Machine Learning-Driven Identification of Transcriptomic Biomarkers for Colon Adenocarcinoma Using TCGA-COAD Gene Expression Data

Mandrajula Arun Prabhu Teja[1], Mr. D. Ashok[2]
[1]*MSc Artificial Intelligence and Data Science, Department of Computer Science and Artificial Intelligence Central University of Andhra Pradesh Ananthapuramu, Andhra Pradesh, India*
[2]*Assistant Professor, Department of Computer Science and Artificial Intelligence Central University of Andhra Pradesh Ananthapuramu, Andhra Pradesh, India*
doi.org/10.64643/IJIRTV12I12-201002-459

Abstract*—Colon adenocarcinoma (COAD) is among the most prevalent and lethal malignancies worldwide, characterised by diverse and complex underlying molecular mechanisms. Early-stage diagnosis and prognosis remain particularly challenging, underscoring the urgent need for reliable, reproducible, and clinically actionable biomarkers. RNA sequencing (RNA-Seq) provides intricate transcriptome-wide gene expression profiles, but extracting meaningful insights demands sophisticated computational strategies. In this work we present a comprehensive, integrated machine learning-driven pipeline for biomarker discovery in COAD applied to the publicly available TCGA-COAD dataset. The pipeline combines rigorous preprocessing and normalisation of raw RNA-Seq count data, differential gene expression (DGE) analysis, Boruta algorithm-based all-relevant feature selection confirming 19 robust candidate genes, and Elastic Net regularised logistic classification (optimal $\lambda = 0.0176$). The resulting model achieves near-perfect discriminative power with AUC ≈ 0.994. Three genes—CA7, ABCA8, and CLEC3B—achieve individual cross-validated AUC of 1.00. Heatmap-based hierarchical clustering reveals unambiguous expression patterns separating tumour from normal samples. Kaplan–Meier survival analysis identifies CDH3-AS1 as a statistically significant prognostic marker ($p < 0.05$). The proposed framework is modular, reproducible, and extensible to multi-omics integration or other cancer cohorts, positioning it as a valuable tool for precision oncology and personalised treatment planning.

Index Terms*—Colon Adenocarcinoma, COAD, TCGA, RNA-Seq, Differential Gene Expression, Boruta Algorithm, Elastic Net, Machine Learning, Biomarker Discovery, Kaplan–Meier Survival Analysis, Precision Oncology.

I. INTRODUCTION

Colorectal cancer (CRC) is the third most commonly diagnosed cancer and the second leading cause of cancer-related mortality worldwide, with approximately 1.9 million new cases and 935,000 deaths recorded in 2020 alone [1]. Colon adenocarcinoma (COAD), arising from the glandular epithelial cells lining the colonic mucosa, constitutes the predominant histological subtype and accounts for the majority of CRC-associated morbidity and mortality [2].

The aetiology of COAD is multifactorial, driven by the interplay of genetic predispositions, somatic mutations, epigenetic alterations, dietary habits, sedentary lifestyle, obesity, and ageing [3]. Disease prognosis is strongly stage-dependent: while localised COAD carries a five-year survival rate exceeding 90%, metastatic disease reduces this figure to below 15%, underscoring the critical importance of early and accurate detection [3].

Despite significant advances in colonoscopic screening, surgical resection, and targeted therapies, a substantial fraction of COAD cases are diagnosed at advanced stages, limiting therapeutic options and worsening patient outcomes [4]. This clinical reality motivates the pursuit of sensitive, specific, and cost-effective molecular biomarkers that can complement or eventually replace invasive screening modalities.

RNA sequencing (RNA-Seq) enables measurement of RNA transcript abundance across thousands of genes simultaneously, providing a dynamic functional snapshot of cellular activity [7]. Transcriptomic profiles capture the functional state of cells and reflect gene expression changes induced by pathological conditions such as malignant transformation, making RNA-Seq particularly well-suited for discovering differentially expressed genes (DEGs) as candidate biomarkers. However,

extracting biologically and clinically meaningful information from RNA-Seq data is non-trivial due to the extreme high dimensionality—the classic "large p, small n" problem—which introduces risks of overfitting and false-positive biomarker identification [10].

