Integrative Machine Learning Reveals Potential Signature Genes Using Transcriptomics in Colon Cancer

Mostafa Amir Hamza[1], and Md. Saiful Islam[2*]

[1]*Department of Biotechnology and Genetic Engineering, University of Development Alternative (UODA), Dhanmondi R/A, Dhaka 1209, Bangladesh.*

[2]*Department of Anatomy, Sher-e-Bangla Medical College, Barishal 8200, Bangladesh*

Abstract

Background

Colon cancer is a significant health burden in the world and the second leading cause of cancer-related deaths. Despite advancements in diagnosis and treatment, identifying robust biomarkers for early detection and therapeutic targets remains imperative.

Materials and methods

This study used an integrative approach combining transcriptomics and machine learning to identify signature genes and pathways associated with colon cancer. RNA-Seq data from The Cancer Genome Atlas- Colon Adenocarcinoma (TCGA-COAD) project, comprising 485 samples (444 tumors and 41 normal tissues), were analyzed.

Results

Differential gene expression analysis revealed 657 upregulated and 8,566 downregulated genes. Notably, EPB41L3, TSPAN7, and ABI3BP were identified as highly upregulated, while LYVE1, PLPP1, and NFE2L3 were significantly downregulated in tumor samples. Gene Set Enrichment Analysis (GSEA) identified dysregulated pathways, including E2F targets, MYC targets, and G2M checkpoints, underscoring cell cycle regulation and metabolic reprogramming alterations in colon cancer. Machine learning models-Random Forest, Neural Networks, and Logistic Regression-achieved high classification accuracy (97–99%) and near-perfect Area Under the Receiver Operating Characteristic Curve (AUC-ROC) values (approximately 1.00), validating their predictive capabilities. Key genes consistently identified across these models highlight their potential translational relevance as biomarkers. This study integrates differential expression analysis, pathway enrichment, and machine learning to uncover critical insights into colon cancer biology.

Conclusion

The findings lay the groundwork for developing diagnostic and therapeutic strategies, with the identified genes and pathways serving as promising candidates for future validation and clinical applications. This approach exemplifies the potential of precision medicine to advance colon cancer research and improve patient outcomes.

Keywords: Cancer Genome Atlas, Colon Cancer, Machine Learning, Transcriptomics,

Introduction

Colon cancer is a major global health concern, ranking as the third most common cancer worldwide and accounting for approximately 10% of all cancer cases. It is also the second leading cause of cancer-related deaths globally. Despite advancements in diagnosis and treatment, the lack of robust biomarkers for early detection and therapeutic targeting remains a critical challenge. In 2020, an estimated 1.9 million new cases of colorectal cancer and over 930,000 related deaths were reported worldwide[1-3]. This highlights the urgent need for innovative approaches to uncover molecular drivers of colon cancer and translate them into clinical applications.

Excluding skin cancers, colon cancer is one of the most frequently diagnosed cancers in both men and women. It ranks third in cancer-related deaths among men and fourth among women, but collectively, it is the second leading cause of cancer mortality. In 2024, the American Cancer Society projects approximately 106,590 new cases of colon cancer (54,210 in men and 52,380 in women)[4]. Alarmingly, the incidence of colon cancer is increasing among younger adults, where it has become the leading cause of cancer-related deaths in men under 50 and the second leading cause in women under 50, following breast cancer. Each generation born since the 1950s faces a higher risk than the previous one. Colon cancer is expected to account for a significant proportion of the 53,010 CRC-related deaths anticipated in the United States in 2024[5, 6].

These statistics underscore the urgent need for regular screening, early detection, and lifestyle modifications to mitigate the risk of colon cancer. Advancements in RNA sequencing (RNA-Seq) have revolutionized transcriptomics, enabling comprehensive profiling of gene expression across various biological conditions, including cancer[7]. The Cancer Genome Atlas (TCGA) provides publicly available RNA-Seq datasets that facilitate in-depth molecular analyses of multiple cancer types[8].

This study focuses on identifying differentially expressed genes (DEGs) and pathways in colon cancer, a malignancy characterized by significant clinical challenges and heterogeneity. Differential expression analysis serves as a foundational step in uncovering genes with altered expression in tumors, while Gene Set Enrichment Analysis (GSEA) offers a pathway-level understanding of systemic changes in tumor biology[9].

Additionally, gene interaction networks provide a systems biology perspective, revealing key nodes and hubs that may serve as regulatory elements or therapeutic targets[10].

With the growing emphasis on precision medicine, the integration of machine learning with transcriptomics has gained momentum. Machine learning approaches, such as logistic regression, artificial neural networks, and random forests, offer powerful tools for feature selection, pattern recognition, and predictive modeling[11]. In this study, we leveraged these techniques using Python's sci-kit-learn library to identify potential marker genes capable of distinguishing tumors from normal samples, intending to enhance diagnostic and prognostic capabilities in colon cancer. By integrating these diverse methodologies, this study provides a holistic approach to biomarker discovery in colon cancer. It identifies potential marker genes and establishes a framework for leveraging RNA-Seq data in translational cancer research. The findings hold promise for advancing our understanding of colon cancer biology and contributing to developing personalized diagnostic and therapeutic strategies.

Materials and Methods

Data Acquisition and Preprocessing

Colon adenocarcinoma (COAD) gene expression data were retrieved from the TCGA database (https://portal.gdc.cancer.gov/repository) using TCGAbiolinks in R. We analyzed 485 samples (444 tumors and 41 normal tissues) after filtering for clinical data availability. Transcriptomic data from the TCGA-COAD project were selected, comprising 485 out of 524 samples: 444 tumor tissue samples and 41 matched normal tissue samples based on clinical data availability. Clinical data for 459 colon cancer patients were also downloaded, with key survival and staging information extracted for analysis.

