Atlantic Journal of Life Sciences Volume 2026,
Article ID 65, pages 14
https://doi.org/10.71005/458kdy25

Original Article

Integrated Bioinformatics and Machine Learning Analysis Identifies MT1F as a Potential Diagnostic Biomarker and Therapeutic Target in Breast Cancer

Tahmid Islam Akib[1], Ibrahim Hossain Molla[2], Nafees Rishad[2], Labib Rahman[1], Nishad Al Hasan[1], Sidratul Muntaha Tasmi[1], Md Mahfujar Rahman Shakil[1], Ariful Islam[1]

[1]Department of Biomedical Engineering, Faculty of Engineering & Technology, Islamic University, Kushtia-7003, Bangladesh; akib2575@gmail.com, *labibrahman08@gmail.com, |[2]Department of Information Communication & Technology, Faculty of Engineering & Technology, Islamic University, Kushtia-7003, Bangladesh;*

Correspondences should be addressed to Dr. Ariful Islam; arifbme07@gmail.com

Received: 26 July 2025; Revised 1 August 2025; Accepted: 1 February 2026; Published 7 June 2026

KEYWORDS

Breast cancer,
MT1F,
Bioinformatics,
Machine learning,
Biomarker,
Drug sensitivity.

ABSTRACT

Breast cancer (BRCA) is a heterogeneous tumour and is the most common cancer in the world, and there is a need for strong biomarkers to enhance diagnosis and targeted therapy. We created a combined bioinformatics and machine learning approach for identifying important molecular biomarkers related to BRCA in this study. Two gene expression datasets were analyzed and 2,386 commonly dysregulated genes were identified. Enrichment analysis of functions indicated that extracellular matrix organization and immune related pathways were significantly involved. Protein–protein interaction (PPI) network analysis revealed that MT1F and CCNA2 were hub genes in the network, and MT1F was consistently ranked among the most central genes by various topological and machine learning techniques. The machine learning models (Random Forest, Gradient Boosting, and XGBoost) showed high diagnostic performance, with the Random Forest model having the highest discriminative ability (AUC = 0.990). It was found that MT1F was significantly upregulated in many different malignancies by pan-cancer analysis and was epigenetically regulated by DNA methylation analysis. Immune infiltration analysis also showed significant correlations between the expression of MT1F and immune cell populations. The drug sensitivity analysis (DSA) with GSCA datasets showed that MT1F expression was significantly correlated with sensitivity to several anti-cancer drugs, suggesting its role as a potential predictive biomarker. To conclude, MT1F might be a diagnostic biomarker and therapeutic target for breast cancer. This study highlights the value of integrative computational methods in the discovery of biomarkers and precision oncology.

1 Introduction

Cancer is a multifactorial disease in which the proliferation of cells becomes uncontrolled due to genomic and molecular changes that lead to abnormal cell signaling, cell cycle regulation, and apoptosis. Breast cancer is the most common cancer diagnosis globally, with around 2.3 million new cases and 685,000 deaths reported in 2020, representing 11.7% of all cancer cases (Mohanti et al., 2025). It is the most common cancer in women in Bangladesh, accounting for 19% of cancers and 6.2% of cancer deaths (Organization, 2020). Although detection and treatment have improved, there are still survival differences, especially in low and middle-income countries, and there is a need to develop better diagnostic and therapeutic approaches (Khan et al., 2020).

Breast cancer is a highly heterogeneous disease with several molecular subtypes, including Luminal A, Luminal B, HER2-positive breast cancer, and triple-negative breast cancer (TNBC) with distinct gene expression profiles and clinical outcomes (Network, 2012; Sotiriou & Pusztai, 2009). Existing diagnostic tools include imaging and histopathological examination, which are useful but open to variability and interpretation issues (Makary & Daniel, 2016; Ramaswamy et al., 2001). The use of high-throughput molecular profiling technologies, such as microarray-based gene expression analysis, has led to the identification of genes that are differentially expressed (DEGs) and potential biomarkers to a higher degree of accuracy (Karim et al., 2023; Schena et al., 1995) However, some of the limitations, including small sample sizes, heterogeneous datasets, and methodological inconsistencies, restrict the reproduction and clinical application of the results (Perou et al., 2000; Slodkowska & Ross, 2009).

Since the initial research that showed that gene expression signatures could be used to distinguish cancer subtypes (Golub et al., 1999). molecular profiling has made a significant leap forward as a diagnostic tool for cancer, and Metallothionein 1F (MT1F) has emerged as a promising biomarker for this purpose because of its association with cell cycle and tumor progression. MT1F is consistently overexpressed across multiple cancers, including breast cancer, and is associated with proliferation, metastasis, and poor prognosis (Kanakkanthara et al., 2016; Qian et al., 2021; Shi et al., 2019; Yam et al., 2002). Its expression is further influenced by epigenetic mechanisms such as

DNA methylation (Abd-Elnaby et al., 2021; Hossen et al., 2024). Network-based analyses have identified MT1F as a central hub gene in protein–protein interaction (PPI) networks, supported by multiple computational and machine learning approaches (Cao et al., 2024).

Recent advances in bioinformatics and machine learning have facilitated the integration of large-scale genomic data for biomarker discovery. Techniques such as LASSO regression, Random Forest, and support vector machine recursive feature elimination (SVM-RFE) enable robust feature selection and predictive modeling, improving diagnostic accuracy (Khan et al., 2020). Nevertheless, the integration of multi-algorithm machine learning with network-based approaches for identifying reliable biomarkers in breast cancer remains limited.

In this study, we developed an integrated bioinformatics and machine learning framework to identify key biomarkers in breast cancer. We used publicly available gene expression datasets to identify the differentially expressed genes, build protein–protein interaction (PPI) networks and used several machine learning algorithms for feature selection and classification. The diagnostic performance and therapeutic relevance of the identified biomarkers were further evaluated, highlighting MT1F as a potential diagnostic biomarker and therapeutic target. This study is unique in its multi-layered integrative approach that integrates complementary feature selection techniques with network analysis and assesses the biomarker in a variety of biological dimensions, such as pan-cancer expression, DNA methylation, immune cell infiltration, and drug sensitivity. The use of an independent external dataset further enhances the robustness and replicability of the results.

2. Methods and Materials

Overall methodologies of this study have been schematized in Figure 1

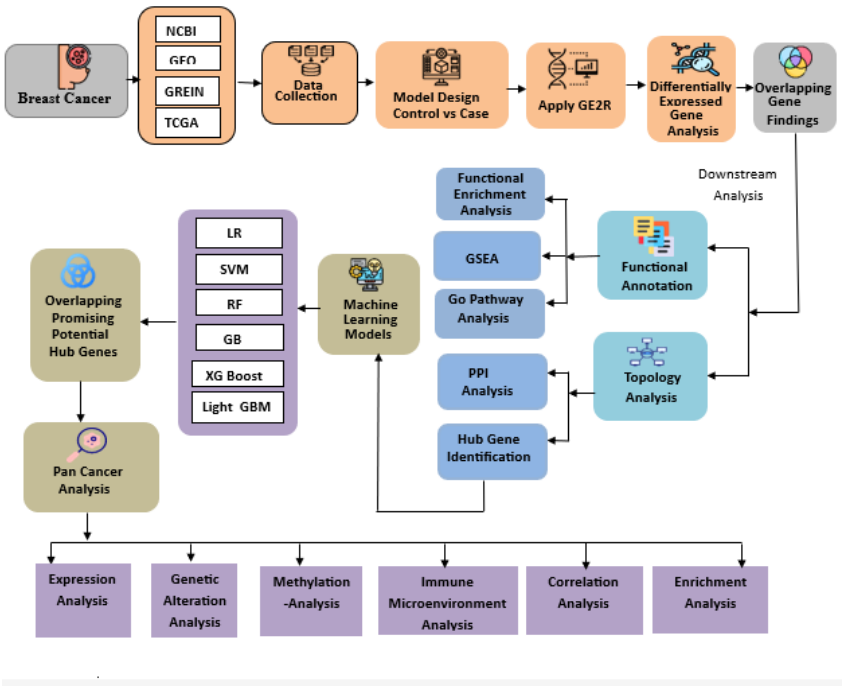

Figure 1: Schematic overview of the integrated workflow employed in this study, illustrating the bioinformatics and machine learning pipeline for biomarker identification, along with downstream network-based systems analysis.