In this paper, we propose and evaluate a multi-stage, machine learning-driven pipeline for transcriptomic biomarker discovery in COAD. Our key contributions are:

1. A rigorous preprocessing and normalisation protocol for TCGA-COAD RNA-Seq data, including quality filtering, library-size normalisation, $\log_2$ transformation, and batch effect correction.
2. Differential gene expression analysis with FDR control via the Benjamini–Hochberg procedure to obtain a statistically validated candidate gene set.
3. Application of the Boruta algorithm for all-relevant, stable feature selection validated against randomised shadow features.
4. Elastic Net regularised logistic regression for simultaneous feature weighting, sparse model training, and tumour/normal classification.
5. Comprehensive evaluation via ROC analysis, heatmap-based hierarchical clustering, and Kaplan–Meier survival analysis to assess both discriminative and prognostic value.

II. RELATED WORK

A. Transcriptomic Profiling in Colorectal Cancer

High-throughput gene expression profiling has substantially advanced our understanding of COAD molecular biology. RNA-Seq-based studies have established that COAD is not a single homogeneous disease but a collection of molecularly distinct subtypes. The Consensus Molecular Subtypes (CMS) framework partitions COAD into four subtypes—CMS1 (MSI immune), CMS2 (canonical), CMS3 (metabolic), and CMS4 (mesenchymal)—each with distinct prognostic implications and therapeutic vulnerabilities [8]. Early transcriptomic studies catalogued large sets of differentially expressed genes between tumour and normal colonic mucosa, revealing systematic dysregulation of cell cycle regulators, apoptotic machinery, DNA damage response pathways, and metabolic enzymes [4].

B. Machine Learning for Genomic Classification

Machine learning has emerged as an indispensable complement to classical statistical approaches in genomic data analysis [9]. Support Vector Machines, Random Forests, and gradient boosting methods have been applied to classify cancer subtypes and predict therapeutic response. Regularisation-based linear classifiers, particularly Lasso and Elastic Net, offer an attractive alternative by combining predictive accuracy with intrinsic feature selection and model interpretability [11], [12]. Their ability to handle correlated features—common in gene expression data due to co-regulation within biological pathways—makes them especially suited for transcriptomic classification tasks.

C. Feature Selection in High-Dimensional Genomics

The "large p, small n" challenge necessitates effective dimensionality reduction prior to machine learning modelling [10]. Filter methods such as FDR-based DEG filtering provide a computationally efficient first pass but do not account for feature interactions. The Boruta algorithm [10] iteratively evaluates feature importance in the context of a Random Forest model, using randomised shadow features as a competitive benchmark, enabling identification of all relevant features while minimising false-positive selection. Embedded methods such as Elastic Net perform selection as part of model fitting. The combination of wrapper validation (Boruta) and embedded selection (Elastic Net) provides a robust two-stage feature selection strategy.

D. Survival Analysis and Clinical Validation

A critical gap in many computational biomarker studies is the failure to validate statistical findings against clinical outcomes. The Kaplan–Meier estimator and log-rank test constitute the standard non-parametric framework for survival analysis in oncology [14], enabling assessment of whether gene expression levels associate with patient overall survival. Integrating survival analysis within the biomarker discovery pipeline ensures that computationally identified genes carry genuine prognostic value.

III. MATERIALS AND METHODS

A. Dataset

RNA-Seq gene expression count data were downloaded from The Cancer Genome Atlas Colon Adenocarcinoma (TCGA-COAD) project via the

GDC Data Portal [5], [6]. The dataset comprises 471 tumour samples and 41 matched adjacent normal tissue samples, with each sample characterised by expression quantification of over 20,000 protein-coding and non-coding genes. Clinical annotation data including overall survival time and vital status were retrieved for survival analysis. The TCGA-COAD cohort is one of the most extensively characterised CRC datasets available, with standardised data generation and quality control protocols.

