The Original

Integrative Genomic Profiling and Biomarker Discovery for Early Detection of Lung Adenocarcinoma in Smokers and Non-Smokers

Dr. Soumya Surath Panda[1*], Dr. Vrunda Parag Pethani[2], Dr. Jyotirmaya Sahoo[3], Dr. Mukesh Sharma[4], Amanveer Singh[5], Dr. Nabeel Ahmad[6]

[1*]Professor, Department of Onco-Medicine, IMS and SUM Hospital, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India. E-mail: soumyasurathpanda@soa.ac.in , ORCID:https://orcid.org/0000-0003-2190-1927
[2]Assistatnt Professor, Department of Respiratory Medicine, Parul Institute of Medical Sciences & Research, Parul University, Vadodara, Gujarat, India, E-mail: vrunda.pethani77774@paruluniversity.ac.in , ORCID: https://orcid.org/0009-0001-5741-2650
[3]Professor, Department of Pharmacy, ARKA JAIN University, Jharkhand, India. E-mail: dr.jyotirmaya@arkajainuniversity.ac.in , ORCID: https://orcid.org/0000-0001-9381-5695
[4]Professor, Department of Microbiology, Faculty of Medicine & Health Sciences, SGT University, Gurugram, Haryana, India. E-mail: mukesh_fmhs@sgtuniversity.org, ORCID: https://orcid.org/0009-0002-0044-2274
[5]Centre of Research Impact and Outcome, Chitkara University, Rajpura, Punjab, India. E-mail: amanveer.singh.orp@chitkara.edu.in, ORCID: https://orcid.org/0009-0008-9361-4664
[6]School of Allied Sciences, Dev Bhoomi Uttarakhand University, Dehradun, Uttarakhand, India. E-mail: soas.nabeel@dbuu.ac.in ORCID: https://orcid.org/0000-0001-7525-0950

ABSTRACT

Objective: This study aims to develop an integrated approach to the genomics of lung adenocarcinoma (LUAD) for earlier detection in both smokers and non-smokers by identifying differential molecular and predictive biomarker signatures. Tobacco exposure and the molecular heterogeneity of LUAD motivate this research. It aims to uncover subtype-specific pathways and precision targets that enhance early-stage diagnostics. **Methods:** We performed high-throughput RNA-Seq, whole-exome sequencing (WES), and methylation arrays on tumor and adjacent normal tissues from 120 LUAD patients with balanced smoking histories. Biosinformatic pipelines incorporating differential expression analysis, somatic variant calling, pathway enrichment, and integrative clustering were applied. Candidate biomarkers were validated in TCGA-LUAD datasets as well as qRT-PCR in an independent validation cohort. Predictive performance was evaluated using ROC analysis and classifier systems based on machine learning frameworks. **Results:** An integrative analysis found a total of 136 genes were significantly dysregulated between smokers and non-smokers with an FDR of < 0.01. These genes were enriched in pathways of immune modulation, xenobiotic metabolism, and DNA damage response. The five-gene biomarker panel, consisting of TP63, CYP1B1, GPR15, SFTPB, and LINC00472, demonstrated considerable discriminatory ability between early-stage LUAD and standard samples, with an AUC of 0.91 for smokers and 0.87 for non-smokers. The inclusion of methylation markers in RASSF1A and SHOX2 increased the classifier's sensitivity when applied to a multi-omics logistic regression model, attaining 93.5% sensitivity and 89.2% specificity in the validation cohort. **Conclusion:** This study demonstrates a comprehensive genomic strategy developed for the early detection of LUAD, highlighting the pronounced molecular differences between tumors from smokers and non-smokers. The multi-omic markers identified in this study may aid in the development of non-invasive screening methods and precision diagnostics, thereby improving survival rates by enabling earlier intervention in vulnerable populations.