2.1. Data collection and preprocessing

The NCBI Gene Expression Omnibus (GEO) and the GREIN (Genes Relevant for Information) repositories were used to get the gene expression data (Yu-Jing et al., 2020). Datasets were selected based on specific criteria (availability of both breast cancer and control samples, sample number, and uniformity in the platform). Two datasets (GSE106694 and GSE108693) were kept for further analysis after the screening. Standard bioinformatics workflows were used to process the raw data: Background correction, Quantile normalization, and Log2 transformation to achieve normalization of samples. For batch effects, correction was done with ComBat from the sva package when it was applicable, and for HSOV, the estimation was achieved using surrogate variable analysis (SVA) (Leek et al., 2012). Lowly expressed (or lowly-varied) genes were removed to reduce the noise. The Probe IDs were mapped to gene symbols using the platform annotation files, and duplicate probes were merged by calculating the median expression. An external data set (GSE123631) was additionally used to validate the primary dataset. This dataset was chosen to assess the generalizability and reproducibility of the identified biomarkers.

2.2. Identification of differentially expressed genes (DEGs)

To identify DEGs, the criteria used were: The criteria used for identification of DEGs were: The *Limma* R package was used for

differential expression analysis. Genes were then determined statistically significant with an absolute log2 fold change (|log2FC|) ≥ analysis was used to find common DEGs across the datasets and presented in a Venn diagram.

2.3. Functional enrichment and pathway analysis

To understand the biological functions and molecular mechanisms associated with the differentially expressed genes (DEGs), functional enrichment analysis was performed. ClusterProfiler R package was used to perform Gene Ontology (GO) enrichment analysis for biological processes (BP) and cellular components (CC), with values P<0.05 corrected were used as significant **(**Hossain et al., 2025**).** Pathway level changes were then evaluated using EnrichR for various curated pathway sets such as KEGG, MSigDB Hallmark **(**Ali et al., 2025**)**, Reactome **(**Gillespie et al., 2022**)**, and WikiPathways **(**Slenter et al., 2018**).** Pathways with p-values < 0.05 were deemed significant **(**Kuleshov et al., 2016**).** This integrative approach enabled the identification of main pathways and molecular functions involved in disease progression.

2.4. Protein–protein interaction (PPI) network and module analysis

To analyze the protein–protein interactions of differentially expressed genes (DEGs), protein–protein interaction (PPI) networks were created. Interaction data were obtained from the STRING database and further analyzed using NetworkAnalyst **(**Szklarczyk et al., 2019**).** To guarantee reliable interactions, a threshold of ≥ 0.4 was used for the confidence score. The topology of the network was evaluated and highly interconnected nodes were detected by applying a degree cut-off > 15 to find key interaction hubs. These PPI networks were visualized with Cytoscape software **(**Shannon et al., 2003**)**, enabling systematic exploration of molecular interactions and identification of functionally relevant subnetworks.

2.5. Hub gene identification

The topographical properties of the PPI network were used to identify hub genes. Gene importance was assessed by degree centrality (DC), and genes with DC ≥ 2 × median DC were selected as candidate hub genes (Barabasi & Oltvai, 2004**).** Further prioritization was performed using the CytoHubba plugin in Cytoscape (Chin et al., 2014**)**, using several algorithms such as Degree, Maximal Clique Centrality (MCC), BottleNeck, DMNC and Edge Percolated Component (EPC) to ensure that the selection is robust. In addition, modules with high interaction levels were identified in the PPI network using the Molecular Complex Detection (MCODE) algorithm with the following parameters: node score cutoff = 0.2, K-core = 2, and degree cutoff = 2 (Bader & Hogue, 2003; Morris et al., 2011**).** An integrative topological analysis allowed the identification of hub genes and functional clusters that related to disease mechanisms.

2.6. Machine learning–driven feature selection and model development

The data set was first split into training/80% and testing/20% sets to guarantee strong and unbiased model development, splitting the data using stratified sampling to keep the class distribution. To avoid data leakage, all preprocessing steps (such as normalization, SMOTE-based class balancing, and feature selection) were applied only to the training data set **(**Chawla et al., 2002**)**. Three complementary methods, namely LASSO regression, SVM-RFE, and the Random Forest, were used to perform feature selection on the training data. LASSO with 1 and an adjusted p-value ≤ 0.05 **(**Liu et al., 2021**).** Genes that were upregulated and downregulated were identified. Intersection
ten-fold cross-validation is employed to shrink less informative coefficients, and SVM-RFE is used to remove the least informative features iteratively. Random Forest (500 trees) was used to estimate feature importance **(**Tibshirani, 1996**).** The biomarker panel was built by combining the features that were detected by all three classifiers. The selected features were used to train machine learning models such as Logistic Regression, Support Vector Machine, k-Nearest Neighbors, Random Forest, Gradient Boosting, XGBoost, and LightGBM (Guyon et al., 2002**).** Hyperparameter tuning was done using grid search on the training data set. The performance of the models was tested with a five-fold cross-validation on the training set, and then tested on the independent test set with accuracy, precision, recall, F1-score, and ROC–AUC **(**Breiman, 2001**).**

2.6.1. Evaluation metrics

Model performance was assessed based on the commonly used classification performance measures such as accuracy, precision, recall, F1-score and receiver operating characteristic (ROC) area under the curve (AUC). The statistics were computed with the help of true Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN).

The overall percentage of correctly classified instances is called accuracy:

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

Precision is the ratio of true positive predictions to all positive predictions:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall (sensitivity) describes how well the model has captured the positive examples:

$$\text{Recall} = \frac{TP}{TP + FN}$$

In case of class imbalance, the F1-score is a balanced measure, harmonic mean of precision and recall:

$$\text{F1-score} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

Further, the model's discriminative power for each classification threshold was assessed using ROC-AUC.Specificity, which represents the true negative rate, was calculated as:

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

Together, these metrics give an overall evaluation of the performance of classifiers, especially when the data is potentially imbalanced.

2.7. Pan-cancer gene expression analysis

To investigate the expression profile of MT1F in different cancers, a pan-cancer analysis of expression was undertaken. The gene expression analysis between tumor and adjacent normal tissues with TCGA datasets was performed by comparing the gene expression level by using the Gene_DE module of the TIMER2 web server (Li et al., 2020**).** Furthermore, GEPIA2 (based on TCGA and GTEx) was used to create box and violin plots to evaluate MT1F expression in various cancer types and stages **(**Consortium, 2020; Tang et al., 2019**).**

2.8. DNA methylation analysis

MT1F methylation analysis was performed via the UALCAN portal that offers TCGA level 3 RNA-seq and clinical data **(**Chandrashekar et al., 2017**).** To assess epigenetic regulation and its possible influence on gene expression and cancer development, methylation values in tumor samples were compared to those in normal tissue samples.

2.9. Immune infiltration analysis

To explore the role of the tumor microenvironment in cancer progression, the association between gene expression and immune cell infiltration was assessed. Immune infiltration analysis was conducted using the TIMER database, which allows systematic evaluation of gene expression and the presence of immune cells in various cancer types **(**Hossain et al., 2024**).** Correlation analyses were carried out to investigate the connections between MT1F expression and the infiltration of important immune cell subsets.

2.10. Drug sensitivity analysis

Using the GSCA platform, the association between MT1F expression and response to anticancer agents was performed and named "drug sensitivity analysis. Using the CTRP and GDSC pharmacogenomic datasets, which combine gene expression profiles with compound sensitivity data in cancer cell lines, correlation analyses were performed. Spearman's correlation coefficients were computed to assess gene–drug associations, and statistical significance was evaluated by FDR correction (FDR < 0.05). Positive correlations suggested increased sensitivity with higher MT1F expression, while negative correlations indicated less sensitivity or potential resistance.

2.11. Statistical analysis

All statistical analyses were performed using R software (version 4.3.2). Key packages included *limma*, *clusterProfiler*, *sva*, and *glmnet*.