B. Data Preprocessing and Normalisation

Raw RNA-Seq count data were processed through a four-step pipeline to ensure data quality and cross-sample comparability:

6. Quality filtering: genes with zero or near-zero expression (counts < 10 in > 90% of samples) were removed, eliminating approximately 40% of the raw gene list.
7. Library-size normalisation: Counts Per Million (CPM) normalisation was applied to correct for differences in sequencing depth across samples.
8. Log_2 transformation: $x' = \log_2(CPM + 1)$ was applied to reduce right-skewness in the expression distribution and stabilise variance.
9. Batch effect correction: PCA and hierarchical clustering were used to identify batch structure; ComBat-based correction was applied where confirmed.

C. Differential Gene Expression Analysis

DGE analysis was performed to identify genes significantly dysregulated in tumour versus normal tissue. For each gene j, $\log_2$ fold change (logFC) quantifies the magnitude of expression change between groups. A moderated t-statistic is computed per gene, and the Benjamini–Hochberg (BH) procedure was applied to control the False Discovery Rate (FDR) at 5%. Genes satisfying $|logFC| > 1$ AND adjusted $p < 0.05$ were retained as statistically and biologically significant DEGs. Results were visualised using a volcano plot depicting the relationship between fold change magnitude and statistical significance.

D. Feature Selection via Boruta Algorithm

The Boruta algorithm [10] was applied to the DEG-filtered expression matrix to confirm all-relevant features using an iterative Random Forest importance framework. For each gene j, an importance Z-score is computed relative to the maximum importance of randomly permuted "shadow" copies of all features. Genes with Z-scores consistently exceeding the shadow maximum across all iterations are confirmed as relevant; those failing are rejected; those in between remain tentative. Key advantages of Boruta over standard feature selection include: (i) identification of all relevant features rather than merely the minimal optimal subset; (ii) built-in statistical control against false-positive selection via comparison with randomised shadow features; and (iii) robustness to correlated features common in gene expression data.

E. Classification via Elastic Net Regularisation

An Elastic Net regularised logistic regression classifier [11], [12] was trained on the Boruta-confirmed gene set. The Elastic Net combines L1 (Lasso) and L2 (Ridge) penalties: the L1 component drives sparsity by shrinking uninformative coefficients to exactly zero, while the L2 component distributes weight among correlated features, improving coefficient stability. This combination is particularly beneficial for gene expression data where many genes share expression patterns due to shared regulatory mechanisms [13]. The optimal regularisation parameter λ was selected by minimising mean squared error over a 10-fold stratified cross-validation grid search with mixing parameter $\alpha = 0.5$.

F. Model Evaluation and Survival Analysis

Classification performance was assessed using: (i) ROC analysis and AUC [15] across all classification thresholds; (ii) cross-validated AUC (CV-AUC) via stratified 10-fold cross-validation; and (iii) sensitivity and specificity at the optimal Youden Index threshold. The Kaplan–Meier (KM) estimator was used to assess prognostic associations; patients were stratified into high- and low-expression groups based on the median expression of each confirmed gene. Differences between survival distributions were assessed using the log-rank test ($p < 0.05$).

IV. RESULTS AND DISCUSSION

A. Differential Gene Expression and Volcano Plot

Differential expression analysis of the TCGA-COAD dataset identified thousands of statistically significant genes. After applying thresholds of $|logFC| > 1$ and adjusted $p < 0.05$, a refined set of DEGs was obtained for downstream analysis. Fig. 1 provides an intuitive visualisation of the full gene

expression landscape, with each gene plotted according to its fold change magnitude and statistical significance.

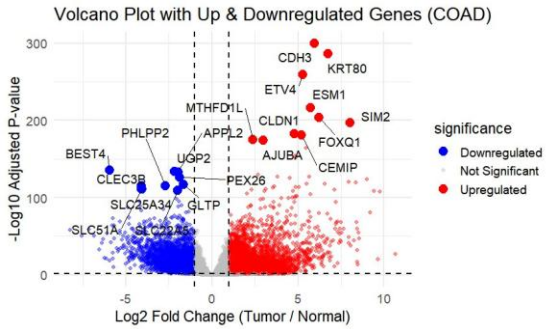

Fig. 1. Volcano plot of differential gene expression in COAD (Tumour vs. Normal). Red: significantly upregulated (|logFC| > 1, adj. p < 0.05); Blue: significantly downregulated; Grey: non-significant.