Gene Expression Pre-Filtering and Normalization

Using TCGAbiolinks, transcriptomics data were downloaded as fragments per kilobase million (FPKM) unstranded normalized data[12, 13]. The dataset initially included 60,660 Ensembl gene identifiers. These Ensembl IDs were converted to gene names, resulting in 42,225 named genes. Genes with zero expression values across all samples were excluded, leaving 10,962 genes for further expression and machine learning analysis.

Since the data was normalized, a log2 transformation was applied to the FPKM data for consistency.

Gene Expression Analysis

Log2-transformed normalized data were used to assess gene expression alterations in tumor patients compared to those in healthy controls. A t-test was performed to identify significant gene expression changes, followed by false discovery rate (FDR) correction to adjust for multiple comparisons. Differentially expressed genes were defined as those with an FDR-adjusted p-value < 0.05. These genes were visualized using a heatmap and volcano plot. Sample clustering patterns were examined through a Principal Component Analysis (PCA) plot to explore group separations.

Gene Set Enrichment Analysis (GSEA)

Based on the differentially expressed genes (p-value < 0.05), hallmark pathway analysis was performed using the Molecular Signatures Database (MSigDB) with GSEA. The identified pathways were visualized using GSEA plots to understand the gene sets associated with specific pathway enrichment patterns[14, 15].

Building the Machine Learning Model:

Machine learning models, including Random Forest (RF), Neural Network (NN), and Logistic Regression (LR), were implemented using the Scikit-learn (sklearn) package in Python to analyze the filtered TCGA RNA-seq dataset[16-18]. The dataset was preprocessed by excluding genes with zero expression values across all samples, resulting in 10,962 genes, and a log2 transformation was applied to the normalized FPKM data. To ensure robust training and evaluation, the dataset was split into training (70%) and testing (30%) sets using stratified sampling to preserve the proportion of tumor and normal samples.

The Random Forest model was constructed using an ensemble learning approach with 100 decision trees, utilizing Gini impurity as the splitting criterion. Hyperparameters such as maximum tree depth and the minimum number of samples required for splits were optimized through grid search and cross-validation. The Neural Network model was designed as a multilayer perceptron with three hidden layers, each consisting of 50

neurons, and employed the rectified linear unit (ReLU) activation function. The Adam optimizer was used for weight updates, and early stopping was applied to mitigate overfitting. Hyperparameters, including the learning rate and batch size, were fine-tuned to maximize performance. The Logistic Regression model employed L2 regularization to prevent overfitting, with the maximum number of iterations set to 1,000 to ensure model convergence.

Model performance was assessed using accuracy, area under the receiver operating characteristic curve (AUC), and F1-score as evaluation metrics. The predictions generated by each model were further analyzed to identify significant marker genes associated with tumor and normal samples, providing insights into their predictive relevance and biological significance.

Results

Differential Gene Expression Analysis

After pre-filtering and normalization, we identified 10,962 genes with non-zero expression values suitable for further analysis (**Supplementary Table 1**). Differential expression analysis revealed 657 upregulated and 8,566 downregulated genes in tumor samples compared to matched normal controls (p-value < 0.05). Notably, EPB41L3, TSPAN7, and ABI3BP were among the most upregulated genes, while LYVE1, PLPP1, and NFE2L3 were significantly downregulated. A heatmap of the differentially expressed genes clearly demonstrated distinct segregation between tumor and normal samples, underscoring significant transcriptomic differences (**Fig. 1A, Supplementary Table 2 and 3**). Principal Component Analysis (PCA) further validated these differences, with tumor and normal samples forming distinct clusters along the principal components, reflecting the unique transcriptional landscapes of each group (**Fig. 1B, Supplementary Table 2**). A volcano plot highlighted the most notable genes, with the top 14 upregulated genes being Bystin like protein (BYSL), Solute Carrier Family 2 Member 13 (SLC2A13), Ectodermal Neural Cortex 1 (ENC1), Ajuba LIM Protein (AJUBA), Tumor Protein p53 Inducible Nuclear Protein 2 (TP53INP2), DEAD Box Helicase 56 (DDX56), Cbp/p300 Interacting Transactivator 2 (CITED2), PTEN Induced Kinase 1 (PINK1), Guanine Nucleotide Binding Protein G(I)/G(S)/G(O) Subunit Gamma 2 (GNG2), Semaphorin 6D (SEMA6D