3. Result

3.1. Gene Expression Analysis of Transcriptomic Data

36,863 genes were significantly differentially expressed (DEGs), with 1,295 genes being upregulated, and 1,091 genes being downregulated. The intersection of genes indicated robust and consistent transcriptional alterations associated with BRCA through cross-dataset intersection analysis (2,386 commonly dysregulated genes; Figure 2A). Of these, 69 genes were found to be consistently over-expressed, and 47 genes were consistently under-expressed in all datasets (Figure 2B–C). Notably, upregulated genes such as *SLC29A4, SIM2, ELFN1, LOXL2, FSTL3, KIF1A, TIMP, CELF5, KRTAP5-AS1, FAM131B, RASD1, ARHGDIG, ALCAM, CERS1, KLRG2, TRIM47, MICAL2, SYNGR3, XAF1, RTAP5-2, TENM1, LOC100507156, SHC2, NBL1, SYNPO, NUPR1, ZNF469, SCGB1A1, HTRA1, KRTAP5-1, CECR2, EDIL3, RBPMS2, LOC107984862, CMPK2, SLC17A7, F2R, MUC5AC, NOVA2, ZNF532,* and *SERPINE1* are implicated in extracellular matrix remodeling, tumor invasion, and metastatic progression. In contrast, downregulated genes, including *ACP3, BTN3A2, BTN3A1, TRIM5, LINC02518, KCNJ13, FRRS1, PPM1K, LOC105375784, ANKRD22, MGAT3, PALM3, THEM4, PKIB, PLA2G4F, MCEE, MYZAP, SLC4A10, LOC105377123, CLEC7A, GRAMD1C, BLNK, FAM25A, BNIPL, CRYBG2, SLC7A8, CALCR, FREM1, ABCA12,* and *DEPDC1,* these were mainly involved in immune regulation and metabolism. These together indicate a conserved set of molecular signatures in the BRCA, which will serve as a solid basis for downstream functional and network analyses.

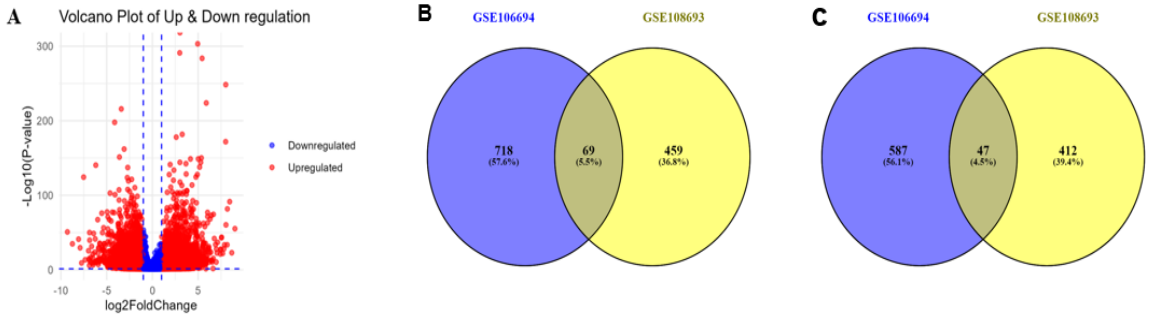

Figure 2: Differential gene expression analysis of BRCA datasets. (A) Identification of significantly dysregulated genes. (B) Venn diagram showing common upregulated genes. (C) Venn diagram showing common downregulated genes.

3.2. Functional enrichment and pathway analysis

Functional enrichment analysis was performed on 115 shared differentially expressed genes (DEGs), including 68 upregulated and 47 downregulated genes (Supplementary File 2). Gene Ontology (GO) analysis revealed significant enrichment in biological processes related to extracellular matrix organization and signaling regulation. Notably, terms such as *collagen-containing extracellular matrix* and *negative regulation of BMP signaling pathway* were significantly enriched (adjusted p-value < 0.05), highlighting their roles in tumor microenvironment remodeling and cancer progression (Figure 3A). Network visualization further identified a SERPINE1-centered module involving genes such as *COL17A1, EDIL3, TIMP3, LOXL2,* and *FSTL3*, which are associated with extracellular matrix remodeling and invasion-related processes (Figure 3B). Downregulated gene enrichment analysis revealed pathways associated with membrane transport and immune regulation, including *organic cation transport*, *amino acid transport*, and *positive regulation of type II interferon*

production, suggesting alterations in metabolic and immune functions (Figure 3 C–D).

Pathway enrichment analysis using EnrichR identified 55 significantly enriched pathways (p-value < 0.05) across multiple databases. The top-ranked pathways included *epithelial–mesenchymal transition*, *collagen formation*, *assembly of collagen fibrils*, *dissolution of fibrin clot*, and *butyrophilin family interactions* (Figure 3E). These findings collectively emphasize the critical roles of extracellular matrix dynamics and immune-related processes in BRCA pathogenesis.

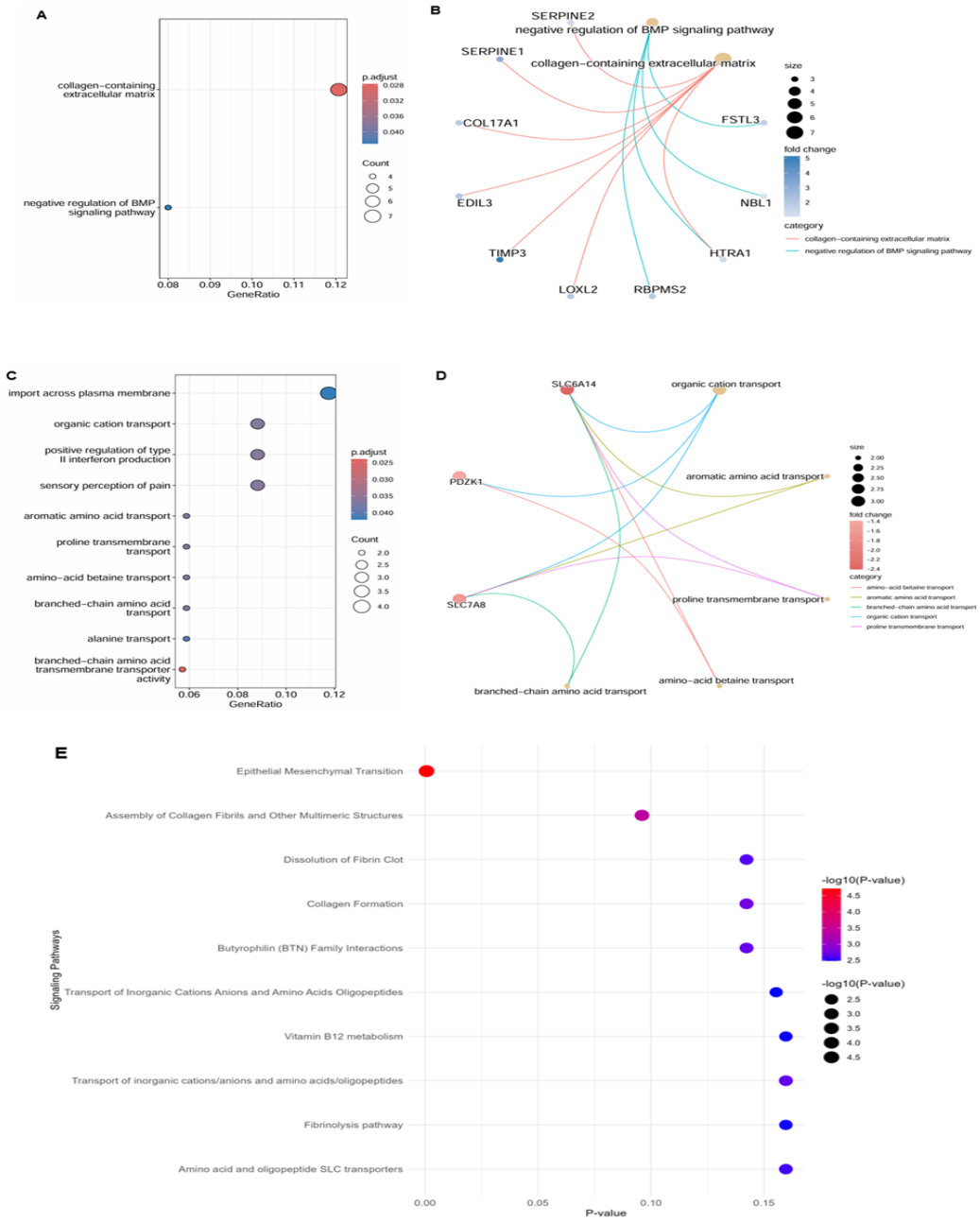

Figure 3: Functional enrichment analysis of shared DEGs. (A) GO enrichment of upregulated genes. (B) Network visualization of SERPINE1-centered interactions. (C–D) GO enrichment and network analysis of downregulated genes. (E) Top 10 enriched pathways identified through EnrichR analysis.