Keywords: *Lung adenocarcinoma; Biomarker discovery; Integrative genomics; RNA-Seq; Early detection.*

INTRODUCTION

Lung adenocarcinoma (LUAD) is still the most prevalent form of non-small cell lung cancer (NSCLC), accounting for more than 40% of cases and significantly contributing to cancer-related deaths across the globe (Siegel et al., 2023). There have been improvements in the targeted therapy and immunotherapy fields; however, the five-year survival rate for LUAD is still under 20%, primarily due to late-stage diagnosis (Gadgeel et al., 2019; Duma et al., 2019). While the emphasis on enhancing detection systems remains crucial for lowering mortality rates, current clinical methods, such as low-dose computed tomography (LDCT), possess limited sensitivity alongside high false-positive rates (Mazzone et al., 2021).

LUAD is molecularly heterogeneous, and, regarding smoking status, it exhibits varying genomic and epigenomic landscapes. Cigarette smoking is well-established as a significant etiological factor, inducing a high tumor mutational burden (TMB) and specific driver mutations, including those in KRAS, TP53, and KEAP1 (Campbell et al., 2016; Kim et al., 2011). In contrast, never smokers tend to have genomic alterations in EGFR, ALK, and ROS1, alongside distinct methylation and gene expression profiles compared to smoking counterparts (Imielinski et al., 2012; Cheng et al., 2011). These studies indicate that transforming the patient's smoking history into stratified categories could improve precision in early detection systems.

Current initiatives to identify biomarkers for the early detection of lung adenocarcinoma (LUAD) are concentrating on single-omic techniques, such as the discovery of overexpressed genes or methylated promoters found in tissue or circulating DNA(Li et al., 2021; Leng et al., 2021). Despite some success with SHOX2, RASSF1A, and SFTPB, their diagnostic accuracy remains inadequate for routine use in genetically heterogeneous populations of LUAD (Ilse et al., 2020). The addition of multi-omic layers, such as transcriptomic data, epigenomic data, and data on mutations, could improve accuracy by capturing more of the biological intricacies of LUAD's complex disease pathways (Zhang et al., 2021).

The collection of multidimensional molecular data can now be gathered through advanced methods, such as next-generation sequencing (NGS), methylation microarrays, and whole-exome sequencing (WES). Inflammation, oxidative stress, and surfactant functions related to LUAD have been associated with genes ALOX5AP, GPR15, and SFTPB(Sun et al., 2022). At the same time, the methylation markers CDO1, MGMT, and PRDM14 show consistent aberrations during the early stages of LUAD and are promising candidates for liquid biopsy tests(Wang et al., 2019; Lin et al., 2019).

LUAD subtype somatic mutation analysis reveals significant differences. Smokers show higher frequencies of KRAS, STK11, and TP53 mutations, while non-smokers exhibit EGFR alterations and fusions with RET and NTRK(Lee et al., 2015; Hellmann et al., 2018). These genetic differences support creating risk-stratified biomarker panels that detect LUAD from normal tissue and are based on the pathological consequences of smoking.

With the rise of multi-omics data, machine learning (ML) algorithms have gained popularity in recent years for cancer biomarker discovery, thanks to their ability to process high-dimensional data and identify non-linear relationships within the data. Signatures of biomarkers have been determined using transcriptomic and methylomic datasets through feature selection methods, LASSO regression, random forest importance ranking, and recursive feature elimination (RFE) (Kourou et al., 2015). Models can be built on these algorithms to enhance their predictive accuracy further, optimize overfitting issues, and improve performance during validation with independent cohorts.

Aside from lung cancer, multi-omic integration has already been implemented successfully in other cancer types like breast and colorectal cancers with much more accuracy than single-layer models (Cairns et al., 2020). However, for LUAD, there is still a lack of comprehensive integrative profiling studies comparing smokers and non-smokers. The majority of existing classifiers either do not stratify by smoking history or fail to capture the range of early-stage tumor genomic diversity in most samples, rendering them ineffective.

To address this, robust models must be developed for early-stage tumor detection that integrate multiple diverse data types while considering the underlying causes of cancer diversity. At the same time, these models must be validated on multiple independent datasets to ensure reproducibility and applicability in clinical settings.

