting might be inferred if the test performance deviates significantly from these results (Fig. 9).

The ROC curves for the test and external test dataset assess the generalization capability of the trained models. XGBoost, LightGBM, and Voting classifiers maintain high AUC values (~ 0.90 to 0.92), demonstrating their robustness on unseen data. In contrast, Decision Trees and KNN show a notable drop in performance, with AUC values ranging between 0.75 and 0.85. The comparison between training and test ROC curves highlights how well each model generalizes, ensuring reliability in real-world applications (Fig. 10). These results confirm that ensemble-based models and transformer-based feature representations yield more stable and robust classification performance, particularly when validated on an independent external dataset.

MCC heatmap provides a more balanced evaluation of model performance, particularly in imbalanced classification settings. The highest MCC scores are observed for LightGBM (0.96 with BioBERT), XGBoost (0.96 with NMF), and Voting (0.95 with DNABERT), indicating their superior ability to maintain correlation between predictions and actual labels. In contrast, Decision Trees exhibit the lowest MCC values (0.73–0.89), reinforcing their lower robustness on the external

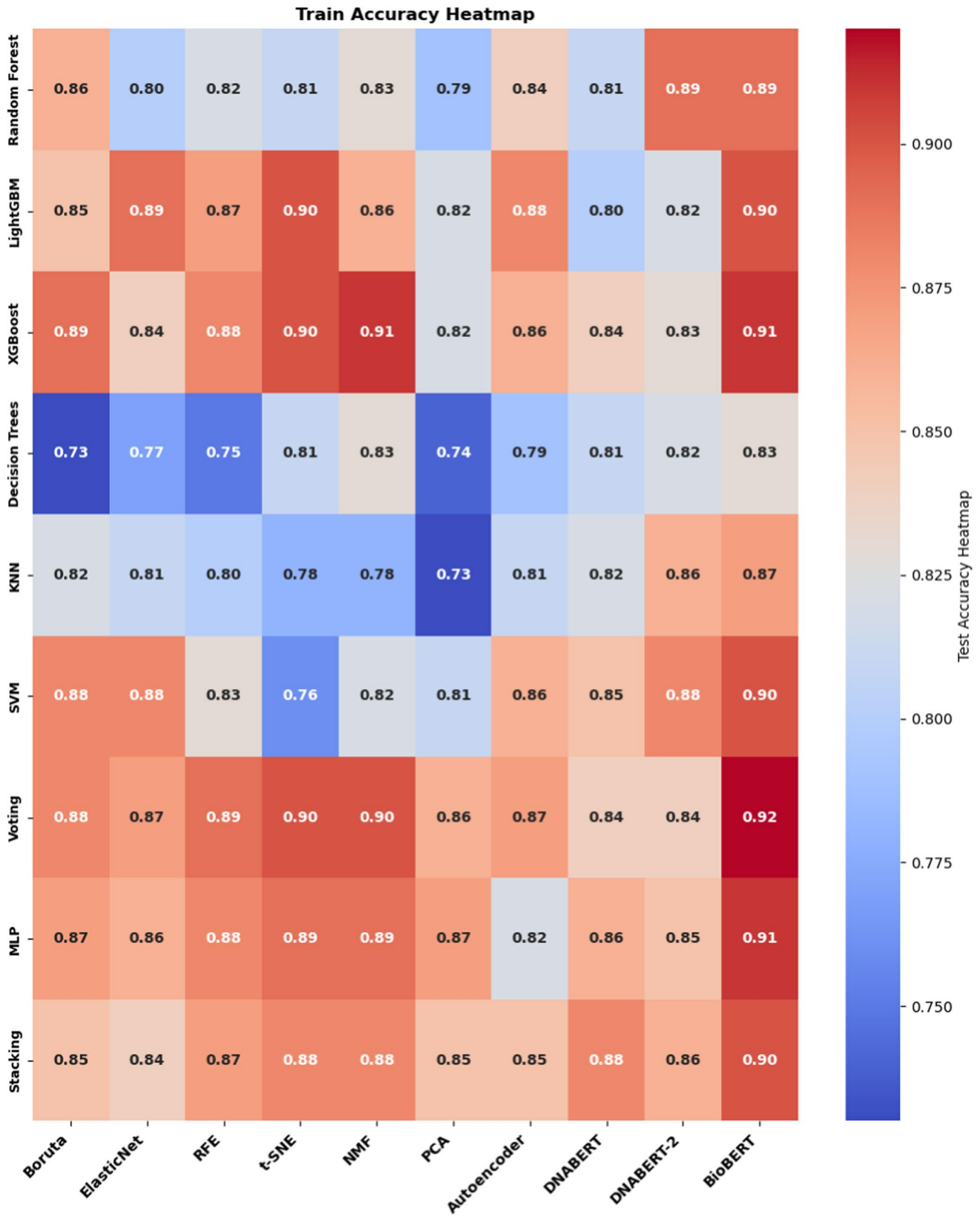

Fig. 4 Train accuracy heatmap

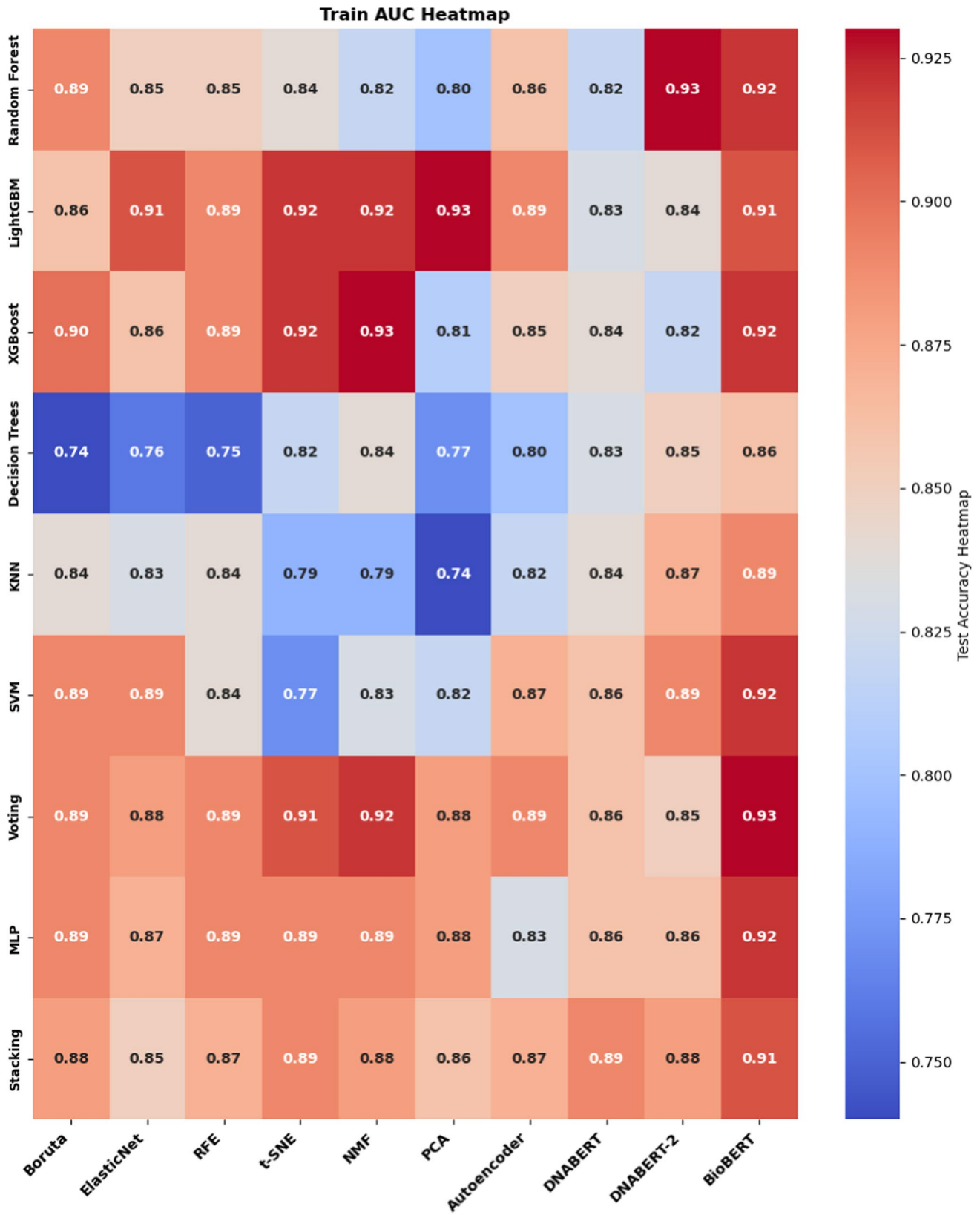

Fig. 5 Train AUC-ROC heatmap

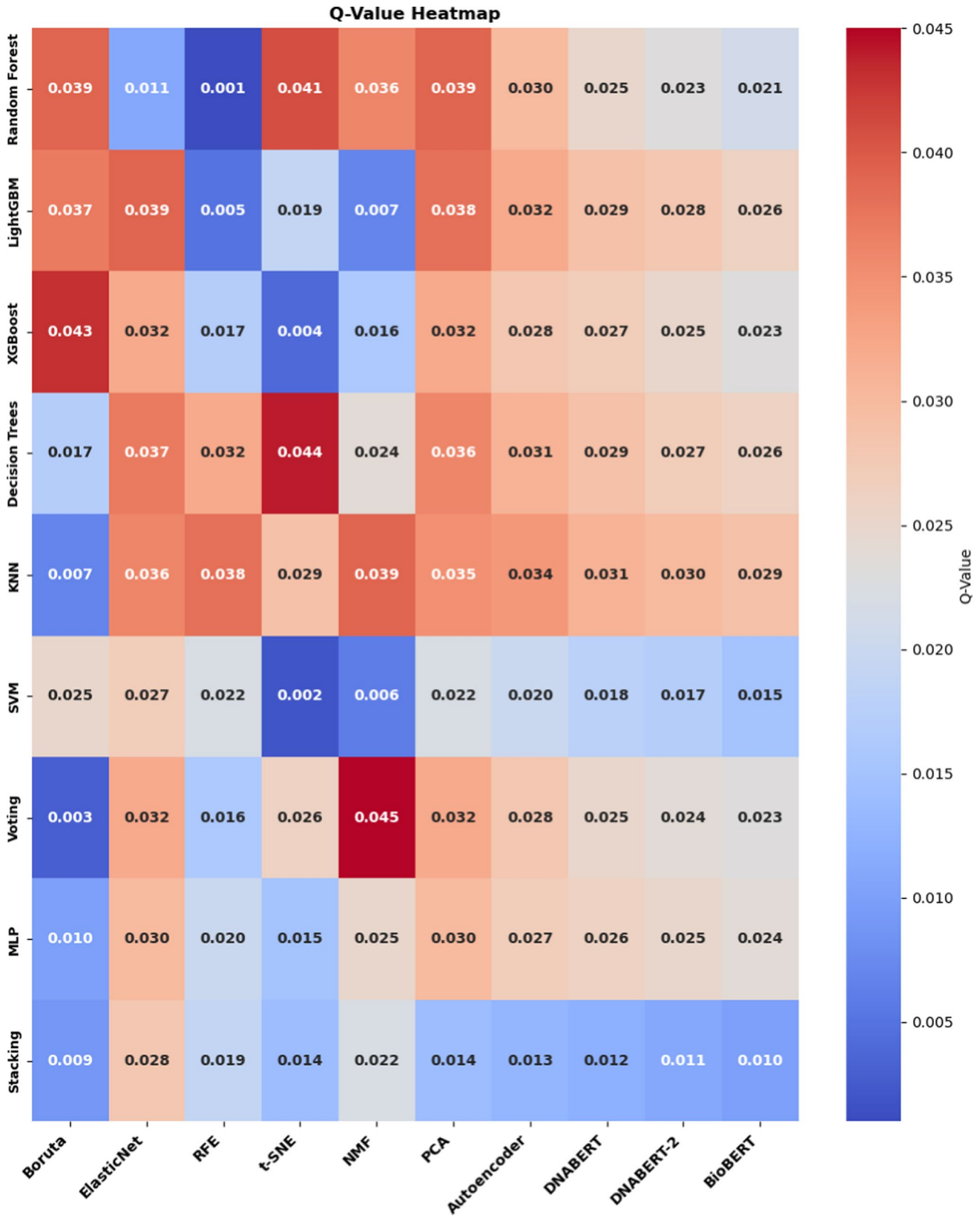

Fig. 6 The Q-value heatmap

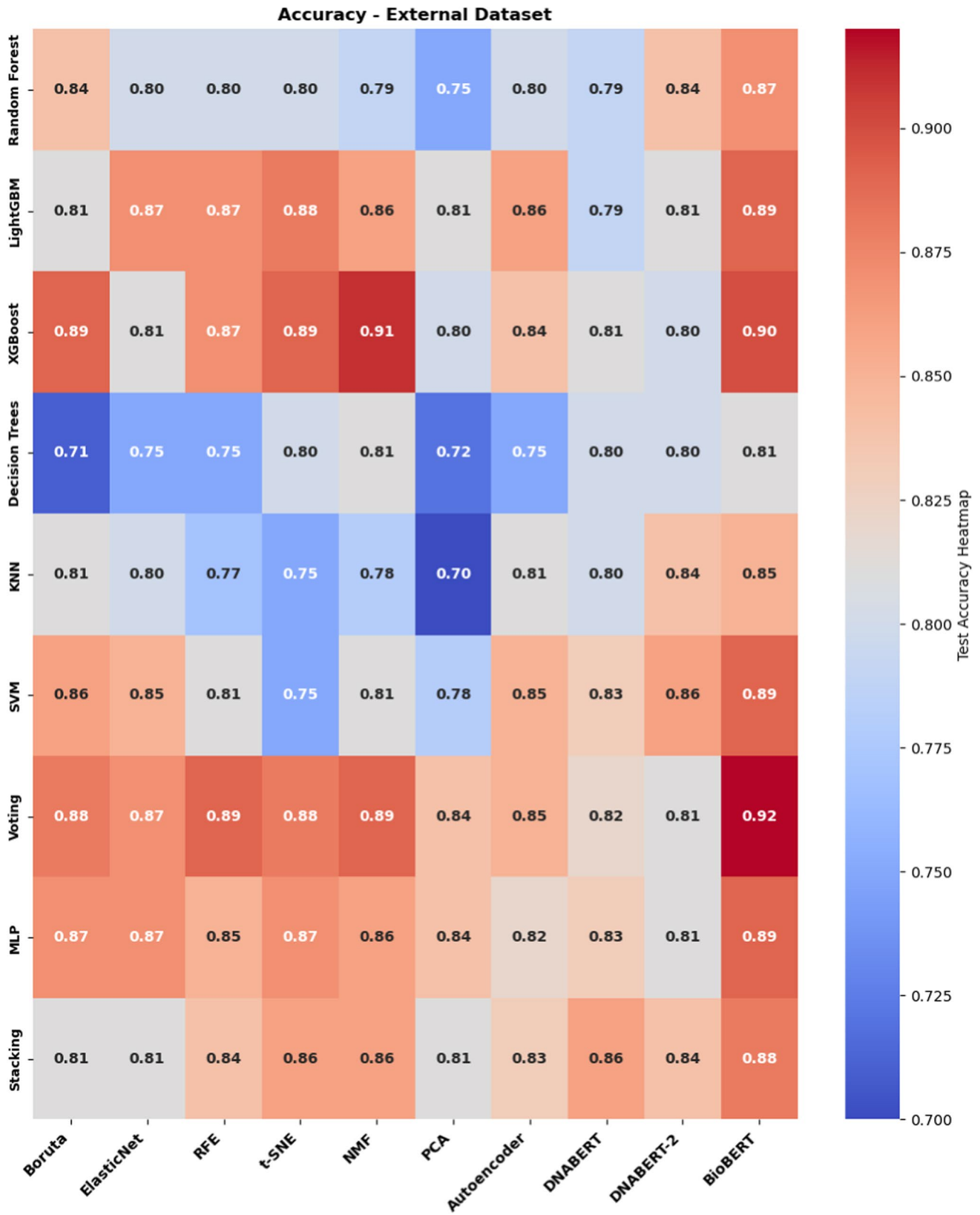

Fig. 7 Model accuracy across feature selection methods (external dataset)

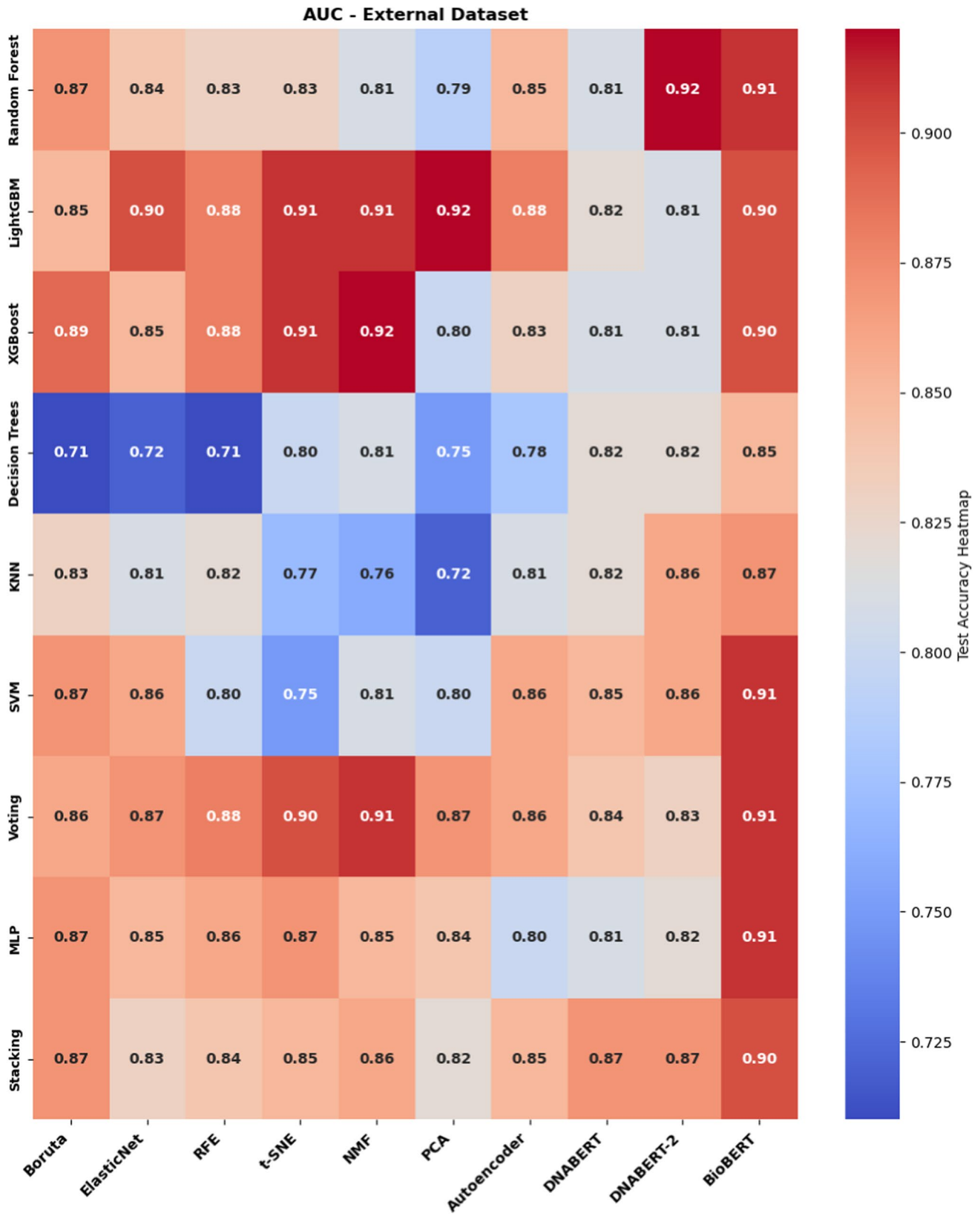

Fig. 8 AUC-ROC performance across feature selection methods (external dataset)