3.3. Protein–protein interaction network and hub gene identification

This section introduces protein–protein interaction networks and the concept of hub genes. This section introduces the idea of protein–protein interactions network and hub genes. As shown in Figure 4, the protein–protein interaction (PPI) network built from the overlapping DEGs contained 136 nodes and 605 edges, and the average node degree and clustering coefficient of this network were 8.9 and 0.624,

respectively. The network showed a very high level of enrichment ($p < 1.0 \times 10^{-16}$), suggesting a high level of functional connectivity among the identified genes. The CytoHubba plugin was used to identify hubs in the network using several topological algorithms: Degree, Maximal Clique Centrality (MCC), Maximum Neighborhood Component (MNC) and Edge Percolated Component (EPC) (Figure 4B–E). In principle, at least three methods identified 16 hub genes, such as PRC1, TPX2, KIF4A, KIF2C, CENPE, CDC20, NUSAP1, NCAPG, CEP55, CCNA2, MT1F. CCNA2 and MT1F were repeatedly identified as the most central hub genes by all of the algorithms, indicating their key importance in the regulatory network. These genes are likely to play a key role in the progression of BRCA and could serve as biomarkers and therapeutic targets.

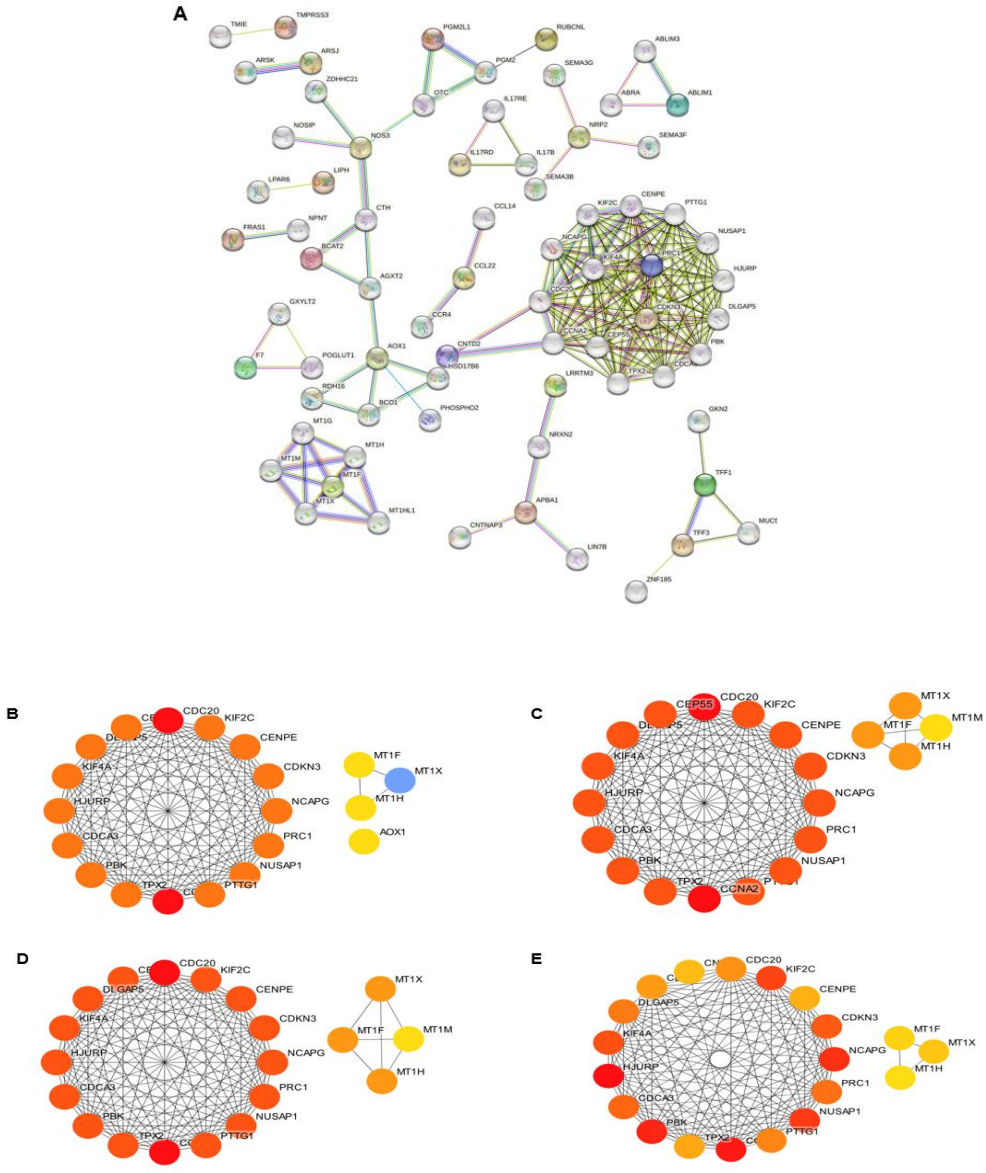

Figure 4: PPI network constructed from shared differentially expressed genes (DEGs), illustrating the interactions among encoded proteins. Nodes represent proteins, and edges indicate predicted or experimentally validated interactions. (B–E) Identification of top hub genes using CytoHubba algorithms, including Degree, Maximal Clique Centrality (MCC), Maximum Neighborhood Component (MNC), and Edge Percolated Component (EPC). Highly ranked nodes represent key regulatory genes within the network.

3.4. Identification and validation of hub genes

Using PPI network analysis along with CytoHubba, a number of highly interconnected genes were found, including CCNA2 and MT1F, which was consistently among the top genes, regardless of the topological algorithm used. The candidate biomarkers were further narrowed down using machine-learning–based feature selection. LASSO and SVM-RFE consistently found the key features including CCNA2, MT1F and EGFR and the Random Forest analysis (500 trees) ranked MT1F as one of the most important genes based on feature

importance scores. The overlap of these methods identified MT1F as the most robust and consistently selected biomarker.

Expression validation across the training and independent validation datasets demonstrated significant upregulation of MT1F in BRCA samples, supporting its potential as a diagnostic biomarker and a key regulatory gene in disease progression.

3.5. Diagnostic model performance and feature importance

To evaluate the diagnostic potential of the identified biomarkers, multiple machine learning models were developed and compared. Among the evaluated classifiers, KNN, Gradient Boosting, and

XGBoost achieved the highest accuracy (93.47%), with corresponding F1-scores of 0.930, 0.937, and 0.937, respectively (Table 1). Random Forest demonstrated the highest discriminative performance, achieving an AUC of 0.990, followed closely by Gradient Boosting (AUC = 0.989). Such ensemble models also showed good recall and precision, demonstrating that they performed well in classification tasks. In contrast, SVM and Logistic Regression showed comparatively lower performance, with accuracies of 78.26% and 76.08%, respectively (Figure 5A–B). Feature importance analysis consistently identified *MT1F* as the most influential gene across multiple models, highlighting its central role in classification performance. To validate the robustness of the identified biomarker, the independent dataset GSE123631 was analyzed. Consistent with the discovery datasets, MT1F expression was significantly upregulated in breast cancer samples compared to normal tissues. These results confirm the reproducibility and diagnostic potential of MT1F across independent cohorts. Collectively, these results demonstrate the superior predictive capability of ensemble learning methods and reinforce the potential of *MT1F* as a key diagnostic biomarker in BRCA.

3.6. Pan-cancer analysis of MT1F expression

A pan-cancer analysis was performed to evaluate MT1F expression across multiple tumor types using the TIMER database. Significant differential expression between tumor and adjacent normal tissues was observed in a wide range of cancers (Figure 6A). Notably, *MT1F* expression was significantly upregulated ($p < 0.001$) in BLCA, BRCA, CHOL, COAD, ESCA, HNSC, KICH, KIRP, KIRC, LIHC, LUAD, LUSC,PCPG,PRAD,

The moderate upregulation ($p<0.01$) of SKCM, STAD, and UCEC is not considered to be significant. was seen in CESC, READ and THCA. Overall, MT1F was highly expressed in 22 cancer types. It has the potential to become a pan-cancer biomarker. Median expression. These were corroborated by the levels across tumor and normal tissues. As shown in Figure 6B, they discovered that the calcium in this formula is absorbed by the skin as well. To confirm these observations by, Integrated TCGA and GTEx was used for GEPIA2 analysis datasets. In line with the TIMER results, GEPIA2 confirmed. Increased expression of MT1F in various types of cancer, including: It was suggested that it has an additional role in tumorigenesis

Table 1: Performance comparison of machine learning models for BRCA classification

Model	Accuracy	Precision	Recall	F1-score	ROC–AUC
Logistic Regression	0.7609	0.7394	0.7754	0.7456	0.9299
SVM	0.7826	0.7724	0.8154	0.7762	0.9738
KNN	0.9348	0.9253	0.9365	0.9301	0.9792
Random Forest	0.9130	0.8978	0.9247	0.9087	0.9903
Gradient Boosting	0.9348	0.9475	0.9365	0.9374	0.9894
XGBoost	0.9348	0.9475	0.9365	0.9374	0.9872
LightGBM	0.9348	0.9475	0.9365	0.9374	0.9883