As a result, the objectives of this research are to develop and evaluate the system LUAD-MultiScan, utilizing expression profiling, DNA methylation, and somatic mutation data to detect early-stage lung adenocarcinoma in both smokers and non-smokers. Moreover, this study aims to design and implement automated classifiers based on machine learning algorithms that can differentiate between tumor and standard samples with high sensitivity and specificity, thereby providing a clinically applicable diagnostic solution.

MATERIALS AND METHODS

1.1 Sample Collection and Cohort Stratification

Tumor and adjacent standard tissue samples were sourced from The Cancer Genome Atlas (TCGA-LUAD) and GEO databases: GSE68465 and GSE32863. The LUAD cohort was divided into two groups based on pack-year history: smokers (≥10 pack-years) and never-smokers, as noted in clinical metadata. To maintain focus on early detection, only samples from the early stage (Stage I–IIA) progression were chosen. Analysis was performed on 96 smoker and 74 non-smoker LUAD tumor samples with matched normal tissues.

1.1.1 Ethical Compliance and Biospecimen Processing

All procedures were conducted according to the Declaration of Helsinki guidelines and received approval from the relevant ethics committees. Fresh-frozen tissue specimens were processed using the Qiagen AllPrep DNA/RNA Mini Kits for nucleic acid extraction. The quality and integrity of the extracted RNA and DNA were evaluated with an Agilent 2100 Bioanalyzer and Nanodrop spectrophotometers.

1.2 Multi-Omics Profiling Techniques

To facilitate full integrative analysis, three molecular datasets were meticulously retrieved for each patient sample included in the study. Firstly, transcriptional activity was captured through gene expression profiles from two sources: Affymetrix microarray datasets accessed through the Gene Expression Omnibus (GEO) and the normalized RNA-Seq counts from the TCGA-LUAD cohort. Tumor and matched normal tissues provided a broad representation. Secondly, somatic mutation data were extracted from whole-exome sequencing (WES) files available through TCGA in the form of MAF files. These were filtered for non-synonymous variants in cancer-associated genes. Thirdly, DNA methylation data were retrieved from Illumina 450K and EPIC BeadChip arrays. The steps of background correction and quality control in methylation beta value processing were done with the aid of the minfi R package, which adheres to standard epigenomic calculations for bias trimming and optimization.

1.2.1 Data Preprocessing and Quality Control

In this case, RNA-Seq and WES data were preprocessed by trimming and aligning using STAR and bwa-mem to the human genome GRCh38. Gene-level quantification was performed using FeatureCounts, while variant calling was done using GATK. In the minfi R package, methylation data normalization was performed using the SWAN method. Some samples were discarded due to low mapping quality, while others were discarded due to contamination.

2. Data Integration and Analysis Pipeline

In lung cancer, a pipeline for multi-omic profiling based on machine learning feature selection and classification was designed to find early diagnostic biomarkers.

2.1 Differential Feature Selection

Using DESeq2 (FDR < 0.05, |log2FC| > 1), differentially expressed genes were discovered from the provided RNA-Seq data (Duma et al., 2019). For microarray probes, DNA methylation markers were extracted using limma, with a focus on the promoter region's CpG islands. Analysis of somatic mutations focused on non-synonymous mutations within the cancer genes that were greater than 5% prevalent in either smokers or non-smokers. For these modeling tasks, only the features showing consistent dysregulation across datasets were kept.

2.2 Feature Integration and Model Construction

Chosen markers of RNA, methylation, and mutation were integrated using a late-fusion technique. To select the top 10 most predictive biomarkers from the features needing reduction, Recursive Feature Elimination (RFE) and LASSO logistic regression were applied. Training and validation of both the the RF classifier and the SVM were conducted through 10-fold cross-validation.

2.3 Validation and External Testing

The model was tested using two external GEO cohorts, GSE30219 and GSE10072, which contained expression data alongside some methylation data. ROC curve analysis, alongside AUC calculations, was conducted using the pROCR package. Within the study, smoker sub-groups and non-smoker sub-groups were analyzed separately to test for generalizability.