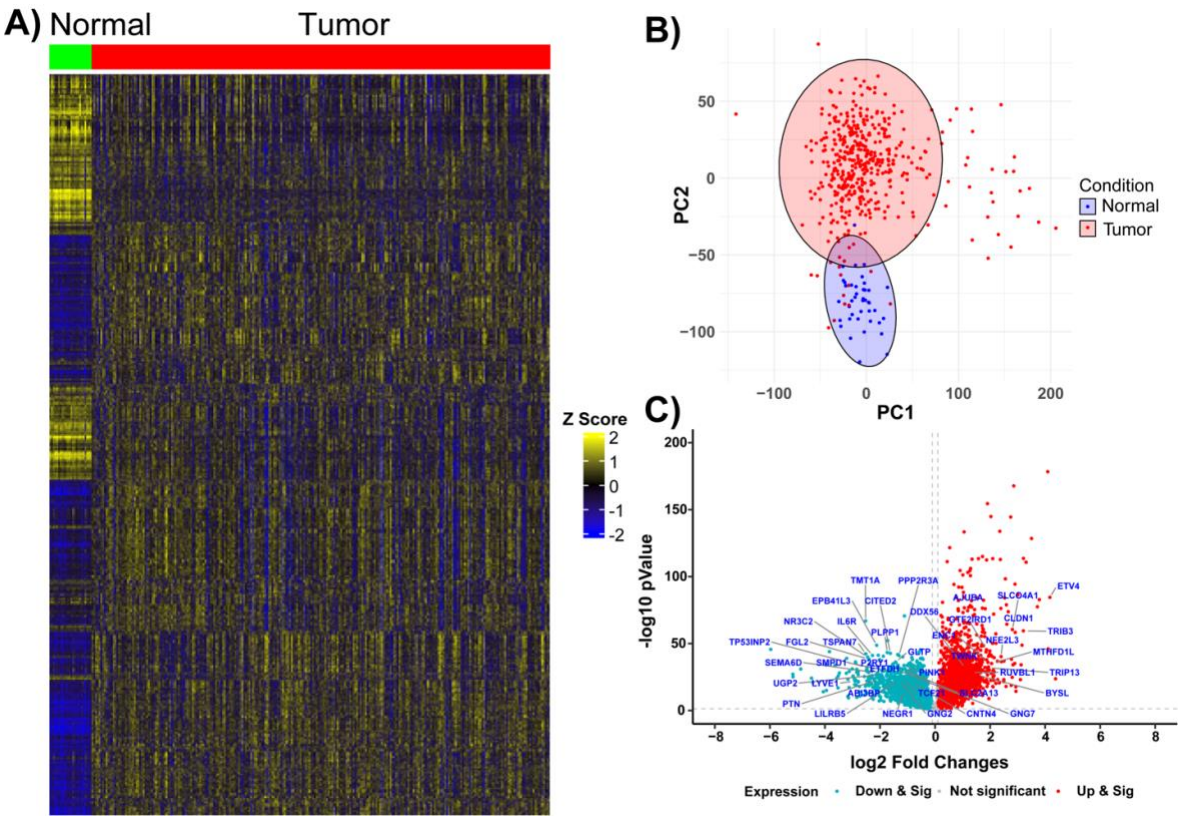


Figure 1. Differentially Expressed Genes in Tumor Tissue Samples of Colon Cancer A). The heatmap illustrates differentially expressed genes in tumor tissues (n = 444) compared to normal tissues (n = 41) from colon cancer patients. The gradient color scale represents increased expression (yellow) and decreased expression (blue). **B).** The PCA plot depicts the clustering of tumor and normal tissue samples from colon cancer patients. **C).** The volcano plot visualizes the distribution of differentially expressed genes between tumor and normal samples. The highlighted genes represent significantly altered genes identified using machine learning models- Random Forest (RF), Neural Network (NN), and Logistic Regression (LR)-all of which achieved AUC values exceeding 97%.

), Claudin 1 (CLDN1), Erythrocyte Membrane Protein Band 4.1 Like 3 (EPB41L3), Tetraspanin 7 (TSPAN7), ABI Family Member 3 Binding Protein (ABI3BP), and the top 27 downregulated genes including Lymphatic Vessel Endothelial Hyaluronan Receptor 1 (LYVE1), Phospholipid Phosphatase 1 (PLPP1), Nuclear Factor, Erythroid 2 Like 3 (NFE2L3), Electron Transfer Flavoprotein Dehydrogenase (ETFDH), Nuclear Receptor Subfamily 3 Group C Member 2 (NR3C2), Solute Carrier Organic Anion Transporter Family Member 4A1 (SLCO4A1), GTF2I Repeat Domain Containing 1 (GTF2IRD1), Twinkle mtDNA Helicase (TWNK), ETS Variant Transcription Factor 4

(ETV4), Transcription Factor 21(TCF21), Protein Phosphatase 2 Regulatory Subunit 3 Alpha (PPP2R3A), Sphingomyelin Phosphodiesterase 1 (SMPD1), Glycolipid Transfer Protein (GLTP), RuvB Like AAA ATPase 1 (RUVBL1), Purinergic Receptor P2Y1 (P2RY1), Thyroid Hormone Receptor Interactor 13 (TRIP13), Contactin 4 (CNTN4),