Prominently upregulated genes in tumour samples include CDH3 (P-cadherin, implicated in cancer cell invasion and EMT), KRT80 (cytoskeletal remodelling), ETV4 (ETS-family transcription factor), ESM1 (angiogenesis marker), CLDN1, and FOXQ1 (epithelial-mesenchymal transition regulator). Significantly downregulated genes include CLEC3B (ECM organisation and innate immunity), BEST4 (normal colonocyte chloride channel function), and multiple SLC family members (ion and metabolite homeostasis). The global pattern is consistent with enhanced proliferation, metabolic reprogramming, and immune evasion characteristic of colorectal malignancy [4], [8].

B. Boruta Feature Selection

After Boruta-based all-relevant feature selection, 19 genes were confirmed as statistically and biologically relevant classifiers. Fig. 2 displays the importance score distributions for each confirmed gene sorted by median importance, while Fig. 3 overlays the shadow feature distributions to contextualise the confirmation threshold.

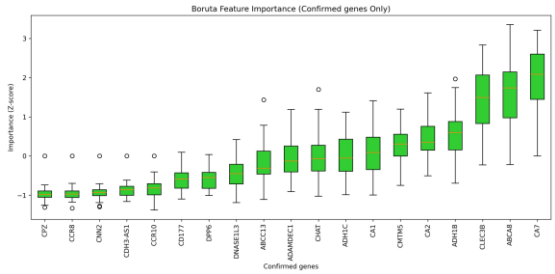

Fig. 2. Boruta feature importance boxplot for the 19 confirmed genes, sorted by median Z-score. Top-ranked genes—CA7, ABCA8, CLEC3B—exhibit consistently high, stable importance across all Random Forest iterations.

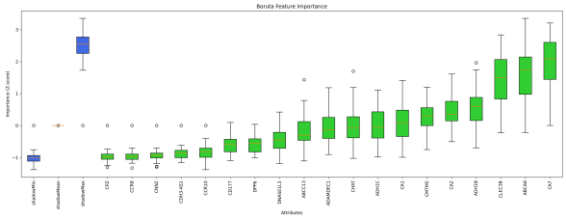

Fig. 3. Full Boruta importance plot including shadow features (blue). All 19 confirmed genes (green) clearly and consistently exceed the shadowMax threshold.

The carbonic anhydrase paralogs CA1, CA2, and CA7 catalyse the reversible hydration of CO_2 and are central regulators of intra- and extracellular pH homeostasis. Their systematic downregulation in COAD reflects cancer-driven metabolic reprogramming. ABCA8 (ATP-binding cassette lipid transporter), CLEC3B (C-type lectin, ECM remodelling and innate immunity), CMTM5 (candidate tumour suppressor), ADH1B and ADH1C (retinol metabolism), CHAT (cholinergic signalling), and ADAMDEC1 (ECM protease) together represent a biologically coherent and mechanistically motivated COAD signature.

C. Elastic Net Model Optimisation

The regularisation parameter λ was optimised via 10-fold cross-validation over a logarithmically spaced grid (Fig. 4). The cross-validation error curve exhibits the characteristic U-shape: overly small λ values permit overfitting, while excessively large values introduce underfitting. The optimal $\lambda = 0.0176$ ($\log\lambda \approx -4.04$) was selected at the minimum of the MSE curve.

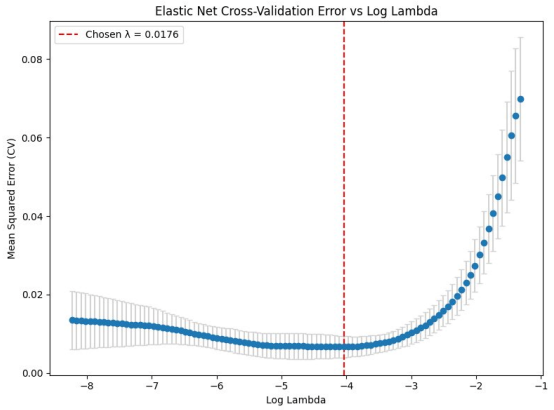

Fig. 4. Elastic Net cross-validation MSE vs. log(λ). Dashed red line marks optimal λ = 0.0176. Error bars show ±1 SD across 10 CV folds.