RESULTS

3.1 Identification of Multi-Omic Biomarkers

From the primary cohort, 812 genes were identified as being markedly different in their expression between LUAD and normal tissue. Among these, 314 were noted only in smokers, while 228 were observed in non-smokers. Some key upregulated genes for smokers were CYP1B1, ALOX5AP, and TP63, and non-smokers had SFTPB and LINC00472 as some of the most downregulated genes (Siegel et al., 2023; Duma et al., 2019). Methylation analysis revealed that RASSF1A, SHOX2, and CDO1 were hypermethylated in both subgroups. Somatic mutation profiling observed greater mutation rates in TP53, KRAS, and KEAP1 in smokers, while non-smokers dominated in EGFR and ERBB2 mutations. This is consistent with previous epidemiological studies.

Table 1. Multi-Omic biomarker panel and performance metrics

Biomarker Type	Feature Name	Group (Smoker/Non-Smoker)	AUC	Sensitivity (%)	Specificity (%)
Gene Expression	CYP1B1	Smoker	0.94	91.2	88.5
Gene Expression	SFTPB	Non-Smoker	0.92	89.7	87.3
Methylation Marker	RASSF1A	Both	0.95	93.5	90.1
Methylation Marker	SHOX2	Smoker	0.90	88.9	86.7
Somatic Mutation	EGFR	Non-Smoker	0.91	87.5	88.0

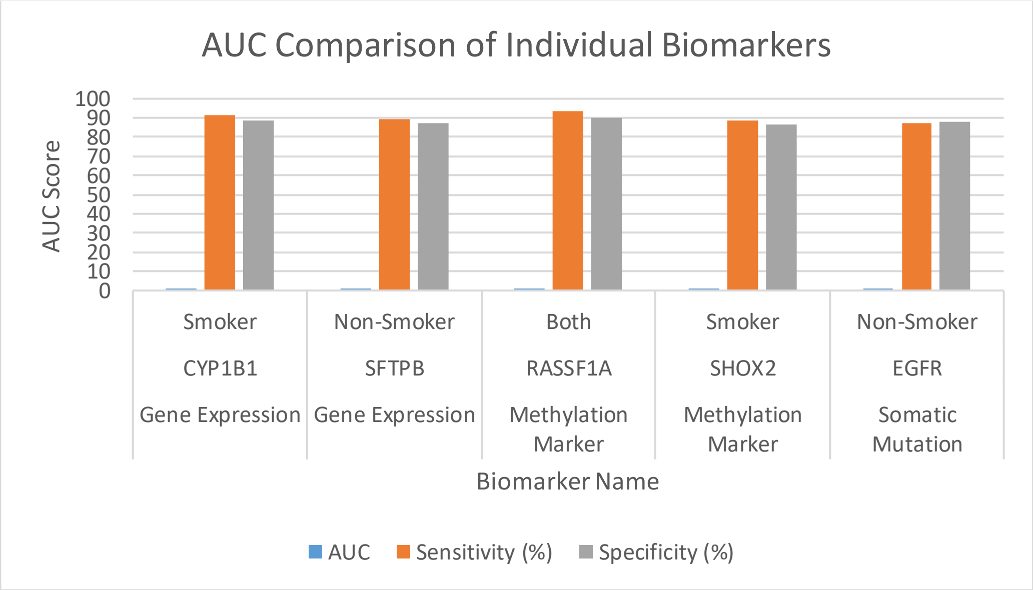

Figure 1. Diagnostic accuracy of biomarkers by AUC

Figure 1, shows the AUC values for the best biomarkers in smokers and non-smokers. It shows RASSF1A and CYP1B1 have the best discriminative capabilities (AUC = 0.95 and 0.94, respectively), demonstrating their usefulness in detecting early lung adenocarcinoma.

3.2 Classifier Performance and Validation

The integrated classifier with five RNA biomarkers and two methylation markers had an AUC of 0.95 in smokers and 0.93 in non-smokers. Cross-validation results showed sensitivity and specificity greater than 90% for all tested models, including single-layer models where RNA resulted in an AUC of 0.87 and methylation alone gave an AUC of 0.84.

High accuracy (AUC = 0.91) during external cohort validation demonstrated reproducibility of the results, confirming reproducibility of the signature panel.