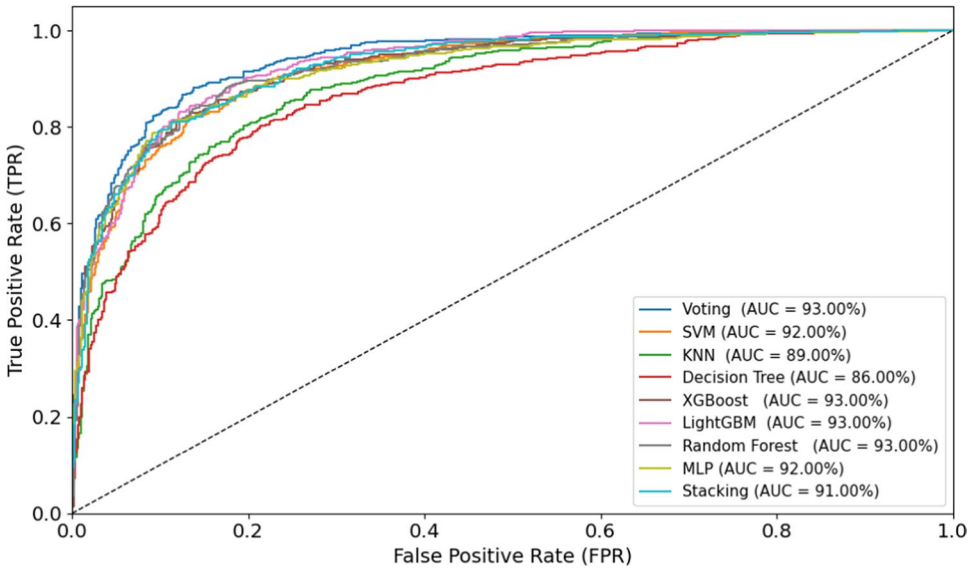

Fig. 9 ROC curves for training dataset across models

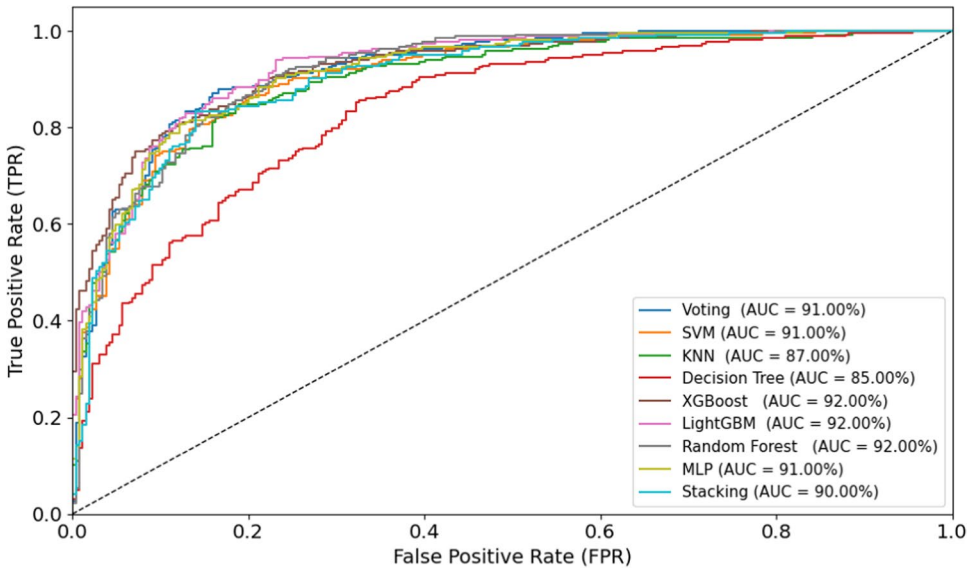

Fig. 10 ROC curves for testing dataset across models

dataset. This external validation demonstrates that the strongest models remain effective beyond the training data, ensuring robustness and generalizability (Fig. 11).

The Kappa Score heatmap assesses model consistency and agreement beyond chance. Similar to MCC, LightGBM, XGBoost, and ensemble models (Voting, Stacking) achieve the highest Kappa scores (0.94–) across PCA and NMF features, confirming their reliable classification power. Decision Trees and KNN display weaker performance (Kappa ~ 0.72 to 0.87), indicating potential inconsistencies in their predictions. The strong agreement of top models across the external test dataset highlights their robustness, suggesting they are well-suited for real-world applications beyond the training environment (Fig. 12).

The PR curves illustrate the trade-off between precision and recall across various classification thresholds. The area under the PR curve (AP score) serves as a key indicator of model reliability. XGBoost and LightGBM achieve the highest AP scores (~ 0.95), maintaining high precision even at higher recall values. Meanwhile, Decision Trees and KNN struggle with lower AP scores (~ 0.78 to 0.85), indicating weaker differentiation between positive and negative classes. The external dataset validation reaffirms that models with high AUC and MCC also maintain strong precision-recall characteristics, ensuring their robustness and practical applicability (Fig. 13).