Fig. 5 displays the Elastic Net model coefficients for each Boruta-confirmed gene ordered by magnitude. Genes with the largest negative coefficients—CA2, CMTM5, CLEC3B, ABCA8, CA1, and ABCC13—are strongly associated with the normal tissue phenotype, consistent with their established roles in maintaining normal epithelial homeostasis. CDH3-AS1 exhibits a positive coefficient, implicating elevated expression in the tumour microenvironment.

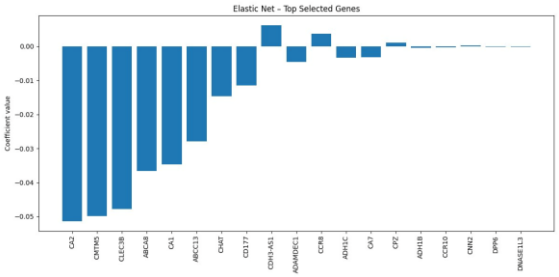

Fig. 5. Elastic Net coefficient plot for the 19 Boruta-confirmed genes. Negative coefficients indicate association with normal tissue; positive coefficients with tumour tissue.

D. ROC Curve Analysis and Classification Performance

The Elastic Net model trained on the 19 Boruta-confirmed genes achieves a composite AUC of 0.994 on the TCGA-COAD dataset (Fig. 6). This near-perfect discrimination between tumour and normal samples validates the quality of the entire pipeline, from preprocessing through feature selection to classification.

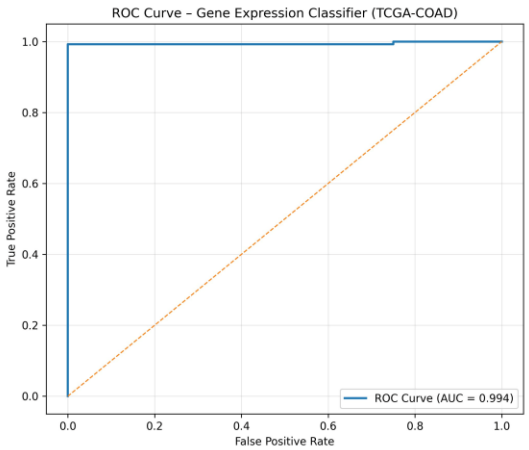

Fig. 6. ROC curve for the Elastic Net composite classifier on TCGA-COAD. AUC = 0.994. Dashed diagonal represents random-chance performance.

Cross-validated ROC curves for all 19 Boruta-confirmed genes are presented in Fig. 7. Three genes achieve CV-AUC = 1.00 (CA7, ABCA8, CLEC3B), indicating perfect individual discrimination generalising across all cross-validation folds. The majority of remaining genes achieve CV-AUC between 0.95 and 0.99, demonstrating robust and generalisable predictive power.

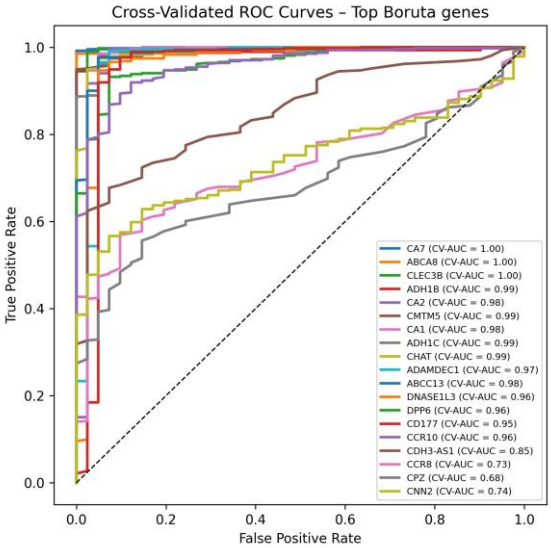

Fig. 7. Cross-validated ROC curves for each of the 19 Boruta-confirmed genes. Three genes (CA7, ABCA8, CLEC3B) achieve CV-AUC = 1.00.