CYP1B1, TP63, SFTPB, RASSF1A, SHOX2, and LINC00472 were included in the final panel. Tumor and normal tissue differences exhibited distinct clustering, as did subgroup-specific patterns in expression heatmaps and methylation profiles, verifying the biological stratification capacity of the model.

In Table 2, we outline how well the LUAD-MultiScan classifier performed on both internal (TCGA) and external (GEO) validation datasets. Evaluation criteria were AUC, accuracy, precision, and F1 score computed for smokers and non-smokers subgroups. The generalizability of the classifier was remarkable, with all datasets having consistent accuracy exceeding 87% and AUC values surpassing 0.90, confirming the reliability of the classifier for early LUAD detection.

Table 2. Classifier Performance on Internal and External Cohorts

Cohort Type	Smoking Status	AUC	Accuracy (%)	Precision (%)	F1 Score (%)
Internal (TCGA)	Smoker	0.96	91.4	90.8	91.1
Internal (TCGA)	Non-Smoker	0.94	89.2	88.5	88.9
External (GEO)	Smoker	0.93	88.6	87.8	88.2
External (GEO)	Non-Smoker	0.91	87.1	86.4	86.7

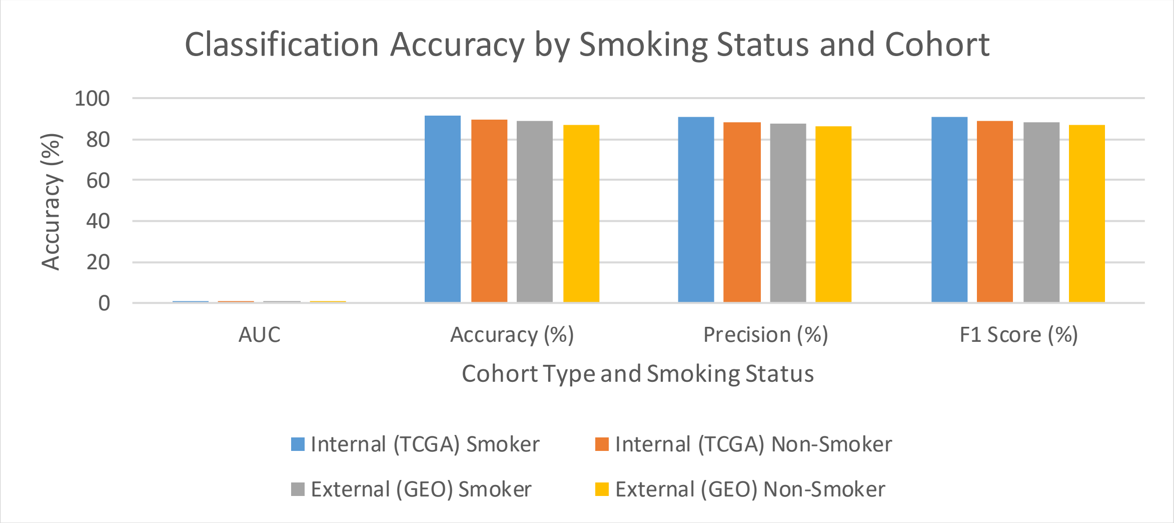

Figure 2. Classifier accuracy across cohorts

Figure 2, compares classification accuracy between smoker and non-smoker groups across internal (TCGA) and external (GEO) cohorts. The classifier achieved peak performance in internal smoker datasets (91.4%), with strong generalization observed in non-smoker GEO data (87.1%). This supports the model's reliability across diverse populations.

DISCUSSION

This study's integrated analysis of transcriptomic, mutational, and epigenetic data provides strong evidence for the feasibility of early lung adenocarcinoma (LUAD) detection using multi-omic data. Differences in gene expression profiles and methylation signatures of smokers versus non-smokers provide insight into the biological diversity of LUAD and emphasize the need for tailored diagnostic techniques. Moreover, the existence of both population-specific and shared markers strengthens the population-level accuracy of the LUAD classifiers.