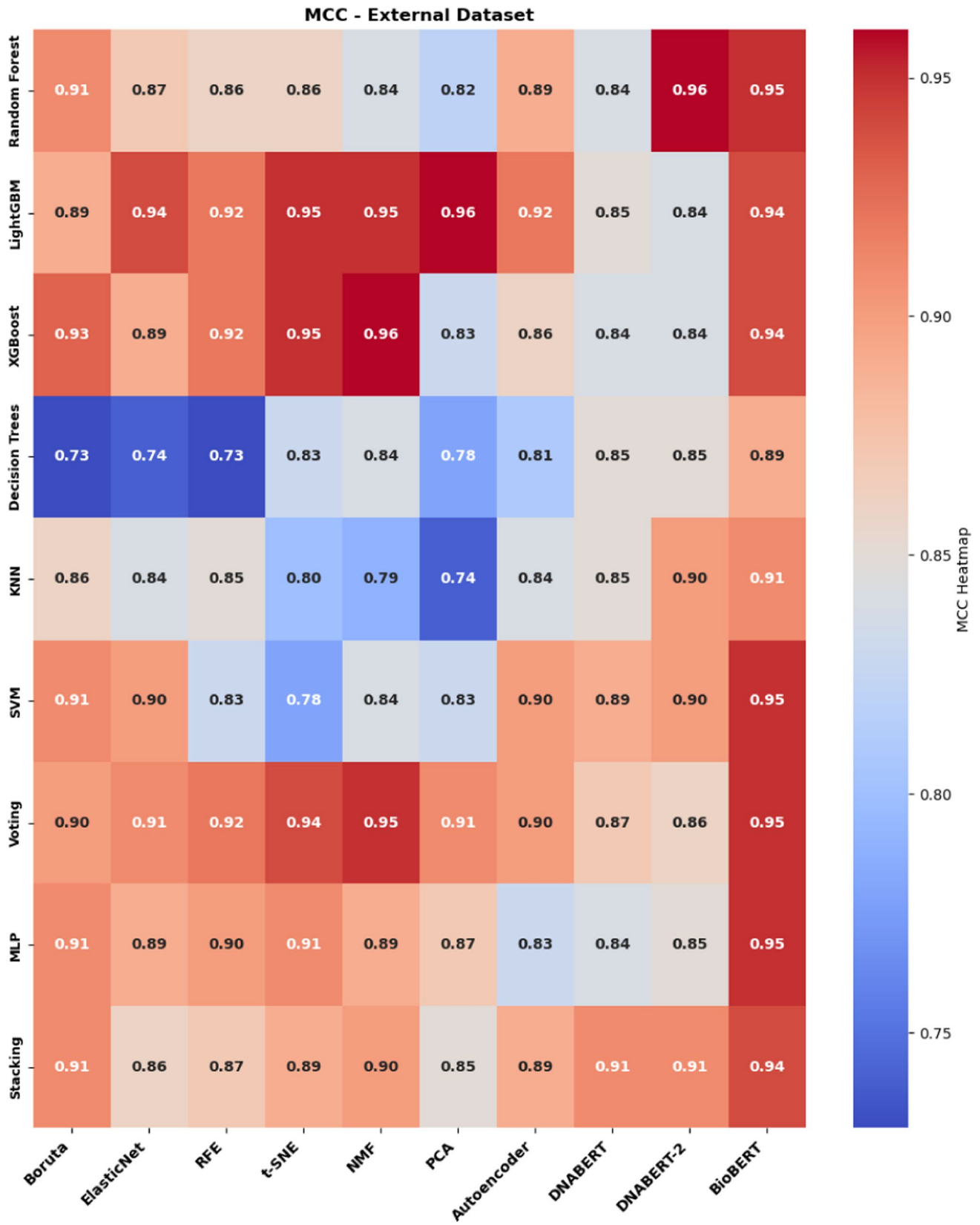

Fig. 11 MCC heatmap for external dataset

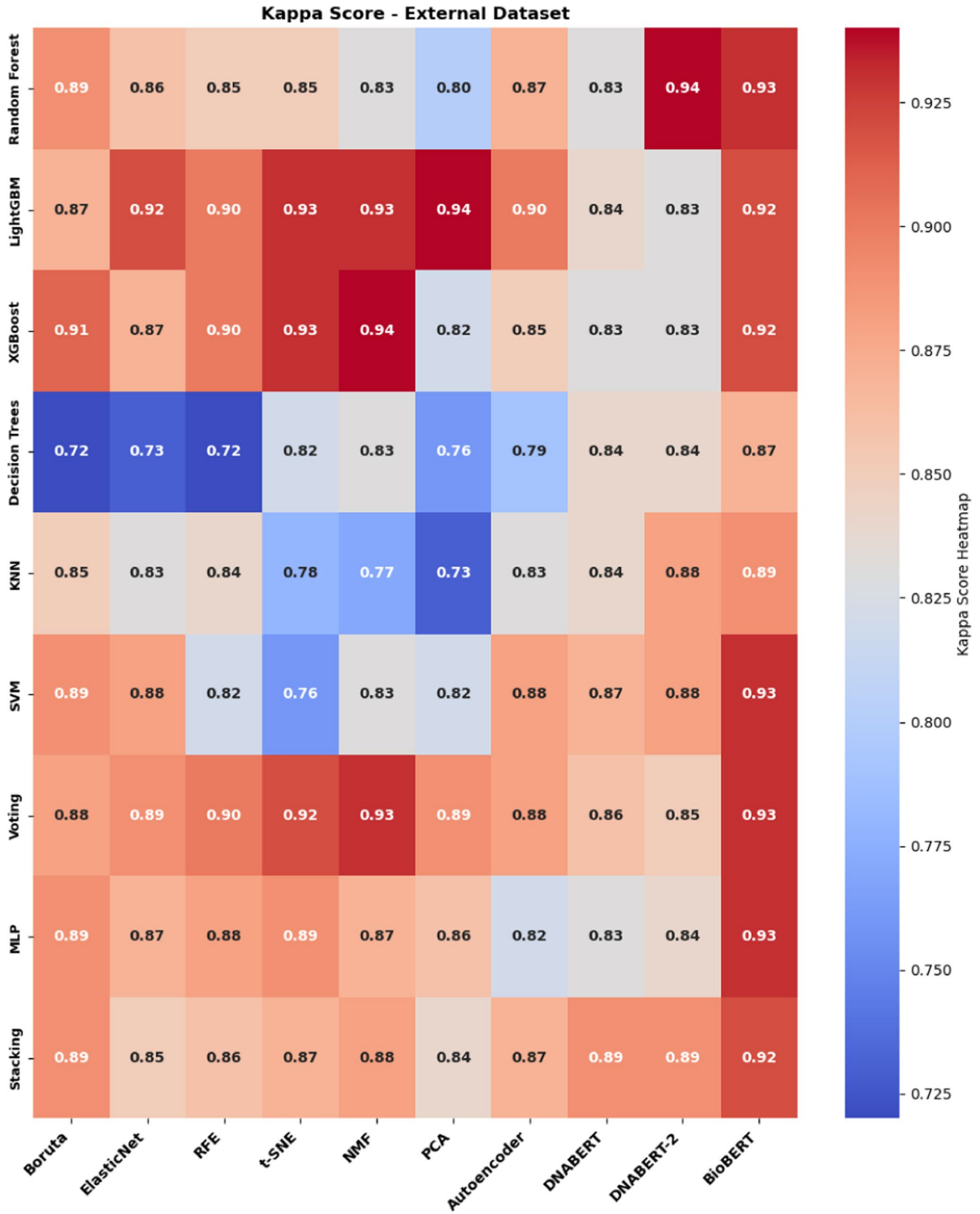

Fig. 12 Kappa score heatmap for external dataset

These results collectively confirm that ensemble methods and transformer-based feature extraction techniques (e.g., BioBERT, DNABERT-2) yield the most consistent, accurate, and robust classification results across multiple performance metrics when tested on unseen external data.

4 Discussion

This study presents an advanced machine learning framework for breast cancer diagnostics, leveraging state-of-the-art feature selection and transcriptomic profiling to enhance predictive performance and robustness. By integrating cutting-edge techniques such as RFE, Boruta, and ElasticNet regularization for feature selection, alongside dimensionality reduction methods including t-SNE, NMF, and transformer-based embeddings like BioBERT and DNABERT, the framework achieved high predictive accuracy and improved model generalizability. Classifiers such as XGBoost, LightGBM, Random Forest, and ensemble voting played a critical role, demonstrating exceptional efficiency in handling high-dimensional transcriptomic data. The primary innovation of this framework lies in its strategic integration of multiple advanced feature selection and dimensionality reduction methods with ensemble-based classifiers, surpassing traditional single-method approaches. This comprehensive strategy enables more effective navigation of complex transcriptomic landscapes, resulting in enhanced accuracy, interpretability, and reliability. This framework offers a powerful diagnostic tool for precision oncology, providing detailed insights into breast cancer pathology while laying the groundwork for more personalized and effective treatments.

In comparison to existing literature, our results outperform or complement previous studies in notable ways. For instance, Alghanim et al.[40] applied multi-omics strategies incorporating DNA methylation and copy number data, using PCA for dimensionality reduction. Although effective for broad omics integration, PCA's linear nature limits its capacity to highlight gene-level feature importance. By leveraging non-linear techniques like NMF, the XGBoost-NMF combination achieved a test accuracy of 0.92, outperforming Alghanim et al.'s results. Additionally, transformer-based models such as BioBERT further boosted performance, with the Voting-BioBERT combination achieving the highest external accuracy of 0.92, highlighting the strength of these advanced methods in breast cancer diagnostics.

Similarly, Mirza et al. [41] employed RFECV-SVM and logistic regression for feature selection, reporting significant diagnostic improvements with a nine-gene signature that achieved balanced accuracy and AUC values of 0.86 and 0.94, respectively. In comparison, our framework, which integrates RFE with advanced ensemble models such as the Voting classifier, not only matched their AUC of 0.94 but also improved test accuracy and generalizability. Furthermore, by incorporating