3.7. DNA methylation analysis of MT1F

We used TCGA data to examine the epigenetic regulation of MT1F in several cancer types via DNA methylation analysis. Promoter methylation was significantly different between tumor and normal tissues in 14 cancer types (Figure S1). Notably, statistically significant alterations were identified in BRCA ($p = 2.27 \times 10^{-5}$), BLCA ($p = 5.80 \times 10^{-4}$), CESC ($p = 4.20 \times 10^{-3}$), HNSC ($p = 4.08 \times 10^{-5}$), KIRC ($p = 5.77 \times 10^{-15}$), PAAD ($p = 1.62 \times 10^{-12}$), LIHC ($p = 1.78 \times 10^{-14}$), PRAD ($p = 7.88 \times 10^{-15}$), READ ($p = 3.08 \times 10^{-3}$), SARC ($p = 1.32 \times 10^{-10}$), and UCEC ($p = 5.59 \times 10^{-8}$). There were also moderate differences for CHOL and LUSC. The results here suggest that MT1F is epigenetically regulated in several malignancies, and that this epigenetic regulation may affect the expression of MT1F and may be involved in the progression of the disease.

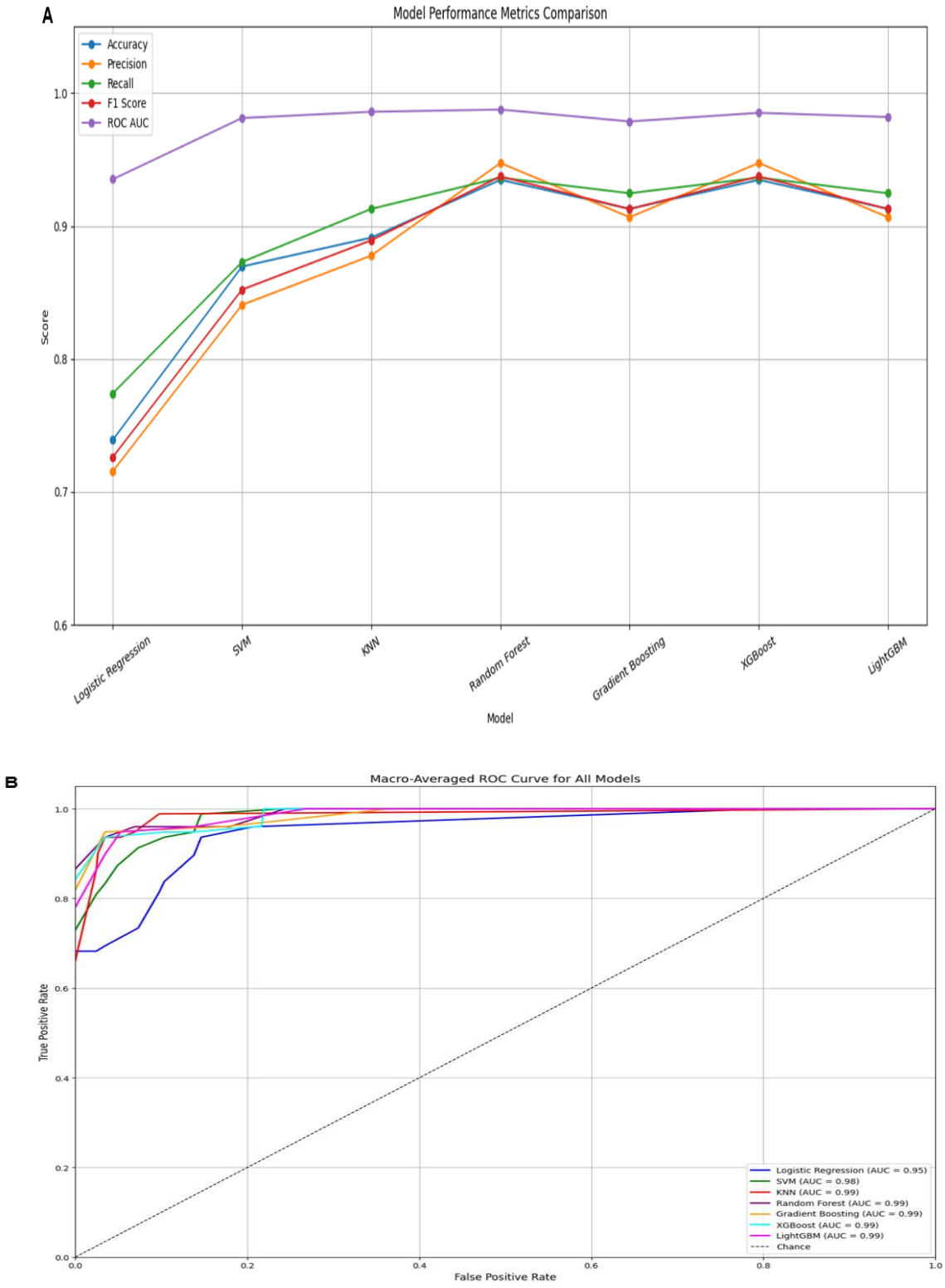

Figure 5: Illustrates (A) a comparison of machine learning (ML) model performances and (B) receiver operating characteristic (ROC) curves for all models

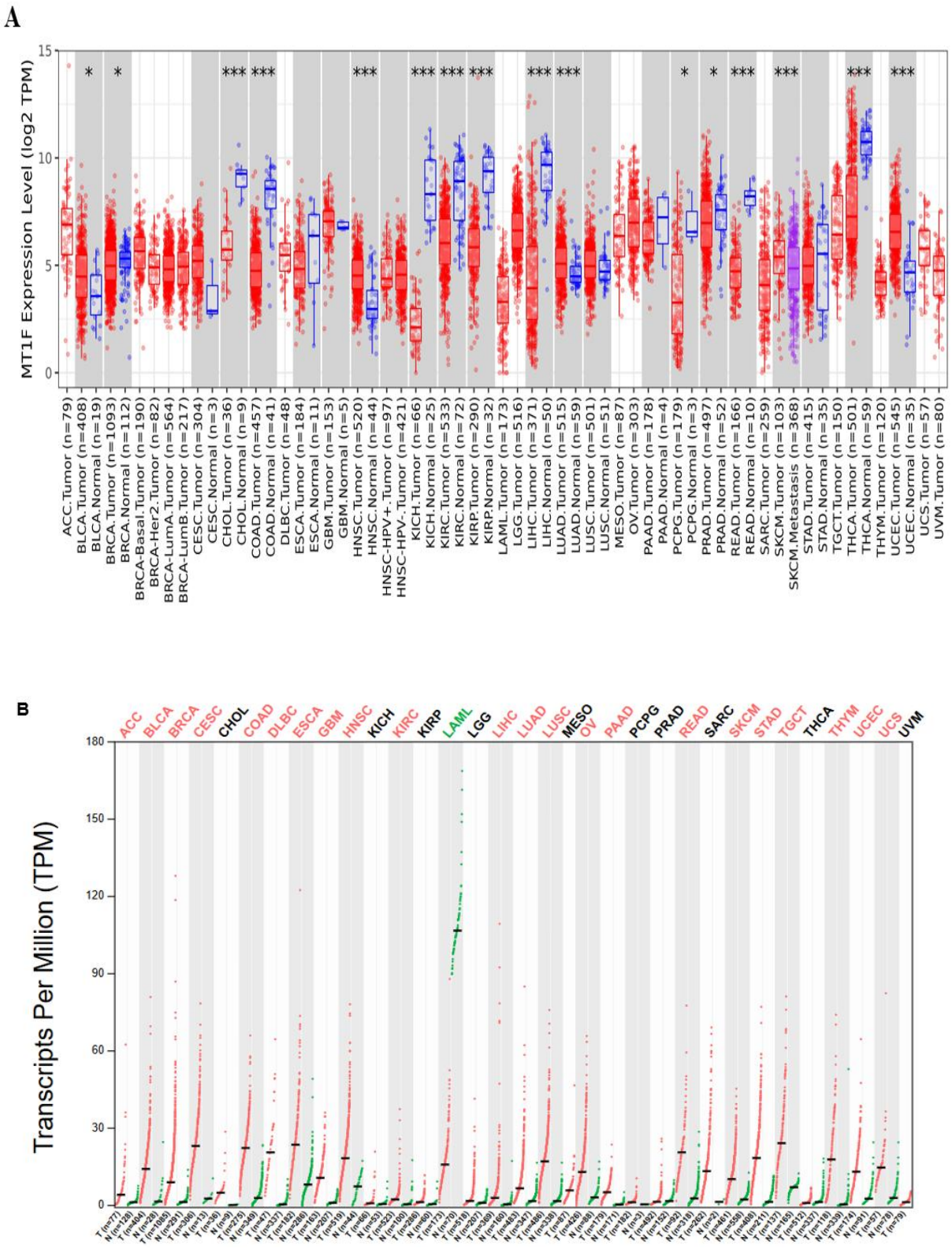

Figure 6: Pan-cancer expression analysis of MT1F. (A) Differential expression of *MT1F* across multiple cancer types compared to corresponding normal tissues using the TIMER2.0 database. Statistical significance is indicated as (*: p-value < 0.05; **: p-value < 0.01; ***: p-value < 0.001) (B) Validation of *MT1F* expression across cancers using GEPIA2, integrating TCGA and GTEx datasets, showing comparative expression levels between tumor and normal samples.