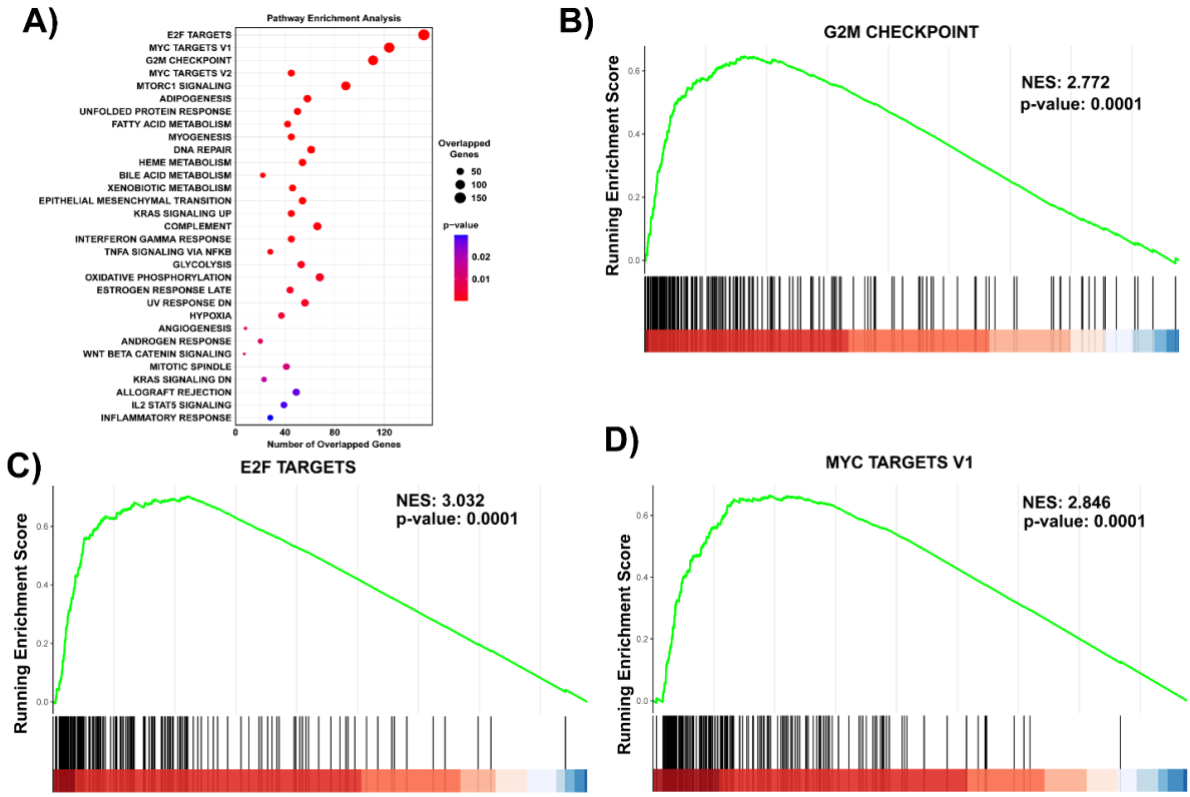


Figure 2. Gene Set Enrichment Analysis (GSEA) Pathway Analysis revealed the enriched biological pathways in colon cancer patients' tumor samples. A). The dot plot represents the enriched pathways identified using GSEA. Pathways are ranked based on enrichment scores, with dot size corresponding to the number of overlapping genes and color indicating statistical significance (adjusted p-value). The enrichment plots display the top three significantly enriched pathways: G2M Checkpoints (B), E2F Targets (C), and MYC Targets (D). The green curve represents the running enrichment score, indicating the accumulation of gene hits as ranked by their differential expression. Vertical black lines denote the positions of pathway genes within the ranked gene list. The normalized enrichment score (NES) and p-values are displayed for each pathway.

Methylenetetrahydrofolate Dehydrogenase (NADP+ Dependent) 1 Like (MTHFD1L), Fibrinogen Like 2 (FGL2), Neuronal Growth Regulator 1 (NEGR1), Interleukin 6 Receptor (IL6R), Thiol Methyltransferase 1A (TMT1A), UDP-Glucose Pyrophosphorylase 2 (UGP2), Tribbles Pseudokinase 3 (TRIB3), Pleiotrophin (PTN), Guanine Nucleotide-Binding Protein G(I)/G(S)/G(O) Subunit Gamma-7 (GNG7), and Leukocyte

Immunoglobulin-Like Receptor Subfamily B Member 5 (LILRB5). These genes exhibited highly significant (p-value < 0.05) (**Fig. 1C, Supplementary Table 2**). The differential expression patterns of these genes provide valuable insights into potential biomarkers and therapeutic targets in colon cancer.

Gene Set Enrichment Analysis (GSEA)

Gene Set Enrichment Analysis (GSEA) identified several hallmark pathways significantly enriched in tumor samples compared to healthy controls. The top 10 pathways included E2F targets, MYC targets V1, G2M checkpoint, MYC targets V2, MTORC1 signaling, adipogenesis, unfolded protein response, fatty acid metabolism, myogenesis, and DNA repair (**Fig. 2A, Supplementary Table 4**). The top-ranked pathway, E2F targets, involved critical cell cycle regulatory genes such as CDC25B, MYBL2, MYC, TRIP13, and UBE2S (**Fig. 2A**). These genes are important for cell cycle progression and are transcriptionally regulated by E2F transcription factors, highlighting their significant role in colon cancer pathogenesis. Enrichment plots of key pathways revealed a distinct concentration of altered genes at one end of the ranked gene list, supporting their biological significance in colon cancer progression (**Fig. 2B-2D**). The top three pathways-E2F targets, MYC targets V1, and G2M checkpoint-showed strong positive enrichment in tumor samples relative to normal controls, reinforcing their critical involvement in tumorigenesis (**Fig. 2B-2D**). These findings underscore the dysregulation of core processes such as cell cycle control, metabolic signaling, and stress responses in colon cancer. The enrichment of pathways like MTORC1 signaling and unfolded protein response points to metabolic rewiring and cellular stress adaptation as key features of tumor biology. Additionally, the dysregulation of fatty acid metabolism and adipogenesis suggests a potential link between lipid metabolism and tumor progression.

Machine Learning Analysis

The machine learning models, including Random Forest (RF), Neural Network (NN), and Logistic Regression (LR), achieved robust classification performance in distinguishing tumors from normal samples. The RF model demonstrated the highest accuracy (97%) and AUC (1.00), followed by NN (99%, AUC = 1.00) and LR (99%, AUC = 1.00). Area under curve (AUC > 0.97) analysis from the RF, NN, and LR model identified 41 top