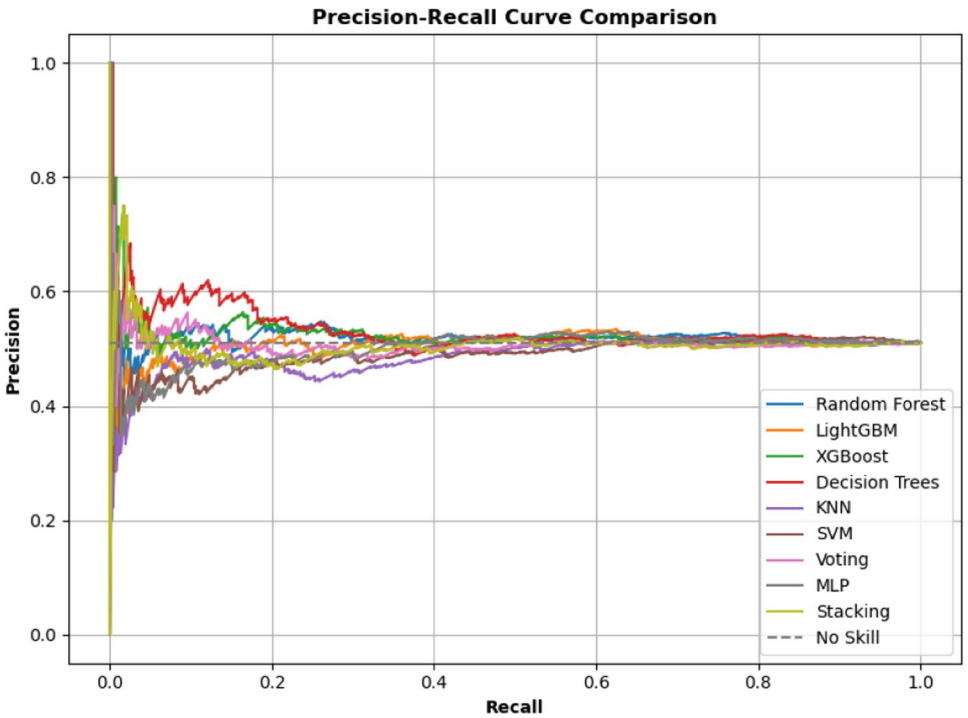

Fig. 13 Precision-recall curve for external dataset

transformer-based models like BioBERT, our approach surpassed the predictive capacity of the singular methods employed by Mirza et al. [41], demonstrating the value of combining diverse feature selection and ensemble strategies for enhanced diagnostic performance. Thalor et al. [74] focused on triple-negative breast cancer (TNBC), using RFE to identify biomarkers and XGBoost for classification with a small subset of genes. Our study expanded its application to a broader transcriptomic dataset and combined it with ElasticNet to enhance feature selection. This integrated approach preserved critical gene interactions, minimized redundancy, and incorporated advanced models such as BioBERT, leading to superior balanced accuracy and AUC metrics compared to those reported by Thalor et al. [57]. These advancements demonstrate the scalability and robustness of our framework for diverse breast cancer subtypes.

In the realm of survival prediction, Doan et al. [75] utilized ML to integrate clinical and transcriptomic data, achieving moderate performance (C-index of 0.688) with survival SVMs. Although our study did not focus on survival analysis, we employed similar classifiers and demonstrated superior predictive validity with the selected gene sets, suggesting that our framework could potentially be adapted for survival studies while maintaining high accuracy. Mehmood et al. [76] applied ElasticNet, LASSO, and Ridge regression for biomarker identification and drug sensitivity, emphasizing gene-drug interactions. However, their approach did not include multi-model integration. Our method, combining BioBERT with ensemble models, allowed for more comprehensive biomarker identification, suggesting practical applications in personalized treatment strategies that surpass Mehmood et al.'s outcomes. Taghizadeh et al. [73] reported a comprehensive ML pipeline involving automated hyperparameter tuning and multiple classifiers, achieving balanced accuracy and AUC values of 0.86 and 0.94 with LGR and MLP models. Our study, incorporating BioBERT and ensemble voting, not only reached similar or better AUC scores but also demonstrated higher balanced accuracy, showcasing the efficacy of combining feature selection and ensemble methods for improved robustness.

The selected gene subsets identified through Recursive Feature Elimination (RFE), Boruta, and ElasticNet were cross-referenced with existing literature and databases to explore their potential biological relevance. Key genes were found to be associated with breast cancer progression, molecular signaling pathways, and tumor behavior, aligning with findings from previous studies. For instance, several genes showed links to cell cycle regulation, apoptosis, and hormone receptor pathways, critical in breast cancer development. While these initial observations provide insights into their potential roles, further experimental validation and pathway enrichment analyses are necessary to confirm their biological significance. Future studies will integrate these findings with clinical annotations to enhance their applicability in personalized medicine.

Despite the strengths and advancements demonstrated in this study, several limitations exist. The reliance on publicly available datasets may limit the heterogeneity and diversity necessary for broader application. This could affect the generalizability when applying these models to more varied real-world populations. Additionally, future integration of clinical variables like age, hormonal status, and patient-specific risk factors could increase the models' clinical relevance and predictive precision. Further research should focus on validating these ML frameworks across multi-institutional cohorts to enhance external validity. Expanding to multi-omics approaches—incorporating proteomic and metabolomic data—would offer deeper insights and comprehensive biomarker panels for breast cancer. Evaluating the models in real-world clinical settings and using interpretable algorithms for clear decision-making explanations would improve their practical application and support wider adoption in personalized medicine. Incorporating patient-specific clinical variables, such as hormonal status, genetic predispositions, and other demographic factors, represents a critical avenue for enhancing the predictive precision and clinical applicability of machine learning frameworks. While this study focused on leveraging transcriptomic data due to dataset limitations, future research will prioritize integrating these clinical variables to provide a more comprehensive diagnostic approach. Current efforts are underway to collect and curate datasets with detailed clinical annotations, which will enable the exploration of combined molecular and clinical models. This integration has the potential to improve personalized diagnostic accuracy and support tailored treatment strategies, advancing the field of precision oncology.

5 Conclusion

This study presents an advanced machine learning framework that significantly improves breast cancer diagnostics by integrating sophisticated feature selection techniques with transcriptomic data analysis. By employing methods such as RFE, Boruta, and ElasticNet for precise feature selection and dimensionality reduction techniques like NMF, t-SNE, and transformer-based embeddings (e.g., BioBERT and DNABERT), the framework identified key gene subsets, enhancing both model interpretability and predictive performance. The use of ensemble models, including XGBoost, LightGBM, and Voting classifiers, demonstrated exceptional accuracy, achieving AUC values up to 0.92 while maintaining robust generalizability across external datasets.

These findings underscore the potential of combining advanced feature selection methods with ensemble-based and transformer-driven machine learning models to boost diagnostic accuracy, interpretability, and clinical relevance. This framework sets a new benchmark in computational oncology by paving the way for personalized treatment strategies and enabling more precise understanding of breast cancer pathology. Future work should explore validation across diverse clinical settings and the integration of multi-omics datasets to further enhance diagnostic capabilities and uncover deeper biological insights.

Acknowledgements Authors are grateful to the Researchers Supporting Project (ANUI2024M111), Alnoor University, Mosul, Iraq.

Author contributions MJS, HHA, RAK, and AY: gave the idea and drafted the manuscript. SG, ASh, GChSh, KSN, AR, HNS, AY, ZHA, and MA: performed the literature search and drafted figures. BF: edited the manuscript and supervised the whole study. All authors read and approved the manuscript.

Funding None.

Data availability The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests The authors declare no competing interests.

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Clinical trial number Not applicable.