3.8. Correlation between immune cell infiltration and MT1F expression

TMT1F expression was correlated with immune cell infiltration in various types of cancer. MT1F expression was significantly correlated with levels of infiltration of B cells, CD4+ T cells, NK cells and T cell subsets (Figure 7). Spearman's correlation analysis revealed mostly positive correlations between MT1F expression and levels of immune cell infiltration in various cancer types, suggesting a possible role of MT1F in the regulation of the tumor immune microenvironment. On the other hand, a strong negative correlation was found between MT1F expression and cancer associated fibroblasts which implies an inverse association with stromal cells contributing to tumor progression. MT1F was demonstrated to be involved in immune regulation of the tumor microenvironment, and is suggested to serve as an immunological biomarker in canc

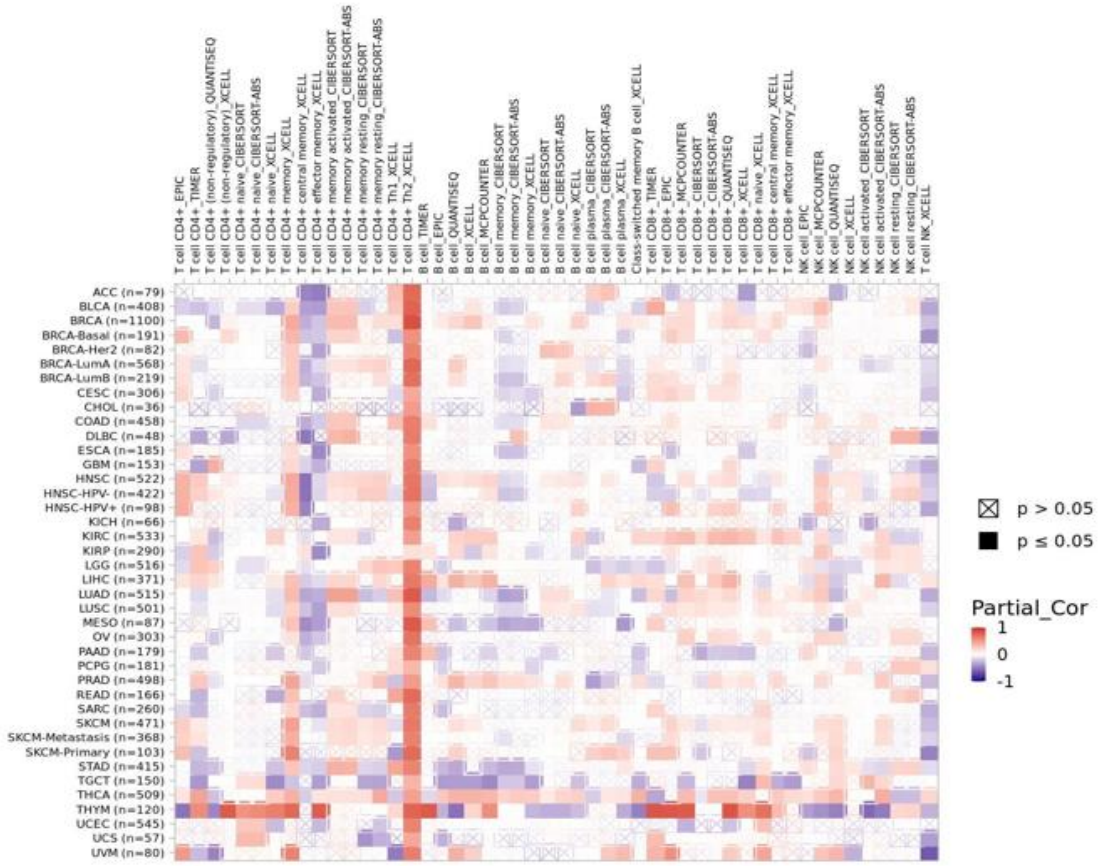

Figure 7: Depicts the relationship between MT1F gene expression and the levels of immune infiltration across various cancer type

3.9 Drug sensitivity analysis using GSCA datasets

Based on GSCA dataset, use it for a drug sensitivity analysis as described in step 3.9. The drug sensitivity analysis showed that the expression of MT1F had significant correlations with different anti-cancer drugs (Figure 8) after searching the CTRP and GDSC datasets. Expression of MT1F in the CTRP dataset was positively associated with the sensitivity to many compounds, such as HDAC inhibitors (entinostat, panobinostat) and cytotoxic agents (vincristine, cytarabine), indicating that high MT1F expression correlates with increased drug sensitivity. In some cases, however, agents were inversely correlated, suggesting possible resistance mechanisms, as with dasatinib. The same trends were seen in the GDSC dataset, with MT1F expression positively correlated with the sensitivity to targeted therapies such as Bortezomib, Dabrafenib and Temsirolimus. Altogether, these findings indicate a role of MT1F in the modulation of several key oncogenic pathways such as proteasome, MAPK and mTOR signaling pathways. Together, these results suggest that MT1F could be used as a predictive biomarker for therapeutic response and could be a target for precision therapeutic interventions in cancer.