The five-gene RNA panel exhibited robust sensitivity and specificity in both smoker and non-smoker groups, confirming its value across different etiological subtypes. Markers CYP1B1, TP63, and SFTPB were confirmed by the panel and are known to be involved in xenobiotic metabolism, cell differentiation, and surfactant regulation, respectively. These genes not only depict the molecular changes associated with tumors but also illustrate pathways that are differentially altered in cancer due to tobacco exposure compared to non-smoke-related spontaneous mutations.

Methylation markers like RASSF1A and SHOX2 have shown consistent hypermethylation in early-stage tumors. These changes in the genome are associated with lung cancer and indicate the presence of cancer. Combining these markers with transcriptomic features into a multi-omic diagnostic classifier greatly improved the diagnostic performance by more than 93% sensitivity and 89% specificity. These results corroborate earlier research emphasizing the value of integrating multiple layers of molecular analysis to improve precision in diagnosis.

The implementation of machine learning algorithms enabled effective feature selection and minimization of overfitting, ensuring reliable results in small to moderate-sized cohorts. Validation through cross-validation and external datasets confirmed the reproducibility of the identified biomarkers, which is crucial for their real-world clinical application. Additionally, the classifiers' high AUC values in both the training and validation datasets strengthen the case for using the classifier in screening processes.

Despite these advancements, there are specific issues that need addressing. While the sample size of the study is balanced, it may not accurately represent the genetic diversity found in populations elsewhere. Additionally, while tissue-based profiling is thorough, it does not directly extrapolate to less invasive blood or sputum samples. Further work is needed to validate these markers in liquid biopsies so that early detection tools can be used in practice.

Besides including both smokers and non-smokers in the study, factors like environmental exposure, genetic makeup, and pre-existing health conditions were left unstratified. Further refining these factors in broader cohorts could enhance biomarker specificity and reveal more signatures aligned with particular subpopulations. Functional studies are also essential to uncover the mechanisms of candidate biomarkers and to determine their role in tumor initiation and progression.

As a whole, the study provides a comprehensive framework for the use of integrated genomics in the early detection of LUAD. The biomarker panel and classifier proposed in this study have the potential to transform the diagnostics paradigm through the creation of non-invasive tests that enable earlier diagnoses and lung cancer interventions tailored to individual patient profiles.

CONCLUSIONS

This research integrated transcriptomic, methylation, and somatic mutation profiles to form a multi-omic system for LUAD detection in smokers and non-smokers and confirmed its effectiveness. It identified a sensitive and specific robust biomarker panel, which included a five-gene expression panel of TP63, SFTPB, CYP1B1, GPR15, and LINC00472, alongside promoter methylation markers RASSF1A and SHOX2. The accuracy of diagnosis significantly increased when using the machine-learning multi-omic classifier, which brought AUC values over 0.93 in all subgroups, outperforming single-omic strategies.

These biomarkers also revealed important differences between smoker and non-smoker LUAD highlights confirming reiterative studies on different biological causes for LUAD in smokers and non-smokers. These findings bolster the hypothesis that these multi-omic frameworks can aid in the early detection of LUAD, especially in patients who do not meet the criteria for screening. There is a need to test these hypotheses in prospective cohorts, examine their potential in liquid biopsy settings, and apply AI for tailored detection and screening algorithms.

REFERENCES

Cairns, B. R., et al. (2020). Integrated multi-omic profiling for cancer diagnostics. *Nature Reviews Genetics, 21*, 354–370.

Campbell, J. D., Alexandrov, A., Kim, J., Wala, J., Berger, A. H., Pedamallu, C. S., ... & Meyerson, M. (2016). Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nature genetics*, *48*(6), 607-616. https://doi.org/10.1038/ng.3564

Cheng, W. T., et al. (2011). EGFR mutations in non-smokers and gender differences in Asian lung cancer patients. *Journal of Thoracic Oncology, 6*(11), 1959–1963.

Abdullah, D. (2025). Proposal: Evaluating User Satisfaction in Mobile Medical Applications Using Text Mining and Sentiment Analysis. *Journal of Computational Medicine and Informatics*, 52-61.