References

1. Erber R, Hartmann A. Histology of luminal breast cancer. Breast Care. 2020;15:327–36.
2. McPherson K, Steel C, Dixon JM. Breast cancer—epidemiology, risk factors, and genetics. BMJ. 2000;321:624–8.
3. Sun Y-S, Zhao Z, Yang Z-N, Xu F, Lu H-J, Zhu Z-Y, et al. Risk factors and preventions of breast cancer. Int J Biol Sci. 2017;13:1387.
4. Wang L. Early diagnosis of breast cancer. Sensors. 2017;17:1572.
5. Moo T-A, Sanford R, Dang C, Morrow M. Overview of breast cancer therapy. PET Clin. 2018;13:339–54.
6. AkbarnezhadSany E, EntezariZarch H, AlipoorKermani M, Shahin B, Cheki M, Karami A, et al. YOLOv8 outperforms traditional CNN models in mammography classification: insights from a multi-institutional dataset. Int J Imaging Syst Technol. 2025;35: e70008.
7. Kimbung S, Loman N, Hedenfalk I. Clinical and molecular complexity of breast cancer metastases. Semin Cancer Biol. 2015;35:85–95.
8. Simpson PT, Reis-Filho JS, Gale T, Lakhani SR. Molecular evolution of breast cancer. J Pathol A J Pathol Soc Gt Britain Irel. 2005;205:248–54.
9. Lin S-X, Chen J, Mazumdar M, Poirier D, Wang C, Azzi A, et al. Molecular therapy of breast cancer: progress and future directions. Nat Rev Endocrinol. 2010;6:485–93.
10. Nafissi N, Heiranizadeh N, Shirinzadeh-Dastgiri A, Vakili-Ojarood M, Naseri A, Danaei M, et al. The application of artificial intelligence in breast cancer. EJMO. 2024;8:235–44.
11. Chugh G, Kumar S, Singh N. Survey on machine learning and deep learning applications in breast cancer diagnosis. Cognit Comput. 2021;13:1451–70.
12. Alshawwa IA, El-Mashharawi HQ, Salman FM, Al-Qumboz MNA, Abunasser BS, Abu-Naser SS. Advancements in early detection of breast cancer: innovations and future directions. 2024.
13. Sahu A, Qazi S, Raza K, Singh A, Verma S. Machine learning-based approach for early diagnosis of breast cancer using biomarkers and gene expression profiles. Comput. Intell. Oncol. Appl. Diagnosis, Progn. Ther. Cancers. Berlin: Springer; 2022. pp. 285–306.
14. Bostanci E, Kocak E, Unal M, Guzel MS, Acici K, Asuroglu T. Machine learning analysis of RNA-seq data for diagnostic and prognostic prediction of colon cancer. Sensors. 2023;23:3080.

15. Abidalkareem A, Ibrahim AK, Abd M, Rehman O, Zhuang H. Identification of gene expression in different stages of breast cancer with machine learning. Cancers (Basel). 2024;16:1864.
16. Cascianelli S, Molineris I, Isella C, Masseroli M, Medico E. Machine learning for RNA sequencing-based intrinsic subtyping of breast cancer. Sci Rep. 2020;10:14071.
17. Rezaeijo SM, Rezaei M, Pourshekhani A, Mohammadkhani S, Goharifar N, Shayankia G, et al. Integrative bioinformatics analysis of miRNA and mRNA expression profiles identified some potential biomarkers for breast cancer. Egypt J Med Hum Genet. 2023;24:62.
18. Bijari S, Sayfollahi S, Mardokh-Rouhani S, Bijari S, Moradian S, Zahiri Z, et al. Radiomics and deep features: robust classification of brain hemorrhages and reproducibility analysis using a 3D autoencoder neural network. Bioengineering. 2024;11:643.
19. Salmanpour MR, Hosseinzadeh M, Akbari A, Borazjani K, Mojallal K, Askari D, et al. Prediction of TNM stage in head and neck cancer using hybrid machine learning systems and radiomics features. Med Imaging Comput Diagn. 2022;12033:648–53.
20. Mahboubisarighieh A, Shahverdi H, Jafarpoor Nesheli S, Alipoor Kermani M, Niknam M, Torkashvand M, et al. Assessing the efficacy of 3D Dual-CycleGAN model for multi-contrast MRI synthesis. Egypt J Radiol Nucl Med. 2024;55:1–12.
21. Afreen S, Bhurjee AK, Aziz RM. Cancer classification using RNA sequencing gene expression data based on Game Shapley local search embedded binary social ski-driver optimization algorithms. Microchem J. 2024;205: 111280.
22. Yaqoob A, Verma NK, Aziz RM. Improving breast cancer classification with mRMR+ SS0+ WSVM: a hybrid approach. Multimed Tools Appl. 2024:1–26.
23. Yaqoob A, Verma NK, Aziz RM. Metaheuristic algorithms and their applications in different fields: a comprehensive review. Metaheurist Mach Learn Algorithms Appl. 2024:1–35.
24. Rahman RU, Singh K, Tomar DS, Musheer R. Building resilient digital forensic frameworks for nosql database: harnessing the blockchain and quantum technology. Sustain. Secur. Pract. Using Blockchain, Quantum Post-Quantum Technol. Real Time Appl. Berlin: Springer; 2024. pp. 205–38.
25. Aziz RM, Hussain A, Sharma P. Cognizable crime rate prediction and analysis under Indian penal code using deep learning with novel optimization approach. Multimed Tools Appl. 2024;83:22663–700.
26. Yaqoob A, Verma NK, Aziz RM, Shah MA. RNA-Seq analysis for breast cancer detection: a study on paired tissue samples using hybrid optimization and deep learning techniques. J Cancer Res Clin Oncol. 2024;150:455.
27. Joshi AA, Aziz RM. Soft computing techniques for cancer classification of gene expression microarray data: a three-phase hybrid approach. In: Comput. Intell. Data Anal., Bentham Science Publishers; 2024. pp. 92–113.
28. Yaqoob A, Verma NK, Aziz RM, Shah MA. Optimizing cancer classification: a hybrid RDO-XGBoost approach for feature selection and predictive insights. Cancer Immunol Immunother. 2024;73:261.
29. Lancashire LJ, Lemetre C, Ball GR. An introduction to artificial neural networks in bioinformatics—application to complex microarray and mass spectrometry datasets in cancer studies. Brief Bioinform. 2009;10:315–29.
30. Upadhyay T. Analysis and prediction of cancer using genome by applying data mining algorithms. Hyderabad: Shineeks Publishers; 2023.
31. Alexe G, Monaco J, Doyle S, Basavanhally A, Reddy A, Seiler M, et al. Towards improved cancer diagnosis and prognosis using analysis of gene expression data and computer aided imaging. Exp Biol Med. 2009;234:860–79.
32. Dlamini Z, Francies FZ, Hull R, Marima R. Artificial intelligence (AI) and big data in cancer and precision oncology. Comput Struct Biotechnol J. 2020;18:2300–11.
33. Dashtban M, Balafar M. Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts. Genomics. 2017;109:91–107.
34. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang C-H, Angelo M, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci. 2001;98:15149–54.
35. Benning L, Peintner A, Peintner L. Advances in and the applicability of machine learning-based screening and early detection approaches for cancer: a primer. Cancers (Basel). 2022;14:623.
36. Alharbi F, Vakanski A. Machine learning methods for cancer classification using gene expression data: a review. Bioengineering. 2023;10:173.
37. Begum S, Dey S, Chakraborty D, Hembrom T, Hazra S, Barman D. Utilizing machine learning techniques for categorizing cancer based on gene expression data: a review. J Electr Syst. 2024;20:1093–112.
38. Islam MT, Xing L. A data-driven dimensionality-reduction algorithm for the exploration of patterns in biomedical data. Nat Biomed Eng. 2021;5:624–35.
39. Ng S, Masarone S, Watson D, Barnes MR. The benefits and pitfalls of machine learning for biomarker discovery. Cell Tissue Res. 2023;394:17–31.
40. Alghanim F, Al-Hurani I, Qattous H, Al-Refai A, Batiha O, Alkhateeb A, et al. Machine learning model for multiomics biomarkers identification for menopause status in breast cancer. Algorithms. 2023;17:13.
41. Mirza Z, Ansari MS, Iqbal MS, Ahmad N, Alganmi N, Banjar H, et al. Identification of novel diagnostic and prognostic gene signature biomarkers for breast cancer using artificial intelligence and machine learning assisted transcriptomics analysis. Cancers (Basel). 2023;15:3237.
42. Pes B, Dessì N, Angioni M. Exploiting the ensemble paradigm for stable feature selection: a case study on high-dimensional genomic data. Inf Fusion. 2017;35:132–47.
43. Hua J, Tembe WD, Dougherty ER. Performance of feature-selection methods in the classification of high-dimension data. Pattern Recognit. 2009;42:409–24.
44. Leclercq M, Vittrant B, Martin-Magniette ML, Scott Boyer MP, Perin O, Bergeron A, et al. Large-scale automatic feature selection for biomarker discovery in high-dimensional OMICs data. Front Genet. 2019;10:452.
45. Chen Q, Zhang M, Xue B. Feature selection to improve generalization of genetic programming for high-dimensional symbolic regression. IEEE Trans Evol Comput. 2017;21:792–806.
46. Tadist K, Najah S, Nikolov NS, Mrabti F, Zahi A. Feature selection methods and genomic big data: a systematic review. J Big Data. 2019;6:1–24.
47. Gnana DAA, Balamurugan SAA, Leavline EJ. Literature review on feature selection methods for high-dimensional data. Int J Comput Appl. 2016;136:9–17.