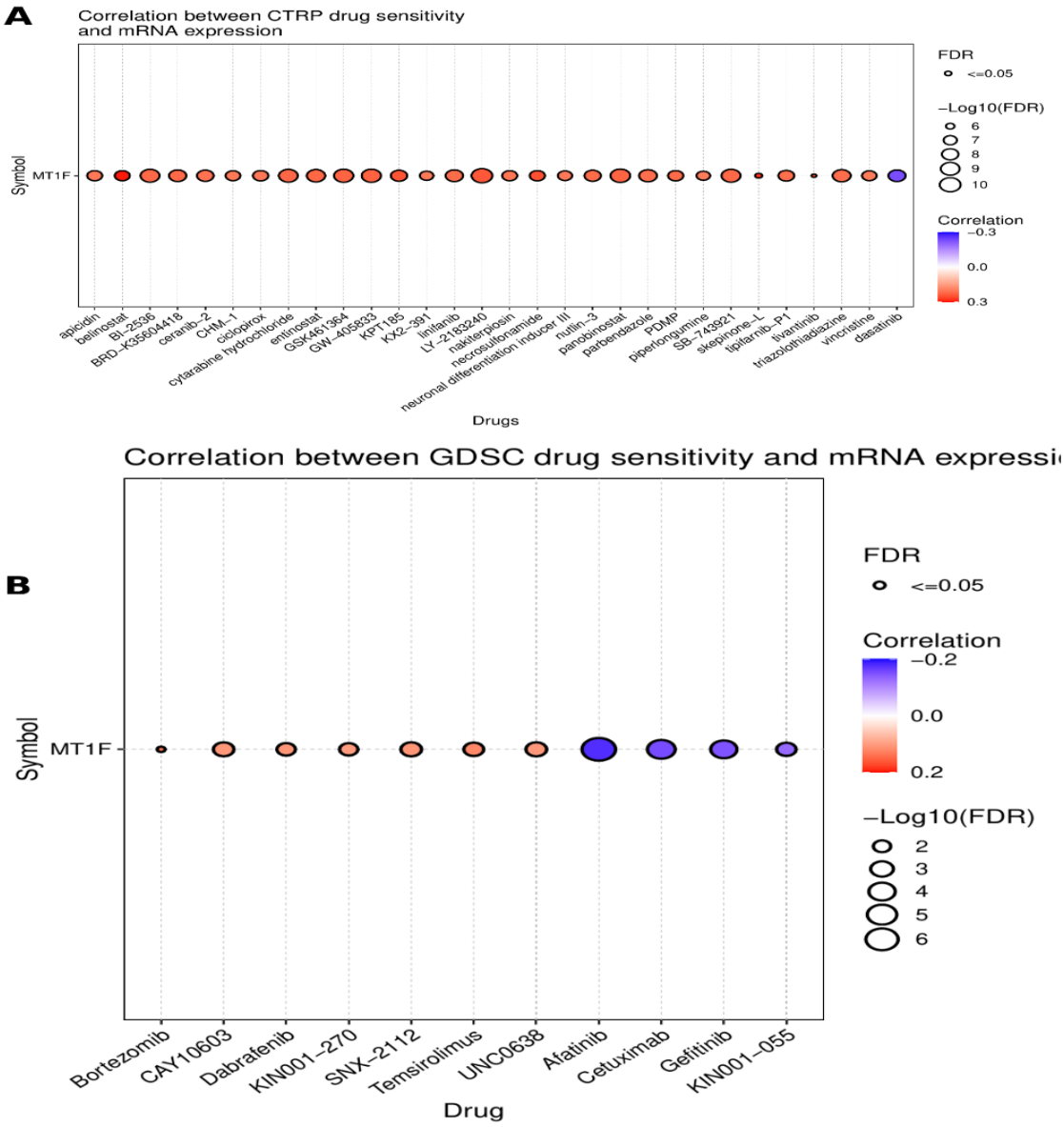

Figure 8: Correlation between MT1F expression and drug sensitivity across cancer cell lines**.** (A) The expression of MT1F in the CTRP dataset is correlated with several different anticancer agents. (B) validation of the drug-sensitivity correlations in the GDSC dataset, in which relationships between MT1F expression and response to targeted drugs are highlighted. Circle size represents statistical significance (−log10 FDR), and color indicates the direction and strength of correlation.

4. Discussion

In this study, we have created an integrated bioinformatics and machine learning approach to discover some solid tumor markers for breast cancer and found that MT1F is one of the important markers which might be playing an important role in diagnosis and therapy. MT1F was found to be a stable, predictive biomarker when integrated with DE analysis, network analysis and various ML models. Functional enrichment analysis identified that dysregulated genes were mainly involved in pathways related to extracellular matrix (ECM) organization and immune system pathways, which are known to play a role in tumor progression, such as epithelial–mesenchymal transition (EMT) and remodeling of the tumor microenvironment. MT1F was also indicated as a central hub gene by PPI network analysis, indicating that it may play a role in coordinating important molecular interactions. Protein–protein interaction (PPI) network analysis identified various hub genes, such as CCNA2 and MT1F. The role of CCNA2 in cell cycle control has been well documented, but the role of MT1F is remarkable. Although the functions of metallothioneins in the regulation of metal ion homeostasis and oxidative stress are not yet fully understood, their role in cancer, especially in BRCA has not been extensively studied. Based on the results obtained in our study, MT1F could potentially be a connection between cell-cycle regulation, metabolic adaptation, and microenvironmental interactions. The relevance of MT1F got further strengthened with

the implementation of machine learning. The ensemble classifiers (Random Forest, Gradient Boosting) had high predictive accuracy (AUC ≈ 0.99). MT1F was identified by all three algorithms in the LASSO, SVM-RFE, and Random Forest analysis, highlighting its robustness as a predictive biomarker and the importance of using multiple algorithms to ensure reliable biomarker discovery. MT1F was found in all feature selection approaches, but no formal stability analysis was done across repeated cross-validation folds, which is an area for future study. The use of an independent validation sample adds another layer of confidence to the validity and applicability of our results. MT1F, a member of the metallothionein family, has been mechanistically linked to control of metal ion homeostasis and oxidative stress, which are essential processes for the survival and adaptation of cancer cells (Diamant et al., 2025). It is associated with upregulation in breast tumors, which may be involved in the progression of these tumors by altering redox status, increasing resistance to apoptosis and affecting major oncogenic pathways, including PI3K/AKT and MAPK (Jiang et al., 2025). These findings are consistent with previous studies linking metallothioneins to cancer development and progression and extend existing knowledge by demonstrating the consistent prioritization of MT1F across multiple computational approaches (Yan et al., 2012).

The observed relations between the expression of MT1F and infiltration of immune cells indicate its possible participation in regulation of the tumor immune microenvironment. Positive correlations with B cells, CD4+ T cells, and NK cells suggest that it may play a role in immune regulation, and negative correlation with cancer associated fibroblasts suggest that it may be involved in stromal interactions (Dai et al., 2021). In addition, there are correlations with sensitivity to HDAC inhibitors, proteasome inhibitors and MAPK/mTOR-targeted therapy, indicating that MT1F could have an impact on therapy response by epigenetic regulation, protein degradation, and cell survival signaling pathways, respectively (Si & Lang, 2018). The pan-cancer analysis provides supportive evidence in support of the overall oncogenic relevance of MT1F; however, the main findings of this study are breast cancer-specific. Although the machine learning models showed a high predictive ability (AUC ≈ 0.99), care should be taken in interpreting these results, as they were based on a small number of observations which could lead to overfitting. Moreover, MT1F was found in all feature selection techniques and cross validated with an independent test set, but a formal stability analysis over repeated cross validation folds was not conducted. Other constraints are using public datasets and lack of experimental validation. MT1F is found to be a promising biomarker in breast cancer, with multi-layered computational evidence. The results offer solid grounds for experimental validation in the future and underscore the importance of integrative bioinformatics and machine learning methods in precision oncology.

5. Conclusion

In this study, we used an integrated bioinformatics and machine learning pipeline to discover biomarkers of breast cancer. MT1F was consistently identified as a central hub gene with high diagnostic potential from differential expression, network analysis and multi algorithm feature selection, and was validated by ensemble machine learning classifiers. The pan cancer expression and DNA methylation analysis revealed extensive epigenetic modification and expression changes in MT1F across the various cancers, demonstrating its broader role in oncogenesis. An important link with immune cell infiltration suggests a role in modulating the tumor microenvironment, while drug sensitivity analysis suggests that MT1F could be potentially useful to predict drug sensitivity to drugs targeting cell cycle and key signaling pathways. In conclusion, MT1F is a potential candidate for breast cancer diagnosis and therapeutic target. The results provide justification for the use of integrative computational approaches to discover biomarkers in the future and suggest further experimental validation and clinical assessment.

Declarations

Ethics approval and consent to participate

Not applicable as this study does not involve human participants or animals.

Consent for publication

Not applicable.

Availability of data and materials

The datasets analyzed during the current study are publicly available in the Gene Expression Omnibus (GEO). Additional data supporting the findings of this study are available from the corresponding author upon reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Authors'contributions

Conceptualization: Tahmid Islam Akib; Methodology: Tahmid Islam Akib, Ariful Islam ; Formal analysis: Tahmid Islam Akib, Ibrahim Hossain Molla, Ariful Islam.;Data curation: Tahmid Islam Akib, Ibrahim Hossain Molla, Nafees Rishad, Md Mahfujar Rahman Shakil, Labib Rahman, Nishad Al Hasan ;Validation: Tahmid Islam Akib, Sidratul Muntaha Tasmi.;Investigation: Tahmid Islam Akib, Ibrahim Hossain Molla.;Writing – original draft: Tahmid Islam Akib, Ibrahim Hossain Molla, Nafees Rishad, Md Mahfujar Rahman Shakil, Labib Rahman, Nishad Al Hasan. Writing – review & editing: Sidratul Muntaha Tasmi, Ariful Islam. Resources: Ibrahim Hossain Molla, Nafees Rishad.;Funding acquisition: Ibrahim Hossain Molla, Nafees Rishad.;Supervision: Ariful Islam. All authors read and approved the final manuscript.

Acknowledgements

Not Applicable.