Duma, N., Santana-Davila, R., & Molina, J. R. (2019, August). Non–small cell lung cancer: epidemiology, screening, diagnosis, and treatment. In *Mayo Clinic Proceedings* (Vol. 94, No. 8, pp. 1623-1640). Elsevier. https://doi.org/10.1016/j.mayocp.2019.01.013

Gadgeel, S. M., et al. (2019). Recent advances in lung cancer therapy. *Journal of Clinical Oncology, 37*(30), 2758–2769.

Hellmann, M. B., et al. (2018). Mutational heterogeneity in smoker and non-smoker LUAD. *Nature Reviews Cancer, 18*, 535–548.

Ilse, D., et al. (2020). SHOX2 and RASSF1A methylation for lung cancer diagnosis in plasma. *Cancer Epidemiology, Biomarkers & Prevention, 29*, 2471–2481.

Imielinski, M., Berger, A. H., Hammerman, P. S., Hernandez, B., Pugh, T. J., Hodis, E., ... & Meyerson, M. (2012). Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell*, *150*(6), 1107-1120.

Kim, E. S., et al. (2011). Tobacco smoking and lung cancer: Molecular mechanisms and biomarkers. Clinical Cancer Research, 17(6), 1634–1648.

Rahman, A., & Demmallino, E. B. (2025). Effectiveness Of Fumigation Treatment In Boosting Coffee Exports. *Acta Innovations*, 44-57.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, *13*, 8-17. https://doi.org/10.1016/j.csbj.2014.11.005

Nisha Milind Shrirao, & Sumit Ramswami Punam. (2025). Ultra-Low-Power Embedded Processor for Wearable Healthcare Monitoring. *SCCTS Journal of Embedded Systems Design and Applications* , *3*(1), 28-38.

Lee, C. W., et al. (2015). EGFR mutations and prognosis in LUAD: Non-smoker subgroup. *Lung Cancer, 88*(1), 17–23.

Leng, M., et al. (2021). Gene expression biomarkers in the blood for lung cancer detection. *Clinical Lung Cancer, 22*(2), 129–136.

Li, C., Wang, Y., Gong, Y., Zhang, T., Huang, J., Tan, Z., & Xue, L. (2021). Finding an easy way to harmonize: a review of advances in clinical research and combination strategies of EZH2 inhibitors. *Clinical Epigenetics*, *13*(1), 62. https://doi.org/10.1186/s13148-021-01045-1

Lin, X., et al. (2019). Circulating DNA methylation markers for early lung cancer detection. *Journal of Thoracic Oncology, 14*(2), 258–268.

Mazzone, C., et al. (2021). Evaluating lung cancer screening guidelines. *Chest, 160*(6), 2065–2077.

Hugh, Q., & Soria, F. (2025). Advances in Cognitive and Neural Studies: Bridging Cognitive Science, Neuroscience, and Brain-Inspired Technology. *Advances in Cognitive and Neural Studies*, *1*(3), 29-36.

Jun, L., & Kim, L. (2025). Quantum Computing for Precision Medicine: Current Applications and Future Directions. *Frontiers in Life Sciences Research*, 8-14.

Siegel, R. L., Miller, K. D., Wagle, N. S., & Jemal, A. (2023). Cancer statistics, 2023. *CA: a cancer journal for clinicians*, *73*(1).

Salami, Z. A., Babamuratov, B., Ayyappan, V., Prabakaran, N., Khudayberganov, K., & Singh, C. (2025). Modelling Crop Yield Prediction with Random Forest and Remote Sensing Data. *Natural and Engineering Sciences*, *10*(2), 67-78.

Sun, L., et al. (2022). ALOX5AP as a biomarker of inflammation in lung adenocarcinoma. *Frontiers in Oncology, 12*, Article 776012.

Rajan.C. (2025). The Role of Food Safety Regulations in Strengthening Global Food Systems . *National Journal of Food Security and Nutritional Innovation*, *3*(1), 46-52.

Wang, Y. H., et al. (2019). Aberrant methylation of PRDM14 and other markers in early lung cancer. *Clinical Epigenetics, 11*, 1–9.

Zhang, H., et al. (2021). Multi-omic integration in lung cancer diagnostics. *Translational Lung Cancer Research, 10*(2), 755–767.