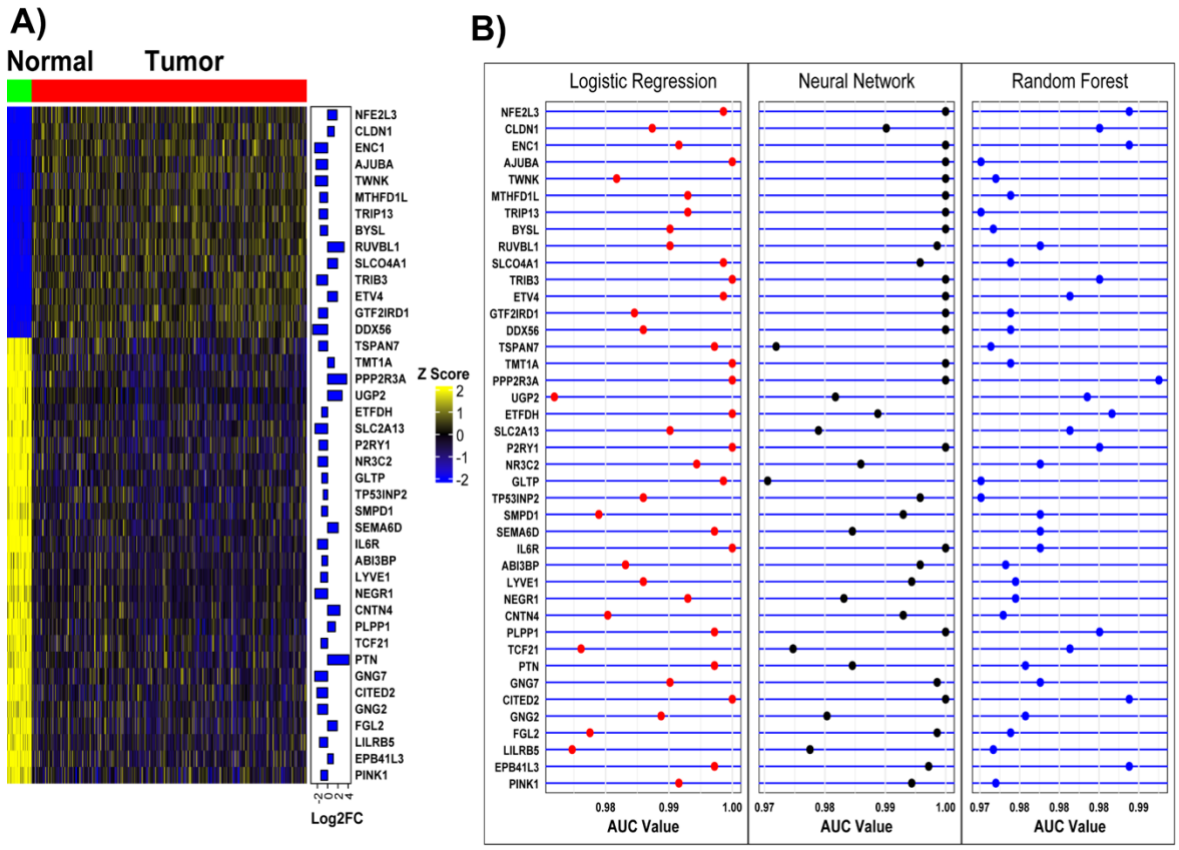


Figure 3. Machine Learning Models Classify Signature Genes in Tumor Patients. A). The heatmap displays differentially expressed genes and their fold changes in tumor tissues compared to normal tissues from colon cancer patients. These genes were identified using machine learning models-Random Forest (RF), Neural Network (NN), and Logistic Regression (LR)-all of which achieved AUC values exceeding 97%. **B).** The dot plot illustrates the AUC distribution across the three machine learning models in tumor tissue samples, highlighting their classification performance.

contributors' genes to classification performance. These genes were consistently highlighted across models, indicating their potential as robust biomarkers and significantly altered in tumor samples compared to healthy samples in colon cancer (**Fig. 3, Supplementary Table 5 and 6**). The AUC-ROC curves demonstrate the models' performance in distinguishing between classes, with the area under the curve (AUC) serving as a metric for classification accuracy. Higher AUC values indicate better model performance. The k-fold cross-validation approach ensures robustness and generalizability by partitioning the dataset into multiple subsets for training and validation (**Supplementary Fig. 1**). The top 3 genes (log2FC >3.00) of EPB41L3, TSPAN7, and

ABI3BP genes were significantly increased, whereas another top 3 genes (log2FC < -2.5) of LYVE1, PLPP1, and NFE2L3 were significantly decreased in tumor samples compared to normal samples in colon cancer (**Fig. 3**).

Identification of Potential Biomarkers

The integrative analysis, leveraging differential gene expression, Gene Set Enrichment Analysis (GSEA), and machine learning approaches, identified a subset of genes with high predictive potential in colon cancer. The top three upregulated genes were EPB41L3, TSPAN7, and ABI3BP, which exhibited significantly increased expression in tumor samples compared to normal samples (**Fig. 3**). These genes are likely involved in tumor progression and could serve as potential targets for therapeutic intervention. Conversely, the top three downregulated genes were LYVE1, PLPP1, and NFE2L3, which were significantly suppressed in tumor samples (**Fig. 3**). Their decreased expression may indicate disruption in pathways critical for maintaining normal cellular homeostasis and immune response. These findings underscore the potential of these genes as robust biomarkers for distinguishing tumors from normal samples. Furthermore, their consistent identification across multiple analytical approaches lays a strong foundation for subsequent validation studies. Ultimately, these biomarkers hold promise for advancing colon cancer diagnostics and therapeutics, paving the way for personalized medicine strategies.

Discussion

This study demonstrates the power of integrating transcriptomics and machine learning to uncover robust biomarkers and pathways in colon cancer. Our findings reveal distinct transcriptional and pathway-level alterations that differentiate tumors from normal samples, providing critical insights into colon cancer biology. By combining differential gene expression analysis, GSEA, and machine learning approaches, we uncovered distinct transcriptional and pathway-level alterations that differentiate tumors from normal samples. These findings not only deepen our understanding of colon cancer biology but also lay a foundation for the development of diagnostic and therapeutic strategies.

The differential gene expression analysis revealed a substantial number of genes with altered expression in tumor samples, including 657 upregulated and 8,566 downregulated

genes. Among these, EPB41L3, TSPAN7, and ABI3BP emerged as the most upregulated genes (log2FC > 3.00), suggesting their potential involvement in tumor progression. These genes have been implicated in cellular adhesion, signaling, and modulation of the tumor microenvironment, all critical processes in cancer[10]. Conversely, the top three downregulated genes, LYVE1, PLPP1, and NFE2L3 (log2FC < -2.5), suggest disrupted immune signaling and lipid metabolism in tumor tissues. For instance, LYVE1 is linked to lymphatic vessel integrity and immune regulation[9], and its suppression may contribute to immune evasion. Similarly, PLPP1 and NFE2L3, associated with lipid metabolism and oxidative stress[8], underscore the metabolic vulnerabilities of tumor cells.

Gene Set Enrichment Analysis (GSEA) provided additional insights into the pathways disrupted in colon cancer. Key pathways, including E2F targets, MYC targets, and G2M checkpoints, were significantly enriched in tumor samples. These findings emphasize dysregulated cell cycle control, proliferation, and metabolic rewiring as hallmarks of colon cancer[7]. The identification of pathways such as MTORC1 signaling and unfolded protein response highlights the ability of colon cancer cells to adapt to metabolic stress and optimize survival in nutrient-limited environments[11]. Enrichment of pathways like fatty acid metabolism and adipogenesis further underscores the interplay between lipid metabolism and tumor progression.