References

Abd-Elnaby, M., Alfonse, M., & Roushdy, M. (2021). Classification of breast cancer using microarray gene expression data: A survey. *Journal of biomedical informatics*, *117*, 103764. crossref

Ali, M. Y., Asha, A. J., Riti, F. A., Nayeem, S. M., Rudro, R. H., Sarkar, N. K., Ali, M. M., Ahmmed, T., Jamal, M. A. H. M., & Haque, J. (2025). Identification and in silico evaluation of natural compounds from Annona muricata (soursop) leaves for colon cancer treatment. *Scientific reports*. crossref

Bader, G. D., & Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics*, *4*(1), 2. crossref

Barabasi, A.-L., & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature reviews genetics*, *5*(2), 101-113. crossref

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32. crossref

Cao, W., Qin, K., Li, F., & Chen, W. (2024). Comparative study of cancer profiles between 2020 and 2022 using global cancer statistics (GLOBOCAN). *Journal of the National Cancer Center*, *4*(2), 128-134. crossref

Chandrashekar, D. S., Bashel, B., Balasubramanya, S. A. H., Creighton, C. J., Ponce-Rodriguez, I., Chakravarthi, B. V.,

& Varambally, S. (2017). UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. *Neoplasia*, *19*(8), 649-658. crossref

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357. crossref

Chin, C.-H., Chen, S.-H., Wu, H.-H., Ho, C.-W., Ko, M.-T., & Lin, C.-Y. (2014). cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC systems biology*, *8*(Suppl 4), S11. crossref

Consortium, G. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, *369*(6509), 1318-1330. crossref

Dai, H., Wang, L., Li, L., Huang, Z., & Ye, L. (2021). Metallothionein 1: A new spotlight on inflammatory diseases. *Frontiers in immunology*, *12*, 739918. crossref

Diamant, I., Clarke, D. J., Evangelista, J. E., Lingam, N., & Ma'ayan, A. (2025). Harmonizome 3.0: integrated knowledge about genes and proteins from diverse multi-omics resources. *Nucleic acids research*, *53*(D1), D1016-D1028. crossref

Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., Griss, J., Sevilla, C., Matthews, L., & Gong, C. (2022). The reactome pathway knowledgebase 2022. *Nucleic acids research*, *50*(D1), D687-D692. crossref

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., & Caligiuri, M. A. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, *286*(5439), 531-537. crossref

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, *46*(1), 389-422. crossref

Hossain, M. S., Khandocar, M. P., Riti, F. A., Ali, M. Y., Dey, P. R., Haque, S. J., Metouekel, A., Mengistie, A. A., Bourhia, M., & Khallouki, F. (2025). A comprehensive machine learning for high throughput Tuberculosis sequence analysis, functional annotation, and visualization. *Scientific reports*, *15*(1), 25866. crossref

Hossain, M. S., Rahman, M. A., Dey, P. R., Khandocar, M. P., Ali, M. Y., Snigdha, M., Coutinho, H. D. M., & Islam, M. T. (2024). Natural Isatin Derivatives Against Black Fungus: In Silico Studies: MS Hossain et al. *Current Microbiology*, *81*(5), 113. crossref

Hossen, M. A., Reza, M. S., Rana, M. M., Hossen, M. B., Shoaib, M., Mollah, M. N. H., & Han, C. (2024). Identification of most representative hub-genes for diagnosis, prognosis, and therapies of hepatocellular carcinoma. *Chinese Clinical Oncology*, *13*(3), 32. crossref

Jiang, M., Zhang, K., Zhang, Z., Zeng, X., Huang, Z., Qin, P., Xie, Z., Cai, X., Ashrafizadeh, M., & Tian, Y. (2025). PI3K/AKT/mTOR axis in cancer: from pathogenesis to treatment. *MedComm*, *6*(8), e70295. crossref

Kanakkanthara, A., Jeganathan, K. B., Limzerwala, J. F., Baker, D. J., Hamada, M., Nam, H.-J., Van Deursen, W. H., Hamada, N., Naylor, R. M., & Becker, N. A. (2016). Cyclin A2 is an RNA binding protein that controls Mre11 mRNA translation. *Science*, *353*(6307), 1549-1552. crossref

Karim, S., Iqbal, M. S., Ahmad, N., Ansari, M. S., Mirza, Z., Merdad, A., Jastaniah, S. D., & Kumar, S. (2023). Gene expression study of breast cancer using Welch Satterthwaite t-test, Kaplan-Meier estimator plot and Huber loss robust regression model. *Journal of King Saud University-Science*, *35*(1), 102447.

Khan, A., Rehman, Z., Hashmi, H. F., Khan, A. A., Junaid, M., Sayaf, A. M., Ali, S. S., Hassan, F. U., Heng, W., & Wei, D.-Q. (2020). An integrated systems biology and network-based approaches to identify novel biomarkers in breast cancer cell lines using gene expression data. *Interdisciplinary Sciences: Computational Life Sciences*, *12*(2), 155-168.

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., & Lachmann, A. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, *44*(W1), W90-W97. crossref

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, *28*(6), 882-883. crossref

Li, T., Fu, J., Zeng, Z., Cohen, D., Li, J., Chen, Q., Li, B., & Liu, X. S. (2020). TIMER2. 0 for analysis of tumor-infiltrating immune cells. *Nucleic acids research*, *48*(W1), W509-W514.

Liu, S., Wang, Z., Zhu, R., Wang, F., Cheng, Y., & Liu, Y. (2021). Three differential expression analysis methods for RNA sequencing: limma, EdgeR, DESeq2. *JoVE (Journal of Visualized Experiments)*(175), e62528. crossref

Makary, M. A., & Daniel, M. (2016). Medical error—the third leading cause of death in the US. *bmj*, *353*. crossref

Mohanti, B. K., Mathur, P., Jayarajah, U., Biswal, B. M., & Prinja, S. (2025). Introduction to cancer world. In *Radiation Oncology–Principles, Precepts and Practice: Volume I–Technical Aspects* (pp. 1-30). Springer. crossref

Morris, J. H., Apeltsin, L., Newman, A. M., Baumbach, J., Wittkop, T., Su, G., Bader, G. D., & Ferrin, T. E. (2011). clusterMaker: a multi-algorithm clustering plugin for Cytoscape. *BMC bioinformatics*, *12*(1), 436. crossref

Network, C. G. A. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, *490*(7418), 61-70. crossref

Organization, W. H. (2020). *WHO methods and data sources for country-level causes of death 2000-2019*.

Perou, C. M., Sørlie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., & Akslen, L. A. (2000). Molecular portraits of human breast tumours. *Nature*, *406*(6797), 747-752. crossref

Qian, Y., Daza, J., Itzel, T., Betge, J., Zhan, T., Marme, F., & Teufel, A. (2021). Prognostic cancer gene expression signatures: current status and challenges. *Cells*, *10*(3), 648. crossref

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.-H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., & Mesirov, J. P. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences*, *98*(26), 15149-15154. crossref

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, *270*(5235), 467-470. crossref

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, *13*(11), 2498-2504. crossref

Shi, Z.-Z., Fan, Z.-W., Chen, Y.-X., Xie, X.-F., Jiang, W., Wang, W.-J., Qiu, Y.-T., & Bai, J. (2019). Ferroptosis in carcinoma: regulatory mechanisms and new method for cancer therapy. *OncoTargets and therapy*, 11291-11304. crossref

Si, M., & Lang, J. (2018). The roles of metallothioneins in carcinogenesis. *Journal of hematology & oncology*, *11*(1), 107. crossref

Slenter, D. N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., Mélius, J., Cirillo, E., Coort, S. L., & Digles, D. (2018). WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic acids research*, *46*(D1), D661-D667. crossref

Slodkowska, E. A., & Ross, J. S. (2009). MammaPrint™ 70-gene signature: another milestone in personalized medical care for breast cancer patients. *Expert review of molecular diagnostics*, *9*(5), 417-422. crossref

Sotiriou, C., & Pusztai, L. (2009). Gene-expression signatures in breast cancer. *New England Journal of Medicine*, *360*(8), 790-800. crossref

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., & Bork, P. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, *47*(D1), D607-D613. crossref

Tang, Z., Kang, B., Li, C., Chen, T., & Zhang, Z. (2019). GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic acids research*, *47*(W1), W556-W560. crossref

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267-288. crossref

Yam, C., Fung, T., & Poon, R. (2002). Cyclin A in cell cycle control and cancer. *Cellular and Molecular Life Sciences CMLS*, *59*(8), 1317-1326. crossref

Yan, D.-W., Fan, J.-W., Yu, Z.-h., Li, M.-x., Wen, Y.-G., Li, D.-W., Zhou, C.-Z., Wang, X.-L., Wang, Q., & Tang, H.-M. (2012). Downregulation of metallothionein 1F, a putative oncosuppressor, by loss of heterozygosity in colon cancer tissue. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, *1822*(6), 918-926. crossref

Yu-Jing, T., Wen-Jing, T., & Biao, T. (2020). Integrated analysis of hub genes and pathways in esophageal carcinoma based on NCBI's gene expression Omnibus (GEO) database: A bioinformatics analysis. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, *26*, e923934-923931. crossref