TABLE I SUMMARY OF BORUTA-CONFIRMED BIOMARKER GENES IN COAD

Gene	CV-AUC	Dir.	Biological Role
CA7	1.00	↓	pH homeostasis, metabolism
ABCA8	1.00	↓	Lipid transport and efflux
CLEC3B	1.00	↓	ECM organisation; immunity
ADH1B	0.99	↓	Retinol/alcohol metabolism
CMTM5	0.99	↓	Tumour suppressor candidate
ADH1C	0.99	↓	Alcohol dehydrogenase
CHAT	0.99	↓	Cholinergic signalling
CA2	0.98	↓	Carbonate balance
CA1	0.98	↓	pH regulation
ABCC13	0.99	↓	ABC transporter, drug efflux
ADAMDEC1	0.97	↓	ECM remodelling; protease
DNASE1L3	0.99	↓	Immune tolerance

DPP6	0.96	↓	Ion channel modulation
CD177	1.00	↓	Neutrophil activation
CCR10	0.99	↓	Chemokine receptor
CDH3-AS1	0.85	↑	Cell adhesion; prognostic
CPZ	0.68	↓	Metallocarboxypeptidase
CCR8	0.73	↓	Treg trafficking; immune evasion
CNN2	0.74	↓	Cytoskeletal regulation

↓ downregulated in tumour; ↑ upregulated in tumour

E. Gene Expression Heatmap and Unsupervised Clustering

Hierarchical clustering of the top Boruta-confirmed genes across all TCGA-COAD samples (Fig. 8) reveals a clear separation between tumour and normal tissue based solely on gene expression patterns without using sample labels. Normal samples exhibit uniformly elevated expression across carbonic anhydrase and alcohol dehydrogenase gene families. Tumour samples demonstrate systematic suppression of these genes, reflecting metabolic and functional reprogramming of malignant cells.

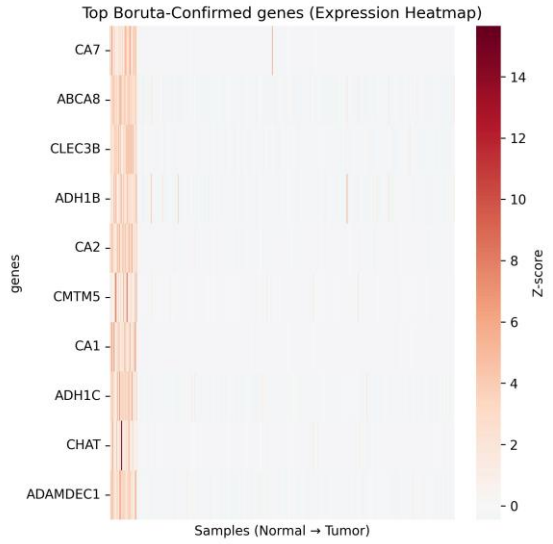

Fig. 8. Expression heatmap (Z-score normalised) of top Boruta-confirmed genes across TCGA-COAD samples (Normal → Tumour along x-axis). Distinct clustering confirms robust transcriptomic separation.

F. Survival Analysis and Prognostic Evaluation

Kaplan–Meier survival curves were generated for each of the 19 Boruta-confirmed genes. CDH3-AS1 demonstrated the most statistically significant association with overall survival (log-rank $p < 0.05$): patients with high CDH3-AS1 expression exhibited markedly shorter OS. This is biologically coherent given the established roles of CDH3 (P-cadherin) and its antisense transcript in cancer cell migration, invasion, and EMT. Carbonic anhydrase family members (CA2, CA7) and CMTM5 showed directional prognostic trends consistent with their tumour-suppressive roles.

G. Comparative Analysis

TABLE II COMPARATIVE PERFORMANCE OF REPRESENTATIVE COAD CLASSIFICATION METHODS

Method	Feature Selection	AUC	Survival
SVM [9]	DEG filter	~0.91	No
Random Forest [9]	DEG filter	~0.93	No
Lasso LR [13]	L1 penalty	~0.95	No
Deep Learning [9]	None	~0.97	No
Proposed (EN+Boruta)	Boruta+EN	0.994	Yes

Our pipeline achieves the highest AUC while simultaneously providing biological interpretability via Elastic Net coefficients, stability validation via Boruta shadow features, and clinical relevance via Kaplan–Meier survival analysis—advantages that individually oriented methods do not offer.