Machine learning analyses reinforced the robustness of the identified biomarkers. The Random Forest, Neural Network, and Logistic Regression models achieved high classification accuracy (97-99%) and AUC values (1.00), demonstrating their predictive power in distinguishing tumors from normal samples. Importantly, identifying key genes across models consistently underscores their translational relevance. Genes such as EPB41L3[19], TSPAN7[20], and ABI3BP[21-24], as well as LYVE1[25-27], PLPP1[28], and NFE2L3[29-32], were consistently highlighted as significant contributors to classification performance. These results confirm the utility of integrating transcriptomics with machine learning for biomarker discovery[11].

The observed transcriptional and pathway alterations align with previous studies while also providing novel insights into the molecular underpinnings of colon cancer[33]. For

instance, the enrichment of E2F and MYC target pathways supports the role of dysregulated transcriptional networks in driving tumor growth[9, 34, 35]. Additionally, the suppression of immune-related genes, such as LYVE1[25-27], highlights potential mechanisms of immune evasion, which may be critical for tumor survival and progression[36].

While this study provides valuable insights, there are limitations that warrant further investigation. Validation in independent cohorts and experimental models is essential to confirm the identified biomarkers and pathways. Additionally, functional studies are needed to elucidate the precise roles of these genes and pathways in colon cancer pathogenesis. Integrating additional omics datasets, such as proteomics or metabolomics, could offer a more comprehensive understanding of tumor biology and uncover additional therapeutic opportunities[10].

In conclusion, our integrative approach identifies signature genes and pathways in colon cancer, offering promising candidates for diagnostic and therapeutic development. Future studies should focus on validating these biomarkers in independent cohorts and exploring their functional roles to advance precision medicine in colon cancer. The identified genes and pathways provide promising candidates for diagnostic and therapeutic applications, paving the way for precision medicine strategies. Future studies focusing on experimental validation and clinical translation will be crucial for leveraging these findings to improve patient outcomes in colon cancer.

Author Contributions:

Conceptualization: M. A. H, and M. S.I; Methodology: M. A. H; Data Collection: M. A. H; Manuscript Preparation: M. A. H, and M. S.I; Writing-review and editing: M. A. H, and M. S.I; All authors have read and agreed to the published version of the manuscript.

Declarations

Conflicts of interest

The authors declare no competing interests.

Ethical approval:

Not applicable

References

1. Xi Y, Xu P. Global colorectal cancer burden in 2020 and projections to 2040. Transl Oncol. 2021;14(10):101174. Epub 20210706. doi: 10.1016/j.tranon.2021.101174. PubMed PMID: 34243011; PMCID: PMC8273208.
2. Sawicki T, Ruszkowska M, Danielewicz A, Niedzwiedzka E, Arlukowicz T, Przybylowicz KE. A Review of Colorectal Cancer in Terms of Epidemiology, Risk Factors, Development, Symptoms and Diagnosis. Cancers (Basel). 2021;13(9). Epub 20210422. doi: 10.3390/cancers13092025. PubMed PMID: 33922197; PMCID: PMC8122718.
3. Siegel RL, Wagle NS, Cercek A, Smith RA, Jemal A. Colorectal cancer statistics, 2023. CA: A Cancer Journal for Clinicians. 2023;73(3):233-54. doi: https://doi.org/10.3322/caac.21772.
4. Walter Reed National Military Medical Center M. Colorectal Cancer Awareness Month: Early detection is the best prevention 2024. Available from: https://walterreed.tricare.mil/News-Gallery/Articles/Article/3719070/colorectal-cancer-awareness-month-early-detection-is-the-best-prevention#:~:text=According%20to%20the%20American%20Cancer,men%20and%2019%2C890%20in%20women).
5. Augustus GJ, Ellis NA. Colorectal Cancer Disparity in African Americans: Risk Factors and Carcinogenic Mechanisms. Am J Pathol. 2018;188(2):291-303. Epub 20171109. doi: 10.1016/j.ajpath.2017.07.023. PubMed PMID: 29128568; PMCID: PMC5785537.
6. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. CA Cancer J Clin. 2024;74(1):12-49. Epub 20240117. doi: 10.3322/caac.21820. PubMed PMID: 38230766.
7. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57-63. doi: 10.1038/nrg2484. PubMed PMID: 19015660; PMCID: PMC2949280.
8. Tomczak K, Czerwinska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn). 2015;19(1A):A68-77. doi: 10.5114/wo.2014.47136. PubMed PMID: 25691825; PMCID: PMC4322527.
9. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545-50. Epub 20050930. doi: 10.1073/pnas.0506580102. PubMed PMID: 16199517; PMCID: PMC1239896.

10. Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12(1):56-68. doi: 10.1038/nrg2918. PubMed PMID: 21164525; PMCID: PMC3140052.

11. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015;16(6):321-32. Epub 20150507. doi: 10.1038/nrg3920. PubMed PMID: 25948244; PMCID: PMC5204302.

12. Mounir M, Lucchetta M, Silva TC, Olsen C, Bontempi G, Chen X, Noushmehr H, Colaprico A, Papaleo E. New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. PLoS Comput Biol. 2019;15(3):e1006701. Epub 20190305. doi: 10.1371/journal.pcbi.1006701. PubMed PMID: 30835723; PMCID: PMC6420023.

13. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot TS, Malta TM, Pagnotta SM, Castiglioni I, Ceccarelli M, Bontempi G, Noushmehr H. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016;44(8):e71. Epub 20151223. doi: 10.1093/nar/gkv1507. PubMed PMID: 26704973; PMCID: PMC4856967.

14. Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, Wadi L, Meyer M, Wong J, Xu C, Merico D, Bader GD. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nat Protoc. 2019;14(2):482-517. doi: 10.1038/s41596-018-0103-9. PubMed PMID: 30664679; PMCID: PMC6607905.

15. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1(6):417-25. doi: 10.1016/j.cels.2015.12.004. PubMed PMID: 26771021; PMCID: PMC4707969.

16. Okoro PC, Schubert R, Guo X, Johnson WC, Rotter JI, Hoeschele I, Liu Y, Im HK, Luke A, Dugas LR, Wheeler HE. Transcriptome prediction performance across machine learning models and diverse ancestries. HGG Adv. 2021;2(2). Epub 20210105. doi: 10.1016/j.xhgg.2020.100019. PubMed PMID: 33937878; PMCID: PMC8087249.

17. Lopez-Cortes A, Cabrera-Andrade A, Vazquez-Naya JM, Pazos A, Gonzales-Diaz H, Paz YMC, Guerrero S, Perez-Castillo Y, Tejera E, Munteanu CR. Prediction of breast cancer proteins involved in immunotherapy, metastasis, and RNA-binding using molecular descriptors and artificial neural networks. Sci Rep. 2020;10(1):8515. Epub 20200522. doi: 10.1038/s41598-020-65584-y. PubMed PMID: 32444848; PMCID: PMC7244564.

18. Ellrott K, Wong CK, Yau C, Castro MAA, Lee JA, Karlberg BJ, Grewal JK, Lagani V, Tercan B, Friedl V, Hinoue T, Uzunangelov V, Westlake L, Loinaz X, Felau I, Wang PI, Kemal A, Caesar-Johnson SJ, Shmulevich I, Lazar AJ, Tsamardinos I, Hoadley KA, Cancer Genome Atlas

Analysis N, Robertson AG, Knijnenburg TA, Benz CC, Stuart JM, Zenklusen JC, Cherniack AD, Laird PW. Classification of non-TCGA cancer samples to TCGA molecular subtypes using compact feature sets. Cancer Cell. 2024. Epub 20241230. doi: 10.1016/j.ccell.2024.12.002. PubMed PMID: 39753139.

19. Son HJ, Choi EJ, Yoo NJ, Lee SH. Mutation and Expression of a Candidate Tumor Suppressor Gene EPB41L3 in Gastric and Colorectal Cancers. Pathol Oncol Res. 2020;26(3):2003-5. Epub 20191211. doi: 10.1007/s12253-019-00787-x. PubMed PMID: 31828581.

20. Qi Y, Li H, Lv J, Qi W, Shen L, Liu S, Ding A, Wang G, Sun L, Qiu W. Expression and function of transmembrane 4 superfamily proteins in digestive system cancers. Cancer Cell Int. 2020;20:314. Epub 20200716. doi: 10.1186/s12935-020-01353-1. PubMed PMID: 32694936; PMCID: PMC7364658.

21. Nong B, Guo M, Wang W, Songyang Z, Xiong Y. Comprehensive Analysis of Large-Scale Transcriptomes from Multiple Cancer Types. Genes (Basel). 2021;12(12). Epub 20211124. doi: 10.3390/genes12121865. PubMed PMID: 34946814; PMCID: PMC8701385.

22. Latini FR, Hemerly JP, Freitas BC, Oler G, Riggins GJ, Cerutti JM. ABI3 ectopic expression reduces in vitro and in vivo cell growth properties while inducing senescence. BMC Cancer. 2011;11:11. Epub 20110111. doi: 10.1186/1471-2407-11-11. PubMed PMID: 21223585; PMCID: PMC3032749.

23. Horpaopan S, Kirfel J, Peters S, Kloth M, Huneburg R, Altmuller J, Drichel D, Odenthal M, Kristiansen G, Strassburg C, Nattermann J, Hoffmann P, Nurnberg P, Buttner R, Thiele H, Kahl P, Spier I, Aretz S. Exome sequencing characterizes the somatic mutation spectrum of early serrated lesions in a patient with serrated polyposis syndrome (SPS). Hered Cancer Clin Pract. 2017;15:22. Epub 20171129. doi: 10.1186/s13053-017-0082-9. PubMed PMID: 29213343; PMCID: PMC5707812.

24. Chen W, Huang J, Xiong J, Fu P, Chen C, Liu Y, Li Z, Jie Z, Cao Y. Identification of a Tumor Microenvironment-Related Gene Signature Indicative of Disease Prognosis and Treatment Response in Colon Cancer. Oxid Med Cell Longev. 2021;2021:6290261. Epub 20210814. doi: 10.1155/2021/6290261. PubMed PMID: 34497681; PMCID: PMC8420973.

25. Sundov Z, Tomic S, Alfirevic S, Sundov A, Capkun V, Nincevic Z, Nincevic J, Kunac N, Kontic M, Poljak N, Druzijanic N. Prognostic value of MVD, LVD and vascular invasion in lymph node-negative colon cancer. Hepatogastroenterology. 2013;60(123):432-8. doi: 10.5754/hge12826. PubMed PMID: 23321007.

26. Parr C, Jiang WG. Quantitative analysis of lymphangiogenic markers in human colorectal cancer. Int J Oncol. 2003;23(2):533-9. doi: 10.3892/ijo.23.2.533. PubMed PMID: 12851706.

27. Capuano A, Pivetta E, Sartori G, Bosisio G, Favero A, Cover E, Andreuzzi E, Colombatti A, Cannizzaro R, Scanziani E, Minoli L, Bucciotti F, Amor Lopez AI, Gaspardo K, Doliana R, Mongiat M, Spessotto P. Abrogation of EMILIN1-beta1 integrin interaction promotes experimental colitis and colon carcinogenesis. Matrix Biol. 2019;83:97-115. Epub 20190831. doi: 10.1016/j.matbio.2019.08.006. PubMed PMID: 31479698.