48. Wang L, Wang Y, Li Y, Zhou L, Liu S, Cao Y, et al. A prospective diagnostic model for breast cancer utilizing machine learning to examine the molecular immune infiltrate in HSPB6. J Cancer Res Clin Oncol. 2024;150:1–16.
49. Chen X, Yi J, Xie L, Liu T, Liu B, Yan M. Integration of transcriptomics and machine learning for insights into breast cancer: exploring lipid metabolism and immune interactions. Front Immunol. 2024;15:1470167.
50. Vadapalli S, Abdelhalim H, Zeeshan S, Ahmed Z. Artificial intelligence and machine learning approaches using gene expression and variant data for personalized medicine. Brief Bioinform. 2022;23:bbac191.
51. Zhang S, Fan R, Liu Y, Chen S, Liu Q, Zeng W. Applications of transformer-based language models in bioinformatics: a survey. Bioinform Adv. 2023;3:vbad001.
52. Jiang J, Ke L, Chen L, Dou B, Zhu Y, Liu J, et al. Transformer technology in molecular science. Wiley Interdiscip Rev Comput Mol Sci. 2024;14: e1725.
53. Le NQK. Leveraging transformers-based language models in proteome bioinformatics. Proteomics. 2023;23:2300011.
54. Zhu J. Comparative study of sequence-to-sequence models: From RNNs to transformers. Appl Comput Eng. 2023;42:67.
55. Choi SR, Lee M. Transformer architecture and attention mechanisms in genome data analysis: a comprehensive review. Biology (Basel). 2023;12:1033.
56. Feng T, Huang Y, Li B. LLM-based DNA promoter prediction. CS582 ML Bioinforma. Work., n.d.
57. Li Q, Hu Z, Wang Y, Li L, Fan Y, King I, et al. Progress and opportunities of foundation models in bioinformatics. Brief Bioinform. 2024;25:bbae548.
58. Ruan W, Lyu Y, Zhang J, Cai J, Shu P, Ge Y, et al. Large language models for bioinformatics. ArXiv Prepr ArXiv:250106271 2025.
59. Sarumi OA, Heider D. Large language models and their applications in bioinformatics. Comput Struct Biotechnol J. 2024;23:3498.
60. Krishnaveni Reddy E, Mohammed TK. Pretrained language model for medical recommendation system (PLM2RS) using biomedical and electronic health record text summarization. In: Int. Conf. Intell. Syst. Sustain. Comput. Berlin: Springer; 2022. pp. 425–33.
61. Reddy EK, Mohammed TK. Pretrained language model for medical recommendation system (PLM 2 RS) using biomedical and electronic health record text summarization. Intell Syst Sustain Comput n.d.:425.
62. Tyagi N, Vahab N, Tyagi S. Genome language modeling (Glm): a beginner's cheat sheet 2024.
63. AlSaad R, Abd-Alrazaq A, Boughorbel S, Ahmed A, Renault M-A, Damseh R, et al. Multimodal large language models in health care: applications, challenges, and future outlook. J Med Internet Res. 2024;26: e59505.
64. Karim T, Shaon MSH, Sultan MF, Hasan MZ, Kafy A-A. ANNprob-ACPs: a novel anticancer peptide identifier based on probabilistic feature fusion approach. Comput Biol Med. 2024;169: 107915.
65. Shaon MSH, Karim T, Sultan MF, Ali MM, Ahmed K, Hasan MZ, et al. AMP-RNNpro: a two-stage approach for identification of antimicrobials using probabilistic features. Sci Rep. 2024;14:12892.
66. Karim T, Shaon MSH, Ali MM, Ahmed K, Bui FM, Chen L. StackAMP: stacking-based ensemble classifier for antimicrobial peptide identification. IEEE Trans Artif Intell. 2024;5:5666.
67. Shaon MSH, Karim T, Shakil MS, Hasan MZ. A comparative study of machine learning models with LASSO and SHAP feature selection for breast cancer prediction. Healthc Anal. 2024;6: 100353.
68. Omondiagbe DA, Veeramani S, Sidhu AS. Machine learning classification techniques for breast cancer diagnosis. In: IOP Conf. Ser. Mater. Sci. Eng., vol. 495. IOP Publishing; 2019. pp. 12033.
69. Rasool A, Bunterngchit C, Tiejian L, Islam MR, Qu Q, Jiang Q. Improved machine learning-based predictive models for breast cancer diagnosis. Int J Environ Res Public Health. 2022;19:3211.
70. Capobianco E. High-dimensional role of AI and machine learning in cancer research. Br J Cancer. 2022;126:523–32.
71. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486:346–52.
72. Tomczak K, Czerwińska P, Wiznerowicz M. Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol Onkol. 2015;2015:68–77.
73. Taghizadeh E, Heydarheydari S, Saberi A, JafarpoorNesheli S, Rezaeijo SM. Breast cancer prediction with transcriptome profiling using feature selection and machine learning methods. BMC Bioinform. 2022;23:410.
74. Thalor A, Joon HK, Singh G, Roy S, Gupta D. Machine learning assisted analysis of breast cancer gene expression profiles reveals novel potential prognostic biomarkers for triple-negative breast cancer. Comput Struct Biotechnol J. 2022;20:1618–31.
75. Doan LMT, Angione C, Occhipinti A. Machine learning methods for survival analysis with clinical and transcriptomics data of breast cancer. In: Comput. Biol. Mach. Learn. Metab. Eng. Synth. Biol. Berlin: Springer; 2022. pp. 325–93.
76. Mehmood A, Nawab S, Jin Y, Hassan H, Kaushik AC, Wei D-Q. Ranking breast cancer drugs and biomarkers identification using machine learning and pharmacogenomics. ACS Pharmacol Transl Sci. 2023;6:399–409.