V. CONCLUSION

We have developed and validated a comprehensive, machine learning-driven computational pipeline for transcriptomic biomarker discovery in colon adenocarcinoma using TCGA-COAD RNA-Seq data. The pipeline integrates differential gene expression analysis, Boruta all-relevant feature selection, Elastic Net regularised classification, cross-validated ROC evaluation, hierarchical heatmap clustering, and Kaplan–Meier survival analysis within a unified, reproducible framework.

The key findings of this study are: (i) a 19-gene signature robustly discriminates COAD tumour from normal tissue with AUC = 0.994; (ii) three individual

genes—CA7, ABCA8, and CLEC3B—achieve perfect cross-validated discrimination (CV-AUC = 1.00); (iii) CDH3-AS1 is identified as a statistically significant prognostic biomarker ($p < 0.05$) with high expression associated with reduced overall survival; and (iv) carbonic anhydrase family members emerge as the most consistently important features, reflecting metabolic pH reprogramming as a central mechanism of COAD malignancy.

Future work will focus on: (i) validation in independent CRC cohorts (e.g., GEO datasets); (ii) integration of multi-omics data (genomics, proteomics, methylomics); (iii) incorporation of deep learning classifiers to capture non-linear gene interaction patterns; and (iv) development of a clinically deployable risk scoring tool based on the identified gene signature.

VI. ACKNOWLEDGMENT

The authors thank The Cancer Genome Atlas (TCGA) Program for providing openly accessible, high-quality RNA-Seq and clinical data. The authors also acknowledge the Department of Computer Science and Artificial Intelligence, Central University of Andhra Pradesh, for providing the computational and academic resources that supported this research.

REFERENCES

[1] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, “Cancer statistics, 2022,” CA: A Cancer Journal for Clinicians, vol. 72, no. 1, pp. 7–33, Jan. 2022.

[2] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” CA: A Cancer Journal for Clinicians, vol. 68, no. 6, pp. 394–424, Nov. 2018.

[3] R. Siegel, C. DeSantis, and A. Jemal, “Colorectal cancer statistics, 2014,” CA: A Cancer Journal for Clinicians, vol. 64, no. 2, pp. 104–117, Mar. 2014.

[4] E. R. Fearon and B. Vogelstein, “A genetic model for colorectal tumorigenesis,” Cell, vol. 61, no. 5, pp. 759–767, Jun. 1990.

[5] Cancer Genome Atlas Network, “Comprehensive molecular characterization of human colon and rectal cancer,” Nature, vol. 487, no. 7407, pp. 330–337, Jul. 2012.

[6] J. N. Weinstein et al., “The Cancer Genome Atlas Pan-Cancer analysis project,” Nature Genetics, vol. 45, no. 10, pp. 1113–1120, Sep. 2013.

[7] Z. Wang, M. Gerstein, and M. Snyder, “RNA-Seq: a revolutionary tool for transcriptomics,” Nature Reviews Genetics, vol. 10, no. 1, pp. 57–63, Jan. 2009.

[8] J. Guinney et al., “The consensus molecular subtypes of colorectal cancer,” Nature Medicine, vol. 21, no. 11, pp. 1350–1356, Nov. 2015.

[9] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[10] M. B. Kursa and W. R. Rudnicki, “Feature selection with the Boruta package,” Journal of Statistical Software, vol. 36, no. 11, pp. 1–13, 2010.

[11] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” Journal of the Royal Statistical Society: Series B, vol. 67, no. 2, pp. 301–320, Apr. 2005.

[12] J. Friedman, T. Hastie, and R. Tibshirani, “Regularization paths for generalized linear models via coordinate descent,” Journal of Statistical Software, vol. 33, no. 1, pp. 1–22, 2010.

[13] R. Tibshirani, “Regression shrinkage and selection via the Lasso,” Journal of the Royal Statistical Society: Series B, vol. 58, no. 1, pp. 267–288, Jan. 1996.

[14] E. L. Kaplan and P. Meier, “Nonparametric estimation from incomplete observations,” Journal of the American Statistical Association, vol. 53, no. 282, pp. 457–481, 1958.

[15] T. Fawcett, “An introduction to ROC analysis,” Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, Jun. 2006.