28. Tang X, Brindley DN. Lipid Phosphate Phosphatases and Cancer. Biomolecules. 2020;10(9). Epub 20200902. doi: 10.3390/biom10091263. PubMed PMID: 32887262; PMCID: PMC7564803.

29. Saliba J, Coutaud B, Makhani K, Epstein Roth N, Jackson J, Park JY, Gagnon N, Costa P, Jeyakumar T, Bury M, Beauchemin N, Mann KK, Blank V. Loss of NFE2L3 protects against inflammation-induced colorectal cancer through modulation of the tumor microenvironment. Oncogene. 2022;41(11):1563-75. Epub 20220128. doi: 10.1038/s41388-022-02192-2. PubMed PMID: 35091681; PMCID: PMC8913363.

30. Palma M, Lopez L, Garcia M, de Roja N, Ruiz T, Garcia J, Rosell E, Vela C, Rueda P, Rodriguez MJ. Detection of collagen triple helix repeat containing-1 and nuclear factor (erythroid-derived 2)-like 3 in colorectal cancer. BMC Clin Pathol. 2012;12:2. Epub 20120209. doi: 10.1186/1472-6890-12-2. PubMed PMID: 22321245; PMCID: PMC3293008.

31. Bury M, Le Calve B, Lessard F, Dal Maso T, Saliba J, Michiels C, Ferbeyre G, Blank V. NFE2L3 Controls Colon Cancer Cell Growth through Regulation of DUX4, a CDK1 Inhibitor. Cell Rep. 2019;29(6):1469-81 e9. doi: 10.1016/j.celrep.2019.09.087. PubMed PMID: 31693889.

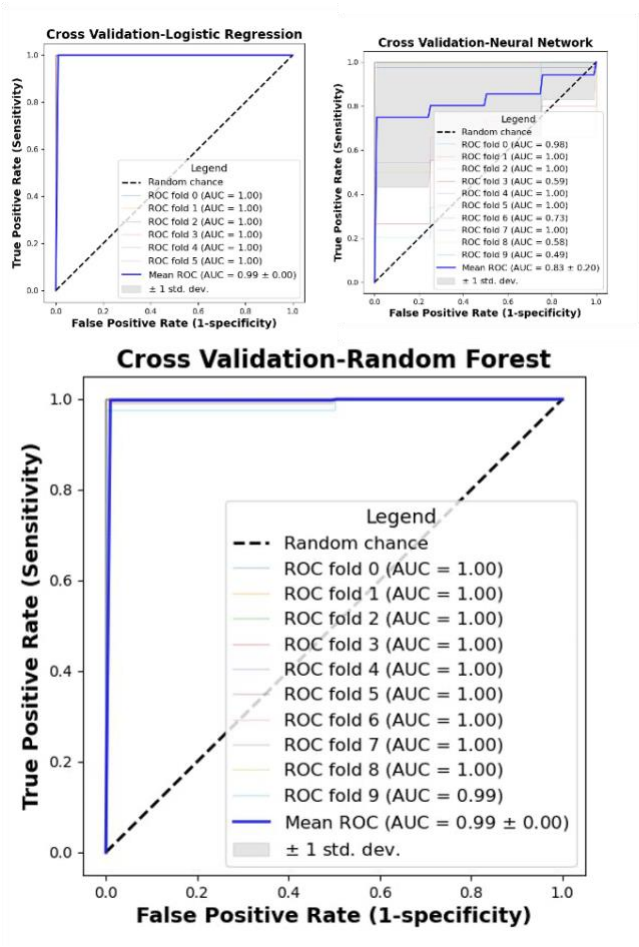
32. Aono S, Hatanaka A, Hatanaka A, Gao Y, Hippo Y, Taketo MM, Waku T, Kobayashi A. beta-Catenin/TCF4 Complex-Mediated Induction of the NRF3 (NFE2L3) Gene in Cancer Cells. Int J Mol Sci. 2019;20(13). Epub 20190708. doi: 10.3390/ijms20133344. PubMed PMID: 31288376; PMCID: PMC6651286.

33. Dunne PD, Arends MJ. Molecular pathological classification of colorectal cancer-an update. Virchows Arch. 2024;484(2):273-85. Epub 20240206. doi: 10.1007/s00428-024-03746-3. PubMed PMID: 38319359; PMCID: PMC10948573.

34. Oshi M, Takahashi H, Tokumaru Y, Yan L, Rashid OM, Nagahashi M, Matsuyama R, Endo I, Takabe K. The E2F Pathway Score as a Predictive Biomarker of Response to Neoadjuvant Therapy in ER+/HER2- Breast Cancer. Cells. 2020;9(7). Epub 20200708. doi: 10.3390/cells9071643. PubMed PMID: 32650578; PMCID: PMC7407968.

35. Johnson J, Thijssen B, McDermott U, Garnett M, Wessels LF, Bernards R. Targeting the RB-E2F pathway in breast cancer. Oncogene. 2016;35(37):4829-35. Epub 20160229. doi: 10.1038/onc.2016.32. PubMed PMID: 26923330; PMCID: PMC4950965.

36. Viudez-Pareja C, Kreft E, Garcia-Caballero M. Immunomodulatory properties of the lymphatic endothelium in the tumor microenvironment. Front Immunol. 2023;14:1235812. Epub 20230907. doi: 10.3389/fimmu.2023.1235812. PubMed PMID: 37744339; PMCID: PMC10512957.



Supplementary Figure 1. AUC-ROC Curves of k-Fold Cross-Validation for Machine Learning Models. The figure illustrates the Area Under the Receiver Operating Characteristic (AUC-ROC) curves for three machine learning models-Logistic Regression (LR), Neural Network (NN), and Random Forest (RF)-evaluated using k-fold cross-validation. The models were trained and tested on the dataset with the following configurations: Logistic Regression (LR): Utilized 5 splits for k-fold cross-validation. Neural Network (NN) and Random Forest (RF): Employed 9 splits for k-fold cross-validation. Random Forest (RF): Also used 9 splits for k-fold cross-validation.